Mining search behavior and user-generated content

Presentation at the Industrial Session
EDBT/ICDT 2012

Carlos Castillo
Yahoo! Research
chato@acm.org

ABSTRACT

In the first part of this presentation, we will overview two systems that gather and display intelligence from search behavior: "Yahoo! Search Clues" and "Yahoo! Political Insights".

In the second part, we will discuss two real-world problems encountered when mining user-generated content: determining which pieces of content are credible, and modeling how users influence each other.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based Services*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

1. MINING SEARCH BEHAVIOR

1.1 Demographics

Search Clues[1] is a system that allow publishers, advertisers, marketers, journalists, etc., as well as simply curious web users, to gather insights about search trends. For a given query, it displays information about the query frequency profile over time, most common previous and next queries, as well as a distribution of locations, age, and gender of the people that issue the query.

Its design incorporates a series of elements, that will be discussed during the presentation:

- Detection of trending topics [4] (Section 4).
- Query-flow-based search session analysis [1, 2]
- Demographic analysis of search [8]

1.2 Political leaning

Beyond frequency and demographic profiles, search behavior can reveal much more about users and their interests.

Political Insights[2] is a prototype of a system to find politically-loaded search queries. A search query is politically loaded if the web sites that are relevant for the query tend to have a clear political orientation (left-leaning or right-leaning). A mutual-reinforcement process starting from a small set of hand-labeled web sites is enough to produce meaningful results.

2. MINING USER-GENERATED CONTENT

2.1 Credibility

The Web is a primary source of news for many people (as high as 61% in the US in 2010, a figure that is still growing [7]). Micro-blogging is a well-established paradigm, as attested by the success of platforms such as Twitter[3]. Whenever a significant event happens in the world, micro-blog posts can provide in a short time valuable information, usually reported by citizens close to the news source.

However, finding trustworthy, credible, information, continues to be challenging. In the specific case of micro-blogging, our problem can be described as (i) detecting newsworthy information, as opposed to just chatter and friendly messages; and (ii) detecting credible information. There are many superficial features from the micro-blog posts and from the way they are propagated by users that can provide evidence to solve these two problems [3].

2.2 Influence

In any user-generated content platform that allows re-posting of content, and in micro-blogging in particular, the concept of *influence* plays an evident role that can be easily observed. This effect can also be measured if an influence model is assumed. A popular model for influence in this setting is the Independent Cascade Model [5], in which an node that performs an action (e.g. a user who has posted a news item) has a certain probability of making each of its neighbors in the graph perform the same action.

Estimating these probabilities accurately given a log of actions (and possibly a graph of connections between nodes) is important to determine the influence of different users.

[1] `http://clues.yahoo.com/`

[2] `http://politicalinsights.sandbox.yahoo.com/`

[3] `http://twitter.com/`

EDBT/ICDT 2012, March 26–30, 2012, Berlin, Germany.
Copyright 2012 ACM 978-1-4503-0790-1/12/03 ...$10.00

The estimated model, however, might be unnecessarily dense, in the sense that it can be easily replaced by another model in which many influence probabilities are set to zero. In other words, in practice only a few connections among users (a "backbone" of influence) is all we need to explain many of the propagations observed in real-world information systems [6].

3. REFERENCES

[1] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 609–618, New York, NY, USA, 2008. ACM.

[2] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 56–63, New York, NY, USA, 2009. ACM.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[4] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 11–20, New York, NY, USA, 2010. ACM.

[5] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[6] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. Sparsification of influence networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 529–537, New York, NY, USA, 2011. ACM.

[7] Pew Research Center. The New News Landscape: Rise of the Internet. http://pewresearch.org/pubs/1508/-internet-cell-phone-users-news-social-expe rience, 2010.

[8] I. Weber and C. Castillo. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 523–530, New York, NY, USA, 2010. ACM.