# An Overview of Microsoft Academic Service (MAS) and Applications

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june (Paul) Hsu, Kuansan Wang
Microsoft Research, Redmond, WA 98052, USA
{arsinha, zhihosh, yangsong, haoma, darrine, paulhsu, kuansanw}@microsoft.com

## ABSTRACT

In this paper we describe a new release of a Web scale entity graph that serves as the backbone of Microsoft Academic Service (MAS), a major production effort with a broadened scope to the namesake vertical search engine that has been publicly available since 2008 as a research prototype. At the core of MAS is a heterogeneous entity graph comprised of six types of entities that model the scholarly activities: field of study, author, institution, paper, venue, and event. In addition to obtaining these entities from the publisher feeds as in the previous effort, we in this version include data mining results from the Web index and an in-house knowledge base from Bing, a major commercial search engine. As a result of the Bing integration, the new MAS graph sees significant increase in size, with fresh information streaming in automatically following their discoveries by the search engine. In addition, the rich entity relations included in the knowledge base provide additional signals to disambiguate and enrich the entities within and beyond the academic domain. The number of papers indexed by MAS, for instance, has grown from low tens of millions to 83 million while maintaining an above 95% accuracy based on test data sets derived from academic activities at Microsoft Research. Based on the data set, we demonstrate two scenarios in this work: a knowledge driven, highly interactive dialog that seamlessly combines reactive search and proactive suggestion experience, and a proactive heterogeneous entity recommendation.

## Keywords

Academic search; Recommender systems; Entity conflation

## 1. INTRODUCTION

Recent years have witnessed a paradigm shift in how the knowledge on the Web is made available to the users. The trend is highly visible in the evolution of the Web search engine. The traditional Web search outcomes often serve the users' need at best in a "hit-or-miss" fashion [4, 7]. A multi-year initiative in the industry, called Bing Dialog in Microsoft [11] and Knowledge Vault in Google [5], addresses this challenge by using statistical inferences to better organize the Web information and support much richer forms of interaction in recognizing and serving the user needs. In addition to reactively retrieving information and answering questions, the model proactively includes additional dialog acts, such as confirmation, disambiguation, refinement and digression. Coupled with statistical user intent inferences, these acts significantly expedite the process of serving users with the knowledge they need [14]. Our work aims at leveraging this model in addressing the information needs in areas where the sheer amount of information available through a multitude of channels has exceeded the human capacity in processing them. Although most search engines have provided advanced operators for users to compose elaborated queries to better filter out unwanted materials, their arcane syntax has relegated their usages to a negligible rate. A goal of the modern dialog approach to Web search is therefore to utilize advanced techniques to enable the search engines to communicate with users in natural language. Because the dialog inferences inevitably require the system to anticipate or predict the needs of the users, another emerging trend in the search engine evolution is to extend the prediction behaviors into system initiated notifications. The growing prevalence of mobile personal assistants serve as a natural vehicle to deliver proactive notifications, potentially preempting the needs of user initiated search for information [10].

In this paper, we present two applications in the area of academic publications to demonstrate the potentials of the emerging search paradigm. The first application, described in Section 3.1, illustrates a natural language powered interactive search experience. By leveraging the relationships among the entities in the academic domain, the natural language processor is able to harvest the syntactic and semantic cues for parsing and predicting user queries. The second application, described in Section 3.2, demonstrates how a recommendation system can take advantage of the relationships across different types of entities to offer heterogeneous suggestions. Noting that the statistical techniques underlying these two applications are by no means perfect, we further decide to make the data used by the two applications publicly available so that the community can jointly attack the challenging unsolved problems. The data set is an update to the corpus previously released for research purposes [2] and will be described in details in Section 2. The two applications also exemplify a commonly encountered scenario in which the results presented to the users should be properly ranked. The ranking algorithms and the measurements for determining the ranking order remain actively research topics. Given the surge in the count of academic entities and observable limitations of citation count based impact metrics, the problem of defining meaningful impact metrics of academic entities (e.g. papers, authors, conferences) is gaining substantial interest among the researchers [8, 3]. We hope this open corpus can contribute not only to advance information technologies for other innovative applications but also trigger a new horizon of


*WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742839.

research efforts towards defining new academic impact metrics as well.

## 2. DATA AGGREGATION AND ENTITY CONFLATION

In this work, we model the real-life academic communication activities as a heterogeneous graph consisting of six types of entities: field of study, author, institution (affiliation of author), paper, venue (journal and conference series, e.g. WWW, SIGIR, KDD etc.) and event (conference instances, e.g. WWW 2015). The relationship between these entities is shown in Fig. 1(a). These entity relationships are rather intuitive. (For instance, the fact that papers get published in journals/conferences justifies the edge between paper and venue nodes in the graph.) We describe how we obtain the raw data and organize them into the connected graph schema in the following subsections.

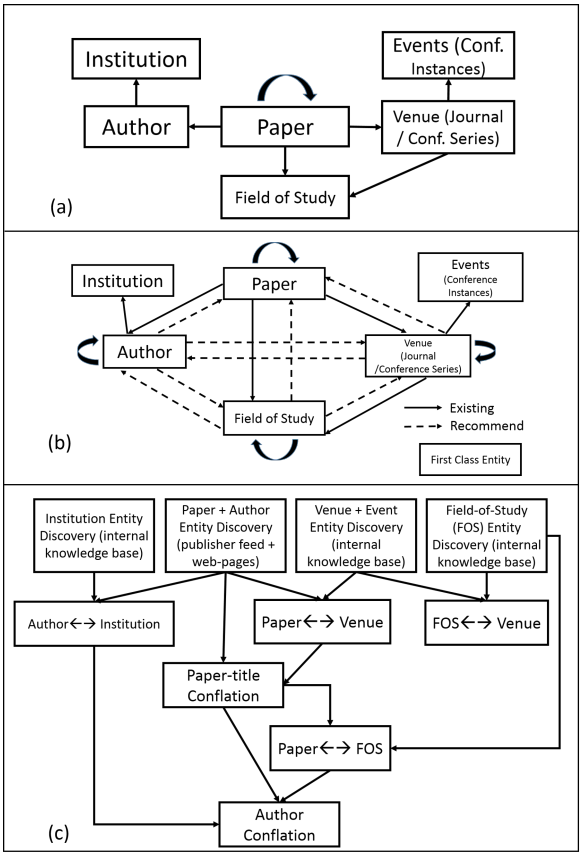


**Figure 1: (a) Academic Entity Relationship Graph, (b) Academic Entity Recommendation Graph, (c) Conflation flow of several entities.**

### 2.1 Paper and Author Entity Discovery

For paper and author entities, we collect data primarily from two types of sources: (1) feeds from publishers (e.g. ACM and IEEE), and (2) web-pages indexed by Bing. Although the majority of our data come from the indexed pages, often the quality of the feeds from publishers are significantly better. In the meantime, there exist a widely practiced convention for annotating the academic Web pages [1]. Pages that follow this recommendation are easier to parse compared to those that do not follow. Once the data is aggregated, the next challenge is to filter out the noise. Certain patterns (such as email address in place of author's name or author name beginning with a number) are easy to tackle while others (such as "Leslie Lamport Microsoft Research") are not. Once we strip off the obvious anomalies in author names and paper titles, we conflate venue, affiliation and keyword attributes of a discovered paper with our identified venue, affiliation and field of study entities (details of these entity types discovery are in Section 2.2 and Section 2.3). Next, we attempt to merge papers having exactly same titles and venues but different sources. For instance, often multiple web-pages mention the same paper but with incomplete information such as missing author full-names and affiliations. These sources, once merged, produce a far more comprehensive information about a paper entity. We refer to this step as title conflation. All the above mentioned information is also considered when we attempt to disambiguate author entities. Author name disambiguation is a well studied problem [6, 12] and we employ various best-effort algorithms. We achieve higher precision for authors with greater context information (e.g. affiliation, coauthors, year and venue of the publication etc.). In addition, the rich entity relations included in the in-house knowledge base, provide reliable signals to disambiguate and enrich the author entities.

### 2.2 Field of Study Entity Discovery

For field of study (FOS) entity, the data are already present in the in-house knowledge base, however, the majority (greater than 95%) are not marked with the "field of study" entity type. Our goal is to label the FOS entities in the in-house knowledge base when their type is missing. The approach is to use some "seed" FOS entities to discover more of them. Two sources are considered for seeding the discovery process: (1) the entities which are currently labelled as FOS type in the knowledge base; (2) the entities that are identified by name-matching the keyword attributes in paper entities. We then leverage the in-house knowledge base related entity relationship, which is calculated based on the entity contents, hyperlinks, and web-click signals, to identify the new FOS candidates. Our intuition is: when an entity is highly related to an existing FOS entity but is not labelled as any type, it is considered as a candidate. At last, we classify the candidates based on the ratio of the number of the same (FOS) type entities in its top N related entities to N, to obtain the final list. This process expands the size of the FOS entities twenty folds and our sample results shows above 98% accuracy of the identified new entities.

### 2.3 Venue, Event and Institution Entity Discovery

The conference-related entities are collected from a few semi-structured websites that are indexed by Bing. These websites serve as hubs of conference organizers posting their latest calls. Such semi-structured data are mostly conference instances (e.g. WWW 2015), although occasional notices for journal special issues are also observed. We conflate the conference instances (events) across different websites, recognize the conference series (venue), and generate the series and instances relationship using various signals obtained from the semi-structured data (e.g. acronym, full name, year, location, etc.). We conflate the category attribute of the conferences with the FOS entities identified in Section 2.2. The discovered academic conference instances and series are later ingested into the in-house knowledge base and conflated with other knowledge base entities with types external to academic domain, such as location, cities and countries. In addition, the journal and institutions are mostly aggregated from the in-house knowledge base.

| Entity name | Entity Count |
|---|---|
| Papers | $>$ 83 million |
| Authors | $>$ 20 million |
| Institutions | $>$ 770,000 |
| Journals | $>$ 22,000 |
| Conference series | $>$ 900 |
| Conference instances | $>$ 26,000 |
| Fields of study | $>$ 50,000 |

**Table 1: Counts of various entities in MAS corpus.**

Following the discovery of six academic domain entity types, these entities are joined to build the heterogeneous entity graph. The flow of conflation of several entity types is shown in Fig. 1(c). Note that the linkage between two entities are denoted by '$\leftrightarrow$' symbol in the diagram. In addition, Table 1 shows the approximate counts of entities that we have in the resulting heterogeneous entity graph based on the snapshot taken in mid January, 2015.

# 3. APPLICATIONS

In this paper, we describe two applications making use of the MAS entity graph. In Section 3.1, we describe the academic search engine based on the Bing Dialog model that can (1) serve constrained academic queries, and, (2) suggests other queries with same prefix. In Section 3.2, we present the academic entity recommendation application that has already been visible in Bing.

## 3.1 Academic Dialog Model

We have leveraged the Bing Dialog for serving academic search queries. In this paper we refer to this as Academic Dialog Model. This model serves as the engine behind a simple interactive website/portal that we have built to demonstrate the power of this model. The screen-shots shown in Fig. 2 are taken from this portal wrapping the Academic Dialog Model output. The data was modeled to showcase the Academic Paper Entity structure (e.g. Paper entity containing Title, Authors, Fields of Study, etc.), with views constructed to give a clean, easy to read format. The website is hosted in a public accessible cloud service[1].

In Fig. 2, we show several screen-shots of the portal serving queries of various degrees of complexity and flavor. The topmost description in largest font-size is the actual user query (e.g. "*fields of study about artificial intel*"). The suggested queries are below the actual query with a '**+**'-sign preceeding them (e.g. "**`+ fields of study about Artificial intelligence`**"). For each suggested query the respective results appear right below them. Moreover, note that the entities are color coded, e.g. author, affiliation, field-of-study and year are highlighted with yellow, red, green and cyan colors respectively.

In Fig. 2(a) the portal is suggesting several fields of study that are related to 'artificial intelligence' even when the actual query is incomplete. Also, in Fig. 2(b) the portal displays authors in a given field. The portal can also display papers in the intersection of two fields (refer to Fig. 2(c)). Besides, an user may be interested in papers authored by a given researcher while the person was at a particular organization (refer to Fig. 2(d)) or during a given range of years (refer to Fig. 2(e)). Lastly, Fig. 2(f) shows a highly constrained query where the format of the query is "*papers citing <author> before <year> about <field-of-study> appearing in <journal>*". We understand that this kind of query is not popular yet,

[1]http://isrc-academic01.cloudapp.net:8080. The demo video is available at https://vimeo.com/117688421.

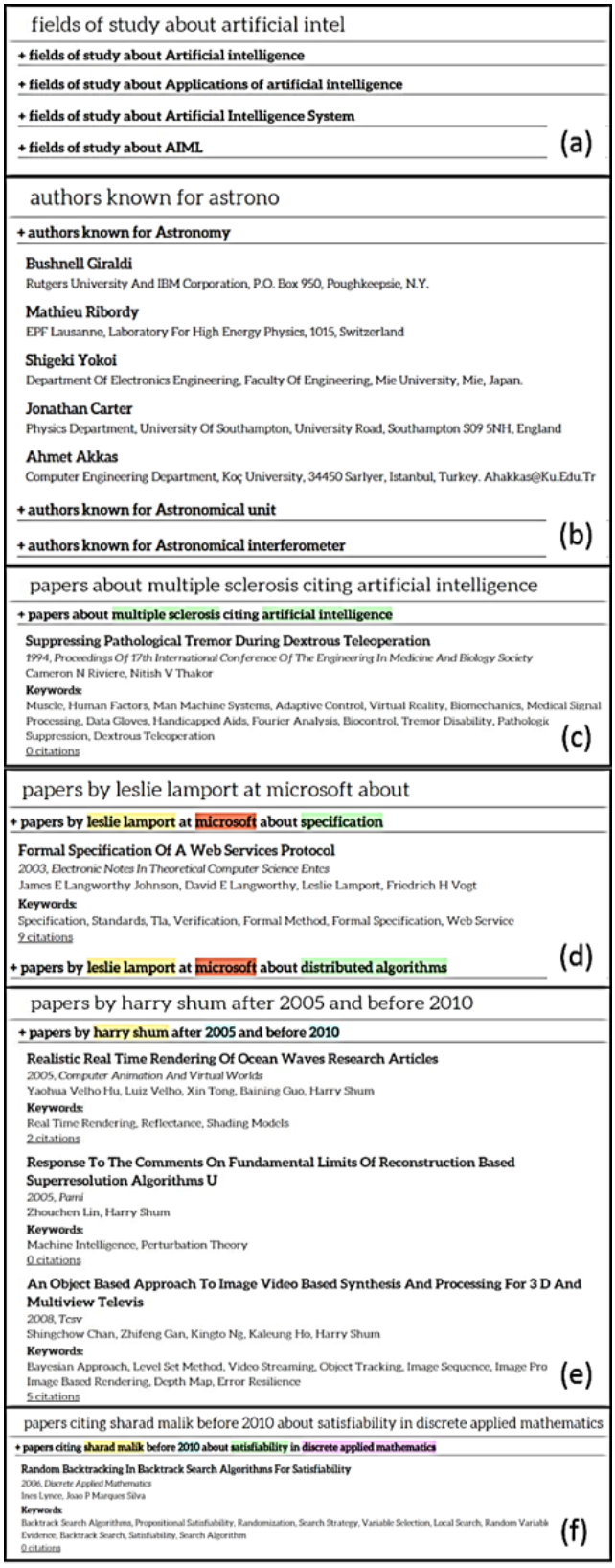


**Figure 2: Examples of academic search queries with varying degrees of complexity. The power of the underlying search engine is not limited to these patterns.**

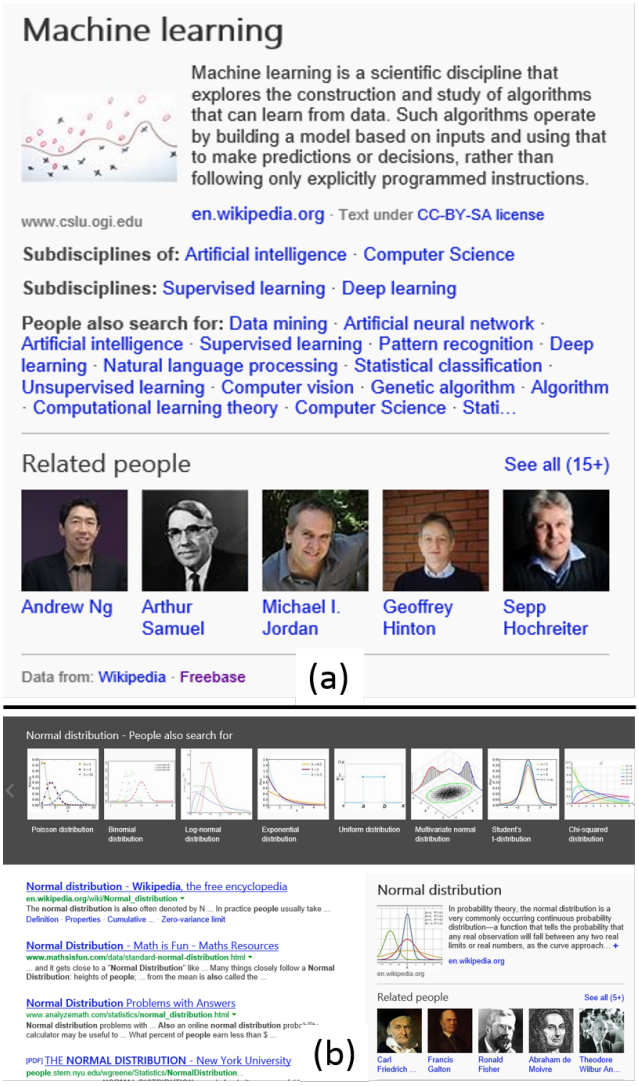


**Figure 3: (a) Bing Entity Pane experience with related field of study and related author recommendation, (b) Bing Carousel experience after expanding the "People also search for" section**

partly due to the incapability of the existing search engines to serve such complex queries. We believe our demo application, once integrated with Bing, will open new frontiers for advanced domain specific search queries. Also note that the power of the underlying engine is not limited to these queries only. These queries are just examples and the engine can handle even more constrained queries.

## 3.2 Academic Recommendation Model

Recommendation in the academic domain is a well researched topic [9, 13]. In this work, our goal is to be able to answer questions generated from a fully connected graph (see Fig. 1(b)) between six types of entities. For example, given a field of study, find out the most prominent authors, the most influential papers, the potential publishing venues and the upcoming events (conferences, workshops). Another example would be, given a venue, find out the scholars with most impact. The cited examples involve heterogeneous types of entities. However, similar problems within homogeneous types of entities can also be of interest, e.g. given a field of study (or conference), find out other relevant fields of study (or conferences).

As we integrate our service into Bing's infrastructure, one strong signal for recommendation is the co-click from the search engine logs. We leverage this result from the search logs to generate the candidate recommendation entities. The co-click signal results have good quality for high frequency query terms in academic domain such as "normal distribution" and "data mining". However, for other less well-known entities with much less query frequency, e.g. scholars who pioneered in a research domain, it is challenging to catch the relationship through sparse web-click signals. In order to discover such connections , we utilize other types of "co-occurrence" in the academic contents: e.g. co-authorship - authors collaborated on the same paper and co-venue - people published in the same sets of conferences/journals etc.. These content-based results generate good quality recommendation entities which complement the click-based results.

Fig. 3(a) shows the deployed Bing entity pane experience of a field of study ("normal distribution") with recommended authors (heterogeneous entity type) and recommended fields of study (homogeneous entity type). Fig. 3(b) shows the Bing carousel experience after expanding the "people also search for" section. This illustrates the rich experience that Bing offers to explore the academic entity relationship in a proactive fashion.

# 4. REFERENCES


[1] Google inclusion guidelines. In *http://www.google.com/intl/en/scholar/inclusion.html#indexing*.

[2] Microsoft academic data. In *http://datamarket.azure.com/dataset/mrc/microsoftacademic*, November 2013.

[3] A. Acharya, A. Verstak, H. Suzuki, S. Henderson, M. Iakhiaev, C. C. Lin, and N. Shetty. Rise of the rest: The growing impact of non-elite journals. *CoRR*, 2014.

[4] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *EDBT 2004 Workshops*, 2005.

[5] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohman, S. Sun, and W. Zhang. Knowledge vault: A web-based approach to probabilistic knowledge fusion. In *KDD 2014*.

[6] J. Huang, S. Ertekin, and C. L. Giles. Efficient name disambiguation for large-scale databases. In *PKDD*, 2006.

[7] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, Apr. 1998.

[8] V. Larivière, G. A. Lozano, and Y. Gingras. Are elite journals declining? *JASIST*, 65(4):649–655, 2014.

[9] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: effective citation recommendation by information network-based clustering. In *SIGKDD'14*. ACM, 2014.

[10] H. Shum. Integrating microsoft academic search into cortana (keynote). In *Microsoft Research Faculty Summit*, 2014.

[11] H. Shum, Y. Kuo, and K. Wang. Bing dialog model: Intent, knowledge and user interaction. In *Microsoft Research Faculty Summit*, July 2010.

[12] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient topic-based unsupervised name disambiguation. In *JCDL*, June 2007.

[13] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *SIGIR*, 2007.

[14] K. Wang. Bing dialog: Towards richer interactions with web search. In *ACM SIGIR*, July 2014.