ROBUST AUTOMATED TOPIC IDENTIFICATION

by

Chin-Yew Lin

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Computer Engineering)

August 1997

Acknowledgements

My greatest thanks go to Eduard Hovy. His insight, vision, and enthusiasm helped make my research a possible venture. His constant support and encouragement keep my faith and interest high in pursuing a research career.

I thank Dan Moldovan for introducing me into the field of natural language processing and I also thank him and the other dissertation committee member, Jean-Luc Gaudiot, for helping me graduate.

My thanks also go to Information Sciences Institute (ISI) for providing a state-of-the-art research environment, and the members of the Natural Language Processing group at ISI who aided and enlightened me during the writing of this dissertation.

To my family back in Taiwan: my parents, brother, sister, and parent-in-laws, many thanks for their physical and mental support.

My sincerest gratitude goes to my wife, Jau-Ching, who more than anyone else, made this all possible with her understanding, patience, and love.

Contents

List Of Tables

List Of Figures

Abstract

As the amount of on-line text keeps growing, it becomes increasing difficult for humans to process the deluge of information in the time available. We need automatic text processing systems to help us scan through huge volume of texts, route them to relevant parties, filter them into prespecified categories, or even summarize them. To achieve this, one crucial step is to identify the major topics of the texts, since summarization, text routing, etc., centrally require knowing the topics. In this research, we investigated several topic identification methods and developed three major results:

(1) We extended existing word-based frequency counting methods to form a new concept-based frequency method based on the assumption 'the more a concept is mentioned in a text, the more important it is.' We used the knowledge base WordNet to generalize words into concepts and showed how to select concepts of the appropriate degree of generalization.

(2) We studied patterns of word co-occurrence (*topic signatures*) consisting of sets of keywords that uniquely identify the topics of interest. We showed how to acquire keywords from texts pre-classified for each topic, using the $tf * idf$ measure. We also demonstrated how to identify topics using topic signatures, introduced confusion sets and multi-level topic signatures, and discussed the problems associated with multiple topics in a text.

(3) We described, implemented, and evaluated a method to learn the *Optimal Position Policy (OPP)* for finding topic-rich sentences in texts. This work is based on the *Position Hypothesis*: in genres with fixed discourse structure, the (ordinal) position of a sentence is related to its importance in a text. We showed how to verify the Position Hypothesis using topic keywords, empirically identify important sentence positions in a genre or domain, and quantitatively evaluate the results with various measures.

This work will eventually form part of an automated text summarization system.

Index Terms - Topic Identification, Text Categorization, Text Classification, Information Retrieval, Information Extraction, Information Filtering, Information Routing, and Natural language processing.

Chapter 1

Introduction

1.1 Need for Topic Identification

As more and more online information services become available, there is an increasing interest in digesting the information they provide. But because the amount of data is so overwhelming, it is simply not possible to rely solely on humans to process all the information. Automatic text processing is an obvious solution to the information overload problem. The Message Understanding Conference (MUC) [13] sponsored by DARPA, and the Text REtrieval Conference (TREC) [60] co-sponsored by NIST and DARPA, have spurred great interest in automatic text processing studies among academic and private research groups.

Automatic text processing techniques, such as automatic text routing and summarization, can help people scan through huge volume of texts, classify them into different categories, route them to relevant parties, and summarize them. To achieve this, one central step is to identify the major topics of the texts. The work in this thesis focuses on one very important text processing technique, *topic identification*.

1.2 What Is a Topic?

According to Brown and Yule ([8], 70) *topic* is the most frequently used, unexplained, term in the discourse analysis literature. For example, Hockett [32] used the term 'topic' as a grammatical constituent to describe sentence structure. We refer this usage of topic as *sentential topic*. To distinguish their notion of topic from sentential topic, Keenan and Schieffelin [39] introduced the term *discourse topic*. The idea

of discourse topic was further explained by Brown and Yule [8], and they defined discourse topic as 'what is being talked/written about'.

Although the notion of discourse topic is clear and intuitive, it is very difficult to pin down the formal procedure of identifying discourse topics. Consider the following sentences (from Schank [79]):

> A: John bought a red car in Baltimore yesterday.
> B: You mean he's not going to buy my car?
> C: John bought a car last year, didn't he?

B and C are both appropriate responses to A, but they seem to address different topics. The pair AB seems to be about "buying a car"; while AC seems to be more about "John's buying a car". This reflects Morgan's [58] idea that 'it is not sentences that have topics, but speakers'. To solve this problem, Schank suggested that instead of concentrating on defining what a topic is, it might be more fruitful to concentrate on the rules for shifting topics. He introduced the notion of *potential topics*, which consists of a subset of the initial conceptualization and a new conceptualization that comprises the new topic shift. For example, sentence A provides a set of initial conceptualization: {*John*, *buy*, *car*, *Baltimore*, *yesterday*}. Sentence B keeps a subset from the initial conceptualization, namely, {*he(John)*, *buy*, *car*}, and a new conceptualization, namely, {*someone(not John)*, *buy*, *car(mine)*}. Brown and Yule proposed a similar idea called *topic framework*, which consists of the objects, events, and states described in the text, with the world knowledge that must be called upon to interpret the text. However, Brown and Yule did not provide an explaination of how the *topic framework* can be computed.

Faced with this dilemma, this thesis takes an empirical approach. We hope that by developing empirical methods to identify topics in texts we will enable others to arrive at a satisfactory a definition.

Our goal is to develop robust automated topic identification modules which can be used not only as a stand alone topic identification unit, but also in other text processing tasks such as text summarization [33], text categorization[1], information routing[2], and so on. Text summarization systems can use topic identification

[1] See Chapter 4.
[2] See Chapter 4.

techiques to select central ideas for summary; text routing systems can categorize texts according to their topics and route texts to the appropriate interested parties.

Ideally, to perform topic identification, we need to first parse a text into some syntactic, semantic, and discoursal representation. We must then analyze the relations among the concepts in the representation to determine which concepts are more central or more peripheral to the subject of discussion. In the end, the most central concepts should be selected as the topic of the text [11]. To implement this solution, we need at least a competent parser which can parse unrestricted sentences into well-formed syntactic and semantic representations, a discourse analyzer which can recognize entities, events, states, and their interrelations across sentences, and a knowledge base that provides all the background knowledge not explicitly mentioned in the text. The problem is that these tools are currently not available, and are unlikely to become available for a long time. Therefore, to perform the topic identification today, we need to look for simpler alternatives.

1.3 Previous Work

One promising alternatives is Automated Text Indexing. Automated indexing is an Information Retrieval task that bears much similarity with the topic identification task. It was proposed by Luhn [53] in 1957 in an attempt to use statistical methods to automatically assign subject index codes to documents. He assumed that the more a word occurs in a text, the more important it is. Therefore word frequency can be used to associate with each word a measure of its significance in the text. Since topic identification must discover what a text is about, the most significant words, as determined by Luhn's method, can be used as an approximation of the 'aboutness' we are looking for.

Luhn's idea was explored further by Edmundson [17] and later by the SMART project led by Salton [72], with successes in Information Retrieval (IR). The success of their frequency-based (statistics-based) methods has been demonstrated in the recent TREC [61] conferencen results, considering the complexity of questions being asked over a 2 gigabyte collection. A sample TREC-4 topic (question) is shown in Figure 1.1. The average performance of the 36 participants were about 50%

breakeven recall[3] and precision[4] in the routing task[5] and 35% breakeven recall and precision in the adhoc task[6]. More details are provided in Chapter 4.

<num> Number: 207
<desc> What are the prospects of the Quebec separatists achieving independence from the rest of Canada?

Figure 1.1: Sample TREC-4 Topic

Since statistics-based approaches have achieved moderate success in real world applications, and powerful parsers, robust discourse analyzers, and adequate background knowledge bases[7] are not yet available, we argue that statistics-based methods that incorporate available robust linguistic tools such as part-of-speech taggers are the best solution currently available for topic identification.

Although similar tasks such as Automated Text Indexing have been investigated by many researchers, topic identification has not been studied as a standalone task. The most related work was done by Hearst [29]. She used a set of topic categories derived from WordNet and an algorithm based on Yarowsky's [88] sense disambiguation algorithm to assign multiple main topic categories to texts. Her approach is similar to our concept generalization method. The major difference is that we perform concept counting and simple sense disambiguation directly on WordNet (see Chapter 3). Her work is summarized in Section 2.4.

1.4 This Thesis

In this thesis, we describe three statistics-based topic identification methods. They are summarized in the following sections.

PROBLEM DRINKING:
Talking it over with a counselor apparently helps people who drink too much but aren't alcoholics, says a study in today's *Journal of American Medical Association*. The study found that two 15-minute counseling sessions from specially trained doctors helped at risk male drinkers cut alcoholic consumption by 14% after a year, at risk women cut back 31%.

Figure 1.2: Sample text from *Lifeline*, *USA Today*, April 2, 1997.

1.4.1 Concept Generalization

Figure 1.2 shows a short text from *USA Today*. As indicated in the heading, it is about *problem drinking*. If we count the frequency of each stemmed word[8], we see that terms *counsel**[9], *drink**, *alcoholic**, *study*, and *risk* all occur twice in the text. Although these 'high' frequency terms show some aspects of what the text is about, a major entity, i.e., *people* (including *male* and *women*) is not treated as a significant term. If a concept taxonomy is used to indicate that *male* and *women* can be generalized as *people*, then the concept *people* has a total generalized frequency 3. This makes *people* the most significant concept. If we have a rich knowledge base, we may be able to generalize "*at risk male drinkers*" and "*at risk women*" into "*at risk drinking people*" which is even *better* than *people*! Therefore we have a better description about what the text is about. The addition of a concept taxonomy provides *external* information not available in the text.

Chapter 3 describes a new method to identify topics, using concept generalization over a concept taxonomy derived from WordNet [57] and the Penman Upper Model [2]. WordNet is a large hand-built lexical database developed at Princeton

[3]Recall is: (the number of correct answers)/(the number of all the possible answers).

[4]Precision is: (the number of correct answers)/(the number of all the answers supplied).

[5]In the routing task it is assumed that the same questions are always being asked, but that new data is being searched.

[6]In the adhoc task, it is assumed that new questions are being asked against a static set of data.

[7]Although large scale general purpose knowledge bases such as CYC [46] are available, no successful natural language or information retrieval applications based on them have been reported.

[8]Stemmed words are words conflated to some canonical forms. Words are treated as the same if they can be conflated into a common canonical form. For example, *counselor* and *counseling* are stemmed into *counsel*, *drinkers* is stemmed into *drink*, and *alcoholics* is stemmed into *alcoholic*.

[9]'*' means any ending.

University. It contains 120,400 word forms organized into 96,760 lexicalized concepts. The Penman Upper Model is a taxonomy of 250 very general abstractions of the objects, processes, and relations in the world, organized to support linguistic processing. This new method can automatically generate *interesting wavefronts* consisting of significant concepts at different level of generalization by setting the *branch ratio threshold* and the *starting depth*. We also demonstrate how this method can perform word sense disambiguation on the fly if enough evidence is present in a text. Evaluation of this method on a collection of 50 *BusinessWeek* articles shows promising results.

1.4.2 Topic Signatures

Clearly, concept generalization is not the final answer. If one can 'conflate' related concepts to their 'central' idea, one can arrive at very accurate topic identifiers. For example, one can derive the term *counsel** from the text in Figure 1.2 from two sources, *counselor* and *counseling sessions*. Although they are not exactly the same thing, they are related in the way that *counselor* is the person who provides *counseling sessions*. In fact, *counselor* should be also conflated with *doctor*, since they refer to the same thing in the text. Ideally, we would like knowledge that can help us generalize *counselor*, *counseling session*, and *doctor* to a concept such as *conseling*. Unfortunately, one obvious candidate for this knowledge, concept taxonomies, is unlikely to help in general. The necessary relations among *counselor*, *counseling session*, doctor, and many other *counseling-related things* are not likely to be provided in a concept taxonomy, and these relations may be domain dependent. This weakness is due to the *incompleteness* of a concept taxonomy.

> John and Bill wanted money. They bought ski-masks and guns and stole an old car from a neighbor. Wearing their ski-masks and waving their guns, the two entered the bank, and within minutes left the bank with several bags of $100 bills. They drove away happy, throwing away the ski-masks and guns in a sidewalk trash can. They were never caught.

Figure 1.3: Robbery story from Hovy and Lin [33].

Trying to avoid prestructured knowledge bases is also not helpful. If one falls back to pure input-based methods such as word counting, one encounters problems of incompleteness in other ways. For example, the text from Hovy and Lin [33] (Figure 1.3) demonstrates a major weakness of all frequency-based methods including word counting and concept counting. Simple word counting would indicate the text is about *ski-mask*. However, we know it is about *robbery*.

To resolve the weaknesses of counting methods in general, we need a way to infer *counseling* from (*counselor*, *counseling session*, *doctor*), and *robbery* from (*money*, *gun*, *ski-mask*, *bank*). In Chapter 4 we describe a new method to identify topics using *topic signatures*. Topic signatures represent topics such as *counseling* as sets of frequently co-occurring words. Each topic signature is trained automatically, constructed using a set of texts pre-categorized as representatives of the topic. The resulting signature is used to identify similar word co-occurrence patterns in new texts. The topic with the most similar word co-occurrence pattern found is assigned to the text. We have trained 32 topic signatures over 16,137 Wall Street Journal texts, and tested them on another 12,906 unseen Wall Street Journal texts. The recall and precision scores of test set are 80% and 73% respectively. This compares well on the average routing task result of TREC-4 participants, although TREC-4 has more complex topics and diverse corpora.

1.4.3 Position Method

Instead of considering words or concepts in the input, it is possible to exploit other kinds of knowledge, such as discourse structure. When the genre of the text exhibits fairly regular discourse structure, one can wonder whether topics tend to occur in certain ordinal or structural locations. Chapter 5 describes a method to automatically identify topic-rich positions in a text according to the *Position Hypothesis*. The Position Hypothesis states that the (ordinal) position of a sentence relates to its importance in a text.

Clearly the association of significance with sentence position is genre dependent. For example, technical papers typically have abstracts, while newspaper articles do not. On the other hand, the first sentence of the first paragraph in a newspaper

article typically contains the most important information. For example, the first sentence of the first paragraph of the *Problem Drinking* text in Figure 1.2 is:

> "Talking it over with a counselor apparently helps people who drink too much but aren't alcoholics, says a study in today's *Journal of American Medical Association*."

This sentence clearly states what the *Problem Drinking* text is about. Although people generally agree about the Position Hypothesis, they do not agree on which positions in a text bearing more information. No systematic studies that provide quantitative results are reported in the literature. Therefore, we have designed an automated genre independent procedure that addresses the following questions:

1. Is the Position Hypothesis applicable on appropriate genres?
2. If the Position Hypothesis is applicable to a specific genre, where are those important positions? Alternatively, what is the relative importance among sentence positions?
3. How can one select sentences from a text according to their relative importance in order to achieve maximal topic identification?
4. What is the quantitative performance of algorithms based on the Position Method? What is their lower bound? What is their upper bound?

We tested this procedure on a set of 13,000 Ziff-Davis news about computers, and discovered the following pattern: text titles always bear the most topical information, followed by the second paragraph, the third paragraph, the first paragraph, and so on. (In the Ziff-Davis collection, the first paragraph is not the most important one, since it usually announces the occurrence of "exciting new event".) We also found that the first sentence of a paragraph bears the most infomation among all sentence positions. We tested this procedure on another set of 2,907 Ziff-Davis texts and reached the same pattern. We also developed a method called the *Optimal Position Policy (OPP)* to guide the selection of sentences from a text considering the overall performance. Evaluation results show that selecting 10% of the sentences from texts according to the OPP achieved recall and precision scores of 35% and 38% respectively when we compared the 10%-sentence extract with manually prepared abstracts for each text in 13,000 Ziff-Davis texts.

1.5 Contributions of This Work

We presented three different methods to empirically identify topics of texts. Since topic is not a well defined concept in discoure analysis, our methods are valuable in that they may help describe various aspects of topics of texts. The contributions of this thesis can be summarized as follows:

Algorithms to Identify Topics

- We extended the idea of counting word to counting concept and showed the use of the *branch ratio threshold* and *starting depth* to select concepts of appropriate level of generalization in a concept taxonomy.
- The concept generalization algorithm achieved current performance without using any linguistic tools. It establishes a performance lower bound for future system.
- We defined topic signatures to capture word co-occurrence patterns. A topic signature consisting of keywords pertaining to a complex concept and provides a simple way to infer the complex concept from its keywords.
- We introduced confusion sets to represent closed related topics and presented multi-level topic signatures to further discriminate closely related topics.
- We used normalizd *idf* to utilize the extra information provided by the known number of texts per topic category.
- We addressed the problem of multiple topics in texts.
- We described an automated method of deriving Optimal Position Policy that utilizes discourse regularity existing in a specific domain using topic keywords.

Empirical, Quantitative Methods to Measure Performance

- We applied the Topic Signature Method in text categorization to evaluate the effectiveness of the topic signatures.
- We provided empirical validation for the Position Hypothesis.

- We quantitatively evaluated the Optimal Position Policy using precision, recall, and coverage scores to measure the performance of the method.

Uses in Other Systems

In this dissertaion, we not only provided a systematic study of topic identification as a stand alone task, we also emphasized that topic identification is one crucial step for different automated text processing task. We showed how topic identification can be used in other text processing tasks such as text categorization in particular. We also mentioned briefly how topic identification can used be in automated text summarization.

Chapter 2

Related Work

2.1 Introduction

In this chapter we provide an overview of the status of topic identification research. Although topic identification is a central step for many automatic text processing tasks, it has not, to date, been studied as a standalone subject. Most of the related work uses topic identification as part of a specific task, such as automatic document indexing, text classification, text categorization, text summarization, and information retrieval. The approaches taken in these various tasks can be summarized in three groups: statistical, knowledge-based, and hybrid.

The statistical approach infers topics of texts from term[1] frequency, term location, term co-ocurrence, etc., without using external knowledge bases such as machine readable dictionaries. The knowledge-based approach relies on a syntactic/semantic parser, knowledge bases such as scripts or machine readable dictionaries, etc., without using any corpus statistics. The hybrid approach combines the statistical and knowledge-based approaches in an attempt to take advantage of the strengths of both approaches and there by to improve the overall system performance.

We describe the related work briefly in the following sections. Some methods are further explained in their related chapters later in this dissertation.

[1] A term can be a word, stemmed word, phrase, or other token, defined by the specific application.

2.2 Statistical Approaches

Early Work: Luhn, Baxendale, and Edmundson

Luhn [53] suggested measuring the significance of a word by its frequency, under the assumption that a writer normally emphasizes an aspect of a subject by repeating certain words related ot it. He also observed that words of very high frequency are too common to be significant, therefore used a statistically determined cutoff frequency to eliminate these words. More about Luhn's work appears in Section 3.2.

At about the same time, Baxendale [3] explored three different methods to extract the essential content of texts. His methods selected terms according to discourse or syntax cues, and then ranked the resulting terms according to their frequency distribution. The first method selected words from *topic sentences*, which are either the first or last sentences of a paragraph, according to references on authoring techniques. We discuss the idea of selecting sentences from content-rich positions in texts from a more general point of view in Chapter 5. The second method simply deleted closed-class words and quantitative adjectives. This is similar to disregarding very high frequency words, as in Luhn's experiments. Recently, Yang and Wilbur [87] used corpus statistics to automatically generate domain specific insignificant words, and claimed better performance. In a similar spirit, we construct common word lists to eliminate words that are too common to carry important information in all our experiments discussed in this thesis. The third method selected prepositional phrases based on the assumption that phrases are likely to reflect the content of an article more closely than any other simple construction. Baxendale found that the index vocabulary extracted by these three methods was highly correlated, although the prepositional phrase selection method was favored because of its reduction of much of the less significant vocabulary and meaning specificity.

Edmundson [18] performed a series of experiments in automatic extracting to assess the effectiveness of four different word significance association methods, namely, *Cue* method, *Key* method, *Title* method, and *Location* method. The Cue method assumes that sentences including pragmatic words such as "significant", "impossible", and "hardly" carry important content. Pragmatic words were complied into a Cue dictionary on the basis of statistical data and refined by linguistic criteria.

The Key method is similar to Luhn's proposal that high frequency content words are significant. The Title method assumes that an author conceives the title as circumscribing the subject matter of the document. Therefore, words of the title and subtitles, etc., are important. The Location method assumes that: (1) certain headings such as "Introduction", "Purpose", and "Conclusion" are good indicators of locations of important contents; and (2) "topic sentences tend to occur very early or very late in a document and its paragraphs." According to his experiments, the performance of these four methods were Location, Cue, Title, and Key, ranked decreasingly; a combination of Location, Cue, and Title methods achieved the best result. The interesting thing is that combining the Key method with the other three methods degraded the performance.

Edmundson's pioneering work in evaluating different word significance association methods is very inspiring. The Concept Generalization method presented in Chapter 3 is an extension of Luhn's frequency method. It uses a concept taxonomy to provide concept generalization/specialization information. The Position method introduced in Chapter 5 develops methods for the empirical validation, automatic acquisition, and quantitative evaluation of Edmundson's Title and Location methods. Recently, the Title method has been used by Apté et al. [1] in text categorization and Nomoto and Matsumoto [62] in topic identification; while Paice [63] re-investigated the Cue method. We leave the Cue method for future exploration.

The Introduction of Statistical Rigor: Sparck Jones

Luhn's idea of significant words only considered within-document term frequency. It does not take within-collection term frequency distribution into account. Within-collection term frequency distribution is important, since terms appearing frequently across different documents within a collection do not provide discrimination power. Closed-class words are good examples. To utilize this observation, Sparck Jones [83] introduced a new word significance assignment scheme called *inverse document frequency* (*idf*) as follows:

$$idf = log(N/n) + 1$$

where N is the number of documents in the collection and n is the number of documents in which the word occurs. The *idf* is smallest, i.e., most insignificant,

for words occur in every document. Such words have no discrimination power over the collection. For words occuring only once in the entire collection, *idf* is maximal. Documents containing such words can be uniquely identified by the presence or absence of these words. Inverse document frequency is a very simple and useful term significance measure. It has been used in conjunction with Luhn's original idea of within-document term frequency and in much other Information Retrieval research [84, 74].

Information Retrieval: Salton et al.

Salton and Lesk's [73] experiments at the end of the 1960s showed that using within-document frequency term weighting provided better performance than not using it. Salton and Yang [77] furthered the idea of using within-document frequency by combining within-document word frequency tf and inverse document frequency idf into a new term weighting scheme, which we now call $tf * idf$. Using SMART [72], they showed significant performance improvement over within-document frequency alone in information retrieval tasks. At the end of the 1980s, Salton and Buckley [76] tried 287 different term weighting assignment methods, and reconfirmed that $tf * idf$ remains the best performer. We used the $tf * idf$ term weighting scheme in selecting topic keywords for constructing topic and document signatures. The results are very encouraging. Details are presented in Chapter 4.

Latent Semantic Indexing (LSI)

Although term weighting schemes such as tf, idf, and $tf * idf$ have been well developed and applied in many practical cases, it has been criticized in the following aspects [12, 56, 71, 35]:

Synonymy: One concept can be expressed by different words. For example, '*cycle*' and '*bicycle*' can both refer to some kind of vehicle [34]. The inability to aggregate the same concept expressed in different word forms handicaps the effectiveness of general topic identification.

Polysemy: One word can have several meanings. For example, '*cycle*' could mean '*life cycle*' or '*bicyle*'. Counting different senses of a word as one sense and

generalizing words under the wrong sense impairs the precision of a topic identification system.

To overcome these difficulties, Deerwester et al. [12] proposed using Latent Semantic Indexing (LSI), which applies singular value decomposition (SVD) to derive a set of uncorrelated factors to represent terms and documents. These factors represent common meaning components extracted from many different words and documents. Each term or document is represented by a vector of weights indicating its strength of association with these factors. No meaning interpretations are given to these factors. LSI has been applied to many domain such as information retrieval, information filtering, and topic spotting [86]. Although the results are modestly promising, the lack of direct intepretation of these factors limits its utility in the topic identification task.

Text Categorization

Text categorization is a text processing task that classifies natural language texts into predefined categories. It has received much attention recently because of the explosive growth of the Internet. Topic identication can be applied to text categorization if a mapping between topics and categories can be established. In Chapter 4, we use the text categorization task to evaluate the effectiveness of one topic identification method, the topic signature method, by assuming that each predefined categories is a topic in our training and test corpora.

Most text categorization methods use pre-categorized training corpora to learn the associations between terms and categories. For example, Lewis and Ringuette [47] compared a Bayesian classifier and a decision tree method. They used an information-theoretic measure to select a set of highly predictive words for each category. They showed that decision tree method achieved better performance and produced rules that were easier to interpret.

Fuhr et al. [19] introduced AIR/X, a rule-based multistage indexing system, which used a special category assignment method called the Darmstadt Indexing Approach (DIA). DIA divided the assignment process into two subtasks: (1) a description step and (2) a decision step. During the description step, the system simply

gathered all the possible relations from terms of a document to all the potential categories. In the decision step, the system then used a probabilistic version of an ID3 classification tree [68] to decide the most likely category for the document. One interesting aspect of the AIR/X system is that it included a second indexing phase that refined the indexing result by considering category interdependencies. This approach is very similar to our multi-level topic signature in Section 4.4.4, which tries to achieve better discriminatory power within closely related topic categories (a so-called *confusion set*).

The topic identification method introduced in Chapter 4 employs only very simple learning techniques. However, it achieves promising results. How to improve our current methods by applying decision trees, Bayesian classifiers, or probabilistic learning methods to our current method is an interesting direction for future research.

One major weakness of most text categorization methods is the requirement of pre-labeled training corpora. Although such corpora are easier to obtain than before, it is desirable to perform topic identification or text categorization without using prelabled corpora. The AutoClass, an unsupervisored classification system developed by Cheeseman [9] and his colleague at NASA Ames Reseach Center, provides a possible solution. We plan to investigate this alternative in the future.

2.3 Knowledge-based Approaches

Lehnert

Lehnert [44, 43] developed a theory called *plot units* for summarizing narratives. Her theory can be applied to identify topics concerning so-called *affect states*, if a plot unit parser is available. Figure 2.1 shows some examples of primitive plot units (A–F) and complex plot units (G–H). Affect states roughly categorize human emotional reactions and states of desire into three states:

positive event (+): events that please

negative event (-): events that displease

mental state (M): mental states with neutral affect

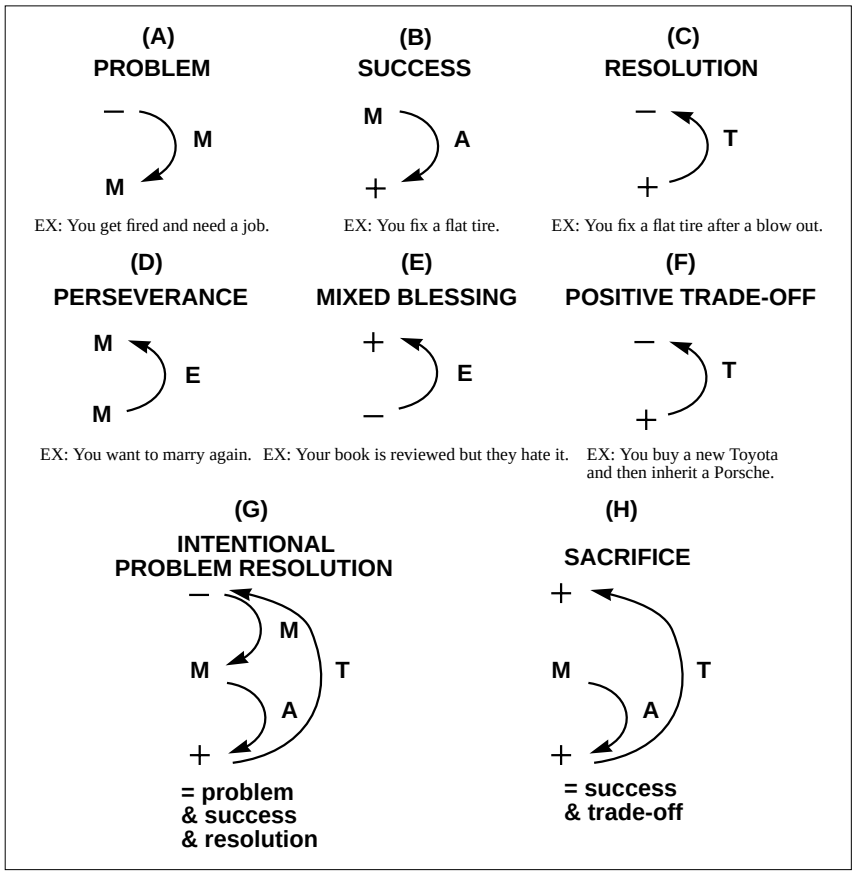


Figure 2.1: Examples of primitive and complex plot units.

Throughout a story each character will be associated with several affect states. These affect states link to each other through four *causal links — motivation (M), actualization (A), termination (T), and equivalent (E)*, which are assigned as the story is progressing. The causal links are defined as following [44]:

motivation: describe causalities behind mental states

actualization: describe intentionalities behind events

termination: describe the affective impact of an event is displaced

equivalent: when multiple perspectives of a single affect state can be separated

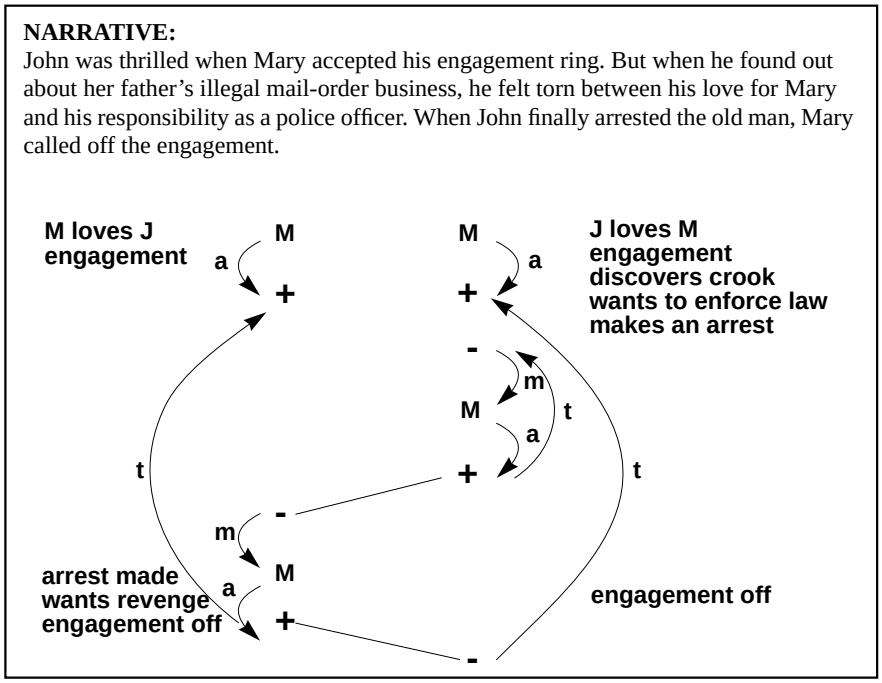


Figure 2.2: A configuration of Plot Units for a short narrative [44] .

Figure 2.2 demonstrates the application of plot units to a short narrative. Lehnert suggested using a predicative knowledge-based story understander and the knowledge structures given in Schank and Abelson [78] to recognize affect states first. After certain affect states are recognized, some *predictive* demons of primitive plot units start to recognize possible plot units.

Consider parsing the example in Figure 2.2. The parser has to infer from the fact *Mary accepted John's engagement ring* that "John loves Mary" and "Mary loves John", as shown. But other inferences are possible, such as "John wants Mary's money" or "Mary wants to use John to get Tom's attention." The parser has to perform many inferences for each proposition in the source text. This translates to slow processing speed. In general, the major weaknesses for plot units or similar systems are their brittleness (they can only apply to affect state and need all the necessary plot units to assure better coverage) and their knowledge-intensive reguirements (a powerful affect state parser is required, and even if built, it cannot be ported to

other domains easily). Although systems with the ability to parse plot units are desirable, they are not practical with today's technology.

A similar situation can also been found in Hahn's [25, 26] TOPIC semantic parser. TOPIC was designed for the summarization of texts on an indicative level. It only considered noun phrases in a text as possible topic candicates. TOPIC consists of three main components, i.e., the Parse Bulletin which keeps a record of the parsing process, the Domain Knowledge Base which contains the domain-specific knowledge needed dor the parse, and the Word Experts which dirves the parse through the text grammar modules they encapsulate. Among these components, the frame-based Domain Knowlege Base and the Word Experts require intensive knowledge engineering. For example, the Domain Knowledge Base contains entries such as "ZetaMachines Inc. is a manufacturer" and "Delta-X is a workstation" [2]. The evaluation of TOPIC has been performed on only 25 documents, although the author claimed the effectiveness of the TOPIC parser. We therefore argue that it is better to start with robust technology such as statistical techniques and then to add knowlege-intensive techniques later to improve system performance. This attitude is not uncommon today; in their recent research in information retrieval, Croft [10], Jacobs [36], and Liddy et al. [48] all share this view.

Riloff and Lehnert

Recently, Riloff and Lehnert [71] employed information extraction techniques in three text categorization algorithms. Their algorithms use *relevancy signatures*, each relevancy signature being pair consisting of a trigger word and a *concept node* that it triggers. Concept nodes are generated by a conceptual sentence analyzer called CIRCUS [45], which is based on a domain-specific dictionary of relevant information extracted from sentences. Concept nodes serve to capture the natural language context surrounding a word. Using relevancy signatures and their algorithms, Riloff and Lehnert achieved over 80% precision with up to 50% recall against baseline precisions of 69% and 55% on two test sets of 100 documents each. Since they claim that "a single relevant sentence is often enough to classify a text as relevant" and

[2]The actual frames for entries such as "ZetaMachines Inc." and "Delta-X" are more complex than the "is a" relation described here. Please refer to Hahn's [25] paper for more detail.

Sketchy Script for Demonstrations	
Event Id	*Predicated Event*
1	The demonstrators arrive at the demonstration location.
2	The demonstrators march.
3	Police arrive on the scene.
4	The demonstrators communicate with the target of the demonstration.
5	The demonstrators attack the target of the demonstration.
6	The police attack the demonstrators.
7	The police arrest the demonstrators.

Figure 2.3: FRUMP's sketchy script for demonstrations in English.

"once a relevant sentence is identified, the remainder of the text can be ignored", their algorithms are tuned to *single-topic* texts. The applicability of using these algorithms to multiple-topic texts has not been demonstrated. However, as we mention in Section 4.5.1, many of our texts contain multiple indices, and even single-indexed texts actually include multiple topics. This fact makes methods such as Riloff and Lehnert's somewhat less appealing than they might be.

DeJong

DeJong [14, 15] developed a system called FRUMP (Fast Reading Understanding and Memory Program), a newspaper skimming program developed at Yale University to skim and summarize news articles. FRUMP uses a data structure called a *sketchy script* to organize its world knowledge. Each sketchy script contains FRUMP's knowledge of what can occur in particular situations such as demonstrations, earthquakes, labor strikes, and so on. FRUMP selected a particular sketchy script based on lexical clues in a news article and then filled the empty slots of the script as FRUMP read the article. A summary was then generated based on what had been captured and filled in the template. Figure 2.3 shows a sketchy script for demonstrations [15] converted into English. (Internally, the events in the sketchy script are represented in conceptual dependency notation.) Figure 2.4 shows an example and summaries produced by FRUMP in three different languages [15].

TEXT:

BY FERNANDO DEL MUNDO MANILA, PHILIPPINES (UPI) - A bomb exploded aboard a Philippine Airline jetliner at 24,000 feet Friday but the only fatality was the bomber, who was sucked out a six-foot-wide hole blased in the wall of the plane's toilet.

The twin-engine British-built BAC-111 jet landed safely in Manila despite loss of pressurization. Three persons aboard the plane suffered minor injuries.

Officals said Rodolfo Salazar, an electrican from Cebu, 350 miles south of Manila, went into the toilet before the blast and was not among the 78 passengers and six crew members accounted for later.

"All circumstances point to the fact that he carried the bomb." and offical said.

Intelligence agents said the explosive may have been a sister banaag. The passengers were held for about four hours for questioning and released.

OUTPUT:

SELECTED SKETCHY SCRIPT $EXLOSION
CPU TIME FOR UNDERSTANDING = 8451 MILLISECONDS
ENGLISH SUMMARY:
A BOMB EXPLOSION IN A PHILIPPINES AIRLINES JET HAS KILLED THE PERSON WHO PLANTED THE BOMB AND INJURED 3 PEOPLE.
CHINESE SUMMARY:
I JIAH FEIHARNG PENNSHEHKEHJI SHANQ DE JAHDANN BAWJAH JAHSYYLE FANQJYH JAHDANN DE REN ERLCHIEE SHANQLE SAN GE REN.
SPANISH SUMMARY:
UNA EXPLOSION DE BOMBA DENTOR DE UN JET DE LA AEROLINIA FILIOINA HA MATADO AL BOMARDERO Y HA HERIDO A 3 PERSONAS.

Figure 2.4: An example *explosion* text and summaries generated by FRUMP.

To use FRUMP to identify complex topic such as demonstration, earthquake, etc., sketchy scripts must be built beforehand, pre-specifying exactly what is interesting or important for each different topic. Systems such as FRUMP cannot deal with input texts which are not included in the repository of the expected events; even when an event in the repository has been identified, the way the event is expressed in the text may not conform to the style used in the repository. In both cases the system will fail. This is also true to our topic signature method, but our topic signatures are acquired automatically by training on domain corpus, which is always better and more portable than the human engineered FRUMP scripts.

FRUMP is a precursor of most of the MUC systems today. Based on FRUMP, Mauldin [56] developed a information retrieval system called FERRET and extended the original system by incorporating an online machine readable dictionary. Although a machine readable dictionary can remedy the synonymy and polysemy problems, it still suffers from the coverage and knowlege-intensive problems that limit its applicability to other domains.

In light of such limitations, methods such as used by FRUMP, FERRET, and our topic signatures need to cooperate with other methods such as word frequency methods, the Concept Generalization method (Chapter 3), and the Position Method (Chapter 5), to ensure broad coverage and robustness.

CONSTRUE

Knowledge-based methods have also been applied to text categorization task. The best known example is the CONSTRUE system [28] developed at Carnegie Group for the the Reuters news service. CONSTRUE can quickly categorize economic and financial news stories into 674 categories, detect 17,000 company names, and route the categorized stories to interested parties. CONSTRUE is a rule-based text categorization system. Its accuracy is over 90% and its speed averages 5 second per message.

Although manually trained rules help CONSTRUE achieve such performance, to redevelop the rulebase of 674 categories takes about 8 person months. However, Apté et al. [1] demonstrated the use of machine learning techniques to acquire text categorization rules automatically and still maintain accuracy at above 75%. One

interesting result of Apté et al.'s experiments is that the system achieved the best performance when words in the headlines are assigned more credit. This is a good justification for our investigation of the Position Method in Chapter 5. Although we have not employed any sophisticated machine learning algorithms or invested much human effort in constructing topic signatures, the results of evaluating topic signatures in text categorization task are comparable to Apté et al.'s. Success stories of applying these techniques in the text categorization task warrant their study for the possible improvement of our current methods.

2.4 Hybrid Approaches

Liddy and Myaeng

DR-LINK, a document retrieval system developed by Liddy and Myaeng [49] at Syracuse University, used conceptal graphs[3] generated by a Conceptual Graph Generator to represent documents and queries, and a Relation-Concept Detector to identify relations among concepts.

DR-LINK works as follows. Sentences in a document are assigned discourse level component labels by a Text Structurer. Sample components are Main Event, Verbal Reaction, Expectation, Evaluation, Previous Event. The matching of queries and documents is carried out by a Conceptual Graph Matcher. The Generator, Detector, Structurer, and Matcher are all knowlege-intensive components: it takes time to perform their intended tasks. Therefore, it is wise to use some fast and reliable mechanisms to filter out clearly non-relevant documents first, and only pass more likely candiate documents to the knowledge-intensive components. Liddy and Myaeng used a Subject Field Coder to produce a subject-based vector representation for each sentence, paragraph, and document, consisting of the normalized frequencies of the Subject Field Codes (SFCs)[4] from the machine-readable version of *Longman's Dictionary of Contemporary English* (LDOCE) to produce an initial focusing on those documents which have real potential for matching a query. Each

[3]Concept graphs used in DR-LINK are a variation of semantic networks with features useful for Information Retrieval.

[4]The Subject Field Codes (SFC) is a classification scheme of 124 broad subject domains used by LDOCE [52].

content word in a text is first disambiguated using psycholinguistically justified sense disambiguation algorithms, then categorized using the SFCs assigned to the disambiguated word in LDOCE.

The SFC vectors used in DR-LINK are very similar to the combination of concept generalization and topic signatures discussed in Chapters 3 and 4 (see Lin [51] for details). However, the Subject Field Coder generalizes each content word using SFC, which is similar to concept generalization using a concept taxonomy within a flat hierarchy. Therefore, the problem of determine the appropriate generalization level, as discussed in Sections 3.4.2 and 3.4.3, does not exist. The Subject Field Coder then generates the vector representation for each intended text unit, which is similar to the construction of document signatures introduced in Section 4.4.1. Although no comprehensive evaluation of DR-LINK is available, we find the similarities between our topic identification approach and DR-LINK very encouraging.

Hearst

Hearst [30] developed an algorithm that automatically assigns multiple topic categories to texts, based on the posterior probability of the topic given its surrounding words (context). Her algorithm is a modification of Yarowsky's [88] sense disambiguation algorithm, which measures the likehood of a category given the terms that occur in the text. Topic categories were derived from WordNet [57], a hand-built thesaurus, automatically by merging concepts in the WordNet hierarchy.

Hearst's method resembles Liddy and Myaeng's Semantic Field Coder, the major difference being that Hearst used probabilistic estimation to detemine the most likely topic categories. Deriving concept categories automatically from existing thesauri is an interesting aspect of Hearst's work. In contrast, the concept generalization algorithm introduced in Chapter 3 uses a combined taxonomy from WordNet and the Penman Upper Model [2] without drastic rearrangement or modification of the original database. Since merging knowledge sources from different origins is common, as demonstrated in Chapter 3, techniques such as Hearst's merit further exploration.

2.5 Discourse Analysis

In the past 10 years, computationally inspired work in discourse analysis has achieved some promising results which can be applied to topic identification. Morris and Hirst [59] introduced a method for finding text structure of "being about the same thing" based on lexical conhesion. They called the manually built lists of related words *lexical chains*. Since lexical chains are text units relating to the same topic, it is natual to use these chains as a way to identify topics.

Passonneau and Litman [66] presented three algorithms using referential noun phrases[5], cue words[6], and pauses to determine whether empirically validated discourse segment boundaries of a test corpus correlate with these lingustic devices. Passoneau and Litman claimed that, although their algorithms did not perform as well as people, the results suggest human performance could be achieved with additional knowledge. This implies that their algorithms are weak methods to approach discourse topics, and an integrated system architecture such as the one described in this the thesis is a promising direction to pursue.

Recently, Marcu [55] described a comprehensive corpus analysis of cue phrases and develped three new algorithms that identified discourse usages of cue phrases, segmented sentences into clauses, and generated valid rhetorical structure trees.

2.6 Summary

In this chapter we described previous work related to topic identification. We described how statistical techniques that do not rely on knowledge-intensive resources and parsing are usually faster, more reliable, and more robust than knowledge-based methods. However, their lack of deep understanding results in lower accuracy of the systems. On the other hand, knowledge-based techniques normally require enormous human effort to build the necessary knowledge. Nonetheless, the effectiveness

[5]The identification of coreference entity relationship is already a standard task in the MUC-6 [13].

[6]Cue words referred here are different from the cue words used in Edmundson's experiments. Here cue words refer to discourse cue words [31].

of the invested human effort is demonstrated in the high accuracy over the intended domains.

How to combine the advantages of both approaches to achieve simultaneous high performance and cost-effectiveness is a major reseach topic [10, 36]. Our position on this problem is to use robust statistical techniques as much as possible, while adapting emerging robust linguistic tools such as Brill's [7] part-of-speech tagger and utilizing available knowledge sources such as WordNet and the Penman Upper Model, to build practically deployable systems and to identify the critical points where linguistic or natural language processing can make great contributions.

The recent exciting developments in statistical and discourse analysis techniques validate our pragmatic approach to solve the topic identification problem. See Figure 6.1 in Chapter 6 for an overview of the proposed integrated topic identification system.

Chapter 3

Using Frequency in Knowledge Based Topic Identification

3.1 Introduction

In this chapter we present a new method for automatically identifying the central ideas in a text based on a knowledge-based concept counting paradigm. Since the knowledge-based concept counting method is an extension of the traditional Information Retrieval word counting method, we first describe how word counting and the $tf * idf$ measure are used in Information Retrieval to identify keywords in texts. We then explain why word counting alone is not adequate; which makes concept counting and generalization necessary. We describe a method of performing concept counting and generalization using the symbols represented as synsets in the hierarchical concept taxonomy WordNet. By setting appropriate cutoff values for such parameters as concept generality and child-to-parent frequency ratio, we control the amount and level of generality of concepts extracted from the text. Three different sentence weight assignment methods based on the concept weights generated by the concept counting method are used to extract sentences from texts for evaluation. We conclude this chapter with a discussion of possible future work.

3.2 Word Frequency and Word Significance

Associating word frequency, i.e., word counting, with word significance was proposed in Luhn's pioneer work in automatic indexing [53] and extracting [54]. His proposal was based on the following assumptions:

1. writers often emphasize an aspect of a subject through the repetition of certain words,
2. writers usually use one sense of a word throughout a text,
3. only a limited number of words are available to express a particular concept, even though writers might choose different words for the same concept for stylistic reasons.

The first assumption enables one to use word frequency to estimate word significance without resorting to linguistic analysis (such as syntactic or semantic methods) that are expensive to implement and not robust enough even at today's scale of technology. The second assumption conforms to the recent finding of Gale et al. [21]: "one sense per discourse"; See Section 3.4.5.2 for a brief discussion of this issue. The third assumption can be addressed by using a thesaurus, which can be automatically acquired if enough sample texts are available [24].

Luhn also recognized that some high frequency words, such as the closed-class words *the*, *a*, *in*, *to*, were too common to be significant. He set up a high cutoff which filtered out high frequency common words, and a low cutoff which eliminated insignificant low frequency words. Words between these two cutoffs he considered as possessing "resolution power" (the ability of words to discriminate text contents). The two cutoffs were determined experimentally.

3.2.1 Inverse Document Frequency (idf) and $tf * idf$

Extending Luhn's idea of insignificant common words to a complete text collection instead of just one document, Sparck Jones [83] proposed a new word significance assignment scheme called *inverse document frequency* (idf) as follows:

$$idf = log(N/n) + 1$$

where N is the number of documents in the collection and n is the number of documents in which the word occurs. It is clear that the *idf* is smallest, i.e., most insignificant, for words that occur in every document, because such words have no discrimination power over the collection. On the other hand, *idf* is maximal for words that occur only once in the entire collection, since each single word can uniquely identify the document that contains it. Inverse document frequency reflects the fact that high frequency content words across documents of a collection are not significant. For example, in the legal domain, words such as *judge*, *lawyer*, and *attorney* are so common that they are not useful in discriminating one document from another in a collection of legal stories.

Although Sparck Jones showed that using *idf* was more effective in retrieval than not using it, *idf* alone does not incorporate the first assumption mentioned in the previous section. If its frequency within one document is abnormally high, a word should possess certain degree of significance for a document, even though the word may occur in every document in the collection. Salton and Yang [77] proposed a new word significance assignment method which considered both within-document word frequency and across collection word frequency, which is usually referred as $tf * idf$, i.e., the product of within-document *term frequency* times *inverse document frequency*. Since words can be canonicalized or stemmed into root forms and combined into phrases, people use *term* frequency instead of *word* frequency. Salton and Buckley [76] recently reconfirmed $tf * idf$ remains the best term significance assigment scheme among 287 different assignment methods.

3.2.2 Term Significance and Context

Fukumoto et al. [20] proposed a method that measures term significance according to *context*. In other words, they used the fact that term significance is context dependent. For example, given a hypothesized document collection shown in Figure 3.1, we want to decide whether a term is significant in discriminating topic category T_1, which consists of documents A_1 and A_2, from topic categories T_2, and T_3. Collection C_1, which encloses T_1, T_2, and T_3, we call the *enclosed* context, since T_1 is enclosed by C_1. Topic category T_1 we call the *enclosing* context, since T_1 encloses A_1 and A_2.

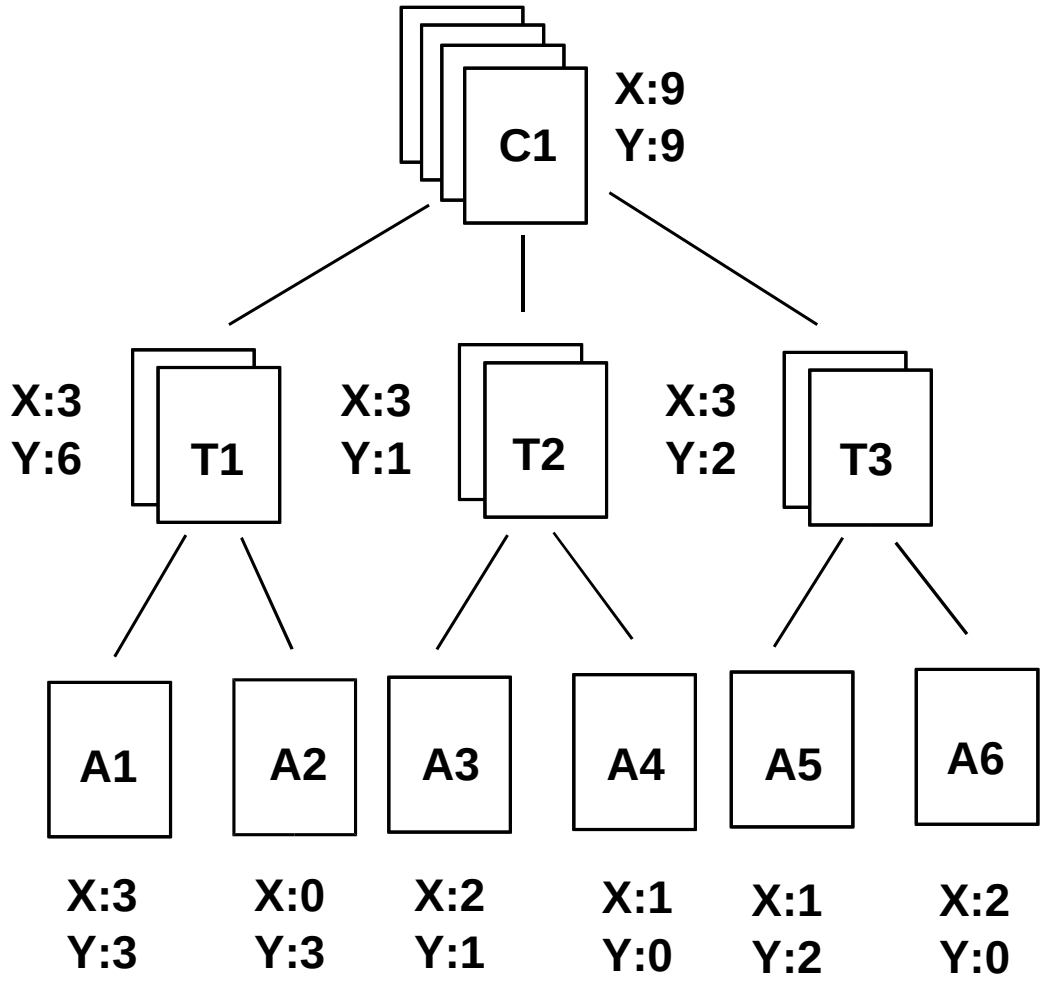


Figure 3.1: Distribution of two terms (X and Y) in a document collection (C_1), which is divided into topic categories (T_1, T_2, T_3), and each topic category is further divided into individual documents (A_1, A_2, ..., A_6). Numbers indicates term frequencies of X and Y at the specific context.

To determine whether term X is significant in context C_1, we have to compute the distributions[1] $D_{Xenclosed}$ and $D_{Xenclosing}$ of term X in its enclosed and enclosing contexts respectively. If $D_{Xenclosed}$ is an even distribution as shown in Figure 3.1 (X has frequency 3 across context C_1 which encloses T_1, T_2, and T_3), X is not significant in that context, since the presence of X does not increase with the likelihood of a text belonging to any one of these categories. On the other hand, Y is significant in context C_1 for identifying T_1, because the enclosed distribution $D_{Yenclosed}$ of Y is skewed (Y has high frequency in T_1 but not T_2 and T_3) and the enclosing distribution $D_{Yenclosing}$ of Y is even (Y has the same frequency across T_1). Notice that *idf* is equivalent to the method of deciding term significance within context C_1 when the topic category context is removed.

3.3 Problems with Word Counting

Information Retrieval (IR) researchers use word counting, cue word, location, and title-keyword techniques [64] to identify topics. Among these techniques, only word counting can be used robustly across different domains; the other techniques rely on stereotypical text structure or the functional structures of specific domains. In Chapter 5 we present a method called *Optimal Position Policy* to tackle the problem of genre- or domain-dependent nature of the location and title-keyword methods. In this section we focus on extending the word counting technique to overcome the following problems.

First, underlying the use of word frequency is the assumption that the more a word is used in a text, the more important it is in that text. This method recognizes only the literal word form and nothing else. Some morphological processing may help, but pronominalization and other forms of coreferentiality generally defeat simple word counting.

Second, lexical ambiguity of words undermines word counts. Should the frequency of "bank" really be 3 in "*I bank my money in the bank on the bank of the Mississippi*"?

[1] Fukumoto et al. used the χ^2 [85] method to compute term distribution.

Third, straightforward word counting can be misleading since it misses conceptual generalizations. For example: *"John bought some vegetables, fruit, bread, and milk."* What would be the topic of this sentence? We can draw no conclusion by using word counting method; where the topic actually should be: "*John bought some groceries.*" The problem is that the word counting method does not consider semantic relations among these words, i.e., *vegetables*, *fruit*, etc., relate to *groceries* at the deeper level of semantics. Similarly:

S 1 *Workplace homicides are rising.*

S 2 *A new cleaning-house on the grisly subject at the University of Oklahoma B-school finds that 1,000 workers were murdered on the job in 1992, vs. a steady 600 in the 1980s.*

Consider sentences **S** 1 and **S** 2. Checking the Collins COBUILD dictionary, we see *homicide* is defined as *the murder of one person by another.* Hence, sentences 1 and 2 are linked by the concept *murder*, which cannot be identified by just counting words; but can be captured by counting concepts.

Recognizing the inherent problem of the word counting method, researchers recently started using Artificial Intelligence techniques [37, 56] and statistical techniques [75, 24] to incorporate sementic relations among words into their applications. Following this trend, we have developed a new way to identify topics by counting *concepts* instead of words.

3.4 Concept Counting

In this section, we extend the word frequency method introduced in Section 3.2 to incorporate knowledge about relations among words recorded in knowledge bases. As mentioned above, a weakness of word counting is that we cannot capture the deeper relations among words: two words may have very different spelling but may be very close semantically. So we need to count not *word* frequency but *concept* frequency. However, what is a concept? According to the *Collins COBUILD Dictionary*, a concept is:

> "an idea or abstract principle which relates to a particular subject or to a particular view of that subject."

In *Webster's 9th New Collegiate Dictionary*, a concept is:

> "an abstract or generic idea generalized from particular instances."

From these definitions, we can say that a *concept* is a generalization of particular instances on the abstract level. For example, a *computer* is a machine that performs computation. In this sense, a pocket calculator is a particular instance of computer. Notebooks, laptops, and workstations are also computers. They could be generalized under one concept — *computer*. If we were to count concepts instead of words, pocket calculators, notebooks, laptops, and workstations would all refer to the concept *computer*.

Figure 3.2 shows a possible hierarchy for the concept *computer*. According to this concept hierarchy, if we find that *laptop*, *hand-held computer*, and *briefcase computer* are mentioned in a text, we can infer that the text is about *portable computers*, which is their parent concept. Moreover, if the text also mentions *workstation*, *mainframe*, and *minicomputer*, it is reasonable to say that the topic of the text is related to *digital computer*.

We still have not addressed any problem of anaphora: "it", "that one", etc. At this time, we simply assume that anaphora would not penalize the system performance too much, and leave the addition of anaphora resolution mechanism to future research. Even so it is interesting to note that using a hierarchy such as the one in Figure 3.2, a system can do a little bit of anaphora resolution. Since an author may refer to *Apple Computer Inc.* in his article as "the company" or "Apple", if we encode *Apple Computer Inc.* and *Apple* as a synonym set, and attach them under the concept *computer company*, this anaphora can be resolved through the relation chain: $\{Apple\ Computer\ Inc.,\ Apple\} \Rightarrow \{computer\ company\} \Rightarrow \{company\}$.

How can we implement concept counting? In the next section we discuss this question in detail.

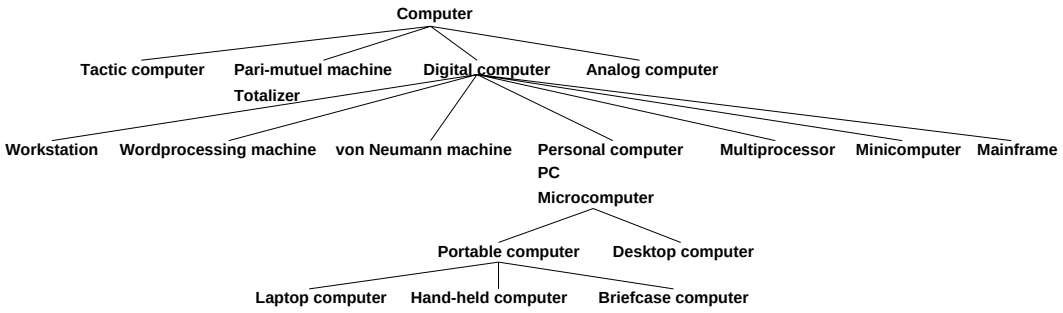


Figure 3.2: A sample hierarchy for *computer*.

3.4.1 The Power of Generalization

The power of the concept frequency method lies in *generalization*. For example, we may have the following three sentences in a text:

S 3 *IBM prices fall.*

S 4 *Machintosh prices fall.*

S 5 *Dell prices rise.*

The word "prices" appears 3 times; "fall" 2 times; "IBM", "Macintosh", "Dell", and "rise" once. But if we count concepts, the concept "PC" appears 3 times; "prices" 3 times; "change" 3 times. Thus these three sentences could be generalized as:

S 6 *PC prices change.*

Now a second problem presents itself: though we may find all the times *IBMs* , *Macs*, etc., are mentioned, we would also like to know that these two concepts can be combined into the concept *computer*; so that we can identify topic of texts using the generalization rather than the particular. To support this generalization, we need a hierarchical thesaurus. For example, assume words *apple*, *orange*, *pear*, *plum*, and *fruit* appear exactly once in the fruit text. Applying the word counting method, those specific instance of fruit will have score 1; while the generalization concept *fruit* will have score 5 — 4 out of 5 coming from the sum of the individual fruit

instances and plus 1 coming from *fruit* itself. Accordingly, this method will identify fruit as the topic of the text, which is not possible by using word counting method.

With such a knowledge base, we then can perform concept counting the right way. Currently, we use the Pangloss Ontology SENSUS [41], a combined knowledge base derived from the Penman Upper Model [2] and WordNet [57] . The example shown in Figures 3.2 is part of the WordNet hierarchy. More details about how we actually implemented our system based on the Penman Upper Model and WordNet appear in Sections 3.5.4.2 and 3.5.4.1 respectively. We discuss a few paramaters used in our knowledge-base concept counting algorithm in the following sections.

3.4.2 Branch Ratio Threshold R_t

Using a hierarchy, the question is now to find the most appropriate generalization. Clearly this is not always the leaf nodes (i.e., the words in the text) since this involves no generalization; in this case concept counting degenerates into word counting. Similarly, not the top node — everything is a *thing*! We need a method of identifying the most appropriate nodes somewhere in the middle of the taxonomy. In order to determine the most appropriate concept(s), we use the *branch ratio*, which we define as:

$$\textbf{Branch Ratio : R} = \frac{MAX(weight\ of\ all\ the\ immediate\ children)}{SUM(weight\ of\ all\ the\ immediate\ children)} \quad (3.1)$$

where *weight* is the number of times a concept is mentioned in the text (for a leaf node), or the sum of the weights of all immediate children (for non-leaf nodes).

It is obvious that the ratio is 1.0 if only one concept is mentioned in the source; while it is 0.0 for any concept not mentioned in the source. We found that the definition of ratio, **R**, is a way to identify the degree of generalization. The higher the ratio is, the less generalization power a parent node has over its immediate children. Why? Considering in Figure 3.3, the parent concept **Business** has weight 6 in case (a). This is the sum of its two immediate children, **Maker** and **Company**. Its ratio is 0.83 according to our definition. In case (b), the parent concept **Computer Company** has weight 10 and ratio 0.20. If we want to identify the topic based on the result in case (a), we should choose concept **Company** as the main idea instead

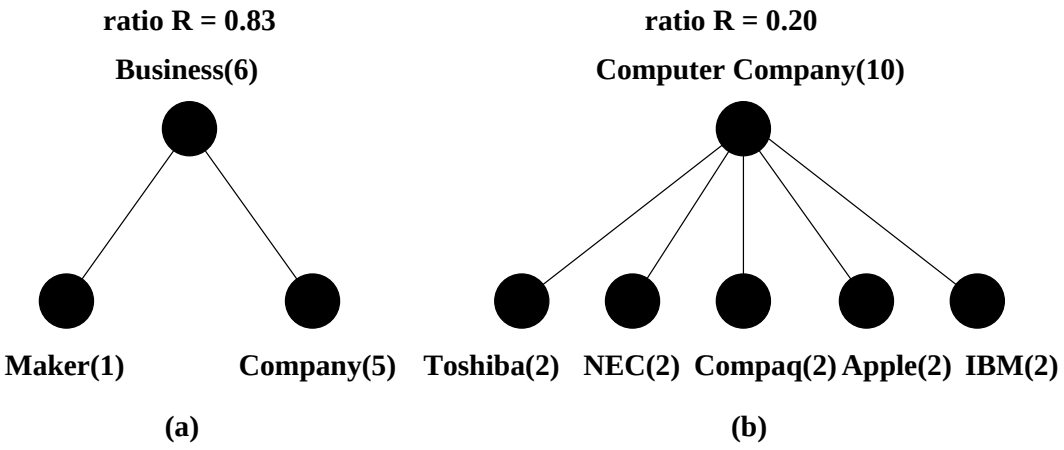


Figure 3.3: A demonstration of the degree of generalization.

of its parent concept **Business**, because the weight difference between **Business** and **Company** is small and **Company** is more specific than **Business**. Eighty three percent of the weight of the concept **Business** comes from its immediate child concept, **Company**.

In contrast, in case (b), we definitely should use the concept **Computer Company** as the topic concept. Its small ratio, 0.20, indicates that moving down to the children level to look for a topic will cause us to fall into too much detail. In fact, there is no clear child to specialize to!

According to this observation, we therefore define the *branch ratio threshold* (R_t) to serve as a cut-off point for the determination of interestingness, i.e., the degree of generalization. We define that an interesting node is a node with its ratio R less than R_t. R_t is experimentally determined, as discussed in Section 3.6. We use $R_t = 0.68$ for now.

Following the examples given in Figure 3.3, we next consider the problems of deciding whether a parent node is more "interesting" for a topic than its child(ren) node(s) in general. First we take the degenerate case in Figure 3.4: the parent concept has the same weight W_i as its child. In this case, we assume that the child is always more interesting for a summary than its parent because it is more specific. Next we consider the following cases. In Figure 3.5, we show several example cases with ratios 1.0, 0.5, 0.5, and 0.0 respectively. In case (a), the ratio 1.0 occurs because

Figure 3.4: A degenerate case shows general to specific relationship.

the weight of node $\mathbf{A_a}$ derives solely from one of its subconcepts, $\mathbf{C_a}$. According to our reasoning that concept $\mathbf{C_a}$ is more interesting than concept $\mathbf{A_a}$ because $\mathbf{C_a}$ is more specific than $\mathbf{A_a}$.

Cases (b), (c), and (d) relate to leaf nodes. Since leaf nodes do not have any children, we have to consider them separately. According to the definition of ratio, we cannot compute the ratio of a leaf node using Equation 3.1. To accommodate this special case, we assume that a *dummy* node is connected to each leaf node, and these dummy nodes carry the same weights as their parents do. In this scenario, each leaf node always has 1 as its ratio. The dummy node modification resolves the issue of computing the ratio at leaf nodes. We now consider another irregularity involving leaf nodes. In Figure 3.6, N_0 is the root concept of N_i, where N_0 has ratio R_0 and depth 0, and N_i has ratio R_i and depth i. R_i is equal to 1.0 because N_i is a leaf node. Assume R_j is greater or equal to R_t, where $0 \leq j < i$. Starting from the root node N_0, we need to decide which node is interesting enough to stop at, having ratio less than the branch ratio threshold R_t. It is clear that we will go down from N_0 to N_i and still not find an interesting node. How do we choose an interesting node in this case? We define that in this case the leaf node is interesting, though its ratio is not less than R_t.

Comparing case (b) and case (c) in Figure 3.5, we find that they have the same ratio, 0.5, and are both interesting by definition. Since for case (b) the absolute value of the sum of concept $\mathbf{A_b}$ is 54 and for case (c) concept $\mathbf{A_c}$ is 4, case (b) is

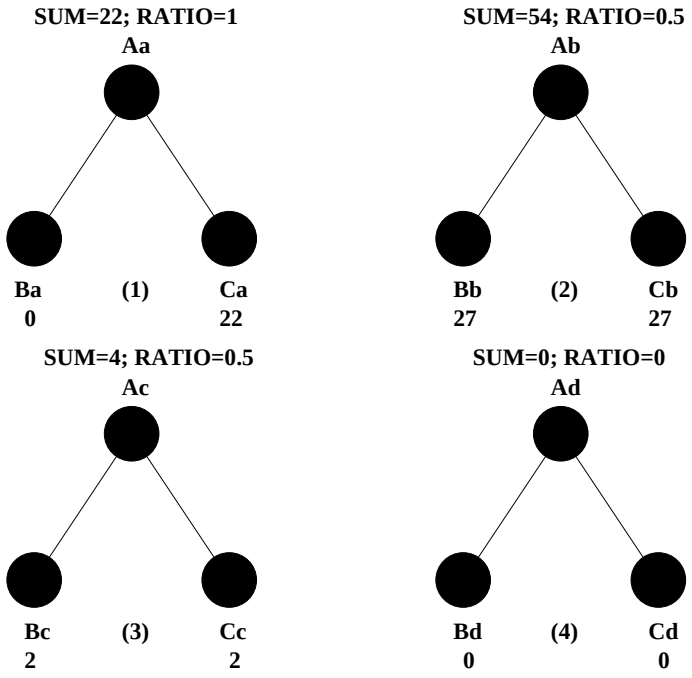


Figure 3.5: The relations between SUM and RATIO.

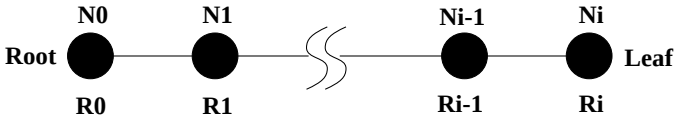


Figure 3.6: A leaf node may be an interesting node.

mentioned more frequently as is therefore more important than case (c). We need a measure to capture this situation. We rank all the "interesting nodes" according to their absolute weights. The larger the weight of a node is, the more interesting it is compared with the other "interesting nodes". In case (d), by definition it is not interesting; because the source does not mention concept $\mathbf{A_d}$ at all. We never consider any nodes with zero weight.

3.4.2.1 Selecting Interesting Concepts Using Branch Ratio Threshold

We have shown how to use ratio to determine the interestingness of a node. In this section, we demonstrate how to collect a list of interesting nodes from a given source by using the branch ratio threshold.

Assume we have two lists, *p_list* and *i_list*, empty at the beginning. The *p_list* stands for *probe list*, which contains all the interesting node candidates. The *i_list* is the interesting list and will contain all the *truly* interesting nodes at the end of processing. We assign weights to all the concept nodes referred to by the source text, and compute all the ratios as shown above. Placing the topmost node in the hierarchy into the *p_list*, we then perform the following, until the *p_list* is empty:

1. take a node n from the *p_list*
2. if the ratio of n is less than R_t or if it is a leaf node, then place n into *i_list*
3. if the ratio of n is greater or equal to R_t, then expand n and place all its immediate children into *p_list*

At the end of this procedure, we have an empty *p_list* and an interesting node list *i_list*. If we repeat this procedure by constructing a new *p_list* from all the direct children of *i_list*, we then produce another *i_list* which consists of interesting nodes with depth lower than the previous *i_list*. We call these generations of *i_lists* the *interesting wavefronts*. The algorithm describing the process of acquiring these interesting wavefronts is given in Figure 3.7. Each wavefront occurs at a progressively deeper level in the hierarchy. This depth is important, since we have to balance the level of specificly (the lower the better) with the level of generality (the higher the more compaction). In the next section we discuss the *depth* of a concept in the hierarchy.

3.4.3 Starting Depth D_s

Starting from the top of a hierarchy and proceeding downward along each child branch whenever the branch ratio is greater than or equal to $\mathcal{R}_t$, we will eventually stop with a list of interesting concepts, namely, the first *interesting wavefront*. Starting another exploration of interesting concepts downward from this interesting

```
 0: CollectInterestingWavefront(SOURCE,KB,RATIOTHRESHOLD)
 1: // Collect interesting wave front from semantic hierarchy
    // SOURCE source text
    // KB semantic knowledge base
    // RATIOTHRESHOLD decides if a node is interesting or not
 2: ActivateConcept(SOURCE,KB) // for each concept appears in the SOURCE,
                                       assign its weight equal to its frequecy
                                       in the SOURCE. Nodes in the KB with
                                       weight greater than 0 are active nodes
 3: PropagateWeight(SOURCE,KB) // propagate weights from the active nodes
                                       to all their ancestor in the KB
 4: ComputeRatio(KB) // compute ratio for all the node in the KB
 5: p_list = TopNode // probe list, initialize with the top most node in
                     // the hierarchical semantic knowledge base
 6: i_list = NULL    // interesting node list
 7: l_list = NULL    // leaf node list
 8: Children = NULL  // store direct children of an expanded node
 9: IF EMPTY(p_list) THEN GOTO 18
10:   GET a node n from p_list
11:   IF RATIO(n) >= RATIOTHRESHOLD OR n is a leaf node THEN GOTO 15
12:     PUT n into i_list
13:     IF n is not a leaf node THEN GOTO 9
14:     PUT n into l_list
15:   Children = EXPAND(n) // get all the direct children of n
16:   PUT Children into p_list
17:   GOTO 9
18: OUTPUT(UNIONOF(i_list,l_list)) // save current interesting wavefront
19: IF EMPTY(p_list = EXPAND(i_list)) THEN GOTO 24
20:   i_list = NULL // reset interesting node list
21:   p_list = UNIONOF(p_list,l_list) // merge with leaf node list
                                          // ensure full coverage
22:   l_list = NULL // reset leaf node list
23:   GOTO 10       // recursive collect interesing wave front
24: STOP
```

Figure 3.7: Algorithm for collecting interesting wavefronts.

wavefront results in a second, lower, wavefront, and so on. By repeating this process until we reach the leaf concepts of the hierarchy, we obtain a set of interesting wavefronts. Among these interesting wavefronts, which one is the most appropriate for generation of topics? Concepts higher in the hierarchy may be too general, while concepts lower in the hierarchy may be too specific.

In order to choose an adequate wavefront with an appropriate level of generalization, we introduce the parameter *starting depth,* $\mathcal{D}_s$. We require that the branch ratio criterion defined in the previous section can only take effect *after* the wavefront exceeds the starting depth; the first subsequent interesting wavefront generated will be our collection of topic concepts. The appropriate $\mathcal{D}_s$ is determined by experimenting with different values and choosing the best one.

The algorithm listed in Figure 3.8 includes both branch ratio and depth threshold criteria. The major difference between algorithms in Figure 3.7 and Figure 3.8 is that the former includes code (lines 9 to 12) to check if the depth threshold criterion is satisfied before testing the ratio threshold criterion, and it only outputs the first *i_list*.

3.4.4 An Example

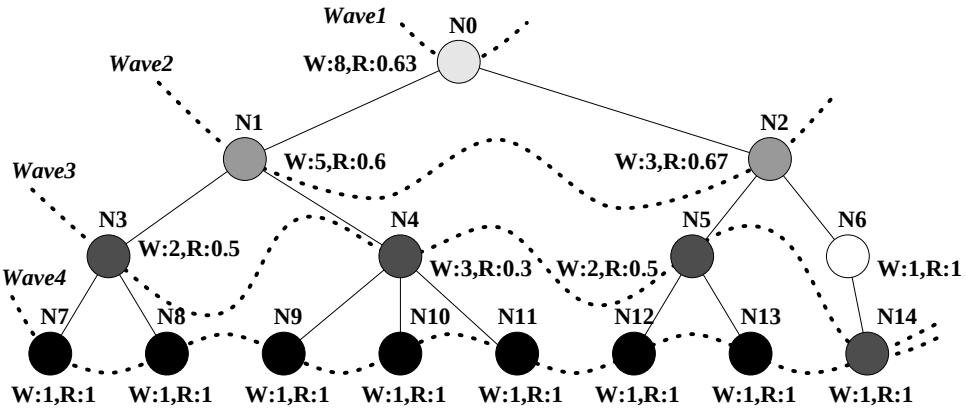
A hierarchical knowledge base provides a pyramidal view of the world. At the base of the structure are very specific concepts such as apple, car, pencil, and so on. The higher a concept is in the pyramidal structure, the more abstract it is. For example, starting from concept *apple* in the WordNet knowledge base, we follow the hypernym relation and move one level higher each time. We connect a path starting from the leaf concept, **apple**, to the root concept, **thing**: **apple** $\Rightarrow$ **fruit** $\Rightarrow$ **produce** $\Rightarrow$ **foodstuff** $\Rightarrow$ **food** $\Rightarrow$ **substance** $\Rightarrow$ **object** $\Rightarrow$ **entity** $\Rightarrow$ **thing**. We assume that this abstraction of world would agree with the general view of the world of various authors. When an author writes about the topic *produce*, he/she might very likely mention *fruit* and its sibling concept *vegetable*. If he/she wants to be more specific, he/she could mention apple, banana, pear, and melon under the concept *fruit*, and radish, pumpkin, mushroom, and celery under the concept *vegetable*. We can apply our algorithm to determine the interesting wavefront of these concepts. The result should be *produce*, or *vegetable* and *fruit*, depending on the text.

```
 0: CollectInterestingWavefrontWithStartingDepth
    (SOURCE,KB,RATIOTHRESHOLD,STARTINGDEPTH)
 1: // Collect interesting wave front from semantic hierarchy
    // SOURCE source text
    // KB semantic knowledge base
    // RATIOTHRESHOLD decides if a node is interesting or not
    // STARTINGDEPTH decides if ratio threshold criterion takes effect
 2: ActivateConcept(SOURCE,KB) // for each concept appears in the SOURCE,
                                  assign its weight equal to its frequecy
                                  in the SOURCE. Nodes in the KB with
                                  weight greater than 0 are active nodes
 3: PropagateWeight(SOURCE,KB) // propagate weights from the active nodes
                                  to all their ancestor in the KB
 4: ComputeRatio(KB) // compute ratio for all the active node in the KB
 5: p_list = TopNode // probe list, initialize with the top most node in
                     // the hierarchical semantic knowledge base
 6: i_list = NULL    // interesting node list
 7: Children = NULL  // store direct children of an expanded node
 8: IF EMPTY(p_list) THEN GOTO 20
 9:   GET a node n from p_list
10:   IF DEPTH(n) > STARTINGDEPTH THEN
11:     Children = EXPAND(n) // get all the direct children of n
12:     PUT Children into p_list
13:     GOTO 8
14:   IF RATIO(n) >= RATIOTHRESHOLD OR n is a leaf node THEN GOTO 17
15:     PUT n into i_list
16:     GOTO 8
17:   Children = EXPAND(n) // get all the direct children of n
18:   PUT Children into p_list
19:   GOTO 8
20: OUTPUT(i_list) as the interesting wavefront
21: STOP
```

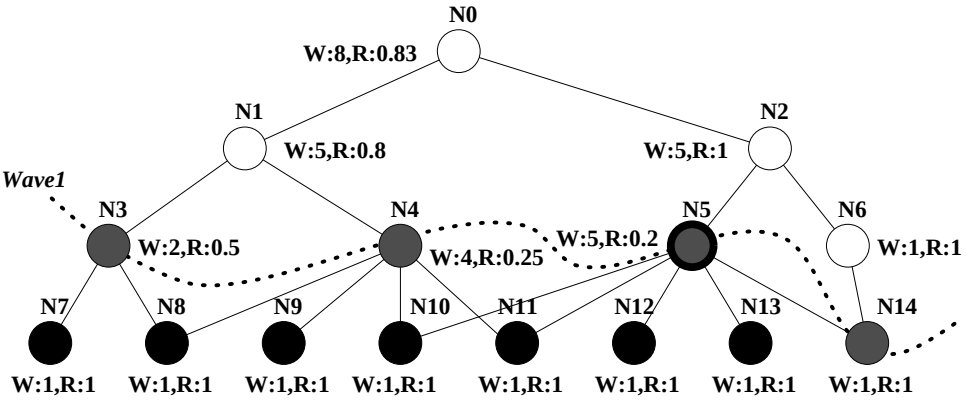
Figure 3.8: Algorithm for collecting interesting wavefronts after Starting Depth.

Figure 3.9 (a) shows a simplified example of the working of the concept counting algorithm. Assume each concept has at most one parent concept. Node N_0 is the root concept of the simple hierarchical knowledge base. Assume that each concept represented by leaf node appears exactly once in the source; hence the weight of each leaf node is one. The weight of their parents are the sum of their weight; for example, N_4 has weight three, N_1 has weight five, and N_0 has weight eight. In this specific case we just apply the simple concept counting technique. Figure 3.9 (a) also shows the interesting wavefronts obtained by applying the algorithm of Figure 3.7. Four interesting wavefronts are generated in this case. The lower the wavefront, the more specific the concepts comprising the wavefront, and the more concepts contained in the wavefront. To select an interesting wavefront for topics, presumably we would like concepts included in the wavefront to be as general as possible so that we can have more compact representation of topics. For example, if we choose *Wave1*, the only topic to mention is N_0; if instead we select *Wave3*, the topic will include N_3, N_4, N_5, and N_{14}.

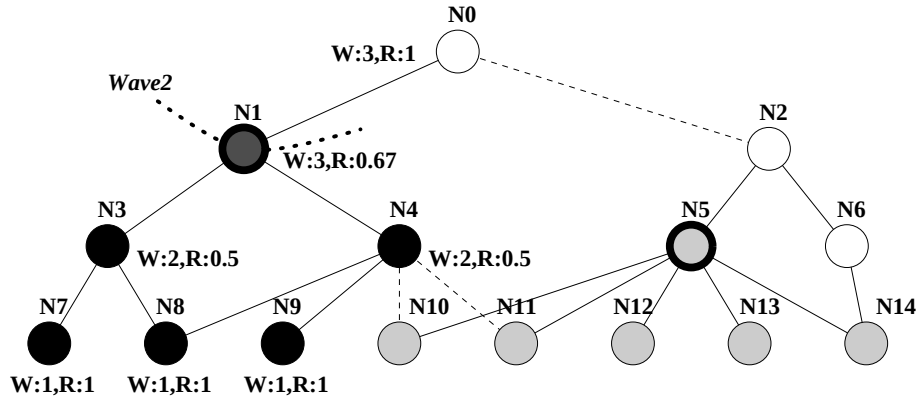
It is interesting to note that no matter which wavefront we select, it covers all the concepts presented in the source. We can verify that in Figure 3.9 (a) where the union of leaf concepts of all the subtrees rooted at the concepts sitting on any wavefront contains all the concepts used in the source; in this case N_7, N_8, N_9, N_{10}, N_{11}, N_{12}, N_{13}, and N_{14}. The dilemma is that the more general a concept is, the more abstract it is. Most of the times N_0 will be **thing**. Therefore if we follow the more general the better rule, we would always get something like *thing* or *object* as the main topic of a text. This kind of result is no better than nothing. A *general* and *specific* interesting wavefront is what we really want. It is a wavefront consisting of nodes located in the *middle* of a hierarchical knowledge base. No rule of thumb is available to decide where the golden middle is, but it justifies the need to introduce the starting depth parameter. We have carried out several experiments to determine an effective starting depth. Details of the experiments are covered in Section 3.6. In next section, using Figures 3.9 (b) and (c) we address the issue of syntactic and semantic ambiguity.



(a) Each concept has one and only one parent concept. W: weight, R: ratio.



(b) Concepts such as N8, N10, N11, and N14 are allowed to have multiple senses.



(c) Concept weight distribution after removing the most instereting node N5

Figure 3.9: Examples of single- and multiple-sense hierarchies when the interesting wavefront collection algorithm is applied.

Bank		
POS	*Hypernym*	*Example*
N	institution	... **Bank** of America
N	building	Turn left at the **bank**.
V	deal	He **banks** with Bank of America.
N	reserve	... access to **banks** of information.
N	slope	She scrambled up the **bank** to the road.
N	ridge	... **bank** of the Mississippi river.
N	row	There is a **bank** of switches.
V	amass	The storm had **banked** sand inside the lagoon.
V	tilt	The airplane turned, **banking** slightly.

Table 3.1: List of multiple syntactic categories and senses for **bank** from Collins COBUILD English Language Dictionary (POS: part of speech; N: noun, V: verb).

3.4.5 Syntactic and Semantic Ambiguity

In this section, we discuss two issues related to the syntatic and semantic ambiguities of words.

The concept counting algorithm shown in Figure 3.8 works well when every concept has only one parent concept, i.e., a single sense. However, a word may have several senses and belong to multiple syntactic categories. Table 3.1 lists some of the possible syntactic and semantic variations for **bank** listed in the Collins COBUILD English Language Dictionary [81]. Two different part of speech categories and nine different senses are shown for **bank**. What are the problems when we encounters such ambiguities? Can we use the concept counting algorithm in this case? The rest of this section is dedicated to answer these questions.

3.4.5.1 Multiple Contributions from a Single Concept

In this section, we illustrate the problem resulting from multiple weight contributions from a single concept. For example, in Figure 3.9 (b), N_8 is the hyponym of N_3 and N_4, N_{10} is the hyponym of N_4 and N_5, N_{11} is the hyponym of N_4 and N_5, and N_{14} is the hyponym of N_5 and N_6. How do we identify the interesting nodes in this example? It may seem that we can apply our interesting wavefront collection algorithm in Figure 3.7 along each sense of a leaf node, with results the same as before. However, this is not the case. Assume that each leaf node appears exactly

	Naive		*Improved*	
Node ID	*Weight*	*Ratio*	*Weight*	*Ratio*
N_0	**12**	0.5	8	0.83
N_1	**6**	0.67	5	0.8
N_2	**6**	0.83	5	1.0
N_3	2	0.5	2	0.5
N_4	4	0.25	4	0.25
N_5	5	0.2	5	0.2
N_6	1	0.1	1	1.0

Table 3.2: Weight and Ratio table for example in Figure 20 (b).

once in the source text. The weight of a parent node is the sum of the weights of its direct children. The result of the calculation for each inner node of the hierarchy is shown in the column labelled *naive* in Table 3.2. Multiple contributions of N_8, N_{10}, N_{11}, and N_{14} inflate the weights of N_0, N_1, and N_2. It is obvious that N_0 should have total weight 8 (sum of all the leaf nodes rooted at N_0), but the inflated weight of N_0 is 12. The difference lies in the ambiguity of a leaf concept that has multiple hypernyms and these hypernyms, which themselves have a common hypernym. The concept counting algorithm shown in Figure 3.8 double-counts the weight passed from the leaf concept at the common hypernym, because multiple paths connect the leaf concept to the common hypernym: N_1 in Figure 3.9 counts the weight from N_8 *twice*. Once from path $N_8 \Rightarrow N_3 \Rightarrow N_1$; once from path $N_8 \Rightarrow N_4 \Rightarrow N_1$. To avoid multiply counting of weight from a single concept, we use an improved weight computation method for the internal concept nodes as follows:

$$w_{internal\_concept_i} = \sum w_{leaf\_concept_{ij}} \tag{3.2}$$

where $w_{leaf\_concept_{ij}}$ is the weight of leaf concept j which has some path to internal concept i, i.e., the weight of a internal concept i is the sum of all the leaf concepts of the subtree rooted at i. The weight and ratio data in the column labelled *improved* in Table 3.2 shows the result without multiple counting. Figure 3.9 (b) shows the result using the improved method.

3.4.5.2 Mutually Related Concepts in a Wavefront

This section addresses the problem of mutually related concepts in an interesting wavefront. For example, the wavefront *Wave1* in Figure 3.9 (b) consists of concepts N_3, N_4, N_5, and N_{14}. Using the algorithm in Figure 3.8, we identify *Wave1* as the interesting wavefront. Therefore, the main points are concepts N_3, N_4, N_5, and N_{14}, where concepts N_3 and N_4 share concept N_8, concept N_5 shares concepts N_{10} and N_{11} with concept N_4, and concept N_{14} is by itself. *Wave1* consists of concepts which are mutually related. The mutual relation originates from concept ambiguity. For example, part of the weights of N_4 and N_5 are from the same concepts N_{10} and N_{11}. In order to represent the identified topics in the most compact form, we do not want to select both N_4 and N_5, since they are very likely referring to similar things. We could either use N_4 or N_5, but which one? Our solution is to select the one with the highest weight.

In this example, N_5 is selected (N_5 is emphasized by a bold border in Figure 3.9 (b)). After the concept with the highest weight in the wavefront is selected, we remove *all* the leaf concepts that contribute weights to that selected concept. We then recompute concept weights and repeat the wavefront algorithm in remainders. At the end, a series of topic concepts with decreasing importance is obtained. Figure 3.9 (c) shows the result of applying the algorithm after removing N_5. It marks N_1 as the topic concept the second time. For this example, the algorithm stops after removing N_1, because the subtree of N_1 covers all the remaining concepts. Removing all the leaf concepts rooted at the selected concept after each run of the algorithm is the same as selecting senses for the leaf concepts. The algorithm assigns a *preferred* sense for each leaf concept. This corresponds well with the observation of Gale et al. [21]:

> ... if a polysemous word such as *sentence* appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense ...

Our algorithm applies a similar principle and performs sense disambiguation at the same time as we try to identify the main points in a text.

In summary, two problems exist when we consider the possibility of syntactic and semantic ambiguities: (1) multiple counting of weight from a single concept, (2)

mutually related concepts in the interesting wavefront. For the first problem, we compute the weight of a internal concept as the sum of all the leaf concepts in the subtree rooted at the internal concept. For the second, the concept with maximum weight is selected from the concepts comprising the interesting wavefront each time, all the leaf concepts in the subtree rooted at the concept of the highest weight are removed, and the algorithm is repeated with the rest of the concepts. Figure 3.10 lists this improved algorithm.

3.4.6 Unknown Words

We discuss the issue of unknown words in this section. This problem is a well known difficulty for any system designed to tackle real world problems. One solution is to use a big dictionary. In his FERRET information retrieval system, Mauldin [56] demonstrated how to use a dictionary to improve system coverage and robustness.

Although using a dictionary helps, words such as people's names, company names, location names, slang, neologisms, and so on may still not in the dictionary. For example, WordNet, the thesaurus used in our experiment, does not contain any proper nouns and company names. Our solution for the unknown word problem is to use the unknown words simply as they are. When an unknown word is encountered for the first time during processing, we simply record the unknown word and set the weight of this word to one. The recurrence of the unknown word will increase the weight of that word. Even though unknown words are not connected to the system's knowledge base, they are treated as leaf nodes. Since leaf nodes will always eventually appear every time in the interesting wavefront, they will be selected as topic concepts when they are the concepts of the highest weight in the interesting wavefronts. The possibility of using machine learning techniques to attach unknown words into the knowledge base according to run time results is a topic for future research.

3.5 Implementation

This section describes the architecture and components of our experimental topic identification system using the knowledge-based concept counting method. We first

```
 0: ImprovedCollectInterestingWavefrontWithStartingDepth
    (SOURCE,KB,RATIOTHRESHOLD,STARTINGDEPTH)
 1: // Collect interesting wave front from semantic hierarchy
    // SOURCE source text
    // KB semantic knowledge base
    // RATIOTHRESHOLD decides if a node is interesting or not
    // STARTINGDEPTH decides if ratio threshold criterion takes effect
 2: ActivateConcept(SOURCE,KB) // for each concept appears in the SOURCE,
                                  assign its weight equal to its frequecy
                                  in the SOURCE. Nodes in the KB with
                                  weight greater than 0 are active nodes
 3: PropagateWeight(SOURCE,KB) // propagate weights from the active nodes
                                  to all their ancestor in the KB
 4: ComputeRatio(KB) // compute ratio for all the node in the KB
 5: p_list = TopNode // probe list, initialize with the top most node in
                     // the hierarchical semantic knowledge base
 6: i_list = NULL    // interesting node list
 7: fi_list = NULL   // final interesting node list
 8: Children = NULL  // store direct children of an expanded node
 9: IF EMPTY(p_list) THEN GOTO 21
10:   GET a concept n from p_list
11:   IF DEPTH(n) > STARTINGDEPTH THEN
12:     Chidren = EXPAND(n) // get all the direct children of n
13:     PUT Children into p_list
14:     GOTO 9
15:   IF RATIO(n) >= RATIOTHRESHOLD OR n is a leaf node THEN GOTO 18
16:     PUT n into i_list
17:     GOTO 9
18:   Children = EXPAND(n) // get all the direct children of n
19:   PUT Children into p_list
20:   GOTO 9
21: IF EMPTY(i_list) THEN GOTO 25
22: MOVE the node with highest weight MAX_N in i_list into fi_list
23: SET leaf nodes in the subtree rooted at MAX_N in KB to 0 weight
           // 0 weight nodes are inactive and do not pariticpate in
           // weight counting
24: GOTO 3
25: OUTPUT(fi_list) as the interesting wavefront
26: STOP
```

Figure 3.10: Improved algorithm, allowing multiple senses and syntactic categories.

describe the system, then introduce the knowledge sources used, and finally show how to use concept counting to identify topics.

The system goes one step beyond topic identification — it performs sentence extraction for evaluation purposes, by extracting each sentence with a count larger than a threshold amount of topics(s) is extrated.

3.5.1 System Overview

Figure 3.11 shows the knowledge-based topic identification system with evaluation modules. The topic identification part contains six major blocks:

- preprocessing
- concept instantiation
- part of speech tagging (optional)
- weight and ratio computation
- topic concepts identification
- concept weight recomputation

The gray boxes are evaluation modules which are described in Section 3.6 and include two blocks:

- sentence extract generation
- result evaluation

The system takes as input an English text and produces a list of interesting concepts with their associated weights. Resources used and user-preset parameters are shown in rounded square boxes and the corresponding data paths are drawn in bold arrow lines. The *hierarchical knowledge base*, *compound word table*, and *stop list* comprise the essential background knowledge for our system. The hierarchical knowledge base is built from *WordNet version 1.4* [57] and the *Penman Upper model* [2], and provides the necessary lexical knowledge and pyramid structure of hypernym/hyponym relations for the weight/ratio computation and topic concepts identification blocks.

For more discussion of the hierarchical knowledge base, see Section 3.5.4. The compound word table contains phrases in the hierarchical knowledge base such as: *chief executive officer (CEO)*, *advertising campaign*, and so on. The phraser uses this table to recognize compound words. A recognized compound word will be treated as a single concept. The stop list consists of words which appear so often in texts that they lose their significance. All the closed-class words such as *a*, *the*, *in*, *to*, *for*, and *on* are in the list. Once a token is identified as a stop list word, it is ignored in later processing.

3.5.2 Preprocessing

The scanner reads input text and produces lexical tokens. One major function of the scanner is to recognize sentence boundaries, paragraph breaks, abbreviations, numbers, and other special tokens. Sentence boundary information is needed in the evaluation process. Abbreviations are restored to their full forms and then used in concept instantiation, for example: *corp.* $\Rightarrow$ *corporation*. Numbers are automatically classified as *price*, *date*, *percentage*, and *number*. For example, tokens *$175* and *$10,000*, are assigned to *price*, token *2/28/1947* is assigned to *date*, tokens ended with *%* are assigned to *percentage*, and other numeric tokens are assigned to *number*. We plan to use this classified information in the future system (it is not used in our current implementation). Thus, numbers are treated as unknown words. Because only the root form of a word is stored in the system's knowledge base, we need a morphology transformation module to carry out some simple inflectional transformation such as *largest* $\Rightarrow$ *large*, *cities* $\Rightarrow$ *city*, and *talked* $\Rightarrow$ *talk*. The morphology transformation module only handles regular transformations. Irregular transformation information (such as *said/say*, *gone/go*, and *children/child*) are stored in an exception table. The phraser accepts tokens from the scanner and performs compound word checking on the sequence of tokens. If a word sequence matches any entry in the compound word table, the word sequence is grouped as a phrase and used as a single unit in the latter processing phase. See Figures 3.17 and 3.18 in Section 3.5.5 for a sample of input and output of the scanner and the phraser.

The user-preset parameters such as starting depth, ratio threshold, and number of sentences to extract (used in evaluation) are provided to the system through a

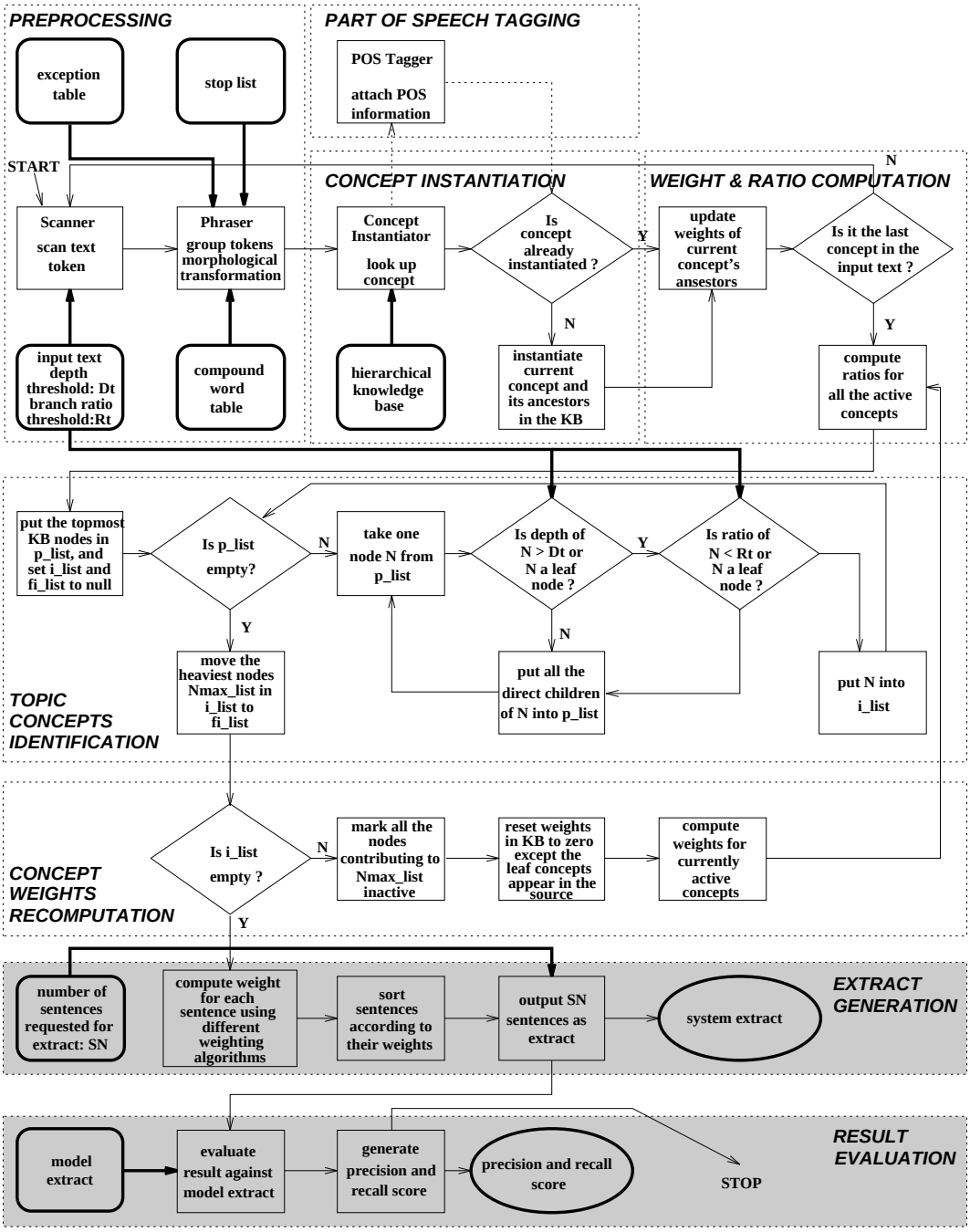


Figure 3.11: Processing flow for current topic identification and sentence extraction system.

Parameter Name	*Value*	*Annotation*
BranchRatio=	0.45	*branch ratio threshold* R_t
StartingDepth=	6	*starting depth* D_s
LWeightThreshold=	4.0	*minimum weight considered significant*
HWeightThreshold=	1000.0	*maximum weight considered significant*
DisplayMode=	1	*various display modes for debugging*
WeightMode=	0	*three weighting methods*
ExtractPercentage=	0.20	*percentage of source extracted*
UsePOSTagInfo=	0	*use POS tag information or not*
OutputFormat=	2	*output as Excel or Matlab format*
ExtractMode=	0	*various extracting modes*
ExceptFile=	allexc.cnt	*exception table file name*

Figure 3.12: A sample system parameter data file.

data file. A sample data file is shown in Figure 3.12. The branch ratio threshold and starting depth are explained in Sections 3.4.3 and 3.4.2. The display mode specifies the level of detail of the output messages. The weight mode parameter is used to select different weight assignment methods to extract sentences for evalutation. See Section 3.6 for more discussion on these different methods, and the minimum and maximum weight thresholds parameters. The other parameters are self-explained in the annotation column.

The whole system is implemented in C++ and runs on an HP 9000/755 RISC machine. The average time for processing a text of 750 words is 8 to 9 seconds.

3.5.3 Part of Speech Tagger

The part of speech tagger is a stand-alone system developed by Brill [6, 7]. If part of speech information is to be used in the processing, the source text is first run through the tagger to generate a tagged file such as the sample shown in Figure 3.13. The tagger attachs a part of speech tag at the end of each word token. For example, *the* is tagged as DT (determiner), and *American* as NNP (proper noun). Because our system only uses four major part of speech categories (noun, verb, adjective, and adverb), the phraser treats all NN* as nouns, VB as verbs, JJ* as adjectives, RB as adverbs, and ignores all the other tags. The tagger cannot assure 100% correctness;

1: The/DT Great/NNP Laptop/char Saga/NN ,/, Chapter/NN One/CD :/: Japanese/JJ personal/JJ computer/NN giants/NNS figure/NN out/IN how/WRB to/TO cram/VB 30/CD pounds/NNS of/IN circuitry/NN into/IN lightweight/JJ machines/NNS ./.
2: Toshiba/NNP and/CC NEC/NNP take/VB the/DT lead/NN in/IN an/DT exploding/VBG market/NN ./.
3: Chapter/NN Two/CD :/: American/NNP rivals/NNS figure/NN it/PRP out/RB ,/, too/RB ./.
4: Chapter/NN Three/CD :/: With/IN better/JJR technology/NN ,/, U.S./NNP laptop/NN makers/NNS Compaq/NNP ,/, Apple/NNP ,/, and/CC IBM/NNP seize/VB control/NN of/IN the/DT market/NN for/IN notebook-size/char computers/NNS ./.

Figure 3.13: A sample output of the part of speech tagger.

for example, in sentence 3 in Figure 3.13, *figure* was tagged as NN instead of VB. This error would direct the propagation of weight from *figure* to a wrong sense (*illustration*, *body*, and *important person* vs. *understand*). Therefore, we also set up an experiment to test the effectiveness of using the tag information generated by Brill's part of speech tagger. Details of the results of the experiment are presented in Section 3.6.1. Concept weight propagation is explained in Section 3.5.5.

3.5.4 Hierarchical Knowledge Base

As described in Section 3.4.1, we need a hierarchical knowledge base to count concepts instead of words. In this section we describe the knowledge sources used in the knowledge-based topic identification system. We first introduce WordNet [57], then discuss the need of expanding the verb hierarchy of WordNet with part of the Penman Upper Model [67]. The augmented knowledge base is called the *Knowledge Kernel* and is described at the end of this section.

3.5.4.1 WordNet

WordNet [57] is a machine-readable thesaurus developed on psycholinguistic principles. It started as a project for developing a lexical database at Princeton University in 1985. It currently contains approximately 95,600 different word forms (51,500 simple words and 44,100 collocations) in 70,100 word meanings, which covers the full range of common vocabulary. WordNet divides lexicons into four categories: noun, verb, adjective, and adverb. We only used the noun and verb sub-hierarchies of WordNet.

WordNet contains approximately 57,000 nouns in 48,800 synonym sets. These are partitioned into twenty-five topical files. Table 3.3 shows these categories with examples and brief explanations. Two major types of relation links connect the noun hierarchy: *hyponymy* and *meronymy*. Hyponymy is also called the "*is_a*" relationship and meronymy the "*part-whole*" relationship. Our system only uses the hyponymy link (and its inverse hypernymy) to move up and down the hierarchy. Although we recognize that the potential benefits of using meronymy, we leave it to future research.

The noun hierarchy of WordNet has an average depth of twelve, under several single topmost root concepts. Because it is desirable for us to have a single root concept, we simply added a top node, *UM-thing*, to the existing WordNet hierarchy. UM-thing means *Upper Model thing*.

WordNet does not contain proper nouns such as company names. But since our corpus is the business domain, company names have to be recognized by the system. Two solutions are posibible. The first one is to build a proper noun recognition program and dynamically attach the recognized proper noun to its proper position in the existing knowledge base. Some kind of learning mechanism can be developed to evaluate the goodness of the attachment and make appropriate adjustment to the knowledge base. Rau [69] has designed an algorithm to extract company names from text and claimed pretty good accuracy. Gallippi [22] applied machine learning technique to identify proper names in multilingual texts and also achieved promising results. The second solution is simply to add the most frequently referenced company names into the knowledge base. Currently, we use the second approach since our test set is a small collection. We plan to use Gallippi's multilingual proper name identification algorithm in the future. Figure 3.2 in Section 3.4 shows part of the noun hierarchy of WordNet rooted at *computer*. The relations between nodes are *is_a* links, from bottom to top.

WordNet includes approximately 21,000 verb word forms in 8,400 synonym sets. The verb hierarchy is divided into fifteen topical categories based on semantic criteria. Table 3.4 shows a short summary of these fifteen categories. Unlike the noun hierarchy, the verb hierarchy is shallow (maximum depth is about 6 levels deep) and organized according to different principle. It is about four levels deep on average; usually, one of the levels is more lexicalized than others. According to WordNet

Twenty-five Topical Categories of Nouns in WordNet		
Category	*Example*	*Explanation*
act, action, activity	{accomplishment, deed}, {behavior, activity}	things that people do
animal, fauna	{vertebrate}, {game, prey, quarry}	a wide variety of animals
artifact	{cloth, fabric, textile}, {medication, medicine}	things made by human beings
attribute, property	{age}, {physiological attribute}	properties of things
body, corpus	{body parts}	major semantic relation is *meronymy*
cognition, knowledge	{ability}, {intellect, mind}	mental contents, states, and processes
communication	{art, fine arts}, {psychic communication}	various aspects of communication
event, happening	{human event}, {natural event}	events
feeling, emotion	{positive emotion}, {mood}	affect states of human
food	{aliment, nourishment}, {beverage, drink, potable}	foods and drinks
group, collection	{array, arrangement}, {group, mathematical group}	a collection of something
location, place	{point, spot}, {region, area}	locations
motive	{motivation, motive, need}, {life}	needs or motives
natural object	{cloud}, {geological formation}	as contrasted with artifact
natural phenomenon	{chemical phenomenon}, {luck, fortune, chance}	natural phenomena
person, human being	{female, female person}, {engineer, technologist}	also includes mythical and supernatural beings
plant, flora	{groundcover, ground cover}	various kinds of plant
possession	{asset}, {liability}	possessions
process	{natural process}, {economic process}	processes
quantity, amount	{definite quantity}, {relative quantity}	numbers, monetary unit, measurement
relation	{connection}, {similarity}	names of relations
shape	{shape, form, contour}	names of shaps
state, condition	{state of matter}, {wholeness, integrity, unity}	names of states
substance	{solid}, {liquid}, {gas}	names of substances
time	{geological time}, {daytime, time of day}	names of times

Table 3.3: Topical categories of nouns in WordNet.

documents [57], the main relation used to organize the verb hierarchy is *entailment*. Four kinds of entailment relations are used in WordNet:

Co-extensive: *limp-walk, lisp-talk*

Proper Inclusion: *snore-sleep, buy-pay*

Backward Presupposition: *succeed-try, untie-tie*

Cause: *raise-rise, give-have*

Tracing into the verb hierarchy, we found the part of hierachy based on the entailment relation is not deep or rich enough to be used for our concept counting method. For example, the entailment chain starting at *snore* is only one level deep, i.e., *snore* $\Rightarrow$ *sleep*. It is even worse that many verbs have no entailment links at all. This probably is a reflection of the claim by Fellbaum [57] that different semantic groups of verbs have distinct structures. The problem is that we need a verb taxonomy at least as deep as the noun hierarchy to make the concept counting technique effective. We found hypernym and hyponym relations in the verb database to be a better alternative. The knowledge structure based on hypernym relation is richer than the one based on the entailment relation. The hypernym relation is widely possessed by verbs. However, the depth of the hierarchy is still shallow. We discuss how to enhance the verb hierarchy of WordNet using part of the Penman Upper Model in the next section.

3.5.4.2 Penman Upper Model

The verb hierarchy of WordNet is divided into 455 separate verb groups with maximum depth 6. They do not have a common root concept. To group them under a single root concept and increase the depth of the verb hierarchy, we created an upper level verb hierarchy on top of these 455 verb groups. The upper level verb hierarchy not only connects these individual verb groups, but also provides the necessary depth the concept counting algorithm. We decided to set up a 4 to 5-level upper verb hierarchy based on Penman's Upper Model [2].

Penman is a text generation system developed at USC/Information Science Institute [67]. The Penman Upper Model is one of the information resources of the Penman system. It is based on *language use* and reflects "the natural organization

Fifteen Topical Categories of Verbs in WordNet		
Category	*Example*	*Explanation*
bodily functions and care	{sweat}, {freeze}, {wash}, {dress}	unaccusative
change	{change, alter}, {change2, turn}, {change3, adjust}	one of the largest verb categories, also *-ify & -ize*
communication	{lisp, stammer}, {fax, telex, e-mail}	verbal and nonverbal, speaking and writing
competition	{face-off, run-off}	in sports, games, and warfare
consumption	{drink}, {eat}	ingesting, using, exploiting, spending, and sharing
contact	{fasten, attach}, {scrub}, {grasp}, {paw}	the largest verb categories degrees of force, holding, touching
cognition	{deduce}, {induce}, {infer, guess}	reasoning, judging, learning memorizing, understanding
creation	{invent, conceive} {engrave, print} {weave, sew}	create by mental act create by artistic means create from raw material
motion	{shake}, {twist} {run}, {crawl}	move, motion-in-place locomotion
emotion or psychology	{fear}, {miss} {amuse}, {charm}	S:Experiencer,O:Source S:Source,O:Experiencer
stative	{connect, link}, {lack, miss}	verbs of being and having
perception	{see}, {smell}, {hear}, {taste}, {touch}	five senses
possession	{have, hold, own}, {give, transfer}, {take, receive}	change of possession and its prior or resultant state
social interaction	{petition}, {veto} {court-martial} {franchise} {ordain}	verbs from different areas of social life: law, politics, economy, education, family, religion, etc.
weather	{rain}, {thunder}	sematically and syntactically distinct from other groups ex: It rains.

Table 3.4: Topical categories of verbs in WordNet.

of terms referring to the world through the language people use to describe and discuss it." The full Penman Upper Model consists of over 200 entities in a generalized hierarchy. We used only some of the top 4 levels of the **process** hierarchy (which contained 20 entities), and added some other entities when necessary. The final version of our upper model is shown in Figure 3.14. The 455 topmost nodes in the WordNet verb hierarchy were manually attached to this upper model. We also developed some simple knowledge base maintenance tools to speed up the process and avoid errors. The final verb hierarchy has an average depth of 6. For example, *snore* has the following two paths: {*snore* ⇒ *breathe* ⇒ *INGEST* ⇒ *ACTIVITY* ⇒ *ACTION-P* ⇒ *PROCESS*} and {*snore* ⇒ *utter* ⇒ *NA-VERBAL-P* ⇒ *VERBAL-P* ⇒ *ACTION-P* ⇒ *PROCESS*}[2]. The concepts shown in capital letters belong to the new upper model.

3.5.4.3 Knowledge Kernel

The enhanced WordNet noun and verb hierarchies comprise the knowledge base for the concept counting topic identification algorithm. We call this knowledge base the *knowledge kernel (KK)*. Figure 3.15 shows a sample entry in the KK. The first eight-digit number, 00007266, is the unique identification number of the specific synonym set. It is followed by a category identification number, 03 for *plant*, a part of speech category identification letter, n for *noun*, and then another number to indicate how many lexical units are included in the synonym set (for example, the three lexical units in this synonym set are *plant, flora, plant life*). The number 016 is the number of relations connecting this synonym set to other synonym sets. This synonym set has sixteen relation links. Link-specific information comes after this indication number. A unit of link-specific information consists of four elements: relation type, pointer to the related set, part of speech category of the related set, and a number that further specifies which one of the lexical units in the related set is targeted. The symbol '@' means hypernym, '#m' member-meronym, '∼' hyponym, '%s' substance-holonym. Meronym and holonym indicate part-whole relations. The relations are reversible: if W_m is a meronym of W_h, then W_h is a holonym of W_m. The member

[2]The suffix "*P*" and "*NA*" are internal coding scheme, where "*P*" means *process* and "*NA*" means *non-addressee*.

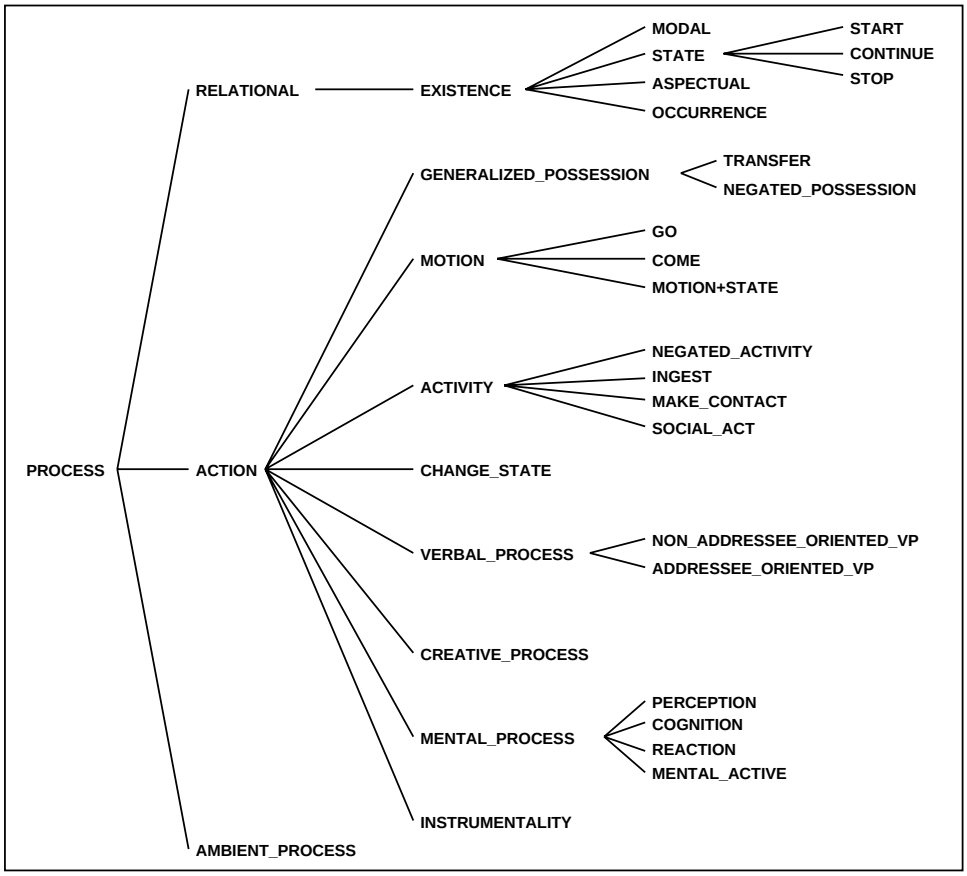


Figure 3.14: The upper model of the topic identificaiton system.

```
00007266 03 n 03 plant 0 flora 0 plant_life 0
016
@ 00002682 n 0000 #m 04887243 n 0000 ~ 04903580 n 0000 ~ 04987460 n 0000
~ 04987581 n 0000 ~ 04987713 n 0000 ~ 04987872 n 0000 ~ 05706620 n 0000
~ 05744257 n 0000 ~ 05744378 n 0000 %s 05750768 n 0000 ~ 05754071 n 0000
~ 05768076 n 0000 ~ 05768307 n 0000 ~ 05768790 n 0000 ~ 05807093 n 0000
| a living organism lacking the power of locomotion
```

Figure 3.15: A sample entry for synonym set {plant, flora, plant_life} in the Knowledge Kernel.

prefix indicates members of a collection such as *tree/forest*, and the substance prefix indicates the stuff of an object such as *protein/muscle*. A brief comment of this synonym set may appear at the end of the data entry. Fox example, flora is *a living organism lacking the power of locomotion*. We also developed a tool, ASK, to access KK interactively. The output of a ASK query session for the word flora is shown in Figure 3.16.

3.5.5 Counting and Merging Techniques

Using the iterative algorithm shown in Figure 3.10, we identified a set of interesting concepts based on concept counting. The topic identification process is also illustrated in Figure 3.11, where concept instantiation, weight and ratio computation, topic concept identification, and concept weight recomputation further demonstrate the interaction among these subtasks.

Figures 3.17 and 3.18 show a trace of the concept counting topic identification process using only the noun hierarchy of the Knowledge Kernel. The system first prints the parameters used to perform the processing. Unknown words are identified, shown, and followed by the text of the input source. The system also records statistics of the text. A sentence-by-sentence listing of the text is generated after the system has finished concept instantiation. Active concepts are marked by angle brackets and separate words are grouped together as a phrase if they are found in the Knowledge Kernel (for example, *NEC Corp.* in sentence S1 and *AST Research Inc* in sentence S2). Two iterations are shown in this example. The concept *computer_company* was identified as the concept with the highest weight in the

Offset	:	00007266
File ID	:	3
Part of speech	:	n
Word count	:	3
Synset	:	plant.0
		flora.0
		plant_life.0
Pointer count	:	16
Pointers	:	[hypernym 00002682 n 0 0]
		[member-meronym 04887243 n 0 0]
		[hyponym 04903580 n 0 0]
		[hyponym 04987460 n 0 0]
		[hyponym 04987581 n 0 0]
		[hyponym 04987713 n 0 0]
		[hyponym 04987872 n 0 0]
		[hyponym 05706620 n 0 0]
		[hyponym 05744257 n 0 0]
		[hyponym 05744378 n 0 0]
		[substance-holonym 05750768 n 0 0]
		[hyponym 05754071 n 0 0]
		[hyponym 05768076 n 0 0]
		[hyponym 05768307 n 0 0]
		[hyponym 05768790 n 0 0]
		[hyponym 05807093 n 0 0]
Gloss	:	a living organism lacking the power of locomotion

Figure 3.16: A sample ASK output for synonym set {plant, flora, plant_life} in the Knowledge Kernel.

first iteration (it carried weight 6). Subconcepts that contributed weights to *computer_company* were also shown. *Computer_company* received contributions from six subconcepts, namely, *NEC Corp., Toshiba Corp., Tandy Corp., AST Research Inc., IBM, Apple*. Notice that the 11th concept in the first iteration was {*plant, flora, plant_life*} from *Apple*. However, by choosing *computer_company* at this iteration, the {*plant, flora, plant_life*} sense of *Apple* was eliminated. In the second iteration, all the remaining concepts had weight 1 since they all occurred only once in the text. At the end, the system selected *computer_company* as the topic for the text.

Although the topic of the text in Figures 3.17 and 3.18 would be better described as "*the competition among computer companies*", the concept counting method still demonstrated its power in generalizing various computer companies into one single concept *computer company*, which is not possible if the pure word counting method is used.

The indication of competition was described using verb phrases: *come back*, *beat a retreat*, *announce a sale*, *nurse a hit*, and *is closing in*, which are very difficult to capture in any knowledge base. Although our knowledge-based concept counting method can generalize *come back*, *beat a retreat*, and *is closing in* into *move*, this is still far from acceptable. It is clear that verb generalization is more difficult than noun generalization, even if a hierarchical verb taxonomy is available. The difficulty is also reflected in the 455 separate verb groups and shallow depth in the verb hierarchy in WordNet. We leave the solution of verb generalization of this kind to future research. In the next section, we describe methods to evaluate the knowledge-based topic identification technique.

3.6 Evaluation

We have implemented a prototype system to test the automatic topic identification algorithm. As the concept hierarchy, we used an extended noun taxonomy from WordNet. The extended noun taxonomy includes computer companies, organizations, etc. WordNet has been used for other similar tasks, such as [70]. For input texts, we selected articles about information processing of average 750 words each out of *BusinessWeek* (1993–94). We ran the algorithm on 50 texts, and for each text extracted eight sentences containing the most interesting concepts.

```
======================================================================
System Parameters:
BranchRatio=0.45
StartingDepth=6
LWeightThreshold=4
HWeightThreshold=1000
DisplayMode=7
WeightMode=0
Percent=0.3
UseTagInfo=0
OutputMode= Matlab Recall(Y)/FalseAlarm(X)
ExtractMode=standard
ExceptFilename=allexcold.cnt
DoEvaluation=0
======================================================================
unknown word -> $175
----------------------------------------
Input Source
----------------------------------------
Now for the latest installment : a free-for-all . Japanese makers , notably
NEC Corp. and Toshiba Corp. , are coming back with new technology .
Tandy Corp. has beat a retreat , announcing the sale May 26 of its GRiD
notebook subsidiary and two plants to AST Research Inc. for $175 million . IBM
— surprise ! — is nursing a hit . Apple , meanwhile , is closing in on the
No. 1 spot .
----------------------------------------
======================================================================
##### Total 5 sentences
##### Total 68 words/phrases
##### Total 43 known words/phrases
##### Total 1 unknown words
##### Total 24 stop list words
##### Total 14 punctuations
======================================================================
Read 1 top noun index
Total usage 27
----------------------------------------
Clustering
----------------------------------------
S0:
<Now> for the latest <installment> : a <free-for-all> .
.....
Active Concepts:
<Now> <installment> <free-for-all>
----------
S1:
<Japanese> <makers> , notably <NEC Corp.> and <Toshiba Corp.> , are coming back
with new <technology> .
.....
Active Concepts:
<Japanese> <makers> <NEC Corp.> <Toshiba Corp.> <technology>
----------
S2:
<Tandy Corp.> has beat a retreat , announcing the <sale> <May> <26> of its
<GRiD> <notebook> <subsidiary> and <two> <plants> to <AST Research Inc.> for
$175 million .
.....
Active Concepts:
<Tandy Corp.> <sale> <May> <26> <GRiD> <notebook> <subsidiary> <two> <plants>
<AST Research Inc.>
----------
S3:
<IBM> — <surprise> ! — is <nursing> a <hit> .
.....
Active Concepts:
<IBM> <surprise> <nursing> <hit>
----------
S4:
<Apple> , <meanwhile> , is closing in on the <No.> <1> <spot> .
.....
Active Concepts:
<Apple> <meanwhile> <No.> <1> <spot>
----------
.....
Active Concepts:
<Now> <installment> <free-for-all> <Japanese> <makers> <NEC Corp.>
<Toshiba Corp.> <technology> <Tandy Corp.> <sale> <May> <26> <GRiD> <notebook>
<subsidiary> <two> <plants> <AST Research Inc.> <IBM> <surprise> <nursing>
<hit> <Apple> <meanwhile> <No.> <1> <spot>
```

Figure 3.17: Trace of a topic identifcation session, part 1.

```
NumMarkerMarked=27
MaximumDepth=10
StartingDepth=6
-------------------------------------------------------------------
current wavefront stops at:
(1) [d=9] 6 computer_company
marker -> [6/NEC Corp.] [7/Toshiba Corp.] [9/Tandy Corp.] [18/AST Research Inc.] [19/IBM] [23/Apple]
(2) [d=4] 1 topographic_point place spot
marker -> [27/spot]
(3) [d=5] 1 installment
marker -> [2/installment]
(4) [d=5] 1 now
marker -> [1/Now]
(5) [d=8] 1 spot blot smear smirch stain
marker -> [27/spot]
(6) [d=5] 1 point spot
marker -> [27/spot]
(7) [d=8] 1 manufacturer maker
marker -> [5/makers]
(8) [d=9] 1 subsidiary_company subsidiary
marker -> [15/subsidiary]
(9) [d=6] 1 blot smear smirch spot stain
marker -> [27/spot]
(10) [d=9] 1 apple
marker -> [23/Apple]
(11) [d=4] 1 plant flora plant_life

concepts with the highest weight:
<computer_company >

.....
Active Concepts:
<Now> <installment> <free-for-all> <Japanese> <makers> <technology> <sale>
<May> <26> <GRiD> <notebook> <subsidiary> <two> <plants> <surprise> <nursing>
<hit> <meanwhile> <No.> <1> <spot>
NumMarkerMarked=21
MaximumDepth=10
StartingDepth=6
-------------------------------------------------------------------
current wavefront stops at:
(1) [d=4] 1 topographic_point place spot
marker -> [21/spot]
(2) [d=5] 1 installment
marker -> [2/installment]
(3) [d=5] 1 now
marker -> [1/Now]
(4) [d=8] 1 spot blot smear smirch stain
marker -> [21/spot]
(5) [d=5] 1 point spot
marker -> [21/spot]
(6) [d=6] 1 blot smear smirch spot stain
marker -> [21/spot]
(7) [d=4] 1 plant flora plant_life
marker -> [14/plants]
(8) [d=8] 1 nurse
marker -> [16/nursing]
(9) [d=6] 1 maker
marker -> [5/makers]
(10) [d=7] 1 nanny nursemaid nurse
marker -> [16/nursing]
(11) [d=7] 1 nanny nursemaid nurse
marker -> [16/nursing]
===================================================================
concepts with the highest weight:
<spot> <installment> <Now> <plants> <nursing> <makers> <Japanese> <subsidiary>
<No.> <26> <GRiD> <notebook> <sale> <technology> <surprise> <hit> <free-for-all>
<1> <two> <May> <meanwhile>

===================================================================
23 concept nodes are generated
Best concepts for Topic
>> [ 6] computer_company
- <NEC Corp./1> <Toshiba Corp./1> <Tandy Corp./1> <AST Research Inc./1> <IBM/1> <Apple/1>

...
```

Figure 3.18: Trace of a topic identifcation session, part 2.

How now to evaluate the results? For each text, we obtained a professional's abstract of that text from an online service. Each abstract contained 7 to 8 sentences on average. In order to compare the system's selection with the professional's, we identified in the text the sentences that contain the main concepts mentioned in the professional's abstract. Sometimes sentences in the abstract combined ideas across several sentences. We included all the sentences involved. We then scored how many sentences were selected by both the system and the professional abstractor. We are aware of this evaluation scheme is not very accurate, but it serves as a rough indicator for our investigation. Figure 3.19 shows one full text, with the identified sentences underlined, and the human's abstract.

To score the text sentences, we developed three measures, varying the combination of weights of the concepts in the interesting wavefront:

1. The weight of a sentence is equal to the weight sum of all parent concepts of words in the sentence.

2. The weight of a sentence is the weight sum of the words in the sentence.

3. Similar to the first measure, but counts only one concept instance per sentence.

To evaluate the system's performance, we defined three counts:

hits: sentences identified by the algorithm *and* referenced by the abstract

mistakes: sentences identified by the algorithm but not referenced by the abstract

misses: sentences in the abstract but not identified by the algorithm

We then borrowed two measures from Information Retrieval research:

$$\textbf{Recall}: \quad hits/(hits + misses)$$
$$\textbf{Precision}: \quad hits/(hits + mistakes)$$

The closer these two measures to unity, the better the algorithm's performance. The precision measure plays a central role in the text summarization problem: the higher the precision score, the more likely that the algorithm identifies the true topics of a text.

Anyone wanna clone a Mac?

Apple Computer Inc. is putting a happy face on its uphill efforts to license software and hardware , attempting to open the door to the first Macintosh clones. But is anybody else smiling ?

On Sept. 19 , the Cupertino (Calif.) company will unveil a fresh licensing strategy and a new smiley logo to be used by Mac software publishers and potential cloners --- much like the one Microsoft Corp. uses to push its Windows operating system. But after nine months of preparing its plan , a critical piece is missing : any big time takers. " I wish [the process] were faster , for sure , " says Apple Chief Executive Michael H. Spindler. " But on the other hand , we want to do it right .

" Since the Mac was introduced in 1984 , Apple has raked in fat a margins on its proprietary technology , even at the expense of market share. In 1984 , Mac hit the market with easy-to-use icons and snazzy graphics. But Apple has introduced no market-wowing changes in the decade since , while Microsoft has been catching up with its Windows operating system. With the Mac-like Windows95 due next year , Microsoft will have all but closed the gap .

Critics say any clone strategy should have been pursued years ago. " Now the window is closing , " says an executive at Acer Inc. , the Taiwanese computer maker that has spent 15 months negotiating a clone deal with Apple --- to no avail. Apple is dithering. Meanwhile , Acer has had a look at Windows95 and is losing interest. Says a former top Apple executive : " Apple had an ice cube in the desert and everybody wanted it. They could have licensed it to everybody. Now all they've got is wet sand .

" Spindler is banking on a crack in the window , negotiating with several small-potato computer manufacturers. He won't name names , but insiders say Apple has been talking to Fujitsu , Toshiba , Olivetti , Vobis Microcomputer , and Motorola. Together , these companies' share of the worldwide PC market adds up to a measly 5%. *Apple seems forced to fish for small fry because companies such as Compaq Computer Corp. and Dell Computer Corp. are simply not interested. IBM , Apple's partner in other software and chip projects , is discussing Mac licensing with Apple but isn't anywhere near an agreement , sources say .*

Whatever the prospects , licensing its technology has become critical to Apple's future. Executives say the company needs to boost the share of computers that use Apple software to 20% of the PC market to keep software developers interested in writing programs for Macintosh over the long haul. With only 10% of the PC market now , Apple is unlikely to reach its goal without the help of clonemakers. The smaller companies , Apple says , can modestly expand the Mac market without harming Apple profits. Fujitsu Ltd. , for example , claims 42% of the education market in Japan --- a segment Apple hasn't cracked. And Olivetti has 20% of the PC market in Italy , where Apple holds a meager 6%.

" What we want in the first pass is market makers , " Spindler says. " Then we'll go beyond this. " To do so , Apple has assembled a 50-person licensing staff , headed by Vice-President Don Strickland. But ask anyone in the computer industry who doesn't work for Apple what they think of the company's latest licensing strategy. *Good odds the answer will be , " Too little , too late. "*

Original text for test document bw092694064.lex

```
Record #13 of 40195
  Author: Rebello, Kathy
   Title: Anybody wanna clone a Mac?
    Year: 1994
 Journal: Business Week
 Company: Apple Computer Inc. - Licenses
Abstract: Apple has announced a new licensing strategy and smiley-face
          logo to foster the first Macintosh clone PCs, but not much
          interest has been generated.  In the early 1980s Apple developed
          easy-to-use icons and handsome graphics for its proprietary system
          at the expense of market share.  However, the company has produced
          no major innovations since, while Microsoft's Windows operating
          system has gained ground.  Industry analysts claim Apple should
          have pursued cloning years ago and are predicting a slim chance
          for success. Industry experts claim Apple has turned to small-time
          PC manufacturers such as Fujitsu, Toshiba and Motorola because
          larger manufacturers such as Dell Computer and Compaq are not
          interested. IBM has expressed interest, but has not reached an
          agreement. Licensing Apple software is crucial to the company's
          success as officials claim Apple must increase its software share
          to 20% to keep developers interested.
Subjects: Licensing
          Marketing Strategy
          Market Share
          Software Development
          Operating System
          Operating systems - Licenses
          Computer industry - Licenses
  Source: Business Week, n3391 (Sept 26 1994): p64(1). 1994
   Issue: n3391
   Pages: p64(1)
Unique ID: 16247666
```

Online abstract for document bw092694064.lex

Figure 3.19: Source text bw092694064.lex and its abstract (*BusinessWeek* 9/26/94, p.64).

We computed the best branch ratio threshold and starting depth by running the system through different parameter settings. We tested ratios = 0.95, 0.68, 0.45, 0.25 and depths = 3, 6, 9, 12. Among them, $\mathcal{R}_t = 0.68$ and $\mathcal{D}_s = 6$ gave the best results. The average results of 50 input texts with branch ratio threshold 0.68 and starting depth 6 are as follows: recall (R) and precision (P) for the three variations are: var1(R=0.32,P=0.37), var2(R=0.30,P=0.34), and var3(R=0.28,P=0.33) when the system picks 8 sentences. The average scores are 0.32 recall and 0.35 precision.

We also randomly selected sentences and computed the recall and precision to ensure that our system did not perform as a random selection system. The result of the random selection method is 0.18 recall and 0.22 precision in the same experimental setting.

Thus although R=0.32 and P=0.35 are not fantastic results, they are enough better than random selection to indicate that concept counting is a promising method.

3.6.1 The Role of the Part Speech Tagger

To evaluate the effectiveness of using the part of speech tagger, we used the same experimental setting, i.e., branch ratio threshold 0.68 and starting depth 6, and computed the recall and precision scores for the three variations after performing part of speech tagging. The results are: var1(R=0.31,P=0.36), var2(R=0.29,P=0.34), and var3(R=0.32,P=0.36). The average recall and precision are 0.30 and 0.35 respectively. Therefore, using part of speech information does not improve the system performance. The small degradation in recall may due to wrong part of speech assignment mentioned in Section 3.5.3.

3.7 Conclusion

In this chapter we reviewed Luhn's idea of using word frequency to estimate term significance, the inverse document frequency, and the $tf * idf$ term significance assignment method used in Information Retrieval. We then demonstrated how to extend word counting method to a knowledge-based topic identification method which counts concepts instead of just words, so that related words (concepts) can be generalized under a single concept.

To generalize related words, we used a Knowledge Kernel, which is a combination of the noun and verb hierarchies of WordNet and part of Penman Upper Model. We used the branch ratio threshold and starting depth to control the level of generality of the identified concepts respectively. We also described how branch ratio threshold and starting depth echo Fukumoto et al.'s idea of context dependent term significance. The possibility of using concept generalization algorithm to perform sense disambiguation was explained.

It is interesting to notice that the idea of context dependent term significance assignment scheme has been used in our knowledge-based topic identification system [50] to select significant concepts in a hierarchical knowledge base in which the *starting depth* is used to set the initial context and the *branch ratio threshold* (vs. χ^2) is used to determine the significance of a concept in the context. We compute the *branch ratio threshold* from the topmost node in the concept hierarchy and move downward in each iteration until we discover all the significant concepts in the context. We then repeat the process from those significant concepts (a new context) and identify the next level significant concepts. Details of these operations are described in Sections 3.4.2 and 3.4.3. It is possible to apply our topic identification algorithm in Fukumoto's experiment to identify significant terms in every context such as collection, topic category, document, and paragraph.

To evaluate the concept counting algorithm, we measured the recall and precision scores of the sentences extracted using three scoring variations based on the identified concepts, comparing against the manually prepared abstracts over 50 *BusinessWeek* magazine articles. Although the system performance has much to improve, it achieved its current performance without using linguistic tools such as a syntactic parser, pronoun resolution algorithm, or discourse analyzer. Hence we feel that the concept counting paradigm is a viable method which can complement other topic identification techniques. The current system establishes a performance lower bound for future systems.

Since the current algorithm does not incorporate information of within-collection concept frequency distribution such as *idf*, which are proven to be effective from word-based Information Retrieval, we plan to explore this possibility in the future.

Furthermore, concept counting has not been used in Information Retrieval and commerically available extraction packages. We hope that this method can make a significant contribution.

Chapter 4

Using Co-occurrence: Topic Signatures

4.1 Introduction

In this chapter, we describe a method of performing topic identification using *topic signatures*, which consist of topic-related key terms identified by the $tf * idf$ term weighting scheme introduced in Section 3.2.1. As mentioned in Section 3.2.1, $tf*idf$ measure takes into account within-document term frequency and within-collection term frequency. A term with high $tf * idf$ is considered as significant for discriminating those documents in which the term occurs offen, from other documents, in which it does not.

The $tf * idf$ measure is the product of term frequency tf and inverse document frequency idf. It is a very common measure used in information retrieval to assess the importance of a term or phrase for a specific document [5, 89]. According to Salton and Buckley [76], who computed 287 different combinations of term-weighting schemes, the best document term-weighting is provided by $tf * idf$.

Pure $tf * idf$ key term selection is very useful for identifing important terms pertaining to a particular document. For example, Figure 4.1 shows a document from the *Wall Street Journal* and top 20 key terms identified by the $tf*idf$ weighting scheme. *Lorenzo, holder, and voting* are the top three terms. According to the simple $tf * idf$ method, we can output these three terms as topics of this document.

In addition, it would certainly be appropriate to say that the topics of this text include *shares*, *corporate control*, and *company leadership*. But since these three terms do not appear in the text, the $tf*idf$ method cannot produce them. What is needed is a method of recognizing the suitability of new, unmentioned topic words,

given the occurrence of sets or patterns of related words. We discuss how to acquire such patterns in the following section.

One method to create concept co-occurrece patterns is to collect documents similar to this text and to identify the co-occurrence pattern of key terms within a specific topic. Since such patterns of related terms, such as *voting*, *power*, *authorize*, *proxies*, etc., each express some aspect of a complex concept, namely *corporate control*, we can then simply list the complex concept as a topic whenever enough of its component terms occur.

This idea resembles knowledge-based concept generalization, described in Section 3.4.1. However, the applicability and power of that method is constrained by two major weakness:

1. the knowledge base does not contain all knowledge;

2. the knowledge base may contain inappropriate knowledge.

For example, the *Wall Street Journal* document in Figure 4.1 contains the terms *preferred_shares*, *share*, *holder*, and *common*. Their WordNet hypernym trees are shown in Figures 4.2, 4.3, 4.4, and 4.5 respectively. According to these results, *holder* shares no higher relation with the other terms, *preferred_shares* and *share* cannot be generalized under the desired sense *stock*[1] (namely *asset* or *possession*), and the correct sense for *common* does not even exist in WordNet! The type of shortcoming occurs frequently, with disastrous consequences for topic search.

4.2 Acquiring Concept Co-occurrence Patterns

In order to use concept co-occurrence patterns to identify topics, we have to answer the following three questions:

1. what is a concept co-occurrence pattern?

2. what does a concept co-occurrence pattern consist of?

3. how can a concept co-occurrence pattern be found?

[1]Although word *stock* appears both in sense 1 of *preferred_shares* and sense 3 of *share*, *stock* in sense 1 of *preferred_shares* does not have common ancestors with *stock* in sense 3 of *share*.

@WSJ870521-0028

TEXAS AIR CORP. holders$_2$ approved a proposal$_4$ that will increase Chairman Frank Lorenzo$_1$'s voting$_3$ power$_{16}$.

The proposal$_4$ doubled the voting$_3$ power$_{16}$ of each Class$_7$ A common$_{12}$ share to 10$_{18}$ votes$_{14}$.

Mr. Lorenzo$_1$ holds 50.7%$_5$ of that class$_7$, which elects$_{20}$ three-quarters$_{13}$ of the Houston airline$_{10}$ holding_company 's directors .

Holders$_2$ also cleared an increase in authorized$_8$ common$_{12}$ to 200 million shares from 75 million .

An amendment to lift$_{19}$ authorized$_8$ preferred_shares$_{17}$ to 50 million from 10$_{18}$ million was withdrawn , however , because the company received less than 50% of the preferred proxies$_{15}$ required to vote$_{14}$ on the proposal$_4$.

Mr. Lorenzo$_1$ told holders$_2$ that the company still is committed to shrinking labor$_9$ costs at its Eastern$_6$ Airlines$_{10}$, despite the unit 's break-even$_{11}$ first quarter .

"We wouldn't expect this type of performance in a bad year because Eastern$_6$ labor$_9$ costs are just too high , " he said .

Rank	1	2	3	4	5
Term	lorenzo	holder	voting	proposal	50.7%
Weight	19.90	9.66	9.05	8.03	7.61
Rank	6	7	8	9	10
Term	eastern	class	authorize	labor	airline
Weight	7.54	7.26	7.08	6.42	6.23
Rank	11	12	13	14	15
Term	break-even	common	three-quarters	vote	proxy
Weight	6.10	5.75	5.74	5.68	5.33
Rank	16	17	18	19	20
Term	power	preferred_shares	10	lift	elect
Weight	5.01	4.83	4.41	4.38	4.29

Figure 4.1: Sample text from the *Wall Street Journal* AIRLINES (AIR) category with top 20 terms and their corresponding $tf * idf$ weights.

Synonyms/hypernyms (ordered by frequency) of noun **preferred_shares**
1 sense of preferred shares

```
Sense 1
preferred stock, preferred shares, preference shares
       => stock
           => capital, working capital
               => asset
                   => possession
```

Figure 4.2: Synonyms/hypernyms of noun **preferred shares** in WordNet.

Synonyms/hypernyms (ordered by frequency) of noun **share**
5 senses of share

```
Sense 1
share, portion, part, percentage
       => asset
           => possession
                      ...
Sense 3
share
       => stock certificate, stock
           => security, certificate
               => legal document, legal instrument, official document,
                  instrument
                   => document, written document, papers
                       => writing, written material
                           => written communication, written language
                               => communication
                                   => social relation
                                       => relation
                                           => abstraction
                      ...
```

Figure 4.3: Synonyms/hypernyms of noun **share** in WordNet; only senses 1 and 3 are shown.

Synonyms/hypernyms (ordered by frequency) of noun **holder**
2 senses of holder

```
Sense 1
holder
       => holding device
           => device
               => instrumentality, instrumentation
                   => artifact, artefact
                       => object, inanimate object, physical object
                           => entity

Sense 2
holder
       => capitalist
           => person, individual, someone, mortal, human, soul
               => life form, organism, being, living thing
                   => entity
               => causal agent, cause, causal agency
                   => entity
```

Figure 4.4: Synonyms/hypernyms of noun **holder** in WordNet.

Synonyms/hypernyms ordered by Frequency) of noun **common**
1 sense of common.

```
Sense 1
park, commons, common, green
       => tract, piece of land, piece of ground,
          parcel of land, parcel
           => geographical area, geographic area,
              geographical region, geographic region
               => area
                   => region
                       => location
```

Figure 4.5: Synonyms/hypernyms of noun **common** in WordNet

For our purpose, a concept co-occurrence pattern is a complex topic in which we are interested. Consider, for example, the topic *Dragon Boat Festival* in Taiwan. A concept co-occurrence pattern for *Dragon Boat Festival* would consist of several key concepts that uniquely identify it. On the Dragon Boat Festival, Taiwanese hang *calamus*[2] and *moxa*[3] on their houses' front doors. They paste up pictures of *Chung Kuei* (a nemesis of evil spirits), and stand *eggs*[4] on end. Adults drink *hsiung-huang wine*[5], children wear *fragrant sachets*[6], and families make *tsung-tzu*[7]. *Dragon boat races* are held around the country [23]. *Calamus*, *moxa*, *Chung Kuei*, *egg*, *Hsiung-Huang wine*, *fragrant sachet*, *tsung-tzu*, and *dragon boat race*, when co-occurring, pertain to an unique concept: *Dragon Boat Festival*. However, these concepts occurring individually would not lead us to the concept *Dragon Boat Festival*.

How can the components of a complex concept such as *Dragon Boat Festival* be acquired? One way is to learn from a teacher. An encyclopedia is a good example: a good one will contain facts about Taiwanese festivals and the *Dragon Boat Festival*. However, what if we want to know about Tibetan Festivals? The problems of incompleteness and inappropriate information still exist.

Another way is to learn from experience: collecting a large enough set of texts about *Dragon Boat Festival* from Taiwan, we can use the $tf * idf$ method to select key terms from each text and identify the frequently co-occurring key terms as a pattern within the collection. If the results indicate that concepts, *calamus*, *moxa*, *Chung Kuei*, *egg*, *hsiung-huang wine*, *fragrant sachet*, *tsung-tzu*, and *dragon boat race* co-occur in most of the documents in the collection but do not co-occur in other texts unrelated to the *Dragon Boat Festival*, we can then construct a *topic signature* of *Dragon Boat Festival* which consists of these key terms. We can use this *topic signature* to identify the topic *Dragon Boat Festival*, or even to augment an

[2]Calamus is believed to have the ability to ward off evil.

[3]Moxa is believed to have the power of preventing pestilence and strengthening health.

[4]It is said to be an auspicious omen if you are able to stand an egg upright at noon on Dragon Boat Festival.

[5]Hsiung-huang wine is made of mineral hsiung-huang and rice wine and is said to cure illness when taken in small amounts.

[6]Fragrant sachets are believed to bring good luck and repel evil.

[7]Tsung-tzu was originally eaten on Dragon Boat Festival only, but gradually envolved into a snack eaten during normal occasions.

existing knowledge base. Learning from a teacher and learning from experience are not competing methods. They complement each other.

In Section 4.3, research related to this work is discussed. We next define document and topic signatures, show how to assess inter-topic relationship using a simple similarity measure, how to generate a *confusion set* for each topic and use it to construct multi-level topic signatures, and how to use *multi-level topic signatures* to identify topics. Implementation and extensive evaluations are presented in Sections 4.5 and 4.6 respectively.

4.3 Related Work

The utility of concept co-occurrence has been demonstrated in other reseach. Artificial Intelligence techniques have been tried in topic identification [14, 56, 71].

In late 1970's, DeJong [14] developed a system called FRUMP (Fast Reading Understanding and Memory Program), which is a newspaper skimming program developed at Yale to skim and summarize news articles. FRUMP uses a data structure called a *sketchy script* to organize its world knowledge. Sketchy scripts represent what can occur in particular situations such as demonstrations, earthquakes, labor strikes, and so on. Given a newspaper article, FRUMP selects the most appropriate sketchy script based on clues found in the article. Typically, these clues are words that reliably indicate the appropriateness of a sketchy script — words like *Richter scale*, *death toll*, and *magnitude* for earthquakes, or *visit*, *dignitary*, and *meeting* for diplomatic visits. A summary can be generated based on what has been activated in the sketchy script, since it has been predefined to contain only the interesting or important aspects of a specific topic.

Recently, Riloff and Lehnert [71] employed the *information extraction* techniques in three text categorization algorithms. Text categorization is the classification of texts into predefined categories. The categorization task is traditionally done by human experts. Two steps are involved: (1) to identify the main topic of a text, and (2) to find an appropriate category for the topic. If topics are uses as categories, text categorization and topic identification are equivalent tasks. Riloff and Lehnert's algorithms use *relevancy signatures*, each relevancy signature being pair consisting of a trigger word and a *concept node* that it triggers. Concept nodes are generated

by a conceptual sentence analyzer called CIRCUS [45], which is based on a domain-specific dictionary of relevant information extracted from sentences. They serve to capture the natural language context surrounding a word. Using relevancy signatures and their algorithms, Riloff and Lehnert achieved over 80% precision with up to 50% recall against baseline precisions of 69% and 55% on two test sets of 100 documents each. Since they claim that "a single relevant sentence is often enough to classify a text as relevant" and "once a relevant sentence is identified, the remainder of the text can be ignored", their algorithms are tuned to *single-topic* texts. The applicability of using these algorithms to multiple-topic texts has not been demonstrated. However, as we mention in Section 4.5.1, many of our texts contain multiple indices, and even single indexed texts actually include multiple topics. This fact makes methods such as Riloff and Lehnert's somewhat less appealing than they might be.

Generally speaking, Artificial Intelligence techniques, such as FRUMP and CIRCUS+revelant signatures, use semantic analyzers and script-based knowledge representations to capture co-occurrence information. They tend to achieve high precision on small collections, but are costly to build, do not ensure good coverage, and have no stated procedures to tackle mutiple-topic texts or identify inter-topic relationships.

The major alternative approach, and the more traditional one, employs statistics gathered over texts in one form or another.

Term-weighting assignment and similarity measure are two key issues involved in the process of determining topics of texts using vector-space model. Term-weighting assignment, as discussed in Section 3.2, is a way to specify the relative importance of a term within a document. This can be achieved in various ways. We described $tf * idf$ at the outset of this chapter. Another method is χ^2, a way of measuring the singularity of a term pertaining to a specific domain or topic relative to all other terms. Watanabe et al. [85] extracted domain specific Janpanese kanji characters using the χ^2 method. They used these key characters to classify three different sets of articles, achieving accuracy levels of 47%, 74%, and 85% respectively.

Statistical word-based techniques have been applied to text categorization for a long time. The vector-space model is the most popular, thanks to its simple definition and ease of implementation. In this model, each document is represented as a vector, where each dimension (component) of the vector states the relative importance of one term. Typically, a vector space contains tens of thousands of

dimensions, one for each words in the text corpus. Each vector represents one document in the space; each vector consists of *terms*. Each term corresponds to a word or word-stem in the document. The value of a term can be *binary* or *weighted*. Binary values indicate the simple presence or absence of a term in the document, while weighted values usually express the number of times a term appears in the document. To identify the category of a text using the vector-space model, one generates a *centroid vector* for each predefined topic from a training corpus. A centroid vector can be regarded as an average representation of that topic over all the texts in the training corpus. To classify a document, one then generates a vector for it, computes the *similarity* of this vector to all controid vectors, and assigns the test document to the most similar topic.

Similarity measures can be computed in several ways [74, 84]. The most popular one is cosine correlation, which measures the cosine of the angle between the document and centroid vectors. Rijsbergen [84] has stated that "the difference in retrieval performance achieved by different measures of association (similarity) is insignificant," we have reached a similar conclusion.

Although word-based techniques have been developed and applied in many practical cases, how to use them to derive inter-topic relations, improve them to differentiate closely related topics, and incorporate domain specific information to normalize *idf* has not been well studied. Being able to assess the inter-topic relatedness and further categorize closely related topics are very important for employing topic identification in other text processing tasks, such as automated text summarization. To resolve the weaknesses of frequency-based word and knowledge-based concept counting methods, we propose a new method, *multi-level topic signatures*, to complement the methods introduced in Chapter 3. Details of this method are presented in the rest of this chapter

4.4 Concept Signatures

In this section we define terms, develop notation, and describe our version of concept signatures.

4.4.1 Document Signature

Each document is associated with a *document signature* defined as follows:

$$idf_j = \log \frac{N}{df_j} \tag{4.1}$$

$$w_{ij} = tf_{ij} \cdot idf_j \tag{4.2}$$

$$DS_i = < (t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \ldots, (t_{in}, w_{in}) >, \ for \ w_{ij} \geq w_{ij+1} \tag{4.3}$$

where N is the number of documents in the training corpus, df_j is the number of documents in which term j is present, idf_j is inverse document frequency, t_{ij} is the number of times term j occurs in document i, tf_{ij} is the term frequency of t_{ij}, w_{ij} is the term weight of t_{ij}, and DS_i is the document signature of document i.

4.4.2 Topic Signature

For k a complex concept, represented by an English word or phrase, a signature for k is a list of pairs:

$$TS_k = < (t_{k1}, u_{k1}), (t_{k2}, u_{k2}), \ldots, (t_{kn}, u_{kn}) >, u_{kj} \geq u_{kj+1} \tag{4.4}$$

$$\bar{tf}_{kj} = \frac{\sum_{i=1}^{N_k} tf_{ij}}{N_k}, \ \forall \ document \ i \ in \ concept \ k \tag{4.5}$$

$$N = \sum_{k=1}^{m} N_k \tag{4.6}$$

$$u_{kj} = \bar{tf}_{kj} \cdot idf_j \tag{4.7}$$

where m is the number of different complex concepts, each t_{kj} is an English word or phrase, and each u_{kj} is its average associated strength/weight, and:

- n – cutoff number of signature terms chosen to represent concepts
- N_k – the number of documents indexed as relating to concept k, and to no other concept
- tf_{ij} – the frequency of term j in document i
- $\bar{tf}_{kj}$ – the average frequency of term j per document single-indexed by concept k
- idf_j – from Equation 4.1

Intuitively, for a complex concept k such as *Dragon Boat Festival*, each signature term $t_{dragon\_boat\_festival\ j}$ (such as *moxa*, *calamus*) appears with a characteristic

relative strength — perhaps twice as many *moxa* as *calamus*, etc. The n most distinguishing words for concept k (say, 300) are determined by $tf * idf$, paired with their frequencies, and listed in order. When a document contains enough of these terms in the right relative proportion of frequency, then one can say that one of the document's topics is the complex concept k.

A topic signature is derived from a collection of documents associated with the topic. It consists of terms pertaining to the specific topic according to the feature selection measure[8] Equations 4.1 and 4.7 show that the term weights u_{kj} of a topic signature serve as term importance weighting factors. These values model the frequency distribution of terms in and across topics. Ideally, we include the high frequency terms whose distributions are concentrated in that specific topic into a topic signature. The inverse document frequency shown in Equation 4.1 is a simple approximation, while the χ^2 method used by Watanabe et al. [85], and probabilistic models used by Larkey et al. [42] and Joachims [38], are other alternatives.

4.4.3 Similarity Measure

Since we have defined both document signature and topic signature in vector form, we can compare them using any of several vector similarity measures [74]. One possible similarity measure is the cosine cofficient measure, defined as:

$$sim(DS_i, TS_j) = \frac{DS_i \cdot TS_j}{\mid DS_i \mid\mid TS_j \mid} \tag{4.8}$$

where "$\cdot$" is the dot product. The result, a value between 0 and 1, is the similarity between document i and topic j. It measures the cosine angle between document and topic signatures: the closer its value is to 1, the more similar the document and topic signatures are. We can compute the similarity between a document and each topic signature, sort the results, and assign the topic with highest similarity value to the document.

[8] We use $tf * idf$, but other alternatives are possible.

4.4.4 Inter-topic Relatedness and Confusion Sets

Equation 4.8 defines the similarity between a document signature and a topic signature. It can also be used to compute the similarity between two topic signatures. Replacing DS_i in Equation 4.8, we have:

$$sim(TS_i, TS_j) = \frac{TS_i \cdot TS_j}{\mid TS_i \mid\mid TS_j \mid} \tag{4.9}$$

The closer $sim(TS_i, TS_j)$ is to 1, the more similar TS_i and TS_j. Equation 4.9 offers one way to estimate the closeness among topics. Ideally, we would like to have $sim(TS_i, TS_j)$ close to zero, so that documents can be assigned to their topics with high confidence. However, TS_i and TS_j are normally not independent, since they may share terms, as we discuss in Section 4.5.2. Therefore, to further analyze the inter-topic relatedness, we do the following:

1. Let $S_{ij} = sim(TS_i, TS_j)$, with k being the number of topics. Compute S_{ij} for all $j \neq i$ and $1 \leq j \leq k$. Call this set S_i, which contains all the similarity values between topic signature TS_i and all the other topic signatures.

2. Compute the maximum (max_i), 1st quartile[9] (Q_{i1}), median (Q_{i2}), 3rd quartile (Q_{i3}), and the interquartile range ($\triangle Q_i$) of the values in S_i. Interquartile range is the difference between Q_{i3} and Q_{i1}. We define *outlier threshold* θ_i as the value that lies at 1.5 times $\triangle Q_i$ from Q_{i3}, and call any values in S_i greater than θ_i the *outliers*. The relations among these values are illustrated in Figure 4.6. The value 1.5 is chosen as a rule of thumb according to the statistics literature [80]. Outliers are values considered *very different* from the rest of values in S_i, since they have values 1.5 times larger than 75% of the values in S_i.

Where in the range of 0 to 1 the median Q_{i2} lies is very important for identifying topic by topic signatures. If Q_{12} is close to 1, the topics are very similar to each other according to their topic signatures. In this case, topic signatures are unlikely to differentiate well, meaning that we must use alternative methods to perform

[9] 1st quartile is the value which has one quarter of values in S_i below it; while 3rd quartile has three quarters. The 2nd quartile is called the *median*.

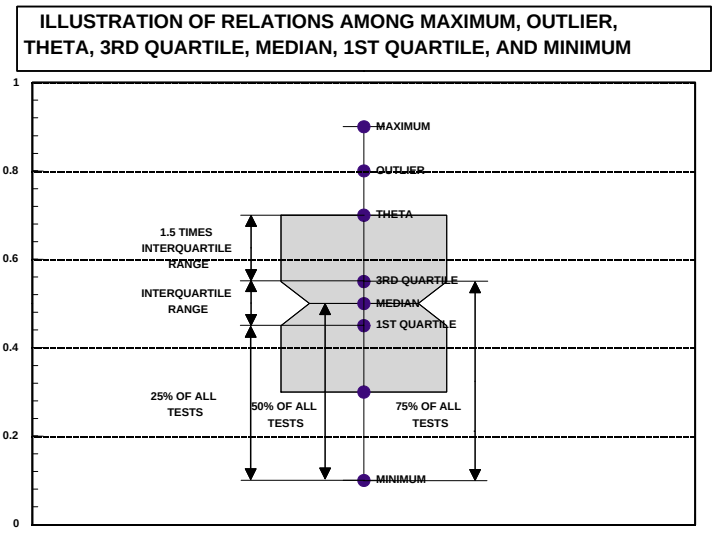


Figure 4.6: Relations among maximum, 1st quartile, median, 3rd quartile, and outliers. Note that maximum is not necessarily greater than outlier limit.

topic identification. On the other hand, if Q_{i2} is close to 0, the topic signatures are mostly independent of each other, and we can expect good identification results. Furthermore, if max_i is less than θ_i then no outliers exist, and we can use θ_i as a cut-off threshold in identification to determine the *goodness* of a match. A good topic identification match has a similarity value computed by Equation 4.8 greater than θ_i, and topic i can confidently be assigned to the test document; if however the result of Equation 4.8 is less than or equal to θ_i, then the match is not a good match and we need to use alternative methods to confidently assign a topic to the test document. One such method is *multi-level topic signatures*, based on the *confusion set*.

Intuitively, the idea is to create a new set of topic signatures, using just terms that differentiated well among the topics in a confusion set. This is a recursive application of signatures, now restricted to just documents within a confusion set. Since the set of such documents alone has different *idf* values for terms, the sub-signatures will differ from the original signatures. That is precisely what we want.

If Q_{i2} is close to 0 and outliers exist, i.e., $max_i > \theta_i$, we define a *confusion set*, CFS_i, as:

$$CFS_i = \{TS_j | \forall j \neq i, sim(TS_i, TS_j) > \theta_i\} \tag{4.10}$$

Notice that each topic T_i has its own θ_i. We then compute additional term weights and *inverse topic signature frequencies* for topic signatures in CFS_i using Equations 4.1, 4.6, and 4.7 with the following modifications, where l is the signature level designator:

$$N_i^{l+1} = |CFS_i^l| \tag{4.11}$$

$$idf_{ij}^{l+1} = \log \frac{N_i^{l+1}}{df_{ij}^{l+1}} \tag{4.12}$$

$$u_{ij}^{l+1} = u_{ij}^{l} \cdot idf_{ij}^{l+1} \tag{4.13}$$

where for confusion set of topic i:

$|CFS_i^l|$ – the number of topic signatures
N_i^{l+1} – the number of topic signatures
df_{ij}^{l+1} – the number of topic signatures in which term j appears at level $l+1$
idf_{ij}^{l+1} – inverse topic signature frequency at level $l+1$
u_{ij}^1 – equal to u_{ij} in Equation 4.7
u_{ij}^{l+1} – new term weight at level $l+1$

Starting with $l = 1$, we generate the original *topic signatures*, as described in Section 4.4.2. We call the orginal topic signatures *first level topic signatures*. We can then compute signature on the next levels $l+1$, $l = 2$, etc., for just those documents in confusion sets, according to the new θ_i at level $l+1$, written as θ_i^{l+1}, and continue this process until no more confusion sets occur, or until topic signatures in confusion sets are no longer distinguishable. The superscript l is used to label each iteration. The level $n+1$ topic signatures are used whenever outliers exist at level n and the similary of the test document and the best matched topic signature is below max_i^n, since a similarity value less than the maximum of its confusion set is not reliable.

Finally, to match documents using multilevel topic signatures, we need to modify the term weight in Equation 4.2 as follows:

$$w_{ij}^{l+1} = w_{ij}^{l} \cdot idf_{ij}^{l+1} \tag{4.14}$$

where w_{ij}^{1} is equal to w_{ij} in Equation 4.2, and idf_{ij}^{l+1} is from Equation 4.12.

4.4.5 The Process of Identifying Topics

Using Equations 4.1 to 4.14, we perform the topic identification task for a document i as follows:

1. Compute document signature DS_i of document i.
2. Compute similarity value $sim(DS_i, TS_j)$ between each document signature DS_i and topic signature TS_j.
3. If $max_j \leq \theta_j$ and $sim(DS_i, TS_j) > \theta_j$ then assign topic j to document i.
4. If $max_j \leq \theta_j$ and $sim(DS_i, TS_j) \leq \theta_j$ then pass this document.
5. If $max_j > \theta_j$ and $sim(DS_i, TS_j) > max_j$ then assign topic j to document i.
6. If $max_j > \theta_j$ and $sim(DS_i, TS_j) \leq max_j$ then compute the next level document and topic signatures, and repeat the procedure from Step 2; if no more confusion sets are available, then assign topic j to document i.

At the end of this process, we have one collection of documents indexed by topics and another collection of documents which have not been assigned any topics because of weak similarity. If true topics of documents assigned by human experts are also available, we then can evaluate our method using *recall* and *precision* measures.

4.5 Implementation

The goal of our study is to develop a simple and robust method to perform topic identification. In service of this goal, we have to treat inter-topic relatedness and

verify our method on a large document collection such as DARPA's TIPSTER corpora [27]. In the previous section, we defined signatures. In this section, we describe their construction.

Several practical problems have to be solved to achieve the main goal. An important early step is the proper treatment of terms. In the text collection, terms appear in various forms, in various morphological variations, sometimes written as phrases using several words, etc. To group words into meaningful phrases and transform words into their root forms, we used WordNet [57] as our dictionary[10]. Documents in the training and test sets were first tagged by Brill's rule-based part of speech tagger [7]. Tagged words were then transformed into their root forms, and then grouped into phrases according to WordNet. We next collected term frequency and inverse document frequency statistics, and then generated topic signatures for all the single-indexed topics listed in the 1987 *Wall Street Journal* texts from the TIPSTER collection. For evaluation (discussed in Section 4.6), we used the 1988 *Wall Street Journal* texts as test set.

Inter-topic relatedness can be measured by appying simple statistical techniques on *similarities* among topic signatures. For each topic signature TS_i, a *confusion set* CFS_i consists of all other topic signatures TS_j whose similiarity with TS_i, written as $sim(TS_i, TS_j)$, is greater then some threshold θ_i. After finding the confusion sets in our data (of which there were 21, approximately 2/3 of the first level signatures), we next computed the *second level* topic signatures for each confusion set, using term weight w_i of TS_i and the frequency of term appearance across current topic signatures as a measure of term specificity, i.e., *inverse topic signature frequency*. Though in principle this process can continue until either no more confusion sets are generated or topic signatures within a confusion set are not distinguishable by the simple $tf * idf$ measure, it was not necessary in our case to go beyond level 2. Primitive inter-topic relatedness was thus managed by simple statistical techniques.

To evaluate the quality of the signatures, we computed the similarity between each test document and the topic signatures, and assigned the document the *topic candidate* that had the maximum similarity. If the similarity between the test document and the topic candidate was greater than a threshold θ_i, then the topic was

[10] Words such as *ate* and *eaten* are transformed into *eat*; consecutive words occur in texts such as *air force* are grouped into a phrase *air_force*, if *ate*, *eaten*, and *air_force* are in WordNet.

output as result; if not, the next level topic signatures were used to determine the final topic for the text. In our experiments, two level topic signatures were used, with very promising results. We discuss these concepts and procedures in more detail in the following sections.

4.5.1 Corpus Statistics

As mentioned above, the *Wall Street Journal* 1987 and 1988 texts in the TIPSTER collection were used as training and test sets respectively. Figure 4.7 is a typical example of a *Wall Steet Journal* text. Each text in the collection contains SGML tags which provide extra information about the text body. As shown in Figure 4.7, a text contains:

- a unique message identification number: WSJ870324-0001.
- a message headline: "John Blair Is Near Accord To Sell Unit, Soures Say."
- a message date: 03/24/87.
- a message source: WALL STREET JOURNAL (J).
- a manually assigned message index section: REL (Reliance Capital Group Inc.), is a company index which is referred to in the message body. TNM, MKT, and TEL are topic categories of the message body. The *Wall Street Journal* indexes used a set prespecified categories. We selected 32 topic categories which contain at least 100 single-indexed texts. These are listed in Table 4.1.
- a message body.

<DOC>
<DOCNO> WSJ870324-0001 </DOCNO>
<HL> John Blair Is Near Accord
To Sell Unit, Sources Say</HL>
<DD> 03/24/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<IN> REL
TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)
MARKETING, ADVERTISING (MKT)
TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>

John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said.
Industry sources put the value of the proposed acquisition at more than $100 million. John Blair was acquired last year by Reliance Capital Group Inc., which has been divesting itself of John Blair's major assets. John Blair represents about 130 local television stations in the placement of national and other advertising.
Mr. Rosenfield stepped down as a senior executive vice president of CBS Broadcasting in December 1985 under a CBS early retirement program. Neither Mr. Rosenfield nor officials of John Blair could be reached for comment.

</TEXT>
</DOC>

Figure 4.7: Sample text WSJ870324-0001 of the *Wall Street Journal* in the TIPSTER collection.

CODE	FULLNAME	FQ	%
air	AIRLINES	743	0.0460
aro	AEROSPACE	373	0.0231
aut	AUTOS, AUTO PARTS	674	0.0418
bbk	BUYBACKS, REDEMPTIONS, SWAP OFFERS	948	0.0587
bcy	BANKRUPTCIES	161	0.0100
bnk	BANKS, THRIFT INSTITUTIONS	459	0.0284
bon	BOND MARKET NEWS	915	0.0567
ceo	DOW JONES INTERVIEW	427	0.0265
cmd	COMMODITY NEWS, FARM PRODUCTS	115	0.0071
div	DIVIDENDS	656	0.0407
eco	ECONOMIC NEWS	291	0.0180
edp	COMPUTERS	300	0.0186
ele	ELECTRIC, ELECTRONICS, APPLIANCES	139	0.0086
env	ENVIRONMENT	139	0.0086
ern	EARNINGS	700	0.0434
fab	FOOD & BEVERAGE, HOUSEHOLD GOODS, SUPERMARKETS, TOBACCO	209	0.0130
fin	FINANCIAL, INSURANCE, MUTUAL FUNDS, ACCOUNTING, LEASING	295	0.0183
lng	NATURAL GAS, PIPELINES	102	0.0063
min	MINING, METALS	225	0.0139
mkt	MARKETING, ADVERTISING	129	0.0080
mon	MONETARY NEWS, FOREIGN EXCHANGE, TRADE	833	0.0516
pet	PETROLEUM	172	0.0107
pha	PHARMACEUTICALS, HOSPITAL SUPPLIES, MANAGEMENT	488	0.0302
pub	PUBLISHING	210	0.0130
rel	REAL ESTATE, REITS, LAND DEVELOPMENT	203	0.0126
ret	RETAILING	107	0.0066
scr	SECURITIES INDUSTRY	266	0.0165
stk	STOCK MARKET, OFFERINGS	493	0.0306
tel	TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH	437	0.0271
tnm	TENDER OFFERS, MERGERS, ACQUISITIONS	4650	0.2882
tra	TRANSPORTATION, TRUCK AND SHIP LINES, RAILROADS	137	0.0085
uti	UTILITIES	141	0.0087
TOTAL		16137	1

Table 4.1: Wall Street Journal 1987 (training set) topic codes, full names, frequencies, and percentages of the number of texts per topic to the total number of texts in the whole collection.

CODE	FULLNAME	FQ	%
air	AIRLINES	527	0.0408
aro	AEROSPACE	438	0.0339
aut	AUTOS, AUTO PARTS	594	0.0460
bbk	BUYBACKS, REDEMPTIONS, SWAP OFFERS	544	0.0422
bcy	BANKRUPTCIES	163	0.0126
bnk	BANKS, THRIFT INSTITUTIONS	298	0.0231
bon	BOND MARKET NEWS	523	0.0405
ceo	DOW JONES INTERVIEW	279	0.0216
cmd	COMMODITY NEWS, FARM PRODUCTS	110	0.0085
div	DIVIDENDS	433	0.0336
eco	ECONOMIC NEWS	230	0.0178
edp	COMPUTERS	365	0.0283
ele	ELECTRIC, ELECTRONICS, APPLIANCES	127	0.0098
env	ENVIRONMENT	120	0.0093
ern	EARNINGS	737	0.0571
fab	FOOD & BEVERAGE, HOUSEHOLD GOODS, SUPERMARKETS, TOBACCO	0	0.0000
fin	FINANCIAL, INSURANCE, MUTUAL FUNDS, ACCOUNTING, LEASING	250	0.0194
lng	NATURAL GAS, PIPELINES	75	0.0058
min	MINING, METALS	158	0.0122
mkt	MARKETING, ADVERTISING	227	0.0176
mon	MONETARY NEWS, FOREIGN EXCHANGE, TRADE	728	0.0564
pet	PETROLEUM	130	0.0101
pha	PHARMACEUTICALS, HOSPITAL SUPPLIES, MANAGEMENT	515	0.0399
pub	PUBLISHING	220	0.0170
rel	REAL ESTATE, REITS, LAND DEVELOPMENT	161	0.0125
ret	RETAILING	64	0.0050
scr	SECURITIES INDUSTRY	229	0.0177
stk	STOCK MARKET, OFFERINGS	204	0.0158
tel	TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH	355	0.0275
tnm	TENDER OFFERS, MERGERS, ACQUISITIONS	3879	0.3006
tra	TRANSPORTATION, TRUCK AND SHIP LINES, RAILROADS	125	0.0097
uti	UTILITIES	98	0.0076
TOTAL		12906	1

Table 4.2: Wall Street Journal 1988 (test set) topic codes, full names, frequencies, and percentages of the number of texts per topic to the total number of texts in the whole collection.

Only the message identification number, the manually assigned message topic indices[11], and the message body were used in our experiments. Messages do not necessarily contain all the SGML tags, and message formats between training set and test set were a little different, but most included a message identification number, topic indices, and a message body. Tables 4.3 and 4.4 show the number of indices per text in the training and test sets. Note that about 30% of the texts do not have any index, and that most of the texts contain a single (35%) or two indices (22.5%), and the average number of indices per text is 1.26. To avoid confusion, we used only single-indexed texts in our experiments. The training set contained 16,137 texts, manually classified into 32 topics. Table 4.1 shows the three-letter topic codes (CODE column), topic full names (FULLNAME column), number of texts in each topic (FQ column), and what percentage each topic contributed to the training set (% column). These 32 topics were selected because each of them has at least 100 texts. Note that the distribution of the number of texts in each topic was not homogeneous: TNM was the most frequent topic type, representing about 28% of the training texts. The test set contained 12,906 texts from 31 topics. Table 4.2 shows topic codes, topic full names, frequencies, and percentages information of the test set. Note that no texts were indexed only by FAB in the test set.

4.5.2 Training Signatures

4.5.2.1 Training and Test Data

It is clear that morphologically stemmed and otherwise canonicalized words may improve the recall of similar documents of the same topic, and phrases may improve the precision of topic identification. To investigate the effect of using morphologically transformed words and phrases as terms instead of just words verbatim from the texts, we set up experiments to construct three sets of topic signatures: (1) using words directly from texts without any modification (designated as **WD**), (2) words with morphological transformation based on WordNet (designated as **TR**), and (3) words plus the phrases recorded in WordNet (designated as **PH**). A stop list which contains high frequency common terms across all topics such as closed-classed words

[11]Company indices are not used.

# of Indices Per Text	# of Instances	Percentage
0	13950	30.03359%
1	16507	35.53867%
2	10366	22.31743%
3	3314	7.13486%
4	1477	3.17990%
5	448	0.96452%
6	207	0.44566%
7	86	0.18515%
8	42	0.09042%
9	17	0.03660%
10	10	0.02153%
11	10	0.02153%
12	9	0.01938%
14	1	0.00215%
15	1	0.00215%
16	1	0.00215%
17	2	0.00431%
Total	46448	100%
Average	1.25	

Table 4.3: Number of indices for the *Wall Street Journal* 1987 collection (training set).

# of Indices Per Text	# of Instances	Percentage
0	11751	29.44818%
1	13813	34.61558%
2	9324	23.36608%
3	2871	7.19477%
4	1216	3.04731%
5	581	1.45599%
6	211	0.52877%
7	77	0.19296%
8	27	0.06766%
9	14	0.03508%
10	4	0.01002%
11	4	0.01002%
12	3	0.00752%
14	3	0.00752%
15	3	0.00752%
16	1	0.00251%
17	1	0.00251%
Total	39904	100%
Average	1.28	

Table 4.4: Number of indices for the *Wall Street Journal* 1988 collection (test set).

(*the*, *a*, *of*, *to*, and so on) was used to filter out terms that had little identification value in our experiments. To obtain the **TR** set, we use Brill's part of speech tagger to assign a part of speech tag to each term in the **WD** set, and a simple morphology analyzer based on WordNet to transform a term into its root form found in WordNet. For example, *ate* and *eaten* are transformed into *eat*. To generate the **PH** set, terms in **WD** that were transformed to root form were grouped into phrases according to the phrase information provided in WordNet. The remaining single words were counted as one-word phrases. Only maximum-length phrases are considered at each word position (for example, *nuclear power plant* will generate two phrases: *nuclear power* and *power plant*, but not *nuclear power plant*. *Nuclear power* is generated from word *nuclear* and *power plant* from word *power*. Since *nuclear power plant* is not recorded as a phrase in WordNet, *nuclear power* is the maximum-length phrase starting with *nuclear*. If *nuclear power plant* is also in WordNet, then *nuclear power plant* and *power plant* will be generated, but not *nuclear power*).

A short summary of the distribution of the number of terms per training set and test set is presented in Table 4.5. The number code 7 indicates training sets (1987), and code 8 indicates test sets (1988). The average number of terms per text is about 237; this is roughly the same across the three sets in the training collection. About 50% of the texts have fewer than 174 terms, and 75% fewer than 274 terms. More detailed information about the distribution of terms in each topic and in each training and test setup is provided in Appendix A. As shown in Tables A.1 through A.6 in Appendix A, the average number of terms in each topic is not uniform. In Table A.1, topic RET has the maximum average number of terms, 414, while topic BBK has the minimum average number of terms, 67. Based on the numbers in Table 4.5, the topic identification results using topic signatures from **WD**, **TR**, and **PH** should be in similar performance range, since they have similar numbers of terms.

4.5.2.2 Training Procedure

Two stages are involved in topic signature training. The first stage is to collect idf_j values for each term j, according to Equation 4.1. The second stage is to compute the average term frequency $\bar{tf}_{kj}$ for each term j in topic k, according to Equation 4.5,

TEXT	MEAN	MIN	1QT	MED	3QT	θ	MAX
WD7	237	24	81	174	274	667	1375
TR7	237	24	81	174	274	666	1372
PH7	232	23	78	170	268	654	1350
WD8	220	24	60	164	196	595	1232
TR8	220	24	60	163	196	594	1230
PH8	215	23	58	160	191	582	1212

MIN: minimum, 1QT: first quartile, MED: median
MAX: maximum, θ: outlier threshold, 3QT: third quartile

Table 4.5: The *Wall Street Journal* 1987, 1988: average number of terms per text per topic.

and its $tf * idf$ term weight u_{kj}, following Equation 4.7. Tables 4.6, 4.7, and 4.8 show the top 5 terms of the computed topic signatures for the three test sets. A brief review of these top 5 signature terms for each topic of the three training sets indicates that these terms are indeed good representatives of concepts associated with their corresponding topics. For example, topic AIR (airlines) in Table 4.6 has *airlines*, *passenger*, *airline*, *air*, and *continetal*. Topic CEO (economic news) has *buget*, *tax*, *spending*, *deficit*, and *congress*.

It also interesting to observe the effect of applying morphological transformation and word grouping. Comparing topic AIR in Tables 4.6 and 4.7, we find that *airlines* and *airline* in Table 4.6 are merged into one term *airline* in Table 4.7 through morphological transformation, and terms *air* and *force* of topic ARO (aerospace) in Table 4.6 are grouped as *air_force* in Table 4.8 through word grouping. This change also makes *contract*, *air_force*, *navy*, *aircraft*, and *army* the top 5 terms of topic ARO in Table 4.8, indicating clearly that most of the texts in topic ARO are about military aircraft contracts. In our experiments, we use the top 300 terms from each test set as topic signatures. The effect of using fewer terms is discussed in Section 4.6.3.

Full topic signatures are listed in Appendix B.

4.5.3 Constructing Confusion Sets

We use Equation 4.9 in Section 4.4.4 and the 32 topic signatures generated using the procedures described in the previous section to compute confusion sets CFS_i and

R	air	aro	aut	bbk
1	airlines	contract	gm	shares
2	passenger	aircraft	cars	debentures
3	airline	navy	ford	common
4	air	air	chrysler	redemption
5	continental	force	auto	outstanding

R	bcy	bnk	bon	ceo
1	bankruptcy	bank	bonds	mr.
2	chapter	banks	issues	quarter
3	wedtech	mr.	bond	share
4	11	banking	issue	expects
5	creditors	loan	debt	cents

R	cmd	div	eco	edp
1	farmers	dividend	budget	ibm
2	farm	split	tax	computer
3	says	payable	spending	computers
4	agriculture	record	deficit	software
5	crop	stock	congress	machines

R	ele	env	ern	fab
1	semiconductor	epa	quarter	says
2	superconductors	waste	net	mr.
3	chips	environmental	million	p&g
4	superconductivity	water	loss	tobacco
5	chip	ozone	share	smoking

R	fin	lng	min	mkt
1	says	gas	steel	ad
2	insurance	pipeline	tons	advertising
3	lawyers	natural	week	says
4	tax	pipelines	capability	thompson
5	law	coastal	usx	saatchi

R	mon	pet	pha	pub
1	trade	texaco	drug	mr.
2	u.s.	oil	patients	magazine
3	japan	pennzoil	aids	says
4	billion	barrels	dr.	editor
5	japanese	getty	drugs	newspaper

R	rel	ret	scr	stk
1	says	stores	securities	shares
2	estate	sales	mr.	offering
3	real	+	firm	stock
4	land	retailers	kidder	common
5	hotel	store	firms	proceeds

R	tel	tnm	tra	uti
1	at&t	shares	rail	utility
2	fcc	filing	railroad	power
3	telephone	group	highway	utilities
4	network	stake	conrail	electric
5	cbs	stock	railroads	commission

Table 4.6: Top 5 terms of each topic signature in set **WD**, the unaltered input words.

R	air	aro	aut	bbk
1	airline	contract	car	share
2	passenger	aircraft	gm	debenture
3	mile	navy	ford	redeem
4	flight	air	chrysler	common
5	air	force	auto	redemption
R	**bcy**	**bnk**	**bon**	**ceo**
1	bankruptcy	bank	bonds	mr.
2	chapter	loan	issue	quarter
3	creditor	mr.	rating	cent
4	wedtech	thrift	bond	earnings
5	11	banking	market	net
R	**cmd**	**div**	**eco**	**edp**
1	farmer	dividend	budget	ibm
2	farm	split	tax	computer
3	crop	payable	spending	software
4	agriculture	stock	deficit	machine
5	grain	record	congress	apple
R	**ele**	**env**	**ern**	**fab**
1	chip	epa	quarter	restaurant
2	semiconductor	waste	net	mr.
3	superconductors	environmental	million	food
4	superconductivity	water	loss	wine
5	intel	state	cent	p&g
R	**fin**	**lng**	**min**	**mkt**
1	lawyer	gas	steel	ad
2	tax	pipeline	tons	advertising
3	firm	natural	week	thompson
4	insurance	coastal	usx	saatchi
5	law	foot	capability	jwt
R	**mon**	**pet**	**pha**	**pub**
1	trade	texaco	drug	mr.
2	u.s.	oil	patient	magazine
3	export	pennzoil	aids	book
4	japan	court	dr.	editor
5	billion	barrels	hospital	newspaper
R	**rel**	**ret**	**scr**	**stk**
1	hotel	stores	firm	share
2	building	+	mr.	stock
3	estate	retailer	security	offering
4	property	sale	kidder	common
5	real	store	broker	underwriter
R	**tel**	**tnm**	**tra**	**uti**
1	at&t	share	railroad	utility
2	network	filing	highway	power
3	fcc	stake	rail	rate
4	telephone	group	ship	electric
5	cbs	acquire	conrail	commission

Table 4.7: Top 5 terms of each topic signature in set **TR**, the morphologically normalized words. Notice that *superconductors* in topic ELE is not transformed into *superconductor*, since *superconductor* is not in WordNet.

R	air	aro	aut	bbk
1	airline	contract	car	share
2	passenger	air_force	gm	debenture
3	mile	navy	ford	redeem
4	flight	aircraft	chrysler	redemption
5	air	army	motor	outstanding

R	bcy	bnk	bon	ceo
1	bankruptcy	bank	bond	mr.
2	chapter	mr.	issue	cent
3	creditor	thrift	rating	quarter
4	wedtech	loan	market	earnings
5	11	banking	debt	expect

R	cmd	div	eco	edp
1	farmer	dividend	budget	ibm
2	crop	payable	tax	computer
3	farm	stock_of_record	spending	machine
4	grower	quarterly	deficit	software
5	grain	declare	congress	personal_computer

R	ele	env	ern	fab
1	chip	epa	loss	restaurant
2	semiconductor	waste	quarter	mr.
3	superconductors	environmental	million	p&g
4	superconductivity	water	cent	food
5	intel	ozone	share	brand

R	fin	lng	min	mkt
1	lawyer	pipeline	steel	ad
2	tax	gas	ton	advertising
3	welfare	natural_gas	week	thompson
4	state	coastal	capability	saatchi
5	firm	cubic_foot	usx	jwt

R	mon	pet	pha	pub
1	trade	texaco	drug	mr.
2	u.s.	pennzoil	patient	magazine
3	export	oil	dr.	book
4	japan	barrel	aid	newspaper
5	billion	getty	hospital	editor

R	rel	ret	scr	stk
1	hotel	store	firm	stock
2	real_estate	+	mr.	offering
3	property	retailer	security	share
4	building	sale	kidder	underwriter
5	city	sears	broker	proceeds

R	tel	tnm	tra	uti
1	at&t	share	railroad	utility
2	network	filing	highway	power
3	fcc	stake	rail	rate
4	cbs	group	union	electric
5	mr.	acquire	conrail	commission

Table 4.8: Top 5 terms of each topic signature in set **PH**, the words joined into phrases if given in WordNet. Of the 160 terms here, 6 are multi-word phrases.

TOPIC	MAX	θ	Confusion Set
air	0.153	0.155	$\emptyset$
aro	0.154	0.160	$\emptyset$
aut	0.158	0.199	$\emptyset$
bbk	0.450	0.196	{tnm/0.450, stk/0.447, div/0.344, bon/0.315, ern/0.220, ceo/0.212}
bcy	0.225	0.212	{fin/0.224, tnm/0.217}
bnk	0.305	0.289	{scr/0.305}
bon	0.346	0.203	{stk/0.346, bbk/0.315, scr/0.208}
ceo	0.811	0.296	{ern/0.811, tnm/0.346, div/0.304}
cmd	0.231	0.199	{mon/0.231, fin/0.209, eco/0.204}
div	0.344	0.140	{bbk/0.344, tnm/0.311, ceo/0.304, stk/0.301, ern/0.290}
eco	0.408	0.279	{mon/0.408, fin/0.316}
edp	0.287	0.210	{ele/0.287}
ele	0.287	0.193	{edp/0.287, mon/0.211}
env	0.227	0.245	$\emptyset$
ern	0.811	0.265	{ceo/0.811, tnm/0.349, div/0.290, stk/0.266}
fab	0.283	0.295	$\emptyset$
fin	0.329	0.387	$\emptyset$
lng	0.214	0.128	{uti/0.214, pet/0.187, tnm/0.136}
min	0.151	0.174	$\emptyset$
mkt	0.288	0.220	{fin/0.288, fab/0.283, pub/0.254, scr/0.221}
mon	0.408	0.280	{eco/0.408}
pet	0.188	0.196	$\emptyset$
pha	0.195	0.162	{fab/0.195, fin/0.170}
pub	0.255	0.242	{mkt/0.255}
rel	0.259	0.265	$\emptyset$
ret	0.240	0.235	{ceo/0.240}
scr	0.329	0.361	$\emptyset$
stk	0.486	0.297	{tnm/0.486, bbk/0.447, bon/0.345, div/0.301}
tel	0.221	0.244	$\emptyset$
tnm	0.486	0.391	{stk/0.486, bbk/0.450}
tra	0.250	0.245	{eco/0.250}
uti	0.214	0.176	{lng/0.214}

Table 4.9: Maximum, outlier threshold, and confusion set for each topic used in the **PH** test set.

outlier threshold θ_i for each topic i. The procedure for identifying confusion sets for each topic i is as follows:

STEP 1: Create 32 topic signatures as described in Section 4.5.2.2

STEP 2: Compute pairwise similarity sim_{ij} between any two topic signatures (for topics i and j), using Equation 4.9

STEP 3: Compute max_i and θ_i for each topic i from the results of the previous step

STEP 4: If $max_i > \theta_i$ and $sim_{ij} > \theta_i$, place topic j, $j \neq i$, into topic i's confusion set

Table 4.9 shows maximim similarity max_i, outlier threshold θ, and confusion sets for 32 topics of the **PH** test set used in our experiments. It is interesting to observe that

we can indeed use confusion sets to identify closely related topics and that $21/32 = 66\%$ actually required this. For example, topic LNG (natural gas, pipelines) is related to topics UTI (utilities), PET (petroleum), and TNM (tender offers, mergers, acquisitions). The inclusion of TNM hints that texts of topic LNG may contain tender and merger offers among natural gas companies.

Note that the confusion sets for any two topics are not symmetric. For example, the confusion set of RET includes CEO, but not vice versa. The asymmetry results from the method which we use to generate the confusion sets. The inclusion of topic X in the confusion set of topic Y is not only determined by the similarity between them, but the outlier threshold θ of topic Y. The outlier threshold θ is determined by the distribution of the similarity between topic Y and all the other topics.

According to Table 4.9, topics CEO (Dow Jones interview) and ERN (earnings) are very similar: They have similarity of 0.811, and of the top 20 terms in each, only one is *not* shared! Such a high similarity predicts our topic identification module will have difficulty in distinguishing texts of CEO from ERN. Table 4.10 lists the top 120 terms of topic signatures CEO and ERN. The top 15 terms of topic ERN can also be found in the top 34 terms of topic CEO. As discussed earlier, this phenomenon requires the use of *multi-level* signatures, which facilitate further specialization of topic signatures within confusion sets.

Figure 4.8 further details distributions of similarity among topics. The ends of the whiskers mark the minimum and maximum similarity, the boxes mark the outlier thresholds[12], the notches mark 1st and 3rd quartiles, and the center of the notches mark medians. Cross marks on the whiskers are similarity value between the topic label on the horizontal axis and other topics. Where the whiskers are extended beyond the boxes, confusion sets exist for the corresponding topics.

4.5.4 Building Second Level Topic Signatures

Although the algorithm introduced in Section 4.4.4 describes how multi-level signatures can be constructed, we needed to use only second level topic signatures in our experiments. Deriving second level topic signatures from confusion sets is straightforward:

[12]The lower box edges mark the values 1.5 times interquartile range from the first quartiles.

CEO Topic signature (PH)							
1	**mr.**$_{66}$	31	fourth	61	second	91	late
2	**cent**$_{4}$	32	end	62	be_about	92	say
3	**quarter**$_{2}$	33	fourth-quarter	63	range	93	$1
4	**earnings**$_{6}$	34	earlier	64	operate	94	gold
5	**expect**$_{56}$	35	acquisition	65	new	95	unit
6	**fiscal**$_{23}$	36	result	66	20%	96	computer
7	**million**$_{3}$	37	business	67	price	97	average
8	**share**$_{5}$	38	gain	68	30	98	restate
9	**sale**$_{9}$	39	per-share	69	forecast	99	inc.
10	**year**$_{15}$	40	operation	70	reflect	100	project
11	**revenue**$_{12}$	41	1988	71	decline	101	ounce
12	**profit**$_{10}$	42	growth	72	third	102	industry
13	**net_income**$_{11}$	43	period	73	year_end	103	expand
14	**net**$_{8}$	44	1985	74	month	104	improve
15	**1987**$_{51}$	45	year-ago	75	nine	105	backlog
16	interview	46	chairman	76	double	106	recent
17	1986	47	post	77	projection	107	current
18	year-earlier	48	president	78	cost	108	dividend
19	company	49	31	79	make	109	pre-tax
20	rise	50	fiscal_year	80	grow	110	at_least
21	report	51	continue	81	charge	111	improvement
22	estimate	52	increase_in	82	total	112	profit_from
23	chief_executive_officer	53	strong	83	indicate	113	six
24	analysts'	54	store	84	billion	114	ago
25	executive_officer	55	more_than	85	maker	115	exceed
26	increase	56	high	86	first-quarter	116	predict
27	earn	57	add	87	restaurant	117	25%
28	compare	58	first_quarter	88	plan	118	figure
29	loss	59	second-quarter	89	10%	119	expansion
30	product	60	market	90	attribute	120	15%

ERN Topic signature (PH)							
1	**loss**	31	discontinue	61	asset	91	yesterday
2	**quarter**$_{3}$	32	end	62	1985	92	fiscal_year
3	**million**$_{7}$	33	fall	63	strong	93	six
4	**cent**$_{2}$	34	first-quarter	64	increase_in	94	jump
5	**share**$_{8}$	35	increase	65	cost	95	double
6	**earnings**$_{4}$	36	fourth	66	mr.$_{1}$	96	provision
7	rise	37	third-quarter	67	pretax	97	growth
8	**net**$_{14}$	38	late	68	second	98	related_to
9	**sale**$_{9}$	39	continue	69	expense	99	investment
10	**profit**$_{12}$	40	second-quarter	70	income	100	low
11	**net_income**$_{13}$	41	write-down	71	new_york_stock_exchange	101	cite
12	**revenue**$_{11}$	42	first_quarter	72	close	102	attribute
13	year-earlier	43	31	73	per-share	103	one-time
14	earlier	44	decline	74	total	104	tiger
15	**year**$_{10}$	45	third	75	reserve	105	first_half
16	billion	46	month	76	stock_exchange	106	unit
17	compare	47	business	77	composite	107	more_than
18	operation	48	profit_from	78	earn	108	bank
19	report	49	nine	79	operating	109	insurance
20	result	50	product	80	chairman	110	tax_credit
21	gain	51	1987$_{15}$	81	improve	111	store
22	fourth-quarter	52	analyst	82	debt	112	maker
23	**fiscal**$_{6}$	53	restate	83	computer	113	new
24	company	54	year-ago	84	pre-tax	114	extraordinary
25	period	55	restructuring	85	chief_executive_officer	115	hurt
26	post	56	expect$_{5}$	86	inc.	116	previously
27	1986	57	trading	87	executive_officer	117	market
28	loan	58	30	88	figure	118	interest
29	charge	59	operate	89	corp.	119	toy
30	reflect	60	high	90	dollar	120	group

Table 4.10: Top 120 terms of **CEO** and **ERN** topic signatures in test set **PH**. Note that the top 15 terms of **CEO** are marked with subscripts of term ranks in topic **ERN**.

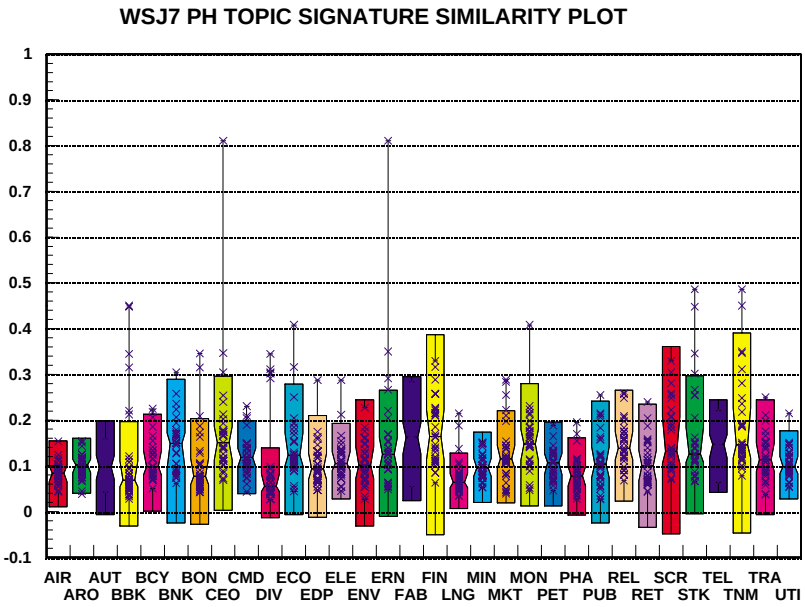


Figure 4.8: Distributions of similarity among topic signatures.

STEP 1: Create a confusion set for each topic, as stated in the previous section.

STEP 2: Compute *inverse topic signature frequency* to measure in how many topic signatures a term occurs.

STEP 3: Compute new term weight, following Equation 4.13.

Table 4.11 shows second level topic signatures of topic CEO. Only topic signatures of CEO and ERN are presented, for comparison with the first level topic signatures shown in Table 4.10. Although some overlap still occurs, the amount in the top 15 terms has dropped dramatically. Among the top 15 terms of the second level topic signature of topic ERN, six of them, *pretax*, *reserve*, *discontinue*, *fall*, *provision*, and *related_to*, do not apper at all in the top 120 terms of the second level topic signature of topic CEO, and the rest are found in the top 71 terms of topic CEO. Only 5 out of the top 15 terms[13] of the first level topic signature stay in the top 120 terms of the second level topic signature of topic ERN. This term

[13]They can be found in the top 23 terms of the second level signature.

rearrangement and drop-out in the second level topic signatures indicates their improved focus on the topic differences in confusion sets. The effectiveness of second level topic signatures is evaluated in Section 4.6.2.

4.6 Evaluations

In this section, we present detailed evaluations of the quality of topic signatures constructed from different term treatments, the application of the second level topic signatures, the performance change relative to the number of terms used in the topic signatures, and the effect of disparities in the number of texts per training topic.

4.6.1 Evaluation of Topic Signatures

Evaluating the effectiveness of topic signatures and topic assignment algorithm is straightforward. It involves the following:

STEP 1: Create document signature (DS_i) for each test document i following Equations 4.2 and 4.3 in Section 4.4.1,

STEP 2: Create topic signature (TS_j) for each topic j from the training corpus, as described in Section 4.5.2.2,

STEP 3: for each topic signature and document pair, compute their similarity sim_{ij} using Equation 4.8,

STEP 4: for each document, sort the similarity values obtained in the previous step, and assign the topic with the highest similarity to the corresponding document,

STEP 5: collect *hit* h_j, *fault* f_j, and *miss* m_j counts for each topic j, and compute *recall* R_j and *precision* P_j as follows:

$$R_j = \frac{h_j}{h_j + m_j} \tag{4.15}$$

$$P_j = \frac{h_j}{h_j + f_j} \tag{4.16}$$

where h_j is the number of documents assigned as topic j that are indeed documents of topic j, m_j is the number of documents which

should be assigned as topic j but are not, and f_i is the number of documents assigned as topic j but should not be.

The results, using three different term treatments on topic signature construction, are summarized in Tables 4.12 and 4.13. Table 4.12 is based on training texts and Table 4.13 on unseen texts. It is clear that topic signatures using terms verbatim (**WD**) from texts provide best performance; topic signatures of terms with morphological normalization (**TR**) provide the worst precision score; and topic signatures of terms with morphological normalization and phrases recorded in WordNet (**PH**) achieved medium performance. According to this results, we prefer the **PH** term treatment, since it includes more meaningful term representation (*air* and *force* are grouped into *air_force*) and performance comparable to the **WD** term treatment. To validate this result, we performed the same test on texts of 31 topics from the *Wall Street Journal* 1988 collection, which were unseen in the previous test. The results are shown in Table 4.13; they conform to the results obtained from the 1987 collection.

Figures 4.9 and 4.10 show the recall and precision scores of each topic of the **PH** sets of the *Wall Street Journal* 1987 and 1988 collections respectively. Detailed results of *hit*, *fault*, *miss*, *recall*, and *precision* for each topic are provided in Appendix C.1. It is clear that topic signatures generated from the training set perform well for most of the topics in the training and test sets: most of them fall in the upper right quarters of the recall/precision graph. The recall and precision scores are more concentrated in the training set than the test set, since the topic signatures are constructed solely from the training set. Note that topics LNG and RET did not perform well in either the training or test set. Their low precision scores result from too many times that other topics are mistakenly assigned topic LNG or RET.

Table 4.14 shows the top 6 most possible fault candidates for each topic. It is interesting that the most common fault candidate for 25 out of 32 (78%) topics is TNM. The number of texts singly indexed as TNM in the training set is 4,650 (29% of total training texts referring to Table 4.1). Most of them are short texts: the average number of terms per text in topic TNM is 95 words according to Table A.3, which is much shorter than the average length of texts in other topics (overall average is 232). We discuss such disparities in Section 4.6.4.

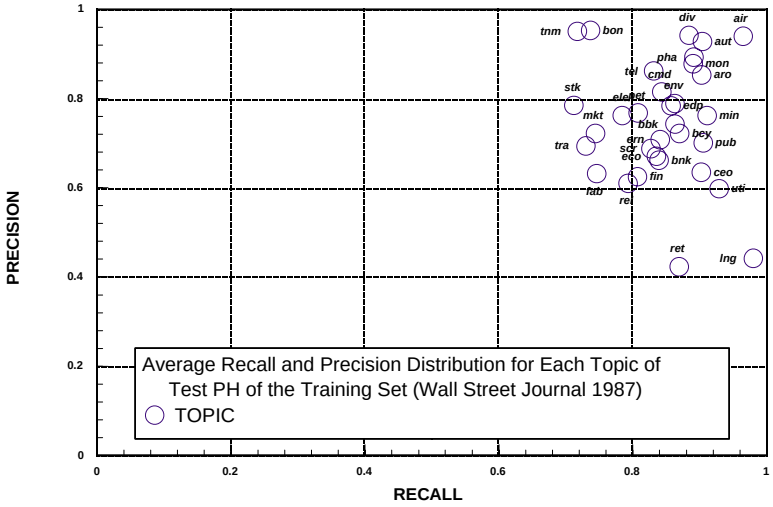


Figure 4.9: Average recall and precision distribution of test **PH** of the training set.

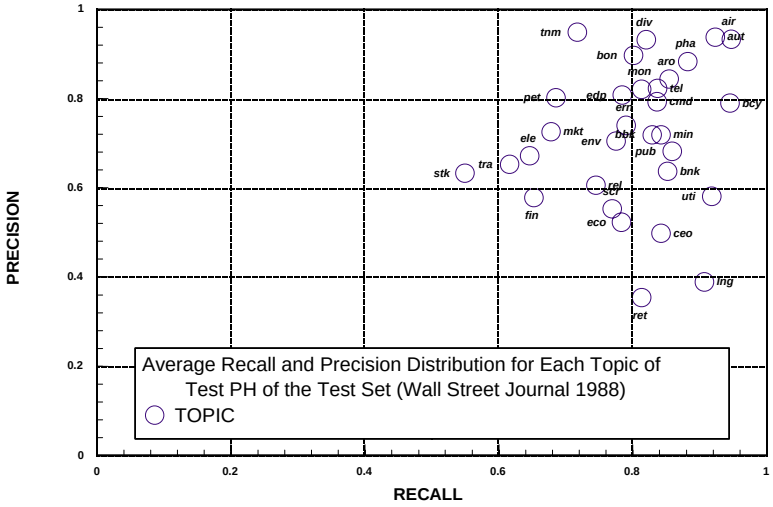


Figure 4.10: Average recall and precision distribution of test **PH** of the test set.

Figure 4.11 shows two texts from topic TNM. Text (1) and (2) were assigned to topics LNG and RET respectively by the topic identifcation algorithm using topic signatures. It is clear that these texts describe *tender-and-merger* events about *natural gas* and *retail* industries. Key terms such as *gas* and *natural gas* in text (1), and *store* and *retailer* in text (2), are terms ranked in the top 5 positions (Tables B.18 and B.26 in Appendix B). On the other hand, terms such as *company*, *sale* (also shared by RET), *acquisition*, *unit*, and so on, are among the top 15 terms in TNM (Table B.30). Although the short message style of TNM texts contributes to the difficulty of topic assignment[14], these examples indicate the need of finer signatures and the existence of multiple topics in texts. Therefore, labeling text (1) as LNG and text (2) as RET is not wrong, but is not complete. One possible solution for this problem is to assign multiple topics to a text whenever the difference of similarity values among competing topics is less than an experimentally determined threshold. Nevertheless, we would like to develop a complete solution for this interesting problem in future research.

4.6.2 Evaluation of Second Level Topic Signatures

Although we achieved very good performance using the first level topic signatures alone, it is still very interesting to know how much the second level topic signatures will help. Table 4.15 shows the results of four tests. Comparing with Tables 4.12 and 4.13, the use of the second level topic signatures did improve precision, but recall degraded a little. However, no dramatic performance increase is obtained. This may be due to the already very good result of just the first level topic signatures. We plan to apply and test the multi-level topic signature topic identification algorthm in the more complex TREC routing task [27] in the future. Detail by topic listings of the *hit*, *fault*, *miss*, *precision*, and *recall* scores are given in Appendix C.2.

[14]The topic identification process relies on only small number of terms, which is not a good example of utilization of concept co-occurrence.

(1) WSJ861203-0146

MONTREAL Noverco Inc. said it agreed to buy the parent of Vermont Gas$_2$ Systems Inc. , a natural gas$_3$ utility based in Burlington , Vt. , for $10.5 million (U.S.) . The acquisition would be the first in the U.S. for Noverco , which is the parent of Gaz Metropolitain Inc. , Quebec 's dominant natural gas$_3$ supplier . Noverco said it will purchase Energy Future L.P. , which in turn owns New England Gas$_2$ Corp. , the parent of Vermont Gas$_2$. Vermont Gas$_2$, which has the exclusive right to distribute natural gas$_3$ in Vermont , earned $1.5 million in 1985 on revenue of $25 million .

(2) WSJ870325-0125

HOUSTON Gordon Jewelry Corp. said it completed the previously announced sale$_4$ of its Catalog-Showroom Stores$_1$ unit for cash and notes to Carlisle Capital Corp. , a closely held Boston-based investment company . The price wasn't disclosed , but it exceeded the book value of the assets being sold , the jewelry retailer$_3$ said .

Figure 4.11: Sample texts from topic TNM which are assigned to topics LNG (1) and RET (2) respectively by the topic identification algorithm. The subscripts indicate the corresponding term ranks in the topic signatures of LNG and RET (Tables B.18 and B.26 in Appendix B).

4.6.3 How Many Terms per Signature?

As mentioned in Section 4.5.2.2, each topic signature consists of 300 top-ranked terms from its corresponding text collection. Can we use fewer terms? What would be the performance loss or gain due to variations in the number of terms used in topic signatures? We address these questions in this section.

To test the effect of using various number of terms in topic signatures, we contructed 60 partial topic signatures from each 300-term signature by using the top 5, 10, 15, ..., 295, 300 (i.e., a 5-term increment) terms from the full length signature. We then performed topic identification using these partial topic signatures, following the procedure described in Section 4.6.1. The results are shown in Figure 4.12; detailed numbers are tabulated in Table 4.16. It is clear that the recall and precision scores improve quickly when more terms are used in the topic signatures, although this rate slows down when the number of terms per topic signature approaches 300.

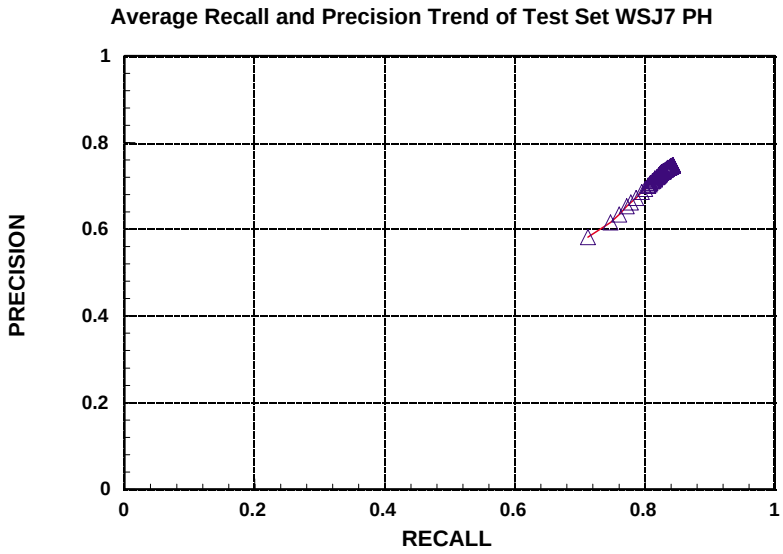


Figure 4.12: Recall and precision trend graph using various number of terms as topic signature (the *Wall Street Journal* 1987 training texts with phrases (**PH**)). This graph shows an increase of both recall and precision when the number of terms per topic signature is increased (center to upper right corner).

This result indicates that using number of terms approximately equal to the third quartile term number is a good choice (see Table 4.5).

4.6.4 *Idf* Normalization

In this section we discuss the issue of disparity of number of texts in each training topics.

The standard *idf* is computed by Equation 4.1 in Section 4.4. We include it here for convenience:

$$idf_j = \log \frac{N}{df_j}$$

In this equation, no concept of prespecified topic categories is involved: the counts relate only to the total number of texts in the training collection and the number of times a term occurs in different documents. Since topic categories are already known during the training phase, and the number of texts in each topic varies widely (see Tables 4.1 and 4.2 in Section 4.6.1), we should modify the original *idf* equation by

including the number of texts per topic as a normalization factor. The idea is: (1) to avoid *idf* value being overfitted to certain topics because some topics have more texts than others, and (2) to use all the available training texts instead of wasting valuable training texts by only using part of them. For example, if term t_j appears in every text of topic TNM, but it rarely occurs in any text of other topics. The unnormalized *idf* of term t_j will be penalized by the much larger number of available training texts in topic TNM, because term t_j appears in about 28% of the training text according to Table 4.1. We made the following adjustment to Equation 4.1:

1. Select a virtual number of texts M. The median number of texts per topic is a good choice, since half of the topics in the training set have number of texts per topic larger than the median, and the other half have less.

2. Compute new $df_j^{'}$ in Equation 4.1 by $\sum_k df_{jk} \cdot M/N_k$, where df_{jk} is the number of documents in which term j appear in topic k and N_k is the number of training texts in topic k. Doing this assumes that if we have M texts of topic K instead of N_k, the distribution of term j will remain the same.

3. Compute the new total number of documents N_t. Assuming each topic containing M texts and total number of topics is T, N_t is equal to $M \times T$.

4. Use the new $idf_j^{'}$ $(= \log[N_t/(df_j^{'}])$ in place of idf_j, and perform the topic identification task accordingly.

Using normalized *idf* for training and test sets are shown in Table 4.17. It is clear that the identification results are both improved in recall and precision. Detailed numbers are provided in Appendix C.3.

4.7 Conclusions

This chapter describes a method to identify topics using topic signatures. Topic signatures extend the traditional $tf * idf$ term weighting method introduced in Section 3.2 by utilizing concept co-occurrence information. The concepts of using mutlilevel topic signatures based on confusion sets and normalized *idf* are also introduced. The results of using normalized *idf* based topic signatures achieved recall and precision scores of 0.848 and 0.764 on the training set and 0.802 and 0.729 on the test set;

this compares very well to the average results of the TREC routing experiment[15], whose participants scored in the 0.5 to 0.6 range. The reasons for this disparity are not clear, but we surmise that they relate to the relative purity of our training set data, and the fact that the TREC routing task has considerably more complex topics. We plan to apply our experience on this investigation to future TREC routing tasks.

The results of our evaluations show clearly that texts can be categorized into different topics. Within the setup of our experiments that used 32 topics from the *Wall Street Journal*, the different nature of topics (e.g., "Tender and Merger", and "Bankrupcy" vs. "Airlines" or "Utilities") will affect the topic identification results. It is very important for a topic identification task to respect the difference in category/topic functions. Some categories can apply to any industries, while others are industries upon whom anything can happen. This phenomenon is fully illustrated in the two examples presented in Section 4.6.1. Therefore, meta-level knowledge about topics is needed to improve the topic identification task. If this kind of knowlege is available, then the two TNM texts can be very well described as texts about tender and merger in the natural gas or retail industries, a much better categorization than just picking the topics with highest similarity.

How to use topic signatures to augment or update existing knowledge bases such as WordNet is also very interesting future reseach. Building topic signatures according to training corpus is likely to discover new or changed relations among concepts. Finding the method to link these newly identified relations back to existing knowlege base would be very beneficial in improving the capability of our topic identification system.

In this and previous chapters, we described the use of $tf * idf$ term frequency, concept counting, and concept co-occurrence to identify topics. We conclude the exploration of frequency-based methods here. In the next chapter, we discuss the use of the regularities of discourse structure in certain genres to identify topics.

[15] Routing task is one kind of text categorization tasks.

CEO 2nd Level Topic Signature (PH)							
1	interview	31	comfortable	61	plastic	91	attribute
2	estimate	32	order	62	anticipate	92	60
3	analysts'	33	high	63	cost	93	previous
4	year-earlier	34	1987	64	$1.1	94	late
5	earnings	35	specific	65	earn	95	roughly
6	be_about	36	contract	66	$1.3	96	in_addition
7	range	37	second-quarter	67	$1.5	97	7%
8	compare	38	$5	68	help	98	item
9	forecast	39	customer	69	grow	99	average
10	fiscal	40	area	70	motel	100	capacity
11	fourth	41	electronics	71	charge	101	spending
12	projection	42	huffy	72	begin	102	restate
13	fourth-quarter	43	rise	73	taylor	103	5%
14	gold	44	40%	74	hecla	104	worthington
15	per-share	45	adjust	75	mattress	105	70
16	project	46	executive	76	better	106	open
17	growth	47	machine	77	$4	107	chief_operating_officer
18	ounce	48	30%	78	$1.6	108	supply
19	expand	49	introduce	79	product_line	109	homestake
20	backlog	50	third	80	around	110	a_little
21	profit	51	go	81	indicate	111	battle
22	period	52	a.	82	automotive	112	look
23	net_income	53	conger	83	center	113	recent
24	at_least	54	financial_officer	84	$2.5	114	earlier
25	1985	55	see	85	small	115	pre-tax
26	year-ago	56	double	86	$30	116	improvement
27	exceed	57	public	87	restaurant	117	profit_from
28	post	58	triple	88	$6	118	ago
29	expansion	59	dilute	89	m.	119	acquisition
30	fiscal_year	60	micropro	90	thrift	120	predict

ERN 2nd Level Topic Signature (PH)							
1	year-earlier	31	climb	61	loan-loss	91	marketing
2	compare	32	12%	62	operating	92	surge
3	fourth-quarter	33	restate	63	nine-month	93	insurance
4	period	34	year-ago	64	11%	94	big
5	post	35	coleco	65	pre-tax	95	mattel
6	pretax	36	drop_in	66	figure	96	tax_credit
7	charge	37	high	67	19%	97	varity
8	reserve	38	result_in	68	widen	98	$2.2
9	earnings	39	result	69	dollar	99	note
10	rise	40	1985	70	14%	100	early
11	discontinue	41	cannon	71	fiscal_year	101	24%
12	fall	42	13%	72	mcorp	102	show
13	profit	43	cost	73	jump	103	domestic
14	provision	44	inventory	74	double	104	personal_computer
15	related_to	45	microsoft	75	growth	105	substantial
16	fourth	46	offset	76	31%	106	extraordinary
17	third-quarter	47	nonperforming	77	profitability	107	hurt
18	tiger	48	16%	78	sharply	108	cpc
19	net_income	49	gain	79	low	109	effect
20	late	50	greyhound	80	nonrecurring	110	first-quarter
21	second-quarter	51	fall_to	81	problem	111	tandy
22	toy	52	fiscal	82	norfolk	112	exclude
23	earlier	53	expense	83	attribute	113	non-accrual
24	accounting	54	income	84	krone	114	17%
25	yen	55	per-share	85	one-time	115	northrop
26	drop	56	kronor	86	first_half	116	shipment
27	after-tax	57	convergent	87	reflect	117	mainly
28	write-down	58	associate_with	88	southern	118	non-interest
29	third	59	commodore	89	gca	119	tiffany
30	profit_from	60	credit	90	gould	120	as_well

Table 4.11: Top 120 terms of **CEO** and **ERN** second level topic signatures in test set **PH**.

TEST	RECALL	PRECISION
WD	0.847	0.752
TR	0.844	0.739
PH	0.843	0.748

Table 4.12: Summary of average recall and precision scores tested on the *Wall Street Journal* 1987 training collection (16,137 texts) with three different term treatments: words without modification (**WD**), words with morphological normalization (**TR**), and words with morphological normalizaion and phrases recorded in WordNet (**PH**).

TEST	RECALL	PRECISION
WD	0.803	0.719
TR	0.802	0.710
PH	0.797	0.716

Table 4.13: Summary of average recall and precision scores tested on the *Wall Street Journal* 1988 test collection (12,906 texts) with three different term treatments.

TOPIC	P	T	1	2	3	4	5	6
air	0.940	46	tnm/26	aro/5	bon/3	tra/3	bbk/1	bcy/1
aro	0.853	58	tnm/15	tra/10	edp/5	aut/4	tel/4	air/3
aut	0.928	47	tnm/26	tra/4	ern/3	mon/3	fin/2	min/2
bbk	0.743	283	bon/136	tnm/99	stk/40	div/6	fab/1	rel/1
bcy	0.722	54	tnm/18	mon/4	bbk/3	bon/3	pha/3	tra/3
bnk	0.662	197	tnm/136	bon/9	scr/8	div/5	fin/5	mon/5
bon	0.953	33	bbk/19	scr/4	fin/3	bnk/2	tnm/2	mon/1
ceo	0.635	221	tnm/79	ern/61	bbk/15	aut/10	div/6	edp/5
cmd	0.815	22	tnm/11	fab/5	mon/2	rel/2	env/1	mkt/1
div	0.942	36	stk/11	bbk/7	ern/6	tnm/5	bon/2	air/1
eco	0.671	119	mon/38	tnm/12	fin/11	env/7	bnk/6	air/5
edp	0.784	71	tnm/44	ern/8	ele/5	tel/5	ceo/4	div/2
ele	0.762	34	tnm/18	pha/5	edp/3	aut/2	stk/2	aro/1
env	0.789	32	tnm/13	cmd/3	aut/2	fab/2	mon/2	pha/2
ern	0.708	243	tnm/121	ceo/18	bnk/13	aut/9	div/9	fab/8
fab	0.632	91	tnm/63	mkt/9	pha/7	ceo/2	cmd/2	air/1
fin	0.625	143	tnm/35	eco/15	pha/13	edp/9	stk/9	bnk/6
lng	0.442	126	tnm/90	bbk/7	pet/7	bon/3	ceo/3	div/3
min	0.762	64	tnm/47	mon/6	rel/3	eco/2	bcy/1	ern/1
mkt	0.722	37	tnm/15	tel/8	pub/3	aut/2	fab/2	scr/2
mon	0.878	103	aut/15	bnk/12	eco/10	tnm/10	aro/8	tel/8
pet	0.768	42	tnm/28	bcy/3	mon/2	aro/1	bnk/1	div/1
pha	0.893	52	tnm/25	fab/7	mon/4	bbk/2	bon/2	cmd/2
pub	0.701	81	tnm/54	tel/7	bnk/2	eco/2	edp/2	fin/2
rel	0.610	103	tnm/70	mon/3	aut/2	bon/2	ern/2	fab/2
ret	0.423	127	tnm/91	bbk/7	ern/7	eco/5	bon/3	stk/3
scr	0.688	100	tnm/33	bnk/10	bon/8	fin/8	stk/5	tel/5
stk	0.785	96	bon/43	tnm/34	bbk/5	div/5	fin/4	cmd/1
tel	0.862	58	tnm/41	edp/4	bon/3	bnk/2	air/1	ele/1
tnm	0.951	171	bbk/47	stk/32	div/29	bnk/11	rel/9	scr/9
tra	0.694	44	tnm/20	aut/5	fab/4	pet/2	air/1	aro/1
uti	0.598	88	tnm/33	tel/11	bon/9	stk/7	bbk/6	div/6

Table 4.14: The top most common fault candidates for each topic of test set **PH** of the training set (Wall Street Journal 1987). Each fault candidate is paired with the number of faults occurring in the topic identification process. The P column lists the precision score for each topic, and the T column lists the total number of faults occurring for a topic.

TEST	RECALL	PRECISION
7WD1b	0.846	0.757
7TR1b	0.843	0.742
7PH1b	0.842	0.752
8PH1b	0.792	0.718

TEST CODE:

7:	1987 WSJ texts, i.e., training set
8:	1988 WSJ texts, i.e., test set
WD:	texts without morphological transformation and word grouping
TR:	texts with morphological transformation but not word grouping
PH:	texts with morphological transformation and word grouping
1b:	using use cosine similarity measure and second level topic signatures

Table 4.15: Second-level signatures: average recall and precision scores for training and test sets.

# of TERMS	RECALL	PRECISION
5	0.660	0.553
10	0.712	0.582
15	0.747	0.616
20	0.760	0.634
25	0.772	0.653
30	0.778	0.662
35	0.786	0.672
40	0.795	0.686
45	0.800	0.693
50	0.805	0.700
55	0.805	0.701
60	0.807	0.703
65	0.808	0.707
70	0.811	0.710
75	0.813	0.712
80	0.815	0.715
85	0.815	0.717
90	0.818	0.720
95	0.820	0.722
100	0.821	0.723
105	0.822	0.726
110	0.822	0.726
115	0.823	0.728
120	0.824	0.729
125	0.825	0.731
130	0.826	0.732
135	0.827	0.732
140	0.828	0.733
145	0.828	0.733
150	0.828	0.734
155	0.829	0.734
160	0.829	0.735
165	0.829	0.735
170	0.831	0.736
175	0.832	0.737
180	0.833	0.739
185	0.833	0.739
190	0.834	0.739
195	0.834	0.740
200	0.835	0.742
205	0.835	0.742
210	0.836	0.742
215	0.837	0.742
220	0.838	0.743
225	0.839	0.744
230	0.839	0.745
235	0.840	0.745
240	0.840	0.745
245	0.840	0.745
250	0.841	0.746
255	0.841	0.746
260	0.841	0.746
265	0.842	0.746
270	0.842	0.747
275	0.843	0.748
280	0.842	0.747
285	0.843	0.747
290	0.843	0.747
295	0.843	0.748
300	0.843	0.748

Table 4.16: Recall and precision trends using different numbers of terms as topic signatures (the *Wall Street Journal* 1987 training texts with phrases (PH)).

TEST	RECALL	PRECISION
7PH1aN	0.848	0.764
8PH1aN	0.802	0.729

TEST CODE:

7:	1987 WSJ texts, i.e., training set
8:	1988 WSJ texts, i.e., test set
PH:	texts with morphological transformation and word grouping
1a:	using cosine similarity measure and first level topic signatures
N:	using normalized *idf*

Table 4.17: First-level signatures: average recall and precision scores for training and test sets using normalized *idf*.

Chapter 5

Using Position: Optimal Position Policy

5.1 Introduction

In this chapter, we investigate another method of performing topic identification. The Position Method was identified in the 1960's, and still remains among the best, frequently outperforming newer methods like word counting. Exploiting regularities of discourse structure in a genre, this method assumes that the (ordinal) positon of a sentence is related to its importance in a text.

Since text genres and subject domains differ significantly, the Position Method is domain- and genre-dependent. For example, scientific articles have abstracts but newspaper articles do not, news articles in newspapers often summarize the story in the first paragraph but editorials do not. This means that the Position Method does not work equally well in all genres and domains, and tends to work better when its data resources and rules are tailored to the nature of the particular texts it is given. How, then, can one develop a method to determine which aspects of a particular genre and/or domain are useful for a Topic Identification system? How can one quickly and easily tune the rules of the system for a new genre and/or domain? How can one measure, or at least estimate, the effectiveness of a specific position method? In this chapter, we describe a Topic Identification module that employs the Position Method, as well as a method for rapidly constructing rules and resources that are tailored to new genres and domains.

5.2 Position as an Indicator of Importance

The idea that the important content of a text tends to appear at specifiable positions in the text is not new. However, no comprehensive experiments have been performed so far to verify this idea. Edmudson's [17] experiment in 1969 is by far the most significant. He introduced four clues for identifying significant words in a text. Among them, Title and Location are related to the Position Method.

Edmundson assumed that "an author conceives the title as circumscribing the subject matter of the document." In other words, words of the title are important. He also assumed that "topic sentences tend to occur very early or very late in a document and its paragraphs." He assigned positive weights to sentences according to their ordinal position in the text. The first sentence in the first paragraph and the last sentence in the last paragraph were most important in his scheme. Edmundson then conducted seventeen experiments in his research to verify the significance of these methods. According to his published results, the Title method and the Location method scored around 40% and 53% accuracy respectively. Accuracy was measured as the coselection rate between sentences selected by Edmundson's program and sentences selected by humans, in which sentences selected by the human were considered as the standard.

Although Edmundson did a good job in identifying important words in a text and providing a convincing experiment to verify his points, his work is not scientific by today's standards for two main reasons. First, his experiment only used 200 documents for training and another 200 documents for testing. As we show later, these numbers are not enough for accurate results. To measure the results satisfactorily, one must train the Title and Location method, which we called the Position Method, on a much bigger document collection. Second, Edmundson used the ordinal position hypothesis on some positions without trying other possible combinations, for example that the second and third paragraphs are more important than the first paragraph. In our experiments, we will also verify his hypothesis that the importance of a sentence in a text is related to its ordinal position in the text, and we will experimentally determine the most relevant ordinal positions.

The second motivation for our study for position method is that researchers have agreed that the position of words in a text has to do with their importance, but they

seem not to agree on where the important words are most likely to be found. Baxendale [3] conducted an investigation of a sample of 200 paragraphs. He found that in 85% of the paragraphs the topic sentence was the first sentence and in 7% it was the final one. Donlan [16] states that a recent study of topic sentences in expository prose showed that only 13% of the paragraphs of contemporary professional writers begin with topic sentences (Braddock [4]). Singer and Donlan [82] maintain that a paragraph's main idea can appear anywhere in the paragraph, or not be stated at all. Paijmans [65] conducted experiments on the relation between word position in a paragraph and its significance and found that

> "words with a high information content according to the *tf.idf*-based weighting schemes do not cluster in the first and the last sentences of paragraphs or in paragraphs that consist of a single sentence, at least not to such an extent that such a feature could be used in the preparation of indices for Information Retrieval purposes."

Kieras [40] confirmed the importance of the position of a mention within a text in psychological studies. It is one purpose of our study to clarify these contradictions and propose a systematic method to identifying important content in a text using position method.

5.3 Optimal Position Policy

In this section, we introduce the position hypothesis, show how to associate importance with the ordinal position of a sentence in a text, and describe how to create the Optimal Position Policy (OPP), which provides the important content positions in a text, by using a training corpus with topic indices.

5.3.1 The Position Hypothesis

The position hypothesis springs from the recognition that texts in a genre generally observe a predictable discourse structure, and that sentences of greater topic centrality tend to occur in certain specifiable locations. The text's title, for example, is a very informative position in most genres, as is the Abstract paragraph in scientific articles. However, the paradigmatic discourse structure differs significantly over

text genres and subject domains. For example, no Abstract paragraph is provided in most news articles, but the first paragraph of newspaper articles normally contains most important information. Figure 5.1 shows a typical *Wall Street Journal* article (after being run through a token segmenter). The first sentence provides a brief summary of the whole message while the other two sentences supply details. Therefore, a topic identification method based on the position hypothesis must take into account these differences and be able to adapt to various text genres or domains.

WSJ870325-0109

MONTREAL Noverco Inc. said it reduced its stake in Sceptre Resources Ltd. , a Calgary , Alberta-based oil and natural gas concern , to 4.8% from 6.3% . Noverco , a natural-gas distribution concern , said in a filing with the U.S. Securities and Exchange Commission that it sold 400,500 Sceptre common shares on March 13 for about $1.6 million (Canadian) . Noverco said it now holds 1,232,000 Sceptre shares .

Figure 5.1: A sample *Wall Street Journal* text.

5.3.2 How to Find Important Positions?

Instead of basing our work on a fixed rule such as "first and last sentences of each paragraph," we need a method to quickly identify the important positions for different genres or subject domains. How do we associate positions of sentences with their relative importance in a text? We use the following steps:

1. Label every sentence in the text with its ordinal paragraph and sentence numbers. For example, we label the first sentence of the first paragraph as (P1,S1) and the last sentence of the last paragraph as (P-1,S-1).

2. Read through each sentence and rank its importance with respect to other sentences.

3. Repeat steps 1 and 2 as required to either establish a pattern or to show that no pattern exists between sentence positions and their rankings. If a pattern exists, we then have an optimal position policy for selecting sentences according to their relative importance in the corresponding subject domain or genre; if

not, then the Position Method cannot be used to select important contents from this particular domain.

Step 1 is straightforward and takes virtually no time, but step 2 may take quite a long time for each text. In order to obtain reliable results, we require significant amounts of ranked sentences. However, currently no such resource is available. The closest approximation is human-made text summaries/abstracts and topic indices. Step 3 requires the repetition of steps 1 and 2 until statistically significant patterns are reached, or the nonexistence of patterns is proved. The biggest problem for automated position determination is the ranking of sentences. If no human-ranked sentences are available, then use summaries or keywords produced by humans. But summaries or keywords are not directly related to text sentences — they only, at best, overlap with some words or phrases. What to do?

Human-made text summaries/abstracts may contain phrases or words which also appear in the original texts. Assuming these phrases or words in summaries are more important than other ones that do not appear in summaries, we can assign sentences with more phrases or words in the summaries a higher importance. However, deciding phrase boundaries in the summaries is not as easy as just using human-made topic indices. A topic index is a phrase or word indicating the subject of a text. Therefore, sentences containing the whole topic index or part of it deserve higher rank than other sentences. Since a topic index has a fixed boundary, using it to rank sentences is easier than using a summary. In the next section, we present how to use topic indices to generate the optimal position policy.

5.3.3 From Topic Indices to the Optimal Position Policy

We determined the optimal position for topic occurrence as follows. Given a text **T** and a list of topics keywords $\mathbf{t}_i$ of **T**, we label each sentence of **T** with its ordinal paragraph and sentence number (P_m,S_n). We then removed all closed-class words from the texts. We did not perform morphological restructuring (such as canonicalization to singular nouns, verb roots, etc.) or anaphoric resolution (replacement of pronouns by originals, etc.), for want of robust enough methods to do so reliably. This makes the results somewhat weaker than they could be.

We defined *sentence yield* as the average number of different topic keywords mentioned in a sentence. We computed the yield of each sentence position in each text essentially by counting the number of different topic keywords contained in the appropriate sentence in each text, and averaging over all texts. Sometimes, however, keywords consist of multiple words, such as "spreadsheet software". In order to reward a full-phrase mention in a sentence over just a partial overlap with a multiword keyword/phrase, we used a formula sensitive to the degree of overlap. In addition, to take into account word position, we based this formula on the Fibonacci function; it monotonically increases with longer matched substrings, and is normalized to produce a score of 1 for a complete phrase match. Our hit function $\mathbf{H}$ measures the similarity between topic keyword $\mathbf{t}_i$ and a window $\mathbf{w}_{ij}$ that moves across each sentence (P_m,S_n) of the text. A window matches when it contains the same words as a topic keyword $\mathbf{t}_i$. The length of the window equals the length of the topic keyword. Moving the window from the beginning of a sentence to the end, we computed all the $\mathbf{H}_s$ scores and added them together to get the total score $\mathbf{H}_s$ for the whole sentence. We acquired the $\mathbf{H}_s$ scores for all sentences in $\mathbf{T}$ and repeated the whole process for the each text in the corpus. After obtaining all the $\mathbf{H}_s$ scores, we sorted all the sentences according to their paragraph and sentence numbers. For each paragraph and sentence number position, we computed the average $\mathbf{H}_{avg}$ score.

These average yields for each position are plotted in Figure 5.14, which shows the highest-yield sentence position to be (P_2,S_1), followed by (P_3,S_1), followed by (P_4,S_1), etc.

Finally, we sorted the paragraph and sentence position by decreasing yield $\mathbf{H}_{avg}$ scores. For positions with equal scores, different policies are possible: one can prefer sentence positions in different paragraphs on the grounds that they are more likely to contains distinctive topics. One should also prefer sentence positions with smaller S_m, since paragraphs are generally short. Thus the Optimal Position Policy for the Ziff-Davis corpus is the list

$$< (T), (P_2, S_1), (P_3, S_1), (P_4, S_1), (P_5, S_1), (P_6, S_1), \ldots >$$

5.4 Experiments

In this section we first describe the corpus which we used to carry out our experiments and detail the procedure of using manually-prepared keywords to associate the ordinal position of a sentence with its importance in a text. We use a Policy Determination Map (PDM) to represent the identified association and show how to use the PDM to generate the Opitmal Position Policy.

5.4.1 Summary of Resources Used in the Experiments

In order to verify the position hypothesis and demonstrate how to construct an optimal position policy in a real-world scale, we use the Ziff-Davis (ZIFF) texts in the TIPSTER collection. We use texts in ZIFF Vol. 1 for training and ZIFF Vol. 2 for one of our evaluation. A typical ZIFF text is shown in Figure 5.2. Each ZIFF text is marked up by SGML tags. Each document contains the DOC, DOCNO, and TEXT fields, with different possible additional fields between the DOCNO and the TEXT markers. The JOURNAL, TITLE, and AUTHOR fields contain respectively the journal, title and author of the material in the TEXT field. The SUMMARY field contains a summary of the full text within the TEXT field. Sometimes only an abstract is available instead of the full text, and in these cases there is no summary, and the abstract is contained in the TEXT field. The DESCRIPT field contains manually-indexed categories for the document. The DOCID field is the identifier used in the original data. Other fields in the data are: ABSTRACT, PRODUCT, ADDRESS, COMPANY, CATEGORY, SPECS, and NOTE. We selected around 13,000 documents from the ZIFF Vol. 1 collection as our training texts. These selected documents all contain SUMMARY and DESCRIPT fields, since a text with SUMMARY field is guaranteed to have the full text within the TEXT field. A text with DESCRIPT field contains human assigned document topic indices.

5.4.2 Preparing to Create the Optimal Position Policy

For each text in the training collection, we do the following:

<DOC>
<DOCNO> ZF109-669-733 </DOCNO>
<DOCID>09 669 733.&M;</DOCID>
<JOURNAL>PC Magazine Dec 25 1990 v9 n22 p46(1)
Full Text COPYRIGHT Ziff-Davis Publishing Co. 1990.&M;
</JOURNAL>
<TITLE>Handy Macro Editor/Debugger for Lotus 1-2-3 2.x. (Personics
Corp.'s Macro Editor/Debugger add-on software) (Software Review)
(First Looks) (evaluation)
</TITLE>
<AUTHOR>Stinson, Craig.&M;
</AUTHOR>
<SUMMARY>Personics Corp.'s $199.95 Macro Editor/Debugger (MED) is an add-on program that finds Lotus 1-2-3 macro problems and allows users to fix them.&P; MED works with Lotus 1-2-3 versions 2.0, 2.01 and 2.2 and uses about 70Kbytes of RAM.&P; MED's display in the lower half of the screen shows two windows in debugger mode.&P; One window shows code while the other displays up to four watchpoint variables and shows addresses of subroutine calls.&P; The editor can use three windows to display range names, code and comments simultaneously.&P; Breakpoints can be set to stop execution when the process reaches a certain cell, when a certain cell's value changes or when a logical expression becomes true.&P; Tracing can step through single instructions or move at full speed, and the next six instructions are shown in a window.&P; MED's error messages are clear and easily understood.&M;
</SUMMARY>
<DESCRIPT>
Company: Personics Corp. (Products).&O;
Product: Lotus 1-2-3 (Spreadsheet software) (Computer programs)
Macro Editor/Debugger (Add-in-on software).&O;
Topic: Macros
Program Editors
Debugging Tools.&M;
</DESCRIPT>
<TEXT>

Handy Macro Editor/Debugger for Lotus 1-2-3 2.x People who write Lotus 1-2-3 work in a language that's inadequately documented, lacks structured programming constructs, and never in its eight-year history has had decent debugging tools.&P; Never until now.&M;
Personics' Macro Editor/Debugger (MED) is a $199.95 add-in that threatens to make 1-2-3 macro development fun.&P; MED is a two-purpose tool.&P; The debugger tracks down problems and the editor fixes them.&P; Since the editor lets you create and modify range names by typing the names into cells, you can use it to create new macros.&M;
MED occupies the lowe half of your screen.&P; In debugger mode, it splits into two windows–one showing the current code context, the other displaying the status of watchpoint variables and the address of the current subroutine's caller.&M;
F2 invokes the editor and up to three windows, allowing you to work with range names, code, and comments at once.&M;
You can monitor up to four watchpoints and set nine breakpoints.&P; The latter can be of three types: "execute" (stop when execution comes to a particular cell), "update" (stop when a particular cell's value changes), and "conditional" (stop when a logical expression becomes true).&P; You can attach a count value to any breakpoint, so execution stops on the breakpoint's nth occurrence.&M;
Macros can be traced in two step modes: one instruction at a time and one cell at a time, or at nearly full speed.&P; In either mode, you can jump through loops at full speed.&P; while tracing, MED's code window shows the macro's next six instructions, even if they're in separate subroutines.&P; Any time execution is paused, you can move the pointer through your code to restart at a new location.&P; You can also invoke the editor, make changes, and restart–all without leaving MED.&M;
On top of these (and other) services, MED provides comprehensive diagnostic and informational messages.&P; It will distinguish between the various types of I/O errors, tell you when arguments are of the wrong type, and report typing errors (such as missing parentheses) in simple English.&M;
If you do serious development in 1-2-3 2.x, MED will preserve the hair on your head.&O;
</TEXT>
</DOC>

Figure 5.2: A typical text form TIPSTER ZIFF collection.

(T0,S0) Handy Macro Editor/Debugger for Lotus 1-2-3 2x Personics Corp Macro Editor/Debugger add-on software Software Review First Looks evaluation
(P1,S1) Handy Macro Editor/Debugger for Lotus 1-2-3 2x People who write Lotus 1-2-3 work in a language that inadequately documented lacks structured programming constructs and never in its eight-year history has had decent debugging tools
(P1,S2) Never until now
(P2,S1) Personics Macro Editor/Debugger MED is a 19995 add-in that threatens to make 1-2-3 macro development fun
(P2,S2) MED is a two-purpose tool
(P2,S3) The debugger tracks down problems and the editor fixes them
(P2,S4) Since the editor lets you create and modify range names by typing the names into cells you can use it to create new macros
(P3,S1) MED occupies the lowe half of your screen
(P3,S2) In debugger mode it splits into two windows one showing the current code context the other displaying the status of watchpoint variables and the address of the current subroutine caller
(P4,S1) F2 invokes the editor and up to three windows allowing you to work with range names code and comments at once
(P5,S1) You can monitor up to four watchpoints and set nine breakpoints
(P5,S2) The latter can be of three types execute stop when execution comes to a particular cell update stop when a particular cell value changes and conditional stop when a logical expression becomes true
(P5,S3) You can attach a count value to any breakpoint so execution stops on the breakpoint nth occurrence
(P6,S1) Macros can be traced in two step modes one instruction at a time and one cell at a time or at nearly full speed
(P6,S2) In either mode you can jump through loops at full speed
(P6,S3) while tracing MED code window shows the macro next six instructions even if theyre in separate subroutines
(P6,S4) Any time execution is paused you can move the pointer through your code to restart at a new location
(P6,S5) You can also invoke the editor make changes and restart all without leaving MED
(P7,S1) On top of these and other services MED provides comprehensive diagnostic and informational messages
(P7,S2) It will distinguish between the various types of I/O errors tell you when arguments are of the wrong type and report typing errors such as missing parentheses in simple English
(P8,S1) If you do serious development in 1-2-3 2x MED will preserve the hair on your head

Figure 5.3: Preprocessed text ZF109-669-733.

1. Preprocessing I:

Extract the text body from the original ZIFF text, remove all punctuation marks except &, _, -, and /. Label each sentence in the text with its ordinal paragraph number in the text and ordinal sentence number within its paragraph. For example, the second sentence in the third paragraph of the text is labeled (P3,S2). The title sentence is labeled (T0,S0). See Figure 5.3 for the preprocessed text ZF109-669-733 whose original text is shown in Figure 5.2.

2. Preprocessing II:

Extract the topic indices from the original ZIFF text. For text ZF109-669-733 in Figure 5.2, we have topics {Macros, Program Editors, and Debugging Tools}, as indicated in the <DESCRIPT> field.

3. Compute Hit Statistics:

Figure 5.4 shows the statistics for text ZF109-669-733. The first line is a unique message identification number for this text. The second line is the number of topic indices listed for the text, followed by the exact listing of those topics. The rest of the block contains statistics for each sentence in the text. According to the theory we have introduced in Section 5.4.2, we have to define a hit function to measure the degree of importance of each sentence in the text. Figure 5.5 illustrates the scoring process. Notice that the matching window moves from the beginning to the end within the sentence and the topic index is matched against the matching window from the last word to the first word. Because the head of a phrase tends to appear in the first word or the last word of the phrase, we give partial credit for a match even when it only matches at these two positions. For example, w_{ij} and t_i in Figure 5.5 are compared starting from A_j and B_i to A_{j-3} and B_{i-3}. If no match is found at the first and the last word position, the matching window moves to the next word A_{j+1} in the sentence and a new comparison between w_{ij+1} and t_i is performed. If a match is found, the matching window moves to w_{ij+L} in the sentence and continues comparison there (L is the number of words in t_i). Our algorithm compares A_{j-1} and B_{i-1} only when A_j and B_i are matched.

Topic indices normally appear in plurals such as Macros and Editors in text ZF109-669-733. In order to improve the matching result, our algorithm always chops off the last 's' in a word. However, it does not perform more sophisticated morphological canonicalization, such as reducing verbs to their root forms. Such manipulations will improve the results.

For this experiment, we defined three different hit functions to describe the matching result:

sentence yield (S): Sentence yield score S_j is computed for each comparison between w_{ij} and t_i. The sentence yield S of the whole sentence is the sum of all S_j computed in the sentence. Sentence yield has the value between 0 and 1. A value of 0 means there is no match between the w_{ij} and t_i under comparison, while a value of 1 means an exact match between the two. Any value between 0 and 1 indicates a partial match. In the case of a partial match, w_{ij} and t_i may match one, two, three, or more words. In order to reward a full-phrase

```
00 ZF109-669-733
01 3
02 1 macros
03 2 program editors
04 3 debugging tools
05 ----------------------------
06 T0 S0 T-0 S-0 2 2 0 0 2 1
07 P1 S1 P-8 S-2 2 1 0 1 2 2
08 P1 S2 P-8 S-1 0 0 0 0 0 0
09 P2 S1 P-7 S-4 2 2 0 0 2 1
10 P2 S2 P-7 S-3 0.5 0 0 0.5 1 1
11 P2 S3 P-7 S-2 0.5 0 0.5 0 1 1
12 P2 S4 P-7 S-1 1.5 1 0.5 0 2 2
13 P3 S1 P-6 S-2 0 0 0 0 0 0
14 P3 S2 P-6 S-1 0 0 0 0 0 0
15 P4 S1 P-5 S-1 0.5 0 0.5 0 1 1
16 P5 S1 P-4 S-3 0 0 0 0 0 0
17 P5 S2 P-4 S-2 0 0 0 0 0 0
18 P5 S3 P-4 S-1 0 0 0 0 0 0
19 P6 S1 P-3 S-5 1 1 0 0 1 1
20 P6 S2 P-3 S-4 0 0 0 0 0 0
21 P6 S3 P-3 S-3 1 1 0 0 1 1
22 P6 S4 P-3 S-2 0 0 0 0 0 0
23 P6 S5 P-3 S-1 0.5 0 0.5 0 1 1
24 P7 S1 P-2 S-2 0 0 0 0 0 0
25 P7 S2 P-2 S-1 0 0 0 0 0 0
26 P8 S1 P-1 S-1 0 0 0 0 0 0
```

Figure 5.4: Sentences and topic indices sentence yield/hit/dhit statistics for text ZF109-669-733. Each sentence is labeled with its forward and backward ordinal paragraph number in the text and sentence number within each paragraph.

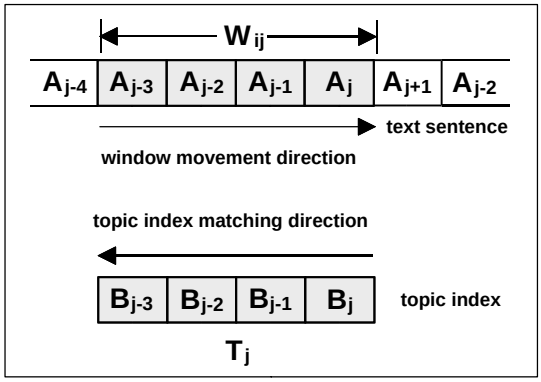


Figure 5.5: Matching a topic index and a sentence.

mention over just a partial match, we used a formula sensitive to the degree of overlap. In addition, to take into account matched word position, we based this formula on the Fibonacci function; it monotonically increases with longer matched substrings, and is normalizes to produce a score of 1 for a complete phrase match. For example, in Figure 5.5 t_i is a four-word topic index and it is compared with the matching window w_{ij}:

Assume $|t_i|$ is the number of words in t_i, $F(k) = F(k-1) + F(k-2)$, and $F(1) = F(2) = 1$. For the example in Figure 5.4, $|t_i|$ is equal to 4.

$$
\begin{array}{llll}
if & A_j = B_i & then & S_j = \sum_{k=|t_i|}^{|t_i|} F(k) / \sum_{k=1}^{|t_i|} F(k) = 0.43 \\
and & A_{j-1} = B_{i-1} & then & S_j = \sum_{k=|t_i|-1}^{|t_i|} F(k) / \sum_{k=1}^{|t_i|} F(k) = 0.71 \\
and & A_{j-2} = B_{i-2} & then & S_j = \sum_{k=|t_i|-2}^{|t_i|} F(k) / \sum_{k=1}^{|t_i|} F(k) = 0.86 \\
and & A_{j-3} = B_{i-3} & then & S_j = \sum_{k=|t_i|-3}^{|t_i|} F(k) / \sum_{k=1}^{|t_i|} F(k) = 1 \\
 & otherwise: & & \\
if & A_j \neq B_i & and & A_{j-|t_i|+1} = B_{i-|t_i|+1} \\
then & S_j = 0.5 & &
\end{array}
$$

Sentence yield S combines the effect of the frequency and quality of a match between a topic index t_i and the corresponding matching window w_{ij}. The frequency factor is one component of S, because one topic may appear many times in a sentence and S sums up all the contributions from each appearance of the topic. The quality factor is the other component, since S takes partial

matches into consideration. Column 5 of lines 6 to 26 in Figure 5.4 is the sentence yield of each sentence; it is the sums of column 6, 7, and 8. Columns 6, 7, and 8 show the sentence yield contributions from each of the three topic indices for each sentence in the text respectively.

hit (H): This measure only monitors the frequency of a topic index occurring in a sentence. A partial match is counted as a hit. The column second from last is the hit score in Figure 5.4. For example, line 12 in Figure 5.4 shows the hit score is 2, resulting from a full match of topic 1, *macro(s)*, and a partial match of topic 2, *program editor(s)*.

dhit (different hit, D): This measure records how many *different* topics are matched in a sentence. In this case, a partial match is also considered as a match of the corresponding topic. This assumes that a partial match identifies an abbreviation of a full match. For example, *editor* in sentence (P2,S4) in Figure 5.4 is a partial match of topic keyword *program editor(s)*. *Dhit* does not keep track of how many times a topic appears in a sentence. Line 9 in Figure 5.4 indicates that the sentence labeled as (P2 S1 P-7 S-4) is hit twice by topic 1, *macro(s)*, but it only registers one Different Hit. *Dhit* is very important in our study, because we are interested in finding as many of the topics as possible. Therefore our position theory must be tuned to identify the most likely positions of sentences which bear many topics instead of positions with high appearance probability of just a few topics.

4. Gather Paragraph And Sentence Statistics:

In this step, we collect facts about the training corpus. This data helps us determine appropriate parameter settings and guides the development of subsequent work on the position policy. These facts include average number of paragraphs per text (PPT), average number of sentences per paragraph (SPP), and average number of sentences per human-made summary (SPS). PPT and SPP prevent us from including a rare rule in the policy (such as the 25th sentence in 100th paragraph when PPT is 17 and SPP is 5). SPS helps us know how far we should go when producing a sentence extraction. If SPS is 6, then it is wise to select the top 6 positions from a text instead of top 20.

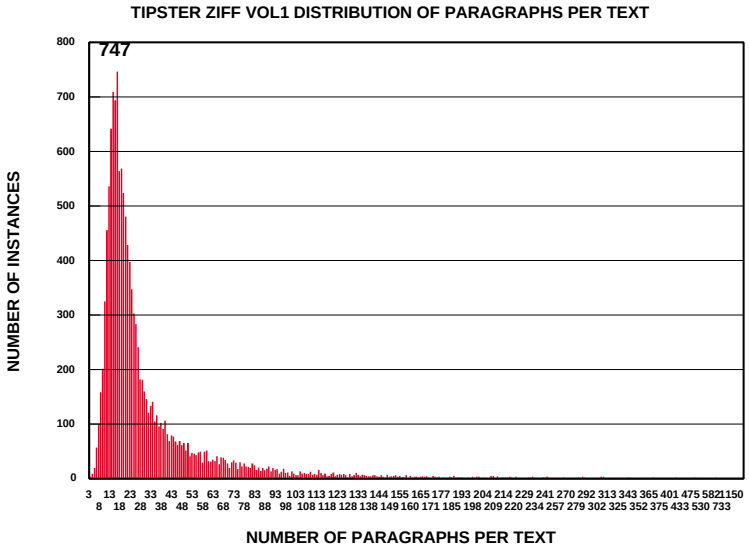


Figure 5.6: Number of paragraphs per text in ZIFF Vol. 1 collection.

For the ZIFF Vol. 1 corpus, PPT is 34.43, SPP is 2.05, and SPS is 5.76. Figure 5.6 shows the number of paragraphs per text. We can read from Figure 5.6 that most texts have fewer than 30 paragraphs. Figure 5.7 is the number of sentences per paragraph. It is obvious that most paragraphs (97.2%) have fewer than 5 sentences. Among them, 47.7% of the paragraph contain only one sentence and 25.2% of the paragraphs contain only two sentences. This means that in 47.7% of cases the first sentence of a paragraph is also the last sentence of the paragraph: in 47.7% of the time choosing the first sentence is the same as choosing the last sentence. Figure 5.8 gives the number of sentences per summary. Most summaries have 5 sentences and over 99.5% have under 10 sentences.

5. Prepare to Explore the Optimal Position Policy:

In order to test the hypothesis of the first/last paragraph and the first/last sentence position for the ZIFF Vol. 1 collection, and to lay the groundwork for building a policy determination map, we computed the average *dhit* score for each paragraph position, counting both forward and backward, and for each sentence position, also forward and backward. Figure 5.10 shows the *dhit* score for the first 50 paragraph

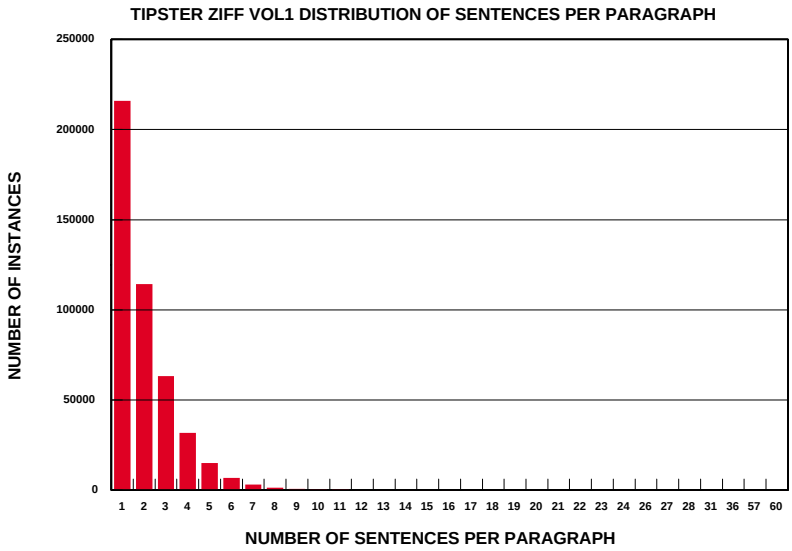


Figure 5.7: Number of sentences per paragraph in ZIFF Vol. 1 collection.

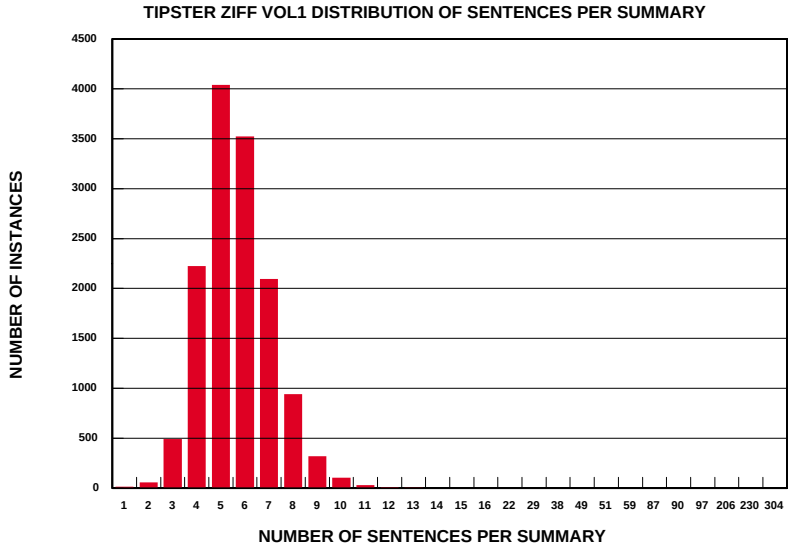


Figure 5.8: Number of sentences per summary in ZIFF Vol. 1 collection.

positions and Figure 5.11 shows the *dhit* score for the last 50 paragraph positions. Since the PPT number is 34.43 as mentioned in step 4, the first 50 and the last 50 positions fully cover most texts and are thus sufficient to illustrate any trends. The figures make clear that the title sentence ($dhit = 1.96$) stands out as a very informative sentence position compared with other paragraph-based positions. Figure 5.10 also indicates that the second ($dhit = 0.75$) and third ($dhit = 0.64$) paragraphs are better positions than the first ($dhit = 0.59$) to find topics. We also see the trend that paragraphs close to the beginning of texts tend to bear more informative contents.

With regard to the ends of texts, Figure 5.11 indicates that paragraph positions close to the ends of texts do not show any particular strong preference for topics. Notice that the maximum *dhit* score in Figure 5.11 does not exceed 0.42. Although it seems to rise from the last paragraph ($dhit = 0.28$) to the 14th-last ($dhit = 0.42$) and then gradually fall again, this phenomenon can be explained as the second paragraph effect mentioned earlier. According to the PPT statistics in Figure 5.9, most texts in the training collection have 13 to 16 paragraphs. When we take into consideration of the 2nd- and 3rd-paragraph effect, the 14th paragraph position peak in Figure 5.11 is just another occurrence of the same effect — counting backward from the end, the 14th-last paragraph *is* the second!

To examine the first/last sentence hypothesis, we computed the average *dhit* scores for the first and the last 10 sentence positions in a paragraph as shown in Figures 5.12 and 5.13 respectively. Figure 5.12 indicates that the closer a sentence is to the beginning of a paragraph, the higher its *dhit* score is. It confirms the *first sentence* theory. On the other hand, Figure 5.13 does not support the *last sentence* theory, since it suggests that the second sentence from the end of a paragraph contains more information. As we mentioned in step 4, 47.7% of the paragraphs in our training collection contain only one sentence and 25.2% of the paragraphs contain two sentences. We also know that the SPP is 2.05. If we use the *first sentence* hypothesis which we have confirmed and the SPP number, the 2nd sentence position from the end of an average paragraph is exactly the first sentence position from the beginning of the paragraph. This explains why a peak of *dhit* score appears at the second to the last sentence.

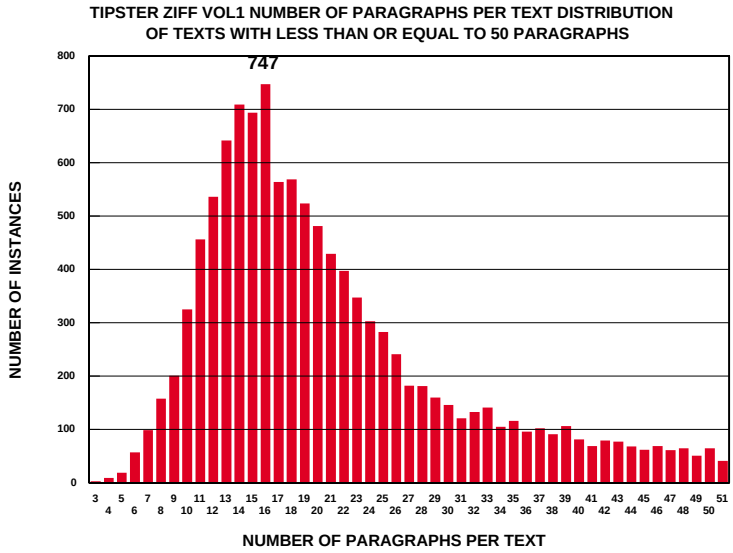


Figure 5.9: Number of paragraphs per text, for texts with fewer than or equal to 50 paragraphs in ZIFF Vol. 1 collection.

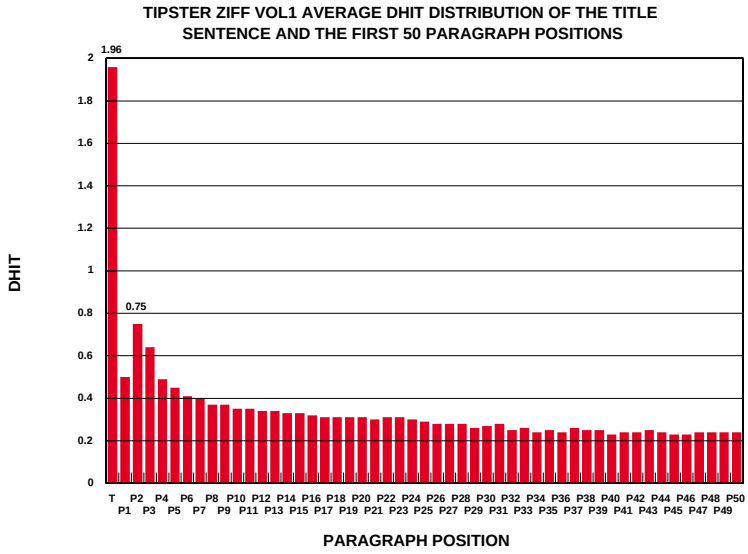


Figure 5.10: ZIFF Vol. 1 *dhit* distribution for the title sentence and the first 50 paragraph positions.

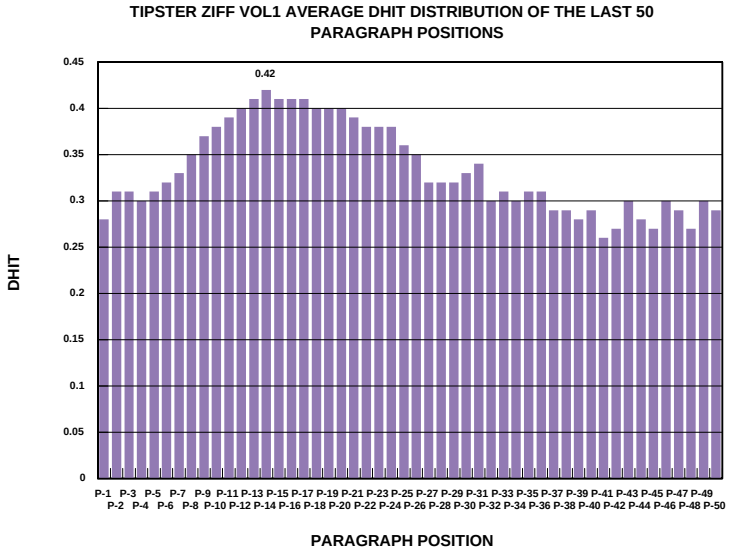


Figure 5.11: ZIFF Vol. 1 *dhit* distribution for the last 50 paragraph positions.

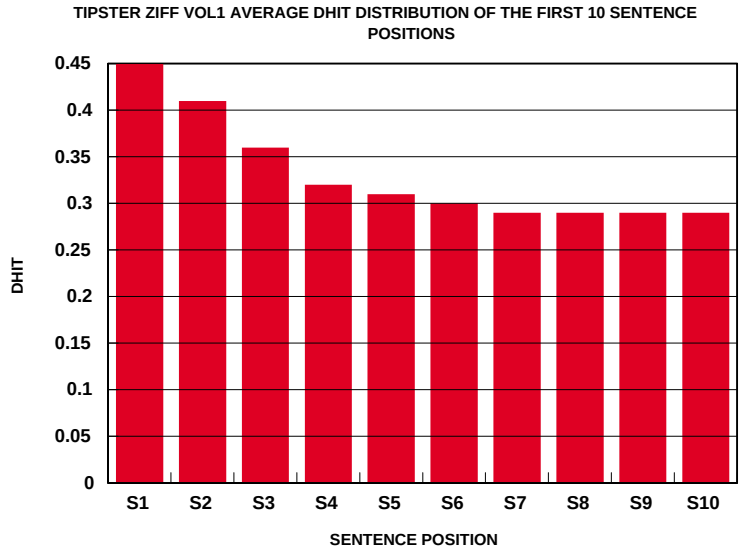


Figure 5.12: ZIFF Vol. 1 *dhit* distribution of the first 10 sentence positions in a paragraph.

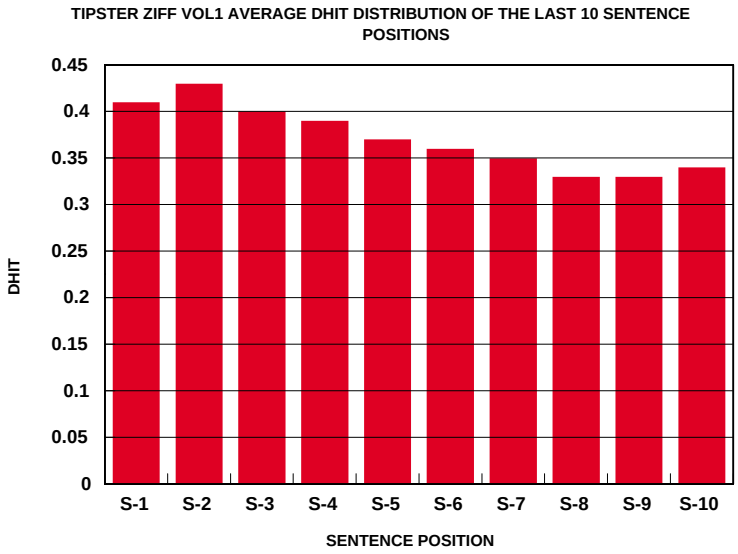


Figure 5.13: ZIFF Vol. 1 *dhit* distribution of the last 10 sentence positions in a paragraph.

6. Construct Policy Determination Map (PDM):

Summing up our findings from the previous steps on the training corpus, we have the following results:

- average number of paragraphs per text (PPT) is 34.43;
- average number of sentences per paragraph (SPP) is 2.05;
- 99.5% of paragraphs include fewer than 10 sentences;
- the title is the most informative position;
- topics are more likely to be found in the first few paragraphs in a text, especially in the second and third paragraphs;
- topics are more likely to be found in the first few sentences inside a paragraph, particularly in the first sentence.

These are corpus-specific results. We may have very different patterns for different corpora, but the process which we used to reach these results will remain the same.

Since the average length of a text is around 34 paragraphs, most paragraphs have fewer than 10 sentences, and topics tend to be found in beginning paragraphs in a text and beginning sentences in a paragraph, we obtained 300 data *dhit* score data points for the sentence positions (P_m, S_n) where $1 \leq m \leq 30$ (the first 30 paragraph positions in a text) and $1 \leq n \leq 10$ (the first 10 sentence positions in a paragraph). The actual *dhit* score for each data point is shown in Table 5.4.2.

7. Create An Optimal Position Policy from the PDM:

The data listed in Table 5.4.2 can be plotted as a contour map, to make it visually accessible. Figure 5.14 shows the contour view of the Policy Determination Map (PDM) with paragraph position in a text as the X axis, sentence position in a paragraph as the Y axis, and the average *dhit* score as the Z axis. The *dhit* score has a peak centered at the position (P_2, S_1) and gradually decreases as positions move away from the peak. *Dhit* score decreases more slowly in the X direction than in the Y direction, which conforms with our results for the *dhit* distributions of paragraph positions and sentence positions shown in Figures 5.10 and 5.12 respectively. An alternative, spectral view of the same data is shown in Figure 5.15.

Creating the Optimal Position Policy:

We define an Optimal Position Policy as follows:

> **DEF**: An Optimal Position Policy is a list of sentence positions, sorted by decreasing likelihood of containing topics in texts of a given genre. It is used as a guideline for choosing important sentences that contain as many topics of a text as possible with high probability.

An OPP does not guarantee that sentences selected following the OPP guideline are always the best choice, since an OPP is based on corpus statistics. However, an OPP does provide statistically-backed good suggestions for selecting important sentences from a text. There are several alternative ways to create an Optimal Position Policy (OPP) from the PDM. The relevant principles are:

1. choose sentence positions with high *dhit* scores early,

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
P1	0.56	0.57	0.55	0.46	0.46	0.52	0.69	0.76	0.5	0.45
P2	0.86	0.74	0.54	0.51	0.46	0.4	0.41	0.33	0.57	0.57
P3	0.81	0.61	0.46	0.45	0.39	0.45	0.42	0.32	0.54	0.23
P4	0.67	0.52	0.44	0.41	0.42	0.37	0.29	0.41	0.32	0.33
P5	0.62	0.48	0.41	0.36	0.36	0.29	0.35	0.4	0.36	0.55
P6	0.58	0.47	0.4	0.38	0.34	0.36	0.37	0.45	0.28	0.33
P7	0.57	0.46	0.39	0.35	0.34	0.41	0.36	0.46	0.5	0.5
P8	0.54	0.45	0.4	0.36	0.34	0.35	0.3	0.37	0.21	0.42
P9	0.54	0.43	0.38	0.32	0.28	0.34	0.23	0.44	0.2	0.16
P10	0.51	0.43	0.37	0.34	0.32	0.32	0.22	0.3	0.5	0.23
P11	0.5	0.4	0.36	0.38	0.36	0.29	0.34	0.23	0.33	0.5
P12	0.5	0.41	0.34	0.32	0.34	0.39	0.35	0.23	0.14	0.26
P13	0.5	0.42	0.37	0.36	0.31	0.3	0.3	0.33	0.33	0.44
P14	0.48	0.41	0.35	0.32	0.31	0.36	0.26	0.49	0.58	0.4
P15	0.49	0.41	0.37	0.37	0.33	0.3	0.36	0.35	0.29	0.25
P16	0.48	0.42	0.37	0.36	0.34	0.35	0.38	0.37	0.16	0.35
P17	0.47	0.4	0.36	0.33	0.33	0.3	0.32	0.41	0.41	0.33
P18	0.46	0.39	0.35	0.33	0.33	0.3	0.28	0.43	0.21	0.2
P19	0.47	0.38	0.39	0.31	0.36	0.37	0.21	0.2	0.09	0.2
P20	0.47	0.43	0.38	0.36	0.34	0.3	0.37	0.23	0.47	0.06
P21	0.46	0.4	0.35	0.36	0.32	0.28	0.32	0.35	0.15	0.18
P22	0.47	0.4	0.34	0.3	0.33	0.26	0.36	0.25	0.22	0.3
P23	0.47	0.42	0.35	0.32	0.28	0.36	0.31	0.43	0.33	0.15
P24	0.46	0.39	0.37	0.35	0.33	0.34	0.31	0.31	0.52	0.8
P25	0.46	0.39	0.39	0.3	0.34	0.32	0.35	0.29	0.15	0.57
P26	0.44	0.4	0.38	0.34	0.32	0.4	0.33	0.33	0.06	0.45
P27	0.43	0.39	0.35	0.28	0.3	0.3	0.18	0.21	0.44	0.33
P28	0.45	0.41	0.36	0.36	0.35	0.3	0.32	0.37	0.33	0.07
P29	0.42	0.38	0.36	0.36	0.3	0.37	0.22	0.06	0.11	0.37
P30	0.44	0.43	0.34	0.3	0.33	0.29	0.2	0.25	0.6	0

Table 5.1: ZIFF Vol. 1 optimal position Policy Determination Map *dhit* scores.

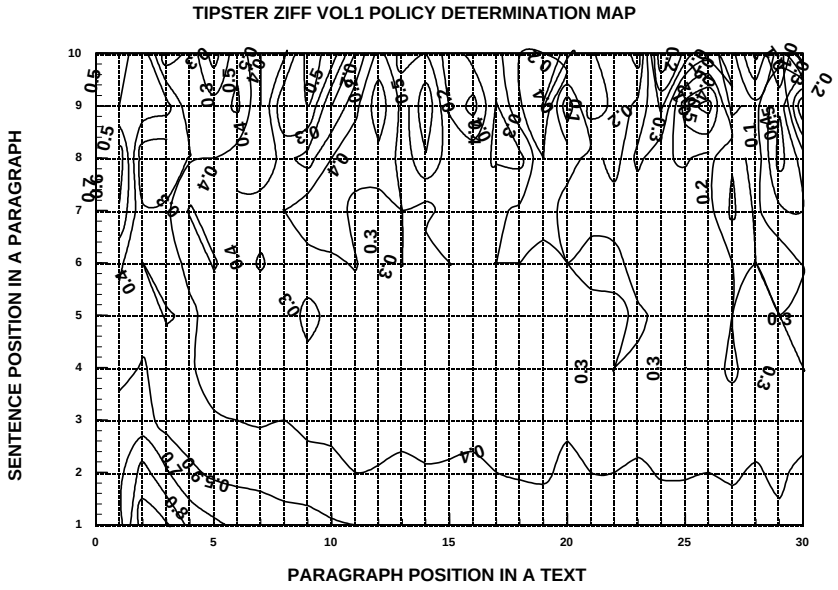


Figure 5.14: ZIFF Vol. 1 optimal position Policy Determination Map in contour view.

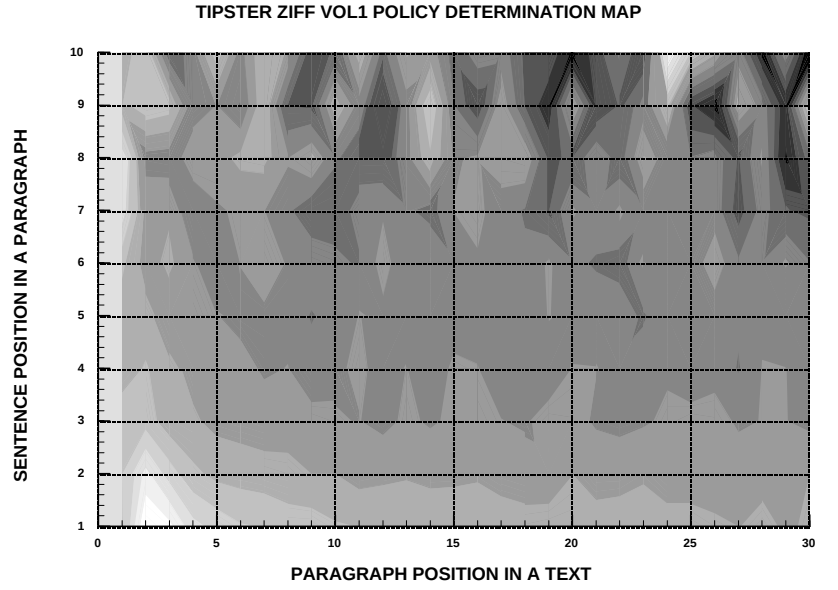


Figure 5.15: ZIFF Vol. 1 optimal position Policy Determination Map in spectral view.

2. avoid choosing sentence positions in the same paragraph, and

3. prefer sentence positions with smaller S_m index.

Sentence positions with high *dhit* score tend to contain more topics. Sentence positions not in the same paragraph are more likely to contains distinctive topics. The chances are better to find sentence positions with smaller S_m, since the average number of sentences per paragraph is only 2.05. A straightforward method for creating an OPP is to sort all the positions according to descending *dhit* score and then simply to pick the top N positions, where N is the desired number of sentences.

However, the contour and gradient graphs in Figures 5.14 and 5.15 only present a comparison of *dhit* scores among individual sentence positions. An extract consisting of sentences of topmost N *dhit* scores cannot guarantee that it has the highest *cumulative dhit* score. To assure that an extract always has the highest cumulative *dhit* score, we have to compute the cumulative *dhit* scores for all possible sentence position combinations and then select the position combination of the highest cumulative *dhit* value. For an N-sentence extract consisting of sentences selected from 300 data points, $C_N^{300} = 300!/(300-N)!N!$ possible combinations must be examined. For extracts of one, two, and three sentences this means that 44,850, 4,455,100, and 330,791,175 trials are required respectively. To compute all these possible combinations is not a formidable task, but the question is: is it worthwhile? Consider the following:

- PDM and OPP are statistics-based results. They give *strong suggestions* instead of *iron rules*.

- The *dhit* score graph shown in Figure 5.10 indicates that the title sentence, the second paragraph, and third paragraph of a text tend to contain much more information than the other parts of the text. Furthermore, the *dhit* score difference of the remaining first 30 paragraph positions lies within a very small range (≈ 0.2).

Therefore, using heuristic rules to develop an OPP seems a reasonable option.

To create the OPP from the PDM, two heuristic rules were tested. Assume E_i is the extract after the *ith* sentence addition, T_i is the remianing sentence pool

from which we can select the next sentence to be added into E_i before the ith sentence removal, N is the number of sentences to be included in the final extract, sentence position is marked as (P_m, S_n), and $\max(T_i)$ gives the sentence position of the highest *dhit* score in T_i. When a tie among *dhit* scores occurs, $max(T_i)$ selects sentence according to the smaller S_n and then the smaller P_m.

Heuristic I: (High *dhit* Only) a greedy algorithm: always select the sentence with the highest *dhit* score. This is a strict application of principle 1 only.

```
1   E_0 = φ
2   T_0 = ∪(P_m, S_n), 1 ≤ m ≤ 30 and 1 ≤ n ≤ 20
3   i = 0
4   while i < N {
5       s_i = max(T_i)
6       T_{i+1} = T_i \ s_i
7       E_{i+1} = E_i ∪ s_i
8       i = i + 1
9   }
10  output E_N
```

Heuristic II: (High *dhit* and First Sentence) also a greedy algorithm, but considers all three principles suggested before.

```
1   E_0 = φ
2   T_0 = ∪(P_m, S_1), 1 ≤ m ≤ 30
3   i = 0
4   while i < N {
5       s_i = max(T_i)
6       T_{i+1} = T_i \ s_i
7       E_{i+1} = E_i ∪ s_i
8       i = i + 1
9   }
10  output E_N
```

In our experiment, we divided the *dhit* score range of 0 to 1 into intervals of 0.05. All the positions with *dhit* score falling in the same interval form an *equivalence group* and are sorted according to principles 2 and 3 within each group. We also concentrated our efforts on sentence position 1 to 5, since the average number of sentences per paragraph is 2.05 and positions with sentence marker S_m higher than

range	members	rank
$dhit > 0.9$	(T), title sentence	1
$0.9 \geq dhit > 0.85$	(P_2, S_1)	2
$0.85 \geq dhit > 0.8$	(P_3, S_1)	3
$0.8 \geq dhit > 0.75$	ϕ	
$0.75 \geq dhit > 0.7$	(P_2, S_2)	4
$0.7 \geq dhit > 0.65$	(P_4, S_1)	5
$0.65 \geq dhit > 0.6$	(P_5, S_1),(P_3, S_2)	6
$0.6 \geq dhit > 0.55$	(P_6, S_1),(P_7, S_1),(P_1, S_1),(P_1, S_2)	7
$0.55 \geq dhit > 0.5$	(P_8, S_1),(P_9, S_1),(P_{10}, S_1),(P_4, S_2), (P_1, S_3),(P_2, S_3),	8
$dhit = 0.5$	(P_{11}, S_1),(P_{12}, S_1),(P_{13}, S_1)	9

Table 5.2: Positions listed according to Heuristic I in 0.05 *dhit* decrement. Only positions whose *dhit* score is greater than or equal to 0.5 and $S_{n \leq 5}$ are listed.

5 are rare. Tables 5.2 and 5.3 show the equivalence groups of positions with *dhit* score greater than or equal to 0.5 following Heuristic I and Heuristic II respectively.

The equivalence group provides another way to classify sentence positions. We assign each equivalence group a rank according to the average *dhit* score of its members. The higher the average *dhit* score of an equivalence group, the higher the rank of the equivalence group. When we apply the OPP to select sentences from a text, we take sentence positions with higher rank first. For example, based on the OPP shown in Table 5.2, a title sentence will always be selected before the first sentence of any second paragraph. Table 5.4 shows the cumulative *dhit* score comparison between Heuristic I and Heuristic II. It is clear that Heuristic II achieves better cumulative *dhit* scores than Heuristic I after the fourth positions where the benefit of selecting sentence from different paragraphs takes off and the higher individual *dhit* score start to have less influence. Therefore, we use Heuristic II to form our OPP. The final OPP for our training corpus is:

$$< (T), (P_2, S_1), (P_3, S_1), (P_4, S_1), (P_5, S_1), (P_6, S_1), \ldots >$$

range	members	rank
$dhit > 0.9$	(T), title sentence	1
$0.9 \geq dhit > 0.85$	(P_2, S_1)	2
$0.85 \geq dhit > 0.8$	(P_3, S_1)	3
$0.8 \geq dhit > 0.75$	ϕ	
$0.75 \geq dhit > 0.7$	ϕ	
$0.7 \geq dhit > 0.65$	(P_4, S_1)	4
$0.65 \geq dhit > 0.6$	(P_5, S_1)	5
$0.6 \geq dhit > 0.55$	(P_6, S_1),(P_7, S_1),(P_1, S_1)	6
$0.55 \geq dhit > 0.5$	(P_8, S_1),(P_9, S_1),(P_{10}, S_1)	7
$dhit = 0.5$	(P_{11}, S_1),(P_{12}, S_1),(P_{13}, S_1)	8

Table 5.3: Positions listed according to Heuristic II in 0.05 *dhit* decrement. Only positions whose *dhit* score is greater than or equal to 0.5 are listed.

Position	Heuristic I	Heuristic II	Difference
1	0.42456	0.42456	0.00000
2	0.49026	0.49026	0.00000
3	0.54154	0.54154	0.00000
4	0.55848	0.57483	0.01635
5	0.58897	0.59955	0.01058
6	0.61211	0.61945	0.00734
7	0.62220	0.63623	0.01403
8	0.62907	0.64335	0.01428
9	0.64547	0.65668	0.01122
10	0.65999	0.66899	0.00900
11	0.67182	0.67930	0.00748
12	0.68295	0.68818	0.00523
13	0.69209	0.69575	0.00365
14	0.69412	0.70338	0.00926
15	0.69862	0.70946	0.01084
16	0.69943	0.71408	0.01465
17	0.70145	0.71902	0.01757
18	0.70206	0.72246	0.02039

Table 5.4: Cumulative *dhit* scores of Heuristic I and II and their difference in the first 18 positions.

5.5 Evaluations

The goal of creating an Optimal Position Policy is to adapt the Position Hypothesis to various domains or genres. In order to determine the adequacy/quality of a specific OPP, two different measures are required:

1. creating a new OPP for another text collection in the same domain should result in a similar OPP (in this way, we prove that the OPP is not simply the result of the collection used, but captures requlariities of the domain or genre), and

2. sentences selected according to the OPP should indeed carry more information than other sentences, as evaluated on unseen texts in the same genre or domain.

Two evaluations to confirm these points are described in the following sections.

5.5.1 Evaluation I

To verify that the procedure described in Section 5.3 for finding an Optimal Position Policy in the TIPSTER ZIFF domain is useful, we selected 2,907 texts from the TIPSTER ZIFF Vol. 2 collection, ZF_251 to ZF_300, as a test set. The average *dhit* scores of 500 positions (P_m, S_n), where $1 \leq m \leq 30$ and $1 \leq n \leq 10$ were computed. The results are shown in Table 5.5.1, Figure 5.16, and Figure 5.17.

Figure 5.16 shows a very similar pattern as was obtained in Figure 5.14. The *dhit* score peaks at position (P_2, S_1) and gradually decreases in the X direction and more rapidly in the Y direction. The OPP, using Heuristic II, for the test texts is:

$$< (T), (P_2, S_1), (P_3, S_1), (P_4, S_1), (P_5, S_1), (P_6, S_1), (P_7, S_1), (P_1, S_1) \ldots >$$

Comparing this to the OPP listed above for the training texts:

$$< (T), (P_2, S_1), (P_3, S_1), (P_4, S_1), (P_5, S_1), (P_6, S_1), (P_7, S_1), (P_1, S_1) \ldots >$$

The similarity between the policy determination maps of training set and test set (they are the same at least at the top 8 positions shown above) confirms two things. First, a correspondence exists between topics and sentence positions in separately

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
P1	0.55	0.58	0.47	0.48	0.44	0.51	0.42	0.4	1	0
P2	0.89	0.68	0.48	0.51	0.51	0.55	0.88	0.44	0.25	1
P3	0.8	0.57	0.46	0.46	0.36	0.39	0.48	0.3	0.5	0
P4	0.67	0.52	0.46	0.38	0.42	0.5	0.4	0.18	0	0.4
P5	0.64	0.45	0.4	0.42	0.36	0.36	0.23	0.23	0.22	0
P6	0.59	0.49	0.4	0.36	0.35	0.24	0.46	0	0	0.33
P7	0.56	0.43	0.37	0.38	0.4	0.38	0.1	0.22	0	*
P8	0.53	0.42	0.35	0.32	0.26	0.25	0.31	0.17	0.14	0
P9	0.53	0.39	0.34	0.34	0.27	0.25	0.16	0.1	0.33	0.75
P10	0.5	0.41	0.38	0.33	0.36	0.36	0.25	0.22	0.33	0
P11	0.48	0.44	0.38	0.35	0.37	0.2	0.34	0.08	0	0
P12	0.5	0.41	0.34	0.29	0.33	0.38	0.35	0.1	0	0.33
P13	0.46	0.39	0.36	0.25	0.23	0.34	0.33	0.4	0.2	0.5
P14	0.49	0.36	0.32	0.29	0.29	0.3	0.25	0.25	0	0
P15	0.46	0.38	0.34	0.32	0.29	0.4	0.11	0	0	0
P16	0.47	0.36	0.32	0.35	0.26	0.3	0.41	0.23	0.3	0.66
P17	0.48	0.42	0.29	0.25	0.24	0.25	0.42	0.23	0.4	*
P18	0.47	0.39	0.35	0.32	0.28	0.2	0.13	0.18	0	0
P19	0.49	0.44	0.33	0.29	0.28	0.3	0.14	0.1	0.6	0
P20	0.44	0.39	0.32	0.3	0.4	0.3	0.47	0	0.28	0.33
P21	0.42	0.44	0.36	0.27	0.27	0.52	0.19	0.43	0.5	0.75
P22	0.43	0.35	0.39	0.3	0.28	0.33	0.35	0	0.33	0
P23	0.46	0.39	0.39	0.36	0.21	0.42	0.42	0.33	0.4	0.5
P24	0.46	0.37	0.34	0.36	0.27	0.39	0.31	0.37	0.5	0.5
P25	0.39	0.4	0.35	0.47	0.54	0.3	0.33	0.33	0	0
P26	0.39	0.38	0.35	0.3	0.26	0.35	0.12	0.11	0.27	0.6
P27	0.41	0.36	0.38	0.22	0.22	0.3	0.56	0.28	1	0
P28	0.42	0.37	0.34	0.31	0.3	0.39	0.15	0.7	0.16	0
P29	0.4	0.3	0.32	0.22	0.15	0.27	0.07	0.6	0	0
P30	0.39	0.32	0.36	0.29	0.31	0.25	0.22	0.2	0.6	1

Table 5.5: ZIFF Vol. 2 (ZF_251 to ZF_300) optimal position Policy Determination Map *dhit* scores. Notice that no *dhit* scores are available at positions (P_7, S_{10}) and (P_{17}, S_{10}), since there are at most nine sentences in paragraph 7 and 10 in the test set. Although some high *dhit* scores are shown in sentence positions S_6 to S_{10}, these data points should be considered as singular points where not enough sentence samples are available.

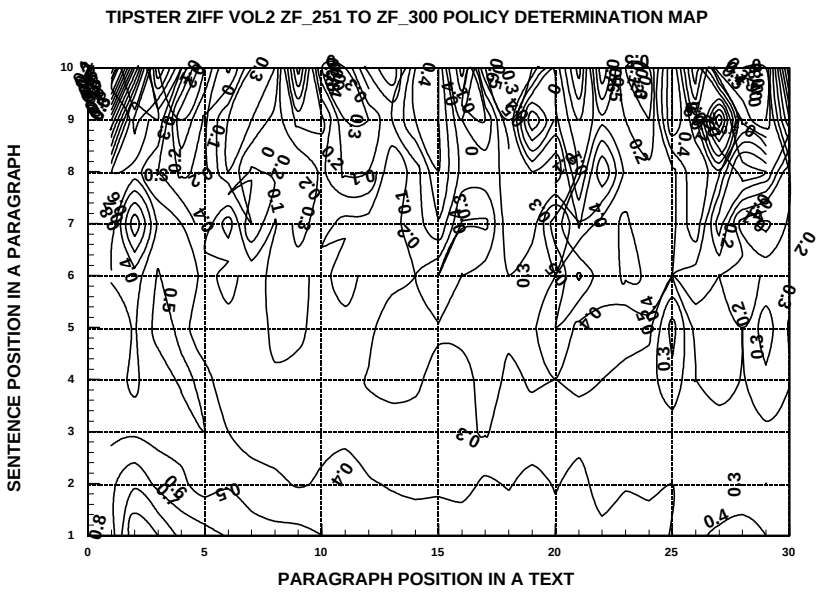


Figure 5.16: ZIFF Vol. 2 (ZF_251 to ZF_300) optimal position Policy Determination Map in contour view.

selected text corpus within a domain such as the TIPSTER ZIFF collection. Second, the regularity between topics and sentence positions can be used to identify topic sentences in texts.

5.5.2 Evaluation II

The evaluation described in the previous section ensures that the procedure used to generate an OPP does produce a similar policy for subsets of the same domain. Now we must verify that the OPP actually selects important sentences and we must measure how well it does so.

One way to evaluate the quality of the OPP-selected sentences is to ask human judges to review them and assign scores. Another way is to measure the similarity between the OPP-selected sentences and the sentences in human-prepared summaries. Both keywords and abstracts contain phrases and words which also appear in the original texts; on the assumption that these phrases or words are more important in the text than other ones, we can assign a higher importance to sentences with

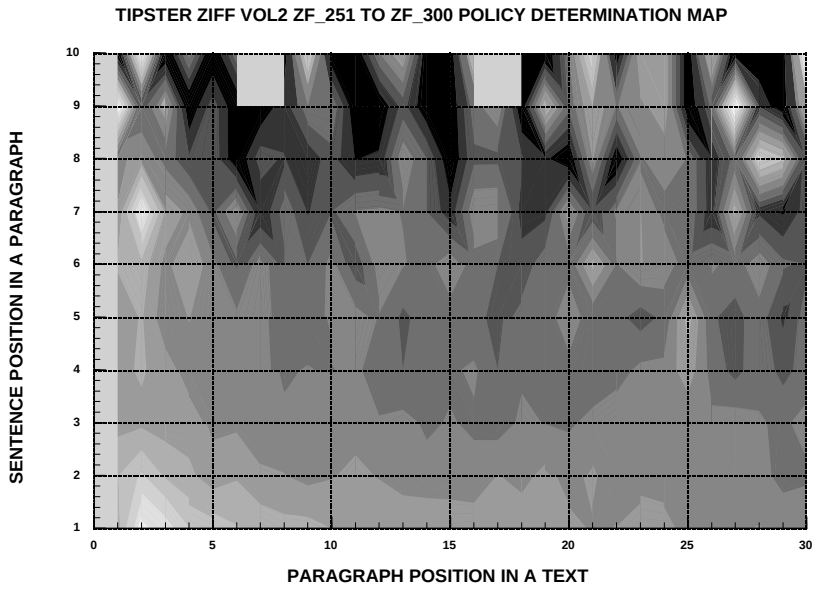


Figure 5.17: ZIFF Vol. 2 (ZF_251 to ZF_300) Policy Determination Map in spectral view. The two white squares at positions (P_7, S_{10}) and (P_{17}, S_{10}) indicate that no data points are available.

more such phrases or words (or parts of them).[1] Since a topic keyword has a fixed boundary, using it to rank sentences is easier than using an abstract.

Although it is more desirable, the first approach is not efficient. It may take days or even weeks to judge all the OPP-selected sentences. The human judge approach is only appropriate when the number of texts in the test set is small. However, since an OPP is a statistics-based result, the effectiveness of an OPP is more likely to show when the number of texts in the test sets is large. Therefore, the second approach is more viable if human-prepared summaries are available.

Most texts of the TIPSTER ZIFF collection include a summary, created by a human. A typical example is shown in Figure 5.2. We used these human-prepared summaries to evaluate the OPP-selected sentences. Unfortunately, the OPP extracts and the human summaries are not immediately comparable. An extract of a text is composed of sentences selected from the text. The major difficulty involved in this evaluation is that human-prepared summaries are not composed of the original sentences in the full texts which cover the topics expressed in these summaries. Sentences of human-prepared summaries often combine parts of two or more sentences in full texts, or phrase the same topics a different ways.

It is more feasible to measure the similarity between the OPP-selected extracts and sub-sentence units of the human-prepared summaries. To achieve this, we broke sentences in the OPP-selected extracts and the human-prepared summaries into windows of different sizes and then compared the windows. A window consists of a certain number of adjacent words in extracts or summaries. Closed-class words such as *the*, *to*, *a*, and some auxiliary verbs such as *be*, *may*, *are* are ignored. Words are all converted into lower case, but no morphological transformation such as canonicalization to verb root or singular noun is used. (Transformation would tend to improve the evalutaion scores). We then computed the similarity between windows of the same size in an extract and their corresponding summary.

Before providing details of the evaluation, we define some terms and three measures used to assess the quality of the OPP-selected extracts. We define, for an extract E and a summary S:

[1] How many topic keywords would be taken over verbatim from the texts, as opposed to generated paraphrastically by the human extractor, was a question for empirical determination—the answer provides an upper bound for the power of the Position Method.

w^E_{mi}: a window i of size m in E.

w^S_{mi}: a window i of size m in S.

$|W^E_m|$: total number of windows of size m in E.

$|W^S_m|$: total number of different windows of size m in S, i.e., how many $w^S_{mi} \neq w^S_{mj}$.

hit : $w^E_{mi} = w^S_{mj}$, i.e., words and word sequencies in w^E_{mi} and w^S_{mj} are exactly the same.

Precision of windows size m:

$$P_m = \frac{number\ of\ hits}{|W^E_m|}$$

Recall of windows size m:

$$R_m = \frac{number\ of\ different\ hits}{|W^S_m|}$$

Coverage of windows size m:

$$C_m = \frac{number\ of\ sentences\ in\ S\ with\ at\ least\ one\ hit}{number\ of\ sentences\ in\ S}$$

5.5.2.1 Precision and Recall

Precision, P_m, measures what percentage of windows of size m in E can also be found in S. In other words, P_m indicates what percentage of E is considered to be important with regard to S. Recall, R_m, measures the diversity of E. A high P_m does not guarantee recovery of all the possible topics in S, but a high R_m does ensure that many different topics in S are covered in E. However, a high R_m alone does not warrant good performance either. For example, an OPP which selects all the sentences in the original text certainly has a very high R_m, but this extract which duplicates the original text is the last thing we want as a summary alternative. It is important to consider P_m and R_m together.

Figures 5.18 and 5.19 show the precision/recall graphs of windows size 1 and 2 respectively. Figure 5.18 indicates that the precision score decreases slowly and

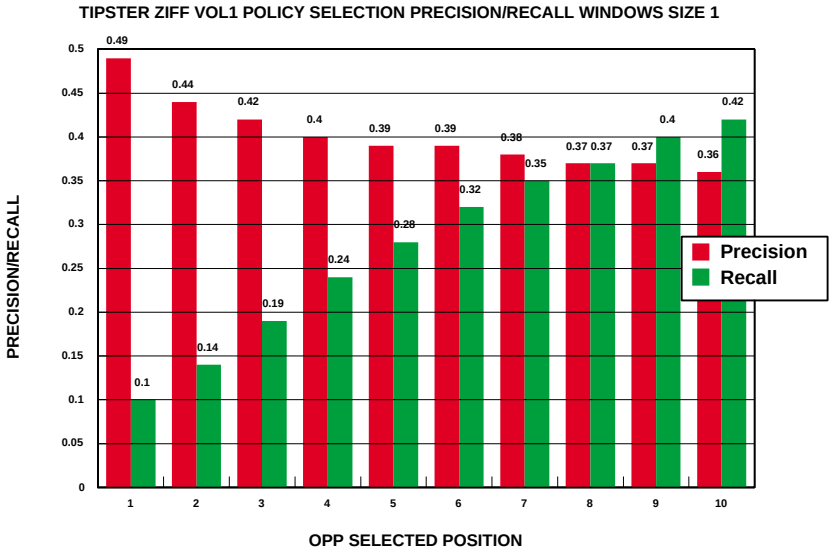


Figure 5.18: Cumulative precision/recall scores of top ten OPP-selected sentence positions of window size 1.

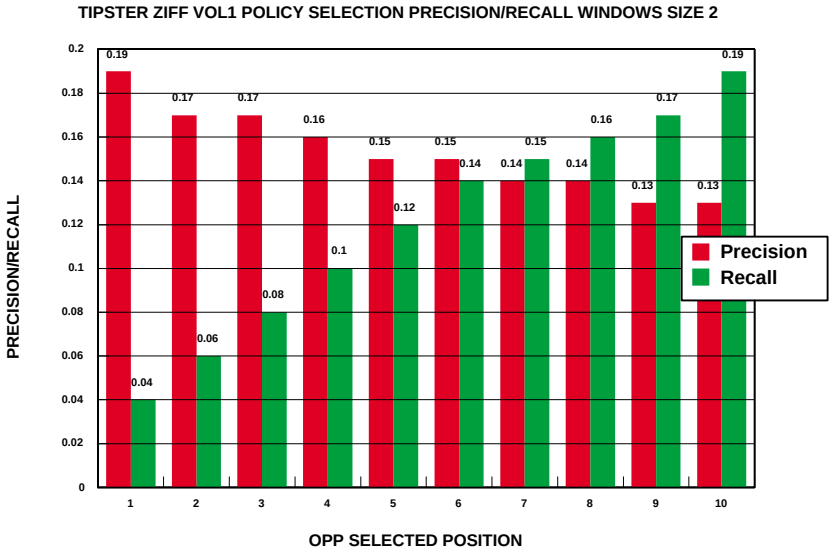


Figure 5.19: Cumulative precision/recall scores of top ten OPP-selected sentence positions of window size 2.

the recall score increases more rapidly as we choose more sentences according to the OPP. When we select 7 sentences (which is 10% of the average length of a ZIFF text), the precision is 0.38 and the recall 0.35. Considering that the matching process requires exact matches and morphological transformation is not performed, this result is very encouraging.

With windows of size 2, precision and recall scores drop seriously. Since long phrases occur less frequently, long windows are more difficult to find matching partners in S and E. As the window size increases, a lot more false combination of words are also introduced; this degrades the performance dramatically. It suggests that ideally it is better to vary the length of windows by always using the right window size to do the matching. In fact, the precision and recall results obtained by using window size 1 include contributions from window sizes longer than 1, since whenever a match is found the individual words in the matching windows have to be the same.

The contribution of precision, P_m^o, and recall, R_m^o, resulting from an m word window *alone* can be computed as:

$$\begin{aligned} P_m^o &\approx P_m - P_{m+1} \\ R_m^o &\approx R_m - R_{m+1} \end{aligned}$$

Figures 5.20 and 5.21 show precision and recall scores with individual contributions from window sizes 1 to 5. P_v and R_v of variable-length windows can be estimated as follows:

$$\begin{aligned} P_v &\approx \sum_{m=1}^{l} P_m^o \\ R_v &\approx \sum_{m=1}^{l} R_m^o \end{aligned}$$

The performance using variable-length windows compared with using windows of size 1 should be within the amount shown in the segments of window size ≥ 5.

5.5.2.2 Coverage

Coverage, C_m, tests the similarity between E and S in a very loose sense. It counts the number of sentences in S for which at least one hit found is in E, i.e., there exists

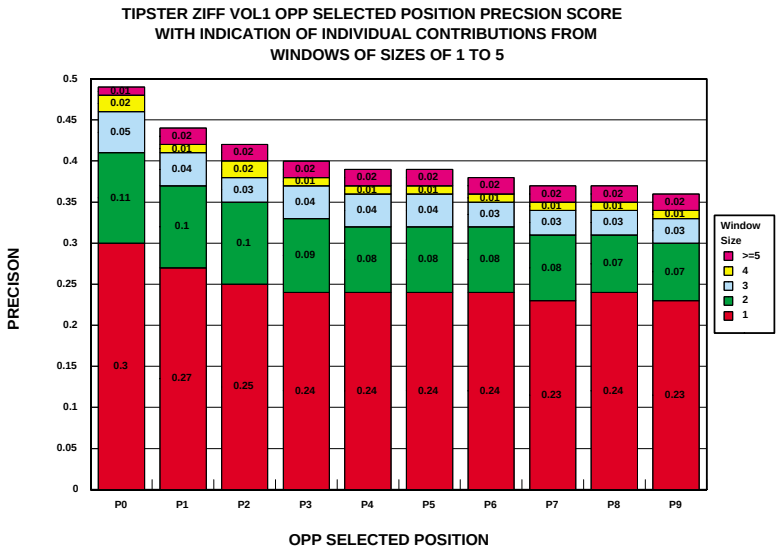


Figure 5.20: Precision scores of individual contributions from windows of sizes 1 to 5.

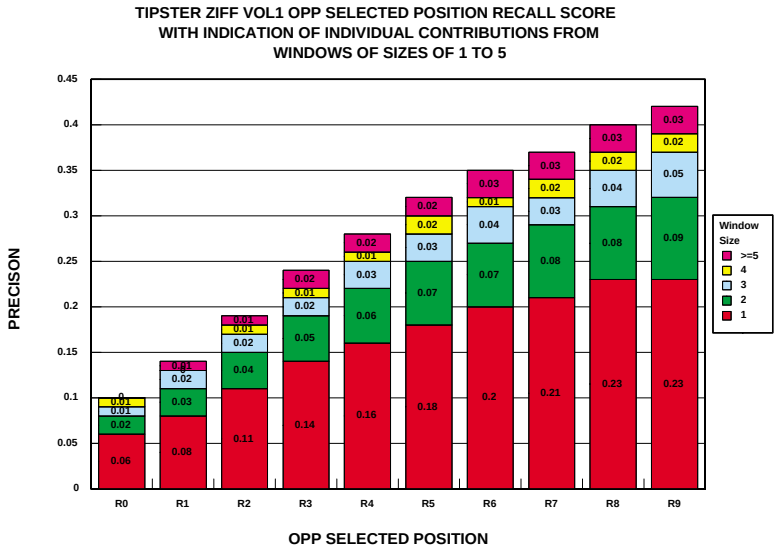


Figure 5.21: Recall scores of individual contributions from windows of sizes 1 to 5.

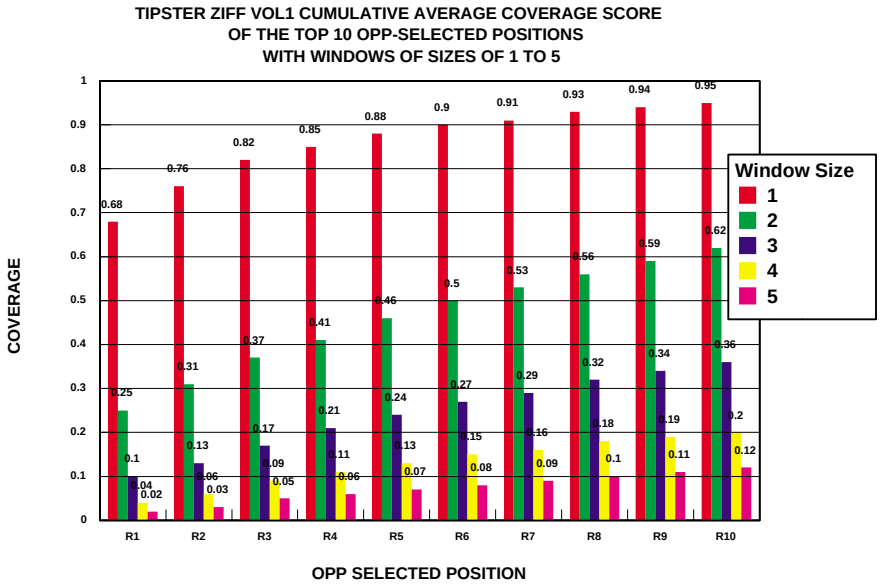


Figure 5.22: Cumulative coverage scores of top ten sentence positions selected by the OPP for windows of sizes 1 to 5.

at least a pair of windows w^S_{mi} and w^E_{mj} such that $w^S_{mi} = w^E_{mj}$. C_m estimates the potential of the OPP procedure. Figure 5.22 shows the cumulative average coverage scores of the top ten sentence positions of the training set following the OPP in Table 5.3. Figure 5.22 indicates that 68% of sentences in S shared with the title sentence at least one word, 25% two words, 10% three words, 4% four words, and 2% five words. The amount of sharing at least one word increases to 88% if we choose the top 5 positions of the OPP, and to 95% if we choose the top 10 positions.

The contribution to C^o_m solely from m-word matches between E and S can be computed as follows:

$$C^o_m = C_m - C_{m-1}$$

The result is shown in Figure 5.23. Notice that the topmost segment of each column in Figure 5.23 represents contribution from matches of *at least* five word long, since we have C_m only up to $m = 5$. The average number of sentences per summary (SPS) is 5.76. If we choose the top 5 sentence positions according to the OPP, Figure 5.23 tells us that this 5-sentence extract E, which is the length of an average summary,

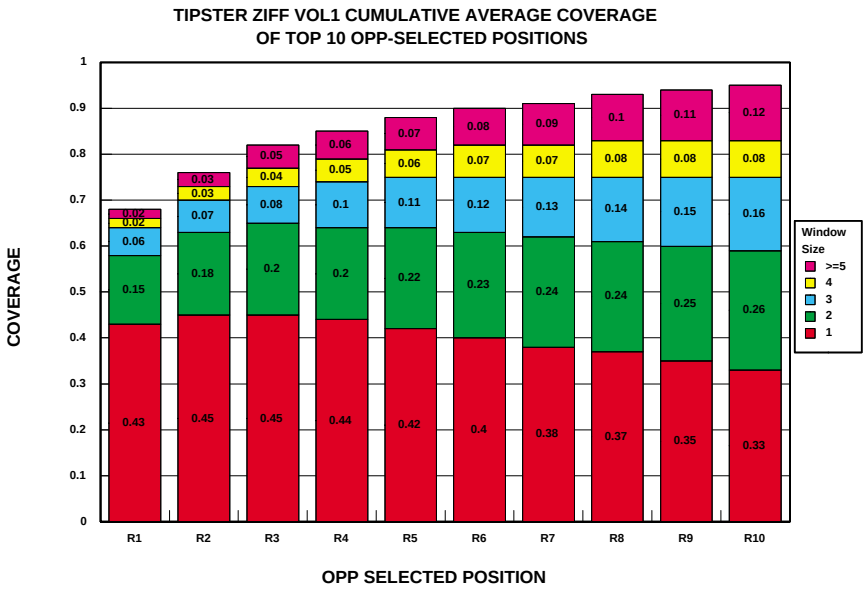


Figure 5.23: Cumulative coverage scores of top ten sentence positions with contribution from each window size given separately.

covers 88% of S, in which 42% solely comes from one-word matches, 22% two-word, 11% three-word, and 6% four-word.

The average number of sentences per text of the TIPSTER ZIFF domain is about 70, since SPP is 2.05 and PPT is 34.43. If we produce an extract of about 10% of the average length of a ZIFF text, i.e., 7 sentences, the coverage score is 0.91. This result is extremely promising and confirms the OPP-selected extract bears important contents.

The coverage measure only requires that sentences in E and S have at least one window in common. It is indeed a very loose way to measure similarity, although we have run the evaluation in different window sizes. It would be interesting to see what percentage of the sentences in S are covered by the OPP-selected extract. Table 5.5.2.2 gives examples of computing coverage scores of three ZIFF texts. The first row for each test run is the position label of sentence in the corresponding summary and the first column is the number of positions selected by the OPP. The C column lists cumulative coverage score up to the number of positions selected indicated in the first column. The values of columns after column C are cumulative

hit scores for each summary sentence. The *hit* scores shown in Table 5.5.2.2 actually are very good, since many of them exceed 50%. This result further confirms the value of the OPP procedure.

5.6 Conclusions

This chapter describes a topic identification method, the Optimal Position Policy, that is based on the position hypothesis. The position hypothesis assumes that the position of a sentence in a text can be used as an indicator of its relative importance in the text. Although the position hypothesis has been used for a long time, there are no systematic studies to verify this idea and evaluate it on large real world text collections. The Optimal Position Policy introduced in this chapter not only presents a complete solution to verify and evaluate the position hypothesis, it also provides a reliable method of how to discover important positions for various domains.

The Optimal Position Policy uses human-prepared topic indices of a set of training texts of the intended domain and identifies the sentence positions where these topics are most likely to be located. For experiments run on 13,000 texts of the TIPSTER ZIFF Vol. 1 and 2,907 texts of the TIPSTER ZIFF Vol. 2 collections, the Optimal Position Policy suggests that the title sentence is extremely important, followed by the first sentences of the second and the third paragraphs. It also confirms that the first sentence of each paragraph really carries more important information than other sentences in a text, as shown in Figures 5.14 and 5.16, and indicates that, for this domain, the last paragraph holds no special importance. One major advantage of using the Optimal Position Policy is that all the sentence positions are ranked according to their *dhit* scores. This provides us the ability to direct the selection of sentences from a text with meaningful guidance whenever a text extract is needed. *Dhit* measures the number of different topics expected in a specific unit of text such as a sentence, a paragraph, or a full text. It is essentially records the recall of topics in these text units. Figures 5.24 and 5.25 indicate that applying the first sentence selection policy without the help of the Optimal Position Policy can reach the same level of performance ($dhit = 62.71\%$) as the Optimal Position Policy ($dhit = 63.62\%$) over 10% (7 sentences) of the length of an average ZIFF text. However, there is a significant difference in performance if only two or three

SZF_001/ZF109-664-691

	C	1	2	3	4	5	6	7	8	9
1	**0.44**	**0.53**	**0.13**	0	0	0	0	**0.22**	0	**0.11**
2	**0.56**	0.53	0.13	0	**0.11**	0	0	0.22	0	0.11
3	**0.67**	**0.59**	0.13	0	0.11	0	0	**0.33**	**0.08**	0.11
4	**0.78**	**0.65**	0.13	**0.63**	**0.22**	0	0	0.33	**0.17**	0.11
5	0.78	**0.71**	0.13	0.63	0.22	0	0	**0.89**	0.17	0.11
6	0.78	0.71	0.13	0.63	0.22	0	0	0.89	0.17	0.11
7	0.78	0.71	0.13	0.63	0.22	0	0	0.89	0.17	0.11
8	0.78	0.71	0.13	0.63	0.22	0	0	0.89	**0.5**	0.11
9	0.78	0.71	0.13	0.63	0.22	0	0	0.89	0.5	0.11
10	0.78	0.71	0.13	0.63	0.22	0	0	0.89	0.5	0.11

SZF_001/ZF109-665-335

	C	1	2	3	4	5	6	7	8	
1	**0.5**	**0.62**	**0.1**	0	**0.09**	0	0	**0.18**	0	
2	**0.88**	**0.77**	**0.3**	**0.08**	0.09	0	**0.17**	0.18	**0.5**	
3	0.88	0.77	0.3	**0.15**	**0.18**	0	0.17	0.18	0.5	
4	0.88	0.77	0.3	0.15	0.18	0	**0.5**	0.18	0.5	
5	0.88	0.77	**0.4**	0.15	0.18	0	0.5	0.18	0.5	
6	0.88	0.77	0.4	0.15	0.18	0	0.5	0.18	0.5	
7	0.88	0.77	0.4	0.15	**0.27**	0	0.5	0.18	0.5	
8	0.88	0.77	0.4	0.15	0.27	0	**0.67**	0.18	0.5	
9	0.88	0.77	0.4	0.15	0.27	0	0.67	0.18	0.5	
10	0.88	0.77	**0.5**	0.15	0.27	0	**0.83**	0.18	**0.75**	

SZF_001/ZF109-665-495

	C	1	2	3	4	5	6	7	8	
1	**0.88**	**0.89**	**0.44**	**0.13**	**0.17**	**0.13**	**1**	0	**0.13**	
2	0.88	0.89	0.44	**0.19**	0.17	0.13	1	0	0.13	
3	0.88	**1**	0.44	**0.63**	0.17	0.13	1	0	0.13	
4	0.88	1	0.44	**0.75**	0.17	0.13	1	0	0.13	
5	0.88	1	0.44	0.75	0.17	0.13	1	0	0.13	
6	0.88	1	0.44	**0.81**	0.17	0.13	1	0	0.13	
7	0.88	1	0.44	0.81	0.17	0.13	1	0	0.13	
8	0.88	1	0.44	0.81	0.17	0.13	1	0	0.13	
9	0.88	1	0.44	0.81	0.17	**0.25**	1	0	0.13	
10	**1**	1	0.44	0.81	**0.83**	0.25	1	**0.07**	0.13	

Table 5.6: Details of computing coverage score of window of size 1. The first row for each sample is the position label of a sentence in the corresponding summary and the first column is the number of positions selected by the OPP. The C column lists cumulative coverage scores up to the number of positions selected indicated in the first column. The values of columns after column C are cumulative *hit* scores for each summary sentence individually. Boldfaced digits indicate positions of new hits. They spread fairly uniformly, so no obvious improvement to OPP strategy is apparent.

sentences are selected. The precision and recall scores of the comparison between the human prepared summaries and the Optimal Position Policy selected extracts have very good balance. The precision score decreases slowly while the recall score increases quickly. These results confirm that the Optimal Position Policy is useful and promising.

One requirement of the Optimal Position Policy approach is that it needs human-prepared keywords of the texts in the domain. Though it may seem more straightforward to use human-selected topic sentences to do the training, human-selected topic sentences are not as widely available as keywords (title indices) are. Furthermore, it seems much easier to automatically handle topics than whole topic sentences when indexed data are not available.

In our experiments, no morphological transformation is performed to canonicalize words to their root forms. According to the *dhit* graphs shown in Figures 5.24 and 5.25, still about 25% of topics remain to be covered even when 31 sentences have been already selected. 31 sentences is 43% of the average length of a ZIFF text. Topics such as *size* and *Small Computer System Interface* are never recovered in the original text. *Size* is used to generalize all the descriptions about the dimensions of a product. *Small Computer System Interface* appears in the original text as *SCSI*. To improve the policy performance, morphological transformation, some kinds of concept taxonomy, and acronym processing module will help.

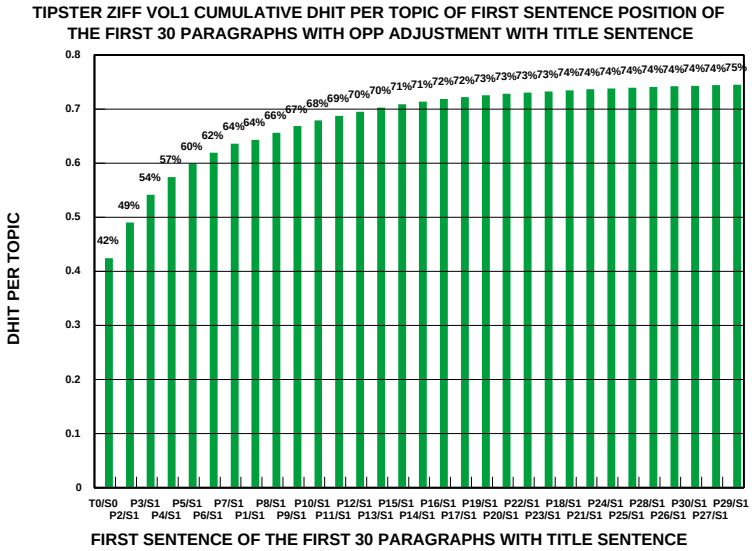


Figure 5.24: Cumulative *dhit* per topic for the first sentence of the first 30 paragraphs, following the OPP.

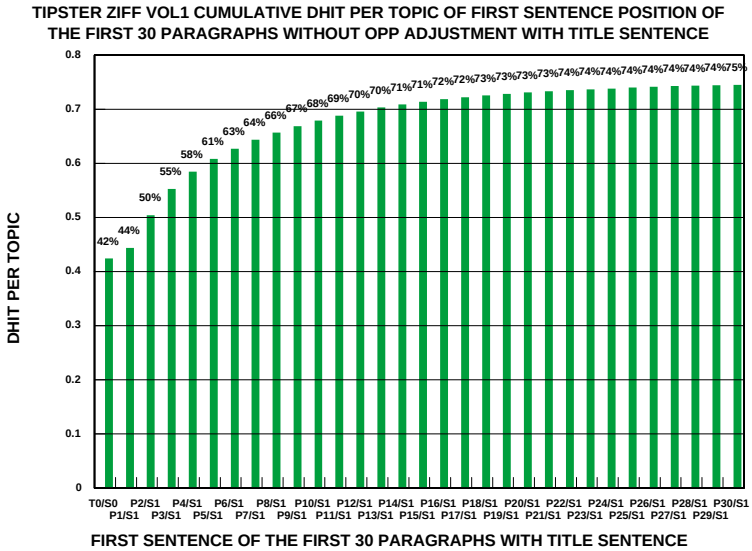


Figure 5.25: Cumulative *dhit* per topic for the first sentence of the first 30 paragraphs, not following the OPP.

Chapter 6

Conclusions

6.1 Review of Thesis

To solve the information overload problem, we increasely rely on automatic text processing techniques such as automated text summarization, text routing, etc. One central step of these techniques is topic identification. A topic is the main idea of what we talk or write about. Although as discussed in Chapter 1 no formal liguistic definition of topic exists in the literature, we described in this thesis three new methods to identify topics of texts in an attempt to empirically determine topicality.

Topic identification is a central and important step for many text processing tasks. For example, Hovy and Lin [33] described an automated text summarization system SUMMARIST that specifically proposed topic identification as the first step to generate summaries. Text categorization is another example. As we demonstrated in Chapter 4, texts can be categorized into their specific topics once their topics are determined by a topic identification method. Other tasks such as text segmentation and automatic text indexing are also straightforward applications. Text segmentation by topic bounaries was attemped by Morris and Hirst [59] and Hearst [29]. Automated text indexing has long been an common practice in Information Retrieval. With current level of success in the topic identification technique, we feel encouraged to continue our research along this line. However, improvements and extensions to individual techniques are required.

The first method described in Chapter 3 extends the word counting method used in Information Retrieval to concept counting. Concept counting uses a concept taxonomy, the Knowledge Kernel, derived from WordNet [57], together with the

Penman Upper Model [2]. The Knowledge Kernel provides external information for generalizing specific concepts into general concepts. This cannot be done by using within-document word frequency information alone. To select appropriate concepts that are both *general* enough and *specific*[1] for generalization, we introduced a new algorithm that generates *interesting wavefronts* consisting of concepts suitable for text topics at different degrees of generality. To measure degree of generality of a concept, we define two new parameters, the *branch ratio threshold* and the *starting depth*.

The concept counting method only generalizes concepts recorded in the concept taxonomy. Therefore, it would miss complex topics such as *counseling* and *robbery* if these are not in the concept taxonomy. The second method discussed in Chapter 4 introduces *multi-level topic signatures* that are automatically trained from text collections, categorized by topics of interest to represent and identify these complex concepts (topics) in texts. To evaluate this method, we trained 32 topic signatures over 16,137 Wall Street Journal texts and tested them on a collection of another set of 12,906 Wall Street Journal texts. The recall and precision results are promising. We also showed that normalizing *idf* according to the number of texts per topic category improved recall and precision scores. In addition, we showed evidence for multiple topics in texts that tends to decrease recall and precision scores.

The third method, the Position Method, is discussed in Chapter 5. It is based on the assumption that, when the text exhibits certain regularites of structure the (ordinal) position of a sentence in the text is related to its importance. Although, the Position Hypothesis has been used implicitly or explictly by many researchers, they do not agree on which sentence positions bear more information content, and they do not quantitatively determine likely positions. It is clear that the association of significant content with sentence position in a text is genre or domain dependent. Therefore, it is necessary to derive the correspondence between significant content and sentence position on a genre by genre basis. This requires automated learning. To achieve this, we demonstrated how to verify the Position Hypothesis for a specific

[1] A overgeneral concept such as *thing* enough does not serve as a good topic for a text, while an overspecific concept that falls back to the orginal concept occurring in a text defeats the purpose of general topic identification.

domain empirically. We also described a method to derive an Optimal Position Policy for selecting sentences according to their *dhit* scores. To evalute the effectiveness the Position Method quantitatively, we computed precision and recall scores, which indicate the peformance lower bound of the Position Method, and coverage score, which indicate the performance upper bound of the Position Method.

6.2 Details of Future Work

While three methods described in Chapters 3, 4, and 5 are interesting, none of them are the last word on the subject. Many extensions, variations, and improvements are possible. It is a rich area for further study. In this section, we outline some of the more immediate extensions that could be performed on each method.

6.2.1 Concept Generalization

The current concept generalization algorithm does not utilize information about the adjective and adverb hierarchies of WordNet. One reason for not using the adjective and adverb hierarchies is that they have different structures from the noun and verb hierarchies. The current concept generalization algorithm needs to be modified to use the adjective and adverb relations in the different structures. More detail about the adjective and adverb hierarchies and a possible solution is described later. Another reason for leaving the adjective and adverb hierarchies out is that we assume that nouns and verbs are more likely to be the topic candidates. Hahn [25] also adopted a similar assumption in implementation of the TOPIC semantic parser.

Although using nouns and verbs alone may be a good and economic decision for our initial exploration of the concept generation method, it is desirable to be able to consider the extra information provided by adjectives and adverbs in the future. For example, an article about *wetness* may use adjectives such as *watery*, *damp*, and *moist*; while using *dried-up*, *sere*, and *anhydrous* to describe *dry*. The relation between *wet* and *dry* is called *antonymy* and adjectives similar to *wet* and *dry* respectively form a *bipolar* cluster. To capture these details about *wet* and *dry*, we need to extend the current concept generalization method to utilize the bipolar structure. A simple solution is to generalize bipolar adjectives by their bipolar

centers such as *wet* and *dry*. How much performance improvement can be achieved by utilizing this extra information is an interesting topic to explore in the future.

Several other straightforward extensions are also necessary. For example, one could include hierachies other than hypernym; use within-collection concept frequency distribution such as *idf*, which is proven to be effective from word-based Information Retrieval; add a sense disambiguation mechanism; test the concept generalization algorithm on a larger collection of texts; automatically add unkown words such as person names, company names, etc., to hierachies; and try the concept generalization algorithm with a richer taxonomy such as CYC or SENSUS. We expect the results of all these kinds of improvements to increase the quality of the output topics and performance of the concept generalization algorithm.

Since the utility of the concept generalization algorithm depends on the supporting concept hierarchies or knowledge bases, the problem of reconfiguring or augmenting the hierarchies is a very important next step. One possible approach is using topic signatures, which we discuss in the next section.

6.2.2 Topic Signatures

Since topic signatures are derived from a training corpus, they carry domain-specific information. This information can be used to expand a domain-independent concept taxonomy such as WordNet. The expanded concept taxonomy will better serve applications designed for that specific domain. Although we have success in building topic signatures and using them in text categorization task, we need thousands or even tens of thousands of signatures to make topic signatures useful for other tasks such as text summarization. It will be very interesting to see how topic signatures scale up. To organize them, topic signature hierarchy may be necessary.

It will be very useful also to test the confusion sets and multi-level topic signatures on TREC topics, so as to perform a direct comparison of these methods with TREC systems.

Another extension is to use relative threshold-based topic assignment methods to determine multiple possible topics for a text. For example, we can use experimenttally determined *cutoff* and *equivalent* thresholds to decide what the most likely topics of a text are. The cutoff threshold θ^c is the lowest acceptable similarity value

for a topic to be considered as a possible topic candidate. The equivalence threshold θ^e is the maximum difference for two topics T_i and T_j to be considered as *equivalent*, i.e., if T_i is selected as a topic then T_j has to be selected and vice versa. Therefore, a set of topics for a text can be determined by first filtering out all the topics with similarity values less than θ^c, then selecting a topic, for example T_0, that has the maximum similarity among the remaining topics, and finally including all the topics that are equivalent to T_0. The relative threshold method may prove to be valuable when we test the topic signature method on TREC collections, since some texts in the TREC collcetions are very long, and long texts intuitively tend to contain multiple topics.

The current topic signature construction method requires knowing topic categories beforehand. We plan to incorporate clustering techniques to generate topic signatures from automatically clustered categories. The probablistic classification algorithm used in AutoClass [9] is a good starting point for future exploration.

6.2.3 The Position Method

As described in Chapter 5, we used Ziff-Davis computer articles to test the Optimal Position Policy. Since the OPP is genre dependent, it will be interesting to perform similar investigation on other genres such as *Wall Street Journal*, *Federal Register*, etc., and to compare them.

We measured sentence yield and *dhit* by absolute paragraph and sentence positions in a text. One question is: will we get similar results when measuring sentence yield and *dhit* by relative paragraph and sentence positons in a text? We need to investigate the effect of normalizing paragraph and sentence positions.

One major weakness of the current OPP is that it relies on topic keywords. If topic keywords are not available, we cannot construct an OPP automatically. To remedy this weakness, we plan to perform experiments similar to Paijmans [65] to learn the possibility of using $tf * idf$ or χ^2 selected keywords as topic keywords. If we find that $tf * idf$ or χ^2 selected keywords cluster at certain paragraph or sentence positions for a genre, we can use these automatically generated topic keywords to generate the OPP for the genre.

6.3 Other Topic Identification Methods

The three topic identification modules described in this thesis use different cues to discover topics of texts. Other alternatives are also possible. *Cue phrases* such as *it is important* ... and *the most effective* ... can be used to pinpoint possible important content.

More sophisticated methods using discourse relations or structures such as *lexical chains* proposed by Morris and Hirst [59], *TextTile* by Hearst [30], and *rhetorical parsing* by Marcu [55] are mentioned in Section 2.6. All these methods could be integrated with the three techniques described in this thesis to provide suggestions and cross-validate various possible topics of a text.

6.4 Integrating Topic Identification Methods

Simply having a set of separate topic identification methods is obviously not enough. Since different methods employ different knowledge and cues, they will tend to complement each other. By integrating them into a single Topic Identification system, they can be used to compensate one another's weaknesses. Figure 6.1 shows a possible configuration of such an integrated multi-evidence topic identification system. A text is first preprocessed into some internal representation and then passed through the different topic identification modules in parallel. Each module delivers the text, in which specific portions have been delimited and given an importance/topicality score. The results from each module are then combined, using some weighting scheme, and the topic(s) of the text is(are) identified by using the integrated decision from multiple sources.

Different engines will delimit different spans of text, and will rate regions on very different grounds. One can thus expect that a span of text rated highly topical by more than one method has a good chance of being so. Although how to weight evidences from multiple sources is not clear, we showed that even using the three topic identification techniques introduced in this thesis alone can achieve promising results. We expect the performance of the integrated system to be better.

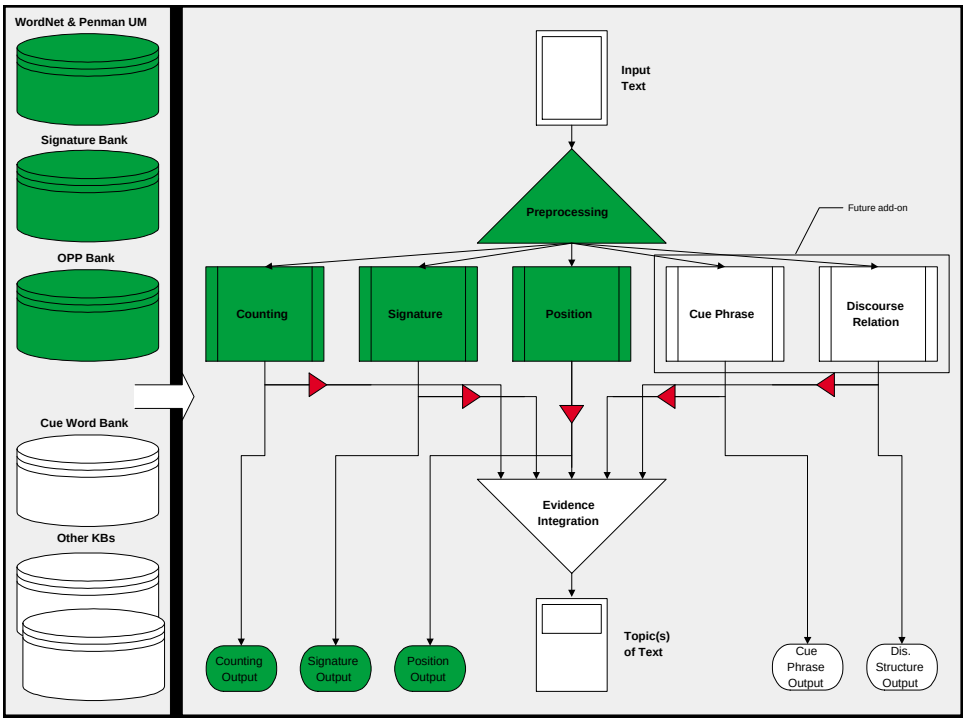


Figure 6.1: Organization of multi-evidence topic identification system.

A simple integration scheme is the following:

$$S = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3 + \cdots$$

where s_i is the score for term i from each topic identification engine and the relative importance for each engine is determined by α_i which can be determined by experiments.

Other alternatives are possible, like running some modules before others, as filters. For example, perform concept generalization on, or construct topic signatures from, important positions selected by the Postion Method. How these combinations affect the system performance and complement each other will be the major reseach focus of future work.

Reference List

[1] Chidanand Apté, Fred Damerau, and Sholom M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–51, July 1994.

[2] John A. Bateman, Kasper, Robert T., Johanna D. Moore, and Richard A. Whitney. A general organization of knowledge for natural language processing: the penman upper model. Technical report, Information Science Institute/USC, Marina del Rey, California, March 1990.

[3] P. B. Baxendale. Machine-made index for technical literature — an experiment. *IBM Journal*, pages 354–61, October 1958.

[4] Richard Braddock. The frequency and placement of topic sentences in expository prose. *Research in The Teaching of English*, 8:287–302, 1974.

[5] R. Brandow, K. Mitze, and L.F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–85, 1995.

[6] E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, University of Pennsylvania, Philadelphia, Pennsylvania, 1993.

[7] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 233–37. Association for Computational Linguistics, 1992.

[8] Gillian Brown and George Yule. *Discourse Analysis*. Cambridge University Press, Cambridge, UK, 1983.

[9] Peter Cheeseman, James Kelly, Matthew Self, John Stutz, Will Taylor, and Don Freeman. Autoclass: A Bayesian classification system. In *Proceedings of the 5th International Conference on Machine Learning*, pages 54–64, 1988.

[10] W. Bruce Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, pages 8–13, April 1993.

[11] Robert de Beaugrande. *Text, Discourse and Process*. Longman, London, UK, 1980.

[12] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, volume 41(6), pages 391–407. ASIS, 1990.

[13] Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995.

[14] Gerald DeJong. *Skimming stories in real time.* PhD thesis, Yale University, New Haven, 1979.

[15] Gerald DeJong. An overview of the frump system. In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for natural language processing*, pages 149–76. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1982.

[16] Dan Dolan. Locating main ideas in history textbooks. *Journal of Reading*, pages 135–40, 1980.

[17] H. P. Edmundson. Problems in automatic abstracting. *Communication of the ACM*, 7(4):259–63, April 1964.

[18] H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–85, 1969.

[19] Norbert Fuhr, Stephan Hartmann, Gerhard Lustig, and Gerhard Knorz. AIR/X — a rule-based mutlistage indexing system for large subject fields. In *Proceedings of the RIAO'91*, pages 606–23, 1991.

[20] Fumiyo Fukumoto, Yoshimi Suzuki, and Jun'ichi Fukumoto. An automatic extraction of key paragraph based on context dependency. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 291–98, 1997.

[21] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the 16th International Joint Conference*, pages 233–37, Los Altos, California, 1992. IJCAI, Morgan Kaufmann.

[22] Anthony Frank Gallippi. *A Machine Learning Approach to Multilingual Proper Name Recognition.* PhD thesis, Department of Electrical Engineering, University of Southern California, Los Angeles, California, Dec 1996.

[23] Government Information Office Taiwan. Dragon boat festival. `http://www.gio.gov.tw/info/festival_c/dragon/dragon_e.htm`, 1996.

[24] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Boston, 1994.

[25] Udo Hahn. Making understanders out of parsers: Semantically driven parsing as a key concept for realistic text understanding applications. *Internation Journal of Intelligent Systems*, 4:345–93, 1989.

[26] Udo Hahn. TOPIC parsing: Accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1):135–70, 1990.

[27] Donna Harman. Data preparation. In *The Proceedings of the TIPSTER Text Program Phase I*, San Mateo, California, 1994. Morgan Kaufmann Publishing Co.

[28] Philip Hayes, Peggy M. Andersen, Irene B. Nirenburg, and Linda M. Schmandt. Tcs: A shell for content-based text categorization. In *Proceedings of the Sixth IEEE CAIA*, pages 320–26, 1990.

[29] Marti Hearst. Using categories to provide context for full-text retrieval results. In *Proceedings of the RIAO '94*, Rockefeller, NY, 1994.

[30] Marti A. Hearst. *Context and Structure in Automated Full-Text Information Access.* PhD thesis, Computer Sicence Division, University of California at Berkeley, Berkeley, California, April 1994.

[31] J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–30, 1993.

[32] Charles F. Hockett. *A Course in Modern Linguistics.* Macmillan, New York, 1958.

[33] Eduard Hovy and Chin-Yew Lin. Automated text summarization in SUMMARIST. In *ACL/EACL97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, 1997.

[34] Richard Hudson. *Word Meaning.* Routledge, London, U.K., 1995.

[35] David A. Hull. *Information Retrieval Using Statiscal Classification.* PhD thesis, Stanford University, Palo Alto, California, 1994.

[36] Paul S. Jacobs. Using statistical methods to improve knowledge-based news categorization. *IEEE Expert*, pages 13–23, April 1993.

[37] Paul S. Jacobs and Lias F. Rau. SCISOR: Extracting information from on-line news. *Communication of the ACM*, 33(11):88–97, November 1990.

[38] Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization. Technical report, Carnegie Mellon University, Pittsburgh, PA, 1996.

[39] Elinor Ochs Keenan and Bambi Schieffelin. Topic as a discourse notion: A study of topic in the conversions of children and adults. In Charles N. Li, editor, *Subject and Topic*, pages 335–84. Academic Press, New York, 1976.

[40] D.E. Kieras. Thematic process in the comprehension of technical prose. In B.K. Britton and J.B. Black, editors, *Understanding Expository Text: A Theoretical And Practical Handbook for Analyzing Explanatory Text*, pages 89–108. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1985.

[41] Kevin Knight and S. Luk. Building a large knowledge base for machine translation. In *Proceedings of the American Association for Artificial Intelligence Conference AAAI-94*, Seattle, WA, 1994.

[42] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Coference on Research and Development in Information Retrieval*, New York, NY, 1996. ACM.

[43] Wendy Lehnert and C. Loiselle. An introduction to plot unit. In David Waltz, editor, *Semantic Structures — Advances in Natural Language Processing*, chapter 3, pages 88–111. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

[44] Wendy G. Lehnert. Plot units: A narrative summarization strategy. In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for natural language processing*, pages 375–412. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1982.

[45] Wendy G. Lehnert. Symbolic/subsymbolic sentence analysis: Exploiting the best of two worlds. In *Advances in Connectionist and Neural Computational Theory*, volume 1, pages 135–64. Ablex Publishers, Norwood, New Jersey, 1991.

[46] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems.* Addison-Wesley, Reading, MA, 1990.

[47] David D. Lewis and M Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annal Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994.

[48] Elizabeth Liddy, Woojin Paik, and Edmund S. Yu. Text categorization for multiple users based on semantic features from a machine-readable dictionary. *ACM Transactions on Information Systems*, 12(3):278–95, July 1994.

[49] Elizabeth D. Liddy and Sung H. Myaeng. DR-LINK's linguistic-conceptual approach to document detection. In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 113–29, 1992.

[50] Chin-Yew Lin. Knowledge-based automated topic identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 308–10. Association for Computational Linguistics, 1995.

[51] Chin-Yew Lin. Identify topics by concept signatures. Technical report, Information Sciences Institute, Marina del Rey, CA, 1997.

[52] *Longman Dictionary of Comtemporary English.* Longman Group, Essex, UK, 1978.

[53] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal*, pages 309–17, October 1957.

[54] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, pages 159–65, April 1958.

[55] Daniel Marcu. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Madrid, Spain, July 1997. Association for Computational Linguistics.

[56] Michael L. Mauldin. *Conceptual Information Retrieval — A Case Study in Adaptive Partial Parsing.* Kluwer Academic Publishers, Boston, 1991.

[57] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. CSL Report 43, Cognitive Science Labortory, Princeton University, Princeton, July 1990.

[58] J.L. Morgan. Some remarks on the nature of sentences. In *Papers from the Parasession on Functionalism.* Chicago Linguistic Society, Chicago, 1975.

[59] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[60] National Institute of Standards and Technology. *The First Text REtrieval Conference (TREC-1)*, March 1993.

[61] National Institute of Standards and Technology. *The Fourth Text REtrieval Conference (TREC-4)*, March 1996.

[62] Tadashi Nomoto and Yuji Matsumoto. Exploring text structure for topic identification. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 101–12. Association for Computational Linguistics, 1996.

[63] Chris D. Paice. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, *Information Retrieval Research*, pages 172–91. Butterworths, 1981.

[64] Chris D. Paice. Constructing literature abstracts by computer: Techinques and prospects. *Information Processing and Management*, 26(1):171–86, 1990.

[65] J.J. Paijmans. Relative weights of words in documents. In L.G.M. Noordman and W.A.M. de Vroomen, editors, *Conference Proceedings of STINFON*, pages 195–208, 1994.

[66] Rebecca J. Passoneau and Diane J. Litman. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices. In E.H. Hovy and D.R. Scott, editors, *Computational and Conversational Discourse: Buring Issues in Discourse*, pages 161–94. Springer Verlag, Heidelberg, Germany, 1996.

[67] PENMAN. The PENMAN documentation and user guide. The PENMAN project, USC/Information Science Institute, Marina del Rey, California, 1989.

[68] J. R. Quinlan. Learning efficient classification procedures and their application to chess and games. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann, San Francisco, CA, 1983.

[69] Lias F. Rau. Extracting company names from text. In *Proceedings of the Conference in AI Applications*, volume 1, pages 29–32. IEEE, 1991.

[70] Philip Stuart Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, Philadelphia, Pennsylvania, 1993.

[71] Ellen Riloff and Wendy Lehnert. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12(3):296–333, July 1994.

[72] Gerald Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, N.J., 1971.

[73] Gerald Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of Association for Computing Machinery*, 15(1):8–36, 1968.

[74] Gerard Salton. *Automatic Text Processing*. Addison-Wesley, Reading, Massachusetts, 1988.

[75] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–26, June 1994.

[76] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, pages 513–23, 1988.

[77] Gerard Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–72, 1973.

[78] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[79] Roger C. Schank. Rules and topics in conversion. *Cognitive Science*, pages 421–42, 1977.

[80] Andrew F. Siegel and Charles J. Morgan. *Statistics and Data Analysis: an introduction.* J. Wiley, New York, 1996.

[81] John Sinclair, Patrick Hanks, Gwyneth Fox, Rosamund Moon, and Penny Stock. *Collins COBUILD English Language Dictionary.* William Collins Sons & Co. Ltd., Glasgow, UK, 1987.

[82] Harry Singer and Dan Dolan. *Reading And Learning from Text.* Little Brown, Boston, Mass., 1980.

[83] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):111–21, 1972.

[84] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2 edition, 1979.

[85] Yasuhiko Watanabe, Masaki Murata, Masahito Takeuchi, and Makoto Nagao. Document classification using domain specific kanji characters extracted by χ^2 method. In *Proceedings of the 16th COLING*, pages 794–99, 1996.

[86] Erik Wiener, Jan O. Pedersen, and Andreas S. Weigend. A neural network approach to topic spotting. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, 1995.

[87] Yiming Yang and John Wilbur. Using corpus statistics to remove redundant words in text categorization. *Journal of American Society for Information Science*, 47(5):357–69, May 1996.

[88] David Yarowsky. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Liguistics*, pages 454–60, Nantes, France, 1992.

[89] Klaus Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of the 16th COLING*, pages 986–89, 1996.

Appendix A

Wall Street Journal Corpus Statistics

TEST CODE:
`MIN`: minimum
`1QT`: first quartile
`MED`: median
`MAX`: maximum
θ: outlier threshold
`3QT`: third quartile

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	194	19	59	99	265	574	2247
aro	207	13	88	146	287	585	1745
aut	284	19	105	197	403	850	1701
bbk	67	18	39	55	77	134	849
bcy	183	15	78	120	174	318	2055
bnk	271	28	115	204	376	767	1780
bon	128	14	51	88	174	358	1006
ceo	180	25	156	175	201	268	455
cmd	323	27	115	247	458	972	1349
div	70	16	40	53	82	145	608
eco	371	31	181	344	485	941	1485
edp	349	30	123	283	460	965	1639
ele	275	46	104	188	420	894	1104
env	266	33	119	226	367	739	1026
ern	147	20	90	126	179	312	664
fab	311	23	119	222	424	881	1521
fin	352	21	146	303	503	1038	1692
lng	148	39	65	116	178	347	724
min	159	20	70	86	163	302	1278
mkt	348	53	145	269	535	1120	1232
mon	210	13	41	88	323	746	1556
pet	250	21	75	130	368	807	1423
pha	332	26	110	244	466	1000	2196
pub	356	33	126	273	533	1143	1242
rel	237	20	72	135	359	789	1114
ret	414	29	129	330	647	1424	1877
scr	252	20	101	182	354	733	1268
stk	107	17	45	58	91	160	2329
tel	282	31	116	216	395	813	1344
tnm	98	17	54	77	114	204	1132
tra	269	23	111	222	349	706	1098
uti	157	28	64	93	167	321	1277
Average	237	24	81	174	274	667	1375

Table A.1: Wall Street Journal 1987 number of terms per text per topic distribution, where terms are words not in the stop list and without any morphological transformation.

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	194	19	59	99	265	574	2242
aro	206	13	88	146	286	583	1743
aut	284	19	105	197	403	850	1712
bbk	67	18	39	55	77	134	850
bcy	183	15	79	120	173	314	2054
bnk	270	28	115	204	376	767	1776
bon	129	14	51	89	174	358	1007
ceo	180	25	156	174	201	268	455
cmd	322	28	115	247	457	970	1333
div	70	16	40	53	82	145	603
eco	370	31	180	342	485	942	1480
edp	349	30	123	283	459	963	1636
ele	275	46	104	188	419	891	1103
env	266	33	119	224	367	739	1023
ern	146	20	90	126	178	310	661
fab	311	23	119	222	423	879	1518
fin	352	21	146	301	502	1036	1687
lng	148	39	66	116	177	343	722
min	159	20	69	86	163	304	1276
mkt	348	53	145	269	531	1110	1226
mon	209	13	41	88	322	743	1555
pet	250	21	75	129	368	807	1421
pha	331	26	110	244	467	1002	2185
pub	356	33	125	272	531	1140	1238
rel	237	20	72	135	358	787	1113
ret	414	29	129	330	648	1426	1876
scr	251	21	101	182	353	731	1266
stk	107	17	45	58	92	162	2331
tel	282	31	115	216	395	815	1343
tnm	98	17	55	77	114	202	1130
tra	268	23	111	222	349	706	1097
uti	157	28	64	93	166	319	1273
Average	237	24	81	174	274	666	1372

Table A.2: Wall Street Journal (1987: morphologically normalized) number of terms per text per topic distribution.

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	192	17	59	98	261	564	2192
aro	201	11	87	141	281	572	1696
aut	281	18	103	194	399	843	1707
bbk	65	16	37	52	73	127	836
bcy	179	15	77	118	168	304	2012
bnk	264	26	113	196	365	743	1732
bon	126	14	50	87	169	347	969
ceo	176	25	152	170	196	262	446
cmd	318	27	113	243	444	940	1309
div	67	16	36	50	79	143	574
eco	362	31	178	330	478	928	1476
edp	342	30	120	276	446	935	1608
ele	270	43	99	189	413	884	1083
env	260	33	120	217	358	715	1000
ern	143	20	87	123	173	302	654
fab	305	23	117	215	414	859	1512
fin	345	21	138	293	496	1033	1660
lng	143	36	63	112	170	330	707
min	156	18	66	85	159	298	1258
mkt	341	52	137	262	523	1102	1210
mon	205	13	39	85	316	731	1517
pet	244	19	73	125	363	798	1375
pha	327	27	106	241	462	996	2180
pub	352	31	121	271	532	1148	1234
rel	231	20	67	129	348	769	1106
ret	406	28	126	315	636	1401	1822
scr	245	21	94	178	345	721	1237
stk	104	15	42	56	88	157	2319
tel	277	29	112	213	384	792	1325
tnm	95	16	53	74	110	195	1103
tra	263	22	106	216	338	686	1084
uti	153	27	61	90	162	313	1260
Average	232	23	78	170	268	654	1350

Table A.3: Wall Street Journal (1987: phrases) number of terms per text per topic distribution.

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	198	23	56	102	281	618	1936
aro	249	10	87	168	333	702	1298
aut	254	17	97	168	313	637	2733
bbk	69	16	39	54	83	149	607
bcy	146	23	78	104	143	240	884
bnk	250	28	107	203	343	697	1149
bon	163	14	58	118	221	465	1288
ceo	204	28	161	194	226	323	652
cmd	301	27	106	243	432	921	1041
div	68	14	35	54	84	157	393
eco	393	25	163	381	526	1070	1812
edp	269	17	101	227	374	783	1555
ele	251	40	95	183	356	747	1726
env	283	33	116	212	396	816	1328
ern	140	26	94	113	163	266	568
fin	298	18	122	236	425	879	1167
lng	117	43	61	92	143	266	454
min	126	18	80	85	126	195	656
mkt	320	37	116	223	460	976	1201
mon	225	13	47	120	359	827	1512
pet	226	26	88	168	302	623	1308
pha	271	26	105	216	353	725	2157
pub	308	34	122	235	454	952	1150
rel	204	21	54	109	261	571	1881
ret	340	34	107	261	442	944	1614
scr	246	20	98	180	343	710	1166
stk	154	17	44	75	148	304	811
tel	276	24	117	219	370	749	1236
tnm	97	15	56	81	108	186	1081
tra	244	27	108	170	322	643	1152
uti	143	36	79	95	170	306	687
Average	220	24	60	164	196	595	1232

Table A.4: Wall Street Journal (1988: unchanged from texts) number of terms per text per topic distribution.

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	198	23	56	102	281	618	1929
aro	248	10	87	168	332	699	1297
aut	254	17	97	168	312	634	2714
bbk	69	16	39	54	83	149	605
bcy	146	23	78	104	143	240	882
bnk	249	28	107	202	343	697	1149
bon	163	14	58	118	221	465	1293
ceo	204	28	161	194	225	321	650
cmd	301	29	106	242	430	916	1042
div	68	14	35	54	84	157	392
eco	393	25	162	380	527	1074	1811
edp	269	17	101	227	373	781	1551
ele	250	40	95	182	354	742	1725
env	283	33	116	211	395	813	1326
ern	140	26	93	113	163	268	569
fin	298	18	122	236	425	879	1165
lng	117	43	61	92	143	266	454
min	126	18	80	85	126	195	655
mkt	320	37	116	222	458	971	1196
mon	224	13	47	119	359	827	1508
pet	226	26	88	168	302	623	1309
pha	270	26	105	216	352	722	2157
pub	307	34	122	235	454	952	1148
rel	204	21	54	108	260	569	1880
ret	340	34	107	262	440	939	1614
scr	245	20	98	180	343	710	1165
stk	154	17	44	75	149	306	809
tel	276	24	117	219	370	749	1232
tnm	97	15	56	81	108	186	1078
tra	243	27	108	170	321	640	1146
uti	143	36	79	95	170	306	688
Average	220	24	60	163	196	594	1230

Table A.5: Wall Street Journal (1988: morphologically normalized) number of terms per text per topic distribution.

TOPIC	MEAN	MIN	1QT	MED	3QT	θ	MAX
air	196	22	56	100	276	606	1914
aro	243	10	82	163	325	689	1275
aut	251	17	95	168	307	625	2742
bbk	66	17	37	52	79	142	601
bcy	143	22	75	103	143	245	875
bnk	243	27	102	192	338	692	1136
bon	160	14	57	117	215	452	1256
ceo	200	28	158	189	219	310	635
cmd	297	29	104	234	425	906	1031
div	65	11	33	51	80	150	397
eco	384	23	158	377	516	1053	1757
edp	264	17	99	225	368	771	1549
ele	246	38	92	177	350	737	1680
env	278	33	113	204	387	798	1312
ern	137	26	91	109	160	263	566
fin	291	18	121	227	409	841	1132
lng	112	41	57	88	137	257	444
min	124	17	78	84	127	200	657
mkt	315	37	112	221	453	964	1196
mon	220	13	45	115	343	790	1486
pet	220	25	87	170	296	609	1266
pha	266	25	103	212	347	713	2098
pub	303	34	118	227	452	953	1136
rel	199	19	51	106	251	551	1861
ret	334	34	101	258	430	923	1594
scr	240	19	93	175	332	690	1127
stk	151	17	42	72	146	302	793
tel	271	24	114	213	357	721	1218
tnm	94	14	54	78	104	179	1049
tra	238	26	104	164	313	626	1112
uti	138	34	76	91	164	296	688
Average	215	23	58	160	191	582	1212

Table A.6: Wall Street Journal (1988: phrases) number of terms per text per topic distribution.

Appendix B

Full Topic Signatures of Test PH

AIR Topic signature (PH)							
1	airline	76	rule	151	way	226	march
2	passenger	77	hour	152	trans	227	dole
3	mile	78	on-time	153	busy	228	department_of_transportation
4	flight	79	official	154	captain	229	whether
5	air	80	consumer	155	board	230	newark
6	fare	81	line	156	administration	231	gibbs
7	airport	82	chicago	157	government	232	vice_president
8	pilot	83	braniff	158	week	233	allow
9	continental	84	fall_to	159	european	234	gate
10	carrier	85	number	160	summer	235	operation
11	traffic	86	federal	161	member	236	must
12	eastern	87	machinist	162	tiger	237	ask
13	faa	88	time	163	reduce	238	represent
14	fly	89	require	164	scheduling	239	takeoff
15	load	90	strike	165	airlines'	240	group
16	union	91	atlanta	166	work	241	midway
17	delta	92	agency	167	add	242	cause
18	american	93	plan	168	go	243	non-refundable
19	revenue	94	day	169	where	244	commuter
20	plane	95	round-trip	170	record	245	july
21	factor	96	city	171	offer	246	one-way
22	ticket	97	hijacker	172	move	247	cancel
23	texas	98	baggage	173	start	248	management
24	aviation	99	runway	174	frequent-flier	249	need
25	united	100	reservation	175	fine	250	force
26	flight_attendant	101	major	176	become	251	accord
27	seat	102	program	177	free	252	job
28	pan	103	change	178	small	253	call
29	twa	104	dallas	179	requirement	254	find
30	northwest	105	she	180	recent	255	ms.
31	delay	106	get	181	for_example	256	airplane
32	controller	107	world	182	near	257	merger
33	am	108	labor	183	bag	258	try
34	year	109	performance	184	april	259	follow
35	rise	110	occupy	185	traffic_control	260	frequent
36	percentage	111	lorenzo	186	proposal	261	recently
37	earlier	112	u.s.	187	cancellation	262	involve
38	department	113	cost	188	departure	263	hartsfield
39	safety	114	crash	189	august	264	handle
40	jet	115	southwest	190	raise	265	within
41	increase	116	report	191	company	266	lose
42	mr.	117	collision	192	more_than	267	would_be
43	pay	118	deregulation	193	high	268	inc.
44	route	119	association	194	mileage	269	ago
45	billion	120	houston	195	control	270	trip
46	month	121	late	196	match	271	trust_fund
47	service	122	flier	197	penalty	272	big
48	travel	123	take	198	maputo	273	september
49	system	124	international	199	president	274	has_been
50	complaint	125	million	200	tell	275	february
51	fill	126	begin	201	hire	276	daily
52	transportation	127	make	202	advance	277	jal
53	aircraft	128	incident	203	o'hare	278	level
54	unit	129	maintenance	204	executive	279	cargo
55	traveler	130	nwa	205	carry	280	come
56	allegis	131	low	206	restriction	281	five
57	employee	132	analyst	207	see	282	surcharge
58	schedule	133	washington	208	1986	283	accident
59	piedmont	134	even	209	denver	284	issue
60	usair	135	attendant	210	say	285	plaskett
61	problem	136	new_york	211	large	286	propose
62	spokesman	137	travel_agent	212	give	287	public
63	industry	138	advance-purchase	213	provide	288	area
64	amr	139	fuel	214	congress	289	several
65	new	140	use	215	alaska	290	price
66	express	141	june	216	procedure	291	practice
67	airway	142	cockpit	217	her	292	cut
68	contract	143	want	218	talk	293	unrestricted
69	discount	144	november	219	improve	294	leave
70	air-traffic	145	period	220	impose	295	negotiation
71	available	146	action	221	market	296	today
72	people	147	corp.	222	nation	297	marketing
73	crew	148	agreement	223	republic	298	regional
74	hub	149	landing	224	four	299	noise
75	maxsaver	150	wage	225	arrive	300	continue

Table B.1: **AIR** Topic signature in set **PH**.

ARO Topic signature (PH)							
1	contract	76	fighter	151	begin	226	time
2	air_force	77	problem	152	titan	227	soviet_union
3	navy	78	trw	153	expect	228	study
4	aircraft	79	order	154	voyager	229	early
5	army	80	gec	155	electronics	230	payment
6	space	81	spokesman	156	committee	231	1988
7	missile	82	raytheon	157	provide	232	chad
8	mcdonnell	83	company	158	ge	233	attorney
9	northrop	84	f-16	159	honeywell	234	engineer
10	pentagon	85	engineering	160	unmanned	235	navigation
11	equipment	86	suit	161	strategic	236	airline
12	nasa	87	martin	162	submarine	237	job
13	defense	88	b-1	163	force	238	more_than
14	receive	89	justice_department	164	spain	239	b-1b
15	shuttle	90	new	165	beggs	240	policy
16	boeing	91	aeronautics	166	booster	241	attack
17	douglas	92	investigation	167	astronaut	242	airplane
18	airbus	93	booster_rocket	168	commercial	243	stinger
19	thiokol	94	whitney	169	claim	244	rep.
20	plane	95	international	170	document	245	explosion
21	engine	96	administration	171	give	246	spare_part
22	award	97	israel	172	work_on	247	iran
23	rocket	98	congress	173	havilland	248	arm
24	lockheed	99	pratt	174	worker	249	whether
25	unit	100	build	175	report	250	second
26	general	101	year	176	communication	251	buy
27	total	102	develop	177	take	252	phase
28	helicopter	103	inc.	178	involve	253	criminal
29	aerospace	104	budget	179	office	254	house
30	program	105	industry	180	hercules	255	investigator
31	win	106	procurement	181	overcharge	256	find
32	dynamic	107	charge	182	base	257	india
33	government	108	design	183	france	258	divad
34	corp.	109	munition	184	schedule	259	guidance
35	military	110	united	185	pakistan	260	testing
36	million	111	make	186	fleet	261	help
37	grumman	112	redesign	187	station	262	canadair
38	system	113	employee	188	yesterday	263	business
39	u.s.	114	issue	189	prosecutor	264	ground
40	launch	115	plan	190	has_been	265	ballistic_missile
41	rockwell	116	division	191	f-18	266	nuclear
42	contractor	117	awacs	192	national	267	technical
43	production	118	research	193	analyst	268	delay
44	cost	119	nato	194	week	269	iranian
45	official	120	vehicle	195	payload	270	beech
46	mr.	121	nimrod	196	former	271	sen.
47	co.	122	case	197	would_be	272	executive
48	service	123	logistics	198	british	273	torpedo
49	weapon	124	use	199	major	274	future
50	test	125	defense_contractor	200	facility	275	review
51	jet	126	md-11	201	joint	276	request
52	project	127	fire	202	effort	277	continue
53	washington	128	improvement	203	profit	278	modification
54	radar	129	a-340	204	armed_service	279	maintenance
55	bomber	130	consortium	205	japanese	280	american
56	part	131	motor	206	center	281	aid
57	morton	132	produce	207	lease	282	f-14
58	technology	133	deliver	208	indictment	283	ariane
59	support	134	delivery	209	itt	284	tell
60	space_station	135	stealth	210	ship	285	advance
61	get	136	tank	211	westinghouse	286	add
62	soviet	137	decision	212	industrie	287	test_equipment
63	electric	138	gao	213	de	288	f-15
64	agency	139	plant	214	need	289	troop
65	challenger	140	classify	215	month	290	maker
66	satellite	141	fuel	216	late	291	grand_jury
67	defense_department	142	accident	217	congressional	292	boisjoly
68	marietta	143	competition	218	agreement	293	safety
69	work	144	estimate	219	bofors	294	allege
70	development	145	japan	220	shipboard	295	orbit
71	flight	146	lavi	221	wing	296	aspin
72	aircraft_engine	147	training	222	subcontractor	297	team
73	billion	148	hughes	223	fraud	298	believe
74	mx	149	european	224	president	299	long-range
75	electronic_equipment	150	textron	225	war	300	campaign

Table B.2: **ARO** Topic signature in set **PH**.

AUT Topic signature (PH)							
1	car	76	cut	151	owner	226	nova
2	gm	77	dodge	152	taurus	227	rover
3	ford	78	period	153	division	228	tomorrow
4	chrysler	79	american	154	local	229	average
5	motor	80	vw	155	1985	230	recent
6	plant	81	make	156	close	231	lincoln-mercury
7	model	82	corp.	157	incentive_program	232	manager
8	auto_maker	83	injury	158	unit	233	automotive
9	+	84	supply	159	high	234	bieber
10	worker	85	consumer	160	crash	235	chry-plym
11	sale	86	month	161	level	236	vice_president
12	vehicle	87	day	162	cover	237	chairman
13	truck	88	audi	163	report	238	acceleration
14	auto	89	cadillac	164	percentage	239	mo
15	union	90	canada	165	major	240	equip
16	toyota	91	hourly	166	score	241	center
17	nissan	92	spokesman	167	nummi	242	brake
18	honda	93	warranty	168	layoff	243	profit
19	production	94	import	169	sable	244	labor
20	total	95	...	170	%	245	certain
21	div.	96	.....	171	people	246	resume
22	uaw	97	overtime	172	dearborn	247	.....
23	assembly	98	.....	173	begin	248	bus
24	u.s.	99	hyundai	174	saturn	249	luxury
25	chevrolet	100	official	175	produce	250	decision
26	mr.	101	.....	176	wage	251	come
27	0	102	canadian	177	affect	252	rather
28	dealer	103	mazda	178	more_than	253	say
29	year	104	co.	179	accident	254	reach
30	recall	105	output	180	corp	255	magna
31	detroit	106	van	181	negotiation	256	replace
32	safety	107	defect	182	move	257	isuzu
33	contract	108	big	183	group	258	take
34	general	109	driver	184	subaru	259	jaguar
35	incentive	110	ago	185	base	260	stempel
36	volkswagen	111	early	186	standard	261	manufacturer
37	buick	112	work	187	result	262	1
38	pontiac	113	drop	188	.....	263	pension
39	1987	114	get	189	decline	264	help
40	build	115	change	190	want	265	government
41	amc	116	.....	191	ambulance	266	pay
42	rebate	117	no.	192	current	267	eight
43	week	118	late	193	operation	268	adjust
44	.....	119	perot	194	even	269	currently
45	new	120	.....	195	talk	270	11-20
46	1986	121	cost	196	quality	271	10
47	engine	122	test	197	end	272	has_been
48	program	123	compare	198	january	273	fuel
49	.....	124	pact	199	1984	274	demand
50	.....	125	offer	200	assemble	275	white
51	company	126	expect	201	extend	276	see
52	.....	127	customer	202	united	277	x-there
53	inventory	128	start	203	ag	278	ohio
54	domestic	129	1988	204	accord	279	1-10
55	job	130	..	205	continue	280	gm-10
56	strike	131	plan	206	daily	281	consider
57	.....	132	iacocca	207	number	282	fx
58	oldsmobile	133	odometer	208	volume	283	reduce
59	renault	134	buyer	209	cause	284	highway
60	price	135	jeep	210	kenosha	285	million
61	japanese	136	low	211	monday	286	settlement
62	problem	137	shift	212	same	287	.....
63	analyst	138	announce	213	december	288	bag
64	agency	139	difference	214	expire	289	introduce
65	schedule	140	fall	215	employee	290	idle
66	increase	141	executive	216	annual	291	use
67	.....	142	.....	217	strong	292	bmw
68	.....	143	investigation	218	porsche	293	billion
69	mich.	144	chg	219	estimate	294	october
70	market	145	earlier	220	top	295	bonus
71	smith	146	fire	221	yesterday	296	tell
72	industry	147	add	222	go	297	think
73	part	148	....	223	rise	298	be_based_on
74	sell	149	financing	224	agreement	299	closing
75	rate	150	japan	225	facility	300	time

Table B.3: **AUT** Topic signature in set **PH**.

BBK Topic signature (PH)							
1	share	76	principal	151	called_for	226	sell
2	debenture	77	announce	152	12	227	eddie
3	redeem	78	shareholder	153	retire	228	arden
4	redemption	79	april	154	american	229	finance
5	outstanding	80	june	155	nov.	230	14
6	due	81	holding_company	156	henley	231	edt
7	common_shares	82	sinking_fund	157	security	232	investor
8	offer	83	14%	158	product	233	golden
9	as_many	84	over-the-counter	159	58%	234	accept
10	tender	85	500,000	160	c	235	international
11	buy_back	86	10	161	additional	236	equity
12	series	87	swap	162	remain	237	spokesman
13	accrue	88	sept.	163	capital	238	store
14	million	89	extend	164	mr.	239	2010
15	purchase	90	financial	165	stock_option	240	bancorp
16	convertible	91	conversion	166	carling	241	southdown
17	buy-back	92	director	167	general	242	a.m.
18	authorize	93	pay	168	1992	243	37.5
19	preferred	94	national	169	market_price	244	oak
20	note	95	chemical	170	1994	245	long-term
21	holder	96	oct.	171	payment	246	16
22	amount	97	9	172	all_of	247	1996
23	plus	98	par	173	hold	248	monday
24	company	99	friday	174	dallas	249	privately
25	stock	100	reduce	175	cost	250	trade
26	common	101	share_in	176	78%	251	earnings
27	plan	102	11	177	100,000	252	basis
28	bond	103	maker	178	$25	253	financing
29	interest	104	antar	179	investment	254	1993
30	common_stock	105	34%	180	expect	255	call
31	class	106	senior	181	right	256	system
32	$1,000	107	march	182	18%	257	believe
33	exchange	108	cash	183	home	258	month
34	buyback	109	gas	184	ohio	259	par_value
35	1	110	begin	185	american_stock_exchange	260	exchangeable
36	trading	111	31	186	13	261	private
37	price	112	say	187	10%	262	more_than
38	preferred_stock	113	bank	188	group	263	freeport-mcmoran
39	close	114	p.m.	189	undervalue	264	medical
40	expire	115	30	190	receive	265	houston
41	stock_exchange	116	make	191	1988	266	increase
42	cent	117	aug.	192	nugget	267	2
43	inc.	118	8	193	current	268	7
44	subordinate	119	energy	194	12:01	269	tektronix
45	corp.	120	nl	195	number	270	food
46	debt	121	july	196	preference	271	repurchased
47	composite	122	jan.	197	ltd.	272	los_angeles
48	dividend	123	5	198	quarter	273	in_private
49	face	124	previously	199	condition	274	comdata
50	yesterday	125	corporate	200	be_part_of	275	asset
51	new_york_stock_exchange	126	approve	201	50	276	may_1
52	preferred_shares	127	much_as	202	part	277	recently
53	board	128	25	203	own	278	6
54	market	129	oil	204	restructuring	279	2012
55	15	130	currently	205	employee	280	inc
56	from_time_to_time	131	total	206	separately	281	200,000
57	co.	132	power	207	july_1	282	1997
58	b	133	purpose	208	would_be	283	end
59	utility	134	12.5	209	suave	284	27
60	convert	135	feb.	210	bank_holding_company	285	2015
61	date	136	calif.	211	authorization	286	result
62	cumulative	137	$100	212	periodically	287	n.j.
63	buy	138	complete	213	recapitalization	288	baldwin
64	first_mortgage	139	value	214	rate	289	retailer
65	issue	140	new	215	negotiate	290	agent
66	tender_offer	141	service	216	chapman	291	chicago
67	program	142	fund	217	$50	292	1998
68	12%	143	odd-lot	218	new_york	293	savin
69	concern	144	est	219	toronto	294	13%
70	in_the_open	145	38%	220	buy-backs	295	$30
71	open	146	trust	221	industry	296	1987
72	transaction	147	warrant	222	year	297	chairman
73	repurchase	148	unpaid	223	edison	298	universal
74	dec.	149	use	224	unchanged	299	itt
75	unit	150	electric	225	sale	300	ramada

Table B.4: **BBK** Topic signature in set **PH**

BCY Topic signature (PH)							
1	bankruptcy	76	month	151	operation	226	plead
2	chapter	77	year	152	settle	227	confirmation
3	creditor	78	bowery	153	committee	228	problem
4	wedtech	79	towle	154	financial	229	maker
5	11	80	poppa	155	spencer	230	get
6	reorganization	81	bedford	156	general	231	damage
7	court	82	triad	157	statement	232	bench
8	minkow	83	president	158	audit	233	fraud
9	protection	84	newedge	159	khashoggi	234	equity
10	zzzz	85	agreement	160	steelmaker	235	1984
11	judge	86	toy	161	decline	236	dalkon
12	mr.	87	protect	162	concern	237	lisp
13	bankruptcy-law	88	reorganize	163	hearing	238	become
14	code	89	secure	164	manhattan	239	product
15	plan	90	failure	165	common_stock	240	august
16	galanis	91	contract	166	note	241	chief_executive_officer
17	file_for	92	consolidated	167	money	242	executive
18	company	93	lawyer	168	grant	243	previously
19	federal	94	tax	169	order	244	carpet
20	lawsuit	95	ruling	170	john	245	no-bid
21	debt	96	management	171	solar	246	village
22	robin	97	tiger	172	merhige	247	investigation
23	unsecured	98	law	173	agency	248	elsinore
24	filing	99	new_york	174	oil	249	lampert
25	claim	100	chairman	175	30	250	firm
26	chargit	101	los_angeles	176	davis	251	agree
27	proceeding	102	kmg	177	federal_court	252	shareholder
28	ir	103	royale	178	moseley	253	michigan
29	file	104	corp.	179	dollar	254	program
30	million	105	steel	180	number	255	say
31	pay	106	fail	181	office	256	industry
32	manville	107	wickes	182	try	257	end
33	creditors'	108	new	183	line	258	white_house
34	former	109	law_firm	184	brook	259	file_in
35	attorney	110	kaiser	185	independent	260	stereo
36	bank	111	make	186	complaint	261	japan
37	best	112	comment	187	japanese	262	small
38	storage	113	labor	188	lubensky	263	four
39	wow	114	seek	189	approve	264	call_for
40	ltv	115	corporate	190	investor	265	chief_executive
41	payment	116	charter	191	credit-card	266	formerly
42	loss	117	counsel	192	earlier	267	meese
43	catain	118	american	193	admit	268	flanigan
44	case	119	approval	194	success	269	merchant
45	operate	120	co.	195	revenue	270	people
46	smith	121	holder	196	accord	271	31
47	asset	122	major	197	certain	272	april
48	liability	123	auditor	198	estimate	273	mclean
49	continental	124	examiner	199	de	274	various
50	nashua	125	appeal	200	interest	275	california
51	continue	126	petition	201	sale	276	vice_president
52	charge	127	inc.	202	list	277	allow
53	owe	128	total	203	jr.	278	lender
54	business	129	employee	204	own	279	work_out
55	resign	130	partnership	205	connally	280	america
56	u.s.	131	texas	206	dylex	281	balboa
57	salant	132	worker	207	amount	282	mavroules
58	technology	133	trust	208	share	283	overmyer
59	settlement	134	net	209	behr	284	texscan
60	nofziger	135	fund	210	pension	285	conner
61	wheeling-pittsburgh	136	energy	211	expect	286	kingsborough
62	suit	137	city	212	yesterday	287	sba
63	cash	138	stock	213	credit	288	teikoku
64	receive	139	officer	214	has_been	289	lumber
65	dispute	140	1986	215	obtain	290	propose
66	trustee	141	group	216	job	291	large
67	emerge	142	airline	217	benefit	292	ask
68	heck	143	loan	218	texaco	293	financing
69	work	144	official	219	sec	294	day
70	dense-pac	145	partner	220	bankruptcy-court	295	pacific
71	name	146	worthen	221	week	296	guilty
72	reach	147	unit	222	cent	297	represent
73	1985	148	director	223	allege	298	face
74	lorean	149	railroad	224	executive_officer	299	march
75	report	150	more_than	225	clean	300	exceed

Table B.5: **BCY** Topic signature in set **PH**.

BNK Topic signature (PH)							
1	bank	76	customer	151	interest	226	try
2	mr.	77	sri	152	change	227	equity
3	thrift	78	banco	153	suit	228	expense
4	loan	79	agency	154	dallas	229	japanese
5	banking	80	source	155	job	230	mortgage
6	deposit	81	case	156	officer	231	document
7	fslic	82	bill	157	get	232	own
8	federal	83	foreign	158	gaubert	233	power
9	fed	84	washington	159	major	234	activity
10	board	85	collapse	160	even	235	man
11	institution	86	million	161	market	236	chemical
12	volcker	87	country	162	member	237	repay
13	fdic	88	business	163	want	238	end
14	banker	89	account	164	act	239	clausen
15	henkel	90	accord	165	yesterday	240	family
16	khoo	91	executive	166	need	241	use
17	billion	92	system	167	california	242	position
18	brunei	93	big	168	d.	243	small
19	regulator	94	troubled	169	financial_institution	244	believe
20	chairman	95	report	170	nation	245	singapore
21	citicorp	96	allow	171	move	246	know
22	greenspan	97	s&l	172	go	247	limited-service
23	asset	98	credit	173	management	248	great
24	vatican	99	association	174	bank_holding_company	249	record
25	financial	100	gebauer	175	check	250	william
26	national_bank	101	worthen	176	reagan	251	control
27	insurance	102	security	177	head	252	add
28	gray	103	marcos	178	director	253	sen.
29	texas	104	banks'	179	ogden	254	cooper
30	official	105	insolvent	180	indictment	255	several
31	home_loan_bank	106	has_been	181	citibank	256	cost
32	capital	107	congress	182	raise	257	sentence
33	money	108	name	183	london	258	loan-loss
34	mellon	109	large	184	week	259	comment
35	fund	110	would_be	185	expect	260	support
36	government	111	barnett	186	action	261	whether
37	committee	112	law	187	leave	262	recent
38	grambling	113	corp.	188	hongkong	263	rate
39	saving	114	proxmire	189	st	264	lend
40	fraud	115	issue	190	require	265	creditor
41	loss	116	pay	191	chief	266	seek
42	savings_and_loan	117	beebe	192	analyst	267	governor
43	year	118	industry	193	month	268	borrow
44	lending	119	atm	194	time	269	poehl
45	bankamerica	120	wright	195	total	270	transfer
46	reserve	121	investigation	196	investigator	271	lawyer
47	fail	122	dollar	197	economic	272	net
48	former	123	give	198	debt	273	meeting
49	charge	124	ka	199	operation	274	home
50	problem	125	popejoy	200	payment	275	standard
51	vernon	126	take	201	close	276	high
52	u.s.	127	1985	202	league	277	vice_president
53	trust	128	amalgamated	203	morgan	278	serve
54	senate	129	new_york	204	regulation	279	find
55	office	130	provision	205	top	280	federally
56	president	131	limit	206	hong	281	criminal
57	s&ls	132	monetary	207	finance	282	baker
58	lincoln	133	decision	208	company	283	propose
59	continental	134	wall	209	become	284	agreement
60	ambrosiano	135	wah	210	ask	285	claim
61	policy	136	1984	211	calvi	286	regional
62	branch	137	real_estate	212	federal_reserve_board	287	commercial_bank
63	depositor	138	butcher	213	proposal	288	profit
64	investment	139	germain	214	day	289	corporate
65	insure	140	international	215	where	290	consider
66	rep.	141	legislation	216	staff	291	unit
67	house	142	american	217	kong	292	part
68	new	143	plan	218	likely	293	city
69	make	144	more_than	219	1986	294	appoint
70	tan	145	swearingen	220	chartered	295	chase
71	attorney	146	hold	221	judge	296	requirement
72	central_bank	147	authority	222	fee	297	agree
73	court	148	borrowing	223	interest_rate	298	begin
74	state	149	group	224	help	299	private
75	rule	150	holding_company	225	keep	300	statement

Table B.6: **BNK** Topic signature in set **PH**.

BON Topic signature (PH)							
1	bond	76	18	151	guarantee	226	reduce
2	issue	77	matthew	152	payment	227	change
3	rating	78	official	153	treaty	228	banker
4	market	79	yen	154	5	229	mae
5	debt	80	1992	155	strong	230	thin
6	due	81	antilles	156	municipal_bond	231	slightly
7	moody	82	euroyen	157	trade	232	15-year
8	debenture	83	european	158	australian_dollar	233	canadian_dollar
9	note	84	west_german	159	7	234	denominate
10	eurodollar	85	tax	160	british	235	raise
11	price	86	101	161	38	236	late
12	point	87	secondary	162	pay	237	acquisition
13	offering	88	maturity	163	large	238	conversion
14	yield	89	percentage_point	164	equity-purchase	239	holiday
15	investor	90	seasoned	165	quote	240	$75
16	bank	91	fixed-coupon	166	corporate	241	2
17	eurobonds	92	gilt	167	sachs	242	authority
18	treasury	93	standard	168	friday	243	pricing
19	par	94	week	169	equity-linked	244	buy
20	u.s.	95	financial	170	continue	245	say
21	trader	96	inc.	171	plan	246	make
22	s&p	97	ltd	172	financing	247	switzerland
23	underwriter	98	low	173	new_york	248	early
24	convertible	99	company	174	creditwatch	249	underwriting
25	london	100	foreign	175	boston	250	three-year
26	proceeds	101	use	176	$150	251	fall
27	subordinate	102	used_to	177	unchanged	252	fund
28	eurobond	103	cite	178	dollar-denominated	253	obligation
29	rate	104	general_purpose	179	stanley	254	shelf
30	launch	105	much_as	180	recent	255	follow
31	dealer	106	fixed-rate	181	18%	256	bondholder
32	dollar	107	morgan	182	amount	257	gain
33	coupon	108	specialist	183	auction	258	deal
34	floating-rate	109	sell	184	short-term	259	manage
35	japanese	110	10-year	185	78%	260	loss
36	borrower	111	15	186	activity	261	purpose
37	lead	112	lower	187	unit	262	electric
38	million	113	maturities	188	major	263	accord
39	security	114	equity	189	2012	264	of_this
40	government	115	wright	190	indicate	265	6
41	five-year	116	day	191	affect	266	interbank
42	billion	117	paper	192	registration	267	group
43	offer	118	$200	193	little	268	june
44	14	119	merrill	194	bid	269	gold
45	12	120	salomon	195	investment	270	2002
46	1	121	service	196	revenue_bond	271	possible
47	credit	122	brother	197	seven-year	272	today
48	senior	123	lynch	198	spread	273	liquidity
49	manager	124	state	199	utility	274	risk
50	warrant	125	nomura	200	quiet	275	4
51	34%	126	yesterday	201	meanwhile	276	house
52	swiss_franc	127	close	202	agency	277	mortgage
53	syndication	128	review	203	premium	278	carry
54	concern	129	participant	204	common	279	dec.
55	downgrade	130	via	205	9	280	burnham
56	12%	131	8	206	general	281	find
57	international	132	tax-exempt	207	38%	282	the_swiss
58	ltd.	133	canadian	208	rise	283	appear
59	interest	134	europe	209	swiss	284	$300
60	basis	135	triple-a	210	remain	285	sale
61	capital	136	long-term	211	$50	286	10
62	interest_rate	137	syndicate	212	1994	287	level
63	1997	138	mature	213	58%	288	swap
64	$100	139	place	214	although	289	underwrite
65	14%	140	total	215	issuers	290	would_be
66	demand	141	sector	216	above	291	inc
67	co.	142	suisse	217	share	292	term
68	trading	143	firm	218	sallie	293	march
69	new	144	34	219	drexel	294	negative
70	corp.	145	tokyo	220	canada	295	gmac
71	currency	146	end	221	year	296	redeem
72	mark	147	session	222	100	297	help
73	high	148	goldman	223	loan	298	position
74	poor	149	decline	224	municipal	299	volume
75	finance	150	expect	225	upgrade	300	lambert

Table B.7: **BON** Topic signature in set **PH**.

CEO Topic signature (PH)							
1	mr.	76	double	151	$2	226	motel
2	cent	77	projection	152	area	227	begin
3	quarter	78	cost	153	extraordinary	228	taylor
4	earnings	79	make	154	ohio	229	write-down
5	expect	80	grow	155	electronics	230	hecla
6	fiscal	81	charge	156	gas	231	mattress
7	million	82	total	157	jump	232	better
8	share	83	indicate	158	analyst	233	$4
9	sale	84	billion	159	huffy	234	$1.6
10	year	85	maker	160	currently	235	despite
11	revenue	86	first-quarter	161	first_half	236	product_line
12	profit	87	restaurant	162	service	237	around
13	net_income	88	plan	163	march	238	20
14	net	89	10%	164	line	239	automotive
15	1987	90	attribute	165	tax_rate	240	full
16	interview	91	late	166	concern	241	primarily
17	1986	92	say	167	restructuring	242	base
18	year-earlier	93	$1	168	40%	243	center
19	company	94	gold	169	10	244	change
20	rise	95	unit	170	second_half	245	full-year
21	report	96	computer	171	major	246	$2.5
22	estimate	97	average	172	adjust	247	share_in
23	chief_executive_officer	98	restate	173	boost	248	industrial
24	analysts'	99	inc.	174	pay	249	18
25	executive_officer	100	project	175	fall	250	largely
26	increase	101	ounce	176	four	251	consumer
27	earn	102	industry	177	executive	252	florida
28	compare	103	expand	178	benefit	253	small
29	loss	104	improve	179	carry-forward	254	$30
30	product	105	backlog	180	sept.	255	$6
31	fourth	106	recent	181	machine	256	hurt
32	end	107	current	182	30%	257	$1.2
33	fourth-quarter	108	dividend	183	introduce	258	m.
34	earlier	109	pre-tax	184	account_for	259	thrift
35	acquisition	110	at_least	185	international	260	operating
36	result	111	improvement	186	good	261	reduce
37	business	112	profit_from	187	several	262	60
38	gain	113	six	188	large	263	level
39	per-share	114	ago	189	go	264	previous
40	operation	115	exceed	190	a.	265	50
41	1988	116	predict	191	june	266	performance
42	growth	117	25%	192	conger	267	roughly
43	period	118	figure	193	financial_officer	268	expense
44	1985	119	expansion	194	see	269	manufacturing
45	year-ago	120	15%	195	has_been	270	in_addition
46	chairman	121	slightly	196	trading	271	u.s.
47	post	122	50%	197	public	272	7%
48	president	123	corp.	198	triple	273	item
49	31	124	equipment	199	tax-loss	274	profitable
50	fiscal_year	125	third-quarter	200	dilute	275	development
51	continue	126	outstanding	201	tax	276	3-for-2
52	increase_in	127	loan	202	flat	277	capacity
53	strong	128	system	203	number	278	25
54	store	129	demand	204	micropro	279	spending
55	more_than	130	dollar	205	volume	280	one-time
56	high	131	stock_split	206	software	281	segment
57	add	132	comfortable	207	plastic	282	$10
58	first_quarter	133	margin	208	co.	283	5%
59	second-quarter	134	capital	209	sell	284	july
60	market	135	discontinue	210	acquire	285	worthington
61	second	136	order	211	annual	286	cite
62	be_about	137	part	212	use	287	40
63	range	138	record	213	$3	288	12
64	operate	139	low	214	28	289	april
65	new	140	specific	215	asset	290	70
66	20%	141	debt	216	anticipate	291	open
67	price	142	plant	217	tax_credit	292	chief_operating_officer
68	30	143	contract	218	$1.1	293	expectation
69	forecast	144	program	219	close	294	supply
70	reflect	145	five	220	$1.3	295	homestake
71	decline	146	all_of	221	$1.5	296	vice_president
72	third	147	$5	222	home	297	a_little
73	year_end	148	division	223	help	298	battle
74	month	149	customer	224	quarterly	299	food
75	nine	150	production	225	profit_margin	300	look

Table B.8: **CEO** Topic signature in set **PH**.

CMD Topic signature (PH)							
1	farmer	76	university	151	problem	226	1986
2	crop	77	new	152	texas	227	beijing
3	farm	78	demand	153	feed	228	time
4	grower	79	buffer	154	metric_ton	229	think
5	grain	80	people	155	winter	230	farmworkers
6	price	81	low	156	order	231	western
7	wheat	82	weaver	157	osha	232	likely
8	corn	83	consumer	158	area	233	research
9	cocoa	84	field	159	billion	234	population
10	immigration	85	hightower	160	africanized	235	output
11	pound	86	china	161	allow	236	reduce
12	agriculture	87	toilet	162	hire	237	major
13	soybean	88	brazil	163	packer	238	economic
14	harvest	89	herd	164	natural	239	know
15	employer	90	federal	165	1985	240	come
16	mr.	91	slaughter	166	average	241	house
17	palm_oil	92	undocumented	167	brock	242	head
18	cotton	93	rule	168	congress	243	local
19	coffee	94	hen	169	ico	244	citrus
20	agriculture_department	95	land-grant	170	big	245	livestock
21	year	96	vanillin	171	policy	246	dealer
22	agricultural	97	begin	172	june	247	continue
23	program	98	world	173	labor	248	almost
24	hog	99	rise	174	case	249	irradiation
25	u.s.	100	philippine	175	help	250	illegals
26	law	101	almond	176	el	251	past
27	worker	102	shortage	177	box	252	way
28	state	103	official	178	standard	253	mexican
29	tomato	104	fruit	179	find	254	political
30	land	105	orange	180	orchard	255	pass
31	bee	106	1	181	health	256	cut
32	cattle	107	mushroom	182	trader	257	change
33	commodity	108	fat	183	delegate	258	million
34	sugar	109	day	184	around	259	special
35	rice	110	consumption	185	weather	260	daniel
36	subsidy	111	dust	186	decline	261	naturalization
37	producer	112	export	187	agency	262	see
38	analyst	113	reform	188	apply	263	sanction
39	bale	114	expect	189	beef	264	rep.
40	vanilla	115	more_than	190	wine	265	try
41	ton	116	support	191	rural	266	measure
42	farm_worker	117	economist	192	her	267	senate
43	farmland	118	report	193	take	268	sell
44	government	119	dairy	194	budget	269	barlowe
45	market	120	honeybee	195	cholesterol	270	rinderer
46	egg	121	rain	196	industry	271	restaurant
47	pork	122	even	197	has_been	272	flood
48	department	123	high	198	make	273	grapefruit
49	gene	124	she	199	crate	274	little
50	immigrant	125	go	200	committee	275	current
51	plambeck	126	want	201	chinese	276	keep
52	acre	127	washington	202	show	277	plan
53	future	128	level	203	the_most	278	iowa
54	amnesty	129	regulation	204	legal	279	accord
55	illegal	130	malaysia	205	ago	280	wage
56	milk	131	payment	206	drop	281	require
57	drought	132	get	207	large	282	increase
58	produce	133	buy	208	pay	283	can't
59	kansas	134	india	209	group	284	provision
60	plant	135	albright	210	marketing	285	run
61	production	136	week	211	palm-oil	286	document
62	supply	137	man	212	spring	287	index
63	bill	138	tropical	213	use	288	lose
64	food	139	would_be	214	manager	289	cartel
65	estimate	140	month	215	deputy	290	attempt
66	quota	141	california	216	total	291	cost
67	country	142	issue	217	family	292	early
68	archbishop	143	girard	218	cent	293	forecast
69	acreage	144	land_reform	219	give	294	sign
70	nation	145	work	220	new_york	295	feed_grain
71	animal	146	oil	221	facility	296	legalization
72	season	147	organization	222	idle	297	peasant
73	surplus	148	mexico	223	freddy	298	set
74	alien	149	fall	224	qureshi	299	labor_department
75	bushel	150	church	225	1988	300	meat

Table B.9: **CMD** Topic signature in set **PH**.

DIV Topic signature (PH)							
1	dividend	76	new_york_stock_exchange	151	14	226	transco
2	payable	77	corp.	152	23	227	industry
3	stock_of_record	78	new	153	24	228	doyle
4	quarterly	79	real_estate	154	end	229	cut
5	declare	80	authorized_shares	155	revenue	230	friday
6	cent	81	approval	156	initial	231	first-quarter
7	share	82	loss	157	fiscal	232	strong
8	stock_split	83	composite	158	hancock	233	over-the-counter
9	split	84	yesterday	159	restructuring	234	has_been
10	2-for-1	85	preferred	160	net	235	ge
11	holder	86	gas	161	investment	236	president
12	shareholder	87	concern	162	operation	237	pacific
13	distribution	88	close	163	21	238	hold
14	pay	89	boost	164	chief_executive_officer	239	earn
15	outstanding	90	16	165	executive_officer	240	techops
16	company	91	payment	166	26	241	allegheny
17	class	92	reflect	167	suspend	242	store
18	record	93	chris-craft	168	reverse_split	243	conn.
19	common	94	product	169	receive	244	own
20	stock_dividend	95	27	170	current	245	improve
21	approve	96	nov.	171	net_income	246	four
22	15	97	1986	172	number	247	33
23	3-for-2	98	12	173	group	248	1988
24	10	99	year	174	nine	249	calif.-based
25	april	100	increase_in	175	rise	250	25%
26	june	101	currently	176	financial	251	certain
27	pre-split	102	mesa	177	preferred_shares	252	capital
28	increase	103	6	178	reverse_stock_split	253	analyst
29	cash	104	expect	179	spokesman	254	$1
30	common_stock	105	18	180	cite	255	40
31	jan.	106	raise	181	3	256	purchase
32	million	107	chairman	182	centex	257	interest
33	right	108	rate	183	loan	258	copperweld
34	march	109	preferred_stock	184	n.j.	259	operate
35	board	110	2	185	earlier	260	additional
36	post-split	111	co.	186	price	261	total
37	feb.	112	realty	187	share_in	262	set
38	dec.	113	bank	188	profit	263	fractional
39	meeting	114	mca	189	intend	264	form
40	distribute	115	omit	190	takeover	265	carter
41	b	116	13	191	result	266	ltd.
42	30	117	property	192	7	267	mail
43	earnings	118	1987	193	stockholder	268	staar
44	annual	119	utility	194	12	269	topps
45	stock	120	may_1	195	announce	270	henley
46	common_shares	121	report	196	bank_holding_company	271	wasserman
47	31	122	would_be	197	continue	272	unchanged
48	payout	123	calif.	198	say	273	texas
49	javelin	124	5	199	propose	274	natural_gas
50	aug.	125	five	200	make	275	kansas_city
51	sept.	126	warner	201	new_york	276	income
52	director	127	effective	202	spin_off	277	outlook
53	july	128	4	203	first_quarter	278	retailer
54	date	129	sale	204	month	279	home
55	1	130	9	205	base	280	security
56	quarter	131	oil	206	todd	281	exercisable
57	plan	132	entitle	207	19	282	electric
58	inc.	133	17	208	doubling	283	billion
59	previously	134	subject_to	209	10%	284	march_2
60	regular	135	american	210	southmark	285	second
61	partnership	136	mr.	211	champion	286	voting
62	special	137	11	212	service	287	broadway
63	trading	138	spinoff	213	3-for-1	288	control
64	20	139	series	214	hawley	289	six
65	vote	140	50	215	resume	290	management
66	unit	141	28	216	proposal	291	adopt
67	authorize	142	lucky	217	acquire	292	care
68	basis	143	limited	218	issue	293	wsj
69	stock_exchange	144	20%	219	rochester	294	require
70	oct.	145	8	220	reduce	295	80
71	25	146	action	221	equipment	296	asset
72	maker	147	investment_trust	222	bayly	297	westcoast
73	29	148	trust	223	investor	298	hostile_takeover
74	22	149	allis	224	write-off	299	canadian
75	holding_company	150	business	225	move	300	holding

Table B.10: **DIV** Topic signature in set **PH**.

ECO Topic signature (PH)							
1	budget	76	automatic	151	pass	226	social
2	tax	77	saving	152	seem	227	enough
3	spending	78	veto	153	unemployment	228	drop
4	deficit	79	index	154	oct.	229	poor
5	congress	80	people	155	begin	230	little
6	billion	81	member	156	month	231	accord
7	gramm-rudman	82	outlay	157	chairman	232	small
8	economic	83	reduce	158	cost	233	fed
9	democrat	84	project	159	keep	234	national
10	reagan	85	miller	160	come	235	labor
11	president	86	baker	161	must	236	create
12	economy	87	reagan_administration	162	james	237	decade
13	tax_increase	88	survey	163	same	238	market
14	house	89	u.s.	164	private	239	enact
15	cut	90	dollar	165	take	240	tell
16	republican	91	crash	166	recent	241	consumption
17	senate	92	washington	167	debt-ceiling	242	oppose
18	economist	93	leader	168	top	243	debt
19	white_house	94	mean	169	figure	244	interest_rate
20	fiscal	95	need	170	capital-gains	245	great
21	recession	96	nonmilitary	171	tax_rate	246	result
22	law	97	show	172	1987	247	period
23	mr.	98	make	173	speaker	248	domestic
24	congressional	99	nation	174	meet	249	debt_ceiling
25	1988	100	low	175	borrowing	250	rather
26	growth	101	grow	176	today	251	office
27	administration	102	week	177	decline	252	gop
28	program	103	political	178	appropriation	253	gray
29	federal	104	minimum_wage	179	limit	254	major
30	income	105	problem	180	family	255	study
31	inflation	106	resolution	181	legislation	256	income_tax
32	committee	107	rostenkowski	182	total	257	call
33	government	108	likely	183	american	258	extension
34	military	109	estimate	184	spend	259	big
35	sen.	110	official	185	sector	260	senator
36	year	111	real	186	indicator	261	corporate
37	rate	112	average	187	conservative	262	presidential
38	raise	113	go	188	panel	263	perhaps
39	state	114	worker	189	university	264	summit
40	increase	115	public	190	money	265	us
41	rep.	116	see	191	early	266	aid
42	deficit-reduction	117	large	192	require	267	put
43	democratic	118	expansion	193	try	268	remain
44	job	119	process	194	whether	269	end
45	budget_deficit	120	number	195	population	270	1986
46	bill	121	want	196	thing	271	talk
47	gnp	122	chile	197	good	272	price
48	policy	123	business	198	issue	273	half
49	new	124	think	199	service	274	far
50	rise	125	education	200	reform	275	of_this
51	target	126	wage	201	time	276	1981
52	reduction	127	cut_in	202	help	277	household
53	package	128	propose	203	find	278	tax_bill
54	d.	129	current	204	future	279	value
55	lawmaker	130	support	205	annual	280	new_hampshire
56	measure	131	forecast	206	nearly	281	benefit
57	high	132	pay	207	effect	282	negotiator
58	change	133	get	208	1	283	investment
59	would_be	134	gain	209	add	284	goal
60	black	135	bentsen	210	statistic	285	burden
61	compromise	136	$108	211	conference	286	indeed
62	social_security	137	increase_in	212	taxpayer	287	consider
63	plan	138	more_than	213	fall	288	reach
64	consumer	139	poverty	214	across-the-board	289	gross_national_product
65	spending_cut	140	welfare	215	debate	290	cause
66	way	141	suggest	216	start	291	supply-side
67	vote	142	work	217	past	292	texas
68	defense	143	levy	218	give	293	day
69	even	144	has_been	219	such_as	294	forecaster
70	level	145	fiscal_year	220	asset	295	again
71	proposal	146	amendment	221	believe	296	argue
72	trillion	147	capital	222	force	297	student
73	wright	148	expect	223	idea	298	individual
74	revenue	149	stock_market	224	question	299	long
75	employment	150	continue	225	country	300	fear

Table B.11: **ECO** Topic signature in set **PH**.

EDP Topic signature (PH)							
1	ibm	76	graphics	151	compete	226	a_lot
2	computer	77	run	152	computing	227	let
3	machine	78	marketing	153	hardware	228	billion
4	software	79	manager	154	she	229	vax
5	personal_computer	80	tandy	155	sculley	230	important
6	apple	81	ibm-compatible	156	quarter	231	desktop
7	digital	82	feature	157	buy	232	cambridge
8	lotus	83	even	158	group	233	rollwagen
9	cray	84	expect	159	cut	234	million
10	microsoft	85	compatible	160	change	235	know
11	pc	86	ncr	161	need	236	able
12	mr.	87	standard	162	performance	237	for_example
13	new	88	develop	163	corporate	238	better
14	product	89	80386	164	gate	239	computer_system
15	model	90	eta	165	processor	240	unit
16	compaq	91	akers	166	president	241	to_do
17	macintosh	92	unisys	167	new_line	242	borland
18	user	93	get	168	share	243	task
19	mainframe	94	take	169	old	244	store
20	workstation	95	available	170	amiga	245	80286
21	program	96	hewlett-packard	171	expert	246	earnings
22	datum	97	rattigan	172	thing	247	around
23	wang	98	faster	173	though	248	run_on
24	customer	99	see	174	u.s.	249	allow
25	operating_system	100	problem	175	network	250	john
26	system	101	move	176	portable	251	engineering
27	use	102	base	177	give	252	try
28	market	103	suit	178	9370	253	sales_force
29	clone	104	announcement	179	product_line	254	cassoni
30	chip	105	think	180	part	255	effort
31	analyst	106	copyright	181	same	256	manufacturer
32	chen	107	job	182	grow	257	order
33	technology	108	disk_drive	183	yesterday	258	team
34	company	109	development	184	strategy	259	prime
35	business	110	small	185	control	260	exist
36	introduce	111	speed	186	midrange	261	firm
37	commodore	112	competitor	187	keep	262	the_most
38	1-2-3	113	executive	188	has_been	263	second
39	design	114	fast	189	estimate	264	hard_disk
40	price	115	equipment	190	xt	265	pont
41	spreadsheet	116	unveil	191	rival	266	recently
42	supercomputer	117	mass.	192	mainframe_computer	267	two_year
43	powerful	118	supercomputers	193	different	268	wordperfect
44	apollo	119	ps2	194	provide	269	word-processing
45	sun	120	inc.	195	easy	270	europe
46	year	121	help	196	growth	271	in_addition
47	work	122	computer_industry	197	find	272	continue
48	version	123	application	198	end	273	set
49	intel	124	screen	199	own	274	student
50	make	125	honeywell	200	plan	275	1985
51	corp.	126	large	201	floppy_disk	276	corp
52	big	127	consultant	202	world	277	canion
53	fujitsu	128	hypercard	203	meeting	278	report
54	disk	129	write	204	start	279	drive
55	maker	130	more_than	205	monitor	280	yet
56	memory	131	ii	206	leave	281	window
57	minicomputer	132	dealer	207	power	282	name
58	international	133	way	208	handle	283	director
59	industry	134	become	209	project	284	gould
60	os2	135	ship	210	ago	285	fall
61	sale	136	copy	211	technical	286	build
62	ashton-tate	137	want	212	micro	287	stock
63	at&t	138	office	213	386	288	today
64	sell	139	time	214	chairman	289	that_is
65	olivetti	140	early	215	computer_business	290	du
66	vice_president	141	add	216	programmer	291	begin
67	people	142	calif.	217	several	292	late
68	microprocessor	143	go	218	emulex	293	hacker
69	cost	144	device	219	megabyte	294	kilobyte
70	information	145	such_as	220	show	295	terminal
71	call	146	worker	221	printer	296	database
72	research	147	announce	222	storage	297	low
73	employee	148	engineer	223	say	298	hitachi
74	personal-computer	149	month	224	major	299	dataquest
75	line	150	revenue	225	charge	300	expensive

Table B.12: **EDP** Topic signature in set **PH**.

ELE Topic signature (PH)							
1	chip	76	consumer	151	call	226	electronics_industry
2	semiconductor	77	high	152	dolby	227	plan
3	superconductors	78	system	153	shipment	228	calif.
4	superconductivity	79	magnetic_field	154	month	229	know
5	intel	80	texas	155	clara	230	introduce
6	material	81	miti	156	sample	231	hemlock
7	japanese	82	electrical	157	major	232	miscarriage
8	scientist	83	manufacturing	158	plant	233	believe
9	technology	84	cd-vs	159	transistor	234	defense
10	superconductor	85	group	160	official	235	copyright
11	superconducting	86	resistance	161	westinghouse	236	radio
12	temperature	87	manufacturer	162	achieve	237	physic
13	dat	88	book-to-bill	163	report	238	enough
14	patent	89	corp.	164	manufacture	239	komag
15	electronics	90	design	165	cost	240	polysilicon
16	u.s.	91	player	166	go	241	engineer
17	machine	92	den	167	rca	242	lead
18	researcher	93	business	168	technique	243	test
19	industry	94	project	169	same	244	customer
20	ceramic	95	record	170	take	245	sell
21	japan	96	production	171	international	246	commercial
22	research	97	find	172	development	247	average
23	mr.	98	memory_chip	173	philipsdu	248	theory
24	robot	99	sale	174	help	249	voice
25	application	100	order	175	yen	250	tiny
26	new	101	analyst	176	power	251	decoder
27	philips	102	fujitsu	177	cool	252	has_been
28	motorola	103	personal_computer	178	several	253	home
29	magnet	104	magnetic	179	thin	254	low_temperature
30	computer	105	nec	180	american	255	yet
31	disk	106	even	181	above	256	inc.
32	sematech	107	refrigerator	182	helium	257	whether
33	make	108	cbs	183	compound	258	accord
34	use	109	argonne	184	danforth	259	build
35	micro	110	show	185	answer	260	bill
36	wire	111	current	186	wafer	261	phenomenon
37	digital	112	kelvin	187	president	262	carry
38	advanced	113	appliance	188	zero	263	though
39	ge	114	venture	189	expect	264	play
40	toshiba	115	schluter	190	add	265	effect
41	ibm	116	cd-v	191	levitate	266	world-wide
42	sound	117	get	192	recent	267	metal
43	instrument	118	national	193	energy	268	ago
44	device	119	big	194	need	269	period
45	company	120	produce	195	whirlpool	270	as_well
46	product	121	world	196	license	271	become
47	maker	122	liquid_nitrogen	197	anti-taping	272	vice_president
48	sony	123	large	198	film	273	growth
49	market	124	study	199	dram	274	increase
50	audio	125	problem	200	marous	275	try
51	microprocessor	126	people	201	similar	276	indicate
52	nishizawa	127	begin	202	think	277	right
53	lab	128	scientific	203	santa	278	price
54	electron	129	advance	204	science	279	fast
55	physicist	130	degree_fahrenheit	205	chip_in	280	race
56	electricity	131	discovery	206	require	281	rather
57	trade	132	matsushita	207	strong	282	congress
58	ratio	133	see	208	agreement	283	something
59	year	134	judge	209	mettur	284	tell
60	develop	135	bell	210	woman	285	miller
61	compact_disk	136	hitachi	211	part	286	come
62	recording	137	government	212	producer	287	work_at
63	university	138	equipment	213	way	288	version
64	video	139	small	214	integrated_circuit	289	foreign
65	silicon	140	near	215	such_as	290	move
66	recorder	141	absolute_zero	216	suit	291	concern
67	room_temperature	142	ingram	217	meissner	292	must
68	electric	143	crystal	218	mitsubishi	293	uart
69	laboratory	144	electronic	219	calif.-based	294	tape_recorder
70	consortium	145	idea	220	phone	295	standard
71	tape	146	80286	221	welch	296	provide
72	association	147	billion	222	form	297	powerful
73	degree	148	worker	223	screen	298	more_than
74	work	149	pont	224	tamura	299	mattel
75	music	150	process	225	for_example	300	below

Table B.13: **ELE** Topic signature in set **PH**.

ENV Topic signature (PH)							
1	epa	76	drinking_water	151	canadian	226	water_system
2	waste	77	sierra	152	spend	227	industrial
3	environmental	78	construction	153	small	228	justice
4	water	79	acid-rain	154	override	229	hunting
5	ozone	80	conservation	155	meet	230	mulroney
6	state	81	washington	156	action	231	ruling
7	incinerator	82	dump	157	effort	232	permit
8	agency	83	salmon	158	issue	233	within
9	clean	84	require	159	system	234	grand_jury
10	landfill	85	rule	160	burn	235	resource
11	hazardous	86	contamination	161	bacteria	236	more_than
12	federal	87	fine	162	even	237	municipal
13	acid_rain	88	county	163	thomas	238	solution
14	garbage	89	new	164	quality	239	bring
15	pollution	90	canal	165	compliance	240	rhine
16	lake	91	measure	166	estimate	241	research
17	law	92	fund	167	funding	242	dispose_of
18	standard	93	florida	168	groundwater	243	material
19	city	94	billion	169	vote	244	reiner
20	site	95	damage	170	industry	245	resident
21	mr.	96	investigation	171	report	246	technology
22	protection	97	wolf	172	federal_government	247	see
23	toxic	98	nation	173	scientist	248	want
24	violation	99	prosecutor	174	limit	249	order
25	reagan	100	veto	175	hodel	250	dirt
26	congress	101	citizen	176	need	251	expect
27	health	102	file	177	whether	252	work
28	emission	103	lead	178	foundation	253	attorney
29	bill	104	plan	179	gas	254	get
30	problem	105	fish	180	lawsuit	255	day
31	air	106	canada	181	yellowstone	256	exposure
32	area	107	new_jersey	182	air-pollution	257	air-quality
33	project	108	asbestos	183	exist	258	national
34	mcmillan	109	lime	184	superfund	259	service
35	act	110	facility	185	private	260	texas
36	plant	111	administrator	186	would_be	261	defendant
37	disposal	112	settlement	187	hazard	262	she
38	public	113	sanction	188	where	263	human
39	river	114	ban	189	department	264	worker
40	management	115	mountain_state	190	spokesman	265	defense
41	program	116	local	191	agree	266	sewage-treatment
42	hauler	117	violate	192	source	267	give
43	suit	118	deadline	193	district_attorney	268	civil
44	chemical	119	outbreak	194	time	269	safe
45	browning-ferris	120	smog	195	call	270	change
46	cleanup	121	risk	196	talc	271	employer
47	dam	122	senate	197	natural_resources	272	pose
48	case	123	people	198	put	273	acid
49	environmentalist	124	contaminate	199	legislation	274	reagan_administration
50	strobel	125	antitrust	200	committee	275	old
51	trash	126	club	201	soil	276	cancer
52	air_pollution	127	use	202	recycling	277	high_court
53	regulation	128	los_angeles	203	provide	278	live
54	u.s.	129	house	204	acidic	279	support
55	study	130	sanitation	205	impose	280	hazardous-waste
56	pcbs	131	sewer	206	part	281	scenic
57	control	132	take	207	claim	282	ton
58	pollutant	133	president	208	office	283	curb
59	charge	134	enforcement	209	science	284	five
60	rollins	135	discharge	210	earth	285	policy
61	propose	136	proposal	211	montana	286	administration
62	government	137	pit	212	test	287	serious
63	cause	138	group	213	kill	288	particle
64	ash	139	protect	214	animal	289	woman
65	forest	140	land	215	reduce	290	opposition
66	kissimmee	141	water_cooler	216	allegation	291	oxygen
67	cost	142	taiwan	217	business	292	substance
68	environment	143	find	218	must	293	occupational
69	level	144	criminal	219	make	294	political
70	indian	145	allege	220	country	295	approve
71	official	146	pollution-control	221	metal	296	congressional
72	year	147	penalty	222	know	297	arco
73	wildlife	148	california	223	authority	298	lawyer
74	clean_up	149	effect	224	grant	299	fiscal
75	court	150	company	225	buffton	300	way

Table B.14: **ENV** Topic signature in set **PH**.

ERN Topic signature (PH)							
1	loss	76	stock_exchange	151	quarterly	226	year_end
2	quarter	77	composite	152	drop_in	227	norfolk
3	million	78	earn	153	performance	228	krone
4	cent	79	operating	154	oil	229	segment
5	share	80	chairman	155	president	230	holding_company
6	earnings	81	improve	156	result_in	231	volume
7	rise	82	debt	157	ago	232	outstanding
8	net	83	computer	158	equipment	233	level
9	sale	84	pre-tax	159	cannon	234	dividend
10	profit	85	chief_executive_officer	160	13%	235	calif.
11	net_income	86	inc.	161	full-year	236	southern
12	revenue	87	executive_officer	162	spokesman	237	consumer
13	year-earlier	88	figure	163	change	238	gca
14	earlier	89	corp.	164	say	239	gould
15	year	90	dollar	165	inventory	240	marketing
16	billion	91	yesterday	166	u.s.	241	surge
17	compare	92	fiscal_year	167	microsoft	242	tax-loss
18	operation	93	six	168	full	243	big
19	report	94	jump	169	offset	244	primarily
20	result	95	double	170	nonperforming	245	profitable
21	gain	96	provision	171	tax	246	mattel
22	fourth-quarter	97	growth	172	16%	247	varity
23	fiscal	98	related_to	173	real_estate	248	part
24	company	99	investment	174	plan	249	texas
25	period	100	low	175	march	250	$1
26	post	101	cite	176	greyhound	251	1988
27	1986	102	attribute	177	10%	252	payment
28	loan	103	one-time	178	fall_to	253	$2.2
29	charge	104	tiger	179	acquisition	254	slightly
30	reflect	105	first_half	180	margin	255	national
31	discontinue	106	unit	181	certain	256	oct.
32	end	107	more_than	182	take	257	note
33	fall	108	bank	183	kronor	258	average
34	first-quarter	109	insurance	184	line	259	production
35	increase	110	tax_credit	185	has_been	260	early
36	fourth	111	store	186	convergent	261	plc
37	third-quarter	112	maker	187	associate_with	262	datum
38	late	113	new	188	property	263	24%
39	continue	114	extraordinary	189	commodore	264	show
40	second-quarter	115	hurt	190	service	265	domestic
41	write-down	116	previously	191	20%	266	personal_computer
42	first_quarter	117	market	192	credit	267	canadian
43	31	118	interest	193	loan-loss	268	substantial
44	decline	119	toy	194	stock_split	269	benefit
45	third	120	group	195	retailer	270	utility
46	month	121	improvement	196	new_york	271	distribution
47	business	122	accounting	197	tax_rate	272	remain
48	profit_from	123	yen	198	nine-month	273	25
49	nine	124	drop	199	15%	274	good
50	product	125	write-off	200	large	275	predict
51	1987	126	price	201	11%	276	plant
52	analyst	127	concern	202	equity	277	cpc
53	restate	128	division	203	additional	278	flat
54	year-ago	129	record	204	capital	279	effect
55	restructuring	130	add	205	19%	280	$3
56	expect	131	after-tax	206	make	281	tandy
57	trading	132	reduce	207	dec.	282	federal
58	30	133	international	208	$2	283	unchanged
59	operate	134	june	209	widen	284	exclude
60	high	135	general	210	14%	285	profit_margin
61	asset	136	sell	211	second_half	286	vice_president
62	1985	137	climb	212	major	287	software
63	strong	138	12%	213	$1.2	288	non-accrual
64	increase_in	139	recent	214	mcorp	289	17%
65	cost	140	grow	215	carry-forward	290	demand
66	mr.	141	security	216	largely	291	northrop
67	pretax	142	stock	217	special	292	shipment
68	second	143	industry	218	31%	293	mainly
69	expense	144	program	219	profitability	294	value
70	income	145	despite	220	account_for	295	non-interest
71	new_york_stock_exchange	146	expectation	221	sharply	296	tiffany
72	close	147	financial	222	april	297	as_well
73	per-share	148	sept.	223	system	298	18%
74	total	149	co.	224	nonrecurring	299	gas
75	reserve	150	coleco	225	problem	300	acquire

Table B.15: **ERN** Topic signature in set **PH**.

FAB Topic signature (PH)							
1	restaurant	76	u.s.	151	spokesman	226	bottler
2	mr.	77	eat	152	lencioni	227	really
3	p&g	78	pepsi	153	researcher	228	become
4	food	79	advertising	154	freeze	229	penaltyaction
5	brand	80	employee	155	colgate	230	head
6	smoking	81	vice_president	156	fda	231	pay
7	wine	82	court	157	unit	232	drinker
8	tobacco	83	call	158	come	233	small
9	beer	84	city	159	local	234	mexican
10	product	85	allegheny	160	go	235	orloff
11	cigarette	86	bread	161	give	236	five
12	anheuser-busch	87	study	162	division	237	place
13	rjr	88	work	163	olive	238	disease
14	coca-cola	89	toothpaste	164	koch	239	scientist
15	company	90	diet	165	time	240	colon
16	executive	91	pillsbury	166	microwave	241	drug
17	guinness	92	think	167	italian	242	man
18	business	93	bottlers	168	art	243	allegation
19	raisin	94	even	169	large	244	attorney
20	drink	95	diner	170	modelo	245	ask
21	unilever	96	use	171	bronfman	246	7-eleven
22	smoke	97	co.	172	1985	247	seem
23	new	98	sell	173	distiller	248	such_as
24	people	99	chain	174	customer	249	problem
25	consumer	100	find	175	operation	250	national
26	edgar	101	store	176	own	251	leave
27	seagram	102	coffee	177	buckley	252	owner
28	get	103	she	178	sweetener	253	continue
29	busch	104	philip	179	inc.	254	grow
30	brewery	105	bottle	180	ingredient	255	her
31	higa	106	fat	181	federal	256	ms.
32	nutrasweet	107	want	182	atlanta	257	more_than
33	market	108	dean	183	allege	258	david
34	bottling	109	brother	184	director	259	meal
35	year	110	tax	185	see	260	produce
36	enrico	111	winery	186	conagra	261	introduce
37	pizza	112	add	187	general	262	thing
38	smoker	113	spe	188	age	263	right
39	southland	114	zagat	189	distributor	264	start
40	pepsico	115	know	190	hill	265	brandy
41	sale	116	try	191	report	266	gallery
42	marketing	117	water	192	stafford	267	martino
43	former	118	johnson	193	brewer	268	pine
44	nabisco	119	chocolate	194	cost	269	official
45	taste	120	claim	195	public	270	soup
46	health	121	liquor	196	beech-nut	271	murree
47	flavor	122	saunders	197	soft-drink	272	worldwide
48	manager	123	fine	198	spoor	273	agency
49	state	124	research	199	thomas	274	supermarket
50	wilson	125	where	200	father	275	ordinance
51	analyst	126	kraft	201	country	276	category
52	worker	127	help	202	move	277	dr.
53	make	128	seven-up	203	risk	278	1983
54	suit	129	morris	204	group	279	ago
55	long	130	way	205	expect	280	several
56	corona	131	new_york	206	top	281	office
57	investigation	132	harford	207	woman	282	restaurateur
58	industry	133	olestra	208	young	283	say
59	beverage	134	price	209	fried	284	competition
60	ice_cream	135	cheese	210	hanley	285	tell
61	soft_drink	136	job	211	has_been	286	cookie
62	take	137	milk	212	great	287	dinner
63	american	138	christian	213	artist	288	idea
64	family	139	big	214	test	289	garden
65	law	140	show	215	billion	290	million
66	case	141	reynolds	216	week	291	policy
67	aspartame	142	slivovitz	217	california	292	anheuser
68	alcohol	143	cancer	218	carnation	293	major
69	name	144	ban	219	beverly	294	bottled
70	charge	145	drinking	220	union	295	giant
71	meat	146	calorie	221	chairman	296	beef
72	lunch	147	though	222	winston-salem	297	although
73	change	148	president	223	develop	298	headquarters
74	plant	149	diaper	224	wendy	299	whether
75	chicken	150	nonsmoker	225	government	300	sauce

Table B.16: **FAB** Topic signature in set **PH**.

FIN Topic signature (PH)							
1	lawyer	76	cpa	151	premium	226	allow
2	tax	77	company	152	education	227	estate
3	welfare	78	college	153	panel	228	national
4	state	79	american	154	expense	229	decision
5	firm	80	make	155	offer	230	private
6	mr.	81	change	156	congress	231	jury
7	insurance	82	bar	157	has_been	232	michigan
8	pension	83	leasing	158	try	233	government
9	retirement	84	rate	159	investor	234	great
10	law	85	executive	160	mason	235	receive
11	judge	86	must	161	support	236	1986
12	client	87	member	162	study	237	question
13	benefit	88	damage	163	corporation	238	kumble
14	accountant	89	problem	164	health	239	lose
15	accounting	90	recipient	165	litigation	240	auditing
16	fasb	91	student	166	salary	241	enough
17	case	92	life	167	asset	242	where
18	employer	93	age	168	counsel	243	house
19	program	94	proposal	169	low	244	know
20	court	95	high	170	charge	245	average
21	accounting_firm	96	payment	171	life_insurance	246	poverty
22	income	97	minimum	172	mandatory	247	call
23	year	98	even	173	several	248	review
24	ir	99	would_be	174	attorney	249	young
25	employee	100	former	175	current	250	touche
26	social_security	101	suit	176	exam	251	mother
27	partner	102	association	177	become	252	financial_statement
28	pay	103	corporate	178	increase	253	agency
29	people	104	profession	179	rampell	254	immigration
30	child	105	standard	180	cover	255	interest
31	plan	106	give	181	law_school	256	penalty
32	cost	107	coverage	182	industry	257	president
33	ira	108	contribution	183	board	258	waterhouse
34	rule	109	help	184	add	259	involve
35	financial	110	provide	185	estimate	260	pcw
36	work	111	practice	186	florida	261	tell
37	cromwell	112	school	187	church	262	defendant
38	money	113	individual	188	organization	263	large
39	worker	114	investment	189	right	264	peat
40	legal	115	major	190	whether	265	ago
41	bill	116	greenfield	191	often	266	vote
42	insurers	117	go	192	see	267	man
43	law_firm	118	committee	193	early	268	isham
44	sullivan	119	write	194	raise	269	aetna
45	institute	120	manager	195	commission	270	today
46	she	121	spouse	196	report	271	consulting
47	new	122	system	197	number	272	apply
48	audit	123	file	198	official	273	finley
49	sec	124	find	199	return	274	peer_review
50	family	125	legislation	200	think	275	criminal
51	public	126	award	201	venture_capital	276	capital
52	policy	127	time	202	rather	277	governor
53	business	128	insurance_company	203	consultant	278	accord
54	job	129	tort	204	settlement	279	five
55	her	130	such_as	205	create	280	annual
56	claim	131	service	206	invest	281	spend
57	get	132	small	207	retire	282	reduce
58	aba	133	university	208	long	283	saving
59	take	134	act	209	can't	284	believe
60	liability	135	trust	210	black	285	washington
61	fund	136	aid	211	amount	286	marwick
62	reform	137	more_than	212	ask	287	force
63	fee	138	woman	213	cuomo	288	keep
64	need	139	parent	214	hyatt	289	account
65	auditor	140	loan	215	poor	290	consider
66	planner	141	risk	216	professional	291	billion
67	new_york	142	williams	217	ms.	292	pass
68	require	143	california	218	limit	293	lawsuit
69	big	144	leave	219	generally	294	form
70	office	145	u.s.	220	1985	295	note
71	deduction	146	group	221	annuity	296	a_lot
72	issue	147	way	222	seek	297	person
73	want	148	bork	223	same	298	for_example
74	use	149	hire	224	living_trust	299	plaintiff
75	federal	150	sue	225	fraud	300	chairman

Table B.17: **FIN** Topic signature in set **PH**.

LNG Topic signature (PH)							
1	pipeline	76	crediting	151	interim	226	damage
2	gas	77	pipelines'	152	public_service	227	tennessee
3	natural_gas	78	problem	153	mile	228	gas-purchase
4	coastal	79	sell	154	pay	229	large
5	cubic_foot	80	judgment	155	sign	230	major
6	contract	81	oil_company	156	development	231	half
7	energy	82	county	157	eastern	232	co
8	producer	83	allow	158	choke	233	concern
9	transco	84	transcontinental	159	award	234	baladi
10	commission	85	force_majeure	160	canadian	235	brae
11	oil	86	million	161	colorado	236	trans-pan
12	well	87	reserve	162	distributor	237	recent
13	natural-gas	88	kasputys	163	austria	238	british
14	exploration	89	ltd.	164	obligate	239	current
15	customer	90	seagull	165	high-priced	240	give
16	regulatory	91	delivery	166	operate	241	wyoming
17	federal	92	operator	167	lawsuit	242	ship
18	valero	93	offshore	168	lasalle	243	breach-of-contract
19	take-or-pay	94	appeal	169	demand	244	pacific
20	occidental	95	amoco	170	production	245	proposal
21	midcon	96	jury	171	pipe	246	southern
22	supply	97	kopp	172	gas_system	247	bring
23	ferc	98	north_sea	173	plan	248	must
24	unit	99	agency	174	provide	249	certain
25	rule	100	mr.	175	capacity	250	refuse
26	anadarko	101	davidson	176	arkla	251	currently
27	panhandle	102	fuel	177	ruhrgas	252	distiller
28	houston	103	settlement	178	charge	253	decontrol
29	transport	104	florida	179	policy	254	northwest
30	price	105	change	180	provision	255	amount
31	foot	106	transcanada	181	hold	256	much_as
32	co.	107	brooklyn	182	credit	257	free
33	petroleum	108	market	183	she	258	commitment
34	court	109	reduce	184	district	259	consortium
35	pipeline_company	110	$412	185	open	260	expensive
36	ruling	111	settle	186	noverco	261	inc.
37	enron	112	year	187	construction	262	1985
38	arco	113	alberta	188	purchase	263	plant
39	transmission	114	u.s.	189	northeast	264	united
40	interstate	115	seek	190	panel	265	antitrust
41	user	116	pogo	191	approval	266	vermont
42	cost	117	association	192	line	267	trial
43	consolidated	118	barrel	193	a_thousand	268	take
44	day	119	1	194	file	269	shipment
45	depth	120	indiana	195	agree	270	spot
46	distrigas	121	construct	196	supplier	271	alaska
47	order	122	discovery	197	canada	272	okla.
48	utility	123	access	198	louisiana	273	chairwoman
49	liability	124	trillion	199	drilling	274	bethlehem
50	corp.	125	shell	200	coast	275	weather
51	anr	126	dispute	201	area	276	has_been
52	buy	127	claim	202	discover	277	low
53	drill	128	opening	203	pay_for	278	make
54	interest	129	case	204	build	279	bubble
55	transportation	130	decision	205	daily	280	fpl
56	tankersley	131	oneok	206	billion	281	expect
57	appeals_court	132	sonatrach	207	esso	282	several
58	field	133	citrus	208	houston-based	283	say
59	rate	134	williams	209	regulation	284	note
60	obligation	135	resource	210	propose	285	on_the_spot
61	primark	136	involve	211	attorney	286	allege
62	industry	137	condensate	212	oklahoma	287	exist
63	project	138	partnership	213	united_kingdom	288	mobil
64	open-access	139	agreement	214	partner	289	independent
65	company	140	suit	215	venture	290	delay
66	flow	141	program	216	plain	291	analyst
67	northern	142	industrial	217	border	292	review
68	new	143	state	218	approve	293	bill
69	directly	144	find	219	peru	294	recover
70	gulf_of_mexico	145	quantity	220	force	295	hope
71	hesse	146	own	221	union	296	per
72	judge	147	columbia	222	source	297	previously
73	spokesman	148	sale	223	additional	298	gas_company
74	phillips	149	system	224	burn	299	produce
75	texas	150	others	225	group	300	begin

Table B.18: **LNG** Topic signature in set **PH**.

MIN Topic signature (PH)							
1	steel	76	japanese	151	local	226	bourke
2	ton	77	rate	152	nucor	227	helton
3	week	78	fall	153	dispute	228	usw
4	capability	79	give	154	member	229	31
5	usx	80	compare	155	steel_plate	230	thyssen
6	mine	81	trumka	156	oberhausen	231	employment
7	coal	82	pact	157	make	232	strong
8	steelmaker	83	inspector	158	current	233	chairman
9	union	84	mining	159	blast_furnace	234	quarter
10	iron	85	copper	160	grievance	235	pound
11	aluminum	86	ruhr	161	big	236	expire
12	utilization	87	report	162	alloy	237	health
13	mill	88	parry	163	island	238	close
14	production	89	rial	164	vote	239	agree
15	msha	90	pittsburgh	165	fetterolf	240	customer
16	to_date	91	hagiwara	166	post	241	seek
17	industry	92	national	167	bargaining	242	hurt
18	alcoa	93	rise	168	action	243	negotiation
19	mr.	94	no.	169	.....	244	early
20	worker	95	agreement	170	investigator	245	know
21	alcan	96	united	171	july	246	ltd.
22	institute	97	talk	172	carbon	247	further
23	bethlehem	98	plan	173	.....	248	raise
24	previous	99	official	174	flat-rolled	249	fire
25	output	100	cost	175	president	250	add
26	year	101	lukens	176	jan.	251	nearly
27	steelworker	102	ago	177	spokesman	252	total
28	strike	103	new	178	facility	253	has_been
29	produce	104	japan	179	china	254	side
30	miner	105	occidental	180	take	255	.....
31	raw-steel	106	toth	181	four	256	say
32	nippon	107	.....	182	pension	257	problem
33	metric_ton	108	reduce	183	cite	258	joint_venture
34	nation	109	product	184	project	259	scrap
35	price	110	o'neill	185	several	260	source
36	earlier	111	.....	186	director	261	committee
37	net_ton	112	expect	187	aluminium	262	boost
38	goode	113	market	188	u.s.	263	canadian
39	job	114	.....	189	accident	264	see
40	contract	115	demand	190	reach	265	oct.
41	calculation	116	end	191	tin	266	sept.
42	percent	117	operation	192	citation	267	virginia
43	moseley	118	wage	193	violation	268	.....
44	capacity	119	reynolds	194	labor_force	269	west_virginia
45	american	120	month	195	hour	270	despite
46	strunk	121	analyst	196	ingot	271	eliminate
47	producer	122	major	197	consolidation	272	pittsburgh-based
48	sheet	123	settlement	198	minimills	273	saturday
49	smelter	124	quebec	199	steel_mill	274	leader
50	increase	125	operator	200	decline	275	ec
51	umw	126	british	201	follow	276	industrial
52	inland	127	creek	202	death	277	community
53	kaiser	128	begin	203	concession	278	ratification
54	agency	129	decrease	204	employee	279	noranda
55	operate	130	continue	205	staff	280	retirement
56	labor	131	graham	206	miners'	281	in_addition
57	metal	132	cut	207	steel_company	282	business
58	company	133	million	208	order	283	long
59	sharon	134	large	209	1	284	brussels
60	plant	135	world	210	yen	285	her
61	work	136	smelters	211	think	286	help
62	ltv	137	washington	212	effective	287	ratify
63	plate	138	maier	213	government	288	3
64	kokan	139	unit	214	represent	289	call
65	steelmaking	140	alumina	215	export	290	furnace
66	num	141	pa.	216	crude-steel	291	2
67	hendry	142	steel_production	217	discount	292	meeting
68	safety	143	corp.	218	1984	293	layoff
69	roderick	144	co.	219	management	294	manager
70	use	145	loss	220	international	295	competitor
71	restart	146	mrs.	221	benefit	296	effort
72	shipment	147	five	222	williams	297	executive
73	indicate	148	aug.	223	high	298	code
74	griever	149	move	224	march	299	schedule
75	design	150	utah	225	lose	300	smelting

Table B.19: **MIN** Topic signature in set **PH**.

MKT Topic signature (PH)							
1	ad	76	king	151	borden	226	head
2	advertising	77	win	152	right	227	music
3	thompson	78	ford	153	beatles	228	jingle
4	saatchi	79	unit	154	feature	229	believe
5	jwt	80	award	155	spot	230	john
6	mr.	81	even	156	local	231	old
7	ad_agency	82	employee	157	explain	232	need
8	commercial	83	army	158	film	233	idea
9	she	84	vice_president	159	viewer	234	executive_vice_president
10	tax	85	television	160	metter	235	sponsor
11	people	86	use	161	kid	236	help
12	her	87	manning	162	firm	237	r
13	client	88	public_relations	163	keep	238	little
14	executive	89	small-business	164	plc	239	begin
15	advertiser	90	small	165	find	240	week
16	campaign	91	see	166	leave	241	move
17	wpp	92	create	167	run	242	ago
18	florida	93	chairman	168	tell	243	n.a
19	agency	94	disney	169	cbs	244	immigrant
20	account	95	take	170	world	245	spokesman
21	elsie	96	bork	171	add	246	probably
22	johnston	97	seem	172	financial	247	johnson
23	yuppie	98	rebate	173	symbol	248	pitch
24	name	99	nike	174	top	249	chevron
25	scanlon	100	try	175	a_few	250	sony
26	actor	101	brand	176	buy	251	pac
27	ogilvy	102	chief_executive	177	us	252	sing
28	marketing	103	look	178	aipac	253	why
29	spielvogel	104	co.	179	leibovitz	254	job
30	group	105	coke	180	more_than	255	advertisement
31	get	106	tyco	181	always	256	own
32	rubicam	107	pay	182	world-wide	257	say
33	business	108	way	183	star	258	mather
34	bates	109	play	184	organization	259	1986
35	walter	110	race	185	age	260	madison
36	o'donnell	111	work	186	statement	261	sell
37	toy	112	blond	187	spend	262	federation
38	young	113	claim	188	ad_campaign	263	can't
39	new_york	114	show	189	become	264	bill
40	celebrity	115	medium	190	cow	265	delfin
41	creative	116	spending	191	corporate	266	pfundstein
42	state	117	magazine	192	consider	267	voice-overs
43	burger	118	contract	193	never	268	trade
44	big	119	car	194	omnicom	269	marketers
45	sorrell	120	woman	195	hard	270	black
46	beer	121	o'neill	196	kirby	271	dixons
47	think	122	feel	197	suit	272	day
48	ayer	123	inc.	198	law	273	ever
49	backer	124	man	199	where	274	start
50	j.	125	event	200	to_do	275	don
51	new	126	voice	201	member	276	ask
52	year	127	president	202	jacoby	277	handle
53	coen	128	singer	203	lawyer	278	accord
54	promotion	129	los_angeles	204	camera	279	happen
55	billing	130	simonds-gooding	205	boy	280	coca-cola
56	make	131	money	206	appear	281	public
57	consumer	132	change	207	red_cross	282	a_little
58	want	133	analyst	208	thing	283	ir
59	service	134	industry	209	association	284	although
60	product	135	design	210	source	285	grey
61	tv	136	lose	211	america	286	rap
62	company	137	good	212	court	287	bring
63	advertising_agency	138	give	213	political	288	live
64	ms.	139	problem	214	management	289	interpublic
65	director	140	pr	215	owner	290	atari
66	u.s.	141	trademark	216	write	291	michael
67	go	142	image	217	former	292	pull
68	usa	143	anheuser-busch	218	city	293	really
69	time	144	marathon	219	press	294	professional
70	judge	145	song	220	review	295	cleese
71	american	146	for_example	221	watch	296	nintendo
72	national	147	has_been	222	house	297	fee
73	worldwide	148	call	223	decision	298	ge
74	know	149	designer	224	the_most	299	mexican
75	office	150	case	225	most_of	300	martinez

Table B.20: **MKT** Topic signature in set **PH**.

MON Topic signature (PH)							
1	trade	76	new	151	west	226	toward
2	u.s.	77	protectionist	152	sen.	227	action
3	export	78	contras	153	become	228	nicaraguan
4	japan	79	tokyo	154	bank	229	manufacturer
5	billion	80	figure	155	service	230	inflation
6	japanese	81	demand	156	must	231	use
7	import	82	legislation	157	exporter	232	program
8	government	83	western	158	current	233	strong
9	surplus	84	germany	159	large	234	private
10	foreign	85	problem	160	australia	235	october
11	country	86	industry	161	end	236	diplomat
12	economic	87	get	162	1987	237	war
13	dollar	88	military	163	spending	238	exchange_rate
14	economy	89	effort	164	state	239	money
15	mr.	90	has_been	165	widen	240	despite
16	soviet	91	amendment	166	central_bank	241	south_african
17	year	92	lira	167	continue	242	mainly
18	good	93	leader	168	transfer	243	economics
19	deficit	94	need	169	mean	244	expect
20	taiwan	95	want	170	rate	245	relation
21	official	96	rep.	171	free-trade	246	case
22	trade_deficit	97	way	172	contra	247	further
23	congress	98	people	173	cost	248	mrs.
24	currency	99	sweden	174	custom	249	raise
25	american	100	u.s	175	big	250	national
26	aid	101	business	176	barrier	251	tell
27	ministry	102	soviet_union	177	negotiation	252	seal
28	world	103	fall	178	trillion	253	border
29	policy	104	change	179	issue	254	lose
30	earlier	105	help	180	report	255	remain
31	tax	106	trade_in	181	provision	256	shipment
32	canada	107	would_be	182	plan	257	retaliation
33	growth	108	worker	183	industrial	258	contrast
34	political	109	seoul	184	reagan_administration	259	businessman
35	domestic	110	abroad	185	such_as	260	yeutter
36	tariff	111	1986	186	company	261	ally
37	president	112	mulroney	187	spain	262	black
38	current_account	113	1985	188	commerce_department	263	september
39	rise	114	quota	189	france	264	communist
40	south_korea	115	kong	190	control	265	the_two
41	nation	116	labor	191	white_house	266	important
42	administration	117	take	192	package	267	capital
43	market	118	prime_minister	193	equivalent	268	november
44	china	119	hong	194	competitiveness	269	iran
45	south_africa	120	million	195	congressional	270	overseas
46	european	121	singapore	196	democratic	271	comprise
47	gephardt	122	washington	197	part	272	february
48	trade_bill	123	reform	198	gold	273	cause
49	measure	124	aquino	199	protectionism	274	baldrige
50	sanction	125	democrat	200	force	275	plant
51	house	126	try	201	week	276	move
52	even	127	go	202	minister	277	whether
53	europe	128	high	203	accord	278	rule
54	franc	129	nakasone	204	more_than	279	finance
55	reagan	130	party	205	think	280	create
56	yen	131	total	206	imbalance	281	where
57	investment	132	oecd	207	find	282	opposition
58	grow	133	work	208	see	283	likely
59	increase	134	reserve	209	member	284	allow
60	month	135	major	210	development	285	kronor
61	korea	136	agreement	211	america	286	sector
62	make	137	show	212	brazil	287	hard
63	bill	138	unilateral	213	begin	288	great
64	price	139	reduce	214	state_department	289	consumer
65	economist	140	pressure	215	cut	290	production
66	nicaragua	141	job	216	proposal	291	italy
67	canadian	142	decline	217	come	292	budget
68	korean	143	low	218	monetary	293	time
69	support	144	mozambique	219	impose	294	ship
70	ec	145	law	220	philippine	295	january
71	senate	146	britain	221	taiwanese	296	buy
72	product	147	talk	222	rand	297	cooperation
73	international	148	give	223	committee	298	small
74	statistic	149	free_trade	224	moscow	299	competitive
75	narrow	150	seem	225	assistance	300	pact

Table B.21: **MON** Topic signature in set **PH**.

PET Topic signature (PH)							
1	texaco	76	find	151	petro-canada	226	triton
2	pennzoil	77	offshore	152	major	227	arctic
3	oil	78	co.	153	think	228	business
4	barrel	79	rawl	154	industry	229	depth
5	getty	80	take	155	sulpetro	230	trust
6	mr.	81	area	156	try	231	ruling
7	exxon	82	government	157	bond-and-lien	232	canada
8	oil_company	83	tanker	158	casseb	233	week
9	mobil	84	royal	159	gas	234	would_be
10	texas	85	liedtke	160	increase	235	legal
11	iran	86	murkowski	161	soviet	236	estimate
12	price	87	executive	162	venezuela	237	agreement
13	gasoline	88	jury	163	settle	238	long
14	petroleum	89	energy_department	164	oil_production	239	discover
15	refinery	90	decision	165	dutchshell	240	property
16	tesoro	91	inventory	166	corp.	241	expect
17	field	92	kpc	167	foreign	242	give
18	well	93	analyst	168	conoco	243	fuel
19	iraq	94	tax	169	settlement	244	capacity
20	crude_oil	95	rig	170	world	245	want
21	arco	96	rule	171	new	246	former
22	saudi	97	big	172	plan	247	water
23	court	98	hunt	173	move	248	refined
24	shell	99	own	174	claim	249	boies
25	opec	100	appeals_court	175	spokesman	250	suggest
26	exploration	101	president	176	demand	251	need
27	billion	102	farris	177	million	252	continue
28	case	103	yemen	178	supply	253	richfield
29	drilling	104	sen.	179	call	254	way
30	judgment	105	1985	180	time	255	plain
31	iranian	106	discovery	181	american	256	member
32	u.s.	107	pemex	182	nearly	257	pletcher
33	judge	108	johnson	183	development	258	q8
34	kinnear	109	tavoulareas	184	sweet	259	tabasco
35	kuwait	110	damage	185	family	260	month
36	energy	111	unit	186	go	261	greece
37	lawyer	112	lien	187	come	262	interfere_with
38	saudi_arabia	113	dollar	188	produce	263	sell
39	gulf	114	seek	189	newfoundland	264	ltd.
40	occidental	115	high	190	federal	265	add
41	chevron	116	marketing	191	board	266	rise
42	state	117	law	192	platform	267	judicial
43	production	118	alaska	193	see	268	oil_business
44	petrofina	119	federal_court	194	obtain	269	choke
45	amoco	120	asset	195	gallon	270	tehran
46	justice	121	where	196	norway	271	norwegian
47	supreme_court	122	country	197	pump	272	although
48	crude	123	frontiers-alaska	198	chairman	273	help
49	refining	124	iraqi	199	more_than	274	never
50	year	125	even	200	issue	275	refiner
51	reserve	126	order	201	political	276	post
52	interest	127	standard	202	output	277	average
53	attack	128	cost	203	several	278	decrane
54	unocal	129	oil_industry	204	tell	279	become
55	day	130	private	205	man	280	though
56	sun	131	get	206	yesterday	281	position
57	kuwaiti	132	right	207	operate	282	spend
58	pipeline	133	trial	208	lease	283	middle_east
59	jamail	134	make	209	large	284	whether
60	drill	135	pay	210	turkey	285	charge
61	west	136	bankruptcy	211	phillips	286	subsidiary
62	north_sea	137	mile	212	contribution	287	bp
63	foot	138	remain	213	persian_gulf	288	fall
64	company	139	flow	214	j.	289	senior
65	contract	140	hall	215	letter	290	late
66	houston	141	new_york	216	deep-water	291	run
67	appeal	142	begin	217	litigation	292	halbur
68	murray	143	has_been	218	operation	293	oil-price
69	statoil	144	mecca	219	total	294	name
70	war	145	refuge	220	minister	295	source
71	award	146	cook	221	grade	296	four
72	bond	147	ask	222	1986	297	partner
73	official	148	north	223	downstream	298	region
74	johnsen	149	slope	224	daily	299	1981
75	project	150	taylor	225	crude-oil	300	allege

Table B.22: **PET** Topic signature in set **PH**.

PHA Topic signature (PH)							
1	drug	76	protein	151	official	226	san_francisco
2	patient	77	smithkline	152	detect	227	within
3	dr.	78	need	153	man	228	type
4	aid	79	genetics	154	public	229	chemical
5	hospital	80	take	155	journal	230	kit
6	fda	81	johnson	156	whether	231	placebo
7	doctor	82	make	157	claim	232	sell
8	test	83	streptokinase	158	science	233	heart-attack
9	tpa	84	get	159	burroughs-wellcome	234	prevent
10	cancer	85	market	160	believe	235	great
11	virus	86	wellcome	161	procedure	236	office
12	disease	87	state	162	effective	237	suggest
13	study	88	call	163	substance	238	euthanasia
14	vaccine	89	body	164	available	239	superoxide
15	medical	90	committee	165	drug_company	240	want
16	treatment	91	lilly	166	squibb	241	safety
17	physician	92	show	167	high	242	aids-related
18	research	93	know	168	medication	243	ama
19	researcher	94	result	169	require	244	method
20	genentech	95	give	170	number	245	technology
21	institute	96	care	171	early	246	roche
22	patent	97	child	172	advisory	247	medical_center
23	health	98	die	173	time	248	smear
24	lab	99	system	174	application	249	prove
25	blood	100	center	175	benefit	250	conduct
26	scientist	101	side_effect	176	development	251	disorder
27	use	102	risk	177	analyst	252	capoten
28	azt	103	biogen	178	expect	253	blood_cell
29	cell	104	approve	179	small	254	for_example
30	cholesterol	105	national	180	decision	255	version
31	clinical	106	receive	181	marketing	256	adult
32	laboratory	107	report	182	recommend	257	as_well
33	treat	108	clinic	183	issue	258	prof.
34	biotechnology	109	victim	184	hope	259	mean
35	develop	110	infect	185	screening	260	lung
36	she	111	clot	186	go	261	severe
37	trial	112	produce	187	society	262	inc.
38	pharmaceutical	113	cost	188	health-care	263	try
39	medicine	114	gene	189	concern	264	illness
40	testing	115	nurse	190	rule	265	potential
41	icn	116	begin	191	technique	266	supply
42	human	117	virazole	192	director	267	that_is
43	merck	118	even	193	more_than	268	1985
44	people	119	several	194	become	269	week
45	new	120	dentist	195	country	270	effect
46	animal	121	acquired_immune_deficiency_syndrome	196	family	271	practice
47	american	122	think	197	suit	272	area
48	her	123	scientific	198	enzyme	273	serious
49	therapy	124	dose	199	involve	274	marion
50	approval	125	level	200	the_most	275	change
51	experimental	126	surgery	201	federal	276	damage
52	pap	127	panel	202	month	277	recommendation
53	heart_attack	128	tell	203	kidney	278	possible
54	alzheimer	129	process	204	day	279	dissolve
55	infection	130	young	205	california	280	inject
56	u.s.	131	must	206	review	281	blood_clot
57	university	132	group	207	important	282	same
58	product	133	death	208	way	283	slide
59	food	134	tha	209	add	284	feel
60	cause	135	form	210	pain	285	a_few
61	find	136	license	211	tumor	286	yet
62	administration	137	government	212	such_as	287	experience
63	woman	138	monoclonal	213	where	288	sale
64	heart	139	life	214	right	289	effort
65	work	140	has_been	215	amgen	290	industry
66	antibody	141	agency	216	transplant	291	name
67	case	142	help	217	effectiveness	292	whose
68	company	143	prescription	218	compound	293	appear
69	upjohn	144	often	219	health_care	294	baxter
70	mr.	145	question	220	finding	295	interferon
71	year	146	provide	221	ask	296	mortality
72	immune	147	see	222	specialist	297	nih
73	datum	148	program	223	surgeon	298	device
74	tissue	149	genetic	224	would_be	299	good
75	problem	150	genetically	225	association	300	fermenta

Table B.23: **PHA** Topic signature in set **PH**.

PUB Topic signature (PH)							
1	mr.	76	know	151	owner	226	help
2	magazine	77	has_been	152	castillo	227	us
3	book	78	trump	153	baseball	228	libel
4	newspaper	79	information	154	advertiser	229	right
5	editor	80	take	155	call	230	use
6	publisher	81	tehran	156	kaminsky	231	mrs.
7	singapore	82	doubleday	157	university	232	hulbert
8	journal	83	pravda	158	big	233	begin
9	herald	84	mcgovern	159	run	234	candidate
10	paper	85	sport	160	family	235	study
11	circulation	86	issue	161	money	236	lyon
12	news	87	textbook	162	employee	237	pay
13	seib	88	entrepreneur	163	loeb	238	manuscript
14	publish	89	make	164	brack	239	propaganda
15	press	90	former	165	claim	240	more_than
16	publication	91	group	166	foreign	241	black
17	publishing	92	journalism	167	literary	242	office
18	she	93	lampoon	168	newsstand	243	publishing_house
19	story	94	novel	169	number	244	crime
20	reader	95	hispanic	170	judge	245	allege
21	time	96	read	171	give	246	hammond
22	reporter	97	knight-ridder	172	small	247	city
23	write	98	day	173	try	248	fact
24	schuster	99	staff	174	entrepreneurship	249	nation
25	article	100	ross	175	ask	250	decide
26	ms.	101	belo	176	law	251	lee
27	iran	102	review	177	bork	252	snyder
28	her	103	miami	178	source	253	five
29	journalist	104	cuban	179	title	254	mcmanus
30	copy	105	u.s.	180	stamp	255	great
31	daily	106	printing	181	way	256	bennett
32	simon	107	executive	182	leave	257	accord
33	soviet	108	even	183	sell	258	later
34	hart	109	country	184	seem	259	top
35	gannett	110	managing_editor	185	change	260	coverage
36	advertising	111	find	186	mirror	261	school
37	government	112	report	187	new_yorker	262	bertelsmann
38	wall_street	113	column	188	awsj	263	columnist
39	author	114	bantam	189	shawn	264	radio
40	asian	115	cover	190	cost	265	organization
41	print	116	political	191	month	266	police
42	page	117	suit	192	word	267	revenue
43	manga	118	world	193	state	268	rath
44	writer	119	never	194	add	269	vitale
45	editorial	120	hofmann	195	ad	270	picture
46	iranian	121	glasnost	196	subject	271	decision
47	neuharth	122	life	197	character	272	friend
48	evans	123	think	198	order	273	casey
49	people	124	job	199	spend	274	zuckerman
50	letter	125	balloon	200	come	275	to_do
51	house	126	become	201	section	276	where
52	ethnos	127	subscription	202	case	277	party
53	woman	128	president	203	long	278	something
54	medium	129	official	204	explain	279	for_example
55	work	130	week	205	show	280	idea
56	today	131	washington	206	upi	281	operation
57	murdoch	132	spitball	207	dispute	282	although
58	post	133	detention	208	young	283	war
59	year	134	see	209	1985	284	charge
60	new_york	135	maxwell	210	company	285	why
61	new	136	question	211	newsletter	286	the_most
62	tell	137	own	212	thing	287	quote
63	usa	138	newhouse	213	prime_minister	288	career
64	south-north	139	man	214	several	289	action
65	dow	140	photographer	215	court	290	text
66	want	141	subscriber	216	edition	291	vanity_fair
67	laxalt	142	appear	217	ago	292	win
68	get	143	american	218	list	293	four
69	jones	144	weekly	219	gorbachev	294	venture
70	name	145	guild	220	professional	295	thai
71	random	146	newsweek	221	market	296	grunwald
72	library	147	western	222	public	297	passport
73	business	148	batuigas	223	need	298	independent
74	correspondent	149	sunday	224	start	299	anglo
75	free	150	go	225	release	300	research

Table B.24: **PUB** Topic signature in set **PH**.

REL Topic signature (PH)							
1	hotel	76	income	151	public	226	mets
2	real_estate	77	headquarters	152	transaction	227	american
3	property	78	county	153	agree	228	industry
4	building	79	find	154	inn	229	several
5	city	80	japanese	155	limited	230	part
6	land	81	rental	156	spend	231	place
7	mall	82	csr	157	expect	232	at_least
8	mr.	83	wagon-lit	158	concern	233	midtown
9	tenant	84	merrill	159	high	234	syndicators
10	tax	85	move	160	amfac	235	offering
11	real-estate	86	gemcraft	161	report	236	village
12	partnership	87	shuwa	162	want	237	program
13	housing	88	more_than	163	metropolitan	238	lynch
14	apartment	89	california	164	former	239	reinsdorf
15	office	90	unit	165	major	240	tonda
16	landlord	91	community	166	management	241	say
17	lease	92	build	167	develop	242	file
18	developer	93	realty	168	way	243	simon
19	project	94	interest	169	bond	244	jointly
20	development	95	chicago	170	consider	245	hill
21	downtown	96	take	171	fair	246	problem
22	rent	97	guest	172	right	247	operation
23	space	98	town	173	service	248	billion
24	crow	99	cost	174	government	249	can't
25	site	100	cushman	175	street	250	proposal
26	manhattan	101	pedestrian	176	construction	251	laguna
27	square_foot	102	sheraton	177	month	252	capital
28	partner	103	use	178	structure	253	old
29	trammell	104	change	179	recently	254	rate
30	acre	105	firm	180	brother	255	denver
31	cave	106	state	181	southland	256	mrs.
32	new_york	107	court	182	control	257	time
33	silverstein	108	pay	183	executive	258	four
34	broker	109	co.	184	purchase	259	form
35	tower	110	for_example	185	vacancy	260	such_as
36	new	111	boston	186	olympia	261	leave
37	year	112	complex	187	number	262	recent
38	sell	113	world	188	campeau	263	gain
39	office_building	114	dollar	189	begin	264	low
40	sale	115	make	190	must	265	would_be
41	owner	116	local	191	luxury	266	hamilton
42	house	117	survey	192	see	267	texas
43	plan	118	buyer	193	open	268	think
44	own	119	harcourt	194	webbs	269	total
45	balcor	120	holiday	195	large	270	loan
46	home	121	where	196	return	271	corporate
47	company	122	coldwell	197	work	272	try
48	park	123	hilton	198	become	273	wakefield
49	investment	124	stay	199	corp.	274	foreign
50	investor	125	criswell	200	niguel	275	payment
51	residential	126	even	201	show	276	lose
52	commercial	127	big	202	near	277	koch
53	area	128	auction	203	couple	278	corporation
54	disney	129	her	204	decide	279	growth
55	center	130	prudential	205	historic	280	private
56	herscu	131	ago	206	charge	281	household
57	room	132	ir	207	value	282	a_lot
58	buy	133	money	208	good	283	los_angeles
59	hall	134	manage	209	foundation	284	penalty
60	price	135	wright	210	lender	285	loss
61	market	136	arvida	211	inc.	286	salomon
62	get	137	source	212	invest	287	york
63	mayor	138	mortgage	213	new_york_city	288	involve
64	urban	139	vacancy_rate	214	complete	289	live
65	resident	140	name	215	boundary	290	six
66	dallas	141	bill	216	measure	291	cdcs
67	people	142	trump	217	berger	292	knab
68	loft	143	surveyor	218	has_been	293	ulundi
69	financial	144	add	219	country	294	international
70	law	145	facility	220	small	295	spokesman
71	group	146	she	221	deal	296	10%
72	million	147	condominium	222	require	297	stanger
73	lisc	148	official	223	vice_president	298	today
74	business	149	manager	224	florida	299	come
75	u.s.	150	go	225	five	300	jersey

Table B.25: **REL** Topic signature in set **PH**.

RET Topic signature (PH)							
1	store	76	week	151	below	226	billion
2	+	77	get	152	survey	227	stock
3	retailer	78	co.	153	plan	228	mervyn
4	sale	79	people	154	change	229	jeffrey
5	sears	80	buy	155	vice_president	230	several
6	mall	81	bloomingdale	156	bookseller	231	macy
7	christmas	82	ward	157	police	232	chicago
8	mart	83	discount	158	executive	233	area
9	she	84	holiday	159	.....	234	discount_store
10	k	85	crime	160	make	235	call
11	department_store	86	spending	161	c-includes	236	level
12	woolworth	87	.....	162	i.	237	field
13	hawley	88	report	163	short	238	8%
14	her	89	merchant	164	u.s.	239	car
15	season	90	gap	165	small	240	a-'87
16	merchandise	91	campeau	166	thanksgiving	241	ralph
17	hudson	92	ikea	167	exact	242	fiscal_year
18	chain	93	stemberg	168	add	243	.....
19	wal-mart	94	zinn	169	expectation	244	note
20	mr.	95	specialty_store	170	crowd	245	troy
21	penney	96	f.w.	171	dec.	246	take
22	consumer	97	inventory	172	.....	247	feiner
23	retail	98	gift	173	market	248	hallmark
24	year	99	1986	174	low	249	bill
25	catalog	100	gibson	175	month	250	drexler
26	dayton	101	saks	176	go	251	goldfeder
27	same-store	102	state	177	antar	252	slow
28	federated	103	increase	178	c-may	253	fiscal
29	shopper	104	expect	179	d-dayton	254	horchow
30	dept.	105	allied	180	d-excludes	255	law
31	carter	106	see	181	fed.	256	b.
32	apparel	107	sweater	182	markdowns	257	department
33	eddie	108	mail-order	183	high	258	1987
34	analyst	109	a-total	184	where	259	think
35	shopping	110	b-based	185	end	260	even
36	woman	111	lifetime	186	associated	261	oct.
37	crazy	112	disney	187	start	262	york-based
38	good	113	ames	188	.....	263	bass
39	sell	114	product	189	above	264	.....
40	gain	115	blue-light	190	designer	265	5%
41	specialty	116	clothes	191	.....	266	benetton
42	retailing	117	crash	192	a_lot	267	million
43	open	118	employee	193	fall	268	weather
44	j.c.	119	roebuck	194	strategy	269	saturday
45	customer	120	new_york	195	burton	270	keep
46	fiorucci	121	best	196	industry	271	trend
47	big	122	marketing	197	growth	272	promotion
48	koslow	123	1986.	198	period	273	special
49	limited	124	nation	199	'86	274	december
50	macke	125	ms.	200	profit	275	general
51	bookstop	126	supermarket	201	corp.	276	management
52	clothing	127	merchandising	202	major	277	five
53	new	128	fashion	203	charge	278	warehouse
54	rise	129	chairman	204	p.m	279	line
55	day	130	.....	205	shirt	280	friday
56	toy	131	chaumet	206	cataloger	281	b-%
57	mercantile	132	tax	207	look	282	unit
58	price	133	wear	208	cinema	283	early
59	item	134	.....	209	october	284	recent
60	bradlees	135	barnard	210	guarantee	285	sales
61	skirt	136	spend	211	chief_executive_officer	286	sir
62	division	137	president	212	schweich	287	say
63	.....	138	dalton	213	cost	288	minneapolis-based
64	hale	139	man	214	steidtmann	289	clerk
65	business	140	manager	215	return	290	large
66	strong	141	november	216	.....	291	weak
67	company	142	chicago-based	217	executive_officer	292	johnson
68	weekend	143	earnings	218	quarter	293	stock_market
69	herrlinger	144	blue	219	move	294	.....
70	exclude	145	inc.	220	run	295	never
71	card	146	sales_tax	221	wanamaker	296	bentonville
72	magnin	147	continue	222	suit	297	late
73	shop	148	miss	223	more_than	298	last-minute
74	result	149	dry	224	operation	299	goldstein
75	montgomery	150	.....	225	post	300	outlet

Table B.26: **RET** Topic signature in set **PH**.

SCR Topic signature (PH)							
1	firm	76	government_security	151	john	226	1985
2	mr.	77	bevill	152	go	227	craven
3	security	78	she	153	prison	228	time
4	kidder	79	investment_banker	154	robert	229	aron
5	broker	80	financial	155	get	230	think
6	merrill	81	department	156	board	231	trial
7	sec	82	fraud	157	retail	232	robinson
8	lynch	83	senior	158	activity	233	ask
9	salomon	84	branch	159	daiwa	234	stock_exchange
10	shearson	85	hardiman	160	position	235	yamaichi
11	hutton	86	tokyo	161	stanley	236	rooney
12	wall_street	87	brokerage_firm	162	several	237	issue
13	morgan	88	bank	163	denunzio	238	four
14	client	89	operation	164	member	239	work
15	former	90	judge	165	has_been	240	hire
16	nomura	91	inc.	166	convict	241	come
17	kane	92	blinder	167	deny	242	prosecutor
18	new_york	93	name	168	takeover	243	underwriting
19	official	94	sloate	169	director	244	april
20	investment	95	jersey	170	million	245	resign
21	market	96	federal	171	equity	246	decline
22	office	97	manager	172	boesky	247	bonus
23	charge	98	unit	173	pay	248	action
24	business	99	guilty	174	spokesman	249	diianni
25	securities_firm	100	allege	175	move	250	fomon
26	banking	101	vice_president	176	fed	251	wrongdoing
27	jefferies	102	managing_director	177	securities_market	252	japan
28	ge	103	staff	178	primary	253	general
29	case	104	finance	179	expect	254	question
30	customer	105	grenfell	180	risk	255	deal
31	rothschild	106	bond	181	admit	256	system
32	loss	107	plead	182	day	257	where
33	brother	108	accord	183	a.	258	volume
34	investor	109	association	184	major	259	vice_chairman
35	painewebber	110	make	185	regulation	260	dean
36	exchange	111	reamer	186	top	261	fine
37	trading	112	investigation	187	indictment	262	order
38	nasd	113	court	188	mortgage-backed	263	recently
39	elliott	114	peabody	189	report	264	gutfreund
40	executive	115	schulman	190	leave	265	schwab
41	stock	116	lehman	191	change	266	the_two
42	trader	117	source	192	buy	267	would_be
43	year	118	corporate	193	merger	268	claim
44	brokerage	119	large	194	witter	269	l.
45	dealer	120	count	195	yesterday	270	job
46	goldman	121	fund	196	billion	271	recent
47	london	122	tell	197	atkins	272	offering
48	arbitration	123	more_than	198	pace	273	feel
49	karger	124	man	199	nagle	274	say
50	ranieri	125	add	200	public	275	plan
51	co.	126	her	201	international	276	manhattan
52	big	127	crash	202	allegedly	277	salesman
53	capital	128	scheme	203	violation	278	municipal_bond
54	eder	129	involve	204	e.f.	279	option
55	rubin	130	kurokawa	205	individual	280	beim
56	commission	131	banker	206	give	281	illegal
57	account	132	cathcart	207	chief_executive	282	five
58	head	133	transaction	208	industry	283	know
59	mortgage	134	take	209	lawyer	284	see
60	securities_industry	135	drexel	210	discount	285	problem
61	new	136	municipal	211	president	286	price
62	employee	137	layoff	212	stock_market	287	compensation
63	chairman	138	become	213	unauthorized	288	control
64	group	139	partner	214	messrs	289	stockbroker
65	u.s.	140	coniston	215	year_old	290	chicago
66	sentence	141	invest	216	settlement	291	nikko
67	money	142	month	217	week	292	way
68	attorney	143	company	218	accuse	293	asset
69	government	144	suit	219	even	294	prudential-bache
70	trade	145	area	220	responsibility	295	bresler
71	management	146	comment	221	guinness	296	whether
72	people	147	future	222	law	297	review
73	japanese	148	insider-trading	223	institutional	298	district
74	rule	149	sell	224	british	299	information
75	boston	150	analyst	225	own	300	cutback

Table B.27: **SCR** Topic signature in set **PH**.

STK Topic signature (PH)							
1	stock	76	b	151	large	226	every
2	offering	77	acquisition	152	14	227	firm
3	share	78	trust	153	make	228	more_than
4	underwriter	79	average	154	grant	229	even
5	proceeds	80	common_stock	155	small	230	head
6	81282	81	preferred_shares	156	invest	231	goldome
7	8787	82	symbol	157	incentive_option	232	high
8	common_shares	83	p-e	158	trader	233	interest
9	sell	84	additional	159	share_in	234	34
10	price	85	new	160	short-term	235	americus
11	investor	86	chart	161	salomon	236	industry
12	+	87	bank	162	equity	237	cleveland-cliffs
13	market	88	toronto	163	get	238	asset
14	change	89	new_york_stock_exchange	164	rise	239	month
15	warrant	90	sale	165	income	240	tend
16	common	91	concern	166	purpose	241	savings_bank
17	trading	92	percent	167	finance	242	chemical
18	issue	93	friday	168	retire	243	lehman
19	company	94	international	169	service	244	wood
20	holder	95	total	170	computer	245	nonqualified
21	net	96	composite	171	hold	246	drop
22	offer	97	analyst	172	product	247	client
23	lead	98	reduce	173	$25	248	five
24	option	99	jones	174	canadian_dollar	249	bp
25	use	100	financial	175	convert	250	cover
26	inc.	101	date	176	mca	251	$50
27	million	102	investment	177	otc_stock	252	resource
28	prechter	103	brother	178	holding_company	253	bargain
29	%	104	maker	179	cash	254	capitalization
30	price*	105	12	180	lynch	255	algoma
31	yesterday	106	general	181	own	256	gw
32	dividend	107	working_capital	182	dutch	257	system
33	class	108	gas	183	stock_option	258	redeem
34	debt	109	pay	184	manage	259	eurotunnel
35	earnings**	110	morgan	185	thereafter	260	31
36	ratio**	111	limited	186	entitle	261	drexel
37	close	112	purchase	187	tokyo	262	raise
38	tax	113	executive	188	money	263	july
39	preferred	114	nasdaq	189	board	264	four
40	buy	115	day	190	real_estate	265	81282*
41	otc	116	index	191	1987	266	target's
42	partnership	117	point	192	rest	267	market_value
43	employees#	118	gold	193	new_york	268	power
44	sales***	119	industrial	194	kong	269	wave
45	stock_exchange	120	earnings	195	say	270	propose
46	outstanding	121	value	196	expect	271	foreign
47	pe	122	year	197	30	272	electric
48	general_purpose	123	group	198	secondary	273	people
49	exercise	124	12.5	199	development	274	estate
50	right	125	u.s.	200	go	275	past
51	used_to	126	billion	201	occidental	276	american
52	co.	127	corporate	202	ratio	277	approve
53	canadian	128	exchange	203	1	278	current
54	mr.	129	merrill	204	crash	279	study
55	preferred_stock	130	utility	205	gundy	280	buyer
56	plan	131	dow	206	overallotments	281	38
57	series	132	base	207	exploration	282	burnham
58	rate	133	goldman	208	25	283	consist_of
59	ltd.	134	underwriting	209	she	284	as_many
60	canada	135	manager	210	$100	285	lambert
61	auction	136	fall	211	inc	286	initial
62	big	137	national	212	depositary	287	commission
63	stock_market	138	partner	213	week	288	find
64	capital	139	co	214	shearson	289	currently
65	cent	140	dollar	215	voting	290	early
66	corp.	141	exchangeable	216	energy	291	recent
67	oil	142	mining	217	portfolio	292	oct.
68	gain	143	boston	218	bic	293	loss
69	security	144	low	219	increase	294	flow
70	public	145	sachs	220	unchanged	295	gulf
71	begin	146	trade	221	annual	296	possible
72	unit	147	file	222	set	297	alex
73	shareholder	148	cumulative	223	elder	298	record
74	over-the-counter	149	stanley	224	institutional	299	technology
75	convertible	150	fund	225	director	300	march

Table B.28: **STK** Topic signature in set **PH**.

TEL Topic signature (PH)							
1	at&t	76	proposal	151	pay	226	way
2	network	77	get	152	court	227	work
3	fcc	78	operator	153	advertiser	228	entertainment
4	cbs	79	analyst	154	continue	229	force
5	mr.	80	channel	155	satellite	230	offer
6	cable	81	provide	156	require	231	sunday
7	bell	82	cut	157	king	232	talk
8	long-distance	83	profit	158	washington	233	become
9	telephone	84	new_york	159	take	234	week
10	telecommunication	85	president	160	prime-time	235	action
11	mci	86	national	161	monopoly	236	current
12	doctrine	87	air	162	home	237	more_than
13	news	88	co.	163	voice	238	keep
14	station	89	consumer	164	inc.	239	wireless
15	service	90	market	165	operation	240	set
16	nbc	91	state	166	right	241	computer
17	sprint	92	employee	167	need	242	name
18	turner	93	deregulation	168	ms.	243	department
19	communication	94	make	169	move	244	find
20	local	95	unit	170	technology	245	concern
21	rate	96	change	171	cap	246	british
22	broadcast	97	revenue	172	even	247	seek
23	television	98	viewer	173	go	248	1984
24	show	99	west	174	northern	249	law
25	broadcasting	100	rule	175	congress	250	issue
26	programming	101	regulator	176	job	251	run
27	abc	102	price	177	own	252	her
28	contract	103	people	178	u.s.	253	manager
29	phone	104	whether	179	television_station	254	joint_venture
30	customer	105	allow	180	replace	255	drop
31	tisch	106	nynex	181	large	256	vote
32	regional	107	give	182	see	257	leave
33	phone_company	108	video	183	schedule	258	siemens
34	gsa	109	taft	184	reduction	259	several
35	greene	110	she	185	pacific	260	marietta
36	itt	111	propose	186	telesis	261	movie
37	company	112	audience	187	public	262	such_as
38	broadcaster	113	vice_president	188	bid	263	long_distance
39	business	114	american	189	operate	264	ruling
40	regulation	115	use	190	argue	265	phone_service
41	new	116	sport	191	spokesman	266	line
42	nfl	117	team	192	reduce	267	sell
43	judge	118	chairman	193	breakup	268	say
44	call	119	billion	194	free	269	telephone_service
45	rating	120	information	195	earnings	270	radio_station
46	executive	121	season	196	former	271	commissioner
47	fairness	122	advertising	197	currently	272	hope
48	charge	123	competitor	198	night	273	end
49	us	124	hbo	199	brook	274	medium
50	fowler	125	expect	200	capital	275	position
51	u	126	venture	201	decree	276	city
52	cost	127	official	202	win	277	toy
53	agency	128	government	203	hour	278	decide
54	espn	129	access	204	million	279	for_example
55	s	130	industry	205	begin	280	carry
56	tv	131	group	206	fall	281	bidding
57	year	132	cge	207	nielsen	282	accord
58	switch	133	big	208	order	283	believe
59	telecom	134	increase	209	viacom	284	story
60	union	135	source	210	director	285	recent
61	radio	136	month	211	fee	286	estimate
62	telephone_company	137	general	212	think	287	international
63	federal	138	patrick	213	biondi	288	araskog
64	fox	139	equipment	214	report	289	help
65	telegraph	140	division	215	consider	290	association
66	gte	141	decision	216	tell	291	martin
67	program	142	world	217	rather	292	case
68	system	143	justice_department	218	major	293	ask
69	competition	144	bill	219	number	294	result
70	game	145	mcgowan	220	regulatory	295	worker
71	commission	146	would_be	221	agreement	296	spelling
72	plan	147	has_been	222	corp.	297	datum
73	time	148	citiesabc	223	strike	298	reach
74	subscriber	149	rep.	224	problem	299	put
75	want	150	try	225	add	300	effort

Table B.29: **TEL** Topic signature in set **PH**.

TNM Topic signature (PH)							
1	share	76	bank	151	more_than	226	inc
2	filing	77	oil	152	accord	227	private
3	stake	78	industry	153	york-based	228	has_been
4	group	79	approval	154	computer	229	loan
5	acquire	80	propose	155	general	230	mining
6	company	81	operate	156	holly	231	finance
7	purchase	82	chairman	157	resource	232	june
8	million	83	reach	158	condition	233	gain
9	common_shares	84	division	159	plant	234	leveraged_buyout
10	sale	85	holding_company	160	financing	235	the_two
11	inc.	86	subsidiary	161	new_world	236	unchanged
12	acquisition	87	director	162	share_in	237	vote
13	transaction	88	new_york_stock_exchange	163	give	238	employee
14	stock	89	expect	164	canada	239	n.y.
15	hold	90	be_subject	165	u.s.	240	possible
16	unit	91	subject_to	166	in_principle	241	chicago
17	shareholder	92	revenue	167	billion	242	los_angeles
18	sell	93	announce	168	nortek	243	gate
19	mr.	94	industrial	169	toronto	244	week
20	investment	95	international	170	home	245	block
21	agreement	96	president	171	31	246	technology
22	corp.	97	say	172	indicate	247	allegheny
23	disclose	98	receive	173	federal	248	current
24	buy	99	decline	174	12.5	249	march
25	offer	100	definitive	175	remain	250	communication
26	exchange	101	energy	176	manpower	251	reject
27	outstanding	102	market	177	western	252	heritage
28	asset	103	equipment	178	restaurant	253	manufacturing
29	concern	104	report	179	annual	254	voting
30	security	105	option	180	increase	255	bell
31	trading	106	gas	181	pacific	256	all_of
32	agree	107	1986	182	purpose	257	buy-out
33	close	108	price	183	insurance	258	issue
34	term	109	friday	184	reliance	259	lead
35	management	110	system	185	restructuring	260	engineering
36	complete	111	washington	186	dec.	261	kincaid
37	investor	112	new	187	30	262	time
38	board	113	common	188	texas	263	preferred_stock
39	own	114	certain	189	letter_of_intent	264	swap
40	maker	115	partner	190	fund	265	quarter
41	commission	116	debt	191	continue	266	reduce
42	product	117	chemical	192	plc	267	distributor
43	business	118	part	193	sec	268	entertainment
44	cent	119	american	194	approve	269	operator
45	co.	120	capital	195	investment_firm	270	convertible
46	previously	121	equity	196	form	271	bidder
47	yesterday	122	american_stock_exchange	197	ohio	272	houston
48	control	123	total	198	n.j.	273	union
49	closely_held	124	calif.	199	provide	274	warrant
50	holding	125	official	200	corporate	275	oct.
51	stock_exchange	126	loss	201	end	276	member
52	national	127	additional	202	large	277	medical
53	bid	128	valued_at	203	currently	278	net
54	tender_offer	129	class	204	delaware	279	city
55	plan	130	state	205	retain	280	banking
56	proposal	131	kenner	206	intend	281	facility
57	service	132	firm	207	add	282	united
58	cash	133	raise	208	name	283	april
59	financial	134	cenergy	209	would_be	284	datum
60	spokesman	135	executive_officer	210	25	285	family
61	partnership	136	chief_executive_officer	211	specialty	286	america
62	composite	137	right	212	combine	287	inn
63	make	138	year	213	property	288	vice_president
64	takeover	139	sign	214	buyer	289	corp
65	base	140	real_estate	215	arrow	290	use
66	operation	141	consider	216	trust	291	stockholder
67	ltd.	142	comment_on	217	law	292	calny
68	interest	143	pay	218	year_end	293	several
69	merger	144	meeting	219	associate	294	florida
70	over-the-counter	145	store	220	shamrock	295	furniture
71	comment	146	tender	221	saving	296	take
72	new_york	147	month	222	hotel	297	$10
73	holder	148	food	223	calif.-based	298	dallas
74	value	149	common_stock	224	conn.	299	court
75	seek	150	canadian	225	development	300	50%

Table B.30: **TNM** Topic signature in set **PH**.

TRA Topic signature (PH)							
1	railroad	76	eurotunnel	151	major	226	overhaul
2	highway	77	use	152	support	227	measure
3	rail	78	mile	153	fisherman	228	sanford
4	union	79	mph	154	large	229	us
5	conrail	80	administration	155	seaman	230	john
6	ship	81	government	156	week	231	test
7	bill	82	vessel	157	company	232	small
8	train	83	hour	158	local	233	65
9	speed_limit	84	require	159	business	234	face
10	amtrak	85	engineer	160	high	235	agreement
11	transit	86	captive	161	agency	236	amendment
12	shipper	87	bus	162	commerce	237	demand
13	vote	88	bahrain	163	1980	238	want
14	labor	89	pass	164	water	239	involve
15	driver	90	wage	165	snow	240	believe
16	strike	91	republican	166	yesterday	241	toyo
17	mr.	92	rate	167	vehicle	242	long
18	truck	93	trucking	168	flowton	243	begin
19	transportation	94	change	169	yard	244	fishing
20	state	95	day	170	get	245	people
21	railway	96	washington	171	spokesman	246	public
22	safety	97	new	172	passenger	247	action
23	navy	98	dole	173	see	248	money
24	teamster	99	sea-land	174	fail	249	coast
25	president	100	crane	175	take	250	subsidy
26	worker	101	fund	176	propose	251	help
27	senate	102	country	177	funding	252	effort
28	csx	103	otsego	178	tanker	253	negotiation
29	congress	104	mine	179	route	254	several
30	sen.	105	construction	180	longshoreman	255	teamsters'
31	contract	106	issue	181	act	256	find
32	o&w	107	brakeman	182	approve	257	continue
33	veto	108	build	183	menka	258	freight_train
34	ila	109	national	184	signal	259	device
35	federal	110	program	185	service	260	chairman
36	shipyard	111	car	186	current	261	add
37	interstate	112	authority	187	run	262	final
38	tunnel	113	group	188	association	263	northern
39	trucker	114	new_jersey	189	r.	264	burlington
40	carrier	115	dispute	190	try	265	keep
41	line	116	member	191	operation	266	walkout
42	freight	117	system	192	limit	267	schmidt
43	afl-cio	118	icc	193	new_york	268	problem
44	legislation	119	make	194	decide	269	repair
45	canadian	120	more_than	195	albee	270	brake
46	project	121	causeway	196	conference	271	drug
47	accident	122	haley	197	crewmen	272	negotiate
48	billion	123	senator	198	whether	273	pay
49	industry	124	private	199	northeast	274	analyst
50	crew	125	represent	200	month	275	has_been
51	house	126	rural	201	commuter	276	coastal
52	law	127	aboard	202	truck_driver	277	must
53	rule	128	55	203	navy_yard	278	investigator
54	road	129	traffic	204	own	279	icebreaker
55	official	130	way	205	budget	280	kwong
56	track	131	even	206	foreign	281	re-flagging
57	override	132	need	207	today	282	charge
58	work	133	chirac	208	shipbuilding	283	picketing
59	cost	134	cp	209	base	284	haul
60	job	135	order	210	nation	285	political
61	employee	136	federation	211	n.y.	286	iranian
62	u.s.	137	southern	212	give	287	commission
63	rich	138	allow	213	gleason	288	five
64	guilford	139	u-haul	214	tung	289	number
65	mass	140	corridor	215	deregulate	290	provision
66	year	141	plan	216	become	291	stop
67	drive	142	hong	217	put	292	low
68	port	143	tax	218	sailor	293	submarine
69	subway	144	shipping	219	d.	294	would_be
70	pacific	145	operate	220	move	295	occur
71	reagan	146	democrat	221	heavy	296	win
72	seafarer	147	where	222	life	297	shipowner
73	locomotive	148	kong	223	leader	298	speed-limit
74	norfolk	149	passenger_train	224	committee	299	spending
75	department	150	raise	225	provide	300	justice

Table B.31: **TRA** Topic signature in set **PH**.

UTI Topic signature (PH)							
1	utility	76	quebec	151	change	226	begin
2	power	77	co-op	152	annual	227	represent
3	rate	78	generating	153	price	228	department
4	electric	79	study	154	states'	229	province
5	commission	80	saving	155	money	230	region
6	electricity	81	construction	156	boost	231	effective
7	public_utility	82	revenue	157	1987	232	make
8	hydro-quebec	83	residential	158	claim	233	average
9	edison	84	osage	159	provide	234	several
10	energy	85	propose	160	tax	235	florida
11	plant	86	capacity	161	equity	236	ruling
12	gas	87	approval	162	efficient	237	15.2%
13	public_service	88	would_be	163	operation	238	waste
14	gulf_states	89	use	164	1	239	jeffries
15	customer	90	agency	165	electric_bill	240	say
16	increase	91	birdsall	166	demand	241	give
17	cogeneration	92	megawatt	167	el	242	heating
18	request	93	group	168	additional	243	force
19	state	94	decision	169	buy	244	recommend
20	regulator	95	bergland	170	own	245	distribution
21	utah	96	nreca	171	high	246	emery
22	co.	97	file	172	recover	247	analyst
23	cost	98	indiana	173	midlantic	248	kansas
24	refund	99	tva	174	get	249	source
25	texas	100	commissioner	175	1986	250	critic
26	project	101	natural_gas	176	law	251	avoid
27	california	102	lighting	177	result	252	allege
28	p	103	grant	178	d.	253	judge
29	georgia	104	generation	179	investor-owned	254	decrease
30	maine	105	clarke	180	$60	255	whether
31	contract	106	gulf	181	new_england	256	gorge
32	pg&e	107	seek	182	penhallegon	257	non-utility
33	power_plant	108	increase_in	183	petrosar	258	e&g
34	million	109	charge	184	pennsylvania	259	silt
35	nuclear	110	commonwealth	185	seabrook	260	delay
36	reduction	111	act	186	u.s.	261	air_conditioner
37	kilowatt	112	reduce	187	new_york	262	$250
38	ratepayer	113	consolidated	188	large	263	deny
39	louisiana	114	nevada	189	public	264	hearing
40	central	115	save	190	ask	265	big
41	conservation	116	coal	191	1988	266	financing
42	regulatory	117	municipal	192	produce	267	limited
43	consumer	118	staff	193	earnings	268	ontario
44	dam	119	nuclear_power	194	bidding	269	technology
45	build	120	river	195	g&e	270	industry
46	return	121	require	196	gas_turbine	271	trustee
47	purpa	122	suit	197	help	272	people
48	pacific	123	receive	198	regulate	273	filing
49	mohawk	124	supply	199	authorize	274	bankruptcy-court
50	light	125	steam	200	proposal	275	gov.
51	spokesman	126	appliance	201	become	276	francisville
52	emergency	127	city	202	industrial	277	holder
53	mr.	128	new_york_state	203	commercial	278	ohio
54	year	129	agreement	204	1989	279	new_hampshire
55	order	130	independent	205	problem	280	future
56	low	131	building	206	has_been	281	consider
57	niagara	132	cut	207	annually	282	1978
58	new_mexico	133	allow	208	puget	283	decide
59	new	134	wisconsin	209	try	284	national
60	efficiency	135	holding_company	210	bend	285	creditor
61	fuel	136	hydroelectric	211	northern	286	purchase
62	generate	137	paso	212	rate-increase	287	her
63	unit	138	lilco	213	action	288	take
64	bill	139	pse	214	partnership	289	county
65	vermont	140	rural	215	mission	290	credit
66	plan	141	yangtze	216	call	291	period
67	need	142	agree	217	vote	292	transmission_line
68	electric_power	143	expect	218	town	293	february
69	approve	144	panel	219	official	294	spending
70	federal	145	oil	220	pay	295	dividend
71	southern	146	generator	221	chicago	296	san_francisco
72	billion	147	amount	222	business	297	environmental
73	company	148	add	223	puget_sound	298	1990
74	rea	149	estimate	224	utilities'	299	at_least
75	facility	150	program	225	base	300	appeal

Table B.32: **UTI** Topic signature in set **PH**.

Appendix C

Recall and Precision Scores for Various Tests

TEST CODE:

`7:` 1987 WSJ texts, i.e., training set

`8:` 1988 WSJ texts, i.e., test set

`WD:` texts without morphological transformation and word grouping

`TR:` texts with morphological transformation but not word grouping

`PH:` texts with morphological transformation and word grouping (phrases)

`1a:` using cosine similarity measure and first level topic signatures

`1b:` using cosine similarity measure and second level topic signatures

`N:` using *idf* term weighting with # of documents per topic normalization

C.1 Results Using Only First Level Topic Signatures

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	716	38	27	0.964	0.950
aro	336	59	37	0.901	0.851
aut	611	49	63	0.907	0.926
bbk	818	310	130	0.863	0.725
bcy	137	50	24	0.851	0.733
bnk	398	267	61	0.867	0.598
bon	722	37	193	0.789	0.951
ceo	391	227	36	0.916	0.633
cmd	98	16	17	0.852	0.860
div	585	30	71	0.892	0.951
eco	239	110	52	0.821	0.685
edp	263	71	37	0.877	0.787
ele	106	22	33	0.763	0.828
env	120	32	19	0.863	0.789
ern	613	224	87	0.876	0.732
fab	154	83	55	0.737	0.650
fin	240	121	55	0.814	0.665
lng	99	178	3	0.971	0.357
min	206	67	19	0.916	0.755
mkt	98	38	31	0.760	0.721
mon	730	112	103	0.876	0.867
pet	147	53	25	0.855	0.735
pha	432	53	56	0.885	0.891
pub	190	76	20	0.905	0.714
rel	161	106	42	0.793	0.603
ret	93	107	14	0.869	0.465
scr	225	101	41	0.846	0.690
stk	341	67	152	0.692	0.836
tel	375	75	62	0.858	0.833
tnm	3225	142	1425	0.694	0.958
tra	100	38	37	0.730	0.725
uti	128	81	13	0.908	0.612
Average				0.847	0.752

Table C.1: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ7; WD 1a.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	715	40	28	0.962	0.947
aro	340	59	33	0.912	0.852
aut	610	48	64	0.905	0.927
bbk	806	329	142	0.850	0.710
bcy	140	52	21	0.870	0.729
bnk	399	300	60	0.869	0.571
bon	677	32	238	0.740	0.955
ceo	381	217	46	0.892	0.637
cmd	97	18	18	0.843	0.843
div	583	41	73	0.889	0.934
eco	240	121	51	0.825	0.665
edp	264	97	36	0.880	0.731
ele	107	31	32	0.770	0.775
env	120	29	19	0.863	0.805
ern	588	233	112	0.840	0.716
fab	153	97	56	0.732	0.612
fin	237	144	58	0.803	0.622
lng	99	167	3	0.971	0.372
min	207	69	18	0.920	0.750
mkt	97	40	32	0.752	0.708
mon	737	105	96	0.885	0.875
pet	147	53	25	0.855	0.735
pha	436	53	52	0.893	0.892
pub	190	82	20	0.905	0.699
rel	165	120	38	0.813	0.579
ret	93	125	14	0.869	0.427
scr	225	96	41	0.846	0.701
stk	337	85	156	0.684	0.799
tel	371	65	66	0.849	0.851
tnm	3116	147	1534	0.670	0.955
tra	101	43	36	0.737	0.701
uti	130	91	11	0.922	0.588
Average				0.844	0.739

Table C.2: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ7; TR 1a.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	717	46	26	0.965	0.940
aro	337	58	36	0.903	0.853
aut	609	47	65	0.904	0.928
bbk	818	283	130	0.863	0.743
bcy	140	54	21	0.870	0.722
bnk	385	197	74	0.839	0.662
bon	674	33	241	0.737	0.953
ceo	385	221	42	0.902	0.635
cmd	97	22	18	0.843	0.815
div	580	36	76	0.884	0.942
eco	243	119	48	0.835	0.671
edp	257	71	43	0.857	0.784
ele	109	34	30	0.784	0.762
env	120	32	19	0.863	0.789
ern	589	243	111	0.841	0.708
fab	156	91	53	0.746	0.632
fin	238	143	57	0.807	0.625
lng	100	126	2	0.980	0.442
min	205	64	20	0.911	0.762
mkt	96	37	33	0.744	0.722
mon	741	103	92	0.890	0.878
pet	139	42	33	0.808	0.768
pha	435	52	53	0.891	0.893
pub	190	81	20	0.905	0.701
rel	161	103	42	0.793	0.610
ret	93	127	14	0.869	0.423
scr	220	100	46	0.827	0.688
stk	351	96	142	0.712	0.785
tel	363	58	74	0.831	0.862
tnm	3336	171	1314	0.717	0.951
tra	100	44	37	0.730	0.694
uti	131	88	10	0.929	0.598
Average				0.843	0.748

Table C.3: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and average *recall* and *precision* of test set WSJ7; PH 1a.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	494	32	33	0.937	0.939
aro	381	78	57	0.870	0.830
aut	545	40	49	0.918	0.932
bbk	455	207	89	0.836	0.687
bcy	148	45	15	0.908	0.767
bnk	259	179	39	0.869	0.591
bon	420	43	103	0.803	0.907
ceo	239	204	40	0.857	0.540
cmd	92	22	18	0.836	0.807
div	357	19	76	0.824	0.949
eco	180	145	50	0.783	0.554
edp	295	73	70	0.808	0.802
ele	82	32	45	0.646	0.719
env	94	34	26	0.783	0.734
ern	648	186	89	0.879	0.777
fin	164	134	86	0.656	0.550
lng	68	137	7	0.907	0.332
min	138	54	20	0.873	0.719
mkt	155	61	72	0.683	0.718
mon	590	115	138	0.810	0.837
pet	100	26	30	0.769	0.794
pha	457	52	58	0.887	0.898
pub	184	90	36	0.836	0.672
rel	119	94	42	0.739	0.559
ret	51	89	13	0.797	0.364
scr	184	134	45	0.803	0.579
stk	110	51	94	0.539	0.683
tel	294	72	61	0.828	0.803
tnm	2707	121	1172	0.698	0.957
tra	79	35	46	0.632	0.693
uti	87	62	11	0.888	0.584
Average				0.803	0.719

Table C.4: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ8; WD 1a.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	496	35	31	0.941	0.934
aro	383	77	55	0.874	0.833
aut	553	41	41	0.931	0.931
bbk	449	201	95	0.825	0.691
bcy	151	43	12	0.926	0.778
bnk	260	203	38	0.872	0.562
bon	402	33	121	0.769	0.924
ceo	233	239	46	0.835	0.494
cmd	92	23	18	0.836	0.800
div	353	21	80	0.815	0.944
eco	183	144	47	0.796	0.560
edp	299	94	66	0.819	0.761
ele	81	35	46	0.638	0.698
env	95	42	25	0.792	0.693
ern	590	197	147	0.801	0.750
fin	163	137	87	0.652	0.543
lng	68	132	7	0.907	0.340
min	136	53	22	0.861	0.720
mkt	158	57	69	0.696	0.735
mon	588	116	140	0.808	0.835
pet	99	24	31	0.762	0.805
pha	457	62	58	0.887	0.881
pub	185	92	35	0.841	0.668
rel	123	96	38	0.764	0.562
ret	50	90	14	0.781	0.357
scr	182	143	47	0.795	0.560
stk	106	57	98	0.520	0.650
tel	299	74	56	0.842	0.802
tnm	2610	139	1269	0.673	0.949
tra	83	43	42	0.664	0.659
uti	92	66	6	0.939	0.582
Average				0.802	0.710

Table C.5: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ8; TR 1a.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	499	36	28	0.947	0.933
aro	374	69	64	0.854	0.844
aut	548	36	46	0.923	0.938
bbk	451	176	93	0.829	0.719
bcy	154	41	9	0.945	0.790
bnk	254	145	44	0.852	0.637
bon	419	48	104	0.801	0.897
ceo	235	237	44	0.842	0.498
cmd	92	24	18	0.836	0.793
div	355	26	78	0.820	0.932
eco	180	164	50	0.783	0.523
edp	286	68	79	0.784	0.808
ele	82	40	45	0.646	0.672
env	93	39	27	0.775	0.705
ern	582	205	155	0.790	0.740
fin	163	119	87	0.652	0.578
lng	68	107	7	0.907	0.389
min	133	52	25	0.842	0.719
mkt	154	58	73	0.678	0.726
mon	592	129	136	0.813	0.821
pet	89	22	41	0.685	0.802
pha	454	60	61	0.882	0.883
pub	189	88	31	0.859	0.682
rel	120	78	41	0.745	0.606
ret	52	95	12	0.813	0.354
scr	176	142	53	0.769	0.553
stk	112	65	92	0.549	0.633
tel	297	64	58	0.837	0.823
tnm	2782	150	1097	0.717	0.949
tra	77	41	48	0.616	0.653
uti	90	65	8	0.918	0.581
Average				0.797	0.716

Table C.6: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and average *recall* and *precision* of test set WSJ8; PH 1a.

C.2 Results Using First and Second Level Topic Signatures

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	716	38	27	0.964	0.950
aro	336	59	37	0.901	0.851
aut	611	49	63	0.907	0.926
bbk	820	271	128	0.865	0.752
bcy	137	46	24	0.851	0.749
bnk	385	246	74	0.839	0.610
bon	707	20	208	0.773	0.972
ceo	367	144	60	0.859	0.718
cmd	98	16	17	0.852	0.860
div	596	54	60	0.909	0.917
eco	240	117	51	0.825	0.672
edp	257	68	43	0.857	0.791
ele	106	31	33	0.763	0.774
env	120	32	19	0.863	0.789
ern	622	216	78	0.889	0.742
fab	157	84	52	0.751	0.651
fin	243	135	52	0.824	0.643
lng	99	103	3	0.971	0.490
min	206	67	19	0.916	0.755
mkt	95	31	34	0.736	0.754
mon	721	106	112	0.866	0.872
pet	147	56	25	0.855	0.724
pha	432	51	56	0.885	0.894
pub	190	76	20	0.905	0.714
rel	161	106	42	0.793	0.603
ret	93	107	14	0.869	0.465
scr	231	136	35	0.868	0.629
stk	330	64	163	0.669	0.838
tel	375	75	62	0.858	0.833
tnm	3391	194	1259	0.729	0.946
tra	100	38	37	0.730	0.725
uti	130	82	11	0.922	0.613
Average				0.846	0.757

Table C.7: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ7; WD 1b.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	716	45	27	0.964	0.941
aro	339	51	34	0.909	0.869
aut	610	48	64	0.905	0.927
bbk	804	320	144	0.848	0.715
bcy	140	50	21	0.870	0.737
bnk	386	256	73	0.841	0.601
bon	656	37	259	0.717	0.947
ceo	343	136	84	0.803	0.716
cmd	97	18	18	0.843	0.843
div	583	64	73	0.889	0.901
eco	245	125	46	0.842	0.662
edp	258	85	42	0.860	0.752
ele	109	46	30	0.784	0.703
env	120	29	19	0.863	0.805
ern	601	250	99	0.859	0.706
fab	158	98	51	0.756	0.617
fin	240	158	55	0.814	0.603
lng	99	102	3	0.971	0.493
min	207	69	18	0.920	0.750
mkt	96	34	33	0.744	0.738
mon	731	94	102	0.878	0.886
pet	147	56	25	0.855	0.724
pha	435	49	53	0.891	0.899
pub	193	79	17	0.919	0.710
rel	165	120	38	0.813	0.579
ret	93	125	14	0.869	0.427
scr	231	147	35	0.868	0.611
stk	317	72	176	0.643	0.815
tel	372	67	65	0.851	0.847
tnm	3256	192	1394	0.700	0.944
tra	101	43	36	0.737	0.701
uti	132	92	9	0.936	0.589
Average				0.843	0.742

Table C.8: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set WSJ7; TR 1b.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	717	46	26	0.965	0.940
aro	337	58	36	0.903	0.853
aut	609	47	65	0.904	0.928
bbk	843	289	105	0.889	0.745
bcy	140	50	21	0.870	0.737
bnk	371	161	88	0.808	0.697
bon	664	22	251	0.726	0.968
ceo	341	149	86	0.799	0.696
cmd	97	21	18	0.843	0.822
div	567	37	89	0.864	0.939
eco	248	123	43	0.852	0.668
edp	251	67	49	0.837	0.789
ele	110	43	29	0.791	0.719
env	120	32	19	0.863	0.789
ern	617	322	83	0.881	0.657
fab	158	91	51	0.756	0.635
fin	239	148	56	0.810	0.618
lng	100	84	2	0.980	0.543
min	205	64	20	0.911	0.762
mkt	95	31	34	0.736	0.754
mon	743	94	90	0.892	0.888
pet	139	45	33	0.808	0.755
pha	434	52	54	0.889	0.893
pub	192	80	18	0.914	0.706
rel	161	103	42	0.793	0.610
ret	93	103	14	0.869	0.474
scr	228	148	38	0.857	0.606
stk	345	113	148	0.700	0.753
tel	363	58	74	0.831	0.862
tnm	3383	186	1267	0.728	0.948
tra	100	41	37	0.730	0.709
uti	131	88	10	0.929	0.598
Average				0.842	0.752

Table C.9: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and average *recall* and *precision* of test set WSJ7; PH 1b.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	499	36	28	0.947	0.933
aro	374	69	64	0.854	0.844
aut	548	36	46	0.923	0.938
bbk	466	201	78	0.857	0.699
bcy	154	36	9	0.945	0.811
bnk	235	119	63	0.789	0.664
bon	404	40	119	0.772	0.910
ceo	215	167	64	0.771	0.563
cmd	92	23	18	0.836	0.800
div	339	20	94	0.783	0.944
eco	188	176	42	0.817	0.516
edp	279	70	86	0.764	0.799
ele	78	49	49	0.614	0.614
env	93	39	27	0.775	0.705
ern	625	250	112	0.848	0.714
fin	167	128	83	0.668	0.566
lng	68	67	7	0.907	0.504
min	133	52	25	0.842	0.719
mkt	148	52	79	0.652	0.740
mon	584	118	144	0.802	0.832
pet	90	23	40	0.692	0.796
pha	453	59	62	0.880	0.885
pub	189	88	31	0.859	0.682
rel	120	78	41	0.745	0.606
ret	52	70	12	0.813	0.426
scr	179	190	50	0.782	0.485
stk	106	89	98	0.520	0.544
tel	297	64	58	0.837	0.823
tnm	2816	161	1063	0.726	0.946
tra	77	36	48	0.616	0.681
uti	91	68	7	0.929	0.572
Average				0.792	0.718

Table C.10: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and average *recall* and *precision* of test set WSJ8; PH 1b.

C.3 Results Using First Topic Signatures and Normalized *idf*

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	721	46	22	0.970	0.940
aro	338	53	35	0.906	0.864
aut	608	46	66	0.902	0.930
bbk	832	289	116	0.878	0.742
bcy	139	45	22	0.863	0.755
bnk	386	178	73	0.841	0.684
bon	702	36	213	0.767	0.951
ceo	380	241	47	0.890	0.612
cmd	97	19	18	0.843	0.836
div	583	40	73	0.889	0.936
eco	243	112	48	0.835	0.685
edp	259	65	41	0.863	0.799
ele	109	33	30	0.784	0.768
env	121	25	18	0.871	0.829
ern	572	242	128	0.817	0.703
fab	158	83	51	0.756	0.656
fin	240	118	55	0.814	0.670
lng	99	108	3	0.971	0.478
min	207	56	18	0.920	0.787
mkt	95	30	34	0.736	0.760
mon	747	95	86	0.897	0.887
pet	141	42	31	0.820	0.770
pha	436	49	52	0.893	0.899
pub	191	69	19	0.910	0.735
rel	162	106	41	0.798	0.604
ret	92	98	15	0.860	0.484
scr	225	83	41	0.846	0.731
stk	365	107	128	0.740	0.773
tel	369	53	68	0.844	0.874
tnm	3443	162	1207	0.740	0.955
tra	102	43	35	0.745	0.703
uti	132	71	9	0.936	0.650
Average				0.848	0.764

Table C.11: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of training set (Wall Street Journal 1987) with phrases (PH) using normalized *idf*.

TOPIC	HIT	FAULT	MISS	RECALL	PRECISION
air	500	38	27	0.949	0.929
aro	380	63	58	0.868	0.858
aut	552	32	42	0.929	0.945
bbk	456	199	88	0.838	0.696
bcy	151	36	12	0.926	0.807
bnk	261	137	37	0.876	0.656
bon	422	50	101	0.807	0.894
ceo	236	239	43	0.846	0.497
cmd	92	18	18	0.836	0.836
div	356	25	77	0.822	0.934
eco	182	164	48	0.791	0.526
edp	286	58	79	0.784	0.831
ele	83	37	44	0.654	0.692
env	93	31	27	0.775	0.750
ern	581	207	156	0.788	0.737
fin	167	107	83	0.668	0.609
lng	68	93	7	0.907	0.422
min	133	50	25	0.842	0.727
mkt	153	54	74	0.674	0.739
mon	589	129	139	0.809	0.820
pet	91	19	39	0.700	0.827
pha	455	58	60	0.883	0.887
pub	187	76	33	0.850	0.711
rel	119	81	42	0.739	0.595
ret	51	70	13	0.797	0.421
scr	187	125	42	0.817	0.599
stk	119	73	85	0.583	0.620
tel	299	60	56	0.842	0.833
tnm	2853	145	1026	0.735	0.952
tra	78	43	47	0.624	0.645
uti	90	56	8	0.918	0.616
Average				0.802	0.729

Table C.12: *Hit*, *fault*, *miss*, *recall*, and *precision* scores for each topic and the average *recall* and *precision* of test set (Wall Street Journal 1988) with phrases (PH) using normalized *idf*.