International Journal of u- and e- Service, Science and Technology
Vol.8, No. 6 (2015), pp.45-54
http://dx.doi.org/10.14257/ijunesst.2015.8.6.05

An Implementation of Intelligent Searching and Curating Technique on Blog Web 2.0 Tool

Harsh Khatter[1], Munesh C. Trivedi[2] and Brij Mohan Kalra[3]

Department of Computer Science
[1,2]ABES Engineering College, Ghaziabad, India
[3]Ajay Kumar Garg Engineering College, Ghaziabad, India
harsh.khatter@abes.ac.in, munesh.trivedi@abes.ac.in, brijkalra@yahoo.com

Abstract

There are numbers of web tools available to spread and explore information on internet. With the increasing use of Blogging sites, people are able to share their opinions, interests, experiences, and their views with others. However, it is not so easy to fetch required information from various Blogs in available time, which normally is very short. From last few years our major concern is on web searching and web mining. In this paper, a searching and curating model is discussed which introduces an approach of fetching the relevant information automatically from various Blog sites and fetches the personalized required information. That exactly is the need of the hour.

Keywords*: Web tools, Blogs, Web searching, Web mining, Curation, Internet, personalize search*

1. Introduction

Blogs are one of the major components of Web 2.0 also known as a Read-Write Web. Blogs are online diaries created by individuals, which provide excellent information on any topic all over the world. In this paper, a searching and curating model is discussed, which introduces a new approach of fetching the relevant information automatically from various Blog sites.

In this paper, a new Blog model is proposed to search the required information. The proposed model includes search module, login and personalization module, content curation module and rating module. Search module aggregates data from various blog sites. Curation module selects the relevant blog information from the searched data. Curation means to select the relevant information and removes the irrelevant data from the searched data. The content curation module automatically curates the blog posts from other blog sites. This proposed model performs searching and then curation on the Blog posts based on user's interest *i.e.,* proposed model is mining blog posts based on user's interest. That is why the model is named as “Blogminer”.

This proposed Blog also works as a blog search engine, which gives blog post results from other blog sites available on World Wide Web. This new method of fetching relevant blog posts automatically called content curation, and this will improve the knowledge experience of a user and reduces the content search time, and utilization of system resources. In this paper, the gaps in current blog posts searching techniques are discussed. Further, the proposed solution with implementation is given. Hope this model will surely improve the searching and blogging experience of the user.

2. Literature Review and Gaps

Singh, *et al.,* (2010) mention that it is easy and simple to create blog posts, which has attracted people and companies across disciplines to exploit it for varied purposes. The

ISSN: 2005-4246 IJUNESST

valuable data contained in posts from a large number of users across the world provide a rich data source. Nasr and Ariffin (2008) also said that Blogging can also be seen as a means of Knowledge Sharing. They stated that the research delves into the use of blogging as a fast, up-and-coming means of knowledge and quality content sharing amongst communities of bloggers with similar interests. Dolinska (2010) discussed a simple blog searching framework and stated that Knowledge gathered on blogs can be used in personal e-Learning. It is a more informal and personal way of learning than the one is offered by traditional e-Learning courses. The framework consists of four modules: BL search tool, SNA analyzer, Network Visualizer, K-Blogs List as shown in Figure 1.

Singh and Joshi (2010) discussed that because of the increasing number of blogs and their unique characteristic, blogs have received much attention from researchers and various studies have been conducted. In addition, optimization of the mechanism is required to obtain the best results from the blogs.

If we discuss about the normal web content searching, over the past decades, various searching techniques are come into existence with the growth of World Wide Web. From the starting of the Web era, various searching methods, techniques and searching algorithms are introduced and as per searching requirement, they come in usage or implemented. There are lots of searching methods in which search can be done on keywords, on queries, on topics, on phrases, on pages, etc. The query based and topic based search is used in forums, whereas the page search or phrase search is used in search engines where the exact finding is required. Rest of all websites use keyword based searching. Keyword based searching provides an easier way to search the contents on internet. In the same way, maximum number of websites use keyword based searching.

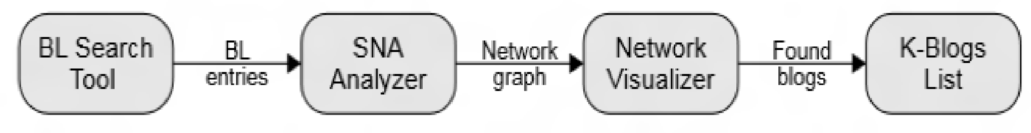


Figure 1. Simple Blog Searching Framework Scheme

Fu and Anyanwu (2011) discuss an effectively interpreting keyword queries on databases. Before this method, heuristics was used for interpreting the keyword queries. Keyword search queries might be in structured, unstructured or semi structured form. On Web, mostly the available data is in unstructured or semi structured form. Guoliang Li, *et al.,* (2008) suggested an efficient 3-in-1 keyword search method which works for all types of data. Zhou (2010) proposed an algorithm of personalized blog information retrieval based on user’s interest model. He discussed the system architecture of representation and the algorithm flow of blog document similarity based on the vector space model.

Some major blog related works are discussed above. This is an evolutionary field. Therefore, everyday a new approach, framework, or model is introduced. Might be some other ensuing algorithm or approach will come with some better idea and great functionalities.

A. Gaps observed:

After the literature survey, various gaps are observed in the current systems.

- There is a large list of available blog search engines but they all are related to specific field only, *i.e.,* they do not cover the entire variety of blog topics.
- To search a relevant post, user has to search it manually on each individual blog site, which itself will consumes lot of time, effort and a large number of clicks. Moreover, the clicks introduce unnecessary traffic due to advertisements in each page, which consumes important system resources unnecessarily.
- At present, there is no method available, which automatically traverses other websites using their URLs and fetches the required information. The important factor is

"What exactly the user wants". That exactly is the need of the hour. There is a need to add some more functionality and service to the blogs.

In terms of blog and blog post searching, numbers of methodologies are already designed and published. After studying the work done in the past in the field of Blogs, Blog architecture, model, framework, its implementation, Blog posts searching, tagging, clustering, ranking, rating, and Blog personalization, it is concluded there is a time to club all in a simplified form. The goal of this paper is to find out the measures to provide user with reliable and accurate blog information conveniently.

3. Proposed Solution

Based on the gaps of current available systems, there is a need for a new approach, which will improve user's searching and knowledge experience with the blogs. The main objective is to create a blog, which will curate the content from various blogs as per the user requirements, and displayed to the user via a blog site named BlogMiner, (Khatter and Kalra, 2012).

As per the requirement, a new method, called Curation has been proposed. Curation means to select the relevant information from the aggregated data, which is aggregated from various other blog sites. The proposed model follows two main steps: searching and then curating. In searching, data from various blog sites is searched and aggregated. The aggregated data is stored into a temporary database. Curation method filters out irrelevant data and fetches the relevant information from the searched data based on user's interest. This data is displayed to the user through a web interface after searching, aggregation and curation.

The proposed model is named BlogMiner and provides following facilities:

- Allows a user to start its own blog.
- Allows a user to perform a local search *i.e.,* search within the blogs of BlogMiner.
- Allows a user to perform a global search *i.e.,* search for blog posts on WWW.

Therefore, the BlogMiner combines the blogging and searching together at one place, which is not available in the present system. (Khatter and Kalra, 2012).

A. Working of curation module:

The curator performs the following steps:

Step 1. Checks for the blog URLs from the database, which is inserted by a user.

Step 2. Traverses each URL and uses wget utilities to retrieve the required blog posts.

Step 3. Fetches all blog posts and hyperlinks on that page and temporarily stored in a database.

Step 4. Selects top rated blog posts from the temporary database within the specified period of blog creation.

Step 5. Stores the selected blog posts into the blog post database and rest of the contents is dropped.

Step 6. Relevant blog posts are displayed to the user through Curator's Interface.

The proposed system works in the manner as discussed below and shown in Figure 2:

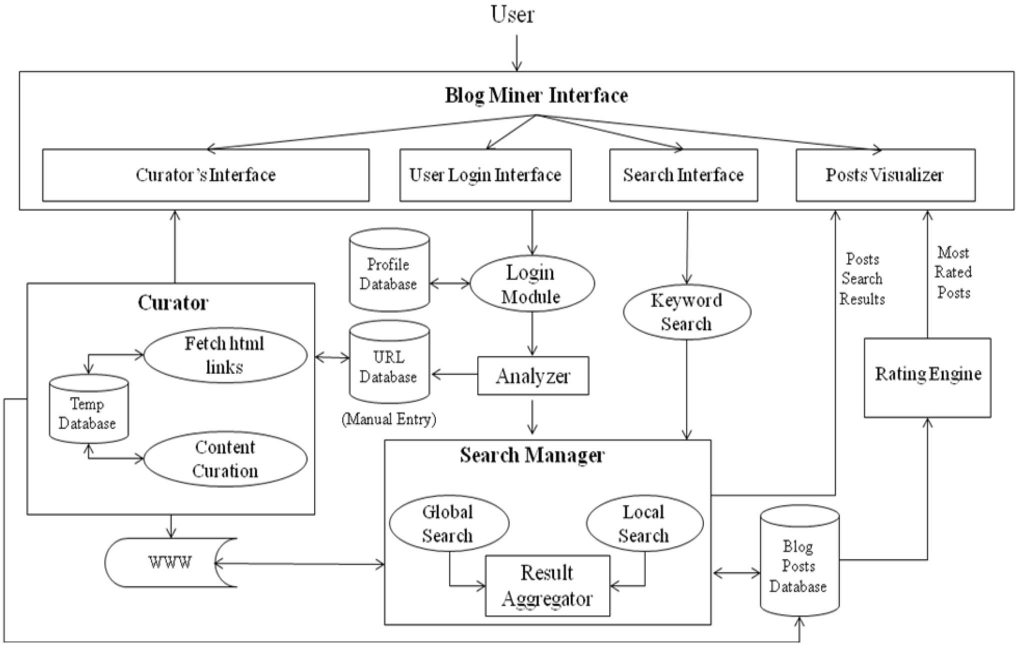


Figure 2. Model of BlogMiner

User will be presented with an opening page in which following facilities are available. User interacts with BlogMiner Interface to retrieve relevant blog posts. User enters a keyword in search bar, *i.e.,* local search bar or global search bar, to fetch the relevant blog posts. The working for local and global search is discussed separately in tabular form in Table 1.

Table 1. Working for Local/Global Search

For Local Search	For Global Search
1. Enters the keyword in local search bar for local blog posts search. **2.** Searches local blog posts for the input keywords (searches within local blog site for the relevant blog posts). **3.** Search is applied to the local blog posts database. **4.** Relevant blog posts are retrieved and presented to the user through BlogMiner Interface. **5.** Now user can read, comment, share and rate the blog posts.	**1.** Enters the keyword in global search bar to search for blog posts in blogosphere. **2.** Searches blog posts for the input keywords (Blogosphere search with the help of Google blog search). **3.** Google blog search results are presented to the user through BlogMiner Interface. **4.** When user clicks on any of the blog post of blogosphere post result, the user is redirected to the webpage of clicked blog post. **5.** User can read that blog post and share, but cannot comment and rate it.

When a user logs in, he/she is presented with blog posts of his interest only. Search Manger searches for the blog posts of user's interest from the blog posts database based on his/her specific interests. Login is of two types: Login for existing users and login for new users. For exiting users, users have to give their login id and password to perform sharing, rating, commenting, or creation of blog posts. For new users, users have to choose their username, password, and field of interests. Based on these, the registration of new users will be completed.

Curator curates the blog posts automatically. It takes the URLs from the URL Database. Periodically, Curator fetches all blog posts from to those URLs. Only the selected blog posts are stored into the blog post database and rest of the content is dropped.

User is permitted to rate the local blog posts by stars and the rating is displayed publicly to all the users in the form of stars. All the blog posts are displayed in reverse chronological order, and user can comment, rate, add and share the post information.

4. Implementation and Results

There are various parameters, which distinguish existing systems to proposed system, BlogMiner. Table 2 highlights the parameters in which existing and proposed system is same whereas Table 3 shows the parameters in which these models are different.

Table 2. Similarities between Existing and Proposed Model

Parameters	Existing system	Proposed system
Blog post search	Yes	Yes
Search engine	Yes	Yes
Top Rated	Yes	Yes
Latest Search/ Updates	Yes	Yes

Table 3. Differences between Existing and Proposed Model

Parameters	Existing system	Proposed system
Combined Search Approach (Blog + Blog Search Engine)	No	Yes
Automatic Content Curation	No	Yes
Personalization	No	Yes
Spams/Advertisement	Yes	No
Resource Consumption	More	Less

A. Testing

Some major blog search engines and blogging sites results are compared with the result of proposed model, BlogMiner. The blog search engines/blogging sites taken are **Google blog search (Earlier, it was blogspot.com), Technorati, Icerocket, and Regator.**

Individually, the resultant posts of these blog post sources are compared with BlogMiner post result on different keywords (*i.e.*, user's interest) as **Olympics, Apple, cricket, conferences, and tablet**. This comparison or test is performed only on the top 20 posts results. While fetching the blog posts results from blog sources, not all results may be blog posts or relevant to the keyword. Therefore, only the relevant blog posts are carried out.

1) Scenario 1 - Google blog search and BlogMiner

Table 4 shows the relevant and irrelevant posts results on education, web, cricket, conferences, and technology keywords of this scenario. Here, irrelevancy shows not all results are posts; some are images, videos, or a news article. Figure 3 shows the graphical representation of Scenario 1.

Table 4. Scenario 1 - Google Blog Search and Blogminer

Keywords	Google blog search		BlogMiner	
	Relevant Posts	Irrelevant Posts	Relevant Posts	Irrelevant Posts
Olympics	18	2	20	0
Special Issue	9	11	20	0
Apple	18	2	19	1
Tablet	16	4	19	1
Cricket	18	2	18	2

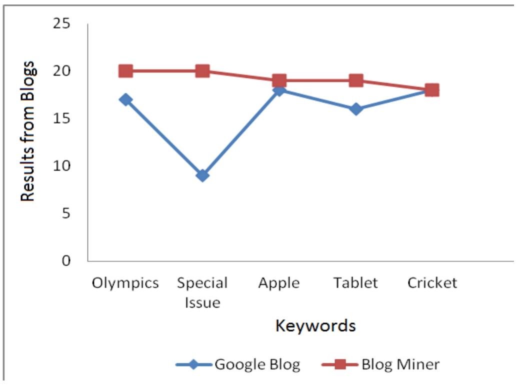


Figure 3. Scenario1 - Google blog search and BlogMiner

2) Scenario 2 - Technorati and BlogMiner

Table 5 shows the relevant and irrelevant posts results on education, web, cricket, conferences, and technology keywords of this scenario. Here, irrelevancy shows the semantically different posts. Figure 4 shows the graphical representation of Scenario 2.

Table 5. Scenario 2 - Technorati and Blogminer

Keywords	Technorati		BlogMiner	
	Relevant Posts	Irrelevant Posts	Relevant Posts	Irrelevant Posts
Olympics	20	0	20	0
Special Issue	1	19	20	0
Apple	18	2	19	1
Tablet	17	3	19	1
Cricket	18	2	18	2

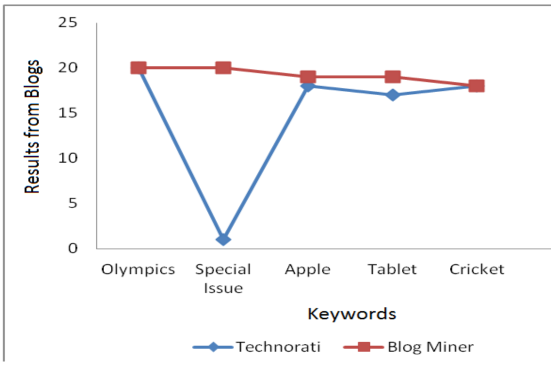


Figure 4. Scenario 2 - Technorati and Blogminer

3) Scenario 3 - Regator and BlogMiner

Table 6 shows the relevant and irrelevant posts results on education, web, cricket, conferences, and technology keywords of this scenario. Here, irrelevancy shows the semantically different posts and ambiguity. Figure 5 shows the graphical representation of Scenario 3.

Table 6. Scenario 3 - Regator and Blogminer

Keywords	Regator		BlogMiner	
	Relevant Posts	Irrelevant Posts	Relevant Posts	Irrelevant Posts
Olympics	16	4	20	0
Special Issue	2	18	20	0
Apple	14	6	19	1
Tablet	9	11	19	1
Cricket	14	6	18	2

4) Scenario 4 - Icerocket and BlogMiner

Table 7 shows the relevant and irrelevant posts results on Olympics, web, cricket, conferences, and technology keywords of this scenario. Here, irrelevancy shows the semantically different posts and language problem (translator must be required). Figure 6 shows the graphical representation of Scenario 4.

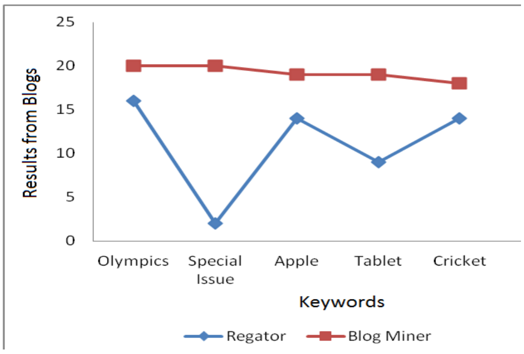


Figure 5. Scenario 3 - Regator and Blogminer

Table 7. Scenario 4 - Icerocket and Blogminer

Keywords	Icerocket		BlogMiner	
	Relevant Posts	Irrelevant Posts	Relevant Posts	Irrelevant Posts
Olympics	13	7	20	0
Special Issue	6	14	20	0
Apple	14	6	19	1
Tablet	8	12	19	1
Cricket	7	13	18	2

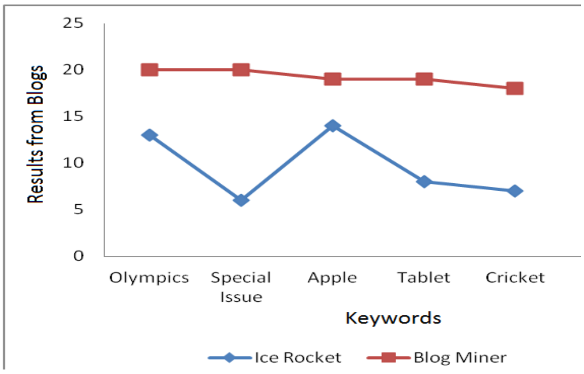


Figure 6. Scenario 4 - Icerocket and BlogMiner

B. Analysis

After observing the results, it is analysed that the proposed model, BlogMiner, is improving the searching experience of user by improving the relevant results. In compare to other blogging sites and blog search engines, results of BlogMiner are much better. An analytical table, Table 8 is shown below.

Table 8. Results Of Average Improvement In Relevant Blog Posts Search

Blog Posts Source	**Average % of relevant blog posts**
Google Blog Search	79%
Technorati	74%
Regator	55%
Icerocket	48%
BlogMiner	**96%**

In Table, the average percentage is calculated based on the relevant blog posts. In Google Blog Search, not all search results are blogs. Some of the results are the video, images, and other content like articles. After analysis, it is found that only 79% of the results are blog posts, or say, a relevant posts. In technorati, based on the sematics, the difference is easily examined between the relevant and irrelevant posts. The average blog posts search results are 74%. In Regator, 55% of the posts are relevant and rest of the posts are irrelevant due to the semantic difference and ambiguity. Icerocket gives 48% of relevant blog posts. BlogMiner curates only the relevant blog posts. Therefore, while searching the keywords only the relevant blog post results are displayed to the user and its average percentage of relevant posts is 96%.

5. Conclusion

BlogMiner is a combination of both, a blog search engine and a blog site. In addition, user can customize their search results based on his interest. To search a relevant post, user does not require wasting his valuable time on clicking here and there for getting the required information. Moreover, with this functionally user will get what exactly he/she wants. This proposed work is an innovative idea in the field of information retrieval from blogs and will surely improve the information searching and the knowledge experience of users.

References

[1] I. Dolinska, “Simple Blog Searching framework based on Social Network Analysis”, Proceedings of the International Multi-conference on Computer Science and Information Technology, Poland, October 18-20, **(2010)**, pp. 611–617.

[2] H. Fu and K. Anyanwu, "Effectively interpreting keyword queries on RDF databases with a rear view", Proceedings of the 10th international conference on The semantic web – Germany, vol. Part I, **(2011)** October 23-27, pp. 193-208.

[3] G. Li, B. C. Ooi, J. Feng, J. Wang and L. Zhou, "EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data", Proceedings of the International conference on Management of data ACM SIGMOD, Canada, **(2008)** June 09-12, pp. 903-914.

[4] N. A. Ahmad and M. A. Mazeyanti, "Blogging as a means of knowledge sharing: Blog communities and informal learning in the blogosphere", Proceedings of the International Symposium on Information Technology, Kuala Lumpur, Malaysia, **(2008)** August 26-28, pp. 1-5.

[5] A. K. Singh and R. C. Joshi, "Semantic tagging and classification of blogs", Proceedings of the International Conference on Computer and Communication Technology (ICCCT), Allahabad, India, **(2010)** September 17-19, pp. 455-459.

[6] V. K. Singh, D. Mahata and R. Adhikari, "Mining the Blogosphere from a Socio-political Perspective", Proceedings of the International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Poland, **(2010)** October 8-10, pp. 365 - 370.

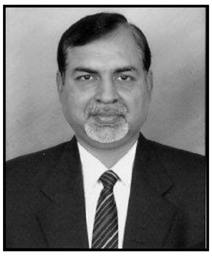
[7] Z. Ping, "Research on Personalized Blog Information Retrieval", Proceedings of the International Conference on Web Information Systems and Mining (WISM), China, **(2010)** October 23-24, pp. 289-292.

[8] K. Harsh and K. B. Mohan, "A New Appraoch to Blog Information Searching and Curating", Proceedings of the CSI 6th International Conference on Software Engineering (CONSEG), IEEE, Indore, **(2012)** September 5-7, pp. 372-377.

Authors

Harsh Khatter, he is working as assistant professor in ABES Engineering College, Ghaziabad, India. He has done Btech and M.tech in Computer Science stream. He is working on Web 2.0 tools, Blogs. His research interests include Web service, Data Mining and Databases, Fuzzy logics. He is a member of IEEE Society, CSI India.

Munesh Chandra Trivedi, he has completed post-graduation and doctorate in Computer Science. He has rich experience in teaching the undergraduate and postgraduate classes. He has published 20 text books and 50 research papers. **IEEE computer society has Sponsored (**Technically & financially) him for organizing IEEE international conference. He has also worked as Member of organizing committee in several IEEE international conferences in India and abroad. He is on the review panel of IEEE Computer Society, International Journal of Network Security, Pattern Recognition letter and Computer & Education (Elsevier's Journal). He is active member of IEEE Computer Society, International Association of Computer Science and Information Technology, Computer Society of India, International Association of Engineers, and life member of ISTE.



Mohan Kalra, he is currently working as a Professor and Head in the Department of Computer Science and Engineering at Ajay Kumar Garg Engineering College, Ghaziabad, India. He has done his B. Tech from Delhi College of Engineering, Delhi in 1977 and completed his M. Tech from IIT, Delhi in 1991. He has vast experience of 35 years of academia and industry in CSE and IT fields. His research interests include eLearning, Computer Networks, and Digital Logic Design. He is also a member of several professional bodies: IEEE, CSI, and IET.