Blog Ranking Based on Bloggers' Knowledge Level for Providing Credible Information

Shinsuke Nakajima[1], Jianwei Zhang[1], Yoichi Inagaki[2], Tomoaki Kusano[2],
and Reyn Nakamoto[2]

[1] Kyoto Sangyo University
nakajima@cse.kyoto-su.ac.jp, zjw@cc.kyoto-su.ac.jp
[2] kizasi Company, Inc
{inagaki, kusano, reyn}@kizasi.jp

Abstract. With the huge increase of recently popular user-generated content on the Web, searching for credible information has become progressively difficult. In this paper, we focus on blogs, one kind of user-generated content, and propose a credibility-focused blog ranking method based on bloggers' knowledge level. This method calculates knowledge scores for bloggers and ranks blog entries based on bloggers' knowledge level. Bloggers' knowledge level is evaluated based on their usage of domain-specific words in their past blog entries. A blogger is given multiple scores with respect to various topic areas. In our method, blog entries written by knowledgeable bloggers have higher rankings than those written by common bloggers. Additionally, our system can present multiple ranking lists of blog entries from the perspectives of different bloggers' groups. This allows users to estimate the trustworthiness of blog contents from multiple aspects. We built a prototype of the proposed system, and our experimental evaluation showed that our method could effectively rank bloggers and blog entries.

1 Introduction

Recently, user-generated content websites such as blogs and social networking services have become established as popular online pastimes. Being user-generated, the amount of data created and subsequently available on the Web has grown exponentially. This, combined with the widely varying quality of user-generated content, makes it increasingly difficult to find credible information. When searching for information on the Web, a user usually makes use of a search engine such as Google. Google's PageRank algorithm [1] does an excellent job of reflecting the general popularity and authority of Web pages. However, traditional search engines do not fare as well with user-generated content due to their differing information characteristics. Particularly, user-generated content is rapidly and frequently updated, thus not effectively evaluated by PageRank, which relies heavily on incoming links. Credibility-focused search and ranking methods for user-generated content are strongly required.

In this paper, we specifically focus on blogs. In this area, there exist several blog-specific search and ranking engines today. These engines usually take one of

G. Vossen, D.D.E. Long, and J.X. Yu (Eds.): WISE 2009, LNCS 5802, pp. 227–234, 2009.

two approaches: (1) ranking the individual entries, (2) ranking the entire blogs as a whole. Approach 1 often uses a keyword-based search and then sorts the results by the entries' post date. Google Blog Search [2] belongs to this type. However, this type of blog search engine does not effectively assess the entry contents, which lowers the credibility of the results. Approach 2 usually uses some combination of link count, access count, as well as voting. Technorati [3] is such a search engine. Although it can find relevant bloggers based on their entry history, it does not guarantee that the latest relevant entries will be returned due to their site-based ranking. Given that blog contents are heavily focused on the latest happenings, this is an unfortunate drawback.

We propose a credibility-focused blog ranking method based on bloggers' knowledge level. Our method assumes that a person who has high knowledge of a certain topic is more credible than a person with a low level of knowledge. Figure 1 shows the relationship between a knowledgeable person and credible information. By analyzing bloggers' past entries and ranking bloggers by their knowledge level, we can provide more credible blog entries to the end user.

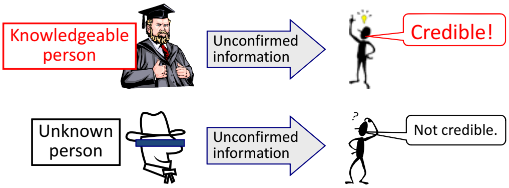

Fig. 1. Relationship between knowledgeable people and credible information

In our method, we first extract topic areas and create a term dictionary which provides domain-specific words for each of these topic areas. We then evaluate a blogger's knowledge level based on the blogger's usage frequency of these domain-specific words. Generally, a blogger's knowledge level varies with respect to different topic areas. A person may be an expert in one field, while (s)he may know little to nothing about another field. For example, a blogger who is familiar with automobiles may lack the basic knowledge on computers. Thus, in our research, a blogger's knowledge scores are calculated for multiple topics. We next rank blog entries based on their bloggers' knowledge scores. Blog entries written by knowledgeable bloggers are highly ranked. In addition, given search keywords, our method can provide multiple ranking lists of blog entries from the perspectives of different bloggers' groups. For example, when a user searches for the "Java" information, our system presents multiple ranking lists from the different aspects, such as "programming", "education", and "examination". Thus, our system is very helpful for users to find credible information by presenting a varied set of topic rankings and highly ranking blog entries written by knowledgeable bloggers.

2 Related Work

Recently, the number of blog search websites, such as Google Blog Search [2] and Technorati [3], has increased with the spread of Web 2.0. These blog search engines index blogs and provide ever useful search functions. These commercial services maintain our motivation on the research of blog ranking.

There is much ongoing academic research into blog search and ranking methods. Fujimura et al. [4] proposed an algorithm for ranking blog entries by weighting the hub and authority scores of a blogger. Kritikopoulos et al. [5] presented a method for ranking blog entries based on a link graph consisting of explicit links and implicit links. Both of these methods focus mainly on link analysis, while our method ranks bloggers and blog entries based on blogs' contents. Links between blog entries are helpful to further improve our blog ranking system, and effective use of this kind of information is a future direction of our current research.

In terms of credibility and trust related research, Gil et al. [6] introduced several factors that users should consider when deciding whether to trust Web contents or not. Adler et al. [7] presented a content-driven reputation system for Wikipedia which could extract authors with a high reputation. Andersen et al. [8] provided a trust-based recommendation system by making use of social network structures. Our work focuses on another type of Web resources, blogs. Our method can identify credible bloggers by analyzing their past blog entries, and subsequently find credible blog entries.

3 Estimating Bloggers' Knowledge Level

In this section, we explain how we estimate a blogger's knowledge level. It is divided into three main parts:

1. Constructing a dictionary of "Knowledgeable Bloggers' Groups (KBGs)", representing topic areas and their domain-specific words.
2. Identifying "Knowledgeable Bloggers (KBs)" for each KBG, assigning bloggers to their relevant topic areas.
3. Calculating KBs' knowledge scores for his/her KBG.

3.1 Constructing a Dictionary of KBGs

Generally, bloggers have their specific interests and post blog entries related to specific topic areas. We call a blogger who is familiar with a topic area as a "Knowledgeable Blogger (KB)" for this topic area. A set of knowledgeable bloggers for a topic area is marked as a "Knowledgeable Bloggers' Group (KBG)" for this topic area. We first extract some keywords representing the topic areas daily discussed in blogs. Each keyword becomes the title of the topic area, and also represents the name of the KBG familiar with the topic area. We then extract frequently used words for each topic area, and create a dictionary, which summarizes the topic areas and their domain-specific words. The detailed process of constructing a dictionary is shown as follows:

1. We perform a regular Web search by using the search keywords such as “expert”, “fan”, and “mania”, and extract the keywords which occur before and after these search keywords.
2. The keywords with occurrence frequency below a certain threshold are filtered.
3. We browse the keywords with occurrence frequency above the threshold, and remove duplicate and inappropriate ones. The remaining keywords are appended to the dictionary as the selected topic areas. In our current implemented system, about 14,000 keywords are registered on the dictionary.
4. For each keyword (i.e., the title of each extracted topic area), we extract the top n words which have high co-occurrence frequency with it from a large blog entry corpus. Specifically, n is 400 in our current system. These words which are domain-specific for their corresponding topic areas are also registered on the dictionary.

Figure 2 is an example of the dictionary. The column g shows the titles of the extracted topic areas (i.e., KBGs’ names). Each row shows each topic area’s domain-specific words and their corresponding co-occurrence frequency β. For example, the keyword “computer” is extracted as the representative title of the “computer” topic. It is also the name of the KBG who has high level of knowledge related to the “computer” topic. This topic area has its domain-specific words, such as “windows”, “desktop” and “company”. Additionally, the dictionary is re-constructed periodically, since blog entries are frequently updated and the words co-occurring with a topic vary with respect to time.

g	j = 1		2		...		400	
computer	windows	$\beta_{1,1}$	desktop	$\beta_{1,2}$	...	...	company	$\beta_{1,400}$
Obama	president	$\beta_{2,1}$	crisis	$\beta_{2,2}$	...	...	white	$\beta_{2,400}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fig. 2. Dictionary of KBGs

3.2 Identifying KBs for Each KBG

For the purpose of finding the relevant set of bloggers for each topic area, we next assign bloggers to their relevant topic areas. It is possible that a blogger is assigned to more than one KBGs. We regard a blogger who continues writing blog entries related to a topic area as a KB of this KBG. We list up several conditions by which it is decided that whether a blogger should be assigned to a KBG. These conditions focus on two aspects: the number of bloggers’ entries containing the title of the topic area or its domain-specific words, and the period in which bloggers continue writing such entries. An example is “the blogger is identified as a KB of a KBG, if (s)he wrote twenty or more entries containing domain-specific words related to the KBG over a three month period”.

3.3 Calculating KBs' Knowledge Scores for the KBG

We now calculate KBs' knowledge scores, which indicates how knowledgeable a KB is for a topic. Basically, scores are calculated based on how often as well as how in-depth a blogger writes blog entries related to a certain topic. If a blogger has an extensive use of the domain-specific words of a topic, high knowledge scores are attached to the blogger.

We first calculate $score_g(e)$, the score of an individual entry e with respect to the topic g as follows:

$$score_g(e) = \sum_{j=1}^{n} \alpha_j \cdot \beta_j \cdot \gamma_j \tag{1}$$

where $n = 400$ is the number of the domain-specific words, $\alpha_j = \frac{n-j}{n}$ is the weight of word j which decreases as j increases, β_j is the co-occurence frequency of word j, and γ_j is a binary value which indicates whether entry e contains word j or not.

Once we have the individual entries' scores, we next calculate $score_g(b)$, the score of a blogger b with respect to the topic g:

$$score_g(b) = \frac{l}{n} \cdot \frac{log(m)}{m} \cdot \sum_{i=1}^{m} score_g(e_i) \tag{2}$$

where e_i is an entry which blogger b wrote, m is the number of entries which blogger b has posted within a given period, $n = 400$ is the number of the domain-specific words, and l is the number of the domain-specific words which occurred in all the entries written by blogger b. $\frac{l}{n}$ indicates the coverage ratio of the domain-specific words which blogger b has used. $\frac{log(m)}{m}$ reduces the effect that a blogger frequently writes a large amount of entries, but most of them are unrelated entries.

4 Ranking Blog Entries Based on Bloggers' Knowledge Level

We now explain the process of blog search and ranking. When an end user enters one or more search keywords, our system does the following steps:

1. Blog entries which contain the search keywords are retrieved. These entries will be grouped and ranked by our system.
2. The system then identifies the bloggers of these entries, and finds the KBGs which the bloggers belong to. As mentioned before, a blogger may belong to multiple KBGs.
3. The KBGs are sorted in the descending order by the numbers of bloggers in each group.
4. Blog entries retrieved by Step 1 are assigned to the corresponding KBGs, according to the affiliation of their bloggers. Since a blogger may belong to multiple KBGs, it is possible that a blog entry is grouped to multiple KBGs.

5. For a KBG, the entries in this group are ranked in the descending order by the knowledge scores of their bloggers. That is to say, the entries written by bloggers with high knowledge scores are given higher rankings than those written by bloggers with low knowledge scores.

We built a prototype of blog search engine [9] based on the previously described method. The beta version of this system was released in September, 2008. As of April 20, 2009, our system contains about 174,000,000 blog entries collected from about 7,422,000 bloggers. The number of KBGs is about 14,000 and the number of KBs for all the KBGs is about 100,000.

Figure 3 is a snapshot of the implemented system prototype. The KBGs are presented in the left part of Figure 3. The end user can freely select the topic they are interested in, and then the entries in this group are presented in the right part of Figure 3. In this way, various aspects from different ranking lists can be provided to the end user. The user can then browse the entries from the aspects they are concerned about. Thus, the user can better acquire knowledge, since entries are organized by different perspectives and more credible information is highly ranked.

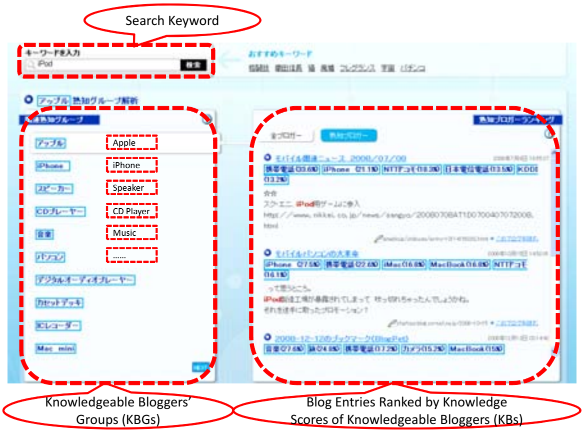

Fig. 3. A snapshot of the system prototype

5 Experiments

We randomly selected 20 search keywords, and four individuals were asked to evaluate our implemented system about the precision of KBG, KB, and blog entries.

5.1 KBG's Precision

Given a search keyword, our system returned some KBGs. Each individual checked the top 5 KBGs and decided whether the name of each KBG was related to the search keyword. The precision is the ratio of the number of relative ones to all the top 5 KBGs. Figure 4 shows the results. k represents a search keyword, each bar for $k1$ to $k20$ is the precision average of four individuals for each search keyword, and the bar for $ave20$ is the precision average of 20 search keywords. The fact that the precision is about 1 indicates that most of the KBGs returned by our system are related to the corresponding search keywords.

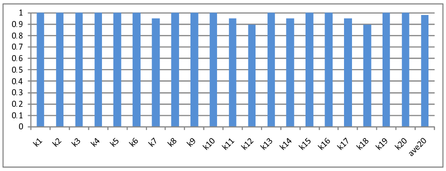

Fig. 4. KBG's precision

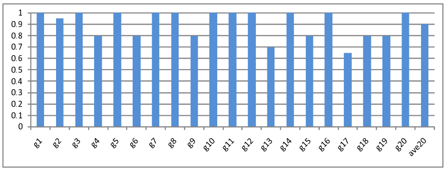

Fig. 5. KB's precision

5.2 KB's Precision

The four individuals were also asked to evaluate KBs. We selected the top KBG for each of 20 search keywords, and subsequently acquired 20 KBGs. For each KBG, the top 5 bloggers were evaluated by each individual. The individual browsed bloggers' entries and judged whether they were appropriate as a KB with respect to the KBG in question. The precision is the ratio of the number of appropriate KBs in all the top 5 bloggers. The results are shown in Figure 5. The average precision of 0.91 indicates most of the bloggers highly ranked by our method are exactly knowledgeable.

5.3 Blog Entries' Precision

For a search keyword and a KBG, the top 5 blog entries in the ranking list of each KBG were browsed by the four individuals. If a blog entry was related to the search keyword and KBG in question, and also regarded as credible, it was marked as appropriate. The precision is the ratio of the number of appropriate ones to the top blog entries. Figure 6 shows the precision considering the top 3 KBGs. In this case, for a search keyword, 15 entries (5 entries in each of 3 KBGs) were evaluated. Figure 7 shows the precision in the case that the 25 entries in the top 5 KBGs were evaluated. The average precision is about 75% and 67%, which are high in the field of blog search. This indicates that our method can extract related and credible entries.

6 Conclusions and Future Work

We proposed a credibilty-focused blog ranking method based on bloggers' knowledge level. The results of this study include: (1) a relationship between knowledgeable people and credible information, (2) a method for estimating bloggers'

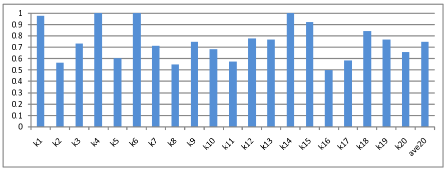

Fig. 6. Blog entries' precision for the top 3 KBGs

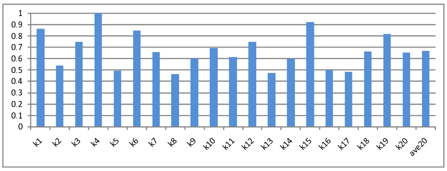

Fig. 7. Blog entries' precision for the top 5 KBGs

knowledge level based on their usage of domain-specific words in his or her past blog entries, (3) a prototype system of blog ranking based on bloggers' knowledge scores, which can provide multiple ranking lists for various topic areas.

In future work, we plan to improve our system by developing more powerful methods for filtering spam, further improving the method for calculating bloggers' knowledge scores, and investigating more aspects for providing information with high credibility.

Acknowledgments

This work was supported in part by the National Institute of Information and Communications Technology, and MEXT (Grant-in-Aid for Young Scientists (B) #20700089).

References

1. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 30(1-7), 107–117 (1998)
2. Google Blog Search, `http://blogsearch.google.co.jp/`
3. Technorati, `http://www.technorati.jp/`
4. Fujimura, K., Inoue, T., Sugisaki, M.: The EigenRumor Alogorithm for Ranking Blogs. In: WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2005)
5. Kritikopoulos, A., Sideri, M., Varlamis, I.: BlogRank: Ranking Weblogs Based on Connectivity and Similarity Features. In: AAA-IDEA (2006)
6. Gil, Y., Artz, D.: Towards Content Trust of Web Resources. In: WWW, pp. 565–574 (2006)
7. Thomas Adler, B., de Alfaro, L.: A Content-driven Reputation System for the Wikipedia. In: WWW, pp. 261–270 (2007)
8. Andersen, R., Borgs, C., Chayes, J.T., Feige, U., Flaxman, A.D., Kalai, A., Mirrokni, V.S., Tennenholtz, M.: Trust-based Recommendation Systems: An Axiomatic Approach. In: WWW, pp. 199–208 (2008)
9. Kizasi Blog Search, `http://kizasi.jp/labo/fansearch/index.py`