International Journal of Advanced Technology & Engineering Research (IJATER)
National Conference on Recent Trends in Science, Technology & Management (NCRTSTM-2018)

BLOGS RECOMMENDATION USING BLOG POPULARITY ALGORITHM AND ITS APPLICATIONS

Krutika Bang , K D Bamane, Arati Gaikwad , Ashish Kumar Saxena,
Pune,India
krutikabang@gmail.com, kalyandbamane@gmail.com, arati aratig.2010@gmail.com, ashishsaxenaofc@gmail.com

Abstract

Social media plays important role in almost every domain. To express the opinions, ideas, emotion and interests there are various resources available on Internet. Amongst all the resources blogs are most popular for the people to express opinion, and ideas. For analyzing the sentiments of public reviews regarding specific products or any other issues web blog mining is the efficient way. Blogs Consist of un-indexed and unprocessed text that reflects the opinion of people. Sentimental opinion mining is the best efficient way to mine this blogs since blog has lots of information.

To evaluate the system, we collect user's feedbacks by experimenting on specific domain blogs. However significant intervals for each Website are computed first (independently) and then the Analysis is performed on it. The output so obtained are the blogs which are of most popular in discussions. Here the approach is to review web blogs using web mining. An experiment has been performed for the analysis of results obtained from the blogs.

Keywords: Blogs, Popularity, Recommendation, Blog Popularity, in-links.

Introduction

Blogs can be defined as frequently modified Web pages in were date entries are listed in reverse chronological sequence. The people write them to express their opinions and emotions, making blogs popular. Analyzing these could provide opportunities for governments and many companies to understand the public demands and way of thinking that was previously costly or difficult to obtain. Many blogs exist, hence manually monitoring and analyzing them is a labor-intensive and time-consuming task. Thus Blog Mining is solution to it where marketers or companies, for example, can get closer to their customers and know about their opinions on certain products. Also political issues can be discussed, disseminate information discussion of ideas, and questions answers can be done through blogs. There are lots many increasing number of blogs and their unique characteristics, there is need for developing techniques for searching and mining them. Due to this kind of mining of blogs, it may be possible for researchers to know what people are thinking about certain things and how they express their feeling at the personal and professional level. People write their true opinions and feeling about something on their blogs. This actual opinion and feeling are usually cannot be expressed in surveys or polls. If such opinions and feeling are understood and dealt at proper time, certain potential problems could be fixed or avoided.

Many challenges exist to apply existing text and Web mining techniques to blogs. The challenges are

- Bloggers update their work much more frequently than Web masters
- Bloggers cover very diverse topics maybe only one paragraph in a particular entry could relate to someone's topic of interest—for example, a product being analyzed

Blogs and Web pages' characteristics differs so much that different mining techniques must be used. For example, you can apply structure mining techniques on hyperlinks between general Web pages, but a hyperlink isn't the only way to link blogs—they can also be linked via comments or subscriptions to other blogs.

Various ranking algorithm such as Page Ranking are available to rank blogs. However, the simple adoption of these algorithms to blogs has some issues such as:

- The links count to a blog entry is generally very small. Scores so calculated by PageRank, are generally too small to permit blog entries to be ranked by importance.
- Some time is needed to develop a number of in-links and thus have a higher PageRank score. Blogs are considered to be Strong communication medium for discussing recent topics, it is desirable to assign a higher score to an entry submitted by a blogger who has been received a lot of attention in the past but do not have more number of in links

System Design

It consists of following modules as:

a) Crawler
b) Parser
c) Blog Popularity algorithm

d) Lucene
e) Output & Opinion Retrieval
f) Crawler

Crawler

Crawlers are typically programmed to visit sites that have been submitted by their owners as new or updated. This sites or some specific pages can be selectively visited and thus are accordingly indexed.

Initially, a Web crawler starts with a list of URLs to visit. During its visit to URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit. A crawler has to take special concern in choosing at each step which pages to visit next.

.

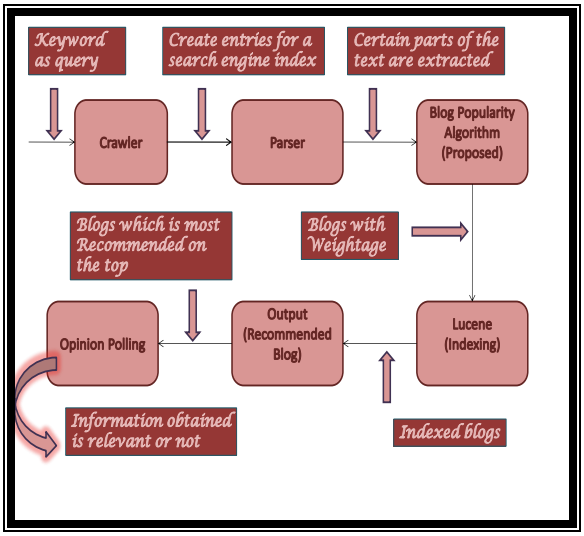


Figure 1. Framework for Blog Recommendation

Parser

A Software component that takes input data (text) and builds a data structure is function of Parser. This Structure can be some kind of parse tree, abstract syntax tree or hierarchical structure. This helps to give a structural representation of the input, checking for correct syntax is in the process. The use of parsers varies by input. In the case of data languages, a parser is often found as the file reading facility of a program, such as reading in HTML or XML text. . In our system blog *parser* extracts information from blogs, including names of people, products, and organizations. It also includes other patterns, such as dates, times, and URLs. Developers can create their own tools to extract information from blogs. In addition to text, blog parsers also extract structural information from blogs, such as comments, posted links, or the bloggers' groups or blogging communities. This data forms the blog linkage information that can be used in blog popularity algorithm.

c) Blog Popularity Algorithm

Step I: Consider the blogs in the set.
Step II: Calculate the surfing probability that blog reader will follow a link in blog A to another blog B

through

$$\mathbf{P}_{\mathbf{A}\to\mathbf{B}} = \frac{\mathbf{R}(A\to B)}{\sum_{\mathbf{X}\epsilon\mathbf{O}(\mathbf{A})}\mathbf{R}(A\to X)}$$ where $O(A)$ means blogs

linked by A

Step III: The relationship score $R_{A\to B}$ represents the relation strength from A to B which is obtained as

$$\mathbf{R}_{\mathbf{A}\to\mathbf{K}} = \sum_{RType}\mathbf{W}(Rtype) * \mathbf{RN}(RType) * \mathbf{B}(Qk)$$

Where $\mathbf{W}(Rtype)$: Type of blog relationship
$\mathbf{RN}(RType)$ The number of the corresponding relationship
$\mathbf{B}(Qk)$: The blog quality score

Step IV: The relationship score is found for each directed

node pair in the social blog network.
Step V: These score is used in probability formula
Step VI: Apply the random walking on the network with

the modification of propagation probability.
Thus the rank of the blog can be find as

$$\mathbf{Brank}\ (A) = \frac{1-d}{n} + d * \sum_{X\in I(A)}\boldsymbol{BRank}\ (\boldsymbol{X}) * \boldsymbol{P}\ (X \to A)$$

Where $I(A)$ is the set of blogs which are linked to A
d is the damping factor

In a social blog network, the algorithm computes popularity scores to rank a single community's blogs. These algorithm modifies the surfing probability in PageRank algorithm,

$$P_{A\to B} = \frac{1}{Outdegree\ of\ blogA},$$

(1)

The probability that visitor visits from NodeA to Node B ($P_{A\to B}$) is decided by the out-degree of A. Now we can adjust the probability that a blog reader

will follow a link in blog A to another blog B using this new formula,

$$P_{A\to B} = \frac{R_{A\to B}}{\sum_{X\in O(A)} R_{A\to X}}, \quad (2)$$

where $O(A)$ means blogs linked by A. In these, the probability is determined by the relationship scores ($R_{A\to B}$). In Equation 2, X indicates the blogs to which blog A links.

The relationship score $R_{A\to B}$ gives the relation strength from A to B. It's decided by three factors.

The first is the type of blog relationship (comment, trackback, blog roll, or citation). Different blog relationships are assigned different weights (W_{Rtype}) because they have distinct meanings for a blogger. In our experiments, $W_{comment}$ is set to 0.25 and others are set to 1. This setting simply represents how easy it is to make a relationship (for example, in our observation, leaving a comment is the easiest way to support a blogger by using the blogging interface).

The second factor is the number of the corresponding relationship. Here the degree of the number (RN_{Rtype}) is to express the relationship's strength. Here Inspite of using the actual numbers, we use the actual numbers natural log. The final factor is the blog quality score (B_{Qk}), which combines the normalized blog features, including the commented post count, the tracked

post count, and the average blog/post life cycle. We use the time span between the last date and first date for all posts to represent a blog's life cycle. These metrics are automatically extracted from data sets.

The blog quality score shows its basic activity. Higher quality score for a blog indicates that the blog's relationships are stronger as compared to ones with a lower score and therefore might receive more support from other bloggers. We assume that the probability of a user moving to a blog with a higher quality score is greater than that of moving to others. This quality score so obtained is converted to the natural log value for future calculation. The relationship score combines all kinds of relationships between two blogs. The relationship score from blog A to blog K is defined as follows:

$$R_{A\to K} = \sum_{Rtype} W_{Rtype} * RN_{Rtype} * BQ_k.$$

Thus we now compute the relationship score for each directed node pair in the social blog network. A directed node pair could be connected by several support edges, a bidirectional interest edge, or both kinds of edges. We then apply the random walking on the network with the modification of the propagation probability. Thus the rank of the blog can be find as below

$$BRank(A) = \frac{1-d}{n} + d * \sum_{X\in I(A)} BRank(X) * P_{X\to A}$$

Where $I(A)$ is the set of blogs linking to A,

d is the damping factor as in the initial PageRank algorithm. Thus using this algorithm Weightage can be given to each blog.

d) Lucene

Lucene is a free or open source information retrieval software library. It is recognized for its utility in the implementation of Internet search engines and single-site searching.

The logical architecture of lucene is the idea of a document containing fields of text. Lucene's API is independent of the file format. Text from PDFs, HTML, Microsoft Word, and Open Document documents, as well as many others (except images), can all be indexed as long as their textual information can be extracted.

e) **Output and Opinion Retrieval**

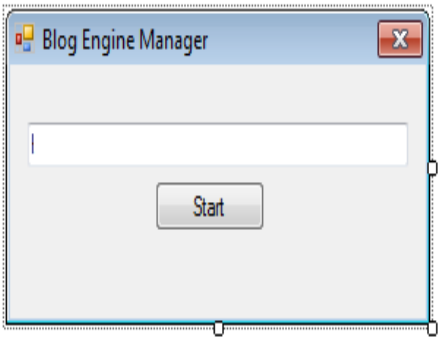
The Output so obtained is the list of the blogs with most recommended blog on the top. There are more blogs available for the reference that will be helpful for user to get or obtained the desired data.

Also opinion is being obtained in the form of positive and negative count in order to determine how the data is relevant to user as per the query he/she have fired on Search Engine.

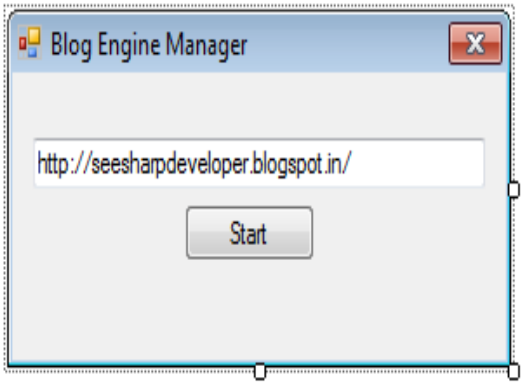
SYSTEM EXECUTION

Following Results appears after implementation of this system design

Following Screen Appears when crawler is debugged. The crawler crawls the web page through world wide web



Following Screen shows the input to the crawler is URL so input is given so that crawler starts its work and retrieve data in database.

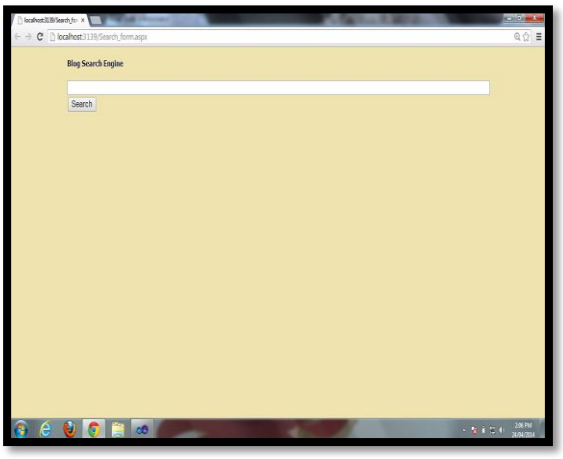


Following Screen shows the entries are stored in data base through proper schema which is crawled by the crawler

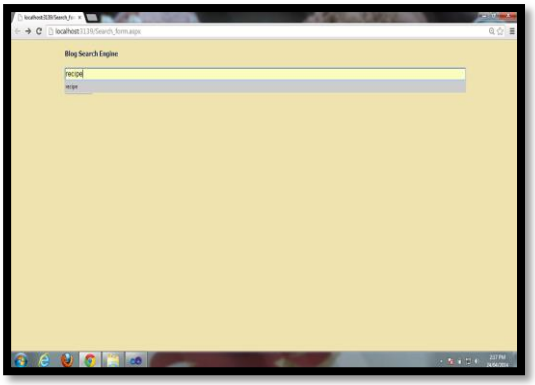
KRUTIKA-PC.BlogR...ng - dbo.UserPost | Summary

Postid	BlogId	PostUrl	Author	PostTitle	Category
3647516248262598	3172831935077...	http://acompute...	zack duncan	Blackhole Exploit...	Blackhole exploit...
9393896701733...	5509307	http://feedprox...	Amy Sherman	Americana Culin...	
9404279868606...	2985150316740...	http://rita-may-r...	Rita May	Pumpkin Waffles	breakfast
2768459354569...	4335524343232...	http://myerecip...	Mythreyi Dilip	Mexican Rice	Mexican Recipes...
3423031620479...	24237728	http://spanishfo...	STRIKER	FC BARCELONA ...	2013 - 2014,201...
3949357750359...	2985150316740...	http://rita-may-r...	Rita May	French Canadia...	Holidays,pork
4041072846237...	3172831935077...	http://acompute...	zack duncan	Windows 8 Not t...	Enderle Group,m...
44167645785 92...	5509307	http://feedprox...	Amy Sherman	More Favorites f...	
4884692349014...	5509307	http://feedprox...	Amy Sherman	New Cookbooks ...	
4889399318837...	4335524343232...	http://myerecip...	Mythreyi Dilip	Vegetable Kothu...	Fusion/ Innovati...
4907574321731...	7351428632842...	http://seesharp...	Ravi Singh	C# Set Impleme...	collections,desig...
4925524024855...	24710794	http://feedprox...	Kongkon Jyoti D...	C++ questions: ...	C++ Questions,...
5154507239784...	4335524343232...	http://myerecip...	Mythreyi Dilip	Vegetable Hari B...	North Indian Re...
56736341814 53...	7351428632842...	http://seesharp...	Ravi Singh	Compiling ASP.N...	asp.net mvc
5992149012419...	24710794	http://feedprox...	Kongkon Jyoti D...	Single Linked List...	C,C Programmin...
5999093787188...	4335524343232...	http://myerecip...	Mythreyi Dilip	Chinese New Ye...	General/Celebra...
6831523209491...	24710794	http://feedprox...	Kongkon Jyoti D...	C++ questions: ...	C++ Questions,...
7799927453560...	24237728	http://spanishfo...	STRIKER	REAL MADRID | ...	2014,Gareth Bal...
7896531296280...	24710794	http://feedprox...	Kongkon Jyoti D...	C++ questions: ...	C++ Questions,...
8006578987450...	4335524343232...	http://myerecip...	Mythreyi Dilip	Idli Kebab / Cock...	Idli and Dosa Sid...
8336363629623...	5085053248658...	http://padmasre...	Padma	Back to Blogging...	
8558315734272...	24237728	http://spanishfo...	STRIKER	SPANISH FOOTB...	2013 - 2014,Lig...
8715123675940...	5085053248658...	http://padmasre...	Padma	RAGI UPMA	Ragi,Upma
9494111779511...	6835777771696...	http://myindia-t...	Vijay	Taj Mahal - Agra	Agra,Uttar Prad...

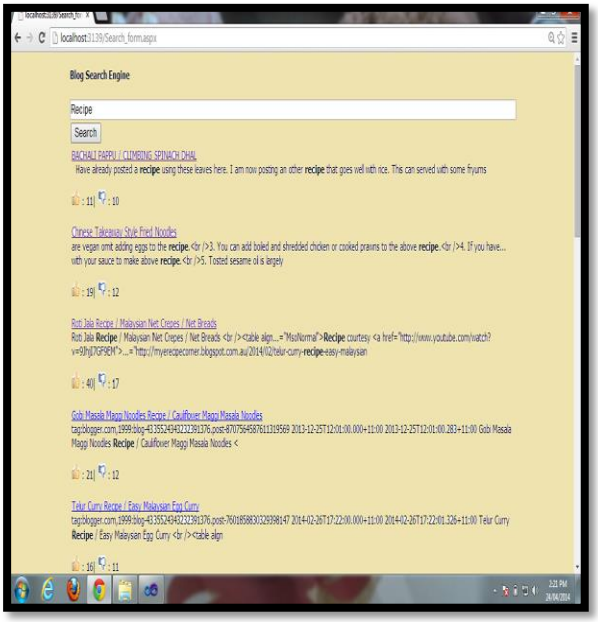
The Following Screen shows the windows which appears first as soon as the Search Engine is run



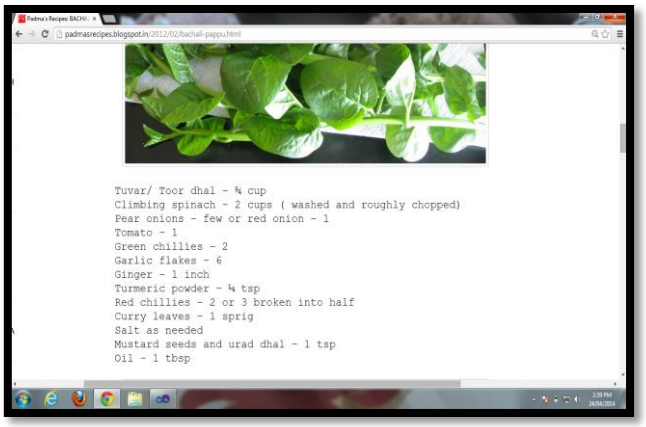
The following Screen shows that user can add a keyword of a particular category of which he want to search blog for eg. Recipe

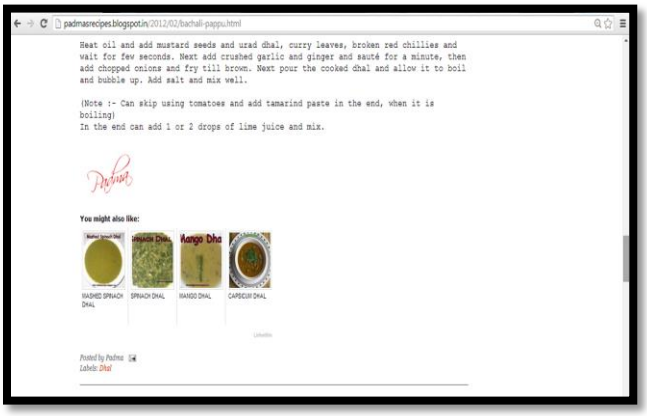


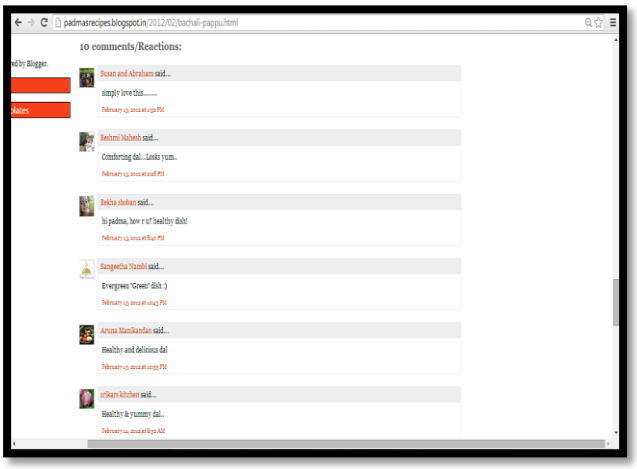
Following Screenshot shows the list of blog that will be obtained as a result with most recommended blog on top as follows



The following screenshots shows the entire blog which is obtained after clicking on top most link. The blog so obtained contain the data which user want to search along with the comments.







APPLICATIONS

Some potential blog mining applications for various domains are business, politics, disaster recovery, social work, cultural studies, and linguistics etc. The general application are as follows:

a) Analysis of Public Awareness

One useful application of blog mining is to evaluate what people say about a company or any organization. After Analyzing this blogs, companies gets a better understanding of their customers' concerns which help them to know about their area of improvement for better decision making.

b) **Analysis of Online Social Activities**

Many online Communities have been formed by bloggers. Opinion and beliefs of these communities on some social activities , sharing of their ideas can be done by reading and commenting on each other's blogs.

c) **Analysis of Public Opinion**

Another important blog mining application is *news monitoring*. Use of blogs has increased for news distribution like anyone can update a blog at any time, blogs represent the views of different individuals without filtering and blogs are interactive. Readers can easily post comments or react their views, or they can write their own blogs.

There are various other sectors in which blogs plays an important roles such as:

- Education
- Business
- Creating Opportunities for Oneself

CONCLUSION

Amongst the various ranking algorithm present for the web pages and blogs, blog popularity ranking algorithm was selected for implementation for the analysis and ranking of blogs. The work implements the Search Engines for the user to search the relevant and most popular blogs amongst

the particular domains.Work is carried out some selected domains of blog like recipe, C# etc. The data set consist of the different web blogs crawled for the selected application domain through world wide web. The algorithm which is implemented uses various attributes for ranking the blogs. The various attributes which are considered are comments, out links which are present in blogs along with the trackbacks In addition the various ranking algorithm consider only hyperlink as their ranking attribute. However considering other attributes than hyperlink is important aspect in Search Engine from users perspective. These framework has extensive contribution to the field of Information Retrieval, Web Intelligence and Information system on various domains through blogging.

LIMITATION

One limitation of blog mining relates to blog quality: some companies pay bloggers to write positive product reviews, thus these blogs don't reflect true user viewpoints. Another problem is splogs—spam blogs that people create to promote a product or another Web site—which are also becoming increasingly popular. Blog mining applications must determine how to distinguish genuine blogs from the others.

REFERENCES

[1] Michael Chau, Porsche Lam, and Boby Shiu, Jennifer Xu and Jinwei Cao, "The blog mining Framework" Social network Application 1520-9202/09 Computer.org /ITPro IEEE Computer Society Jan/ Feb 2009.

[2] Geetika T. Lakshmanan & Martin A. Oberhofer "Knowledge Discovery in Blogosphere "Social computing in Blogosphere 1089-7801/10 IEEE Computer Society Mar/Apr 2010.

[3] Chih-Lu Lin and Hung-Yu Kao "Blog Popularity Mining Using Social Interconnection Analysis" Social Networking 1089-7801/10 IEEE Computer Society Mar/Apr 2010.

[4] H. Qian and C.R. Scott, "Anonymity and Self-Disclosure on Weblogs," J. Computer-Mediated Comm., vol. 12, no. 4, p. 1.

[5] N. Glance et al., "Analyzing Online Discussion for Marketing Intelligence," Proc. 14th Int'l Conf. WWW (WWW 2005), ACM Press, 2005, pp. 1172–1173.

[6] A. Qamra, B. Tseng, and E.Y. Chang, "Mining Blog Stories Using Community-Based and Temporal Clustering," Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM 2006), ACM Press, 2006, pp. 58–67.

[7] B.Chen et al., "Predicting Blogging Behavior Using Temporal and Social Networks," Proc. 7th IEEE Int'l Conf. Data Mining (ICDM 2007), IEEE CS Press, 2007, pp. 439–444.

[8] T. Nanno et al., "Automatically Collecting, Monitoring, and Mining Japanese Weblogs," Proc. 13th Int'l Conf. WWW, (WWW 2004), ACM Press, 2004, 320–321.

[9] B. Nardi et al., "Why We Blog," Comm. ACM, vol. 47,no. 12, 2004, pp. 41–46.

[10] R. Blood, R., "How Blogging Software Reshapes the Online Community," Comm. ACM, vol. 47, no. 12, 2004, pp. 53–55.

[11] R.Kumar et al., "Trawling the Web for Emerging Cyber communities," Computer Networks, vol. 31, nos. 11–16, 1999, pp. 1481–1493.

[12] S. Baker and H. Green,. "Blogs Will Change Your Business," Business Week, 2 May 2005, pp. 44–53.

[13] M. Chau and H. Chen, "Personalized and Focused Web Spiders," Web Intelligence, eds., N. Zhong, J. Liu, and Y. Yao, eds., Springer-Verlag, 2003.

[14] K. Fujimura, T. Inoue, and M. Sugisaki, "The EigenRumor Algorithm for Ranking Blogs," Trusting Agents for Trusting Electronic Societies, LNCS 3577, Springer, 2005, pp. 59–74.

[15] M. Chau and J. Xu, "Studying Customer Groups from Blogs," Proc. 6th WeB 2007, (WEB2007),2007.