Homepage Search in Blog Collections

Jangwon Seo
jangwon@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

ABSTRACT

A blog homepage consists of many individual blog postings. Current blog search services focus on retrieving postings but there is also a need to identify relevant blog homepages. In this paper, we investigate the properties of blog collections and describe the differences between blog homepage searches and general web page searches. We also introduce and evaluate a variety of approaches for blog homepage search. Our results show that noise reduction and the appropriate combination of techniques can achieve significant improvements in retrieval performance compared to a baseline approach and a traditional named page finding approach for general web pages.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: General

General Terms

Algorithms, Measurement, Experimentation

Keywords

Information Retrieval

1. INTRODUCTION

Weblogs or blogs are an increasingly popular method of transmitting personal opinions and views, and the scale of the "blogosphere" has grown dramatically. While blogs share some similar features with traditional web pages, they also have distinct characteristics in that they have structural features to help users continuously generate content as well as generally subjective content created by expressing personal opinions with no editing. As the blogosphere grows, search techniques customized for blogs are needed to identify relevant material amongst the enormous amount of blog "noise". The creation of the TREC 2006 Blog track represents an effort in that direction.

Most research on blog retrieval has focused on blog postings. For example, although Google provides search results like 'Related Blog' in addition to postings in its 'BlogSearch'[1] service, postings are still considered as the main target of blog retrieval by most of the search engine services. One reason for this is that most researchers and service providers view blog searches and general web page searches as being essentially the same thing.

On the other hand, as blog subscription methods such as RSS and ATOM have become more prevalent, it is important to be able to identify relevant blogs as well as blog postings. Further, many blogs address a small number of specific topics related to the user's interests rather than being completely general. Therefore, if there is a relevant blog related to a specific topic, then that blog is likely to consistently generate good quality postings about the topic.

In this paper, we will focus on search techniques for complete blogs rather than postings. Since the term 'blog search' often means 'posting search', we refer to each blog composed of its own postings as a blog 'homepage'. Furthermore, 'Blog homepage search' will be used in order to refer to techniques for retrieving blog homepages. We show that there are considerable differences between blog collections and general web collections, and that customized techniques for blog homepage search are the most effective.

In the next section, we briefly survey related work on blog search. Next, we review the properties of the blog collections used in our experiments. This includes an investigation of the amount of "noise" and duplication in these collections, and the results of applying techniques to remove spurious text. Then, we will introduce several techniques for blog homepage search and compare them to standard retrieval techniques for web pages and for text passages. We conclude with a discussion of the results.

2. RELATED WORK

We are not aware of any previous research on blog homepage search. However, there has been some work focusing on blog posting search. Mishne and de Rijke[16] showed that blog posting searches have different goals than general web searches by analysing blog search engine query logs. In order to help search engines index postings, Glance[8] introduced a blog segmentation method using a combination of blog feeds and model-based wrapper segmentation. In the TREC 2006, the blog track appeared as a new task. The main task of the track was "opinion retrieval", i.e. to locate blog postings that express an opinion about a given target.

[1] http://blogsearch.google.com

3. PROPERTIES OF BLOG COLLECTIONS

3.1 The Collections

We used two blog collections for our experiments. The primary one for our retrieval experiments is the TREC Blogs06 Collection [13]. The collection was crawled by the University of Glasgow from December 6, 2005 to February 21, 2006 and contains 3,215,171 postings and 46,001 unique blog homepages. In this collection, there are a large number of spam blogs intentionally included, and each blog page contains "noise" or spurious text such as menu frames and advertisements. For the retrieval experiments, we used queries from the TREC Web track and did our own relevance judgments, as described in Section 5.1.

The second collection is the ICWSM Blog Collection[2]. This was collected by Nielsen BuzzMetrics for participants in the International Conference on Weblogs and Social Media. It contains 14 million postings and 3 million blog homepages. This collection was only used in our investigation of noise and duplication in blog collections. It is relatively "clean" compared to the TREC Blogs06 Collection in that it contains much less spurious text, as will be shown in our experiments.

3.2 Structural Noise

Any set of documents created by a large number of individuals will contain various kinds of noise and spurious text that will deteriorate retrieval performance. Blog collections are a good example of this. In blog collections, we can broadly identify two kinds of noise. One is spam. In fact, since blog pages are also web pages, it is natural that blog collections will inherit the noise features of web pages including spam. A blog itself can be a spam blog composed only of spam postings, or one posting could be spam. Given that spam in blogs will be similar to spam in the general web, we assume that this noise can be handled by general spam handling techniques.

The second type of noise is more specific to blogs in that it comes from the page structures of blogs. In most blogs, menubar frames or header material for the blog are located on the top or the side of each page, whereas the main body of each posting is located in the center of each page. Furthermore, we can easily discover that advertising links like Google AdSense[3] occupy the corners of each page. Although it is risky to over-generalize the structure of blog pages, most blogs resemble each other because they are created by blog publication software or blog service providers. The page structure of a blog is repeated over most of the postings of the blog, and as a result, similar texts in the menubar or the advertisements are frequently repeated. Therefore, such patterns can become important noise in blog retrieval because retrieval techniques are generally based on statistical features of text such as the term frequency count (tf) or inverse document frequency (idf). We refer to this noise as a structural noise.

3.2.1 Structural Noise Reduction

Although repeated text over postings can be an important source of noise, it is not desirable to simply remove it all because some duplication is not noise, such as quotes or copying news articles. In order to reduce structural noise we focus on removing the non-content parts of the blog. A content selection algorithm that extracts the content body from an HTML page can be a good tool for structural noise reduction [7, 17]. The algorithm exploits the fact that there are fewer HTML tags in the content part than in the remaining parts of a web page. By mapping the number of tags versus text tokens, we can draw a document slope curve(DSC) as shown in Figure 1.

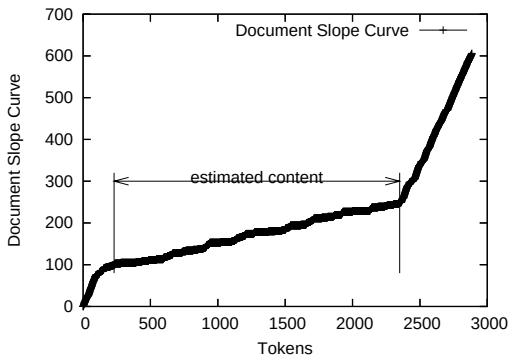

Figure 1: Document Slope Curve for Content Selection. The low slope area is presumed to be 'Content'

$$DSC[k] = \begin{cases} 0 & \text{if } k = 0, \\ DSC[k-1] + 1 & \text{else if } T[k] \text{ is a tag}, \\ DSC[k-1] & \text{otherwise} \end{cases}$$

where $T[k]$ is the k_{th} token in an HTML page. We can estimate the longest low slope area from the DSC to be the content part of the page. We evaluated how many tokens overlap between the content estimated by the content selection algorithm and the content that we manually extracted with a random selection of 30 posting pages from the TREC Blogs06 Collection. The corresponding average recall and precision results were 0.7006 and 0.7302, respectively. This relatively simple content selection algorithm achieves reasonably good performance even though each blog page contains various complicated structures.

3.2.2 Structural Noise Estimation

A plagiarism detection technique can help us estimate the amount of structural noise coming from repeated text. Bernstein and Zobel [2] introduced the SPEX algorithm for efficient near-duplicate document detection. The algorithm calculates how many chunks are shared among documents, where a chunk is a fingerprint of a sequence of the predetermined number of word tokens. This novel algorithm allows us to efficiently use time and space by discarding unnecessary chunks for duplicate document detection.

To verify the relation between duplicate text and noise reduction, we used SPEX on our blog collections. First, we randomly selected 850,000 postings from the TREC Blogs06 Collection. We refer to the set of the selected documents as Blogs06-C. Second, we filtered Blogs06-C by the content selection algorithm. We refer to the filtered collection as Blogs06-C-NR. Last, we randomly selected 1,300,000 postings from the ICWSM Blog collection. We refer to the set of the selected documents as ICWSM-C. Because the ICWSM-C collection has been much more heavily filtered than the

[2]http://www.icwsm.org/data.html
[3]http://www.google.com/adsense

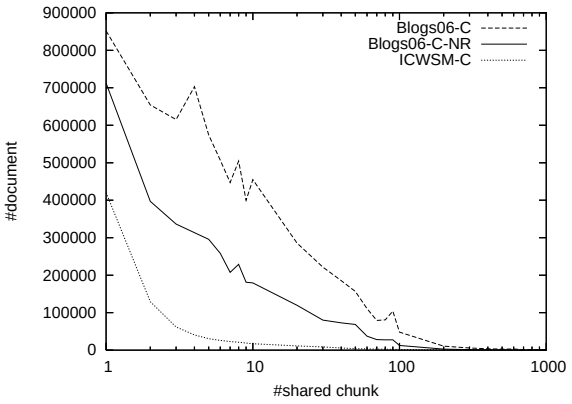

Figure 2: The number of documents sharing text chunks for 3 blog collections; the TREC Blogs06 collection (Blogs06-C), a filtered version of the TREC collection (Blogs06-C-NR), and the ICWSM collection.

TREC collection, we expect the characteristics of this collection to be very different. Nevertheless, we will show the results together for comparison. Figure 2 shows the results of the SPEX experiment.

The graph clearly shows that blog postings from the Blogs06-C-NR shared fewer chunks than did those of Blogs06-C. The difference between the curves of the Blogs06-C and the Blogs06-C-NR is presumed to be the amount of near-duplicate text in the structural noise part. That is, there is a considerable amount of near-duplicate text in the structural noise part of blogs. Note that, although the leftmost peak of the graph is basically caused by real near-duplicate text, we still cannot rule out the possibility of fingerprint collision.

There are considerably fewer text chunks shared in the ICWSM Blog collection which does not have any menubar or advertisement links, i.e. any structural noise. In addition, the large difference between the curves of the two collections can be also explained by the following fact. The Blogs06-C-NR collection still contains duplicate text like boilerplate text, e.g., phrases for copyright or sentences to encourage visitors to leave comments and to subscribe to RSS, whereas the ICWSM collection does not contain such text. In general, such phrases are automatically inserted by blog publication software and service providers. Unfortunately, our content selection algorithm cannot remove them since such text is considered as a part of the content in that there are few tags in the text. However, we don't need to remove all boilerplate text because such text will have little effect on the retrieval, in contrary to structural noise, in that a very limited range of words tends to be used.

Consequently, while structural noise is related to unnecessarily duplicate text, the noise can be reduced by methods such as content selection. The effect of noise reduction by the content selection algorithm on the retrieval performance will be discussed in Section 5.4.3.

3.3 Link Structure

Methods exploiting the link structure of hyperlinks of web pages have been universally used by search engines since PageRank[3] and HITS[10] appeared. The link structure

Table 1: Average document length according to collection in terms of the number of words

	TREC Blogs06	TREC Blogs06 after Noise Reduction	ICWSM
#word	658	342	143

not only effectively reflects dynamics of authority or popularity among web pages but also efficiently makes up for weak aspects of the content-based text search techniques. As a result, link structure analysis has come to be an indispensable feature that should be considered for any web retrieval task.

We analyzed the link structures of the TREC Blogs06 Collection as shown in Figure 3. Further, the graph also contains the link structure analysis for the TREC wt10g Collection, which is a small general web collection, for the sake of comparison. In case of the wt10g Collection, although the inlink counts and the outlink counts are distributed differently, the ranges of the distribution are almost the same. In contrast to this, the distributions of the inlink counts and the outlink counts of the TREC Blogs06 Collection are separated by a big gap. Generally, the inlink counts are distributed over smaller values compared to the outlink counts. This means that most of the outlinks in a given posting page rarely point to other posting pages in the collection. In the real world as well as in our results, there is a tendency that blogs frequently depend on authoritative web pages such as news articles, specifications of products, or the profiles of people rather on the pages of other blogs. This result causes some doubt about the usefulness of the explicit link structure in blog retrieval, especially in a collection limited to blog postings. The impact of using link structure on retrieval performance will be discussed in Section 5.4.4.

4. SEARCHING TECHNIQUES FOR BLOG HOMEPAGES

In this work we use language modeling-based retrieval [18]. This approach has been shown to give good results for a variety of retrieval tasks and is relatively easy to customize with new features or weights.

4.1 Baseline

The simplest and most widely-used technique for blog homepage searches is to retrieve based on the titles of blogs. Generally, a title represents the contents of a blog although we would not expect it to be as descriptive as the full text of postings.

4.2 Passage Retrieval Approach

The relation between a blog homepage and its postings can be viewed as the same as the relation between a document and its passages if a blog homepage is defined as an aggregate of the texts of its postings. Under this definition, we can apply passage retrieval techniques to blog homepage search. In past research on passage retrieval, the decision of how to split a document into a plurality of passages was critical. In blog homepage retrieval however, only a simple merge process is required to construct a blog homepage document from its postings because the postings can be regarded as passages perfectly split from a virtual original

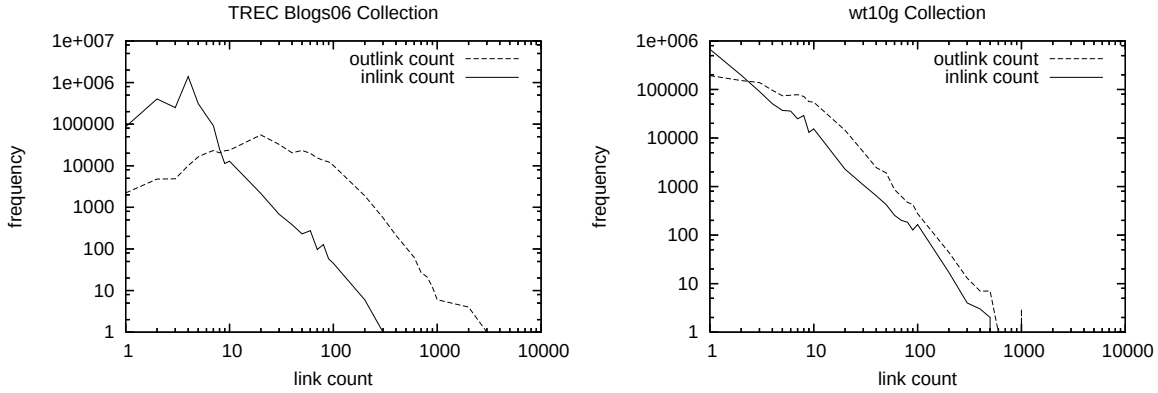

Figure 3: Distribution of the inlink counts and the outlink counts for a blog collection (TREC Blogs06) and a general web collection (wt10g)

document, i.e. a blog homepage document by the author of the document.

Note that the TREC Blogs06 Collection contains the crawled main pages of blog homepages. We assume that the main page of a blog cannot represent the contents of all postings in the blog because it typically is only a snapshot of recent postings. Therefore, we use the collection of main pages only for our baseline approach. Further, to avoid confusion, we refer to a main page of a blog homepage as a blog main page, whereas we refer to a constructed blog homepage document by the aggregation of its postings as a blog homepage document.

For the passage retrieval approach to blog homepage retrieval, we consider the following methods suggested in previous research by Salton et al. [19], Callan [5] and Kaszkiel and Zobel [9].

4.2.1 Global Evidence

One method for blog homepage search is to generate ranking scores based on a language model of the entire blog homepage. We call this the global evidence strategy. We refer to the global evidence score for the i_{th} blog homepage document as $E_G[i]$. We also refer to the technique that constructs the global evidence (the aggregated homepages) and computes the ranking with them as GE.

4.2.2 Local Evidence

The other method for blog homepage search is to generate ranking scores based on language models of the postings. We call this the local evidence strategy. However, because what we want is not the score of the postings but the score of the blog homepages, conversion from the local evidence to a global score is necessary. We introduce two simple conversion methods.

First, the conversion can be accomplished by using the score of the highest-ranked posting of each blog homepage.

$$E_L[i] = \max_j score_L[i][j] \tag{1}$$

where E_L represents the converted local evidence and j is the index of the N top ranked posting of the blog homepage. We refer to the technique that constructs the local evidence, and converts it into a global ranking based on the highest-ranked postings as Local Evidence Highest, or LEH.

The second conversion method is to use a summation of the scores of the N top ranked-postings.

$$E_L[i] = \sum_j score_L[i][j] \tag{2}$$

We refer to this technique as Local Evidence Summation, or LES.

4.2.3 Combination of Global Evidence and Local Evidence

It is known that combination of evidence from different levels of document representation often produces better retrieval results than those from any single level [5, 20]. Typically, a linear combination of the global evidence and the local evidence is used.

$$E_C[i] = \mu E_G[i] + (1-\mu)E_L[i] \tag{3}$$

where μ is a weight parameter. We refer to the combination technique that uses global evidence and the local evidence strategy LEH as CGLEH, and to the combination of the global evidence and the local evidence strategy LES as CGLES.

4.3 Query Expansion

Queries searching for homepages rather than for individual postings are likely to be short. Such a query does not sufficiently describe the subject that the user want to search on. Accordingly, we expect that query expansion could play an important role in achieving more accurate retrieval.

4.3.1 Relevance Models

The relevance model that was proposed by Lavrenko and Croft [12] provides a good framework for query expansion. According to a Bayesian approach, we can estimate a query model, $\hat{\theta}_Q$, over possible query terms, w, given a query, Q.

$$P(w|\hat{\theta}_Q) \propto \sum_{D \in \mathcal{R}} P(w|\theta_D)P(Q|\theta_D)P(\theta_D) \tag{4}$$

where θ_D is a document language model and $\mathcal{R}$ is a set of the N top ranked documents returned with a query likelihood language model. The assumption of a uniform distribution can be used for a prior, $P(\theta_D)$.

Previous research has shown that a linear combination of the relevance model with the initial query model produces

average performance that is higher than using the relevance model alone.[1].

$$P(w|\theta_Q) = \lambda P(w|\tilde{\theta}_Q) + (1-\lambda)P(w|\hat{\theta}_Q) \quad (5)$$

where $\tilde{\theta}_Q$ is the original query model by the maximum query likelihood.

In practice, we expand the query by combining the original query and the expanded query with a weight parameter λ. We refer to the query expansion based on the relevance model as RM.

4.3.2 Mixture of Relevance Models

Diaz and Metzler [6] showed that the relevance model can be improved by using any external collection that contains more relevant documents than does the original collection. We can get an estimated query model based on a mixture of relevance models by modifying Equation 4.

$$P(w|\hat{\theta}_Q) = \sum_{c \in \mathcal{C}} P(c)P(w|\theta_Q, c) \quad (6)$$

where $\mathcal{C}$ is a set of collections. A prior, $P(c)$ can be used as a weight factor of the collections.

Our target is the collection of blog homepage documents. An author of a blog is generally interested in one or more topics. Accordingly, a plurality of topics is usually mixed in a blog homepage document. On the other hand, a posting is relatively topic-oriented. Since query expansion based on the relevance model depends on $P(w|\theta_D)$, we can guess that query expansion from the set of the postings can produce more accurate results than that from the set of blog homepage documents. This relation is depicted in Figure 4. We define a sample space of a relevance model for query expansion as a set of words of the N top ranked documents of the initial retrieval. Then, the sample space of the posting level is smaller and more definite in terms of the relation with the underlying relevance model than that of the blog homepage document level. That is, we can estimate a relevance model with fewer total words and more topical cohesiveness when using the posting level. Accordingly, we predict that query expansion at the posting level will be relatively free from noise compared to the blog homepage document level. Therefore, we expand queries based on the mixture of relevance models by using the set of postings as the external collection. We refer to the query expansion based on the mixture of relevance models as MRM.

4.4 Named Page Finding Approach

Named page finding is analogous to the blog homepage search in that both of them are designed to find representative pages among many individual pages and are based on web pages. Techniques for named page finding were suggested by Metzler et al. [15] in their work for the TREC 2005 Terabyte Track. To apply the techniques to blog homepage search, we consider the following features. First, we use the link structure as a feature. For example, we use anchor text or prior probabilities related to inlink counts or PageRank. Second, we exploit the structure of HTML pages such as titles, headings and bodies as another feature. Last, the proximity of query terms can be used as a feature.

To combine the link features and the structural features,

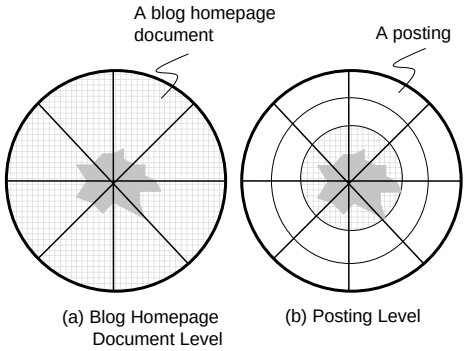

Figure 4: The relation between sample spaces of the underlying relevance model and document models according to each level. The outer circle means a sample space of relevant blog homepages. The gray area represents the sample space of the underlying relevance model. The area filled with grids indicates the sample space for query expansion, i.e. the N top ranked documents set of the initial retrieval.

we use a language model developed by Kraaij et al. [11].

$$P(D|Q) \propto P(D) \prod_{q \in \mathcal{Q}} \sum_{f \in \mathcal{F}} w_f P(q|D, f) \quad (7)$$

where $P(D)$ is a prior, $\mathcal{F}$ is a set of fields, e.g., a title and a body in HTML pages, w_f is a weight for a field, f, and $P(q|D, f)$ represents a language model of a field f of a document, D.

To exploit the proximity of query terms, we make use of the dependence model proposed by Metzler and Croft [14].

$$P(D|Q) \propto \sum_{q \in \mathcal{Q}} \psi_Q f_Q(q, D) + \sum_{q \in \mathcal{U}} \psi_U f_U(q, D) \quad (8)$$

where $\mathcal{Q}$ is a set of individual terms of a query, $\mathcal{U}$ is a set of combinations of the terms, and ψ_Q and ψ_U are weight factors. While f_Q is a function to output the score based on the occurrence of individual terms, f_U is a function to compute the score based on the unordered proximity of terms.

Since the results of this approach used with the collection of blog postings can be considered as local evidence, we can use the conversion method described in Section 4.2 to convert the scores into global evidence. When it is converted using the highest ranked posting, we refer to the approach as NPFH. When it is converted using summation, we refer to the approach as NPFS.

5. EXPERIMENTS AND RESULTS

5.1 Test Collection

We selected 50 queries for our experiments from queries of the Topic Distillation Task of the TREC 2002 Web Track and the TREC 2003 Web Track. We assumed that, when people try a certain blog homepage search, they want to find blogs that contain the postings relevant to some specific topics. The queries of the Topic Distillation Task are with a mixture of abstract queries and explicit queries, and thus we felt that they fit well with the experiments. The relevance judgments for each query were made by ourselves using a pooling method [22]. The criteria used for relevance was as follows.

- Does the blog consistently create postings relevant to the topic?
- Do more than half of the postings in the blog deal with the topic?

When both of the criteria were satisfied, we marked the blog relevant.

Of the 50 queries, 25 were randomly selected and used for parameter training. The remaining 25 queries were used for evaluation.

5.2 Retrieval System and Evaluation

For our experiments, we used Indri [23] as the retrieval system. Indri is a search engine based on both the language modeling and the inference network frameworks. It supports structured queries and pseudo relevance feedback based on relevance models. Furthermore, prior probabilities can be integrated into the language model probability at query time. These powerful features allowed us to efficiently experiment with various approaches.

To evaluate the performance of the retrieval runs, we used MAP, GMAP, MRR and bpref as measures [4, 24]. GMAP is a geometric mean of per-query average precision, while the MAP is an arithmetic mean. GMAP provides a more robust measure to reflect the improvement of low performance queries than MAP[24].

$$\text{MAP} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} AP(q) \tag{9}$$

$$\text{GMAP} = \sqrt[|\mathcal{Q}|]{\prod_{q \in \mathcal{Q}} AP(q)} \tag{10}$$

where $AP(q)$ is an average precision for a query, q in a set of topics, $\mathcal{Q}$. MRR is a mean reciprocal rank of the first relevant retrieved document. MRR is useful for evaluating high accuracy retrieval situations [21]. The bpref measure proposed by Buckley and Voorhees [4] provides an accurate measure when there are many unjudged documents.

$$\text{bpref} = \frac{1}{R} \sum_{r} (1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}) \tag{11}$$

where R is the number of judged relevant documents, N is the number of judged irrelevant documents, r is a relevant retrieved documents, and n is a member of the first R irrelevant retrieved documents.

5.3 Experimental Design

5.3.1 Experiment 1

We tested the baseline approach described in Section 4.1 with titles were extracted from the blog main pages of the TREC Blogs06 Collection. Further, we also did the experiment using the whole content of the blog main pages.

5.3.2 Experiment 2

We did an experiment on the combination of GE, LE, CGLE, RM and MRM described in Section 4.2 and Section 4.3 with the blog postings of the TREC Blogs06 Collection.

For the LEH or LES + RM experiments, after the retrieval run is carried out with queries expanded from the set of the postings, the scores of the result are converted into global evidence by the defined conversion method. GE + RM produces the result using the pseudo relevance feedback against the set of the blog homepage documents. For GE + MRM, the queries combined with the queries expanded from the set of the postings and the original queries are used against the set of blog homepage documents.

5.3.3 Experiment 3

We experimented on the combination of GE, LE, CGLE, RM and MRM described in Sections 4.2 and 4.3 with the TREC Blogs06 Collection filtered by the content selection algorithm introduced in Section 3.2.1.

5.3.4 Experiment 4

We did an experiment using the named page finding approach described in Section 4.4 against the TREC Blogs06 Collection. To study the effects of the three features of this approach, i.e. the link structure, the proximity of query terms, and the HTML structure, we compared the effectiveness of each feature separately.

To get the prior probabilities from PageRank and the inlink count, we estimated $P(\text{Relevance}|\text{inlink count})$ and $P(\text{Relevance}|\text{PageRank})$ from the relevance judgments of the wt10g collection. Although the statistics and characteristics of two collections are totally different, we made the (strong) assumption that their posterior would resemble each other on the grounds that Figure 2 shows that the range of the distributed inlink counts are somewhat similar. Then, we made sets of the priors by mapping the estimated probabilities to the real feature values.

5.4 Results and Discussion

5.4.1 Result of Experiment 1

Table 2 presents the result of the baseline approach. It performed poorly as we expected. But, when using the whole content of the main pages instead of using the titles only, a dramatic improvement occurred. This means that a main page may be a reasonable representation of a whole blog, contrary to our assumption. However, it is not still as good as the results using the blog homepage documents. The reason is that there are different types of main pages in blogs. Some main pages are designed to summarize the contents of the whole blog, while others just give a snapshot of recent postings. In the latter case, using only the main page may fail for some queries.

5.4.2 Result of Experiment 2

The result of the combination of the passage retrieval approach and query expansion is demonstrated in Table 3. It is obvious that the summation technique is much better than using the highest-ranked posting when converting local evidence to the global evidence. In addition, the results show that the combination of local evidence and global evidence outperformed the other methods. On the other hand, when using only local evidence or global evidence, it is hard to tell which one performed better. Furthermore, the result of query expansion is somewhat confusing. When query expansion was applied to retrieval, it hurt performance in most of the evaluation measures.

5.4.3 Result of Experiment 3

Table 4 presents the result of the combination of the passage retrieval approach and query expansion after noise re-

Table 2: Result of Experiment 1

Method	MAP	GMAP	MRR	bpref
Baseline (title)	0.1329	0.0074	0.4503	0.2088
Baseline (whole)	**0.3035**	**0.2171**	**0.5660**	**0.3822**

Table 3: Result of Experiment 2

Method	MAP	GMAP	MRR	bpref
LEH	0.2905	0.2310	0.4542	0.1996
LEH + RM	0.2951	0.2160	0.4560	0.2216
LES	0.3922	**0.2889**	0.5532	0.3992
LES + RM	0.3826	0.2760	0.5492	0.4098
GE	0.3489	0.1913	0.5894	0.4126
GE + RM	0.3464	0.1785	0.5795	0.4061
GE + MRM	0.3550	0.1776	0.5307	0.4030
CGLEH	0.3866	0.2236	**0.6060**	0.4077
CGLES	**0.4025**	0.2309	0.5979	**0.4380**

Table 4: Result of Experiment 3. Noise reduction is applied to the collection.

Method	MAP	GMAP	MRR	bpref
LEH	0.2393	0.1865	0.5273	0.1694
LEH + RM	0.2471	0.1848	0.4879	0.1852
LES	0.3542	0.2625	0.5902	0.4029
LES + RM	0.3315	0.2408	0.5380	0.4024
GE	0.3326	0.2267	0.5634	0.4037
GE + RM	0.3268	0.2160	0.5222	0.4123
GE + MRM	0.3588	0.2489	0.5976	0.4434
CGLEH	0.3562	0.2608	0.5861	0.3934
CGLES	**0.3994**	**0.3040**	**0.6936**	**0.4593**

duction. Although the performance was somewhat worse in the case of using the simple methods (compared to the unfiltered collection), it is more efficient because the size of the collection was reduced to almost half as shown in Table 1. However, in the case of the enhanced methods, i.e. the combination methods, it shows significant improvements. Especially, contrary to the Experiment 2, it clarifies that query expansion can contribute to the improvement of retrieval performance. Moreover, as we expected, query expansion by the mixture of relevance models performed better than that by the relevance model. The reason is presumed to be that noise reduction decreased the probability that irrelevant words are contained in the expanded queries. Consequently, this result shows that the content selection algorithm substantially reduces the noise that have an effect on retrieval performance. Based on the filtered collection, the advanced retrieval methods demonstrated the best results in our experiments.

5.4.4 Result of Experiment 4

Table 5 presents the result of the named page finding approach. In spite of being one of the state of the art methods for general web page collections, the approach performed worse than the others. The result might be explained by the difference between general web pages and blog pages. We examined which of the features of web pages is more useful for blog homepage search through experiments with various combinations of the features.

The most useful feature is the structure of the HTML page. The experiment using this feature exhibited a considerable improvement, whereas there was noticeable degradation when it was not used. That is, blog pages inherit the advantage of being able to exploit page structure from general web pages. In the case of proximity features, they do not seem to improve performance. This may be a result of the characteristics of our queries since most of them are not names, but instead are abstract concepts where the explicit relationship of query terms can be somewhat ignored.

Table 5: Result of Experiment 4. 'LK' indicated that the features related to the link structure is used, 'PR' represent that the proximity of query terms is used, and 'ST' means that the structure of the HTML page is used.

	Method	MAP	GMAP	MRR	bpref
	LK	0.1116	0.0412	0.2947	0.2188
N	PR	0.1151	0.0539	0.3188	0.1951
P	ST	0.1435	0.0402	0.3045	0.1723
F	LK+PR	0.0955	0.0390	0.3206	0.2098
H	PR+ST	0.1361	0.0368	0.3158	0.1651
	ST+LK	0.1337	0.0460	0.3191	0.2188
	LK+PR+ST	0.1088	0.0420	0.2923	0.2172
	LK	0.3257	0.1806	**0.6194**	0.4064
N	PR	0.3421	0.1749	0.5816	0.3783
P	ST	**0.3866**	**0.2183**	0.6137	**0.4078**
F	LK+PR	0.3010	0.1525	0.5966	0.3714
S	PR+ST	0.3561	0.1878	0.5849	0.3554
	ST+LK	0.3246	0.1757	0.5859	0.4017
	LK+PR+ST	0.2962	0.1530	0.6018	0.3539

In contrast, the link structure feature was definitely unhelpful. The result when the link feature was not used outperformed the result using links. Indeed, this result is predicted according to the analysis of Section 3.3. Because the number of inlinks and outlinks in our blog collection are out of balance, it is difficult to say whether the link structure analysis algorithm based on inlinks rather than outlinks works well. That is, we might conclude that using link structure is not helpful for blog homepage search. Nevertheless, this result just reflects a single blog collection and a fairly small collection of queries. The situation may be different if all related web pages were part of the collection. Also, the structure of pages in the blogosphere is constantly evolving and, as more good quality blog postings are created, more links to these postings may occur.

6. CONCLUSION AND FUTURE WORK

We showed that homepage search in blog collections is somewhat different from general web page search. We introduced various techniques for blog homepage search, and demonstrated that the passage retrieval approach fits well with this task. Furthermore, our work shows that blog homepage retrieval requires high quality noise reduction to achieve accurate results. A content selection algorithm based on tag distribution is a candidate method for noise reduction.

There is a crucial issue for future work. Although global evidence from whole blog homepage documents works well in our experiments, there remains a scale problem to be solved in order to apply it in a commercial search engine. The blog homepage documents generated from the small collection we used were of reasonable size. In contrast, blog homepage documents constructed in this way by commercial engines

may be too large and the number of the postings grows continuously. We will continue to work with this problem to identify techniques for concise representation of the collection of postings. Further, we will study the implicit link structure of blogs. Since the explicit link structure of blogs for this collection was shown to be not helpful, we are planning to replace this feature with an analysis of the implicit link structure, e.g., investigating co-derived documents of blogs.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NHN Corp. and in part by NSF grant #IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, and X. Li. Umass at trec 2004: Novelty and hard. In *Online Proceedings of 2004 Text REtrieval Conference (TREC 2004)*, 2005.

[2] Y. Bernstein and J. Zobel. Accurate discovery of co-derivative documents via duplicate text detection. *Information Systems*, 31:595–609, 2006.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, 1998.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[5] J. Callan. Passage-level evidence in document retrieval. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, 1994.

[6] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, 2006.

[7] A. Finn, N. Kushmerick, and B. Smyth. Fact or fiction: Content classification for digital libraries. In *Joint DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[8] N. Glance. Indexing weblogs one post at a time. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 43–46, 2006.

[9] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 1997.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, 1998.

[11] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, 2002.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[13] C. Maconald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

[14] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, 2005.

[15] D. Metzler, T. Strohman, Y. Zhou, and W. B. Croft. Indri at trec 2005: Terabyte track. In *Online Proceedings of 2005 Text REtrieval Conference (TREC 2005)*, 2006.

[16] G. Mishne and M. de Rijke. A study of blog search. In *ECIR*, pages 289–301, 2006.

[17] D. Pinto, M. Branstein, R. Coleman, M. King, W. Li, X. Wei, and W. B. Croft. Quasm: A system for question answering using semi-structured data. In *the JCDL 2002 Joint Conference on Digital Libraries*, pages 46–55, 2002.

[18] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, 1998.

[19] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *SIGIR '93: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, 1993.

[20] G. Salton and C. Buckley. Automatic text structuring and retrieval experiments in automatic encyclopedia searching. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–30, 1991.

[21] C. Shah and W. B. Croft. Evaluating high accuracy retrieval techniques. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–9, 2004.

[22] K. Sparck Jones and C. J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[23] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[24] E. M. Voorhees. The trec robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.