Summarization System by Identifying Influential Blogs

Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng
NEC Laboratories America, 10080 N. Wolfe Road, SW3-350, Cupertino, CA 95014, USA
{xiaodan, ychi, hino, belle}@sv.nec-labs.com

ABSTRACT

We demonstrate a system that summarizes the opinions in the massive and complex blogosphere by finding the most influential blogs with highly innovative opinions.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Information Filtering*

Keywords

Summarization, Ranking, Social Influence

1. INTRODUCTION

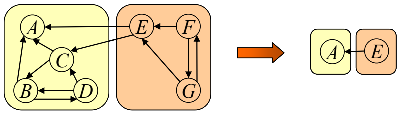
Blog is a self-publishing media on the Web that has been growing quickly and becoming more and more popular [1]. Blogs allow millions of people to easily publish, read, respond, and share their ideas, experiences and expertise. The blogosphere is a fruitful media for us to understand people's responses to events, and gather customers' opinions on products and services. Thus it is important to have a tool to find and summarize the most informative and influential opinions from the massive and complex blogosphere.

Social influence describes the phenomenon by which the behavior of an individual is directly or indirectly affected by the thoughts, feelings, and actions of others in a population [2]. Such influence is present and plays an important role in the blogosphere. The conversation in the blogosphere usually starts from innovators, who initiate ideas and opinions; then followers are primarily influenced by the opinions of these innovators. Thus the opinions of the influential innovators represent the millions of blogs and thousands of conversations on any given topic.

As an example, Figure 1(a) illustrates how seven blogs link to each other when they publish their opinions. Blogs *A*, *B*, *C*, and *D* discuss the same topic – e.g. how to share videos in YouTube. Then Blog *E* initiates the discussion of Google's acquiring YouTube, and links to blogs *A* and *C* to describe the background of YouTube. Following blog *E*, blogs *F* and *G* start to discuss this acquisition news. In this example, blog *A* and blog *E* are the influential blogs with innovative opinions. They also highly influence the opinions of other blogs. The opinions from blog *A* and *E* are the most representative and provide an overview of the opinions in the blog network. A summarization of this blog network in Figure 1(b) captures these influential blogs with innovative opinions.

In this demo, we summarize the blogosphere by capturing the most influential blogs with highly innovative opinions.

ICWSM 2007, March 26-28, 2007, Boulder, Colorado, U.S.A.



(a) The blog network (b) The summarization network

Figure 1: A motivating example

2. SYSTEM

Figure 2 provides an overall block diagram of our system. For a given query, the related blogs are first retrieved. Then retrieved blogs are ranked by how influential each blog to others and how innovative the opinions are in it. Following that, summarization is obtained by utilizing the ranking results.

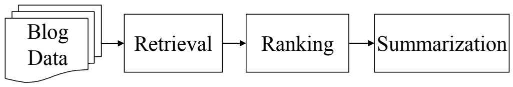
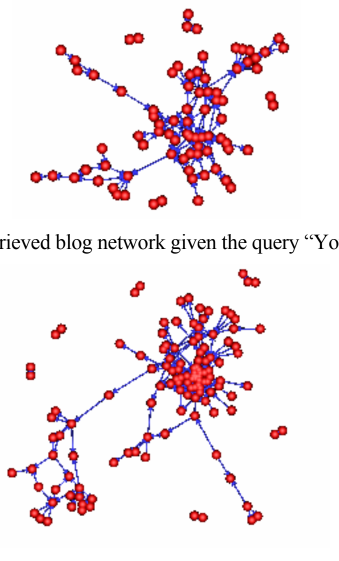


Figure 2: The overall block diagram of the system

In our system, given a query, we first show a blog network where nodes represent the blogs that discuss this query and edges represent the links among blogs embedded in entries. Figures 3(a) and (b) show the results when queries are "YouTube" and "Iraq" respectively.

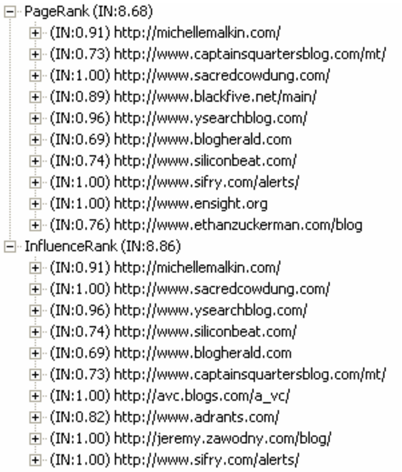


(a) The retrieved blog network given the query "YouTube"

(b) The retrieved blog network given the query "Iraq"

Figure 3: Retrieved blog networks: nodes represent blogs, edges represent entry links

After retrieving the blog network, we rank the top blogs using our InfluenceRank algorithm [4], in which blogs are ranked by how important they are to other blogs as well as the novelty of the information they contribute to the network. Information novelty of one blog is measured as the average information novelty of the entries in this blog compared to other entries which link to them. Figure 4 illustrates the top 10 blogs detected by our algorithm and PageRank [3]. Compared to PageRank, our InfluenceRank algorithm selects more influential blogs with novel information. As an example, given query "YouTube", http://www.captainsquartersblog.com/mt/ is ranked as 2nd by PageRank, while it gets demoted to 6th by our InfluenceRank algorithm because its information novelty is relatively low (*0.73*).

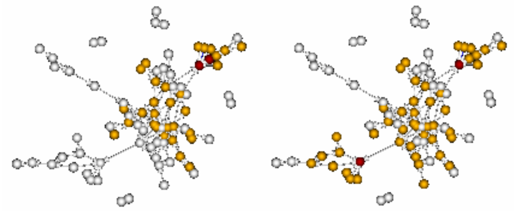


⊟ PageRank (IN:8.68)
⊞ (IN:0.91) http://michellemalkin.com/
⊞ (IN:0.73) http://www.captainsquartersblog.com/mt/
⊞ (IN:1.00) http://www.sacredcowdung.com/
⊞ (IN:0.89) http://www.blackfive.net/main/
⊞ (IN:0.96) http://www.ysearchblog.com/
⊞ (IN:0.69) http://www.blogherald.com
⊞ (IN:0.74) http://www.siliconbeat.com/
⊞ (IN:1.00) http://www.sifry.com/alerts/
⊞ (IN:1.00) http://www.ensight.org
⊞ (IN:0.76) http://www.ethanzuckerman.com/blog
⊟ InfluenceRank (IN:8.86)
⊞ (IN:0.91) http://michellemalkin.com/
⊞ (IN:1.00) http://www.sacredcowdung.com/
⊞ (IN:0.96) http://www.ysearchblog.com/
⊞ (IN:0.74) http://www.siliconbeat.com/
⊞ (IN:0.69) http://www.blogherald.com
⊞ (IN:0.73) http://www.captainsquartersblog.com/mt/
⊞ (IN:1.00) http://avc.blogs.com/a_vc/
⊞ (IN:0.82) http://www.adrants.com/
⊞ (IN:1.00) http://jeremy.zawodny.com/blog/
⊞ (IN:1.00) http://www.sifry.com/alerts/

Figure 4: Top-ranked blogs given query "YouTube", where *IN* denotes "Information Novelty" in the range of [0, 1]

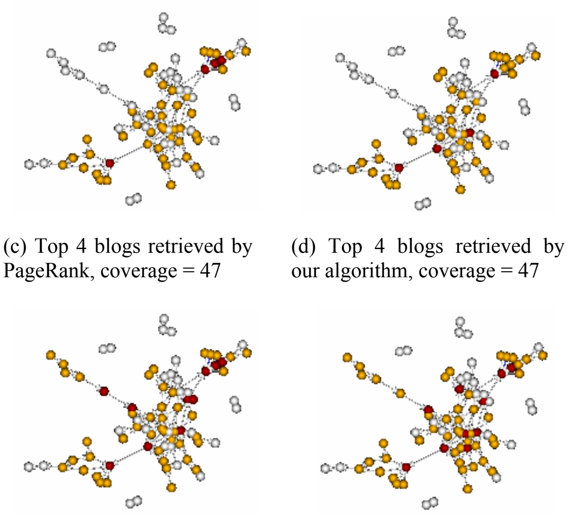
The top-ranked blogs capture the influential opinions thus provide a summary of the opinions regarding a query in the blogosphere. The number of blogs that these top-ranked blogs could influence in the network should be a good indicator of how good the summary is. We introduce the concept of *coverage* due to this intuition. In the blogosphere, coverage of a set of blogs is measured as how many blogs can either *directly* or *indirectly* reach the selected blogs by following any path of links. Intuitively, top-ranked blogs detected by InfluenceRank are novel information contributors, and thus they tend to be scattered instead of focused in the network. As a result, important blogs detected by InfluenceRank tend to have a better coverage comparing to those detected by PageRank. Figure 5 confirms these points and demonstrates that the top-ranked blogs detected by InfluenceRank tend to be more scattered in the network than those detected by PageRank. Furthermore, top-ranked blogs detected by InfluenceRank tend to have a higher coverage comparing to those detected by PageRank.

After obtaining the top-ranked blogs, our system summarizes the blog network by using the opinions from the important entries in top-ranked blogs to capture the influential and innovative opinions on a given query in the blogosphere. Figure 6 provides our story summarization result given query "YouTube". We also highlight the text mentioning this query.



(a) Top 2 blogs retrieved by PageRank, coverage = 32

(b) Top 2 blogs retrieved by our algorithm, coverage = 47



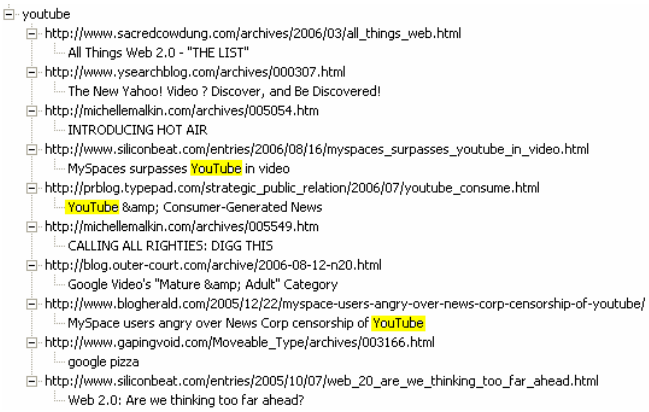
(c) Top 4 blogs retrieved by PageRank, coverage = 47

(d) Top 4 blogs retrieved by our algorithm, coverage = 47

(e) Top 10 blogs retrieved by PageRank, coverage = 54

(f) Top 10 blogs retrieved by our algorithm, coverage = 57

Figure 5: Blog network summarization results given query "YouTube". Dark red nodes represent top-ranked blogs, orange nodes represent blogs that directly or indirectly linked to the top-ranked blogs, and white nodes represent others.



⊟ youtube
⊟ http://www.sacredcowdung.com/archives/2006/03/all_things_web.html
All Things Web 2.0 - "THE LIST"
⊟ http://www.ysearchblog.com/archives/000307.html
The New Yahoo! Video ? Discover, and Be Discovered!
⊟ http://michellemalkin.com/archives/005054.htm
INTRODUCING HOT AIR
⊟ http://www.siliconbeat.com/entries/2006/08/16/myspaces_surpasses_youtube_in_video.html
MySpaces surpasses YouTube in video
⊟ http://prblog.typepad.com/strategic_public_relation/2006/07/youtube_consume.html
YouTube &amp; Consumer-Generated News
⊟ http://michellemalkin.com/archives/005549.htm
CALLING ALL RIGHTIES: DIGG THIS
⊟ http://blog.outer-court.com/archive/2006-08-12-n20.html
Google Video's "Mature &amp; Adult" Category
⊟ http://www.blogherald.com/2005/12/22/myspace-users-angry-over-news-corp-censorship-of-youtube/
MySpace users angry over News Corp censorship of YouTube
⊟ http://www.gapingvoid.com/Moveable_Type/archives/003166.html
google pizza
⊟ http://www.siliconbeat.com/entries/2005/10/07/web_20_are_we_thinking_too_far_ahead.html
Web 2.0: Are we thinking too far ahead?

Figure 6: Opinion summarization result given query "YouTube"

In our demonstration, a blog summarization system is presented that leverages influential blogs to capture the opinions in the massive and complicated blogosphere. Influential blogs create innovative opinions that affect other blogs, and thus are a good source for summarization. Our system uses an InfluenceRank algorithm to identify the influential blogs. The summarization derived from the influential blogs provides a more diverse and comprehensive summary of opinions in the blogosphere.

3. REFERENCES

[1] http://www.technorati.com/weblog/2006/11/161.html

[2] E. Katz and P. Lazarsfeld, Personal Influence, New York: The Free Press, 1955

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 30(1-7):107-117, 1998.

[4] X. Song, Y. Chi, K. Hino, and B. L. Tseng, InfluenceRank: Finding Opinion Leaders in the Blogosphere, submitted to SIGIR, January 2007.