CMP-5036A - Information Retrieval Report

Coursework 2

100108964

Contents

Description

This report outlines the assignment completed and the techniques used to evaluate the information retrieval system.

1 Introduction

The assignment was to produce a search engine to work with uea.ac.uk/computing domain. Using the crawler provided and creating an indexer to strip out the formatting of the pages crawled. Left with lowercase characters and no punctuation. This was then output into files for processing with the system. Next the basic system was created to run queries based on TF*IDF. Running the queries allowed the system to collect a list of ranked results in the form on URLS.
After the first system - another improved system was implemented to include stemming and stop words. This report will discuss and present the techniques and examinations of the results acquired by the experiments.

2 Techniques

Queries entered by the user are statements of information that the user would want results for. Queries can identify several objects that could match the entered query with different degrees of relevancy.
Some information retrieval systems compute a numeric score to show the top ranking objects. These are presented to the user in a ranked list depending on how each object matched the users query. This is done in this system by calculating the normalised TF*IDF score and cosine similarity.
Once the first system was completed, another system was implemented to improve results by using stemming. This process reduced the words to root such as words that can end in a suffix i.e."Running". Stemming is used to improve the effectiveness of the system and reduce size of the index files. Another improvement is to use stop words. These are predefined words that are filtered during the process. These can be common words used such as "the, a, is".

2.1 TF-IDF Formula

During research of the TF-IDF formula there were different methods for calculating the normalised TF. One formula

$$TF = 1 + log(termFrequency) \tag{1}$$

Another used:

$$TF = \frac{termFrequency}{corpusTotal} \tag{2}$$

Term frequency is the specific term in a document and corpus is the total number of documents within the corpus. The formula was simplified using to 1 + log(tf) and freq/P. After implementing each of the TF*IDF - then calculated the cosine similarity to retrieve a list of ranked results. After implementing both formulas - the consine similarity was calculated to retrieve a ranked list of results. Calculated the similarity metric between each entry in a query and associated document vector. The metric co-related the weighted words A with a document vector B representing the normalised term frequencies.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A}.\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}}$$

A high cosine value indicates that an entry is closely related to the query and is a good candidate for being relevant.

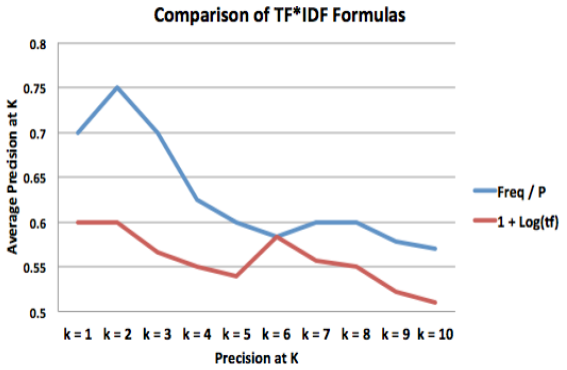

Figure 1: Figure 1. Comparison of TF*IDF formulas

3 Experiments

When indexing the punctuation and HTML code were stripped, then converted all the text to lowercase. This is to help with indexing the site with getting the words and the term frequencies. Once indexing has finished could then query each system with a list of suitable queries provided. From the results could then calculate the precision at k

$$P(k) = \frac{relavantDoc}{k} \tag{3}$$

The average precision at k.

$$avg.P(k) = \frac{\sum_{i=1}^{q} P(k)}{queries} \tag{4}$$

Experiment showed that both formulas have the same average precision at 6. This was the closest 1 + Log(tf) formula got to surpassing freq/P, however in comparison of the two TF formulas, the freq/P overall returned most relevant results.
The relevancy was decided in a consistent manner across the experiments conducted as relevancy is subjective to the user. After selecting the best TF formula, next was to improve the the results and indexing time, as it took a long time to complete and would become overwhelmed with requests.
Stemming was used next to improve results. This is done by stripping unnecessary words of their suffix. This showed significant improvement.

Another improvement implemented was the edition of stop words. These are common words used usually short and predefined. The list contained 30 common words used in search

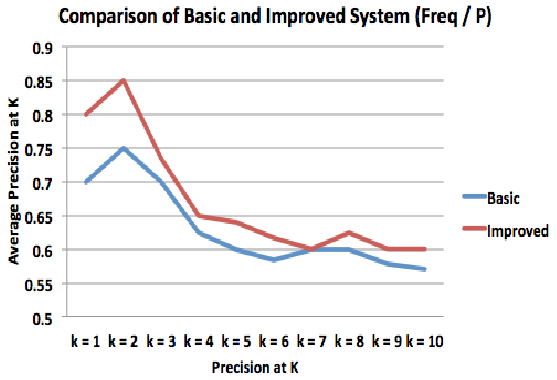

Figure 2: Comparison of basic system against improved stemming and stop words. Using freq/P formula

querying.
When comparing the cosine similarity of the results for the system with stop words against the system with stemming there is not any noticeable difference in URL ranking. Figure 2 shows the improvements made has increased the relevance of returned URLs.

4 Results

During the experiment the TF formula chosen and the changes has positive effect on the system which the results show. The overall ranked URLs retrieved from the queries has an average precision at 10 of 0.6 (figure 3)

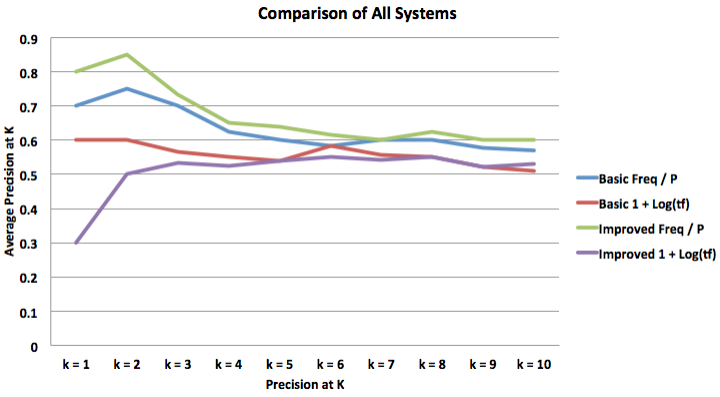

Figure 3: Comparison of all systems

5 Conclusion

This report examines the effectiveness of the information retrieval system created in order to characterise the distribution of retrieval effectiveness.
If more improvements were to be done - possible implementation of title weighting as this would be a method to increase the amount of relevant results returned to the user. Title weighing favours results that have the query terms in the title.