Content Based Recommendation and Summarization in the Blogosphere

Ahmed Hassan and **Dragomir Radev**
Department of EECS
University of Michigan
Ann Arbor, MI 48109
hassanam,radev@umich.edu

Junghoo Cho and **Amruta Joshi**
Department of Computer Science
University of California
Los Angeles, CA 90095
cho,amrutaj@cs.ucla.edu

Abstract

This paper presents a stochastic graph based method for recommending or selecting a small subset of blogs that best represents a much larger set. within a certain topic. Each blog is assigned a score that reflects how representative it is. Blog scores are calculated recursively in terms of the scores of their neighbors in a lexical similarity graph. A random walk is performed on a graph where nodes represent blogs and edges link lexically similar blogs. Lexical similarity is measured using either the cosine similarity measure, or the Kullback-Leibler (KL) divergence. In addition, the presented method combines lexical centrality with information novelty to reduce redundancy in ranked blogs. Blogs similar to highly ranked blogs are discounted to make sure that diversity is maintained in the final rank. The presented method also allows to include additional initial quality priors to assess the quality of the blogs, such as frequency of new posts per day and the text fluency measured by n-gram model probabilities, etc. We evaluate our approach using data from two large blog datasets. We measure the selection quality by the number of blogs covered in the network as calculated by an information diffusion model. We compare our method to other heuristic and greedy selection methods and show that it significantly outperforms them.

Introduction

Recent years have witnessed a tremendous growth of the blogosphere. The size of the collection of blogs on the World Wide Web has been lately exhibiting an exponential increase. Blogs are now one of the main means for spread of ideas and information throughout the Web. They discuss different trends, ideas, events, and so on. This gave rise to an increasing interest in analyzing the blogosphere by the Information Retrieval (IR) community.

A weblog (blog) is a website maintained by an individual who uses it as a self-publishing media by regularly publishing posts commenting on or describing some event or topic. Blogs made it easy for everybody to publish, read, comment, and share ideas.

Blogs are different in style when compared to traditional web pages. A blog is usually written by an individual person and organized in a set of posts. Blogs tend to be affected by each others. Some blogs start introducing new information and ideas that spreads down to other blogs.

One of the most interesting problems in the Blogosphere is how to provide the Internet users with a list of particularly important blogs with recurring interest in a specific topic. Several blog search engines are used to search blog contents. They return a list of blogs relevant to a particular topic. However, this list of relevant blogs is usually very large making it hard to select a small set of blogs with interest in a particular topic.

Information needs in blog searching and ranking differ substantially from that of conventional Web search (Mishne and de Rijke 2006). This gave rise to targeting research efforts towards new methods for ranking blogs. (Mishne and de Rijke 2006) reports that, unlike Web search, most blog search queries are informational (looking for information about some topic) and that most blog search queries aim at locating blogs and blog posts which focus on a given concept or topic.

In this work, we propose a graph based method based on random walks and lexical centrality for identifying a small representative set of blogs with interest in a specific topic.

One approach to ranking blogs and blog posts is to use the same techniques applied to traditional web pages. However, techniques developed for traditional web pages might not be appropriate for blogs.

The most well established techniques for ranking traditional web pages are the link popularity based algorithms like PageRank (Page *et al.* 1999) and Hypertext Induced Topic Selection (HITS) (Kleinberg 1998) algorithms. PageRank assigns a numerical weight to each web page according to the links it receives and the weights of the pages that link to it. HITS determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

One of the reasons why link popularity based algorithms might not work well for blogs is the weakly linked nature of blog pages. For example, Leskovec et al. report that the average number of links per blog post is only 1.6 links (Leskovec *et al.* 2007). They collected a total of 45,000 blogs and 10 million posts, among which they found only 16.2 million links. This small number of links per blog post results in a very sparse network especially when trying to build a network of blogs that are relevant to a particular topic.

Another reason is that blog posts have a relative short life span when compared to traditional web pages. For example, a good blog post may eventually receive a lot of good links pointing to it. However, we will not able to assess how good this post is until it receives those links and that will need time, and by that time, users may not be interested in it anymore.

Yet another reason for why link popularity based algorithms might not work well with blogs is that bloggers try to exploit the system to boost the rank of their blogs. For example, they trade and even sell and buy links to each other to make sure that their blogs get as high rank as possible.

Instead of considering hyperlinks to measure the authority of blogs, the proposed method uses text similarity between blog posts to select a small representative set of blogs given a larger number of blogs relevant to a particular topic. The method is based on the hypothesis that important or representative blogs tend to lexically similar to other important blogs. Hence, we may define the blog score as a recursive function of the scores of its neighbors. Iteratively updating those scores will lead to a score that assigns higher ranks to more representative blogs. This is equivalent to a random walk over the network of blogs. The score of each blog is equivalent to the amount of time the random walk surfer spends on each node. The proposed method also takes diversity into consideration while assigning ranks to different blogs. The graph based methods discount blogs similar to highly ranked blogs by decreasing their influence on the rest of nodes in the graph. This reduces the transition probability to those nodes in the random walk and hence allows other diverse node to receive higher ranks. Although graph based centrality is a very good measure for predicting the importance of blogs, it is not the only measure. Other measures that address the quality of the blog itself, regardless of other blogs, may also affect the blog rank. To take this into consideration, we introduce a modification to our graph based approach to take initial quality priors into consideration when calculating the node centrality. This allows the method to bias the selection to blogs with certain properity. For example, we can bias the selection to blogs with higher rate of posts, more fluent text,...etc.

The rest of the paper will proceed as follows. We first present an overview of related work. We then describe our graph based approach for blog ranking. Next, we describe the different measures we use to measure text similairty, how diversity can be taken into consideration during the ranking process, and how initial quality priors can be used to capture the base quality of nodes. The following section describes experiments and results. Finally, we present conclusions in the last section.

Related Work

The problem of weblog (blog) ranking or selecting which blogs to read has been lately receiving increasing attention. The problem is different from from traditional document ranking in ad-hoc information retrieval in several ways (Arguello *et al.* 2008). For example, a blog is not a single document, rather it is composed of a collection of documents. Moreover, blog queries always represent an interest in a particular topic, rather than a passing information need (Arguello *et al.* 2008). Hence, specific methods has been developed to target this problem.

Leskovec et al. (Leskovec *et al.* 2007) present a method detecting outbreaks in a network. To detect outbreaks, they select nodes in anetwork that can detect the spread of a virus or information as quickly as possible. They propose an algorithm that can detect outbreaks based on optimization of submodular function. They use their method to find the best locations for sensor placement in water distribution networks to quickly detect contaminants. They also extend their method to detecting which blogs one should read to catch all important stories.

TREC 2007 Blog track (Macdonald, Ounis, and Soboroff 2007) presented a new blog distillation track which is quite related to the problem we are addressing. In their formulation, the problem is to find blog feeds with a principal, recurring interest in X, where X is some information need expressed as a query (Macdonald, Ounis, and Soboroff 2007). The input to such a system is a query and the output is a ranked list of blog feeds. The CMU system (Elsas *et al.* 2007) achieved the best performance in the 2007 track. Their approach depends on indexing both feeds and individual posts. They also use a Wikipedia-based query expansion approach to improve the results.

Lin et al. (Lin and Cohen 2008) present a semi-supervised method for classifying political blogs in a blog network into liberal or conservative and ranking them within each predicted class. They use PageRank (Brin and Page 1998) to determine the importance or authority of a blogs. However, they modify the algorithm such that authority scores propagate only within blogs belonging to the same community.

Arguello et al. (Arguello *et al.* 2008; Elsas *et al.* 2007) present a system for blog ranking and recommendation. Their system compares different blog retrieval models that view either the blogs or the posts as retrieval units. They also use Wikipedia for query expansion to further improve queries.

Java et al. (Java and Oates 2007) study the feeds subscribed by a set of publicly listed Bloglines users. They use the subscription information to come up with feeds topics and feeds that matter for particular topics. Using the Bloglines subscription data, they identify feeds that are popular to a given topic. Topics are approximated by the folders names and merging related folders together.

Song et al. (Song *et al.* 2007) proposes a method for identifying opinion leaders in the Blogosphere. They define opinion leaders as those who bring in new ideas and opinions. They rank blogs according to how novel the information they contribute to the network.

Marlow et al. (Marlow 2007) uses blogroll links and permalinks to predict the authority and influence of blogs. The study shows that hyperlinks between blogs can be used to track influence, however it does not well when the problem is restricted to finding important blogs within a particular topic.

Several methods have been proposed for identifying the most central nodes in a network. Some of the most popular methods, for measuring centrality of nodes in a network, are degree centrality, closeness, and betweenness (Newman 2003).

Another powerful method for measuring centrality in a network is eigenvector centrality. This method has been successfully applied to several types of networks. For example it has been used to measure centrality in hyperlinked web pages networks (Brin and Page 1998; Kleinberg 1998), lexical networks (Erkan and Radev 2004; Mihalcea and Tarau 2004; Kurland and Lee 2005; 2006), and semantic networks (Mihalcea, Tarau, and Figa 2004).

BlogRank

In this section we describe how to assign relative weights or ranks to a set of blogs based on the content of the blogs. The method we used is similar to the methods described in (Erkan and Radev 2004; Mihalcea and Tarau 2004; Kurland and Lee 2005), which were originally used for ranking sentences and documents in extractive summarization and information retrieval systems.

The main hypothesis behind this method is that important or representative blogs tend to be lexically similar to other important or representative blogs. Hence, we can use text similarity to link blogs or blog posts to each other.

Consider two blogs or blog posts p, and q such that p and q are lexically similar to each other. That will result in a link between p and q. This link is suggesting that p and q share a common topic of interest. And it is also suggesting that they may have been affected by each others and that the textual similarity is a way of conferring authority between them. Note that this is different from hyperlink based authority where p may simply ignore to add a link to q or add a non informative link to q based on some link-exchange agreements. Another advantage is that the assessment of q's quality is independent of the textual content of q. Hence, this assessment is completely out of q's control which would make the technique more immune to spamming.

A collection of blogs can be represented as a network where similar blogs are linked to each other. The proposed method is based on the premise that important blogs tend to be lexically similar to other important blogs. Or in a finer level of granularity, important blog posts tend to be lexically similar to other important posts, and important posts tend to belong to important blogs.

Hence given a collection of blog posts that are related to a specific topic and a similarity measure, we can build a network a network where each blog is represented by a node and edges link blogs if their textual similarity exceeds some threshold. The edges of the network are weighted with the weight representing how similar the text of the two blogs is to one another.

Given this network, we can define the importance score of a blog recursively in terms of the scores of other similar blogs. This can also be implemented in a lower level of granularity where nodes of the graph represent a single blog post, rather than a blog. In this case, each post is assigned an importance score. The importance score of a blog can then be calculated by taking the average of the scores of all its posts. In the former case where nodes represent blogs, the importance score is directly assigned to the blog.

When building a network of posts, we only consider two posts similar if they belong to two different feeds. This makes sure that posts within the same feed are not connected. Hence, a blog feed cannot gain credit by having several posts similar to each others.

The recursive definition of the score of any blog b in the blogs network is given by:

$$p(b) = \sum_{t \in adj[b]} \frac{p(t)}{deg(t)} \tag{1}$$

where $deg(t)$ is the degree of node t, and $adj[b]$ is the set of all blogs adjacent to b in the network. This can be rewritten in matrix notation as:

$$\mathbf{p} = \mathbf{pB} \tag{2}$$

where $\mathbf{p} = (p(b_1), p(b_2), \ldots, p(b_N))$ and the matrix $\mathbf{B}$ is the row normalized similarity matrix of the graph

$$\mathbf{B}(i,j) = \frac{\mathbf{S}(i,j)}{\sum_k \mathbf{S}(i,k)} \tag{3}$$

where $\mathbf{S}(i,j) = \sim (b_i, b_j)$. Equation (2) shows that the vector of salience scores $\mathbf{p}$ is the left eigenvector of $\mathbf{B}$ with eigenvalue 1.

The matrix $\mathbf{B}$ can be thought of as a stochastic matrix that acts as the transition matrix of a Markov chain. An element $\mathbf{X(i,j)}$ of a stochastic matrix specifies the transition probability from state i to state j in the corresponding Markov chain. And the whole process can be seen as a Markovian random walk on the speeches graph. To help the random walker escape from periodic or disconnected components, (Brin and Page 1998) suggests reserving a small escape probability at each node that represents a chance of jumping to any node in the graph, making the Markov chain irreducible and aperiodic, which guarantees the existence of the eigenvector.

Equation (2) can then be rewritten, assuming a uniform escape probability, as:

$$\mathbf{p} = \mathbf{p}[d\mathbf{U} + (1-d)\mathbf{B}] \tag{4}$$

where N is the total number of nodes, $\mathbf{U}$ is a square matrix with $\mathbf{U}(i,j) = 1/N$ for all i, j, and d is the escape probability chosen in the interval $[0.1, 0.2]$ (Brin and Page 1998).

Similarity Functions

The most popular similarity function used to measure document similarity is the well-known cosine measure defined on the document vectors in the tf or tf-idf weighted term space. Some other possible similarity measures are edit distance, Kullback-Leibler (KL) divergence (Lafferty and Zhai 2001), language models (Kurland and Lee 2005), or generation probabilities (Erkan 2006).

To measure the similarity between two blogs, we can use the bag-of-words model to represent each sentence as an N-dimensional vector of tf-idf scores, where N is the number of all possible words in the target language. The similarity between two blogs is then computed using the cosine similarity between the two vectors.

Kullback-Leibler (KL) divergence (Lafferty and Zhai 2001) is another popular measure used in information retrieval. KL divergence is a measure of the distance between two distributions. In information retrieval, the KL divergences between the language model of a query and the language models of each document are computed to find the similarity between the query and each document. KL is a distance measure. Hence, to measure similarity we have to use the negative or the inverse of KL.

Diversity Ranking

Suppose we already identified a node x as the most important node in the graph. Now, we would like to identify the second important node such that it is important and meanwhile as diverse as possible with respect to the first selected node. The importance of a node is usually calculated as:

$$p(b) = \sum_{t \in adj[b]} \frac{p(t)}{deg(t)} \tag{5}$$

where $p(u)$ is the importance of node u, $adj[u]$ is the set of nodes that are adjacent to u, and $deg(v)$ is the degree of the node v.

The problem with this formula is that blogs very similar to x will benefit from their connection to x and hence receive high ranks. However those blogs are probably quite redundant and do not have any new information.

To solve this problem, we will modify the above formula as follows:

$$p(b) = d(b) \sum_{t \in adj[b]} \frac{p(t)}{deg(t)} \tag{6}$$

$$d(b) = 1 - sim(b, x) \tag{7}$$

Where $d(u)$ is a discounting factor to penalize nodes that are similar to the already selected node, and $sim(u, x)$ is the similarity between nodes u and x.

How about the case of more than one important node? For example, what should we do when we are trying to select the third or the fourth or even the nth node? Or when x is set of nodes rather than a single node.

The formula for calculating d(u) can be modified as follows:

$$d(b) = 1 - max_{\forall x_i} sim(b, x_i) \tag{8}$$

In this way, each node is penalized with respect to the closest node to it that was already selected.

Adding Priors

Node importance calculated from the a lexical network is a good measure for determining importance or coverage of blogs. However, it might not be the only attribute that affects blog quality. Other attributes that are more related to the blog itself, rather than to its position in a network of blogs, might also be involved. Some of those attributes are the fluency of the text (e.g., n-gram model probabilities), the formatting,the frequency of updates, the use of particular vocabulary, the average length of posts, the number of posts, etc.

To incorporate these attributes into our graph based approach, we propose a modification to method to allow it to take initial node quality priors into consideration.

Assume we have a certain node quality measure that uses one or more of the features we mentioned above. Let's define a vector$Q$ of priors where $Q = q_1, \ldots, q_n$, where n is the number of nodes in the network. Before going on we further normalize Q such that the sum of all entries in Q is 1.

$$q_{i_{norm}} = \frac{q_i}{\sum_j q_j} \tag{9}$$

Let's also define a trade-off factor β, where $0 \leq \beta \geq 1$, that controls the weight we assign to the priors vs the weight we assign to the centrality scores.

Now we can redefine the formula as follows:

$$p(b) = (1 - \beta) \sum_{t \in adj[b]} \frac{p(t)}{deg(t)} + \beta * q_{norm} \tag{10}$$

Several attributes can be used toward calculating the initial quality prior. Each of those attributes may be used to bias the solution towards a specific property. As mentioned earlier, we can use priors to favor blogs with higher number of posts, more fluent text, etc.

Experiments

Data

We used two large test collections through our experiments, the BLOG06 dataset created by the University of Glasgow, and the UCLA Blogocenter dataset.

BLOG06 is a TREC test collection, created and distributed by the University of Glasgow. It contains a crawl of Feeds, and associated Permalink and homepage documents (from late 2005 and early 2006)(Macdonald and Ounis 2006). The dataset contains 100,649 feeds covering an 11 weeks period. Permalink documents and blog homepages were also collected. The dataset contains a total of 3,215,171 permalink documents and 324,880 home page documents (Macdonald and Ounis 2006). To make the collection more realistic, a list of known spam blogs was also included. 17,969 spam blogs were added causing the spam component to form a reasonable component of the collection (Macdonald and Ounis 2006).

The other dataset is a massive dataset built by the The Blogocenter group at UCLA. They have been retrieving RSS feeds from the Bloglines, Blogspot, Microsoft Live Spaces, and syndic8 aggregators for the past several years. The dataset contains over 192 million blog posts (Ka Cheung Sia 2008).

Evaluations Metrics

We borrow an idea from the studies of the spread of influence in social networks to evaluate our method's results.

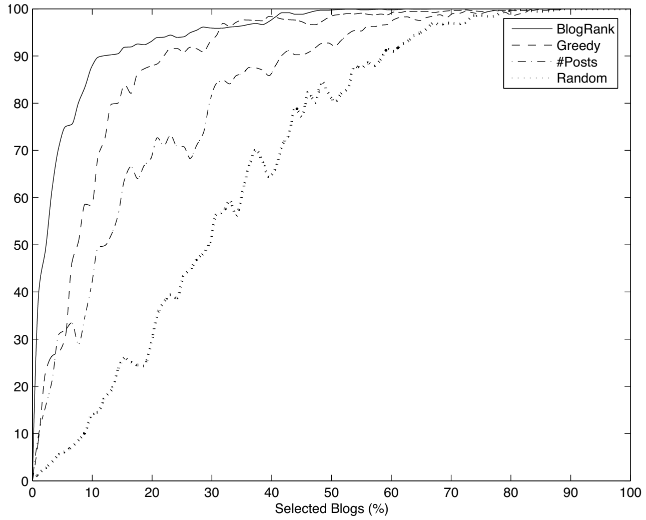

Figure 1: Percentage of covered blogs vs percentage of selected blogs for BlogRank order, greedy algorithm order, number of posts order, and random order - Topic: Global Warming.

Diffusion models for the spread of an idea in a social network consider each node as either active or inactive. Inactive nodes tend to become active as more of its neighbors become active. One of the earliest models that captures this process was proposed in (Granovetter. 1978). Several other models was then presented to capture the same process. At the core of all those models lies the Linear Threshold Model (Kempe, Kleinberg, and Tardos 2003). This model assumes that the influence of each node by any of its neighbors is a function of the weight of the edge connecting the node to that neighbor. The diffusion process starts by a set of active nodes and a threshold θ_v for each node v selected uniformly at random. At each step, any node v, for which the sum of the weights of the edges connecting it to its active neighbors exceeds its threshold, is activated.

$$\sum_{w \in adj[v]} b_{v,w} \geq \theta_v \tag{11}$$

The thresholds θ_v are randomly selected to reflect the lack of knowledge of their values (Kempe, Kleinberg, and Tardos 2003)

As stated in previous sections, the goal of the proposed method is to select a small set of blogs M that summarizes or covers most of the information content in a larger set of blogs N. Each node (blog) in the network will be considered as either active or inactive. An active node is a node that has been covered by the small selected set M. An inactive node is a node that has not yet been covered. We also assume that the node's tendency to become covered increases as more of its neighbors become covered. Or in other words, a node's information content is more likely to become covered as more and more of the information contents of its neighbors is covered.

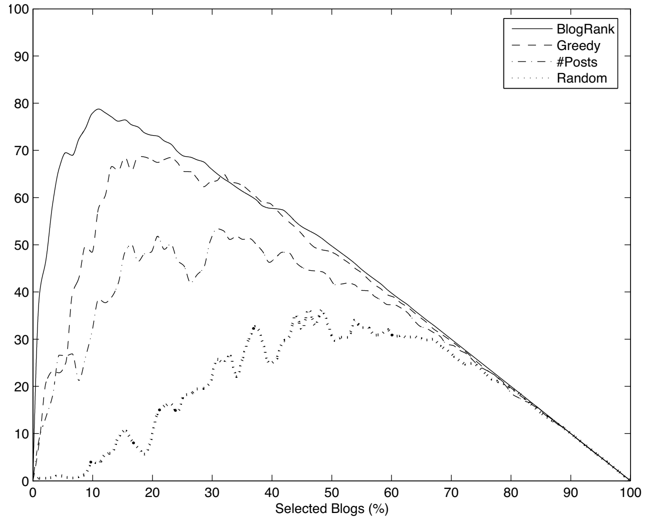

Figure 2: Percentage of activated blogs vs percentage of selected blogs for BlogRank order, greedy algorithm order, number of posts order, and random order - Topic: Global Warming.

We can use this model to evaluate the selection of the smaller set M with respect to the bigger set N. The quality of a selected set of feeds is evaluated by the number of nodes that become active in the blogs network when the selected feeds are initially designated as active. The output of the proposed method is a ranked set of blogs such that a blog with higher rank is a better candidate for inclusion in the selected set than a blog with a lower rank. Given a ranked list of blogs R, we evaluate all subsets M_i where $M_i = R[1-i]$ for all i in $[1 \text{ to } |R|]$.

Due to the randomness inherent in selecting the nodes activation threshold, we repeat the evaluation n times and report the average of the results of the n runs.

Comparison to Other Systems

We compare our method to several other methods of blog selection. The first method is a simple method that select blogs for inclusion in the selected set uniformly at random.

The second method is one of a family of heuristic selection techniques that try to find the most popular blogs by using some fixed "goodness" criteria. The criterion we use here is the number of posts in a blog. Whenever we want to select a subset of blogs M with $|M|$ blogs, we select the $|M|$ blogs with the highest number of posts.

The third method is based on a greedy algorithm that uses hill climbing search. The method greedily adds the node that maximizes the marginal gain to the selected set. Initially we start with an empty set $M_0 = \phi$. At each step i, we add the node that maximizes the marginal gain $b = \arg\max_{\text{non-selected blogs}} C(M_{i-1} \cup b) - C(M_{i-1})$ where $C(M_i)$ is the quality of the subset M_i measured by the number of nodes it covers in the blog network. So the method tries to maximizes the gain in the evaluation metric described in the previous section.

Results

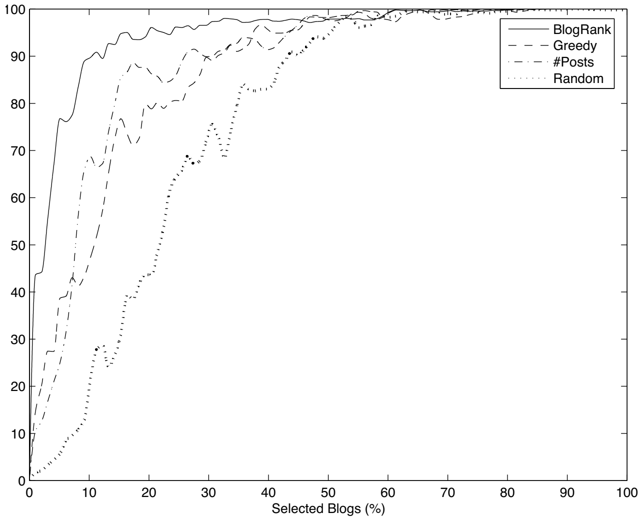

Figure 3: Percentage of covered blogs vs percentage of selected blogs for BlogRank order, greedy algorithm order, number of posts order, and random order - Topic: iPhone.

We compare the performance of the proposed method to a random selection method, a heuristic selection method and a greedy method. Figure 1 shows the percentage of covered blogs vs. percentage of selected blogs for the proposed method (BlogRank), the greedy method, the heuristic (number of posts) method, and a random method for the ”Global Warming” topic. Figure 2 compares the performance of the same methods on the same topic with respect to the percentage of activated blogs vs percentage of selected blogs. The difference between Figure 1 and Figure 2 is that Figure 2 does not count the selected blogs when calculated the number of covered blogs (i.e. it only shows blogs that were not selected yet were covered by other selected nodes). Figure 3 and Figure 4 show similar results for the ”iPhone” topic.

We notice that random blog selection performs the worst. We also notice that heuristic blog selection based on the number of posts does not perform very well either. Selection based on the greedy based method outperforms random and heuristic selection. The figures also show that the proposed method outperforms all other methods with a considerable amount of improvement.

These results show than random selection and heuristic based selection for blogs that summarize or cover the information content of a particular topic do not work well.Greedy methods are better than random and heuristic methods, however they are very computationally costly ($O(|V|^4)$, where $|V|$ is the number of nodes). The proposed method outperforms all other methods and at the same time it is much faster than the greedy method.

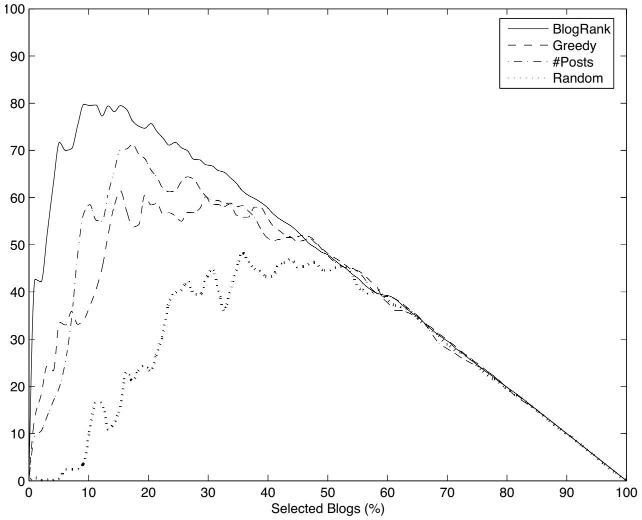

Figure 4: Percentage of activated blogs vs percentage of selected blogs for BlogRank order, greedy algorithm order, number of posts order, and random order - Topic: iPhone.

Generalization to Future Data

The quality of a blog in summarizing or covering the information content in the Blogosphere with respect to some particular topic may change with time. Hence, we need to evaluate how good our method is in predicting the quality of a particular blog. We split our data into two smaller datasets. The first part represents ”History” and covers the first half of the time period. The second part represents ”Future” and covers the second half. We now have two networks for each topic (history and future). We use our method to rank blogs using only the history network and evaluate it on the future network. We also use our method to rank blogs using all the data set, but evaluate it on the future network. Finally we use the greedy method to rank blogs using only the history network and evaluate it on the future network.

Figure 5 compares the performance of the proposed method when tested on known data, the proposed method when tested on unknown future data, and the greedy method when tested on unknown future data for the ”Global Warming” topic. Figure 6 shows the same comparison for the ”iPhone” topic. We notice that the gaps between the curves of the proposed method evaluated on known and unknown data is small. This suggests that the proposed method generalizes well for the future data. On the other hand, we see that the greedy method seems to overfit when evaluated on the future data.

Other Experiments

All the experiments reported in this section use the cosine similarity metric to measure text similarity. This is an arbitrary choice and any other text similarity measure may be used. For example, we tried using the KL divergence to measure text similarity. We compared the performance of the method when using cosine similarity and KL based similar-

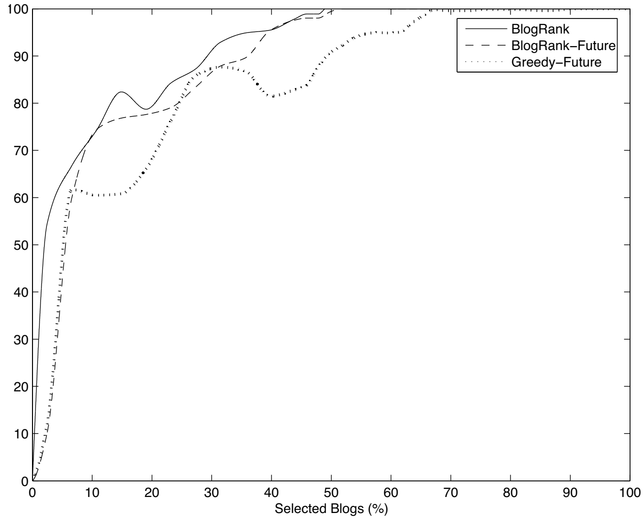

Figure 5: Percentage of activated blogs vs percentage of selected blogs for BlogRank order (learned from all data and evaluated on future data),BlogRank-Future order (learned from history data and evaluated on future data), and Greedy-Future order (learned from history data and evaluated on future data) - Topic: Global Warming.

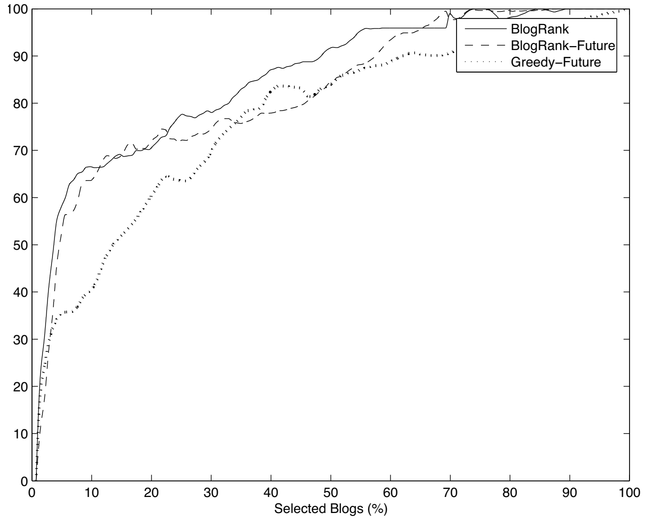

Figure 6: Percentage of activated blogs vs percentage of selected blogs for BlogRank order (learned from all data and evaluated on future data),BlogRank-Future order (learned from history data and evaluated on future data), and Greedy-Future order (learned from history data and evaluated on future data) - Topic: iPhone.

ity and the difference in performance was negligible. One advantage that cosine similarity has over KL based similarity is that it is a symmetric measure which reduces the number of similarity measures by half.

Another experiment that we performed was using priors to bias the blog selection towards some property. For example, we tried using the number of posts as a prior. This led to a solution with almost the same quality as shown in Figure 7. If we examine Figure 8 which compares the average number of posts for the two solutions with respect to the percentage of selected blogs, we notice that the new solution has larger average number of posts, especially for selected set with smaller sizes. This difference decreases till they become equal when all blogs are selected.

This shows that we can bias our solution to a specific property using the initial priors without losing much of the solution quality. For example, we can bias our solution to blogs with larger number of posts, smaller number of posts, longer posts, fluent text (as measured by language model scores), etc.

Conclusions

The size of the collection of blogs on the World Wide Web has been lately exhibiting an exponential increase. This gave rise to an increasing interest in methods for blog ranking that can provide the Internet users with a list of particularly important blogs with recurring interest in a specific topic. In this work, we presented a stochastic graph based approach for selecting a small set of blog feeds that best represent a larger set of feeds within a given topic. The approach is based on lexical similarity and random walks. The proposed method takes diversity into consideration when measuring blog importance by penalizing blogs quite similar to already selected blogs. It may also benefit from additional initial priors to bias the selection towards blogs with a specific property such as frequency of new posts per day and the text fluency . We showed that the proposed method achieves promising results and outperforms other random, heurisitic, and greedy selection methods. We also showed that the method performs well when tested on unseen future data.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. 0534323, ”Collaborative Research: BlogoCenter - Infrastructure for Collecting, Mining and Accessing Blogs”. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Arguello, J.; Elsas, J. L.; Callan, J.; and Carbonell, J. G. 2008. Document representation and query expansion models for blog recommendation. In *International Conference on Weblogs and Social Media 2008*.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. 30(1–7):107–117.

Elsas, J.; Arguello, J.; Callan, J.; and Carbonell, J. 2007. Retrieval and feedback models for blog distillation. In *The Sixteenth Text REtrieval Conference (TREC 2007)*.

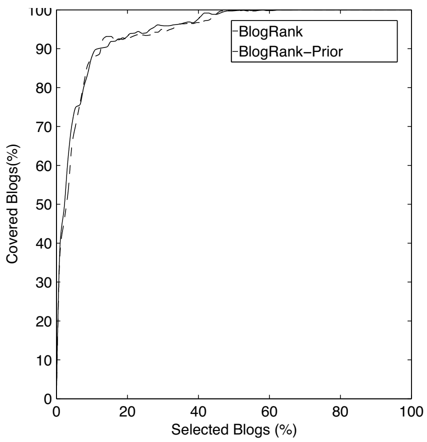

Figure 7: Percentage of covered blogs vs percentage of selected blogs for BlogRank order and BlogRank with priors order - Topic: Global Warming.

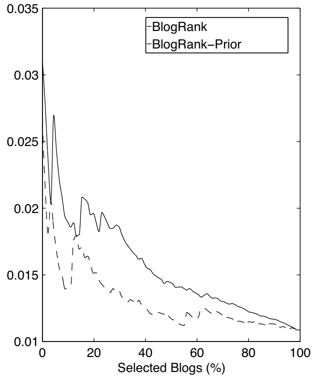

Figure 8: Average normalized number of posts vs percentage of selected blogs for BlogRank order and BlogRank with priors order - Topic: Global Warming.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Erkan, G. 2006. Language model-based document clustering using random walks. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 479–486. New York City, USA: Association for Computational Linguistics.

Granovetter., M. 1978. Threshold models of collective behavior. *American Journal of Sociology*.

Java, A.; Kolari, P. F. T. J. A., and Oates, T. 2007. Feeds that matter: A study of bloglines subscriptions. In *ICWSM*.

Ka Cheung Sia, Junghoo Cho, Y. C. B. L. T. 2008. Efficient computation of personal aggregate queries on blogs. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD 2003*.

Kleinberg, J. 1998. Authoritative sources in a hyperlinked environment. *9th ACM-SIAM Symposium on Discrete Algorithms*.

Kurland, O., and Lee, L. 2005. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of SIGIR*, 306–313.

Kurland, O., and Lee, L. 2006. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of SIGIR*, 83–90.

Lafferty, J., and Zhai, C. 2001. Document language models, query models, and risk minimization for information retrieval. In *The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; Glance, N.; and BuzzMetrics, N. 2007. Cost-effective outbreak detection in networks. In *The 13th International Conference on Knowledge Discovery and Data Mining (KDD) 2007*.

Lin, F., and Cohen, W. 2008. The multirank bootstrap algorithm: Semi-supervised political blog classification and ranking using semi-supervised link classification. In *International Conference on Weblogs and Social Media 2008*.

Macdonald, C., and Ounis, I. 2006. The trec blogs06 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow.

Macdonald, C.; Ounis, I.; and Soboroff, I. 2007. Trec 2007 blog track. In *The Sixteenth Text REtrieval Conference (TREC 2007)*.

Marlow, C. 2007. Audience, structure, and authority in weblog community. In *The 54th Annual Conference of the International Communication Association*.

Mihalcea, R., and Tarau, P. 2004. TextRank: Bringing order into texts. In *EMNLP2004*.

Mihalcea, R.; Tarau, P.; and Figa, E. 2004. Pagerank on semantic networks, with application to word sense disambiguation. 1126–1132.

Mishne, G., and de Rijke, M. 2006. A study of blog search. In *ECIR 2006*.

Newman, M. E. J. 2003. A measure of betweenness centrality based on random walks. Technical Report condmat/0309045, Arxiv.org.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford Digital Library Technologies Project, Stanford University.

Song, X.; Chi, Y.; Hino, K.; and Tseng, B. 2007. Identifying opinion leaders in the blogosphere. In *The sixteenth ACM conference on Conference on information and knowledge management*.