

“Blog08 Track” - A report

Austin Wood
Student Number = s3083208
Royal Melbourne Institute of Technology
s3083208@student.rmit.edu.au

Mujtaba Hussain
Student Number = s3093175
Royal Melbourne Institute of Technology
s3093175@student.rmit.edu.au

ABSTRACT

In the age of information that we live in today, more and more information now comes from public domain. The pioneer of the information age is now contained in the form of web logs, or blogs. In 2006, the university of Glasgow in Scotland initiated a TREC track to investigate the opportunities that the said source can provide in terms of information retrieval. The source was 160GB's of uncompressed logs from the web. This research aims to find the result of ad-hoc search and variations of ad-hoc search on this large collection. We investigate effect of searching the collection based on variations of query types and evaluate the experiments by using various metrics like Mean Average Precision, R Precision , Precision etc.

Categories and Subject Descriptors

Blog08 Tracks - Search Technology [

General Terms

]: Search Engines

Keywords

Blogs, Search Engines, Opinions, ad-hoc

1. INTRODUCTION

Before the advent of free information on the internet, most of the information needs had to be satisfied from specific sources like a library and these information resources were written by specific people from that area. But as the information moved away from a specific core and spread across the internet. More and more people started being the source of knowledge and the notion of specific locations of information was removed. The advent of social networking promoted the idea of casual information sharing. Initially, the idea of a web log was to share casual details promoting a healthy casual experience but as more and more people

all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2008 RMIT **Blog08 Track Report** ...

started sharing their knowledge, more and more specific information started spreading across the internet. Blogs are of particular interest to the researchers as their is no format binding a blogger(Blog User) from posting her/his information any way they want. There is also no limit or prohibition on how this information is shared or how much can be shared and this allows the blogger the freedom to post information in a form they think is most relevant.

This provides a problem for researchers as there is no known pattern to search information in across these blogs. They must be treated as large collection of texts with the possibility of relevant information sporadically spread. This leads to ad-hoc search being the largest area of research currently prevalent among blog-researchers with interest recently spreading into opinion finding.

2. TERMS AND DEFINITIONS

2.1 Mean Reciprocal Rank

Whenever a search engine produces a set of resources in response to a query, the position of the resource in that query is defined as its rank. The inverse of a resources' rank or its reciprocal rank is a metric used in Information Retrieval to determine whether an algorithm or an experiment was successful in elevating a desired resource. A mean of reciprocal ranks taken across a set of queries is called *Mean Reciprocal Rank* and is also a major metric used for evaluation of experiments and algorithms.[4]

2.2 Average Precision

Whenever we retrieve a set of resources from a search engine, it is not necessary that all the retrieved resources will be relevant to the query. The ratio of relevant resources retrieved to the total retrieved resources is called Precision. The average of such a ratio over a set of queries is called *Average Precision* and is also a metric used for evaluation in Information Retrieval.[4]

$$AvgP = \frac{\sum_{r=1}^N (P(r) \times rel(r))}{\#relevant\ documents}$$

2.3 R-Precision

R-Precision is another form of precision with the main aim of generating a single value metric of the ranking by computing the precision at the R-th position in the ranking, where R is the total number of relevant documents for the current query. Even though it is useful to see the effect

after each query, it can be calculated over an entire set of queries.[1]

3. BACKGROUND

The TREC Blog track was first introduced in 2006 by the University of Glasgow. It is made up of the following tasks;

1. Baseline adhoc (blog post) retrieval task
2. Opinion finding (blog post) retrieval task
3. Polarised opinion finding (blog post) retrieval task
4. Blog finding distillation task

For each of the three years it has been running, the track has been using the Blog06 test collection that was compiled by the University of Glasgow in 2006 [3]. It consists of 38.6GB of feeds, 88.8GB of permalink documents, and 28.8GB of homepages; and comprises of over 3.2 million unique documents.

The first year it run, in 2006, there were only two tasks. The first being the opinion finding task, and the second being an open task to decide what next years tasks would be. The distillation task and the polarised opinion finding task came out of this open task in 2006.

The opinion finding tasks involves finding blog posts that express an opinion on the given topic. The polarised opinion finding task is similar, except results returned should either be all positive opinions or all negative opinions.

The blog finding distillation task is similar to the adhoc retrieval task, but instead of returns posts about that given topic, the user should be given the key blogs about a given topic.

4. EXPERIMENTS

4.0.1 Data

The experiments in this research were developed to reflect the nature of ad-hoc search. We received three sets of data.

4.0.2 Zipped Content

The content supplied to the researchers in this case was zipped archives which contained a large number of HTML files downloaded from the internet, named permalinks. A sample permalink file is shown below:

```
<DOCNO>BLOG06-20051211-116-0000022346</DOCNO>
<FEEDNO>BLOG06-feed-057904</FEEDNO>
<FEEDURL>http://www.simplephotography.de/journal/
/rss</FEEDURL>
<BLOGHPNO>BLOG06-bloghp-057904</BLOGHPNO>
<BLOGHPURL>http://www.simplephotography.de/journal/
</BLOGHPURL> <DOCHDR>
0.0.0.0 20051230304756 20311 Date: Fri, 30 Dec 2005
04:47:54 GMT Pragma:
no-cache Server: Apache Vary: Host Content-Type:
text/html Expires: Mon, 26
Jul 1997 05:00:00 GMT Last-Modified: Fri, 30 Dec
2005 04:47:55 GMT
Client-Date: Fri, 30 Dec 2005 04:47:55 GMT Client-
Peer: 130.209.241.223:8080
```

Each individual document inside these permalinks is surrounded by TREC *DOC* and *DOCNO* tags to enable TREC evaluation.

4.0.3 Relevance Judgements

We also received a relevance judgements file. The relevance judgements in this case had been done by the TREC examiners from 2007 submitted runs. A sample is shown below:

```
852 0 BLOG06-20060119-021-0021655938 0
852 0 BLOG06-20060119-026-0011977846 0
852 0 BLOG06-20060119-026-0012093605 0
852 0 BLOG06-20060119-034-0022116811 0
852 0 BLOG06-20060119-039-0007368143 0
852 0 BLOG06-20060119-039-0022247719 0
852 0 BLOG06-20060119-042-0010559998 0
852 0 BLOG06-20060119-044-0006953092 0
852 0 BLOG06-20060119-051-0006573896 0
852 0 BLOG06-20060119-057-0029245453 0
852 0 BLOG06-20060119-059-0012442192 0
852 0 BLOG06-20060119-060-0029271164 0
852 0 BLOG06-20060119-061-0025903046 0
852 0 BLOG06-20060119-061-0025922065 0
852 0 BLOG06-20060119-062-0001919847 1
852 0 BLOG06-20060119-062-0004812129 1
852 0 BLOG06-20060119-062-0004848775 0
852 0 BLOG06-20060119-064-0003170737 0
852 0 BLOG06-20060119-064-0016818084 0
852 0 BLOG06-20060119-065-0010720152 1
852 0 BLOG06-20060119-066-0008454256 0
```

These relevance judgements provided up with a text book example of how to structure our judgements so that they could be compared with ones done by TREC evaluation experts.

4.0.4 Topic File

The third and last piece of data that was provided to us was the data file containing all the topics that would be used in the ad-hoc search. These contained three main fields that were of interest with respect to this research experiment; namely *title*, *Description* and *Narrative*. Description and Narrative are expanded definitions of the title field. An example is shown below:

```
<top>
<num> Number: 851

<title> "March of the Penguins"

<desc> Description:
Provide opinion of the film documentary "March
of the Penguins".@

<narr> Narrative:
Relevant documents should include opinions con-
cerning the film
documentary "March of the Penguins". Articles
or comments about
penguins outside the context of this film docu-
mentary are not
```

relevant.

</top>

4.1 Software

The main task in this experiment was creating the index of uncompressed zipped archives, which was in total 160GB. The search engine that the researchers chose for that purpose is the called Zettair. Zettair is an in-house search engine for RMIT, built by the Search Engine Group (SEG). Zettair provides us with the *zet_trec* binary, which given the proper arguments, evaluates a data set based on the topic file and generates IR metrics for evaluation.

4.2 Equipment

The experiments were run on a GNU/Linux based system called *rocky2* which has 1.5GB of memory and a standard Intel(R) Pentium(R) 2.80GHz processor. The researchers had access to 1.2 TB of main disk.

4.3 Experiments

The main aim of this research is to analyse the effects of ad-hoc search of search metrics like Mean Average Precision (MAP) etc. We decided to evaluate ad-hoc search based on the three parameters provided in the Topic File.

4.3.1 Indexing

Indexing the permalinks file took some time as random archives were corrupt and had to be at times manually weeded out of the collection. Using zettair, the entire collection was indexed without omitting any of the HTML tags and the eventual size of the index was approximately 2GB.

4.3.2 Experimental Runs

The researchers ran three sets of experiments, identical but using different input parameters. In the first set, the index was queried against the title field of the given Topic, the second set queried the index using the Narrative and lastly using the description.

These sets of experiments would give us a fair idea of how the amount of detail in the query string affects the final IR metrics. The results of the three evaluations are shown below.

4.3.3 Evaluation based on title

Average precision for all rel docs 0.1253
R-Precision docs retrieved): 0.2023

As we can see from Figure 1, as the number of documents increases, the precision decreases rapidly.

4.3.4 Evaluation based on Description

Average precision for all rel docs: 0.0827
R-Precision for docs retrieved: 0.1514

Once again, as shown in Figure 2, when the index is queried with the description field, increase in the number of documents shows a decrease in precision.

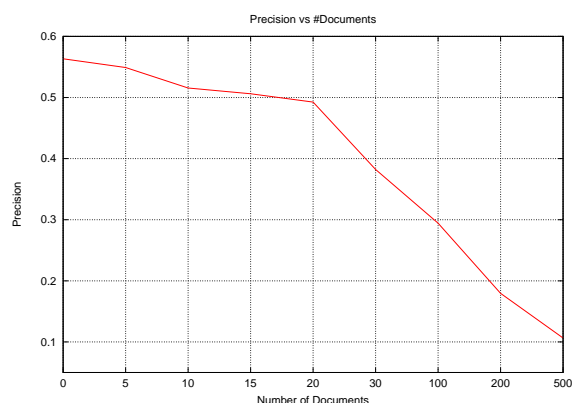


Figure 1: Precision vs #Documents

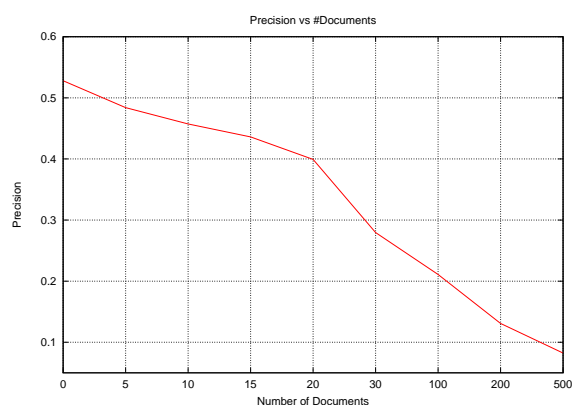


Figure 2: Precision vs #Documents

4.3.5 Evaluation based on Narrative

Average precision for all rel docs: 0.0376
R-Precision for docs retrieved: 0.0863

Even when the query string is at its largest in the narrative form, there is a decrease in the precision, as can be seen in Figure 3.

4.3.6 Evaluation based on pruning

An honours paper written by one of the authors (Mujtaba Hussain) [2] of this research found that certain amount of pruning of the query improved the Mean Reciprocal Rank (MRR) metric of the evaluation. This research decided to make use of those findings and implement them as a separate experiment.

The paper in question ranks query terms important based on their Inverse Document Frequency (IDF). Experiments conducted by the author of the paper have shown that a query with less IDF words performs better with regards to MRR than a query with more IDF words. This principle, when applied to the narrative field of the topic file provided, yielded the following result.

Average Precision for all rel docs: 0.0256
R-Precision for docs retrieved: 0.0562

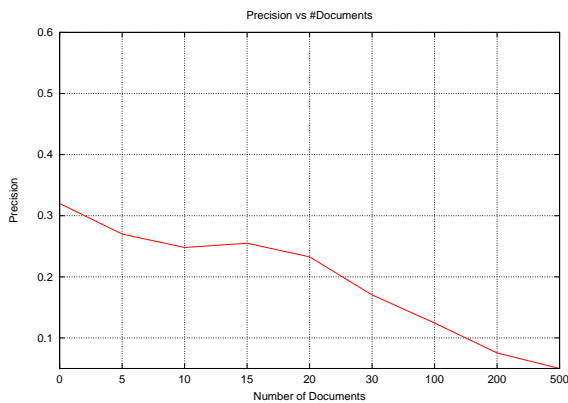


Figure 3: Precision vs #Documents

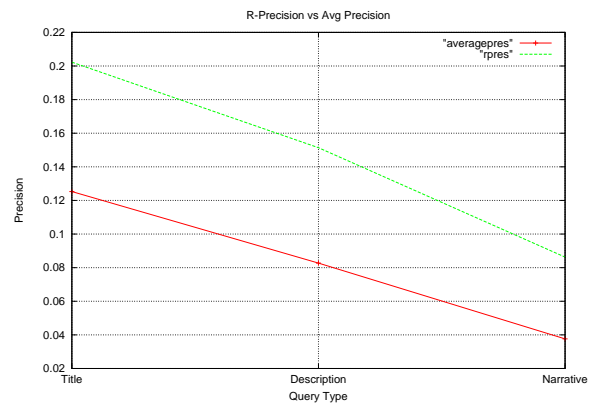


Figure 5: R-Precision vs Avg Precision

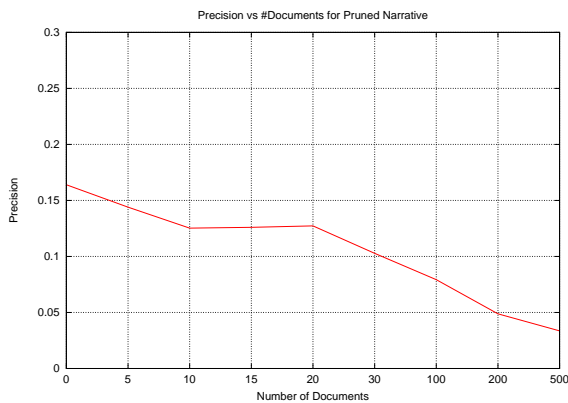


Figure 4: Precision vs #Documents for Pruned Narrative

As shown in Figure 4, the precision per document number decreases, following the trend set by non-pruned input.

4.3.7 Evaluation based on pruning

4.4 Results

When we plot the average precision of the three runs and generate a graph, we observe that there is a general decline in both R and Average precision as they query type goes from less descriptive to most descriptive, as shown by Figure 5.

This shows that a short number of precise, descriptive words returns a more accurate result. This shows that with more words added to the query the greater the likelihood of a non-relevant document being chosen, which is the case for large collections.

Even when the data was pruned based on a research result, the experiment proved to have no effect. This is a result of the query and the corresponding collection being very vague in content and hence decisive pruning had no real effect.

5. REFERENCES

- [1] IR Evaluation.
http://www.cs.armstrong.edu/leizhu/iir/modules/c5/IR_Evaluation.pdf.

- [2] M. Hussain. Anticipating search needs in real time help desk environments. Royal Melbourne Institute of Technology, Computer Science and Information Technology, 2008.
- [3] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. *DCS Technical Report Series*, 2006.
- [4] I. H. Witten, I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, May 1999.