Performance Evaluation of Desktop Search Engines

Chang-Tien Lu, Manu Shukla, Siri H. Subramanya, Yamin Wu
Department of Computer Science, Virginia Polytechnic Institute and State University, USA
ctlu@vt.edu, mashukla@vt.edu, siris@vt.edu, ywu4@vt.edu

ABSTRACT

With the rapid increase in computer hard drive capacity, the amount of information stored on personal computers as digital photos, text files, and multimedia has increased significantly. It has become time consuming to search for a particular file in the sea of files on hard drives. This has led to the development of several desktop search engines that help locate files on a desktop effectively. In this paper, the performance of five desktop search engines, Yahoo, Copernic, Archivarius, Google, and Windows are evaluated. An established dataset, TREC 2004 Robust track, and a set of files representing a typical desktop have been used to perform comprehension experiments. A standard set of evaluation measures including recall-precision averages, document level precision and recall, and exact precision and recall over retrieved set are used. The evaluations performed by a standard evaluation program provide an exhaustive performance comparison of the desktop search engines by representative information retrieval measures.

1. INTRODUCTION

Desktop search engines, also called localized search engines, index and search files in a personal computer (PC). They retrieve references to files on the computer's hard drives based on keywords, file types, or designated folders. Simple text match search capabilities are not sufficient for the amount of information in PCs today. To conduct file searches on the PC's hard drive, performance of the desktop search engine in terms of Information Retrieval (IR) measures, e.g. precision and recall, play an important role in measuring the accuracy of the search results.

Various companies have released their versions of desktop search engines like Microsoft Windows desktop search[1], Yahoo desktop search[2], Copernic desktop search[3], Google desktop search[4], Archivarius 3000[5], and Ask Jeeves[6]. Of all these available tools, the performance of five, Windows desktop search, Google desktop search, Archivarius, Yahoo desktop search and Copernic desktop search are evaluated and analyzed in this paper using standard Information Retrieval evaluation measures.

The tools selected are evaluated with the help of an evaluation program (trec_eval) provided by Text REtrieval Conference (TREC) for assessing an ad hoc retrieval run[7]. Information retrieval measures like recall-level precision averages, document-level precision, document-level recall and fallout-recall average are used to evaluate the search engines. The Robust Track 2004 dataset[8] from TREC 2004 is used as benchmark for our experiments. The queries and data are used for each of the desktop search engines under consideration and the experiment results are compared and analyzed. We also compare the desktop search engines against a set of documents reflecting a typical desktop with a mix of text documents, spreadsheet, images, video and audio files to make the comparison more relevant and inclusive. TREC is used for the evaluation as it's a standard set of documents with baselined queries which makes the comparisons fair and repeatable.

Desktop search tools are relatively new and little work has been published with respect to comparing their performance evaluation. Preliminary comparisons have been conducted on the top search engines[9]. An evaluation of 12 leading desktop search tools has been performed based on a number of criteria including usability, versatility, accuracy, efficiency, security, and enterprise readiness[10]. While previous studies have evaluated on several different criteria, they have not focused on using standard information retrieval evaluation measures[11].

This paper is organized as follows. Section 2 introduces the five desktop search engines and provides information regarding the evaluation measures; Section 3 describes the evaluation procedures including the benchmark data, evaluation measures, and the experiment conducted; Section 4 presents the experimental results and analysis; and final conclusions and future directions are summarized in Section 5.

2. EVALUATION

The five desktop search engines are evaluated on the following criterion and measures on TREC documents:

- Recall-precision average
- Document-level precision
- R-Precision
- Document-level recall
- Mean Average Precision (MAP)
- Exact precision and recall over retrieved set
- Fallout-recall average
- Document-level relative precision
- R-based precision

These measures are chosen for evaluation as they provide insights into how desktop search engines incrementally retrieve documents and build their result sets for a group of queries and the impact that has on the accuracy of final query result sets. The evaluation on typical user desktop documents is done with average recall and average precision measures over all queries. We evaluate the following desktop search engines based on the above criteria.

Microsoft's **Windows desktop search** (WDS) application [1] is closely integrated with Windows. The tool provides options to index particular folders or file types on the computer. The search also allows results to be returned as ranked in order of

1-4244-1500-4/07/$25.00 ©2007 IEEE

relevance or unranked.

Yahoo desktop search (YDS) is based on X1 desktop search[2]. YDS takes a "reductive" approach to displaying results. It helps selectively index only the content that is chosen like files, emails, IMs, contacts and to set individual indexing options for each type of content. YDS provides fine grained control over indexing options like specifying the folders that should be indexed or the file types that can be indexed. YDS allows saving queries for later use, and organizing these searches alongside the generic queries in the search pane.

Copernic desktop search (CDS) allows files types to be selectively indexed. User can choose to index video, audio, images, and documents. It allows third-party developers to create plug-ins that enable new file type indexing[12]. For business use, Coveo[13], a spin-off company from Copernic, provides enterprise desktop search products with enhanced security, manageability, and network capability.

Google desktop search tool allows users to scan their own computers for information much the same way as they do for using Google to search the Web. Out of the many features this tool provides[14], noteworthy features include returning search results summarized and categorized into different supported file types with a total count of matches associated with each type.

Archivarius desktop search is a full-feature application designed to search documents and e-mails on the desktop computer as well as network and removable drives[5]. It allows files to be searched on many advanced attributes like modification date, file size, and encoding.

3. EVALUATION PROCEDURE

The benchmark dataset, evaluation programs and procedures used to perform the experiment are described in detail in this section.

3.1 TREC Documents Evaluation

The Robust 2004 track is one of the tracks from the TREC 2004 dataset of raw document collections and consists of 4 sub-collections, namely the Foreign Broadcast Information Service (FBIS), Federal Register (FR94), Financial Times (FT) and LA-Times collections. These documents are mainly newspaper or newswire articles and government documents. The total size of the track is 1.904 GB and contains approximately 528,000 documents on 250 topics. Robust track collection is chosen for our experiments as its queries are designed to perform poorly on the document set and are best suited to measure the consistency of the information retrieval ability of search engines.

The “Robust Queries” or topics, which are also parts of Robust track, are used as sampling queries to asses the performance of the desktop search engines. There are several types of queries in this query collection, including long query, short query, and title query. The title query, a total of 100 queries is used to perform the experiment and evaluate the engines. The evaluation program used to perform the experiment is the TREC evaluation program (trec_eval)[7]. The TREC evaluation program is the standard tool for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results[15, 16]. Retrieval tasks whose results are a ranked list of documents can be evaluated from the TREC evaluation program.

The experiments consist of following steps as illustrated in Figure 1. The TREC dataset used for the evaluation (TREC 2004 Robust track) was first downloaded on to the hard drive of the computer to a separate folder, and only this folder was indexed for each of the systems under evaluation, so that other files already present in the computer would not interfere with the experiments and the results. After downloading the desktop search tool, all the queries are executed on this test dataset for each of the desktop search tools, and the retrieved documents are stored in a file (results file) to be used as one of the input files to the TREC evaluation program.

Each of the 100 queries is supplied to the desktop search tools, WDS, Yahoo, Archivarius, Google and Copernic, one at a time and the results, documents retrieved, are stored in a file (result files are formatted as described in the readme file of the evaluation program that accompanies the dataset) and is given as one of the inputs to the evaluation program. The query relevant documents file (qrels) provided by TREC is used as the other input file.

The three result files including the query results file, the query file, and the query relevant documents file, are obtained after executing the 100 queries on each of the five tools under consideration. The evaluation program is then executed once for each set of the results files to obtain the results for each run. The final evaluation files obtained are then compared and analyzed to produce resulting tables and plots that help in evaluating and comparing the tools.

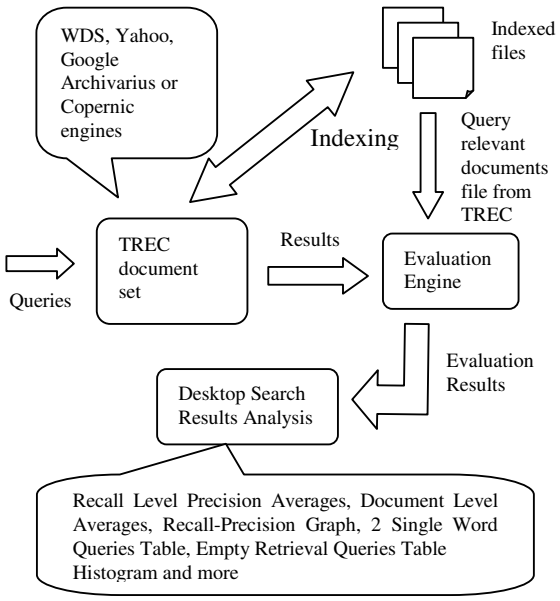

Figure 1: TREC Corpus Experiment Design.

3.2 User Desktop Documents Evaluation

In this part of the experiment, we focus on testing the desktop search engines using the representative user desktop document set. These documents are what a typical user will have on their desktop. This document set consists of 692 files and 2 subfolders on a Windows XP desktop taking 259MB of hard

drive space. These files are chosen to represent a typical desktop environment and comprise of Windows Word documents, Windows Excel spreadsheets, image files of jpeg, bmp and gif formats, Adobe pdf files, simple text files created using Windows Notepad, html files, Windows PowerPoint presentation files, and xml files. Most files are in the main folder and some are in the 2 subfolders within the main folder. A set of queries created against the document set, 15 in all are ran against the document set and the number of documents retrieved are used to calculate the precision and recall rates for each desktop search engine. The architecture of the experiment is similar to one shown in Figure 1.

3.3 Evaluation Environment

The TREC experiment is performed on a laptop with environment resembling a typical desktop machine used by a user at home or work. It has 40GB hard drive, 128 MB RAM, 1GHz process speed and running Windows XP Home Edition operating system. The user desktop experiments are performed on a machine with Intel dual core 920D processor with two 2.4GHz cores, 1GB of RAM and 80GB hard drive. All the experiments for each document set are conducted on the same machine so that the results obtained are uniform.

4. Experiment Results and Analysis

We now present the results of the experiments conducted on the TREC and desktop document set.

4.1 TREC Corpus Results

The evaluation results automatically generated by the TREC evaluation program are based on standard information retrieval precision recall based measures. Statistical results for total number of documents over all 100 title queries are in Table 1.

Desktop search engines	Total documents retrieved	Total relevant retrieved
Yahoo	334235	504
WDS	368024	467
Google	28182	70
Copernic	535957	695
Archivarius	2071186	973

Table 1: Retrieval Summary

Table 1 shows that in total relevant document retrieved, Copernic has a slight edge over all other desktop search engines and Google retrieves the least total relevant documents. However it should be noticed that Google retrieves far less total documents than others. That indicates that Google is more selective in the documents it accumulates in the result set.

Google and WDS allow results to be returned as ranked in order of relevance or unranked. The results for the two are shown separately. The recall-precision average is illustrated in Figure 2. The graph shows the interpolated recall-precision average at recall levels from 0 to 0.3. After 0.3 recall, the average starts to approach 0 for all search engines. This measure shows the dramatic fall in the recall–precision average as the recall rate increases. This could be due to the high degree of precision-recall when the recall rate is low and subsequent decrease as lower ranked documents tend to deteriorate the precision average rapidly. The recall-precision average summary as illustrated in Figure 2 shows that Google performs much better than the rest of desktops with WDS ranking the second when the percentage of all the relevant docs for all queries that have results retrieved is within 10%. Yahoo and Copernic have almost the same performance curve. The ranked results of both Google and WDS have higher precision-recall averages than their unranked results.

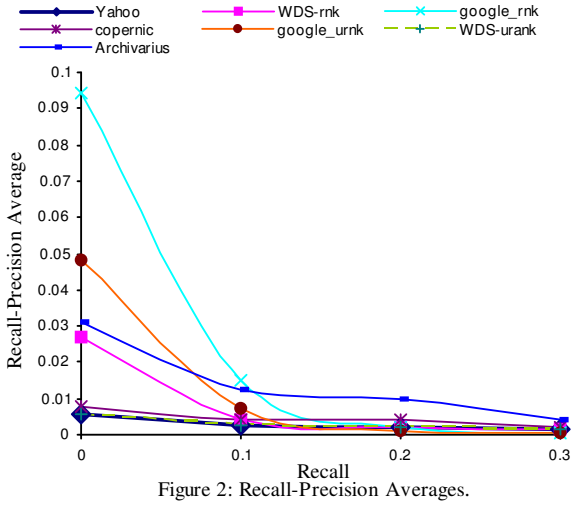

Figure 2: Recall-Precision Averages.

The evaluation program results for document level precision are shown in Figure 3. As expected, the precision of the results decreases with the increase in number of documents in the result, as the ratio of number of relevant documents retrieved to the total retrieved documents decreases as the number of documents in the result increases. As shown in Figure 3, ranked results for Google and WDS have much better document level precision rates than the rest of engines, especially when the document count is low. Precision rates are also high for Google's unranked results, showing Google is generally more precise than other engines in retrieving documents.

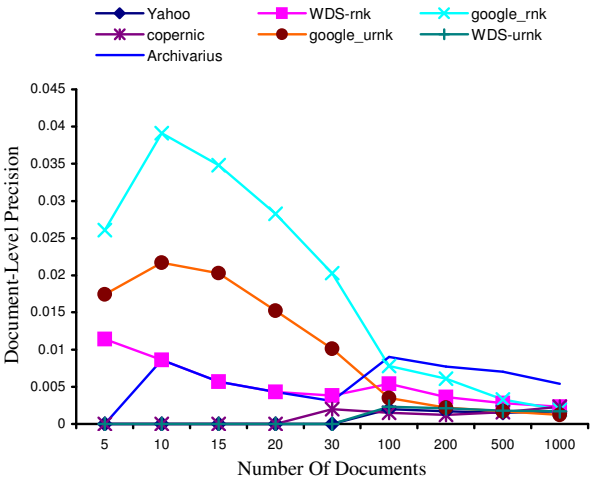

Figure 3 - Document-Level Precision.

R-Precision measures in Figure 4 show high precision rate for Google's ranked results. This shows that Google's ranked document in the result set have the highest precision or the number of relevant documents compared to the total number of documents in the result set while Copernic and Yahoo have the lowest. This shows that Google is better at measuring relevancy of the document before it is added to the result set compared to other search engines.

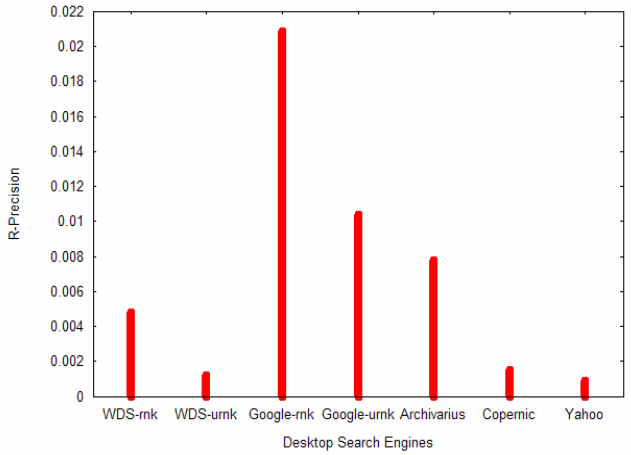

Figure 4 – R-Precision.

Figure 5 shows the average precision of retrieved document sets over returned documents level of recall for all the queries ran against the dataset using the 5 evaluated search engines. It shows the overall average precision of the results at various revels of recall. Google and WDS ranked and unranked results evaluations by TREC evaluation program are shown separately as the number of relevant documents at various levels of recall for unranked results will be different from ranked results. The figure clearly shows high MAP rates for Google's ranked results and Archivarius. This measure shows that Google and Archivarius tend to keep the relevancy levels high at various levels of recall or the number of relevant documents in the result set the highest. The R-Precision value for Google's unranked results are poor indicating that if unranked results are retrieved, then the precision over queries deteriorates due to most relevant documents not appearing early in the result set.

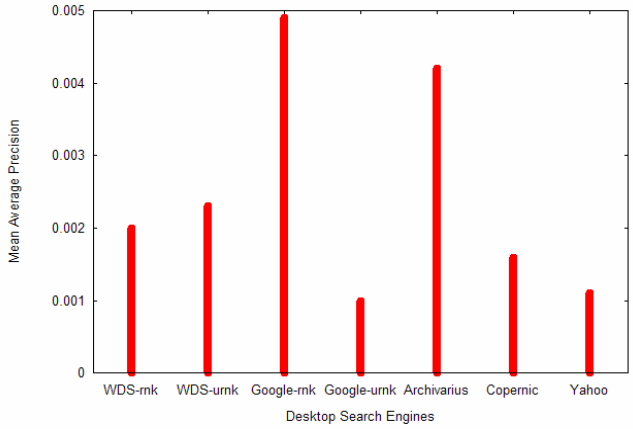

Figure 5 – Mean Average Precision (MAP)

Figure 6 shows the recall level for each of the search engines when incrementally increasing number of documents returned in the query results. As expected, all search engines show increasing number of relevant documents added to the result set with increase in the number of documents in the result set. It shows that for all search engines, the recall level increases with the increase in number of documents retrieved. The document level recall values rise fastest for Archivarius with increase in documents. This indicates Archivarius is less discriminating in adding a document to the result set.

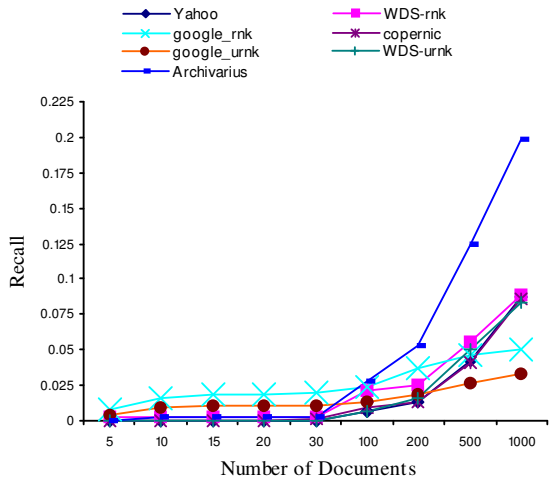

Figure 6: Document-Level Recall.

Figure 7 shows results of exact precision rate for each search engine for the entire result set as evaluated by the TREC evaluation program. Both Google's ranked and unranked results have high Exact precision over retrieved sets. This shows that Google recalls the most number of relevant documents overall in the final result set but it is not markedly better in this measure than all other search engines.

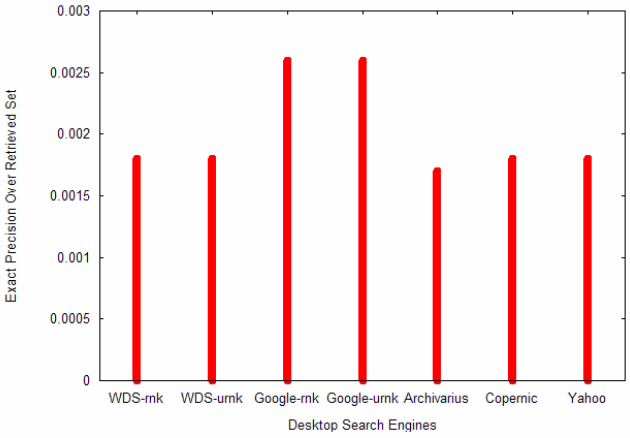

Figure 7 – Exact Precision Over Retrieved Set.

Exact recall rate is the recall of relevant documents compared to total retuned documents for the entire result set. Figure 8 shows the exact recall rate for the search engines. Recall is highest for Archivarius as figure shows while the result for Google is poor. This indicates that in terms of retrieving the

most number of relevant documents, Archivarius and Copernic perform better than others. As previous results indicated, Google tends to perform much better in retrieving the most number of relevant documents in the first set of returned documents in the result but overall it sacrifices recall rate for improved precision.

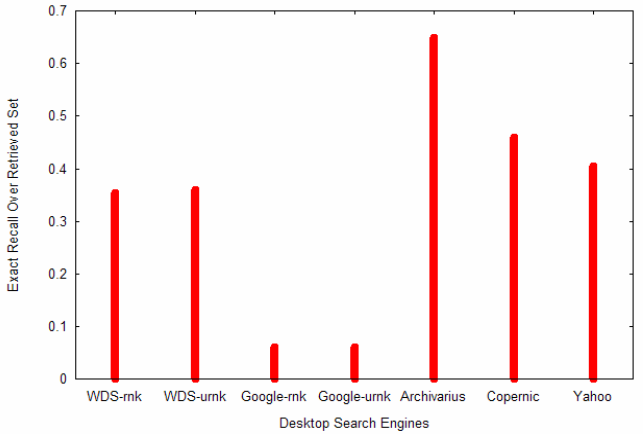

Figure 8 – Exact Recall.

The fallout recall average shown in Figure 9 provides average fallout rate at various levels of recall for all queries. As shown in Figure 9, the fallout-recall average is highest for Archivarius and Google. This indicates that Archivarius and Google keep out most number of relevant documents in the result set as they return incrementally larger document sets to avoid precision loss unlike other search engines from the result.

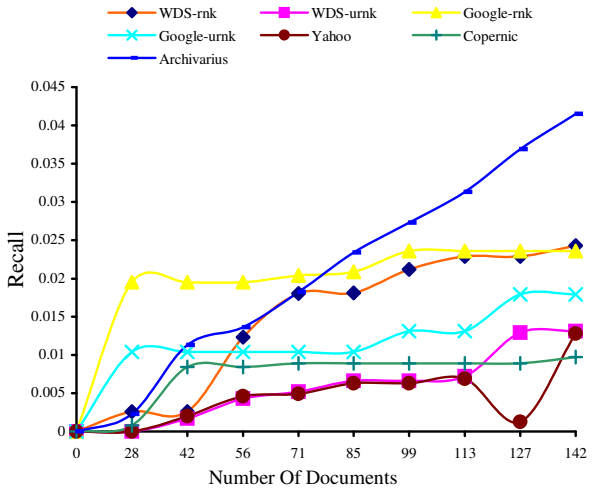

Figure 9: Fallout-Recall Average.

Figure 10 shows the relative precision after certain numbers of documents are retrieved. The general trend is different from document-level absolute precision for top 1000 documents as shown in Figure 3. The top performers are Archivarius, WDS and Yahoo with Archivarius precision increasing rapidly after 100 documents. This shows that while Google's precision stays about the same as more documents are incrementally returned in result set, precision of Archivarius starts slowly but increases rapidly as more documents are added to the result set. Since overall precision of Google is better, it shows that Archivarius is simply getting better at adding more relevant documents with increase in number of documents in result. Google's algorithm does that from the very beginning when building the resulting document set.

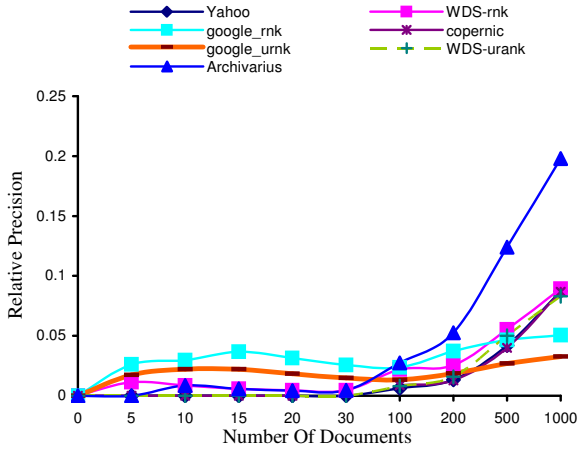

Figure 10: Document-Level Relative Precision.

R-based precision measures are shown in Figure 11. The top performers are Google, WDS with relevance ranking options and Archivarius. Google search engine outperforms the rest of tools. This indicates that Google is very precise in returning relevant documents in the first set of relevant documents returned while other search engines tend to have similar precision levels at various levels of recall. The high precision of Google's results at early levels of recall stand in stark contrast to tools like Yahoo and Copernic that have very low precision and it improves only slightly indicating these engines are weak in identifying relevant documents.

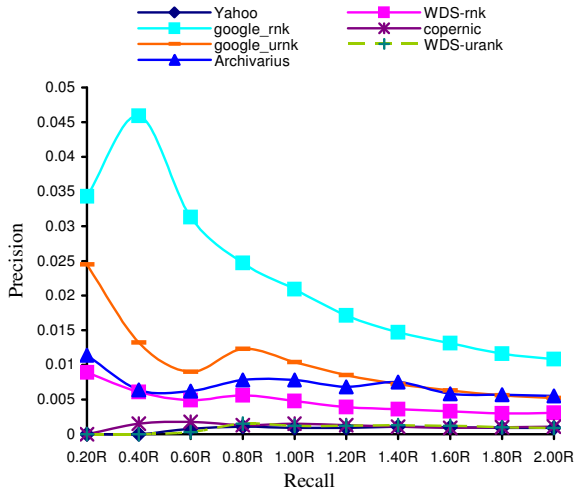

Figure 11: R-Based Precision.

From the analysis performed on the results obtained from the evaluation program, it can be concluded that Google and Archivarius desktop search performed better than the rest in most information retrieval specific evaluation measures,

although the precision is low for Archivarius and recall is low for Google. The low precision of most engines could be attributed to the manner in which the relevance file containing relevant documents for queries (qrels file, one of the input files to the evaluation program) is created. It is created by taking into account the related information of the queries to retrieve documents in addition to straight keyword searches. Hence the documents expected to be retrieved do not have to contain the exact keywords in the query but any word related to its meaning. Most desktop search engines perform word matches rather than look for the meaning of the query resulting in low precision. This also results in single word queries performing better than multiword queries and some difficult words not retrieving any results at all.

From these results, we can conclude that users should choose Archivarius if they prefer highest recall that is most number of relevant documents retrieved for a query. For highest precision, i.e., for most number of relevant documents and least number of non-relevant documents, Google DTS is the tool of choice. The underlying trend in the result points to the uniformly poor performance for a standard set of documents by all search engines evaluated using traditional IR measures.

4.2 User Desktop Experiment Results

We performed experiments with desktop search engines on a set of documents that reflect a typical user desktop. The results of the experiment for the five desktop search engines as average recall and precision rates for all the queries ran against the document set and the documents retrieved are shown in Table 2.

Desktop Search Engine	Average Precision	Average Recall	F-Measure
Google	0.77	0.93	0.84
Copernic	0.54	0.95	0.69
Archivarius	0.56	0.84	0.67
Yahoo	0.56	0.85	0.68
WDS	0.55	0.93	0.69

Table 2: User Desktop Documents Retrieval Results

User desktop search results in Table 2 show Google DTS performs better in precision compared to others. The average recall is highest for Copernic with lower precision indicating that it retrieves large number of documents that contain many irrelevant documents, whereas Google has high recall and also keeps precision level high. The recall rate of Archivarius though higher than other searches engines for the TREC document set is lower than others for the representative desktop document set. F-measure values show overall better performance of Google DTS compared to other engines. These results show that the search engines perform much better on actual desktop documents in terms of precision and recall than they do on the TREC document set, but the relative precision and recall levels for the different search engines stay approximately the same. This confirms the relevance of the results over TREC documents as a way to measure relative effectiveness of the search engines.

5. CONCLUSIONS

The results obtained in our study provide revelations in the available desktop search engines retrieval capabilities not identified in previous studies. The information retrieval measures based comparison is unique to our evaluation. The results also differ from previous evaluations when functionality of evaluated applications is combined with the information retrieval measures results. The TREC Robust 2004 track provides suitable test data in terms of the size of data and the types of queries provided cover a wide range of areas. The typical user document set results validate conclusions from TREC document set as applicable for searches on a typical desktop.

In the future we plan to extend the evaluation on user desktop document set to all the IR measures provided by TREC evaluation engine. We plan to include more of available tools in the marketplace for comparison in the evaluation experiment to provide the users insight into the relative strengths and weaknesses of the available tools. Desktop search tools being relatively new applications will be improved over time and comparison of all the tools will play a major role in users selecting the right tool for their desktop computers.

6. REFERENCES

1. Microsoft, http://www.microsoft.com/windows/desktopsearch. 2007.
2. Yahoo, http://desktop.yahoo.com/. 2007.
3. Copernic, http://www.copernic.com/en/products/desktop-search/. 2007.
4. Google, http://desktop.google.com/features.html. 2007.
5. Likasoft, http://www.likasoft.com/document-search/. 2006.
6. Jeeves, A., http://sp.ask.com/en/docs/desktop/overview.shtml. 2007.
7. Buckley, C., http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz. 2005.
8. Voorhees, E., *Overview of the TREC 2004 Robust Retrieval Track.* Proceedings of the Thirteenth Text Retrieval Conference (TREC-13), 2005.
9. Cole, B., *Search Engines Tackle the Desktop*, in *IEEE Computer*. 2005. p. 14-17.
10. Noda, T. and S. Helwig, *Benchmark Study of Desktop Search Tools*, in *Best Practice Reports, UW E-Business Consortium*. 2005, University of Wisconsin-Madison.
11. Beaza-Yates and Ribeiro-Neto, *Modern Information Retrieval*. first ed. 1999: Addison Wesley.
12. Copernic, http://www.copernic.com/en/products/desktop-search/. 2005.
13. Coveo, http://www.coveo.com/en/Products/CES.aspx. 2006.
14. Google, http://desktop.google.com/features.html. 2006.
15. TREC-Evaluation, http://trec.nist.gov/pubs/trec13/t13_proceedings.html. 2005.
16. TREC, http://trec.nist.gov/. 2006.