

Blog Map: Social Similarity of Blogs

Meng-Yang Lee, Shiang-Wen Cheng and Rung-Huei Liang

Dept. of Computer and Communication Engineering

Ming Chuan University

Taoyuan, Taiwan

Crazyion2@gmail.com zheng.koobe@gmail.com liang@mcu.edu.tw

ABSTRACT

Blogs become more and more popular in the recent years. However, there is not an effective tool providing us to find blogs to read. Bloggers have done much effort in searching and recommending valuable blogs. If we can treat blogs as instances of people in the virtual world, we can use the idea of interpersonal-attraction and take the advantage of these bloggers. In this paper, we use the metaphor of a map to visualize social distances of blogs. Transforming interpersonal- attraction and the computed blog similarity into clear spatial relation, we can mine socio-similar blogs easily.

Keywords: Social computing, blog, visualization.

1. Introduction

More and more people use blogs to broadcast their ideas, thinking, and what they discovered in the real or virtual space to the world. According to the report of Technorati¹, there are 50 million blogs in the world. In such large amount of blogs, is there any technology helping us to find blogs which fit our needs? And is it really help?

If we try to recall how we find a good blog in traditional way, you will find out that we often trust someone's recommendation. Therefore, why don't we use this idea in blog search area?

Web search engines divide a blog into pages. If you just want to find an answer of something, this technology was proved that it really works. However, if you want to find a blog to read, it can't help you.

In this research, we will visualize our results (Figure 1). If we use the idea of traditional social visualization, the scene may be look like this (Figure 2). Orange-colored blogs inside the white circle are more similar to the centered blog. We think that not only the distance, the radius, is significant but also two coordinates are.

¹ <http://www.technorati.com>



Figure 1. The scene of blog map.

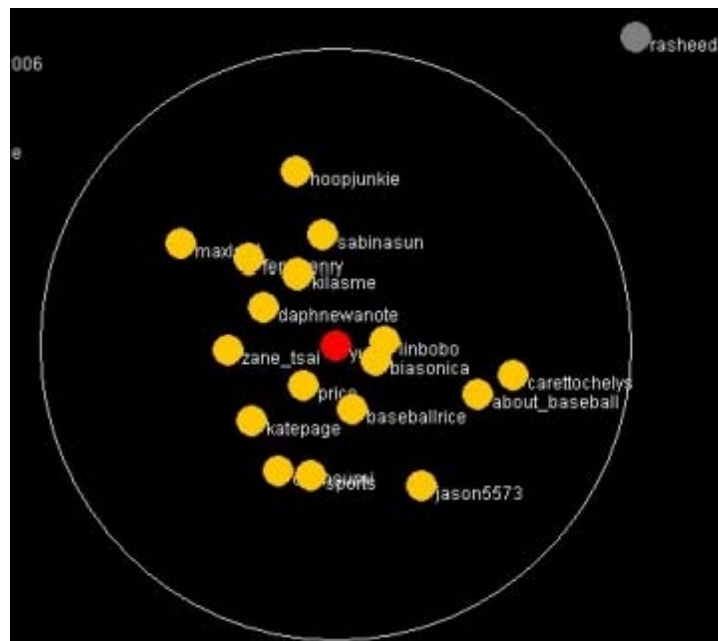


Figure 2. Traditional social visualization.

2. Background

In this section, we will describe methods of current blog-search engines. Most of blog search engines focus on articles instead of blogs. Hence if you want to find some blogs you're interested in by these blog search engines, you must pick up some good articles and then read the recent articles of the blogs. After that, you finally can decide if you want to subscribe this blog or not.

All of blog search engines use keywords to classify blogs. Words sometimes can't describe the meaning of something very precisely. Thus, it must have some semantic problems in the essence of using keywords.

2.1. Blog Search Engine by Google

Google's blog search engine² focuses on articles, instead of blogs. You can't get the assistance as much as the Google web search engine. They think that the titles of blogs are the most important part, and therefore, most of titles of the relative blogs contain the same keywords. However, as we all know, this hypothesis is not strong enough. Since Google focuses on articles, you can't get much help in searching blogs.

2.2. Blog Search Engine by Technorati

Technorati thinks differently. They take advantage of the effort of the readers. They provide a system which allows readers to tag blogs with keywords. Thus, others can use keywords to find matched blogs which are tagged with the same keywords.

This way avoids the misunderstanding of the articles by keywords measurement, and provides an overview of the blog. Nevertheless, using keywords to describe blog still loses some information to readers. Although it avoids some drawbacks of keyword-search, we still only can get the information of the popularity. Is every blog you like in the top 100? The answer may be not. Therefore, this attribute is not strong enough.

2.3. Blog Search Engine by BlogPulse

BlogPulse³ can compare links to find out which blogs cite the same links within the blog you keyed in. However, it ignores or mixes some import links; blogrolls, for example. This function does help sometime, but it needs large amounts of links, because it only counts the links in articles. In this research we think that even if a blog has few links, we still can find some similar blogs.

3. Methodology

3.1. Research problems

Whether a blog fits our needs or not is nothing about the frequency of keywords, which may be about the popularity. Since the popularity is not absolute, we are much interested in **how to calculate the similarity of blogs?**

Blogs are different from web pages. In general, blogs are a set of web pages which follow special structures. Many users think blogs as an information source. Since a blog is an

² <http://blogsearch.google.com>

³ <http://www.blogpulse.com>

information source, you won't want all your information focusing in only one area, just like books you read. Bloggers will often recommend some good blogs within which their contents might be very different, and these blogs must have some good feature worthy of reading. Therefore, **could we stand on the shoulders of giants by reading blogs recommended by experts?**

Blogs become much more easy to use, so that many bloggers update their blogs frequently. Blogs also contain many functions that could reveal lately interest of bloggers, i.e., blogroll. For these reasons, a blog can be thought as an **instance of person** in the virtual world. Therefore, the problem of finding blogs you like can be reduced to the problem of finding people you like. How can you find people you like? Based on recommendations of the people we trust, we often get satisfactory result.

3.2. Interpersonal-attraction

It is said that "Birds of a feather really do flock together." In the social psychology, it is proved that the people like the ones who are similar to them. We called that "interpersonal-attraction." We are often attracted to some people who have attitudinal similarity of us, because these people will give positive feedbacks than negative ones. This phenomenon is known as the matching hypothesis 2.

Nevertheless, this doesn't mean that opposites cannot find affinity. Sometimes we are attracted to the complementary others. For example, although we can't play musical instruments, we still might be attracted to others who can play. Moreover, if you like some people, you might like their friends.

In the blogosphere, interpersonal-attraction is presented by the action of citing links, and especially, these links in the blogrolls.

3.3. Links Analysis

Links play the important roles in blog matching. For example, BlogPulse compares the links to find the similar blogs. Links are more precise than keywords to determine interests of a blog. Generally, blogs have the common structure, although blogs can be present in other forms.

Links in a blog can be classified into several categories:

1. Links in the blogroll.
2. Links in the articles
3. Links in the comments

Links in blogroll are the strongest part which could represent as the trend of the blogger. Bloggers will often put their friends' blogs or some valuable links there. Another type of links is in the articles while they disappear very quickly by frequent update. On the contrary,

blogrolls are more stable and strong. Therefore, if a blogger puts a blog link in his blogroll, we can say that these two blogs have some strong similarity.

Eytan Adar et al.³ discover that if two blogs link each other, the probability of citing the same links by these two blogs is higher than those two blogs which don't link each other (Figure 3). They call this probability of citing the same links as the similarity of blogs.

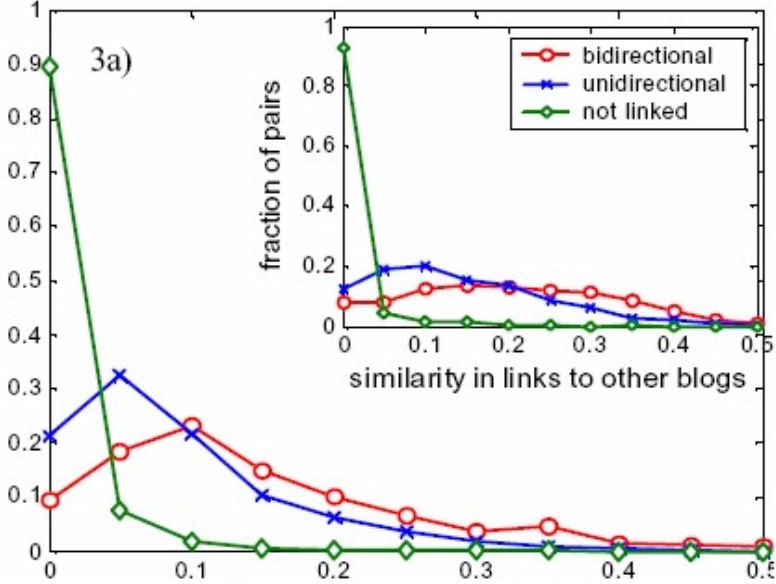


Figure 3. Similarity in links to non-blog URLs. 3

3.4. Blog-to-Blog instead of keyword- to-keyword

If we want to find some data of “Java”, we can type “Java” into Google. However, a search engine doesn't know what we mean is about program or coffee? If we feel confused, Google does, too. The key point is that if you use keywords to search, what you get is specified. If the result is mixed up within two different areas, it is the ambiguity of keywords.

It works if you want to find some answers, but if you want to find some blogs to read, it helps less. The first reason is that results are always sorted by popularity. Unless you can check top 100 or more results, this order helps you little, because the top blogs are very stable. Nevertheless, there are many blogs worthy to read, even if their ranks are more than 1000.

It is said that tools limit your thinking, and so the way, keyword-to-keyword, does. All of tools use keywords, and hence users are hard to image that they can find information sources by blog-to-blog just as they do in the real word. In the real world, we open our mind by talking with our friends, teachers, and etc to get recommendations and suggestions. What we get is not only focus on a fixed area. Implementing this idea to blog search engine, it must help us to expand our information sources.

When you type a blog and want to find other blogs by this key blog, it means that you like or trust the information this blog provide you. Once again, if you want to find some answers, try keyword-search. Otherwise, our similarity matching might work better.

4. Implementation

4.1. Research restrictions

There is not any API that can help us to retrieve accurately every part of blogs. However, if we focus on a specific BSP(Blog Service provider), we can almost retrieve every part. In this research, we choose yam-roodo BSP⁴, because it only allows users to modify the CSS, and all of themes it supports have the common structure. Although we retrieve all links, we still can't determine which links are blogs unless these links are the blogs in the yam-roodo.

4.2. Evaluation of the similarity of blogs

In this research, we divide the similarity into two parts: the similarity of links, and the similarity of readers. The similarity of links is to determine the similarity of content. The other attribute could tell users which blogs have the same readers as the key blog does. This attribute can help users to expand their information source.

4.3. Candidates in blog map

The candidates could be divided into two levels. This is decided by the relation of the key blog with them.

1. A User selects a key blog in yam-roodo.
2. If yam-roodo blogs are listed in the blogroll of key blog, put these blogs into Candidate1.
3. If those in yam-roodo blogs have at least one same blogroll as key blog does, put these blogs into Candidate1.
4. Put blogs which put the key blog in their blogroll into Candidate1.
5. Put blogs which cite the same links as the key blog does into Candidate1.
6. Put blogs which have at least the one same reader or trackback as the key blog does into Candidate1.
7. Retrieve the blogs which appear in the blogroll of a blog in Candidate1, and put them into Candidate2.

When we get all candidates, we can evaluate their ranks of the similarity, and then, draw them in the scene.

4.4. Similarity of links

Hypothesis 1:

If two articles cite the same links, we can say that their content must have some common points.

⁴ <http://blog.yam.com>

Nevertheless, this hypothesis is not enough to describe the phenomenon of links. Looking at the link structure carefully, you will find out that links contain some type of the hierarchy. It means that www.nytimes.com and www.nytimes.com/pages/arts/ must have some relation and it really does. Properly, this relation is weaker than the fixed links. Therefore, here is another hypothesis.

Hypothesis 2:

If links in the articles have the same domain name, we can say that content of two articles have some common points.

Base on these two hypotheses, we can model the relation of links in two articles of different blogs. The relation of links in two blogs could be modeled as:

$$Rank(A, B) = \left(\begin{array}{l} \frac{(freq(BR_1) + freq(BR_2) + \dots + freq(BR_n)) * C_1}{\min(|BR_A|, |BR_B|)} \\ + \frac{(freq(L_1) + freq(L_2) + \dots + freq(L_n)) * C_2}{\min(|L_A|, |L_B|)} \\ + \frac{(freq(FL_1) + freq(FL_2) + \dots + freq(FL_n)) * C_3}{\min(|FL_A|, |FL_B|)} \end{array} \right) * W_{ci}$$

BR_n means the series of the same blogroll.

L_n means the series of the same links which appear in the articles.

FL_n means the series of the same Domain Name which appear in the articles.

$freq(x)$ means the inverse of the times of link-x appear in all candidates.

C_1, C_2, C_3 are constants which are used to weight the different parts.

W_{ci} is a constant which is used to weight the blogs.

The $freq$ of links are very important. It can reveal how important a link is. There are not only links which bloggers support will exist in the blogroll. Blogroll also contain some other links, like counters or some tools.

4.5. Similarity of Readers

This research will also consider the readers who left comments in blogs. Most of the search engines don't consider this part or mix it with the blog entries. Readers are very important parts while search engines don't any pay attention.

What can comments do? Image that: there are two good blogs but their contents are different, i.e. one is about IT, and the other is about photograph. However, they both have the same readers, and in this situation, it means that many readers are interested in both these two blogs. We implement this idea, since our goal is to find the good blog to read.

Thus, the rank of readers could be model as:

$$Rank(A, B) = \left(\frac{(freq(R_1) + freq(R_2) + \dots + freq(R_n))}{\min(|R_A|, |R_B|)} \right) * W_{Ci}$$

R_i means the i th reader.

$Freq(x)$ means the inverse of the times of reader- x appear in all candidates.

In this research, readers mean that some people left their emails, blogs, or website in the comments. To preventing the false readers, we will ignore the authors of comments who only left their names.

5. Results

5.1. Statistics

We currently have clawed 3344 yam-roodo blogs from June 15, 2006 to July 10, 2006, and saved all their links in the mySQL database. There are total 520,590 links in our database.

5.2. User Interface

In this research, we use a map of the town to represent the key blog and candidates. X-coordinate indicates the similarity of links and Y-coordinate indicates the similarity of readers (Figure 4). We put a house in the left-top corner to represent the key blog, and we group the other candidates into different sizes of building. If users want to know about more detail, they can click the house and then the list of blogs is shown (Figure 5).

Notice that Candidate blogs at location A and B in Figure.4 are worthy of further discussion. Blogs at location A share many similar links with the key blog while they don't have many common readers. This might indicate that these blogs focus on similar topics by different groups of readers. For some reason, these groups of people might not know each other who share such similar interests. Thus, our work can help a reader to reach another group of people who potentially want to know each other. On the other hand, blogs at location B and the key blog are read by almost the same readers although these blogs focus on much different topics. A reader can expand his interest by finding out blogs read by others in the same reader group. Nevertheless, blogs at location C are the most similar blogs with readers who might already know each other.

5.3. Conclusion and Discussion

Bloggers often read and filter information they read, and then they will post on their blogs, and they do much effort on recommendation. If blog search engines could take advantage of their effort, the result would be better than just counting the keywords.

Blogs are very different from traditional web pages. Therefore we need to use different thinking to search blogs. Using blog-to-blog can help us to expand our information sources.



Figure 4. The UI of blog map.



Figure 5. The blog list of blog map.

Our Algorithm doesn't tend to replace the keyword-searching. This Algorithm tends to provide another thinking that can help users to find out more blogs to read. However, it is pity that there is not a common interface which can let us retrieve every part of a blog precisely. We hope that it will come true someday, and then this algorithm can be applied to more blogs.

REFERENCES

1. Dave Sifry (2006), In Sifry's Alerts, Retrieved from <http://www.sifry.com/alerts/archives/000432.html>, June 7, 2006.
2. Stephen L. Franzoi (2005), Social Psychology 4th edition, McGRAW.HILL.
3. Eytan Adar and Lada A. Adamic (2005), Tracking Information Epidemics in Blogspace, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05).

部落格地圖：部落格上的社交相似性

李孟陽, 鄭翔文 和 梁容輝

銘傳大學資訊傳播工程系所

Crazyion2@gmail.com zheng.koobe@gmail.com liang@mcu.edu.tw

摘要

本論文針對部落格的搜尋、比對以及分類問題提出了一套新的分類思考。此方法將部落格視為個人在網路上的實體，基於有機體的社交性和物以類聚的概念找出與關鍵部落格相近且適合讀者閱讀的部落格。

部落格會不斷地與網路世界融合產生新的面貌。這面貌最顯著的表現方式就是部落格中所包含的連結，其展現了部落格作者當下的意志。而部落格與部落格之間相連性更是一種強化的社交網絡。

本方式並不透過關鍵字找關鍵字的無機式思考，改採透過部落格的有機式思考。