# Identifying and Ranking Topic Clusters in the Blogosphere

**M. Atif Qureshi**
Korea Advanced Institute of Science and Technology
atifms@kaist.ac.kr

**Arjumand Younus**
Korea Advanced Institute of Science and Technology
arjumandms@kaist.ac.kr

**Muhammad Saeed**
University of Karachi
saeed@uok.edu.pk

**Nasir Touheed**
Institute of Business Administration
ntouheed@iba.edu.pk

## Abstract

The blogosphere is a huge collaboratively constructed resource containing diverse and rich information. This diversity and richness presents a significant research challenge to the Information Retrieval community. This paper addresses this challenge by proposing a method for identification of "topic clusters" within the blogosphere where topic clusters represent the concept of grouping together blogs sharing a common interest i.e. topic, the algorithm takes into account both the hyperlinked social network of blogs along with the content in the blog posts. Additionally we use various forms and parts-of-speech of the topic to provide a broader coverage of the blogosphere. The next step of the method is to assign topic-specific ranks to each blog in the cluster using a metric called "Topic Discussion Rank," that helps in identifying the most influential blog for a specific topic. We also perform an experimental evaluation of our method on real blog data and show that the proposed method reaches a high level of accuracy.

## 1 Introduction

With a proliferation of Web 2.0 services and applications there has been a major paradigm shift in the way we envision the World Wide Web (Anderson, 2007; O'Reilly, 2005). Previously the Web was considered as a medium to access information in a read-only fashion. Weblogs or blogs is one such application that has played an effective role in making the Web a social gathering point for masses. The most appealing aspect of blogs is the empowerment they provide to people on the World Wide Web by enabling them to publish their own opinions, ideas, and thoughts on many diverse topics of their own interest generally falling into politics, economics, sports, technology etc. A blog is usually like a personal diary (Sorapure, 2003) with the difference that it's now online and accessible to remote people, it consists of posts arranged chronologically by date and it can be updated on a regular basis by the author of the blog known as blogger. Moreover bloggers have the option to link to other blogs thereby creating a social network within the world of blogs called the blogosphere – in short the blogosphere is a collaboratively constructed resource with rich information on a wide spectrum of topics having characteristics very different from the traditional Web.

However with these differing characteristics of blogs arise many research challenges and this is in particular the case for the Information Retrieval domain. One important problem that arises within this huge blogosphere (Sifry, 2009) is with respect to identification of topic clusters. Such a task involves identification of the key blog clusters that share a common interest point (i.e., topic) reflected quite frequently through their blog posts. This is a special type of cluster-



Beijing, August 2010
*Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010*, pages 55–62,

ing problem with useful applications in the domain of blog search as Mishne and de Rijke (2006) point out in their study of blog search about the *concept queries* submitted by users of blog search systems.

Moreover ranking these bloggers with respect to their interest in the topic is also a crucial task in order to recognize the most influential blogger for that specific topic. However the blog ranking problem has a completely different nature than the web page ranking problem and link popularity based algorithms cannot be applied for ranking blogs. The reasons for why link based methods cannot be used for blog ranking are as follows:

- Blogs have very few links when compared to web pages; Leskovec et al. report that average number of links per blog post is only 1.6 links (2007). This small number of links per blog results in formation of very sparse network especially when trying to find blogs relevant to a particular topic.
- Blog posts are associated with a timestamp and they need some time for getting in-links. In most of the cases when they receive the links the topics which they talk about die out.
- When link based ranking techniques are used for blogs, bloggers at times assume the role of spammers and try to exploit the system to boost rank of their blogs.

In this paper we propose a solution for identification of topic clusters from within the blogosphere for any topic of interest. We also devise a way to assign topic-specific ranks for each identified blog within the topic cluster. The cluster is identified by the calculation of a metric called "Topic Discussion Isolation Rank (TDIR)." Each blog in the cluster is also assigned a topic rank by further calculation of another metric "Topic Discussion Rank (TDR)." The first metric "TDIR" is applied to a blog in isolation for the topic under consideration and the second metric "TDR" takes into account the blog's role in its neighborhood for that specific topic. Our work differs from past approaches (Kumar et al., 2003; Gruhl et al., 2004; Chi et al., 2007; Li et al., 2009) in that it takes into consideration both the links between the blogs as well as the content in the blog posts whereas a majority of the past methods follow only link structure. Furthermore we make use of some natural language processing techniques to ensure better coverage of our cluster-finding and ranking methodology. We also perform an experimental evaluation of our proposed solution and release the resultant data of blog clusters and the ranks as an XML corpus.

The remainder of this paper is organized as follows. Section 2 presents a brief summary of related work in this dimension and explains how our proposed methodology differs from these works. Section 3 explains the concept of "topic clusters" in detail along with a description of our solution for clustering and ranking blogs on basis of topics. Section 4 explains our experimental methodology and presents our experimental evaluations on a corpus of 50,471 blog posts gathered from 102 blogs. Section 5 concludes the paper with a discussion of future work in this direction.

## 2 Related Work

Given the vast amount of useful information in the blogosphere there have been many research efforts for mining and analysis of the blogosphere. This section reviews some of the works that are relevant to our study.

There have been several works with respect to community detection in the blogosphere: one of the oldest works in this dimension is by Kumar et al. who studied the bursty nature of the blogosphere by extracting communities using the hyperlinks between the blogs (2003). Gruhl et al. proposed a transmission graph to study the flow of information in the blogosphere and the proposed model is based on disease-propagation model in epidemic studies (2004). Chi et al. studied the evolution of blog communities over time and introduced the concept of *community factorization* (2007). A fairly recent work is by Li et al. that studies the information propagation pattern in the blogosphere through *cascade affinity* which is an inclination of a blogger to join a particular blog community (2009). Apart from detection of communities within the blogosphere another related study which has recently attracted much interest is of identifying influentials within a "blog community" (Nakajima et al., 2005; Agarwal et al., 2008). All

these works base their analysis on link structure of the blogosphere whereas our analytical model differs from these works in that it assigns topic based ranks to the blogs by taking into account both links and blog post's contents.

Along with the community detection problem in the blogosphere there has also been an increasing interest in ranking blogs. Fujimura et al. point out the weak nature of hyperlinks in the web blogs and due to that nature they devise a ranking algorithm for blog entries that uses the structural characteristic of blogs; the algorithm enables a new blog entry or other entries that have no in-links to be rated according to the past performance of the blogger (2005). There is a fairly recent work closely related to ours performed by Hassan et al (2009) and this work identifies the list of particularly important blogs with recurring interest in a specific topic; their approach is based on lexical similarity and random walks.

## 3 Cluster Finding and Ranking Methodology

In this section we explain the concept of "topic clusters" in detail and go into the details of why we deviate from the traditional term of "blog community" in the literature. After this significant discussion we then move on to explain our proposed method for identification and ranking of the "topic clusters" in the blogosphere: two metrics "topic discussion isolation rank" and "topic discussion rank" are used for this purpose.

### 3.1 Topic Clusters

As explained in section 2 the problem of grouping together blogs has been referred to as the "community detection problem" in the literature. However an aspect ignored by most of these works is the contents of the blogs. Additionally most of the works in this dimension find a blog community by following blog threads' discussions/conversations (Nakajima et al., 2005; Agarwal et al., 2008) which may not always be the case as blogs linking to each other are not necessarily part of communications or threads.

With the advent of micro blogging tools such as Twitter (Honeycutt and Herring, 2009) the role of blogs as a conversational medium has diminished and bloggers link to each other as a socially networked cluster by linking to their most favorite blogs on their home page as is shown in the snapshot of a blog in Figure 1:


**Figure 1: Blog Showing the List of Blogs it Follows**

Normally those bloggers link to each other that have similar interests and importantly talk about same topics. Hence the idea of topic cluster is used to extract those clusters from the blogosphere that have strong interest in some specific topics which they mention frequently in their blog posts and additionally they form a linked cluster of blogs. As pointed out by Hassan et al. the "task of providing users with a list of particularly important blogs with a recurring interest in a specific topic is a problem that is very significant in the Information Retrieval domain" (2009). For the purpose of solving this problem we propose the notion of "topic clusters." The task is much different from traditional communi-

ty detection in the blogosphere as it utilizes both content and link based analysis. The process of finding topic clusters is carried out by calculating a metric "Topic Discussion Isolation Rank" which we explain in detail in section 3.3.

### 3.2 Rank Assignment to Topic Clusters

As we explained in section 1, due to the unique nature of the blogosphere, traditional link-based methods such as PageRank (Page et al., 1998) may not be appropriate for the ranking task in blogs. This is the main reason that we use the content of blog posts and lexical similarity in blog posts along with links for the rank assignment function that we propose. Furthermore we take a blog as aggregate of all its posts for the retrieval task.

### 3.3 Topic Discussion Isolation Rank

Topic Discussion Isolation Rank is a metric that is used to find the cluster of blogs for a specific topic. It takes each blog in isolation and analyses the contents of its posts to discover its interest in a queried topic. We consider a blog along three dimensions as Figure 2 shows:

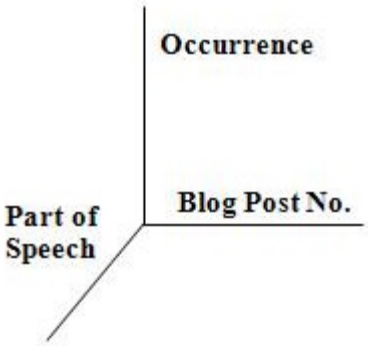


**Figure 2: Blog TDIR Dimensions**

As mentioned in section 1 of this paper we utilize some natural language processing techniques to ensure better coverage of our cluster-finding and ranking methodology: those techniques are applied along the part of speech dimension shown in Figure 1, for a given topic we analyze blog post contents not only for that particular topic but also for its associated adjectives and adverbs i.e. the topic itself is treated as a noun and its adjectives and adverbs are also used. For example if the topic of interest is "democracy" we will also analyze the blog post contents for adjective "democratic" and adverb "democratically." Furthermore, a weight in descending order is assigned to the noun (denoted as $w_n$), adjective (denoted as $w_{adj}$) and adverb (denoted as $w_{adv}$) of the queried topic where $w_n > w_{adj} > w_{adv}$. This approach guarantees better coverage of the blogosphere and the chances of missing out blogs that have interest in the queried topic are minimal. The blog post number denotes the number of the post in which the word is found and occurrence is a true/false parameter denoting whether or not the word exists in the blog post. Based on these three dimensions we formulated the TDIR metric as follows:

$$1 + \frac{(n_{noun} x\ w_n) + (n_{adjective} x\ w_{adj}) + (n_{adverb} x\ w_{adv})}{Number\ of\ total\ posts}$$

Here $w_n$, $w_{adj}$ and $w_{adv}$ are as explained previously in this section and $n_{noun}$ denotes the number of times nouns are found in all the blog posts, $n_{adjective}$ denotes the number of times adjectives are found in all the blog posts and $n_{adverb}$ denotes the number of times adverbs are found in all the blog posts. This metric is calculated for each blog in isolation and the blogs that have TDIR value of greater than 1 are considered part of the topic cluster.

Additionally we also use various forms of the queried topic in the calculation of TDIR as this also ensures better coverage during the cluster detection process. In the world of the blogosphere, bloggers have all the freedom to use whatever terms they want to use for a particular topic and it is this freedom which adds to the difficulty of the Information Retrieval community. Within the TDIR metric we propose use of alternate terms/spellings/phrases for a given topic – an example being the use of "Obama" by some bloggers and "United States first Black President" or "United States' Black President" by others. Such ambiguity with respect to posts talking about same topic but using different phrases/spellings/terms can be resolved by using a corpus-based approach with listing of alternate phrases and terms for the broad topics. Moreover the weights used for each of the part of speech "noun", "adjective" and "adverb" in the TDIR metric can be adjusted differently for different topics with some topics having a stronger indica-

tion of discussion of that topic through occurrence of noun and some through occurrence of adjective or adverb. Some examples of these various measures are shown in our experimental evaluations that are explained in section 4.

### 3.4 Topic Discussion Rank

After the cluster-finding phase we perform the ranking step by means of Topic Discussion Rank. It is in this phase that the socially networked and linked blogs play a role in boosting each other's ranks. It is reasonable to assign a higher topic rank to a blog that has interest in the specific topic and is also a follower of many blogs with similar topic discussions than one that mentions the topic under consideration but does not link to other similar blogs: Topic Discussion Rank does that by taking into account both link structure and TDIR explained in previous section. This has the advantage of taking into account both factors: the content of the blog posts and the link structure of its neighborhood.

The following piecewise function shows how the metric Topic Discussion Rank is calculated:

$$TDR[b] = \begin{cases} TDIR; \text{ if zero outlinks from blog} \\ TDIR + \frac{Matching\_Outlinks}{Total\_Outlinks} \times \sum_{o:(o,b)} TDIR \times damp; \text{ otherwise} \end{cases}$$

*Explanation of notations used:*

*b - blog*

*o : (o,b) – outlinks from blog b*

The TDR is same as the TDIR in case of the blog having zero outlinks as such a blog exists in isolation and does not have a strong participation within the social network of the blogosphere. In the case of a blog having one or more outlinks to other blogs we add its own TDIR to the factor

$$\left(\frac{Matching\_Outlinks}{Total\_Outlinks} \times \sum_{o:(o,b)} TDIR \times damp\right)$$

.

Here matching links represent blogs that are part of topic cluster for a given topic (i.e. those having TDIR greater than 1 as explained in section 3.3) and each matching link's TDIR is summed up and multiplied by a factor called damp. Note that summation of TDIR is used in the first iteration only, in the other iterations it is replaced by TDR of the blogs.

Furthermore it is important to note that the process of TDR computation is an iterative one similar to PageRank (Page et al., 1998) computation, however the termination condition is unlike PageRank in that PageRank terminates when rank values are normalized whereas our approach uses the blog depth as a termination condition which is an adjustable parameter. Due to the changed termination condition the role of spam blogs is minimized.

The damping factor *damp* is introduced to minimize biasness as is explained below. Consider the two blogs as shown with the link structure represented by arrows:

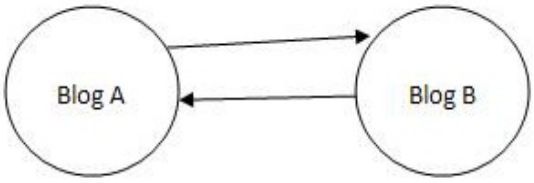


**Figure 3: Example for Damping Factor Explanation**

In this case let's assume the TDIR of blog A is 2 and the TDIR of blog B is 1. Using the formulation for TDR without the damping factor we would have 2+(1/1x1)=3 for blog A and 1+(1/1x2)=3 for blog B which is not the true reflection of their topic discussion ranks. However when we use the damping factor the resultant TDR's are 2+(1/1x1x0.9)=2.9 for blog A and 1+(1/1x2x0.9)=2.8 for blog B and this more correctly represents the topic discussion ranks of both the blogs.

## 4 Experimental Evaluations

This section presents details of our experiments on real blog data. We use precision and recall to measure the effectiveness of our approach of cluster-finding. The experimental data is released as an XML corpus which can be downloaded from:
http://unhp.com.pk/blogosphereResearch/data.tar.gz.

### 4.1 Data and Methodology

The data used in the experiments was gathered from 102 blog sites which comprised of 50,471 blog posts. Currently we have restricted the data set to only the blogspot domain (blogger.com service by Google).We used four blog sites as seeds and from them the link structure of the blogs was extracted after which the crawl (Qureshi et al., 2010) was performed using the XML feeds of the blogs to retrieve all the posts in each blog. Each blog had an average of 494 posts.

The topics for which we perform the experiments of finding TDIR and TDR were taken to be *"compute"*, *"democracy"*, *"secularism"*, *"bioinformatics"*, *"haiti"* and *"obama."*

The measures that we use to assess the accuracy of our method are precision and recall which are widely used statistical classification measures for the Information Retrieval domain. The two measures are calculated using equations 4.1 and 4.2:

$$\text{Precision} = \frac{|Ct \cap Ca|}{|Ca|} \qquad (4.1)$$

$$\text{Recall} = \frac{|Ct \cap Ca|}{|Ct|} \qquad (4.2)$$

Here Ca represents the topic cluster set found using our algorithm i.e. the set of blogs that have interest in the queried topic, in other words it is the set of the blogs that have TDIR greater than 1. Ct represents the true topic cluster set meaning the set of those blogs that not just mention the topic but are really interested in it. The reason for distinguishing between true cluster set Ct and algorithmic cluster set Ca is that our method just searches for the given keyword i.e. topic in all the posts and since natural language is so rich that just mentioning the topic does not represent the fact that the blog is a part of that topic cluster. Hence we use a human annotator/labeler for identification of the true cluster set from the set of the 102 blogs for each of the 6 topics that we used in our experiments.

### 4.2 Results

We plot the precision and recall graphs for the topics chosen. Figure 4 shows the graph for precision:

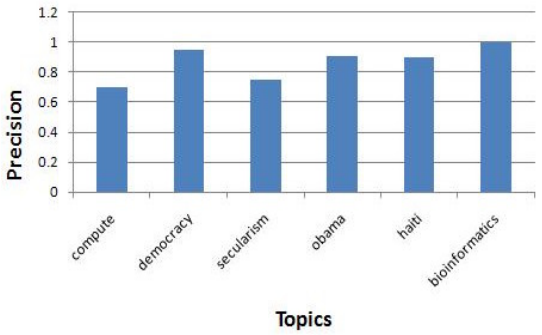


**Figure 4: Precision Graph for Chosen Topics**

The average precision was found to be 0.87 which reflects the accurate relevance of our method. As can be seen from the graph in figure 4 the precision falls below the 0.8 mark only for the topics compute and secularism – the reason for this is that for these two topics a higher proportion of false positives were discovered. Not all the posts having the word "compute" were actually related to computing as found by human annotator. Same was the case for the word secularism – since our method searches for adjective secular and adverb secularly in case of secularism not being found hence there were some blogs in which secular was used but the blog's focus was not in secularism as an idea. On the other hand precision measures for the topics "democracy", "obama", "haiti" and "bioinformatics" were quite good because these words are likely to be found in the blogs that actually focus on them as a topic hence reducing the chances of false positives.

Figure 5 shows the graph for recall:

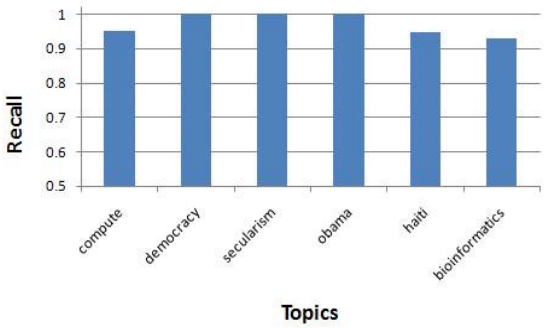


**Figure 5: Recall Graph for Chosen Topics**

The average recall was found to be 0.971 which reflects the high coverage of our method. As the graph in figure 5 shows the recall value is

mostly close to 1 for the chosen topics. This high coverage is attributed to the part of speech dimension as discussed in section 3.3; this technique rules out the chances of false negatives and hence we obtain a high recall for our method.

### 4.3 Additional Experiments

In addition to experiments on the six coarse-grained topics mentioned above we performed some additional experiments on two fine-grained topics and also repeated the experiment performed on topic "Obama" with an additional term "Democrats." On formulating the cluster with these two terms the precision increased from 0.907 to 0.95 which clearly shows that incorporation of extra linguistic features into the TDIR formulation ensures better results. Moreover the ranks of some blogs were found to be higher than the ranks obtained previously and this increase in rank was due to the fact that many posts had subject theme "Obama" but they used the term "Democrats" – when we used this alternate term the ranks i.e. TDR more correctly represented the role of the blogs in the cluster.

The two fine grained topics for which we repeated our experiments were: healthcare bill and avatar. Additional terms were also included in the TDIR and TDR computation process which were as follows:

*healthcare bill – obamacare*
*avatar- sky people, jake sully*

These alternate terms were chosen as these are the commonly associated terms when these topics are discussed. At this point we provided them as query topics but for future work our plan is to use a machine learning approach for learning these alternate phrases for each topic, and knowledge bases such as Wikipedia may also be used to gather the alternate terms for different topics.

The precision for the topic healthcare bill was found to be 0.857 which had a negligible effect on excluding "obamacare"; however recall suffered more on exclusion of alternate term "obamacare" as it fell from 1 to 0.667. Results for the topic "avatar" however were quite different with a precision of 0.47 and a recall of 1; this was due to the large number of false positives that were retrieved for the term avatar and we found reason for this to be that our approach does not take into consideration case-sensitivity at this point hence it failed to distinguish between the term "avatar" and movie "Avatar". Also in the case of topic "avatar" the alternate phrases did not have any effect and hence there is a need to refine the approach for fine-grained topics such as this one – we present future directions for refinement of our approach in section 5.

## 5 Conclusions and Future Work

In this paper we proposed the concept of "topic clusters" to solve the blog categorization task for the Information Retrieval domain. The proposed method offers a new dimension in blog community detection and blog ranking by taking into account both link structure and contents of blog posts. Furthermore the natural language processing techniques we use provide a higher coverage thereby leading to a high average recall value of 0.971 in the experiments we performed. At the same time we achieved a good accuracy as was reflected by an average precision of 0.87.

For future work we aim to combine our proposed solution into a framework for auto generation of useful content on a variety of topics such as "blogopedia"; the content can be obtained automatically from the blog posts and in this way manual effort may be saved. We also plan to refine our approach by taking into account the temporal aspects of blog posts such as time interval between blog posts, start post date and time, end post data and time into our formulation for "Topic Discussion Isolation Rank" and "Topic Discussion Rank". Moreover as future directions of this work we plan to incorporate a machine learning framework for the assignment of the weights corresponding to each topic and for the additional phrases to use for each of the topics that we wish to cluster.

## References


Agarwal, Nitin, Huan Liu, Lei Tang, and Philip S. Yu, 2008. *Identifying the influential bloggers in a community*. In Proceedings of the international Conference on Web Search and Web Data Mining (Palo Alto, California, USA, February 11 - 12, 2008). WSDM '08. ACM.

Anderson, Paul, 2007. *What is Web 2.0? Ideas, technologies and implications for education.* Technical report, JISC.

Chi, Yun, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng, 2007. *Structural and temporal analysis of the blogosphere through community factorization.* In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Jose, California, USA, August 12 - 15, 2007). KDD '07. ACM.

Fujimura,Ko, Takafumi Inoue, and Masayuki Sugizaki, 2005. *The EigenRumor Algorithm for Ranking Blogs.* In Proceedings of the WWW 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Gruhl, Daniel, R. Guha, David Liben-Nowell, and Andrew Tomkins, 2004. *Information diffusion through blogspace.* In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM.

Hassan, Ahmed, Dragomir Radev, Junghoo Cho and Amruta Joshi, 2009. *Content Based Recommendation and Summarization in the Blogosphere*. Third International AAAI Conference on Weblogs and Social Media, AAAI Publications.

Honeycutt, Courtenay, and Susan C. Herring, 2009. *Beyond microblogging: Conversation and collaboration via Twitter*. In Proceedings Hawaii International Conference on System Sciences, IEEE Press

Kumar, Ravi, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins, 2003. *On the bursty evolution of blogspace.* In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM.

Leskovec, Jure, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Van-Briesen, and Natalie Glance, 2007. *Costeffective outbreak detection in networks*. In The 13th International Conference on Knowledge Discovery and Data Mining (KDD) 2007. ACM.

Li, Hui, Sourav S. Bhowmick, and Aixin Sun, 2009. *Blog cascade affinity: analysis and prediction*. In Proceeding of the 18th ACM Conference on information and Knowledge Management (Hong Kong, China, November 02 - 06, 2009). CIKM '09. ACM.

Mishne, G. and Maarten de Rijke, 2006. *A Study of Blog Search*. In Proceedings of ECIR-2006. LNCS vol 3936. Springer.

Nakajima,Shinsuke, Junichi Tatemura, Yoichiroara Hino, Yoshinori Hara and Katsumi Tanaka, 2005. *Discovering Important Bloggers based on Analyzing Blog Threads.* In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM.

O'Reilly, Tim, 2005. *What is Web 2.0: Design Patterns and Business Models for the next generation of software.*

Page, Larry, Sergey Brin, Rajeev Motwani and Terry Winograd, 1999. *The PageRank citation ranking: Bringing order to the Web*, Technical Report, Stanford University.

Qureshi, M. Atif, Arjumand Younus and Francisco Rojas, 2010. *Analyzing Web Crawler as Feed Forward Engine for Efficient Solution to Search Problem in the Minimum Amount of Time through a Distributed Framework.* In Proceedings of 1st International Conference on Information Science and Applications, IEEE Publications.

Sifry, David, 2009 Sifry's Alerts. *http://www.sifry.com/alerts/*

Sorapure, Madeleine. 2003. *Screening moments, scrolling lives: Diary writing on the web.* Biography: An Interdisciplinary Quarterly, 26(1), 1-23.