The Knowledge Web Meets Big Scholars

Kuansan Wang
Inernet Service Research Center
Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{firstname.lastname}@microsoft.com

ABSTRACT

Human is the only species on earth that has mastered the technologies in writing and printing to capture ephemeral thoughts and scientific discoveries. The capabilities to pass along knowledge, not only geographically but also generationally, have formed the bedrock of our civilizations. We are in the midst of a silent revolution driven by the technological advancements: no longer are computers just a fixture of our physical world but have they been so deeply woven into our daily routines that they are now occupying the center of our lives. No where are the phenomena more prominent than our reliance on the World Wide Web. More and more often, the web has become the primary source of fresh information and knowledge. In addition to general consumption, the availability of large amount of contents and behavioral data has also instigated new interdisciplinary research activities in the areas of information retrieval, natural language processing, machine learning, behavioral studies, social computing and data mining. This talk will use web search as an example to demonstrate how these new research activities and technologies have help the web evolve from a collection of documents to becoming the largest knowledge base in our history. During this evolution, the web is transformed from merely reacting to our needs to a living entity that can anticipate and push timely information to wherever and whenever we need it. How the scholarly activities and communications can be impacted will also be illustrated and elaborated, and some observations derived from a web scale data set, newly release to the public, will also be shared.

1. INTRODUCTION

Two and a half decades ago, the World Wide Web was born as a collection of hypermedia documents first envisioned by Bush during World War II [3] and Nelson in 1963 [7]. Just like a brick-and-mortar library where various forms of publications are classified and retrieved through a master catalog, the web quickly saw directory services, such as Yet Another Hierarchical Officious Oracle (Yahoo) [15], become a major means for users to discover and navigate to web documents. Information retrieval technologies that had been developed to mimic the services provided by a human librarian soon were applied [10]. They give rise to commercial search engines that, till today, are still a dominant and indispensable tool as web usages are being ever more deeply ingrained into our daily lives. Despite the strong demands and unrelenting investments, however, web search technologies have yet to achieve the goal of becoming as intelligent and helpful as a human librarian. For over two decades search engines have largely remained as a term matching machine. The web documents are indexed by keywords, which are then matched against the terms in the user query. It is mainly the responsibility of the users to include the "right" words and phrases into the query in order to obtain satisfying results. Search engines lend minimal assistance in query formulation other than reciting search history or quoting queries frequently issued by other users. The search results are presented as a list of hyperlinks to web documents that the users have to fetch and read through individually in order to get the desired answers. As the search engines index more contents and selling advertisements becomes a major means to sustain the technology investments, query terms are often misinterpreted for better or worse. A student who wants to brush up the knowledge of a linear algebra concept *matrix* will often find the search results peppered by commercial products that are selling namesake movie downloads or hair products. If the student has just obtained a driver's license and shared the information through personal communication service offered by the search engine company, it is very likely that the search results will include advertisements of a namesake car model from Toyota Motor Corporation that is quite popular among the students. The user experience can turn so gruesome when the spam contents seep through the search results that systematic and algorithmic breakthroughs to rid off unwanted and malicious contents have become an important and actively researched area (e.g., [6]).

The knowledge web is a recent effort to return the search technology development back to its original aspiration of reaching the level of intelligence comparable to or even exceeding a human librarian. First initiated in [9] and by the industry in general [5, 12], the goal of the knowledge web is to go beyond the lexical analysis and elevate all aspects of search algorithms to the semantic level. A knowledge web search engine analyzes and indexes documents in meaning-bearing units, called *entities* that model the concepts of the world, and the *relationships* among them that characterize the knowledge of the entities. Similarly, a knowledge web processes user queries in entities and relationships with which the documents passages are assessed for relevance. Both aspects of the semantic analysis are the subject of the Entity Recognition and Disambiguation (ERD) Challenge jointly held by major web search companies [4]. These research efforts enable the knowledge web to distill meaning from numerous synonymous expressions, alleviating the

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741739.

burden of users in composing the literal constructs to express their intents.

The work described in [14] goes a step further to redefine the knowledge web as a *human-computer dialog* problem, of which information retrieval is a key but not the only ingredient. More specifically, the criteria to optimize the system are no longer just the relevance per query but the user satisfaction over the entire dialog session. Recasting the problem as dialog lends itself immediate impacts on the user experience. First, recognizing the search intent cannot always be deciphered in a single exchange, the knowledge web is equipped with Bayesian inference capabilities [1, 13] to compute the best course of interactions. In addition to answer, the knowledge web can execute dialog acts such as confirmation and disambiguation if such actions can probabilistically satisfy user faster. Aside from reactively responding to user, the system can proactively suggest recourse, refinement or even digression. When the user has trouble expressing intents, the system can now proactively tap into the indexed knowledge and offer meaningful suggestions. When the intent expression is ambiguous, the system can confirm or disambiguate rather than presumptuously jump into actions. Most importantly, the system can now be more confident in the timing of recommending commercial offers, turning them from annoyances to opportunities for the user.

In terms of evolving the web from a collection of documents intended for human to read to a collection of knowledge and information that can be manipulated by machine, the knowledge web and the semantic web proposed by Berners-Lee *et al* [2, 8] share the same vision. The development of knowledge web, however, adopts a tactics that we feel more pragmatic. Instead of imposing a standard, Resource Description Framework (RDF), as a means for machine to share and exchange semantics, we prioritize our efforts to teach machine to read in human languages so that semantics can be retained and shared among systems in its original natural language forms. This approach enables the machine to converse not only with one another but also with the human and bring benefits in assisting human to acquire knowledge from the web. We believe the knowledge web approach is better in sidestepping the adoption problem that the semantic web community is facing, while not precluding the possibility of ingesting RDF contents from the semantic web.

2. APPLICATIONS FOR THE ACADEMIC DOMAIN

Many search engines dedicated to the academic domain exist, but the search logs indicate that these engines have been most effective as an electronic library without librarians, namely, the search activities are dominated by paper or author retrievals that their names are already known to the users. There do exist search sessions that the users want to conduct research on subject matters or domain experts in particular areas. The search results have been poor and traffic for such sessions has been very low, probably because users have learned not to use search engines under these scenarios.

Microsoft Academic Services (MAS) is a recent attempt to introduce an electronic librarian to the digital libary by applying the notion of knowledge web to the academic domain. As described in [11], the domain is modeled by six entity types surrounding the scholarly communication activities: publication (e.g., papers, research grant proposals), author, institution, field of study, venue (e.g., journal, conference series), and event (e.g., an issue of the journal, an instance of a conference, academic news), and the "librarian" is trained to conduct statistical inferences on such a heterogeneous graph. Although still in training, MAS has shown promise and been very helpful in many new use cases that are hopelessly difficult to do with a traditional search engine. For example, MAS can easily list the prominent authors for a field of study based on their publication records, making it suitable for researchers to identify potential project collaborators or paper reviewers. Conflict of interests can be resolved with little effort as MAS is aware of the affiliation history of each recommended author. Even publication search can be augmented with filtering capabilities based on venue of publication or author affiliations. Co-author relationships can be quickly discovered in queries much closer to natural language, and MAS is often seen to provide serendipitous information to further characterize the relationships based on other properties of the authors and their joint work. As academic researchers ourselves, we are very encouraged by the potential of MAS, and look forward to engage with the community to jointly develop MAS and take it to the next level of intelligence.

3. REFERENCES

[1] J. Allen, C. I. Guinn, and E. Horvtz. Mixed-initiative interaction. *Intelligent Systems and their Applications, IEEE*, 14(5):14–23, 1999.

[2] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific American*, 284(5):28–37, 2001.

[3] V. Bush. As we may think. *SIGPC Note.*, 1(4):36–44, April 1979.

[4] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, and K. Wang. ERD'14: entity recognition and disambiguation challenge. In *ACM SIGIR Forum*, volume 48, pages 63–77, 2014.

[5] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, pages 601–610, New York, NY, 2014.

[6] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1189–1190, 2007.

[7] T. H. Nelson. *Literary Machines*. Mindful Press, 1994.

[8] N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

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

[10] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

[11] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of Microsoft academic service (MAS) and applications. In *Proceedings of the 24th International World Wide Web Conference*, WWW '15, Florence, Italy, May 2015.

[12] H. Wang. Chinese search engine – Baidu's practice. In *SIRIP, The 37th annual international ACM SIGIR conference*, Gold Coast, Australia, July 2014.

[13] K. Wang. A plan-based dialog system with probabilistic inferences. In *INTERSPEECH*, ICSLP-2000, pages 644–647, Beijing, China, 2000.

[14] K. Wang. Bing dialog: toward richer interactions with web search. In *SIRIP, The 37th annual international ACM SIGIR conference*, Gold Coast, Australia, July 2014.

[15] Yahoo! media relations. *The History of Yahoo! – How it all started...*, 2005.