Weblog Search Engine Based on Quality Criteria

F. Azimzadeh[1], A. R. Ramli[2], B. M. Ali[3] and H. Ibrahim[4]
[1] Islamic Azad University, Maybod Branch, Department of Computer Engineering, Iran
[2]Institute of Advanced Technology, University Putra Malaysia, 43400 Serdang, Malaysia
[3]Department of Engineering, 43400 Serdang, Malaysia
[4]Department of Computer Science and Information Technology, 43400 Serdang, Malaysia

1. Abstract

Nowadays, increasing amount of human knowledge is placed in computerized repositories such as the World Wide Web. This gives rise to the problem of how to locate specific pieces of information in these often quite unstructured repositories. Search engines is the best solved. Some studied show that, almost half of the traffic to the blog server comes from search engines.
The more outgoing and informal social nature of the blogosphere opens the opportunity for exploiting more socially-oriented features. The nature of blogs, which are usually characterized by their personal and informal nature, dynamically and constructed on the new relational links required new quality measurement for blog search engine.
Link analysis algorithms that exploit the Web graph may not work well in the blogosphere in general. (Gonçalves et al 2010) indicated that most of the popular blogs in the dataset (70%) have a PageRank value equal -1, being thus almost invisible to the search engine.
We expected that incorporated the special blogs quality criteria would be more desirably retrieved by search engines.

2. INTRODUCTION

Over the last decade, the World Wide Web (WWW) has become in a place for the exchange and publication of information. In recent years, we have seen a growing amount of semi structured and unstructured data as well as social media (Madnick, Wang et al. 2009).
Weblogs, and other forms of social media, differ from traditional web content in many ways (Hurst and Maykov 2009). A Weblog or blog is a personal page maintained by its owner as a single author which is updated based on his opinions in the chronological order. Blogs have simple content management tools enabling non-experts to build easily updatable web diaries or online journals, whereas some blogger update the personal weblog sometimes within a day.
Unlike books and journals, most of this information on Weblog is unfiltered, i.e. not subject to editing or peer review by experts. This lack of quality control of Weblogs makes the task of finding quality information on the web especially critical. An effective search engine on weblog should satisfy the user‘s needs in achieve high quality information.
Technorati which works in blogs field indexed 133 million blogs records since 2002-2008, and approximates the number of blog posts about 900,000 per day (Http://technorati.com 2008).
However, as we shall see, the quality metrics discussed here are more sophisticated than simple in-degree counts. The most obvious use of significance metrics is in web search and retrieval where the most relevant and high-quality set of pages must be selected from a vast index in response to a user query (Dhyani, Ng et al. 2002).
This research pays attention to extract high quality information based on the user’s query in Weblog. Some of the most significant current discussions are extract high quality information from a vast number of no filtering information, consider special connection and structure in Weblog and pay attention to selected language for Weblog.

3. PROBLEM STATEMENTS

The information explosion on the Internet has placed high demands on search engines (Wen, Nie et al. 2001). Based on the large search engine strategy, business is more important than quality so in the search result, the advertisement come upper than high quality pages. This is not desirable for user, usually user looking for high quality information.

In this context, the concepts of information quality are highly relevant to user achieve his or her purpose in the process of Information Retrieval. There is no general agreement about the concept of quality in information retrieval by search engine. **A** major issue is the inability of search engine technology to wade through the vast expanse of questionable content and return "quality" results to a user's query (Knight and Burn 2005). However, two major problems with quality search engine in the Weblog are incorporating of information quality criteria in the search result and considering some differences between Weblog and other web page in the crawler algorithm.

Defining what Information Quality (IQ) is within the context of the Weblog and its search engines then will depend greatly on whether dimensions are being identified for the producers of information (blogger), the crawler and ranking systems used for information retrieval, or for the searchers and users of information.

This research attempt to solve the problem of how to consider information quality criteria for search engine in Weblog specifically as it pertains to information retrieval on the Weblog.

There are a number of issues efficient to selection suitable information quality criteria and implementation the general search engine and Weblog search engine. These problems will be briefly addressed here.

Applying QoI commonly to the World Wide Web has its own set of problems. Firstly, there are no quality control procedures for information uploaded onto the Web and secondly, users of the information have to make judgments about its quality for themselves (Rieh 2002), creating a uniquely subjective environment where one user's quality could be of little or no value to another user. This makes quality dimensions such as relevancy and usefulness not only enormously important but also extremely difficult to gauge (Knight and Burn 2005).

Another problem is that many search engines like as Weblog search engine are not index able by traditional methods. Only surface web documents have a hyper linking system that supports easy indexing (King, Li et al. 2007). Typical ranking methods based on simple link graphs (i.e. PageRank) favour A-list weblogs, with many reciprocal links (Herring, Kouper et al. 2005). Although the graph formed by the hyperlinks between weblog posts is part of the web graph, the ranking algorithms for web pages seem to be insufficient (Kritikopoulos, Sideri et al. 2006) while there are particular and important link between Weblog, like written comments, received comments or list of friends. These links have recently been called (number of written comments, number of received comments, comment per entry and number of friends), also are information quality criteria in the considered QoI model. However, a major problem with this kind of criteria is, how can consider that in the significant part of search engine algorithm.

4. OBJECTIVES

Nowadays Weblogs play an important role in social interactions. People who maintain blogs and update them, so called bloggers, involve in a series of interactions and interconnections with other people (Blood 2004). Within this activity arise an enormous data and information repository, however in this process information quality is usually ignored, this lead to the problem that document is retrieved without regard to their quality.

The aim of this study is to implement a search engine for Weblog, based on IQ criteria. For this, a number of IQ criteria should be selected. The selected criteria must be suitable for Weblog environment. The criteria also should be incorporated into the search engine. Search engines download parts of the existing Weblog and offer Weblogs' users access to this database through keyword search.

The search engines at least should contain two fundamental components: web crawlers, which find, download, and parse content in the Weblog, and data miners, which extract keywords from pages, rank document importance, and answer user queries(Lee, Leonard et al. 2008).

In this research, we must implement the crawler for Weblogs so that consider special attributes of Weblogs such as comments and friends' link and other needed criteria. Finding high quality page priority is the main task for crawler in the quality based Weblog search engine.

We must also design the search engine data miners to rank document importance which formulate by selected IQ criteria.

Kargar et al (Kargar, Ramli et al. 2008) found seven IQ dimensions which contain subjective score, authority, link popularity, timeliness, latency, maturity and redundancy, each dimension include some IQ criteria on weblog demonstrated in the table1. This work captures a number of IQ criteria from the next table to incorporate in the search engine.

5. LITERATURE REVIEW

There is opening a view on quality approach search engine in general. However, Weblog search engine is a new field, then a little amount of literature has been published on this, the second part of this section review some attempts on quality in Weblog search engine.

5.1 General Search Engine

The introduction of quality metrics has been a recent development for public search engines, most of that constructed earlier on purely textual comparisons of keyword queries with indexed pages for assigning relevance scores. Engines such as Google use a combination of relevance and quality metrics in ranking the responses to user queries (Dhyani, Ng et al. 2002).

Table1. IQ Dimensions and IQ Criteria Suggested by Kargar et al

IQ DIMENSION	IQ CRITERIA
Subjective Score	Cohesiveness, Concise, Believability, Understandability, Completeness, Objectiveness, Accuracy, Informativeness, Presentation
Authority	Num. Received Comments, Num. Written Comments, Num. Entries, Num. Referred, Num. Visitors, Comment Per Entry
Link Popularity	Num. Links, Num. Visited Links, Num. Friends
Timeliness	Last Login, Last Update, Availability
Latency	First Load Time, Full load time
Maturity	Meta Tag, Age
Redundancy	Multimedia Rate, Weblog Size

Low quality and spam are some of the most serious challenges for any Web search engine. Search engines and research reacted with a heuristic approach to approximating quality (Mandl 2006). In the field of search engine based on quality we identify three main approach, link analysis, IQ criteria approach and topic and content relevancy.

5.1.1 Link Analysis

Link analysis is the approach most often discussed for quality assessment in information retrieval (Mandl 2006). Link structure analysis is based on the notion that a link from a page p to page q can be viewed as an endorsement of q by p. Previous research in Web search such as PageRank has demonstrated that the quality of a Web page is dependent on the hyperlink structure in which it is embedded.

The basic assumption of PageRank and similar approaches is that the number of in- or back-links of a Web page can be used as a measure for the popularity and consequently for the quality of a page. PageRank assigns an authority value to each web page which is primarily a function of its back links. Additionally, it assumes that links from pages with high authority should be weighed higher and should result in a higher authority for the receiving page. To account for the different values each page has to distribute, the algorithm is carried out iteratively until the result converges (Borodin, Roberts et al. 2005). Therefore, PageRank is a characteristic of the web page itself; it is higher if more web pages link to this page, as well as if these web pages have high PageRank.

Two important types of techniques in link-structure analysis are co-citation based schemes and random-walk based schemes (Dhyani, Ng et al. 2002). The main idea behind co-citation based schemes is the notion that when two pages p1 and p2 both point to some page q, it is reasonable to assume that p1 and p2 share a mutual topic of interest. Likewise, when p links to both q1 and q2, it is probable that q1 and q2 share some mutual topic. On the other hand, random-walk based schemes model the Web (or part of it) as a graph where pages are nodes and links are edges, and apply some random-walk model to the graph. Pages are then ranked by the probability of visiting them in the modelled random walk.

5.1.2 IQ Criteria approach

The approach implemented by (Zhu and Gauch 2000) explicitly integrates quality assessment into an information retrieval system. Six criteria for quality are extracted from Web pages: currency, availability, information–to-noise ratio, authority, popularity, and cohesiveness (Zhu and Gauch 2000). These criteria are

used to influence the ranking of documents in retrieval. In retrieval experiments, it was shown that the average precision of the rankings including quality information retrieval measures in some cases.

Dyhani et al have been suggested that to measure the usefulness of a page several other criteria could be combined with the PageRank (or any quality metrics in general) (Dhyani, Ng et al. 2002), there is mention the criteria; search engine leads (whether users actually select the page when presented as a query result), number of visits (referral and popularity), time spent by visitors (usefulness of content), interaction with dynamic content and user interface (implies utility) and local navigation (implies user's interest through intent to explore further). This research has been identified; each of the criteria above is vulnerable to misinterpretation, a large number of visits to a Web page occur for the same reason as a high PageRank, i.e., high visibility. Similarly, time spent is influenced not just because how interesting the content is but also by Web page layout, design and readability.

Understanding IQ from the point of view of the user (or searcher) of Information, involves understanding the processes of information retrieval on the Internet. More often than not, information retrieval involves using a Search Engine, a specific set of keywords or concepts, which make up a user's query, followed by a decision process where the user makes value judgements concerning the results returned by the search engine to their query. These value judgements involve the user making choices according to concepts such as accuracy, currency and usefulness (Rose and Levinson 2004).

Naumann and Rolker's approach is more complex, using a three-fold assessment for the quality of an information source, according to the subjects, objects and processes involved in information retrieval. The premise of this model is based on two basic assumptions: 1. the quality of information and 2. the information retrieval process, involves both the influences and the processes involved with information quality and retrieval are used to assign quality scores within three contexts, Subject, Process or Object criteria(Naumann and Rolker 2000). The scores are used to create metadata that is used to assign a PageRank for the information source when it is listed in the results of a user's query. The criteria that classified by them demonstrates in the table 2.

Table2. Subject, Object and Process Criteria considered by (Naumann and Rolker 2000)

Subject Criteria	**Object Criteria**	**Process Criteria**
Believability	Completeness	Accuracy
Concise	Customer Support	Amount of data
Interpretability	Documentation	Availability
Relevancy	Objectivity	Latency
Reputation	Price	Response time
Understandability	Reliability	Consistent representation
Value-Added	Security	
	Timeliness	
	Verifiability	

IQIP Model proposed approach involved four steps (IQIP); Identify, Quantify, Implement and Perfect. Knight performed the IQIP project that consider(Knight and Burn 2005) ; Identify (the user, environment and task), Quantify (prioritise appropriate dimensions of Information Quality using a 'Dimension Score'), Implement (the chosen IQ dimensions into the Web Crawler) and Perfect (improve the crawler through system and user feedback).

5.1.3 Topic and Content Relevancy Approach

Several attempts have been made to increase the quality of information retrieval; one method is Web clustering engines that organize search results by topic. Search a result clustering is clearly related to the field of document clustering but it poses unique challenges concerning both the effectiveness and the efficiency of the underlying algorithms that cannot be addressed by conventional techniques. The main difference is the emphasis on the quality of cluster labels, whereas this issue was of somewhat lesser importance in earlier research on document clustering(Batsakis, Petrakisa et al. 2009).

As the primary aim of a search results clustering engine is to decrease the effort required to find relevant information, user experience of clustering-enabled search engines is of crucial importance.

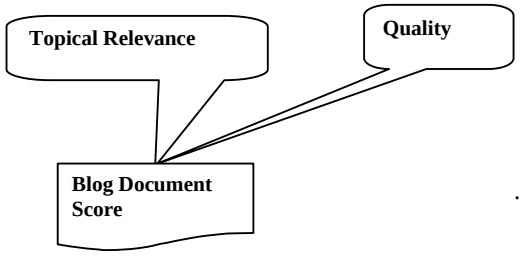
Since clustering engines are meant to overcome the limitations of plain search engines, we need to evaluate whether the use of clustered results does yield a gain in retrieval performance over ranked lists (Batsakis, Petrakisa et al. 2009).
A new approach to learning paths reaching web pages relevant to a given topic is proposed inspired by a research (Liu, Janssen et al. 2006). In particular, this work is enhanced with ideas from classic focused crawlers (for better estimation of page content relevance with the content of the topic) combined with ideas from document clustering for learning promising paths leading to target pages.

5.2 Weblog Search Engines

A large part of the hidden web resides in weblog servers; traditional search engines perform poorly on blogs. As our mention above link analyse is a method employed to achieve high quality list in search engine result list. In weblog arias some different link, (Kritikopoulos, Sideri et al. 2006) present a method for ranking weblogs utilizing both link graph and similarity, and based on an enhanced and weighted graph of weblogs capturing crucial weblog features. The explicit hyperlinks between weblog entries, and the implicit links between weblogs that emanate from the similarity in the users that post or in the topics discussed, turn the blogosphere into an interesting graph. In this weighted graph the nodes are either the weblogs or the contained posts and the edges are either hyperlinks or similarity links. The weights express the strength of relation between nodes. Rankings are then assigned using their algorithm, BlogRank, which is a modified version of PageRank.
(Cohen and Krishnamurthy 2006) analysed that communities of weblogs can be related by type or topic, or connected via hyperlinks. Every weblog consists of a series of entries, expressing uniform or contradictory opinions and link to other entries or web pages. To determine the link between Weblog Gill's definition is: Links are used as suggestions, as a means to express agreement or disagreement (Gill 2004). According to (Zhou and Davis 2007), Bloggers generally have read and respond relations with each other. These are indicated by hyperlinks between weblogs. The blogroll lists the read relations while hyperlinks appear in the entry bodies or entry comments represent the respond relations. Each hyperlink is a communication instance between two bloggers. The number of instances would indicate the strength of tie between bloggers.
Hurts and Maykove discusses the challenges and strategies for weblog (Hurst and Maykov 2009); they mention that the most important difference between traditional webpage and Weblog is the highly temporal nature of the content. Applications that leverage social media content must, to be effective, have access to this data with minimal publication/acquisition latency. Hurst and Msaykove emphases to specifics of social media aggregation pose some unique challenges. Real-time (The information in blogs is time-sensitive), Coverage (It is important to fetch the entire blogosphere), Scale (As of now, the size of the blogosphere is on the order of few hundred millions blogs) and Data Quality (The crawler should output good quality, uncorrupted data).
In this recent study the challenges of building a blog crawler that finding last modification in a short time overcome by considering a ping server that is a service which aggregates blog updates. Whenever a blog receives a new post, its blog host sends this information to a ping server. A few public ping servers publish a list of blogs which were updated every 5 minutes.

6. OUR WORK

This research should be done to make a new search engine based on IQ criteria on Weblog. For this aim will be considered two scores for each post; one of them show that relevancy between search engine query and Blog post, the other one is the score based on quality of information criteria.
A blog search engine may receive a search query. The blog search engine may determine scores for group of blog documents in response to the search query, where the score are based on a relevance of the group of blog documents to the search query and a quality of the group of blog documents, which is independent of the query terms A blog search engine may receive a search query. The blog search engine may determine scores for group of blog documents in response to the search query, where the score are based on a relevance of the group of blog documents to the search query and a quality of the Blog documents, which is independent of the query terms. Figure 1, demonstrates relational between relevancy and quality in blog's score.



Figures 1: demonstrate relational between relevancy and quality in blog's score**.**

In the Relevancy phase we rank measures the relevance between a query and a document based some special Weblog characterized. Many models have been proposed to estimate the similarity between the query and the document in general search engine. We need to consider some special text in the Blog such as comments issued for a post and apply relevancy algorithm on it.
In the quality phase we apply total quality score for this post based on the quality criteria.
(Andersson and Silvestrov 2008) exploited the next formula to employed quality and relevancy in the general search engine.

$$\mathrm{Rel}\omega(\tau) = P(\tau,\omega)q(\tau), \quad (1)$$

Where $P(\tau,\omega)$ is the on-page score of Blog post τ for query ω, i.e. how relevant the information on τ is thought to be to query ω, and $q(\tau)$ is the *quality* function of τ , as calculated from factors not directly present on the page. The quality function q is include the total score for Blog post τ that evaluated by quality of information criteria. Note that q is not query-dependent, but rather assigns a general quality weight to each page regardless of the query. The range of q is usually taken to be [0, 1], and thus multiplication by q acts as a kind of damping on the document scores.

7. CONCLUSION

The search engine must facilitate user to find high quality information when submitting a query. Applied both of important users need; relevancy and quality in search engine return results, will help to user that fined more useful information by search engine.
We incorporate quality metric for Weblog search engine and consider special relational in Weblog environment.

REFERENCES

[1] Andersson, F., Silvestrov, S. (2008). "The mathematics of Internet search engine "Acta Appl Math (2008) 104: 211–242
[2] Batsakis, S., E. G. M. Petrakisa, et al. (2009). "Improving the performance of focused web crawlers " Data & Knowledge Engineering, Elsevier **68**(10): 1001-1013.
[3] Blood, R. (2004). "How blogging software reshapes the online community." Commun. ACM **47**(12): 53-55.
[4] Borodin, A., G. O. Roberts, et al. (2005). "Link analysis ranking: algorithms, theory, and experiments." ACM Trans. Internet Technol. **5**(1): 231-297.
[5] Cohen, E. and B. Krishnamurthy (2006). "A short walk in the Blogistan " Computer Networks, Elsevier **50**(5): 615-630.
[6] Dhyani, D., W. K. Ng, et al. (2002). "A survey of Web metrics." ACM Comput. Surv. **34**(4): 469-503.
[7] Gill, K. E. (2004). How can we measure the influence of the blogosphere? Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2004.. New York, WWW2004.
[8] Gonçalves, M. A., G. A. Almeida, et al. (2010). "On Popularity in the Blogosphere". Digital Object Indentifier 10.1109/MIC.2010.54 1089-7801) 2010 IEEE.
[9] Herring, S. C., I. Kouper, et al. (2005). Conversations in the Blogosphere: An Analysis "From the Bottom Up". awaii International Conference onSystem Sciences (HICSS-38). Los Alamitos, IEEE Press.
[10] Http://technorati.com. (2008). "http://technorati.com/blogging/state-of-the-blogosphere/."
[11] Hurst, M. and A. Maykov (2009). Social Streams Blog Crawler. Proceedings of the 2009 IEEE International Conference on Data Engineering, IEEE Computer Society.
[12] Kargar, M. J., A. R. Ramli, et al. (2008). "Formulating Priory of Information Quality Criteria on the Blog." World Applied Sciences **4**(4): 586-593.
[13] King, J. D., Y. Li, et al. (2007). "Mining world knowledge for analysis of search engine content." Web Intelli. and Agent Sys. **5**(3): 233-253.
[14] Knight, S. a. and J. Burn (2005). "Developing a Framework for Assessing Information Quality on the World Wide Web." Informing Science Journal **8**(10): 159-172.
[15] Kritikopoulos, A., M. Sideri, et al. (2006). BlogRank: ranking weblogs based on connectivity and similarity features. Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications. Pisa, Italy, ACM.

[16] Lee, H. T., D. Leonard, et al. (2008). IRLbot: scaling to 6 billion pages and beyond. Proceeding of the 17th international conference on World Wide Web. Beijing, China, ACM.
[17] Liu, H., J. Janssen, et al. (2006). "Using HMM to learn user browsing patterns for focused Web crawling " Data & Knowledge Engineering, Elsevier **59**(2): 270-291.
[18] Madnick, S. E., R. Y. Wang, et al. (2009). "Overview and Framework for Data and Information Quality Research." J. Data and Information Quality **1**(1): 1-22.
[19] Mandl, T. (2006). Implementation and evaluation of a quality-based search engine. Proceedings of the seventeenth conference on Hypertext and hypermedia. Odense, Denmark, ACM.
[20] Naumann, F. and C. Rolker (2000). Assessment methods for information quality criteria. International Conference on Information Quality.
[21] Rieh, S. Y. (2002). "Judgment of information quality and cognitive authority in the Web. , 53 (2), ." Journal of the American Society for Information Science and Technology **53**(2): 145-161.
[22] Rose, D. E. and D. Levinson (2004). Understanding user goals in web search. Proceedings of the 13th international conference on World Wide Web. New York, NY, USA, ACM.
[23] Wen, J. R., J. Y. Nie, et al. (2001). Clustering user queries of a search engine. Proceedings of the 10th international conference on World Wide Web. Hong Kong, Hong Kong, ACM.
[24] Zhou, Y. and J. Davis (2007). Analysis of Weblog Link Structure – A Community Perspective. Web Information Systems and Technologies, Springer. **1:** 307-320.
[25] Zhu, X. and S. Gauch (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Athens, Greece, ACM.