

# Direct Comparison of Commercial and Academic Retrieval System: an initial study

Yefei Peng  
School of Information Sciences  
University of Pittsburgh  
Pittsburgh, PA 15260, USA  
ypeng@mail.sis.pitt.edu

Daqing He  
School of Information Sciences  
University of Pittsburgh  
Pittsburgh, PA 15260, USA  
daqing@mail.sis.pitt.edu

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Systems and Software—*performance evaluation*

## General Terms

Experimentation, Measurement

## Keywords

Google Desktop, Indri, Baseline, Enterprise email retrieval

## 1. INTRODUCTION

Searching for information has increasingly prevailed in people's life. A new trend in the study of information retrieval is the amplified interest in search over corporation and personal information collections. Recently released commercial desktop search engines are the achievements due to this trend. Comparing to the Web search engines, these desktop search engines are working on collections with relatively small size, and often rely on the machine power of a single desktop computer. These new developments provide an interesting opportunity for the evaluation of academic retrieval systems. Academic retrieval systems are systems developed in academic environment for research use. Many of them are open source systems. For many years, due to various reasons like collection size and machine resources, commercial search engines and academic retrieval systems are studied and evaluated in non-interrelated environments. Now, since commercial desktop search engines and academic retrieval systems are working on collections with similar size and on comparable machine power, it becomes possible to directly compare their performance.

In this paper, we want to compare two off-shelf retrieval systems. They are Google Desktop (GDS) v20051208 and Indri 2.0. GDS is a popular desktop search engine, which provides full text search on various types of files in a personal computer, including emails<sup>1</sup>. Google didn't disclose what is the retrieval model used in GDS. Indri 2.0, is a state-of-art retrieval system developed by University of Massachusetts Amherst and Carnegie Mellon University<sup>2</sup>. It combines an

<sup>1</sup><http://desktop.google.com/>

<sup>2</sup><http://www.lemurproject.org/indri/>

inference network with a language- modeling approach, and has been widely used in TREC experiments. The reason that we use these system in off-shelf manner without any tuning is because that is often what happens when most users use retrieval systems.

Our research questions were: 1) Can commercial desktop search engines like GDS be employed in direct comparison with the state of art academic retrieval systems like Indri on TREC like platform? 2) How does GDS's performance comparing to that of Indri in off-shelf manner? 3) Is there any limitation of GDS could be identified through our study?

## 2. EXPERIMENTS ON EMAIL RETRIEVAL

Our study was built on the email search tasks defined in Enterprise Track of TREC 2005 [1]. The email search task had two subtasks: searching for known emails in the collection (called "known item search") and searching for emails discussing certain topics (called "discussion topic search"). The collection was W3C email collection, which was based on the crawls at w3c.com website in June 2004. After removing empty or duplicated emails, the W3C email collection consisted of 174,294 emails with average size 9.8KB. In our experiments, Indri was running on a Dell 8300 computer, CPU 3.2GHz, Fedora 4 OS. GDS was running on a Dell 3000 computer, CPU 3.0GHz, Windows XP Pro OS.

In the known item search task, there were 125 evaluation topics, and the average length of queries was 5.42 words. Top 100 returned emails were evaluated against the relevant judgments provided by TREC. The measures included the mean reciprocal rank (MRR) of the correct answer, the fraction of topics with the correct answer in the top 10 returned emails (Success at 10 or S@10), the fraction of topics that found the correct answer in the returned 100 emails (S@inf).

In the discussion topic search task, there were 59 topics, each of which has a short title and a longer description. The average length of title was 3.65 words, whereas that of the description was 21.3 words. Top 1000 returned emails were evaluated against the ground truth provided by TREC. The measures were the mean average precision (MAP), R Precision (R-Prec), and Precision at top 10 (P@10).

We performed in total five different retrieval runs for discussion topic search using GDS and Indri: *Indri-TitDes*: queries used the titles and the descriptions, and the searches were on Indri; *Indri-Tit*: queries used the titles only, and the searches were on Indri; *GDS-TitDes*: queries used the titles and descriptions, and the searches were on GDS. *GDS-Tit*: queries used the titles only, and the searches were on GDS. *GDS-Tit-Rev*: queries used the terms in the titles but in

**Table 1: The Results of Known Item Searches**

Runs	MRR	S@10	S@inf	Time(s)
Indri	0.2544	0.392	0.544	7.592
GDS	0.2994	0.448	0.528	1.092
GDS-Rev	0.3154	0.480	0.528	0.999
TREC-Best	0.621	0.784	0.920	-

**Table 2: The Results of Discussion Topic Searches**

Runs	MAP	R-Prec	P@10	Time(s)
Indri-TitDes	0.1846	0.2137	0.2831	9.621
Indri-Tit	0.1871	0.2287	0.3356	9.829
GDS-TitDes	0.0002	0.0002	0.0034	0.701
GDS-Tit	0.1887	0.2375	0.3627	0.702
GDS-Tit-Rev	0.1874	0.2308	0.3763	0.811
TREC-Best	0.3782	0.4051	0.5000	-

reversed order, and the searches were on GDS.

To stick to the off-shelf approach, we used Indri as a straightforward plain retrieval system without employing any extra technique to boost the performance of Indri system, including various query expansion techniques. However, after noticing that the word order of a query had impact in GDS searches, we performed all email searches on GDS with query terms in their original and reverse order respectively. We also list best result of TREC-2005 on the same tasks as "TREC-Best" in tables to indicate the true state of art of the academic email search.

### 3. RESULTS AND DISCUSSION

As shown in Table 1, in known item search, the two GDS runs (i.e.,GDS: with original query term order; GDS-Rev: with reverse term order) outperformed the Indri run. The relative MRR improvement of the two GDS runs over that of Indri run were 17.7% and 24.0% respectively. The difference was significant between GDS-Rev and Indri (two-tailed t-test resulted in  $p$  values of 0.046), and nearly significant between GDS and Indri ( $p = 0.076$ ).

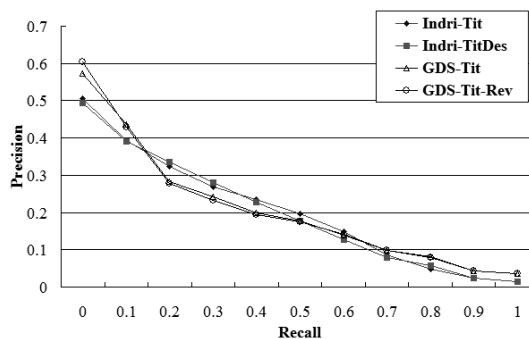
Interestingly, GDS-Rev achieved significant improvement ( $p=0.012$ ) over GDS when measured by S@10. This might indicate that the word order in GDS is really important. However, without knowing the actually query processing techniques in GDS, we cannot tell for sure.

The results for discussion topic search are in Table 2. It seems that GDS has trouble handling long queries. GDS-TitDes generated really low performance. The other two GDS runs (GDS-Tit and GDS-Tit-Rev) outperformed either of the two Indri runs, but the improvement was only significant when comparing the two GDS runs with Indri-TitDes under the measure of P@10. The  $p$ -values in two tailed paired t-test were 0.008 and 0.003 respectively.

We also plotted the results of four runs Indri-Tit, Indri-TitDes, GDS-Tit, GDS-Tit-Rev into 11-points precision/recall graph (see Figure 1). The Figure shows that GDS preformed relative superior at high precision end over Indri runs, but almost no difference at high recall end.

In both tasks, TREC-Best greatly outperformed all GDS and Indri runs. This shows the effect from training, and tuning, and also demonstrates the limitation of using off-shelf systems directly without any modification.

GDS did demonstrate some limitations through our study.

**Figure 1: 11-points precision graphs showing performance on discussion topic search.**

Firstly, it has problems in handling long queries. When searching for discussion topic with queries containing terms from titles and descriptions, which in average are about 25 words long, GDS was struggled in performing the search, and returned averagely only 0.07 relevant documents per topic. Secondly, although not being documented anywhere, it is apparent that the order of query terms does make noticeable or even significant difference on search results in GDS. Therefore, it is important to study further to establish the optimal order of query terms in GDS.

Although not directly from our study, we identified several other issues with GDS. First, it doesn't support explicitly assigning weights to query terms as those in Indri. GDS only allow user use repetitive words to emphasize some query words [2]. This limits GDS' ability to provide more advanced control of the retrieval system. Second, according to GDS Help Center, GDS only indexes the first 10,000 words in a document. Although this did not make great impact to rather short email documents, it will limit GDS's ability to handle long news documents that are common in TREC.

### 4. CONCLUSION

In this paper, based on TREC Enterprise email search tasks, we performed initial comparison between GDS, a commercial desktop search engine, and Indri, a widely used academic retrieval system. Although GDS still have issues regarding long queries, long documents, and the optimal sequence of query terms, it generated comparable results to the plain untuned Indri system in both known email search task and discussion email search task. As a representative from popular commercial desktop search applications, GDS is a good baseline for establishing the direct comparison between commercial and academic retrieval systems.

### 5. REFERENCES

- [1] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *Proceeding of TREC 2005*, 2005.
- [2] T. Calishain and R. Dornfest. *Google Hacks*. O'Reilly, December 2004.