

# Large Scale Analysis of Search Engine Content

John D. King, j5.king@qut.edu.au

**Abstract**—We mine a large taxonomic dataset for subject classification rules. We then use these rules to perform an extensive analysis of the subject matter of the largest general purpose internet search engines in use today.

## I. INTRODUCTION

In recent history it has become impossible for one person to have full knowledge about every domain of human endeavour. The previous way of accessing and discovering information was to manually search of a set of books or journals. However the introduction of search engines has forever changed the way people access and discover information. By using a search engine, people can find information about almost any subject in seconds, and as more material becomes electronically available the influence of search engines will continue to grow.

However, little is known about the content of the largest general purpose search engines. We introduce a new world knowledge assisted method which we use for the classification of large search engines, even those which contain many billions of documents<sup>1</sup>. Table I shows the search engines used in this work. The search engines were compared across hundreds of subjects, and the similarities and differences between the engines were analysed. As far as the author is aware this is the first time a study of this size and scope has been carried out.

## II. METHOD

We generate subject classification rules from a large human expert classified training set. The training set is a large collection of expert human classified documents across many different subjects which are parsed and added to a database. For each subject a set of classification terms are selected using statistical analysis. These terms should preferably be subject-specific (occurring within few or no other subjects) and

<sup>1</sup>At present Google lists its index size as just over 8 billion pages.

Title	Abbreviation	URL
Altavista	AV	<a href="http://www.altavista.com/">http://www.altavista.com/</a>
America Online Search	AOL	<a href="http://search.aol.com/">http://search.aol.com/</a>
Ask Jeeves	ASK	<a href="http://webk.ask.com/">http://webk.ask.com/</a>
Google	Google	<a href="http://www.google.com/">http://www.google.com/</a>
MSN Search	MSN	<a href="http://search.msn.com/">http://search.msn.com/</a>
Teoma	Teoma	<a href="http://www.teoma.com/">http://www.teoma.com/</a>
WiseNut	Wisenut	<a href="http://www.wisenut.com/">http://www.wisenut.com/</a>
Yahoo Search	Yahoo	<a href="http://www.yahoo.com/">http://www.yahoo.com/</a>

TABLE I  
THE SEARCH ENGINES USED IN THIS PAPER

should occur frequently within the subject and infrequently in other subjects. It is difficult to decide which terms to select as there are many possible terms to describe a subject. Many terms may not occur in common English dictionaries yet are still valuable for classification. Many of these terms are technical and subject specific such as conference names, acronyms and names of specialist technology. Some examples from computing are *RMI*<sup>2</sup>, *SMIL*<sup>3</sup>, and *XSLT*<sup>4</sup>. Few standard English dictionaries include these terms, yet if any of these acronyms occur in a document it is likely the document covers a subject related to computing. For each subject we use statistical analysis to extract a set of terms that are most representative of the subject. The statistical analysis includes finding patterns between subject nodes and terms which are used to extract classification terms.

Once the subject classification terms are extracted we use them to query the search engines. From the number of results returned from each query we are able to find the representation of each subject in each search engine.

## III. RESULTS

The subject distributions of each search engine were analysed and it was found that some search engines had a bias towards the sciences and others toward the arts. The analysis also showed that Teoma and ASK use the same index for their results. Each search engine was compared to Google's index and it was found the search engine that was most similar to Google was AOL (which uses a different version of Google's index) and the search engine that was the most different to Google was WiseNut.

Figure 1 shows the content distributions of the search engines for each of the top-level subject groupings. We can only show the results for the ten highest level subjects because there is not enough space to show the results of the hundreds of lower level subjects.

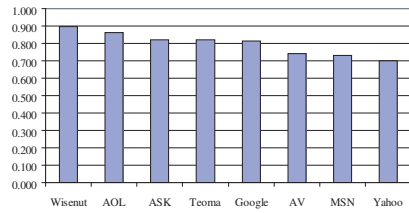
## IV. CONCLUSION

We have developed a new search engine classification method which is highly scalable and easily distributed. The method shows the subject matter of each search engine. We use this method to analyse the subject matter of the largest internet search engines in use today.

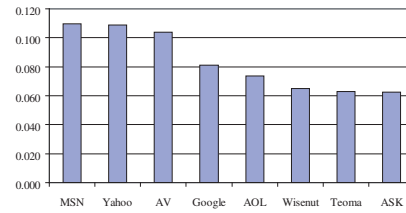
<sup>2</sup>Remote Method Invocation.

<sup>3</sup>Synchronized Multimedia Integration Language

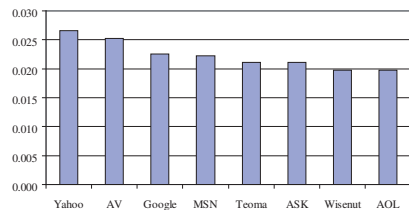
<sup>4</sup>Extensible Stylesheet Language Transformation.



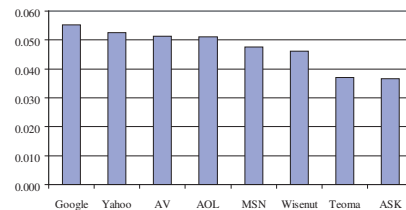
(a) Generalities



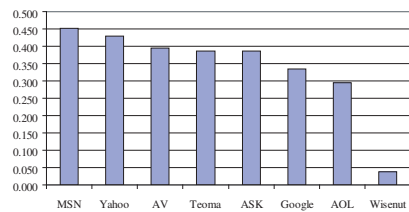
(b) Philosophy &amp; Psychology



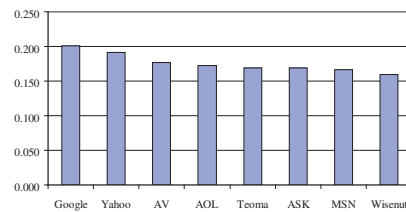
(c) Religion



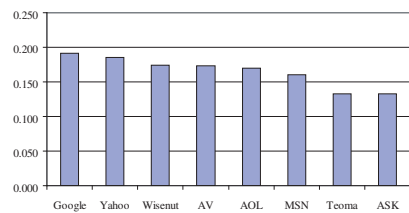
(d) Social Sciences



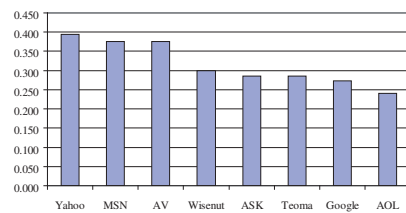
(e) Language



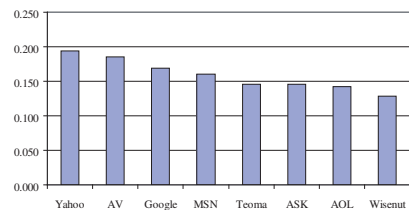
(f) Natural sciences &amp; mathematics



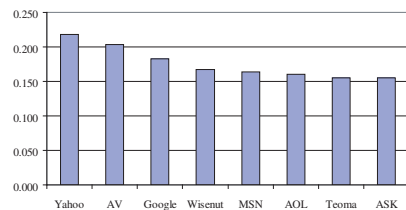
(g) Technology (Applied sciences)



(h) The Arts



(i) Literature &amp; rhetoric



(j) Geography &amp; history

Fig. 1. Normalised Distribution of Search Engine Content