Inverse Document Frequency and Web Search Engines

Kevin Prey, James C. French, Allison L. Powell
Department of Computer Science*
University of Virginia
Charlottesville, VA
{kjp4f | french | alp4g}@cs.virginia.edu

Charles L. Viles
School of Information and Library Science*
University of North Carolina, Chapel Hill
Chapel Hill, NC
viles@ils.unc.edu

Technical Report CS-2001-07
February 5, 2001
Analysis based on data collected January 11, 1999

INTRODUCTION

Full text searching over a database of moderate size often uses the inverse document frequency, $idf = \log(N/df)$, as a component in term weighting functions used for document indexing and retrieval. However, in very large databases (e.g. internet search engines), there is the potential that the collection size (N) could dominate the *idf* value, decreasing the usefulness of *idf* as a term weighting component. In this short paper we examine the properties of *idf* in the context of internet search engines. The observed *idf* values may also shed light upon the indexed content of the WWW. For example, if the internet search engines we survey index random samples of the WWW, we would expect similar *idf* values for the same term across the different search engines.

ABOUT THE EXPERIMENT

To examine the usefulness of *idf*, we devised the following experiment. We submitted a number of one-word queries to three different full-text internet search engines. We used the "hit count" as the document frequency, and the search engine's size as the collection size. This enabled us to reasonably estimate each term's *idf* value.

About the Queries

Jansen et al. [1] examined log files from the Excite search engine and found that 74 terms occurred more than 100 times in their sample space. We used those 74 terms then omitted very common terms (e.g. "the", "and") as well as an expletive. This left 56 terms for our query set (see Table 1). Using this query set, we examined the Altavista, Excite, and Infoseek search engines. These three search engines were selected because they are among the largest and most commonly used full-text sites on the internet. At the time of this experiment Altavista was the largest of the sites with about 140 million documents indexed; Excite and Infoseek indexed about 55 and 43 million documents, respectively.

THE RESULTS

When examining the document frequency values of the query terms, we noticed two unusual behaviors of the search engines. The most peculiar was Excite's "hit count". Each time that we polled Excite, we found that many of the hit counts were not unique. Stated more clearly, there were far too many ties for us to believe that the results Excite was returning were actually accurate. The January 11, 1999 poll, which included 36 non-unique entries, is shown in Table 2. This behavior was present in all polls, even after Excite had

adult	american	anal	art
basketball	big	black	business
car	celebrities	chat	city
college	company	computer	employment
erotic	estate	florida	free
games	gay	girls	high
history	internet	jobs	john
magazine	men	music	naked
ncaa	news	nude	photos
pics	pictures	porn	porno
real	school	service	sex
software	state	stock	stories
texas	university	video	war
women	wrestling	xxx	young

Table 1: Listing of Query Terms

celebrities	65317	pictures	834806
wrestling	74962	games	1026376
porno	86031	black	1099546
erotic	105773	men	1099546
ncaa	105773	women	1099546
anal	113313	young	1177933
porn	121391	video	1261907
nude	139316	art	1448243
naked	171287	big	1448243
gay	210593	college	1448243
xxx	210593	music	1448243
pics	241690	history	1780584
basketball	297153	american	1907522
girls	419291	john	1907522
estate	552260	real	2043509
florida	552260	computer	2189190
photos	552260	school	2189190
adult	591630	software	2189190
jobs	591630	city	2345257
employment	633808	company	2512450
chat	678992	university	2691563
sex	678992	business	3089005
stock	678992	high	3089005
stories	678992	internet	3309219
texas	678992	news	3309219
car	779253	state	3309219
magazine	779253	free	3545133
war	779253	service	3545133

Table 2: Excite Hit Counts on Jan. 11, 1999 (Non-unique results are shaded.

*This work supported in part by DARPA contract N6601-97-C-8542 and NASA GSRP NGT5-50062.

made updates to the collection. It appears that Excite is using some method of "binning" the results. Note that Excite uses "concept-based clustering," [2] but this does not seem to be the source of the problem. Neither Altavista nor Infoseek exhibited this behavior.

Altavista also had an unusual behavior. When query results were returned, in most cases Altavista would display:

AltaVista found about 1,276,203 pages for you.

While for some queries Altavista would return:

AltaVista found 7,156,760 pages for you.

We looked closely at the queries that produced each result, but were unable to determine why this occurred. It seemed unusual that the number of hits did not determine which output was returned. While Infoseek seemed to behave well, given the results from Altavista and Excite, we cannot be sure that any of the results from any of the sites are entirely accurate.

Even though we have doubts about how accurate the results are, we feel confident that the numbers returned are reasonable approximations of the actual values. Also, since *idf* is computed as a logarithm, any error in the document frequency is reduced significantly. We ran the polls of the search engines five times over a period of about one and a half months. The results of each poll were unique; however, the differences were always small and were probably due to updates to the collections. We ranked the results by the number of documents that were returned and calculated the corresponding values for *idf*. Using the results from January 11, 1999 (our most recent poll), we found that the values of *idf* varied significantly between the search engines, but when ranked by *df*, a given term was ranked similarly at each search engine. For example, the word "internet" was ranked 3, 2, and 5 by the three search engines in the study. Table 3 shows the 10 queries that returned the most documents and the 5 that returned the least, plus the *idf* values for each query. Note that 8 of the top 10 queries are the same for all three search engines.

We calculated the Spearman coefficient of rank correlation, *R*, for the three complete rankings of 56 queries. The values of *R* range from -1 to 1, where *R=1* when the two rankings are in perfect agreement and *R=-1* when they are in perfect disagreement. The three *R* values we calculated were all above .92, which confirms a strong similarity in the rankings. Table 4 shows the three *R* values.

DISCUSSION

If each search engine collection were a random sample of the web, we would expect very high agreement among the calculated values for *idf*. There is no such agreement. The *idf* values differ widely across collections, however the search engines are remarkably consistent in their ranking of the terms. For a given search engine, there was also variability in the *idf* values, suggesting that the *idf* value is not dominated by *N* in very large collections and can therefore be a useful indexing component.

REFERENCES

[1] B. Jansen, A. Spink, J. Bateman, T. Saracevic, "Real Life Information Retrieval: A Study Of User Queries On The Web," SIGIR Forum, vol. 32, no. 1, Spring 1998, pp. 5-17.

[2] S. Lawrence and C. L. Giles, "Searching the World Wide Web," Science, vol. 280, no. 5360, 3 April 1998

	Excite			Altavista			Infoseek		
Rank	**Query**	**idf**	**df**	**Query**	**idf**	**df**	**Query**	**idf**	**df**
1	free	1.19	3545133	news	0.43	52210282	service	0.03	39956021
2	service	1.19	3545133	internet	0.44	51014980	free	0.05	37598516
3	internet	1.22	3309219	university	0.49	45682076	state	0.06	37324414
4	state	1.22	3309219	service	0.50	44092028	news	0.06	36755003
5	news	1.22	3309219	business	0.54	40021805	internet	0.07	36589013
6	business	1.25	3089005	company	0.64	32422031	high	0.07	36291280
7	high	1.25	3089005	free	0.65	31691435	business	0.07	36092258
8	university	1.31	2691563	state	0.65	31225414	company	0.08	35835654
9	company	1.34	2512450	school	0.67	29648515	computer	0.09	34962201
10	city	1.37	2345257	software	0.68	29205900	university	0.10	33685205
52	erotic	2.72	105773	erotic	1.75	2478870	anal	1.58	1120808
53	ncaa	2.72	105773	porno	1.77	2379450	erotic	1.69	872598
54	porno	2.81	86031	celebrities	1.93	1646730	ncaa	1.75	757629
55	wrestling	2.87	74962	ncaa	1.98	1473177	porno	1.86	586782
56	celebrities	2.93	65317	wrestling	2.18	914820	wrestling	2.24	243397

Table 3: Top 10 and Bottom 5 Queries for Each Search Engine (*df* = document frequency)

Altavista and Excite	0.9576	Altavista and Infoseek	0.9260	Excite and Infoseek	0.9649

Table 4: Spearman's R Values for Each Pair of Rankings