USER ASSISTED KEYWORD SEARCH ON RSS SYHTHESIZER FOR RELEVANT WEBLOGS

TEH PHOEY LEE, DR. AZIM A.A.G., DR HAMIDAH I., DR RODZIAH A.

Faculty of Computer Science and Information Technology
University Putra Malaysia
phoeylee@hotmail.com, azim@fsktm.upm.edu.my, hamidah@fsktm.upm.edu.my, rodziah@fsktm.upm.edu.my

ABSTRACT

Information transfers through web has plays a significant role in the utilization of its resources. The major parts that involve in information transfer recently are through sharing and publishing weblogs. Weblogs that normally published in chronological order by bloggers mostly attracts the frequent returned blogs readers unless it is shared in a public new feeds or perhaps reach from the other platforms. Although RSS aggregators able to aggregate the latest update in timely manner, weblogs that was written in chronological order by multiple bloggers have caused the conflicting of result aggregated from different weblogs ended up an irrelevant or unrelated lists of weblogs collection. It is refers mainly due to the ambiguity of synonym (different but similar meaning) and polysemous (similar but multiple meaning) words used by different bloggers. The orientation of this paper is at first to review the techniques used on the weblogs search options, or any other available assistants to reach a weblog over the internet. Continue with an explanation on the ambiguity problem among the weblogs. Next, identify if any RSS aggregator comes with a solution. Besides reviewing the available techniques on relevancy measurement, an experiment is performed to test if user assisting in keyword search able to improve the aggregated results to a more relevant of collection. Subsequently, the result concluded a satisfactory list of weblogs collected after benchmarking the list of weblogs collected from Feed Demon. Paper ends with the future works.

Keywords: *RSS, weblogs, tag, folksonomy, keyword*

1.0 INTRODUCTION

A weblog (blog) is a simplified form of web publishing that allows anyone with a computer and internet connection to post content online (Gurzick 2006). It often provides commentary or news on a particular subject, such as food, politics, or local news. Some function as more personal online diaries (Qamra 2006). Blogs are dated, unedited, highly opinionated personal online commentary including hyperlinks to others resources. They are reverse chronological sequences of journal-like entries, maintained and published on the web (Li 2007). Blogs are difference from e-forum whereby from the definition of oxford dictionary, forum is a meeting or medium for an exchange of views. Categorized of all types of blogs has been done by Teh (2007).

RSS. The Rich Site Summary, also known as Really Simple Syndication or RDF Site Summary Format, is an eXtensible Markup Language (XML) of text-based format for sharing and distributing web content to the RSS readers. RSS allows the syndication of list of hyperlinks, which work together with other information or metadata (Jose 2005) that used to publish frequently updated digital content, such as news feeds, podcasts or blogs. It was originally uses for subscribing news headlines, to provide up to date information and news summaries (UIC 2000). RSS aggregators or also named as RSS readers are used to collect the updated digital content from the sites users subscribed.

Next part of this paper is starts with a review on the available assistants in weblogs search options or the techniques used to reach a certain weblogs. Part 2 investigates if any aggregators or others platforms solve the word ambiguity problem. Part 3 identifies the techniques used in relevancy measurement, and the last part discussed the experiment and evaluation.

2.0 WHAT ASSISTING USERS SEARCH FOR A WEBLOG?

This question included answering "How users normally reach to a weblog" and "How do they search to link to a weblog". Furthermore "Are the weblogs that users collected relevant to the search of topic?" or "It is just a random hit?" The reviews on this part mainly discuss the all the features and functions available on those weblogs "search" and "reach" options. There are few options assisting users.

2.1 RSS Aggregators

RSS feed has become a "should-have" function among the weblogs. The reason is to assist user returns to the weblogs without "memorising" the URL (Universal Resource Locator), "searching" through the web browsers, or "bookmarking" on favourites for the updated. Users just need to read the headline on the RSS aggregators before deciding to read the weblogs in details. Many online information sources, including web sites and news services, are broadcasting content to the web using "syndicated feeds" or "news feeds". Most of the weblogs and some social networking sites that offer weblogs features are actively using RSS feeds to assist user reading weblogs. For social networking sites, there mostly collaborated the weblog feature with other applications to a shared-feed. Table 1 listed several weblogs sites and social networking sites that using RSS feeds.

Table 1. Weblogs and Social Networking Sites with Weblogs using RSS feeds.

Sites for Weblogs	Type
technorati.com	-
blogger.com	+
friendster.com	-
livejournal.com	+
multiply.com	-
typepad.com	+
wordpress.org	+
myspace.com	+
blogs.com	+
antzblog.com	+
twitter.com/blogmarks	+

For sites that purely provided for weblogs sharing, it is marked with type (+), for those social networking sites which offer weblogs applications it is mark with (-). The differences between them are that for social networking blogs, users may need to use the aggregators provided from the specific sites due to the RSS share-feed with other application's updates.

For the (+) type, blogs readers may at first subscribe to RSS aggregators. There are two platforms designed for RSS aggregators, which consists of browser-based (online) and client-based (desktop) aggregators. Aggregators are dedicated programs which allow users to read RSS files (Fagan Finder, 2004). Heinz (2007) has graded the top 10 used browser-based and client-based aggregators in year 2007. Table 2 listed the Top 10 highly used RSS aggregators.

Table 2. Top 10 highly used RSS Aggregators

Top 10	Browser-based	Client-based
1	GoogleReader	FeedDemon
2	iGoogle	OmeaReade
3	Netvibes	Awasu
4	Bloglines	Snarfer
5	Newsgator	RSS Bandit
6	Feeds 2.0	RSSOwl
7	Rojo	BottomFeeder
8	NewsAlloy	BlogBridge
9	MySindicaat	PixelNews
10	Fwicki	AgileRSS

Referring to the statistic of Heinz (2007), Google Readers stand the highest point for browser based RSS aggregator while Feed Demon stands the highest point for client-based aggregators. Although Bloglines.com was the older aggregator then Google Reader from the online community, Google Reader still won the highest points because of the Google tradition of allowing users to do things very cleverly, quickly and easily. (Mark ,F. 2003).

2.2 Tagging and Tag Cloud

Tag clouds are visually-weighted renditions of collections of words (tags) that can be used to represent the concepts present in large collections of information. It is a very useful tool for classifying items and can retrieve weblogs easily by just clicking on a keyword.

However, question arises on how tags can be utilized and how will it produce results to users that searches for the items? Today, social networking sites such as Flickr uses collaborative tagging method to allow users tag their content and also recommending tags to other users that uses their service; therefore, collaborative tagging plays an important role in today's social services.

Some weblogs use tag clouds on their site to assist readers reaching the a relevant weblogs, for examples, BlogMarks, Technorati and etc. A sample of Technorati tags are shown in the Figure 1.

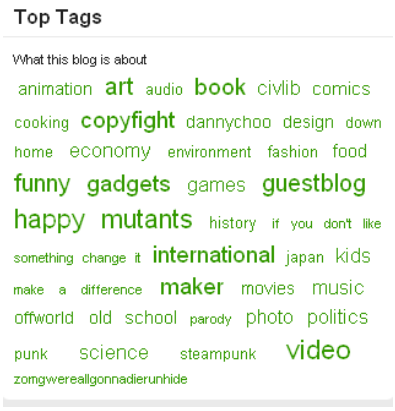

Figure 1. Technorati's Keyword Tag Cloud

The concept of tag is similar to folksonomy. The word Folksonomy has been used as a system of classification derived from the method of collaboratively creating and managing tags to annotate and categorize content. This technique helps in assisting user search with the pre-categorized tags available, working as social bookmarking on mainly photograph annotation.

Since year 2004, tagging was successfully used mainly on most of the popular site as the 'reach' functions. For instance, Facebook the most popular social networking site on the Internet supported the photo tagging. YouTube a video sharing website that allows words tag. The del.icio.us allows users tagging the bookmarks as they surf along the Internet, this will guide users reaching the site. All the tag can be categorized and listed as the content for user's references. Is this technique increases the "hit" numbers on the relevant "story" beside the traditional keyword search functions? This issue is remains to be discussed.

2.3 Traditional Web Search

This is the traditional search functions. Users just need to go to a search engine and type in a keyword onto the search function to reach a certain topic of weblogs.

There are many available internet web browsers search tools that readers may use to reach a weblog. Table 3 summarizes several types of search engines available to search for weblogs.

Table 3. Search Engine to Search for Weblogs

Name	Sites	Type
Google	Google.com	A
Yahoo	Yahoo.com	A
Windows Life Search (MSN)	Live.com	A
All The Web	Alltheweb.com	A
AltaVista	Altavista.com	A
AOL Search	Aolsearch.aol.com	A
Ask Jeeves	Ask.com	A
Nescape	Search.netscape.com	A
gigablast	Gigablast.com	A
Open Directory	Dmoz.org	A
HotBot	Hotbot.com	B
Dogpile	Dogpile.com	B
Padia Power search	Pandia.com/powersearch/	B
Search	Search.com	B
Psychcrawler	Psycherawler.com	C
Political Infromation	Politicalinformation.com	C
Inomics	Inomics.com	C

The types of the search engine are categories as A, the general purpose type, using crawler based or directory based. B, the Meta type, the 'search' are send to several other search engines simultaneously and weblogs are aggregated in one site. And C, the Topic specific type, the sites search has its specialized topic of information, for instance, a site only return search on psychology, politics or just economics weblogs.

The search options by most of the search engine have its own search syntax making their link retrieves by users. Mohanty and Chudamani (2008) compared Googles and Yahoo web search on a search term. Googles assisting the search options include the following options; 1) goes directly to top ranked site in query, 2) bring up list of related

sites, 3) searches by govt. and milsites, 4) search within results, 5) field searching with link only. While yahoo labeled 'search options' by 1) all features by main page search, 2) subject categories, 3) date range searches and etc. All of these functions are assisting users with the general search options.

2.4 Universal Resource Locator (URL)

This is the most common way to reach a weblogs. Users might just need to type the URL given, send, publish, display or link from bloggers or from other platforms. URL might be distributed from a chat, e-mail, link from a weblog of friends or perhaps Shorts Message Service (SMS). Some weblog writers use their own domain name from their own server, for instance skycoral.com or yourname.org. These options benefited them. These bloggers can also add-on the RSS feeds to attract a return reader. Somehow, they advertise themselves by increasing the commercial hit rate onto their weblogs if their weblogs are meant for that.

3.0 WHAT WENT WRONG?

The relevancies of the weblogs collected are from RSS aggregators are found not semantically related due to its chronological order of publication. Golder and Huberman (2006) stated that tagging systems (also known as folksonomies) are also beset by many problems that exist as a result of the necessarily imperfect, yet natural and evolving process of creating semantic relations between words and their referents.

The problems that have been identified are *polysemy*, *synonymy* and *basic level variation*. These are usually the main problems that occur not only in RSS aggregators, tagging systems but also in normal search engine. The explanations of the problems are described below:

3.1 Polysemy

A polysemous word is one that has many related senses. For example, "mouse" may refer to the mammal or pointing device of the computer. Which is a word that has multiple and unrelated meanings. In practice, polysemy limits query results by returning related but potentially inapplicable items.

3.2 Synonymy

This refers to multiple words having the same or closely related meanings. It is a greater problem in tagging systems as inconsistency among the terms used in tagging can make it very difficult for one to be sure that all the relevant terms has been found. For example, when a user wants to look for the keyword "television", tags can be either classified as "television" or "tv", therefore both tags will also have different results of query, and this confuse the users on whether they should search "television" or "tv" to get the results. Further examples include plurals or part of a sentence in the item (e.g. cat and cats). In this case, taggers will either need to widely agree on a convention, or add on multiple or more complex queries to cover many possibilities.

3.3 Basic level variation

Related terms that describe an item vary along a continuum of specificity ranging from general to specific. Some people may consider terms at different levels of specificity to be most useful or appropriate for describing the item in question. For example, a person might consider "MP4 players" as basic and "iPod Touch" as specific. The underlying factor behind this variation may be that basic levels vary in specificity to the degree that such specificity makes a difference in the lives of an individual. Therefore, conflicting basic levels can also prove disastrous, whereby some tagged terms may seem too specific for a user or vice versa.

Despite all that we have discussed above are all the platforms that provided with functions that assisting users reaching their weblogs. Are the weblogs that users reach really related to a relevant topic user wanted? If users become the focus in assisting in the search functions, will it improve the relevant hits of weblogs based on user interest?

4.0 KEI AND KEYWORD RELEVANCY

KEI (Keyword Effectiveness Indicator) is a technique used in search engine marketing Angel (2008). A measure composed of a constant factor and a simple division is used.

KEI = (Search Count for Term ^ factor) / Result Count for Term

The search count is a count of how often a word is entered as a search term. The Result Count is the count of the returned result. Thus the higher the search counts the higher the KEI. Unfortunately, there are no shorthand mathematical tricks for establishing relevancy.

Angel (2008) established a business-specific measure of word relevance. The collection of terms that hit a certain site was calculated and the terms with the highest score was collected from potential customers. An overall relevancy weighted score from all sources are stored to be used with the future search. The whole processes were complicated that user passes for keyword discovery, but the idea was exactly getting the users "search" terms to assist the obtaining of the relevant result. T

Brooks (1998) stated that the unresolved state of theorizing about relevance inhibits the conduct of empirical experimental. Several typical of experiments were discussed in his paper. First was by constructed a descriptor tree or hierarchy, a search that located a record captured by the top terms in a hierarchy are measured and checked. Another type of experiment is by using sliding bars that permitted subjects to express degrees of relevance.

Based both Angel (2008) and Brooks (1998)'s idea, another experiment is performed synthesizing the RSS aggregators to identify if the user assisted in keyword search may improve the relevant weblogs count.

5.0 USER ASSISTED KEYWORD SEARCH

An experiment was performed to identify if allowing users assisting in keyword search will gain a better relevant weblogs on the search result. This idea is built to the synthesis of RSS aggregators based on Angel (2008)'s idea. Synthesizer is built with the add-on thesaurus stored in the database assisting in the search functions. Users will need to provide their own keywords on similar or relevant keyword in assisting the search of weblogs aggregated onto the RSS aggregators.

Before the experiment, 100 questionnaires were distributed to collect keywords from users. Weblogs are commonly written in user-defined-language, or perhaps dialect. This is the reason a certain keywords collected could not be found from a normal dictionary or thesaurus corpus. In order to reduce the topic coverage, the collection of keywords or terms on a specific topic of weblogs are collected, categorized, and stored in the synthesizer, integrated in the search process.

This experiment was participated by 20 computer science students and 10 Computer Science lecturers. Most of them have at least a 2 years experience in reading weblogs and 80% of them using RSS aggregator. The results obtained in the search are benchmarking with the top used desktop-based Feed Demon.

The variables for the measurements from the experiment's feedback are listed below:

- Total numbers of links returned for each category, (Tt).
- Total numbers of relevant links that user found relevant for each keyword, (Tr).
- Rate of satisfaction for each of the result measured in five-point Likert scale (one is strongly not satisfied and five is strongly satisfied).
- Total numbers of links return by Feed Demon, (Td).
- Frequency of Keywords used, (Fk).
- Average of total numbers of relevant link that user found relevant. (Ta = Tr + Tr +... Tn/ n).
- Relevancy (Ta/Tt).

The relevancy measurement was conducted. Results obtained on the opinion of relevancy of participant are analyzed in next section.

6.0 RESULTS

The result in Table 4 shown the overall results obtain based on the categories of foods. Each category was assisted with the keywords provided by users that were stored in the database. Satisfaction on the highly returned relevance result fallen on *dessert* group. It is averagely rated as 4.07 of satisfaction. The *Malaysian foods* also returned a high relevant result. It is rated with 4 point satisfaction.

Table 4. Results of the Relevancy and Rate of Satisfaction from the Experiment

	Relevancy Ta/Tt		**Rate of Satisfaction on Likert Scale from experiment (Mean)**
	Experiment	**Feed Demon**	
Noodles	0.994	0.932	3.32
Seafood	0.976	0.960	3.13
Sauces	0.573	0.853	3.47
Desserts	0.975	0.906	4.07
Fast Foods	0.962	0.937	3.80
Breads	0.965	0.89	3.13

Japanese Foods	0.999	0.378	3.47
Malaysian Foods	0.996	0.961	4.00
Western Foods	0.765	0.934	3.17
Drinks	0.993	0.844	3.35

Comparing to Feed Demon which do not come with the "users assisted keyword search option", the overall results shows an increasing of average of relevancy from the "users assisted keyword search option" collected result.

7.0 CONCLUSION AND FUTURE WORKS

The experiment has illustrated in improving the relevant search result by user assisted in keyword search compare to Feed Demon. It has proven that by giving and extra feature applying to RSS aggregator or any other platform, it may helps users obtained a better relevant weblogs.

With the similar idea, tag cloud is to apply the similar concept, based on user interest; displaying the tag cloud from a batch of relevant words collected by users. It is predicted to increase the speed of search with visualization weighted-tag cloud for instance, bigger text for higher relevant of keywords being weight by users. Similar experiment can also be performed with some application, for instance, for vegetarian groups of people, allergic patient, medical food cattery systems and etc.

REFERENCES:

[1]. Angel, G. (2008). KEI and Keyword Relevancy. SEMphonic.

[2]. Brooks, T.A (1998) Distance Model of Relevance Assessment. University of Washington, Seattle,

[3]. Fagan, F. (2004). RSS aggregator, *http://www.faganfinder.com/search/rss.php.*

[4]. Golder, S.A and Huberman, B.A. (2006). The Structure of Collaborative Tagging Systems. Springer.

[5]. Gurzick, D., Lutters W. G. (2006). From the personal to the profound: Understanding the blog life cycle. ACM.

[6]. Heinz, T. (2007). About.com: Top 10 Windows RSS Feed Readers and News. *http://www.about.com*

[7]. Jose, A. C. Alessandra, A. M. (2005). A Software Infrastructure for RSS Deployment and Linking on the web. Proceedings 11th Brazilian Symposium on Multimedia and the web. vol 125. ACM Portal. New York, NY, USA.

[8]. Li, B.B., Xu, S.T., Zhang, J. (2007). Enhancing clustering blog documents by utilizing author/reader comments". ACM Press.

[9]. Mark, F. (2003). bloglines.com *http://www.bloglines.com*

[10]. Mohanty R. and Chudamani K.S. (2008). A Comparative Study of Google and Yahoo Web resources on the Search term "Physics India". 6th international CALIBER – 2008, University of Allahabad, Allahabad.

[11]. Qamra A., Tseng B., Chang E.,Y., (2006) "Mining Blog Stories Using Community-Based and Temporal Clustering" ACM Press.

[12]. Teh P.L. (2007). An Insight of blogopshere over the Internet. Journal for the Advancement of Science & Arts. Volume III, page: 65 – 68.

[13]. UIC University of Illinous at Chicago (2000). RSS at UIC: RSS Standards. *http://www.uic.edu/depts/accc/ecomm/rss/rss.html#2.*