

Tagging and Searching:

Search Retrieval Effectiveness of Folksonomies on the World Wide Web

P. Jason Morrison, M.S.

Information Architecture and Knowledge Management Program (IAKM),  
School of Library and Information Science,  
Kent State University,  
PO Box 5190, Kent, Ohio, United States 44240

[pjmorris@kent.edu](mailto:pjmorris@kent.edu)

Present Address: 3585 Daleford Road, Shaker Heights OH 44120

## **Abstract**

Many Web sites have begun allowing users to submit items to a collection and tag them with keywords. The folksonomies built from these tags are an interesting topic that has seen little empirical research. This study compared the search information retrieval (IR) performance of folksonomies from social bookmarking web sites against search engines and subject directories. Thirty-four participants created 103 queries for various information needs. Results from each IR system were collected and participants judged relevance. Folksonomy search results overlapped with those from the other systems, and documents found by both search engines and folksonomies were significantly more likely to be judged relevant than those returned by any single IR system type. The search engines in the study had the highest precision and recall, but the folksonomies fared surprisingly well. Del.icio.us was statistically indistinguishable from the directories in many cases. Overall the directories were more precise than the folksonomies but they had similar recall scores. Better query handling may enhance folksonomy IR performance further. The folksonomies studied were promising, and may be able to improve Web search performance.

## **Keywords**

information retrieval, folksonomy, social tagging, social bookmarking, World Wide Web, search

## 1 Introduction

Since the early days of the World Wide Web users have had a choice of competing information retrieval (IR) systems to satisfy their information needs. Two types of systems were prevalent: search engines, with automated methods to collect documents on the Web and full-text search, and directories, with documents collected and categorized by human experts. After years of competition, a small number of search engines with advanced algorithms dominate information seeking on the Web (Sullivan, 2006). This dominance might suggest that the question of which IR system was best for the Web had been settled. Despite advertisements asking, “do you Yahoo?” (Kaser, 2003), users have settled on using “to Google” as the verb for Web search (Quint, 2002).

Recently sites have begun to employ new methods to make Web surfing a social experience. Users of social bookmarking sites like Del.icio.us (<http://del.icio.us/>) can add Web documents to a collection and “tag” them with key words. The documents, tags, relationships, and other user-supplied information are compiled into what is called a “folksonomy” (Gordon-Murnane, 2006). Users are able to browse or search the folksonomies in order to find documents of interest. Most have with mechanisms to share items with others and browse other users' commonly used tags.

The word "folksonomy" is derived from taxonomy. Taxonomies can be found on the Web in the form of links arranged in a hierarchical system of exclusive categories (Rosenfeld and Morville, 2002, p. 65-66). Subject directories such as Yahoo

(<http://dir.yahoo.com>) serve as taxonomies cataloging links to millions of documents.

Taxonomies are usually controlled by experts and are fairly static, tending to use official terminology rather than vernacular phrases. Folksonomies, in contrast, are distributed systems of classification, created by individual users (Marieke Guy, 2006).

Most academic papers on the use and effectiveness of folksonomies have been descriptive (Dye, 2006; Fitcher, 2006; Al-Khalifa and Davis, 2006; Chudnov, Barnett, Prasad and Wilcox, 2005). To place folksonomies in context with familiar topics of study like search engines and subject directories, a study was done to examine the effectiveness or performance of systems employing folksonomies compared to more traditional Web IR systems.

User strategies for information seeking on the Web can be put into two categories: browsing and searching (Bodoff, 2006). Although it would be very interesting to study the effectiveness of folksonomies versus traditional hierarchical taxonomies when users browse a catalog of Web documents, studying search performance is more straightforward. Traditionally IR performance is measured in terms of speed, precision, and recall, and these measures can be extended to Web IR systems (Kobayashi and Takeda, 2000, p 149). Precision and recall are the primary measures of interest. Precision is defined as the number of relevant results retrieved by a particular IR system divided by the total number of items retrieved. Recall is traditionally found by dividing the number of relevant documents retrieved by an IR system by the number of relevant

documents in the collection as a whole. Web IR systems have very large collections and the actual number of relevant items for a given query is unknown, so it is impossible to calculate absolute recall. A relative recall measure can be defined as the number of relevant items returned by one IR system divided by the total number of relevant items returned by all IR systems under study for the same search.

It is important to note that Web sites that employ folksonomies, even those included in this study, are not necessarily designed to have search as the primary goal. Studying folksonomies in this way is valuable, however, because these systems are designed to organize information and search is an important and commonly used IR function on the Web.

There is no single widely accepted definition of folksonomy, so it is important to explain how the term is used in this study. A very strict definition of folksonomy might only include the tags and their relationships, ruling out any system that included document titles and descriptions, rating and recommendation systems, etc. One folksonomy might rely heavily on social network connections between users while another ignores them. For the purposes of this study, a broad definition is used:

1. The collection is built from user contributions.
2. Users participate in distributed classification or evaluation.
3. There is a social networking aspect to the addition, classification, or evaluation of

items.

Item two is an important point when considering sites like Reddit (<http://www.reddit.com>) and Digg (<http://www.digg.com>). These systems allow users to contribute document titles, categories, “up” and “down” votes to affect the ranking of items in the collection, and other information. A more strict definition that requires explicit tagging would leave such sites out of consideration.

## **2 Review of related literature**

Folksonomy performance has not been extensively investigated so this study relied on previous literature to determine how users search the Web and how to best evaluate IR performance.

### **2.1 How users search the Web**

Any study of Web IR should support normal search behavior and involve the kinds of queries users generally enter. Real-world search logs have been studied extensively. In a 1999 study, Silverstein, Henzinger, Marais, and Moricz examined query logs from AltaVista that included over one billion queries and 285 million user sessions. The study had three key findings:

1. Users generally enter short queries;
2. Users don't usually modify their queries; and
3. Users don't usually look at more than the first 10 results.

Jansen, Spink, and Saracevic (2000) and Spink, Dietmar, Jansen, and Saracevic (2001) studied queries from Excite users and had very similar findings. In addition, they found that relevance feedback and Boolean operators were rarely used and were as likely to be employed incorrectly as correctly.

Jansen and Spink (2006) compared the results of nine different large-scale search engine query log studies from 1997 through 2002. They found that in U.S. search

engines around 50% of sessions involved just one query. Query length was fairly uniform with between 20% and 29% of queries containing just one term. Studies of single-site search engines (Chau, Fang, and Liu Sheng, 2005) have found that searches were similar to general Web searches in the number of terms per query and the number of results pages viewed.

## **2.2 Measuring IR performance on the Web**

Studies of IR performance can examine systems with a defined database or systems that retrieve information from the Internet as a whole, with the latter more relevant to the topic at hand. Performance studies generally use some measure of relevance to compare search engines. Greisdorf and Spink (2001) gave a good overview of the various ways in which relevance can be measured.

Web search engines have been studied for more than a decade. In one relatively early study, Leighton and Srivastava (1999) compared the relevance of the first 20 results from five search engines for 15 queries. Although earlier studies of search engine effectiveness exist, the authors went to lengths to describe and use a consistent, controlled methodology. Queries came from university library reference desk questions and an earlier study and were submitted in natural language, making no use of Boolean or other operators. The researchers were prevented from knowing which engine a particular result came from when judging relevance and performance differences were tested for statistical significance. Result documents were placed into categories based on Mizzaro's



(1997) framework for relevance. Overall relevance was measured by “first 20 precision” with an added factor to account for the effectiveness of ranking. The study found differences in relevance scores based on which relevance category was used, and found the best search engines performed significantly better than the worst. Table 1 presents additional relevant details about this and other comparable studies.

A 1999 study by Gordon and Pathak looked at eight search engines and calculated recall and precision measures. The researchers found that "shootout" studies that pit search engines against each other often only considered the first 10 to 20 results, fewer than many traditional IR studies. They developed a framework of seven features thought to contribute toward the usefulness of such a shootout:

1. Searchers with genuine information needs should be the source of the searches;
2. In addition to queries, information needs should be captured fully with as much context as possible;
3. The number of searches performed must be large enough to allow meaningful evaluations of effectiveness;
4. Most major search engines should be included;
5. Special features of each search engine should be used, even if that means the actual queries submitted to different engines will not be identical;
6. The participants that created the information need should make relevance judgments; and
7. The experiments must be designed and conducted properly, using accepted IR

measurements and statistical tests to accurately measure differences (p. 146-147).

The researchers, following a procedure suggested by Hull (1993), calculated the precision and relative recall of each engine for the first 15 documents, first 16, etc. up to the first 20 and averaged the measurements to generate the average at document cut-off value 15-20 or DCV(15-20). They found statistically significant differences in precision and recall at all document cut-off numbers studied.

A later study by Hawking, Craswell, Bailey and Griffiths (2001) studied effectiveness using queries culled from Web server logs. They generally agreed with Gordon and Pathak's (1999) list of seven features, but found the requirement that those providing the information need evaluate the results too restrictive and thought it reasonable to present the same query to each engine. The authors also proposed an eighth desirable feature:

8. The searches should cover information needs with different topics and with different types of desired results. (p. 35)

They presented four different types of information needs based on the desired results (Table 2). The study found search engine performance results surprisingly consistent with Gordon and Pathak (1999). For most engines, the precision decreased slowly as the number of results considered increased.

Further evaluations of search engine IR performance have continued. In a study

published in 2004, Can, Nuray, and Sevdik devised and tested an automatic Web search engine evaluation method (called AWSEEM) against human relevance judgments. The number of search engines retrieving a document was used as a measure of relevance, like an earlier study by Mowshowitz and Kawaguchi (2002), but AWSEEM also took into account the intersection of the retrieved documents' content. The study found a strong, statistically significant correlation between AWSEEM and human results when looking at the top five results or more for each engine.

### **2.3 The overlap of search results between systems**

Previous studies that have examined the overlap of results returned by different search engines inspired the design of this study as well. Gordon and Pathak (1999) examined overlap in addition to precision and recall and found that approximately 93% of relevant results appeared in the result set on just one search engine. This percentage was fairly stable even at higher document cut-off values. Overlap was higher for results judged to be relevant than for all results.

Spink, Jansen, Blakely, and Koshman (2006) conducted a large-scale study of overlap in search results between four major Web search engines and one metasearch engine. Two large sets of queries were randomly selected from user-entered queries and submitted. They found that the majority of the results returned on the first results page were unique to one search engine, with only 1.1% of results shared across all engines.

Overlap has even been used as a measure of relevance in itself. Can et al. (2004) and Mowshowitz and Kawaguchi (2002) both used the appearance of a URL in the results of multiple search engines as a measure of relevance.

#### **2.4 Query factors and performance**

The studies already described varied in their use of logical operators and other search engine-specific features but these factors have been studied in depth. In a study by Lucas (2002), participants created queries on eight topics for a search engine of their choice. The best-performing query was compared to all others to see how the use of operators related to performance. The number of terms in the query and the percentage of terms matching between queries on the same topic had the highest relevance correlation. Perhaps most interesting, users did not often consider which operators were supported by the engine of their choice, resulting in worse performance.

Eastman and Jansen (2003) sampled queries from a search engine that allowed query operators. They created a new set of duplicate queries with the operators removed and submitted both sets to three search engines. The use of logical operators in queries did not significantly improve IR performance overall, although results varied by engine.

#### **2.5 Other measures of performance**

Precision and recall are not the only measure of IR performance in the literature.

One issue to consider in comparing IR systems on the Web is the size of their indexes. Hawking et al. (2001) found no positive correlation between index size and performance. In a later study, Hawking and Robertson (2003) found that increased index size could improve performance but that study did not examine live search engines indexing the Web as a whole.

Search result ranking can be used to evaluate IR performance. In Vaughan's (2004) study, for example, the correlation between engine ranking and human ranking was used to calculate the "quality of result ranking." Rather than calculating recall, "ability to retrieve top ranked pages" was calculated by comparing the result set with the set of the top 75% sites as ranked by human judges.

Performance has also been measured without explicit relevance judgments. Beg (2005), for example, defined a measure of user satisfaction called the search quality measure (SQM). In this measure, participants did not directly judge relevance. Instead, a number of implicit factors were observed including the order in which the participant clicked on results, the time spent examining documents, and other behaviors.

### **3 Methodology**

#### **3.1 Research design**

In order to better understand the effectiveness of folksonomies at information retrieval, a shootout-style study was conducted between three different kinds of Web IR system: search engines, directories, and folksonomies. The precision and recall performance of the systems was measured and compared and the overlap of results between different IR systems was examined.

Based on the arguments presented in Hawking et al. (2001) and Gordon and Pathak (1999), participants were asked to generate the queries themselves using their own information needs and then judge the relevance of the search results. Just 20 results were collected from each IR system for each query, since many users stop after the first set of results when searching (Silverstein et al, 1999). Previous studies have found strong correlations between different judges (Voorhees 2000), so including additional judges did not seem necessary.

Relevance judgments were made on a binary, yes-or-no basis similar to the methods used in Hawking et al. (2001) and Can et al. (2003). Greisdorf and Spink (2001) found that when the frequency of relevance judgments was plotted on a scale from not relevant to relevant, the highest frequencies were found at the ends. Binary judgments capture the ends of the scale and require less participant effort.

Table 1 shows relevant information about four of the comparable studies mentioned in the literature review along with the present study. The number of participants, queries, and IR systems tested for the present study were within the range of the earlier studies. The binary relevance scale and the precision and recall measures were also comparable.

### **3.1.1 Information needs and queries**

In order to generate a range of information needs, participants were randomly prompted to create queries that fell under the topics listed in Table 2. Topic 1 and topic 2 each address one of the information need types from Hawking et al. (2001), with the rest requiring a selection of relevant documents. The fourth type, the need for an exhaustive collection, would be very difficult to test on the Web and was not studied.

### **3.1.2 Participants**

Participants were drawn from students in the School of Library and Information Science (SLIS) and Information Architecture Knowledge Management (IAKM) graduate programs at Kent State University. Although this population might not completely reflect all Internet users, most studies in the literature have used academics to provide queries and judge relevance.

### 3.1.3 IR Systems in This Study

Studies in the literature review looked at as few as five and as many as 20 search engines. This study examined just eight search systems in order to keep the time required to participants low. Examples of search engines, directories and folksonomies were needed.

Many Web sites employ tagging and allow users to search their folksonomies. Some, such as Flickr (<http://www.flickr.com/>) and YouTube (<http://www.youtube.com>), are restricted to one domain and would not be easy to compare to general search engines or directories. Social bookmarking sites, which allow users to make note of, share, and search Web documents were thought to be most comparable to general search engines and directories.

One difficulty in choosing comparable systems is the fact that Web IR systems vary widely in collection size. Google indexes billions of pages, for example, whereas the Open Directory Project (<http://www.dmoz.com/>) catalog is in the millions. Collection sizes were hard to come by for social bookmarking sites, although there are methods to make estimates (Agarwall, 2006). Because size estimates were not available for all systems considered the impact on IR performance was not examined.

Only social bookmarking systems that allowed searches constrained their own collections were used. Furl (<http://www.furl.net>), for example, includes a search



interface powered by the search engine Looksmart, but users may constrain their search to Furl's folksonomy. Some folksonomies also limit the number of items retrieved by the search interface. These factors eliminated some folksonomies originally considered for study such as StumbleUpon (<http://www.stumbleupon.com>). Furl, Del.icio.us and Reddit's search interfaces functioned much like a traditional search engine, allowing automatic retrieval and parsing of 20 results.

The search systems were chosen for their popularity and large user base, ability to be reliably parsed by an automatic testing interface, and comparability to search engines examined in previous studies. Taking these factors into account, Google, Microsoft Live (formerly MSN), Alta Vista, Yahoo Directory, the Open Directory Project, Del.icio.us, Furl, and Reddit were chosen for this study.

### **3.2 Data collection and procedures**

Participants were able to access the test interface at their convenience. Each participant was asked to create and rate the results for three queries. Figure 1 shows the search page for the first query. The query prompts from Table 2 were assigned randomly to each search and put in the text of the search page. Participants were asked to describe their information need in the text area and then type a query into the "Search" field.

The test interface submitted the query to each of the IR systems in the study and retrieved the results pages. The raw HTML code of the result pages was parsed using

rules tailored for each system. The document title, URL, description, and rank was retrieved for the results from each search and stored in the database. If a description was not available the test interface attempted to follow the document's URL and retrieve a description from the document itself. The contents of the “description” meta tag was used if available. If not, the content of the first paragraph tag was used. Despite these efforts no description could be found for some documents and was left blank.

The search interface collected up to 20 results from each IR system and then randomized their order and identified duplicated results via a case-insensitive comparison of URLs. Results were then presented like common Web search result pages (Figure 2), with the document title displayed as a link to the document followed by the description and URL. Participants could click on the links to see each document in a new browser window, but were not required to explicitly view each document. Results from different IR systems were presented in the same way so that participants did not know which system returned a particular result.

Participants were asked to check a box to the right of each result if they thought the document was relevant to their search and the relevance judgments were saved to a database. The “Submit” button was placed at the very bottom of the results page, so participants had to see all the results before moving on. Participants were prompted to submit a query and rate the results three times, although one user submitted an additional query (most likely after hitting the back button).

### **3.3 Measures of effectiveness**

Precision, relative recall, and retrieval rate were used to measure the effectiveness of the various IR systems. The measures were calculated at a cut-off of 20 and then at the average for cut-offs 1 through 5. The first measure was the maximum number of results collected by the interface while the second both weighed early relevant results more strongly than latter ones and smoothed out effects particular to one cut-off value (Gordon and Pathak, 1999, p. 154).

When calculating relative recall at lower cut-offs, the number of relevant items retrieved by one IR system up to that cut-off was divided by the number of relevant documents retrieved for all engines at the maximum cut-off. Counting just the results up to the lower cutoff yields an unrealistically low estimate for the total number of relevant documents. Cases where no documents were judged to be relevant are excluded from recall measurements. Calculations of precision do not take into account searches that did not retrieve any documents.

## 4 Results

A total of 34 participants completed the experiment, completing 103 searches. Each search was submitted to 8 different systems and the top 20 results were collected. For many searches, one or more of the search systems returned fewer than 20 documents – the total number actually returned was 9266. Out of that number, 2021 were documents returned multiple times for one query, about 22% of the total.

### 4.1 Statistical analysis

A comparison of the queries entered by the participants with those in other shootout studies and Web query log studies can give some sense of this study's external validity (Table 3). The queries entered by the participants had more words and used more operators than generally found in query logs, but fell within the range found in other shootout studies.

The documents listed on the results page were not explicitly ranked but most search engines attempt to return results with the most relevant documents near the top. Eisenberg and Barry (1988) found significant differences in relevance judgments when ordering randomly, from high to low, and from low to high. Their recommendation to present results randomly was followed and very little bias was found in the participants' relevance judgments due to display rank. When considering the relevance of results at each position within each search, Spearman's rho for document position was  $-.076$  (significant at the .01 level), meaning that for any given search the ordering effect was

minor.

Many tests for statistical significance rely on an assumption of a normal distribution and equal variances (Kerlinger and Lee, 2000, p. 414). Since Shapiro-Wilk tests for normality found significant scores for all measures, the null hypothesis that the distributions were symmetric and normal was rejected. The Levene test for homogeneity of variance also yielded significant statistics so equal variances could not be assumed. Testing at different document cut-off values or when cases were grouped by IR system yielded similar results. When data is not normally distributed nonparametric statistical tests are preferable to parametric tests (Kerlinger and Lee, 2000, p. 415). The Kruskal-Wallis test can be used instead of a normal analysis of variance by considering ranks instead of the values from the observations (Hollander and Wolfe, 1973 p. 194). This study largely followed the literature by testing significance with Kruskal-Wallis tests, grouping using Tukey's HSD, and using Spearman's Rank for correlation analysis.

## **4.2 Information retrieval effectiveness**

### **4.2.1 Precision and recall at DCV 20**

Precision, recall, and retrieval rate were calculated and significant differences were found ( $p < 0.01$ ) among the IR systems and system types at all document cut-off values examined.

Precision results for each engine can be seen in Table 4. Google, Yahoo, Alta Vista and Live have been competing against each other for years, so it is not surprising that they performed fairly well. Del.icio.us returned results that were almost as likely to be relevant as those from Live, which shows that a folksonomy can be as precise as a major search engine. Reddit's very low precision score may be due to the lack of tags and reliance on user-submitted titles. Table 5 shows the results for recall. Alta Vista had the highest relative recall with Google and Live following. The rest of the search systems fell much lower. The traditional search engines, with automatic collection gathering, obviously had the advantage.

Tukey's HSD ( $\alpha = 0.05$ ) were calculated to determine the subsets of IR systems that were statistically indistinguishable from each other by precision (Table 4). Three overlapping groups were found with the folksonomies are represented in each group. Del.icio.us actually fell into the highest performance group with search engines and directories. Grouping the search systems on recall performance (Table 5) showed more clear distinctions. All the directories and folksonomies fell within the same group, search engines filling out the two higher performance groups.

When searches were grouped by IR system type, the search engines had both the highest precision and retrieval rate (Table 6). The directories clearly led the folksonomies in precision while recall scores were very similar. The folksonomies were much more likely to retrieve documents than the directories, perhaps because folksonomies impose

less control over submissions. This increase in quantity saw a corresponding decrease in quality.

#### **4.2.2 Precision and Recall at DCV(1-5)**

Figure 3 shows how the precision of each IR system varied as more documents were considered. Most of the systems' precision scores fell as the cut-off increased, with Alta Vista, Google, Live and Yahoo showing particularly pronounced drops early on. This decline indicates that documents were ranked effectively. Del.icio.us, oddly enough, tended to have higher precision as the DCV increased. The rest of the systems showed an early increase in precision after the first result followed by a steady decline.

In Figure 4 a similar chart is shown for recall. The gap between the search engines and all other systems is striking. It is also interesting to note that the search engines continued to increase recall performance at a relatively stable rate at the higher cut-off values. The directories and folksonomies had much lower retrieval rates, so it stands to reason that they would not see recall gains at high document cut-off values where few searches retrieved additional results.

Only the search engines reliably retrieved all 20 documents so a lower DCV was used for further analysis. Table 7 shows the results at DCV 1-5. Alta Vista led both precision and recall while Reddit had the poorest performance in both measures. Compared with the values at DCV 20, all the IR systems except Del.icio.us had a higher

precision score, with higher-ranked documents more likely to be relevant.

Unsurprisingly, recall scores were lower for all IR systems. A Tukey's HSD analysis showed very similar groupings to the values at DCV 20, with precision groups showing a great deal of overlap.

Table 8 shows the breakdown by IR system type. Search engine searches again had the best performance in all three measures. Directory searches had the next best precision and recall, with folksonomy searches taking second place in terms of retrieval rate.

### **4.3 Performance for different information needs**

#### **4.3.1 Specific information needs**

An IR system might be particularly well suited to a particular information need. Since folksonomies have social collection and classification methods they might have an advantage for entertainment and news searches. Search engines may miss novel items between spider visits and directories may be slow to add new items, but users can find and tag items as quickly as news can spread. Conversely folksonomies may perform poorly at factual or exact site searches because they lack both the structure of a directory and the breadth and full-text processing of search engines. Table 9 shows the results by search system type and information need. Differences between groups were significant ( $p < 0.01$ ).



The folksonomies did outperform directories for news searches in both precision and recall but fell well behind the search engines. The directories retrieved very few documents for news searches, none of them relevant. For entertainment searches the folksonomies lost out to the other IR systems but multiple comparisons (Tukey HSD,  $\alpha = .05$ ) showed that performance difference between folksonomies and directories was not significant.

As expected the folksonomies underperformed the other IR systems for factual and exact site queries. These were the worst performing information needs for the folksonomies. Comparing the scores for the different groups in for factual and exact site information needs (Tukey HSD,  $\alpha = .05$ ) showed that the difference between folksonomies and search engines was significant for both precision and recall. Although the folksonomies consistently scored worse than the directories, the difference in scores was not significant.

#### **4.3.2 Categories of information needs**

When searches were subdivided by information need category and IR system type (Table 10), the precision, recall and retrieval performance varied significantly ( $p < 0.01$ ) between groups. The search engines had the highest precision and recall scores for all information need categories, with particularly good recall performance for specific item searches. The directories and folksonomies saw their best precision performance in

searches for a selection of items. The folksonomies seemed to be least suited to searches for a specific item.

The precision scores of the folksonomies and directories fell into the same group when compared by a Tukey HSD. The same was true for recall scores. So although the directories did outperform the folksonomies in precision and recall in each category of information need, there is little confidence in this difference.

#### **4.4 Overlap and relevance of common results**

In many cases a particular document appeared in the results of multiple IR systems in a single search. Reviewing the literature has revealed that documents returned by more than one search engine are more likely to be relevant. Table 11 shows that this is clearly the case in this study as well. The percentage of results deemed relevant by participants for a document returned by two IR systems was almost double the relevance rate for documents returned by just one. Significant differences were found ( $p < 0.01$ ) whether grouped by individual IR system or system type.

Nearly 90%, of documents only appeared in the results of one IR system. Table 12 compares the results of the current study with previous studies discussed in the literature review. Although the similar Gordon and Pathak study (1999) saw less overlap, the current study falls comfortably between those results and the results of Spink et al. (2006). Overlap rates went up when only considering results that were judged to be

relevant, much like Gordon and Pathak's (1999) results. These similarities point to the validity of comparing results from the directories and folksonomies in the same ways previous studies have compared just search engines.

A positive effect was found when a document was returned by more than one type of search system. Documents that appeared in the results of just one type were only about half as likely to be relevant (17.5%) as those that appeared in two types (34.7%), and those that appeared in all three were even more likely to be relevant (42.3%). A breakdown of these results by specific IR system type is shown in Table 13. Documents that were returned by all three types were most likely to be relevant, followed closely by those returned by directories and search engines and those returned by folksonomies and search engines. This is very interesting because it suggests that meta-searching a folksonomy could significantly improve search engine results. Although documents that appeared in both directory and search engine results scored even better, the difference between that set and the folksonomy/search engine set was not statistically significant. On the other hand, incorporating a folksonomy into an existing directory might not significantly improve performance.

#### **4.5 Query characteristics affecting performance**

An examination of the characteristics and content of the participant's queries uncovered some reasons why the folksonomies so often under performed the other IR systems. Only a few query characteristics correlated significantly with recall and

retrieval rate ( $p < .01$ ), all very weakly. For example, the more words in the query the lower the retrieval rate (-.141). The presence of operators had a significant negative correlation with retrieval rate (-.110) especially when excluding Boolean operators (-.151) and correlated significantly with recall (-.103) as well. When just considering folksonomy searches the word count correlation was no longer significant, but the query operator factors had larger negative correlations with retrieval rate and recall. Some of the folksonomies' poor performance may have been due to poor support of query operators.

A few example queries illustrate other causes for poor performance. In one search the participant entered the query “showtimes 45248 borat” and explained their information need as “I want information on the showtimes for the movie Borat.” Google, Live and AltaVista performed very well with this query, with precisions of 100%, 100%, and 62.5% and recall values of 22.2%, 33.3%, and 55.6%, respectively (DCV 20). None of the folksonomies returned any results at all. Removing the zip code from the query produced results in Del.icio.us and Furl and searches for “showtimes” or “borat” produced thousands. The folksonomies seem to have treated all words in the query as absolutely required. Search engines had the advantages of more liberal policies and much larger collections to pick from. What are the chances of another user in a social bookmarking site drilling their way down to a relevant page and then tagging it with the zip code?

The query “Louis Hennepin” was likewise successfully executed by the search engines but had no results in the folksonomies. The participant's information need was biography, bibliography, and criticism of Hennepin for a class paper. Had another Del.icio.us, Furl, or Reddit user needed to write a similar paper, the participant could have benefited from that user's research. In cases like this, however, where the subject of study is specialized, the searcher may be the first user using the site for this subject. Folksonomies were thought to be well suited to timely searches because users could easily add new items of interest to the collection. Conversely, for older topics or items falling outside of mass interest, folksonomies might perform poorly. As these systems grow larger and gain more users, though, this could become less of an issue.

## 5 Discussion

### 5.1 Conclusions

This study demonstrated that folksonomies from social bookmarking sites could be effective tools for IR on the Web and can be studied in that same way that search engines have been studied in the past. The folksonomies' results overlapped with the results from search engines and directories at rates similar to previous IR studies with documents returned by more IR systems more likely to be judged relevant. A document that was returned by both a folksonomy and a search engine was more likely to be relevant than a document that only appeared in the results of search engines.

Significant performance differences were found among the various IR systems and the IR system types. In general the folksonomies had lower precision than directories and search engines, but Del.icio.us did perform better than Yahoo and about as well as Live at DCV 20. In recall, there were few statistically significant differences between the directories and the folksonomies.

The social methods used by folksonomies may be helpful for some information needs when compared to expert-controlled directories, but in general the search engines with their automated collection methods were more effective. Folksonomies performed better for news searches than the directories but the search engines had much higher performance in all categories. The folksonomies did particularly poorly with searches for an exact site and searches with a short, factual answer.

This study discovered a number of possible causes for the differences in performance among the various IR systems and IR system types. The use of query operators was found to have a small but significant negative correlation with recall and retrieval rate. The use of operators correlated more strongly with poor retrieval rates for folksonomies. The folksonomies seemed to handle queries differently than the other IR systems, in some cases requiring all terms in the search string be present in order to return a result.

Despite the fact that the search engines had consistently higher performance, folksonomies show a great deal of promise. First, this study showed that search results from folksonomies could be used to improve the IR performance of search engines. Second, this study demonstrated a number of ways in which existing folksonomies might be able to improve IR performance, for example by better handling query operators and not requiring all terms. Finally, folksonomies are a relatively new technology that may improve as more users are added and techniques are fine-tuned.

## **5.2 Suggestions for future research**

The current study is one of the first studies to examine folksonomies on the Web empirically and barely scratches the surface of this interesting subject. Having completed this study, a number of ways to improve the methodology have presented themselves. Additional information, such as whether or not participants followed links before making relevance judgments, would be very interesting. This study used a relatively small

number of IR systems and more participants and searches would also be beneficial.

The folksonomies differed in precision and recall, and it is likely that there are specific characteristics of folksonomies that would illuminate these differences. Methods of collection building, socialization, tagging, and ranking may differ from one folksonomy to the next. The poor precision and recall performance of Reddit suggests that there might be a very important difference between systems that employ tagging and those that search user-submitted document titles and rankings.

Although this study examined the use of query operators, query length, and similar factors, the methods by which folksonomies can increase the effectiveness of their internal searching functions deserve further study. If a social bookmarking system allows users to assign categories, tag with keywords, change the title, contribute a description, or input comments for an item, which of these fields should be given the most weight? If one document has been given a highly positive average rating by users, while another has a lower rating but better matches the query text, how should they be ranked in the results? Concepts that are well-known in the IR literature like Boolean logic, wildcards, stemming, correcting for spelling errors, and use of synonym rings are not consistently applied (or even applied at all) in folksonomies.

This study did not address any information needs that could only be satisfied by an exhaustive list of documents. Although this is a difficult proposition on a large,



uncontrolled collection like the Web, there are possible ways to address these information needs. One possibility would be to set up an artificial scenario where new documents were made available on the Internet referencing a nonsense word or name that does not currently return any results in any of the IR systems under consideration. It would be important to find ways to ensure that such an artificial situation matches real user information needs and IR tasks.

Further studies with IR systems that cover different domains are also needed. With the growth of blogs and systems dedicated to finding blogs and articles posted to blogs, it would be interesting to perform a similar study with blog search systems such as Google Blog Search and Technorati. A shootout-style study pitting an expert-built classification system against Flickr or YouTube for multimedia would be very interesting.

Search is just one small aspect of the use of social classification, collaborative tagging, and folksonomies. Studies could be done comparing navigation path length, task completion rate and time, and other measures when browsing a conventional, hierarchical directory as opposed to a tag cloud with only “similar-to” or “see-also” relationships. It would also be interesting to study the many other ways in which users might use folksonomies and social bookmarking systems, for example browsing the newest items, looking for random items out of curiosity or for entertainment, organizing their own often-used resources, or socializing with other users.

**Acknowledgments**

The researcher would like to thank Professor David Robins for all his input and advise during the course of this study. Professors Yin Zhang and Marcia Zeng also provided valuable advice and direction.

**Vitae**

P. Jason Morrison is a senior analyst with AT&T and an M.S. in information architecture and knowledge management from Kent State University. His interests include usability, social tagging and folksonomies, the use of social software on the Web.

## 6 References

- Agarwal, Amit. (2006). Search Engine Index Sizes: Google vs Yahoo vs MSN. *Digital Inspiration*. Retrieved August 2, 2006 from <http://labnol.blogspot.com/2006/07/search-engine-index-sizes-google-vs.html>.
- Al-Khalifa, H. S. and Davis, H. C. (2006). FOLKSONOMIES VERSUS AUTOMATIC KEYWORD EXTRACTION: AN EMPIRICAL STUDY. *In Proceedings of IADIS Web Applications and Research 2006 (WAR2006)*. Retrieved January 20, 2007, from <http://eprints.ecs.soton.ac.uk/12547/>.
- Beg, M. (2005). A subjective measure of web search quality. *Information Sciences*, 169(3-4), 365-381.
- Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science and Technology*, 57(1), 69-86.
- Can, Fazli, Nuray, Rabia, and Sevdik, Ayisigi B. (2004). Automatic performance evaluation of Web search engines. *Information Processing and Management*, 40, 495-514.
- Chau, Michael, Fang, Xiao, and Liu Sheng, Olivia R. (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363-1376.
- Chudnov, D, Barnett, J, Prasad, R and Wilcox, M. (2005). Experiments in academic social book marking with Unalog. *Library Hi Tech*.
- Dye, Jessica. (2006, April). Folksonomy: A Game of High-tech (and High-stakes) Tag.

*EContent*, Vol. 29 Issue 3, 38-43.

Eastman, Caroline M. and Jansen, Bernard J. (2003, October). Coverage, relevance, and ranking: The impact of query operators on Web search engine results. *ACM Transactions on Information Systems (TOIS)*, Volume 21, Issue 4, 383 – 411.

Eisenberg, Michael and Barry, Carol (1988). Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39(5), 293-300.

Fichter, Darlene. (2006, May-June). Intranet Applications for Tagging and Folksonomies. *Online*, Vol. 30 Issue 3, 43-45.

Gordon, M and Pathak, P. (1999). Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. *Information Processing and Management*, 35, 141-180. Retrieved August 27, 2006 from <http://www.cindoc.csic.es/cybermetrics/pdf/60.pdf>.

Gordon-Murnane, Laura. (2006, June). Social Bookmarking, Folksonomies, and Web 2.0 Tools. *Searcher*, 14(6), 26-38.

Greisdorf, H., & Spink, A. (2001). Median measure: an approach to IR systems evaluation. *Information Processing and Management*, 37(6), 843-857

Guy, Marieke and Tonkin, Emma. (2006, January). Folksonomies: Tidying Up Tags? *D-Lib Magazine*. Volume 12, No. 1. Retrieved August 27, 2006, from <http://www.dlib.org/dlib/january06/guy/01guy.html>.

- Hawking, D., Craswell, N., Bailey, P., Griffiths, K. (2001). Measuring Search Engine Quality. *Information Retrieval*, 4, 33-59. Kluwer Academic Publishers.
- Hawking, David, and Robertson, Stephen. (2003, January). On Collection Size and Retrieval Effectiveness. *Information Retrieval*, v.6 n.1, 99-105.
- Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *ACM SIGIR 1993*, Pittsburgh, PA.
- Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2), 207-227.
- Jansen, B., and Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1), 248-263.
- Kaser, D. (2003). Do You Yahoo!? *Information Today*, 20(4), 16.
- Kerlinger, Fred N. and Lee, Howard B. (2000). *Foundations of Behavioral Research*. 4th Edition, Wadsworth / Thomson Learning.
- Kobayashi, Mei and Takeda, Koichi. (June 2000). Information retrieval on the web. *ACM Computing Surveys (CSUR)* Volume 32, Issue 2 pp. 144 - 173
- Leighton, H. V. and Srivastava, J. (1999). First 20 precision among world wide web search services (search engines). *Journal of the American Society for Information Science*, 50(10), 882-889.

- Lucas, W. (2002). Form and function: The impact of query term and operator usage on Web search results. *Journal of the American Society for Information Science and Technology*, 53(2), 95-108.
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48, 810–832.
- Mowshowitz, A., and Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing and Management*, 35(2), 141–156.
- Quint, B. (2002). 'Google: (v.)...'. *Searcher*, 10(2), 6.
- Rosenfeld, Louis, and Morville, Peter. (2002). *Information Architecture for the World Wide Web*. Sebastopol, CA: O'Reilly.
- Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12. Previously available as Digital Systems Research Center TR 1998–014 at <http://www.research.digital.com/SRC>.
- Spink, A., Dietmar, W., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226–234.
- Spink, A., Jansen, B., Blakely, C., and Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. *Information Processing and Management*, 42(5), 1379-1391.

- Sullivan, Danny. (2006). ComScore Media Metrix Search Engine Ratings. *Search Engine Watch*. Retrieved March 14, 2007, from <http://searchenginewatch.com/showPage.html?page=2156431>.
- Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, 40(4), 677-691.
- Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5), 697-716.

### **Figure legends**

Figure 1. Search screen from test interface.

Figure 2. Results screen from test interface.

Figure 3. IR system precision at cut-offs 1-20.

Figure 4. IR system recall at cutoffs 1-20.



Table 1

*Characteristics of Previous Studies*

	<b>Leighton and Srivastava (1997)</b>	<b>Gordon and Pathak (1999)</b>	<b>Hawking et al (2001)</b>	<b>Can et al (2003)</b>	<b>The Present Study</b>
<b>Information needs provided by</b>	Library reference desk, other studies	Faculty members	Queries from web logs	Students and professors	Graduate students
<b>Queries created by</b>	The researchers	Skilled searchers	Queries from web logs	Same	Same
<b>Relevance judged by</b>	Same researchers (by consensus)	Same faculty members	Research Assistants	Same	Same
<b>Participants</b>	2	33	6	19	34
<b>Queries per participant</b>	15	1	9	1 to 2	3 to 4
<b>Total queries</b>	15	33	54	25	103
<b>Engines tested</b>	5	8	20	8	8
<b>Results evaluated per engine</b>	20	20 <sup>a</sup>	20	20	20
<b>Total results evaluated per evaluator</b>	1500	160	3600	160 or 320	About 160
<b>Relevancy scale</b>	4 categories <sup>b</sup>	4-point scale	Binary	Binary	Binary
<b>Precision measures</b>	P(20), weighted groups by rank	P(1-5), P(1-10), P(5-10), P(15-20)	P(1), P(1-5), P(5) P(20)	P(10), P(20) <sup>c</sup>	P(20), P(1-5)
<b>Recall measures:</b>	none	Relative recall; R(15-20), R(15-25), R(40-60), R(90-110), R(180-200)	none	Relative recall: R(10), R(20) <sup>c</sup>	Relative recall: R(20), R(1-5)

a. Relevancy of items in the top 20 results in each engine was used to evaluate the top 200 results in the other engines.

b. Converted to binary for analysis

c. P(1-10), P(1-20), R(1-10), R(1-20) used to compare human and automatic evaluation but not to compare engines

Table 2

*Information Need Query Prompts*

	<b>Topic</b>	<b>Query prompt</b>	<b>Information need category</b> (Hawking, et. al., 2001, p. 55)
1.	Factual information	“Think of a factual question that you know the answer to.”	A. Short factual statements that answer a question
2.	Exact site	“Find the home page of your current (or most recent) employer.”	B. Specific document or web site
3.	Academic research	“Think of a topic you are doing research on, or have researched in the past.”	C. A selection of documents that pertain to an area of interest
4.	News	“Think of something that was in the news recently.”	
5.	General	“Think of a topic you have searched for in the past out of curiosity.”	
6.	Entertainment	“Think of something you might search for if you were looking for entertainment.”	
	-	-	D. Every document in the collection that matches the information need

Table 3

*Query Characteristics of Search Engine Log Analysis Studies*

	Query Log Studies					Shootout Studies			
	S et al (1999)	Jansen et al (2000)	Spink et al (2001)	Chau et al (2005)	L & S (1997)	H et al (2001)	Lucas (2002)	Can et al (2003)	Current Study
<b>Words per query</b>	2.35	2.21	2.4	2.25	4.9	5.9	2.76	3.80	4.10
<b>Operator use rate</b>	20.4%	<10% x	4% x 5% y	34.4%	-	-	41.4%	-	31.1%
<b>Operators per query</b>	0.41	-	-	-	0	0	0.63	1.68	0.72

S et al - Silverstein et al. (1999); L&S - Leighton and Srivastava (1997); G&P - Gordon and Pathak (1999); H et al - Hawking et al (2001); E&J - Eastman and Jansen (2003); V - Vaughan (2004)

x - Boolean operators only

y - Plus (+) operator only

Table 4

*Search Systems Grouped by Precision(20) Tukey HSD*

IR System	N	Subset for alpha = .05		
		1	2	3
Reddit	62	.0413		
Furl	75	.0938	.0938	
Open Directory	37	.1723	.1723	.1723
Del.icio.us	43		.2109	.2109
Live	103			.2354
Alta Vista	102			.2630
Yahoo Directory	36			.2706
Google	93			.2860
<b>Sig.</b>		.091	.188	.218

Table 5

*Search Systems Grouped by Recall(20) Tukey HSD*

IR System	N	Subset for alpha = .05		
		1	2	3
Open Directory	98	.0239		
Del.icio.us	98	.0412		
Reddit	98	.0420		
Furl	98	.0450		
Yahoo Directory	98	.0638		
Live	98		.3413	
Google	98		.3517	
Alta Vista	98			.4312
<b>Sig.</b>		.765	1.000	1.000

Table 6

*Precision(20) and Recall(20) by Search System Type*

<b>IR System Type</b>	<b>Precision</b>	<b>Recall</b>	<b>Retrieval Rate</b>
<b>Directory</b>	0.2208	0.0438	0.1757
<b>Folksonomy</b>	0.1037	0.0427	0.4278
<b>Search Engine</b>	0.2607	0.3747	0.9544

Table 7

*Search System Average Precision(1-5) Recall(1-5) and Retrieval Rate(1-5)*

<b>IR system</b>	<b>Avg precision</b>	<b>Avg recall</b>	<b>Avg retrieval rate</b>
	<b>Value</b>	<b>Value</b>	
<b>Open Directory</b>	0.2045	0.0120	0.2970
<b>Yahoo Directory</b>	0.3266	0.0245	0.2828
<b>Del.icio.us</b>	0.1988	0.0137	0.3349
<b>Furl</b>	0.1210	0.0157	0.6611
<b>Reddit</b>	0.0682	0.0062	0.5911
<b>Google</b>	0.4327	0.1060	0.9001
<b>Live</b>	0.3889	0.1305	0.9987
<b>Alta Vista</b>	0.4381	0.1517	0.9903
<b>Overall</b>	0.3015	0.0576	0.6320

Table 8

*Average Precision(1-5) Recall(1-5) and Retrieval Rate(1-5) by Search System Type*

IR system type	Avg precision	Avg recall	Avg retrieval rate
Directory	0.2647	0.0183	0.2899
Folksonomy	0.1214	0.0119	0.5290
Search Engine	0.4194	0.1294	0.9631

Table 9

*Precision(1-5) Recall(1-5) and Retrieval Rate(1-5) by Information Need and Search System Type*

Information need	IR system type	Avg precision	Avg recall	Avg retrieval rate
Research	Directory	0.2104	0.0142	0.2604
	Folksonomy	0.1760	0.0104	0.4378
	Search Engine	0.4626	0.0910	0.9474
	Overall	0.3629	0.0416	0.5845
News	Directory	0.0000	0.0000	0.0604
	Folksonomy	0.1547	0.0168	0.5734
	Search Engine	0.3724	0.0969	0.9616
	Overall	0.2752	0.0427	0.5930
General	Directory	0.2272	0.0215	0.4057
	Folksonomy	0.1659	0.0133	0.6295
	Search Engine	0.5371	0.0872	1.0000
	Overall	0.3805	0.0456	0.7350
Factual	Directory	0.2186	0.0005	0.2104
	Folksonomy	0.0601	0.0071	0.6013
	Search Engine	0.4405	0.0952	0.9524
	Overall	0.2741	0.0407	0.6700
Entertainment	Directory	0.2022	0.0212	0.2111
	Folksonomy	0.1362	0.0163	0.4835
	Search Engine	0.2991	0.1276	0.9259
	Overall	0.2464	0.0593	0.5888
Exact Site	Directory	0.1022	0.0222	0.2225
	Folksonomy	0.0272	0.0084	0.4475
	Search Engine	0.3536	0.2682	0.9683
	Overall	0.2338	0.112	0.6140

Table 10

*Precision(1-5) Recall(1-5) and Retrieval Rate(1-5) by Information Need Category and Search System Type*

Information need category	IR system type	Avg precision	Avg recall	Avg retrieval rate
<b>Short Factual Answer</b>	<b>Directory</b>	0.2186	0.0095	0.3494
	<b>Folksonomy</b>	0.0601	0.0071	0.6013
	<b>Search Engine</b>	0.4405	0.0952	0.9524
	<b>Overall</b>	0.2741	0.0407	0.6700
<b>Specific Item</b>	<b>Directory</b>	0.1033	0.0332	0.3325
	<b>Folksonomy</b>	0.0272	0.0084	0.4475
	<b>Search Engine</b>	0.3536	0.2682	0.9683
	<b>Overall</b>	0.2338	0.1120	0.6140
<b>Selection of Relevant Items</b>	<b>Directory</b>	0.3048	0.0158	0.2645
	<b>Folksonomy</b>	0.1608	0.0139	0.5393
	<b>Search Engine</b>	0.4355	0.0962	0.9636
	<b>Overall</b>	0.3282	0.0453	0.6297

Table 11

*Relevance of URL by IR System Overlap*

<b>Number of engines returning the URL</b>	<b>Number of unique results</b>	<b>Relevance rate</b>
1	7223	.1631
2	617	.2950
3	176	.3580
4	43	.4884
5	15	.4667
6	2	.0000
Total	8076	.1797

Table 12

*Overlap Rates in Previous Studies*

<b>Number of engines</b>	<b>Gordon and Pathak (1999), @20</b>		<b>Spink et al (2006)</b>	<b>Current Study</b>	
	All results	Relevant	All results	All results	Relevant
1	96.8%	93.4%	84.9%	89.44%	81.18%
2	2.9%	6.0%	11.4%	7.64%	12.54%
3	0.3%	0.4%	2.6%	2.18%	4.34%
4	0.1%	0.3%	1.1%	0.53%	1.45%
5	0.0%	0.0%		0.19%	0.48%
6	-	-		0.02%	0.00%
7	-	-		-	-
8	-	-		-	-



Table 13

*Relevance by Search System Type Permutation*

Engine types returning same URL			N	Relevance rate
Directory	Folksonomy	Search Engine		
no	no	yes	4801	.2350
no	yes	no	2484	.0676
yes	no	no	592	.1419
no	yes	yes	94	.3191
yes	no	yes	67	.4179
yes	yes	no	12	.1667
yes	yes	yes	26	.4231
Total			8076	.1797

# Step 2: Search Exercise #1

## Create a Query

Think of something you might search for if you were looking for entertainment. Describe the information you are looking for in the "Information Need" field below. Next, type a query into the "Search" field. Try to enter the type of search terms or phrases you would normally use if you were using a search engine on the web.

Information Need :

Search :

Submit :

*(please note: searches make take as long as 3 minutes to return results).*

# Step 2: Search Exercise #1

## Judge the Search Results

Please be patient while the search is run. Your browser stops loading and no results are displayed, please reload the page to try again.

Performing search ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...  
Searching ...

You searched for **Colbert Report**. Below are your results.

Please look through the list below. For each link, check the check box if you think the web site is relevent to your search. Please judge the web sites the same way you would if you were using a normal web search engine like Google.

If you click on any of the links the web site will pop up in a new window - close the window to return to the list.

[An inside look: The production of the Stephen Colbert SAHD Report](#)

<http://www.angelfire.com/zine2/athomedad/index.blog/1255487/an-inside-look-at-the-production-of-the->

Relevant?

[Guest Goods » Blog Archive » \[Colbert Report\] Toby Keith](#)

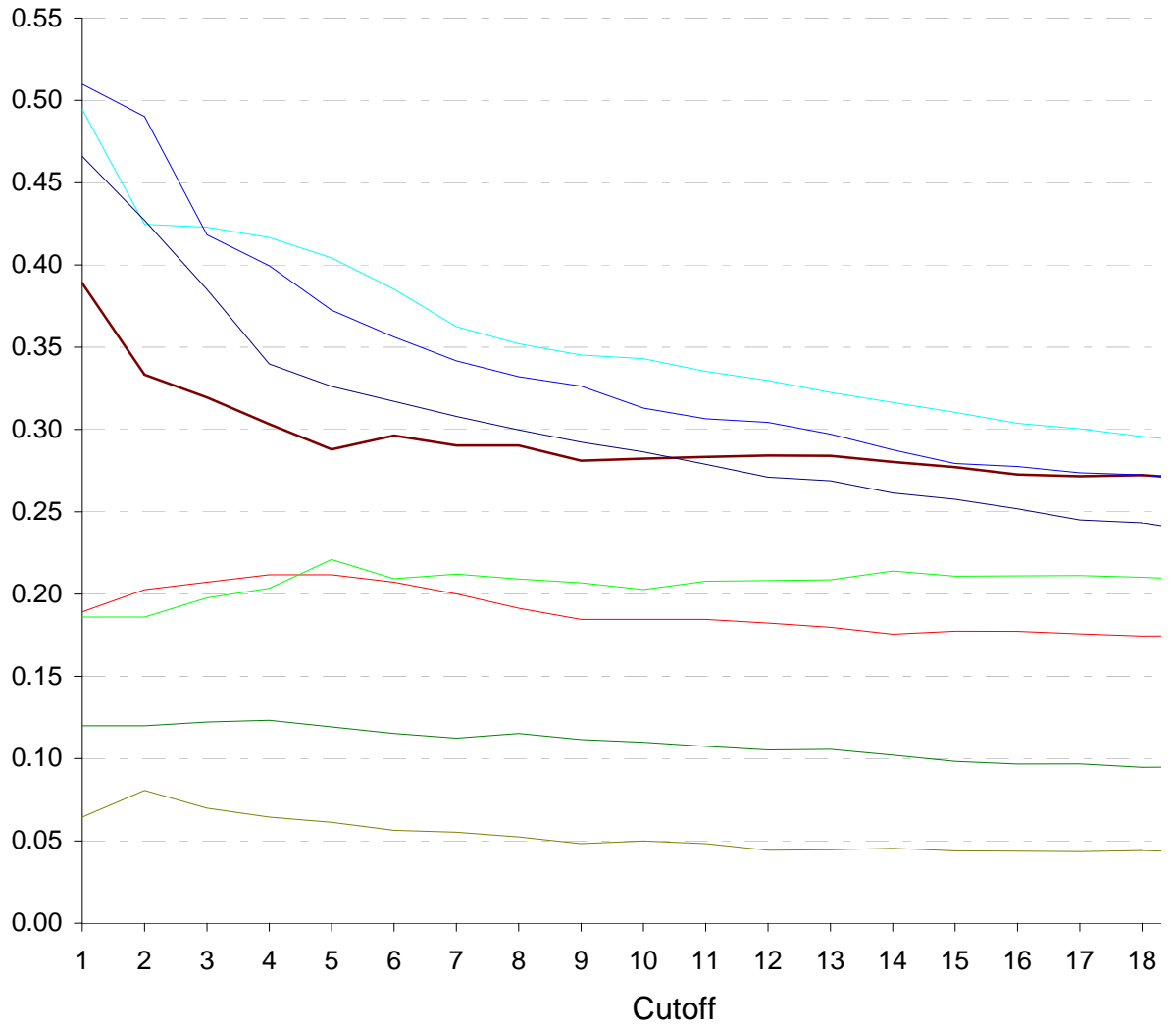
by Toby Keith Tagged with: Colbert Report , The Colbert Report , Toby Keith , Keith This entry Guest Goods » Blog Archive » [  
<http://guestgoods.com/2006/09/12/colbert-report-toby-keith/>

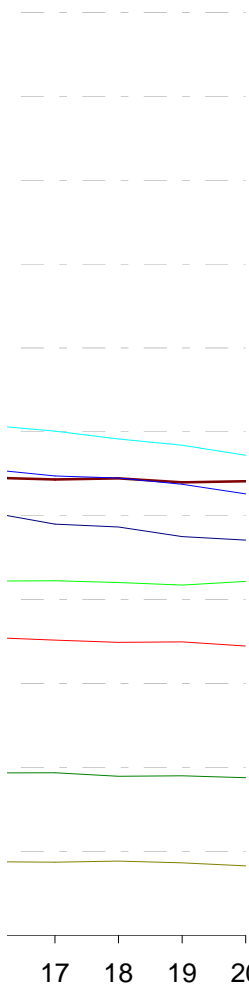
Relevant?

[IFILM - Video, User Video, Movies, Trailers, Music and Viral Videos](#)

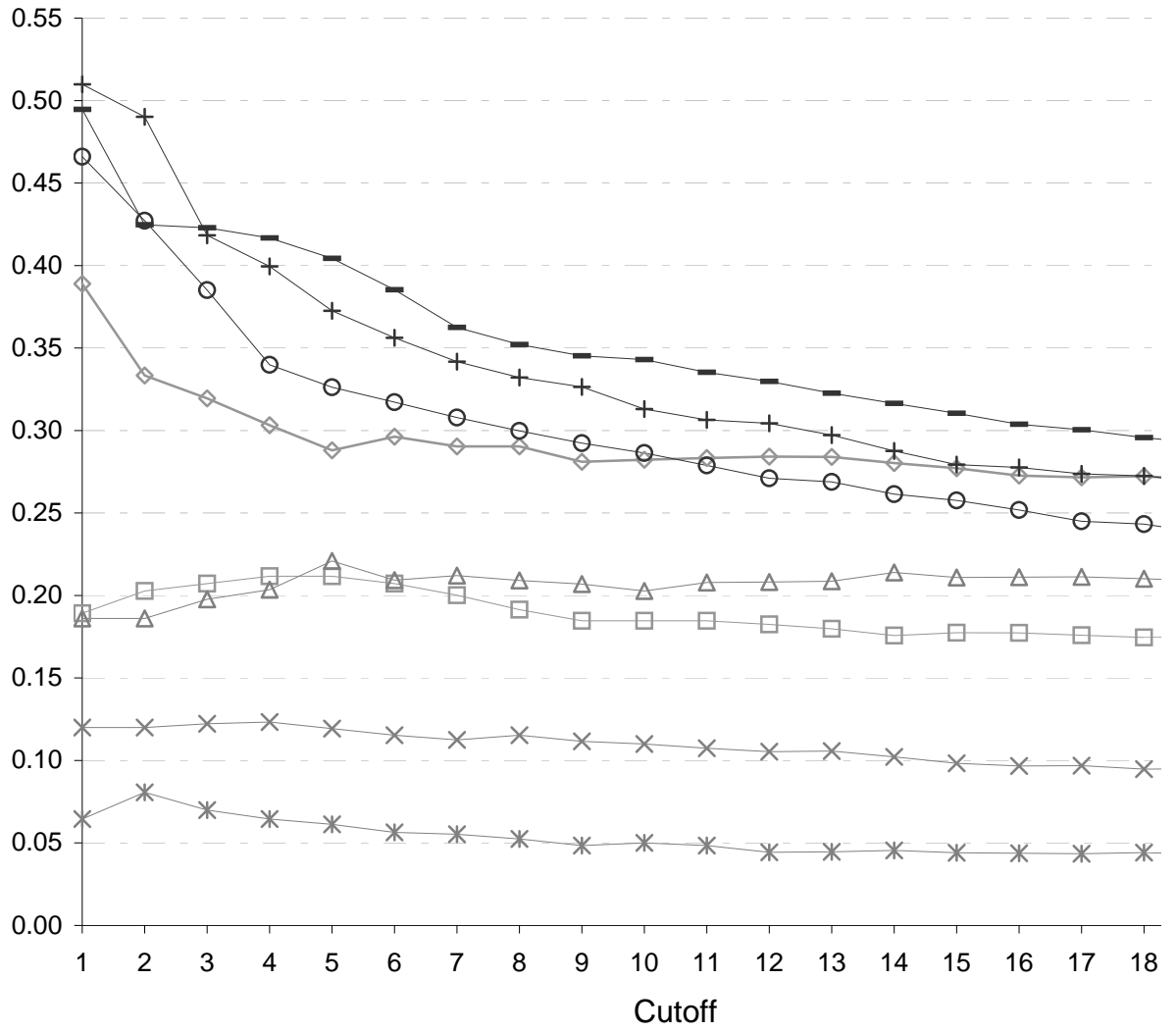
Program new! The Colbert Report Comedy Central » Pimp My Ride Black History Month • BMW Vodcast • Colbert  
<http://ifilm.com/>

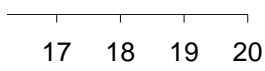
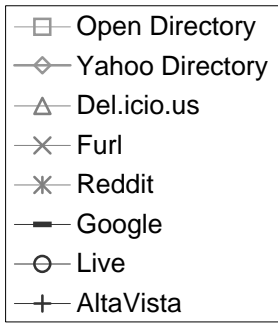
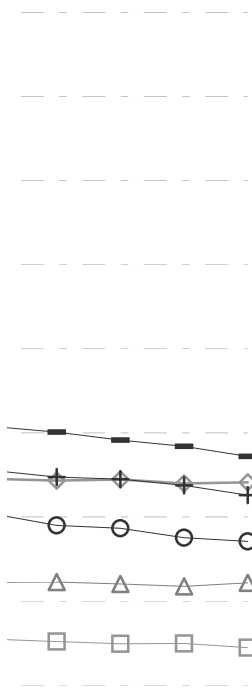
Relevant?





- Open Directory
- Yahoo Directory
- Del.icio.us
- Furl
- Reddit
- Google
- Live
- AltaVista





data

cutoff	engine_name	engine_precision		Open Director	Yahoo Directo	Del.icio.us
1	Open Directory	0.18918919	1	0.18918919	0.38888889	0.18604651
2	Open Directory	0.2027027	2	0.20270270	0.33333333	0.18604651
3	Open Directory	0.20720541	3	0.20720541	0.31944444	0.19766977
4	Open Directory	0.21171081	4	0.21171081	0.30324167	0.20348837
5	Open Directory	0.21171081	5	0.21171081	0.28796389	0.22093023
6	Open Directory	0.2072027	6	0.20720270	0.29629722	0.20930465
7	Open Directory	0.20012703	7	0.20012703	0.29034722	0.21207442
8	Open Directory	0.19144054	8	0.19144054	0.29034444	0.20916279
9	Open Directory	0.18468108	9	0.18468108	0.28108056	0.20690000
10	Open Directory	0.18468378	10	0.18468378	0.28231944	0.20276744
11	Open Directory	0.18468108	11	0.18468108	0.28332500	0.20783953
12	Open Directory	0.18243784	12	0.18243784	0.28416667	0.20819302
13	Open Directory	0.17982703	13	0.17982703	0.28399444	0.20863953
14	Open Directory	0.17567838	14	0.17567838	0.28032778	0.21400698
15	Open Directory	0.17747568	15	0.17747568	0.27715556	0.21090930
16	Open Directory	0.17736486	16	0.17736486	0.27264167	0.21110000
17	Open Directory	0.17587297	17	0.17587297	0.27161944	0.21126977
18	Open Directory	0.17455135	18	0.17455135	0.27226111	0.21020930
19	Open Directory	0.17478919	19	0.17478919	0.26990000	0.20870930
20	Open Directory	0.1722973	20	0.17229730	0.27055833	0.21085349
1	Yahoo Directory	0.38888889				
2	Yahoo Directory	0.33333333				
3	Yahoo Directory	0.31944444				
4	Yahoo Directory	0.30324167				
5	Yahoo Directory	0.28796389				
6	Yahoo Directory	0.29629722				
7	Yahoo Directory	0.29034722				
8	Yahoo Directory	0.29034444				
9	Yahoo Directory	0.28108056				
10	Yahoo Directory	0.28231944				
11	Yahoo Directory	0.283325				
12	Yahoo Directory	0.28416667				
13	Yahoo Directory	0.28399444				
14	Yahoo Directory	0.28032778				
15	Yahoo Directory	0.27715556				
16	Yahoo Directory	0.27264167				
17	Yahoo Directory	0.27161944				
18	Yahoo Directory	0.27226111				
19	Yahoo Directory	0.2699				
20	Yahoo Directory	0.27055833				
1	Del.icio.us	0.18604651				
2	Del.icio.us	0.18604651				
3	Del.icio.us	0.19766977				
4	Del.icio.us	0.20348837				
5	Del.icio.us	0.22093023				
6	Del.icio.us	0.20930465				
7	Del.icio.us	0.21207442				
8	Del.icio.us	0.20916279				
9	Del.icio.us	0.2069				



data

10 Del.icio.us	0.20276744
11 Del.icio.us	0.20783953
12 Del.icio.us	0.20819302
13 Del.icio.us	0.20863953
14 Del.icio.us	0.21400698
15 Del.icio.us	0.2109093
16 Del.icio.us	0.2111
17 Del.icio.us	0.21126977
18 Del.icio.us	0.2102093
19 Del.icio.us	0.2087093
20 Del.icio.us	0.21085349
1 Furl	0.12
2 Furl	0.12
3 Furl	0.12221733
4 Furl	0.12333333
5 Furl	0.11933333
6 Furl	0.11533733
7 Furl	0.11248267
8 Furl	0.11533333
9 Furl	0.111624
10 Furl	0.11
11 Furl	0.10744933
12 Furl	0.105332
13 Furl	0.105668
14 Furl	0.10215867
15 Furl	0.09828533
16 Furl	0.09667333
17 Furl	0.096816
18 Furl	0.094732
19 Furl	0.094964
20 Furl	0.09384
1 Reddit	0.06451613
2 Reddit	0.08064516
3 Reddit	0.06989194
4 Reddit	0.06451613
5 Reddit	0.06129032
6 Reddit	0.05645161
7 Reddit	0.05530323
8 Reddit	0.05241935
9 Reddit	0.04838387
10 Reddit	0.05
11 Reddit	0.04838387
12 Reddit	0.04435484
13 Reddit	0.04466129
14 Reddit	0.04546452
15 Reddit	0.04400323
16 Reddit	0.04373387
17 Reddit	0.04349194
18 Reddit	0.04418548
19 Reddit	0.04309677
20 Reddit	0.04131452

data

1 Google	0.49462366
2 Google	0.42473118
3 Google	0.42293763
4 Google	0.41666667
5 Google	0.40430108
6 Google	0.3853043
7 Google	0.36252473
8 Google	0.35215054
9 Google	0.34527204
10 Google	0.34301075
11 Google	0.33528065
12 Google	0.32974516
13 Google	0.32257957
14 Google	0.31643656
15 Google	0.31039892
16 Google	0.30376344
17 Google	0.30043656
18 Google	0.29570215
19 Google	0.29202043
20 Google	0.28602151
1 Live	0.46601942
2 Live	0.42718447
3 Live	0.38511165
4 Live	0.33980583
5 Live	0.32621359
6 Live	0.31715049
7 Live	0.30791165
8 Live	0.29975728
9 Live	0.29233204
10 Live	0.28640777
11 Live	0.27889709
12 Live	0.27103204
13 Live	0.26884854
14 Live	0.26144272
15 Live	0.25760388
16 Live	0.25182039
17 Live	0.2449932
18 Live	0.24326019
19 Live	0.23761262
20 Live	0.23543689
1 AltaVista	0.50980392
2 AltaVista	0.49019608
3 AltaVista	0.41829706
4 AltaVista	0.3995098
5 AltaVista	0.37254902
6 AltaVista	0.3562098
7 AltaVista	0.3417451
8 AltaVista	0.33210784
9 AltaVista	0.32638039
10 AltaVista	0.3129902
11 AltaVista	0.30647647

data

12 AltaVista	0.30432941
13 AltaVista	0.29721863
14 AltaVista	0.28764314
15 AltaVista	0.27933137
16 AltaVista	0.27757353
17 AltaVista	0.27370686
18 AltaVista	0.27247157
19 AltaVista	0.26877157
20 AltaVista	0.2629902

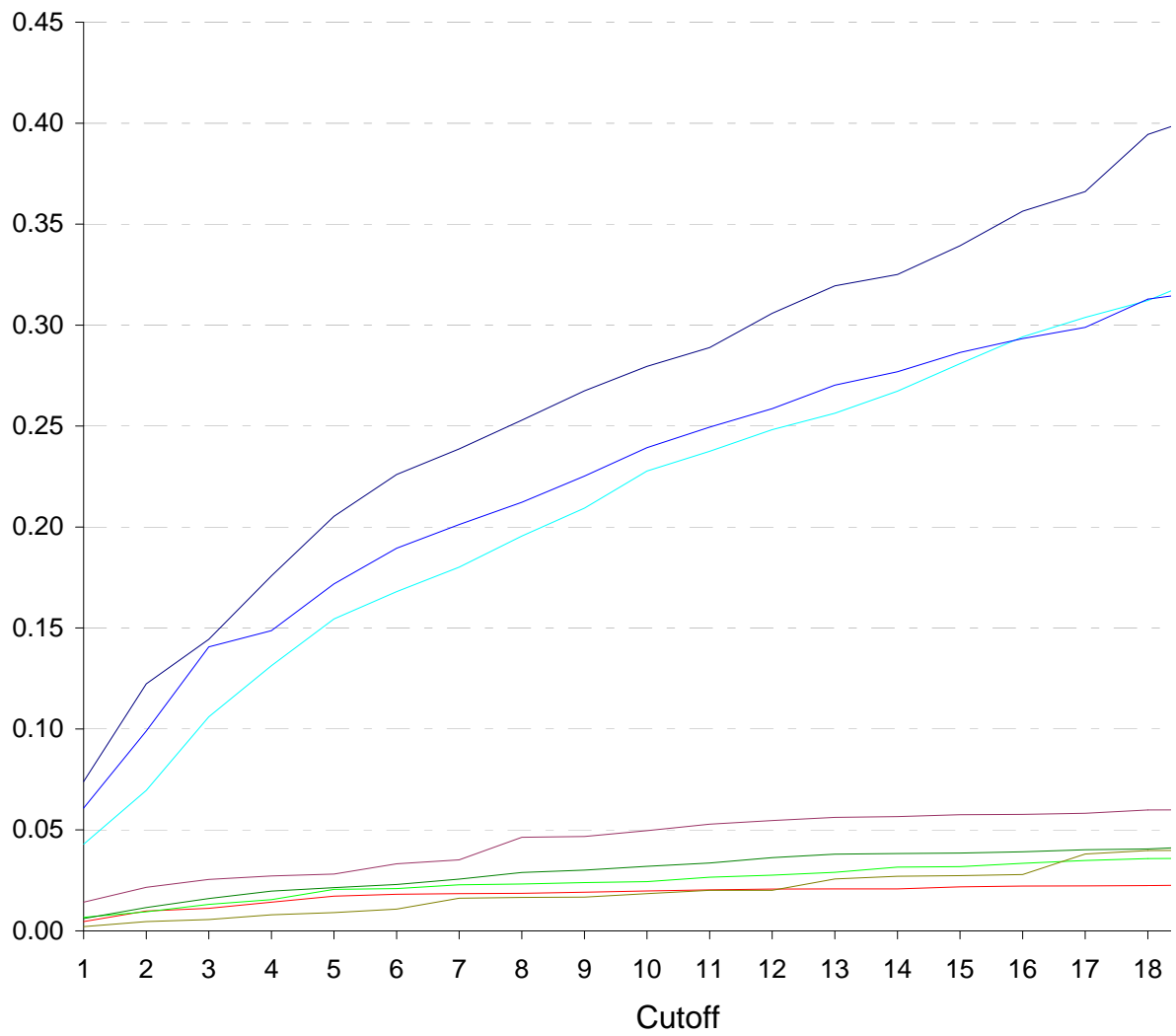
data

Furl	Reddit	Google	Live	AltaVista
0.12000000	0.06451613	0.49462366	0.46601942	0.50980392
0.12000000	0.08064516	0.42473118	0.42718447	0.49019608
0.12221733	0.06989194	0.42293763	0.38511165	0.41829706
0.12333333	0.06451613	0.41666667	0.33980583	0.39950980
0.11933333	0.06129032	0.40430108	0.32621359	0.37254902
0.11533733	0.05645161	0.38530430	0.31715049	0.35620980
0.11248267	0.05530323	0.36252473	0.30791165	0.34174510
0.11533333	0.05241935	0.35215054	0.29975728	0.33210784
0.11162400	0.04838387	0.34527204	0.29233204	0.32638039
0.11000000	0.05000000	0.34301075	0.28640777	0.31299020
0.10744933	0.04838387	0.33528065	0.27889709	0.30647647
0.10533200	0.04435484	0.32974516	0.27103204	0.30432941
0.10566800	0.04466129	0.32257957	0.26884854	0.29721863
0.10215867	0.04546452	0.31643656	0.26144272	0.28764314
0.09828533	0.04400323	0.31039892	0.25760388	0.27933137
0.09667333	0.04373387	0.30376344	0.25182039	0.27757353
0.09681600	0.04349194	0.30043656	0.24499320	0.27370686
0.09473200	0.04418548	0.29570215	0.24326019	0.27247157
0.09496400	0.04309677	0.29202043	0.23761262	0.26877157
0.09384000	0.04131452	0.28602151	0.23543689	0.26299020

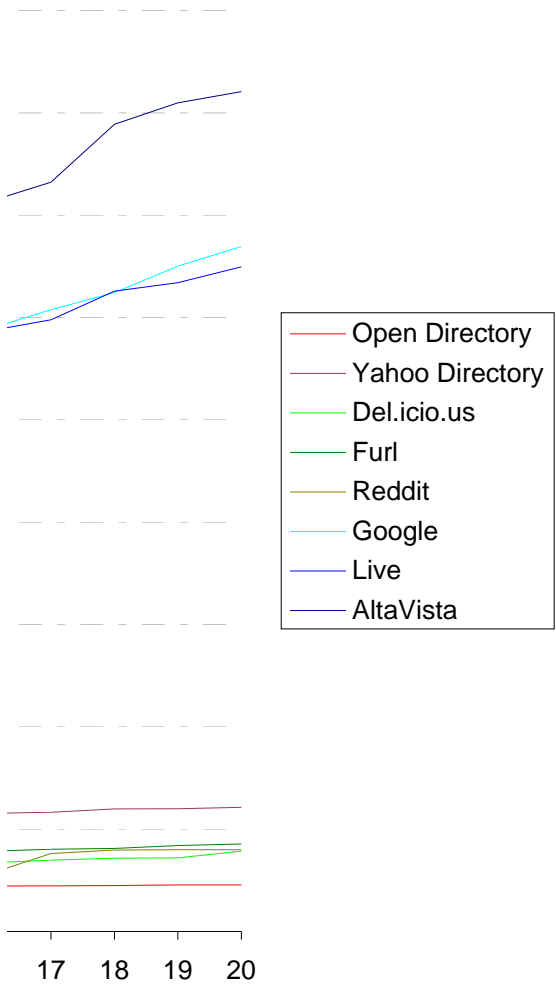
data

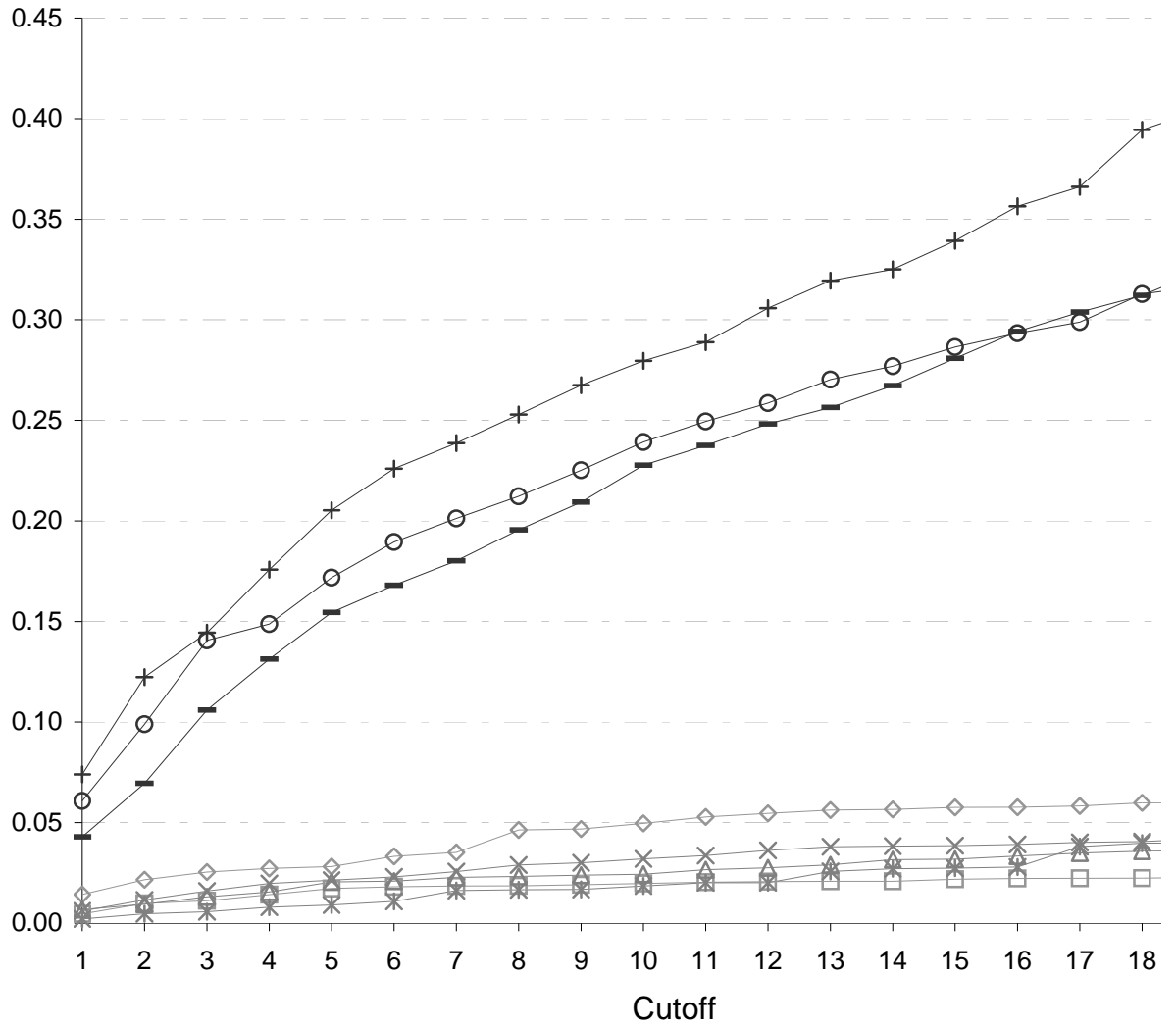
data

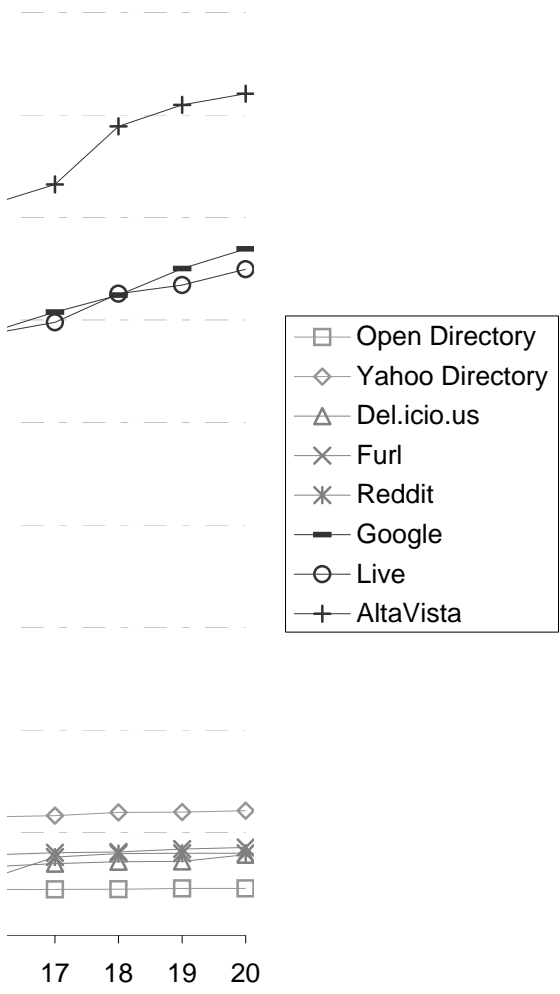
data











data

cutoff	engine_name	engine_relativ	engine_relative_recall_20	
	1 AltaVista	0.29271262	0.07396019	Open Director
	2 AltaVista	0.34281553	0.1223466	1 0.00461456
	3 AltaVista	0.35873786	0.14440777	2 0.00983495
	4 AltaVista	0.36493301	0.17582524	3 0.01117573
	5 AltaVista	0.36959806	0.2052932	4 0.01423786
	6 AltaVista	0.39351456	0.22600388	5 0.0172165
	7 AltaVista	0.39456505	0.23867087	6 0.01813107
	8 AltaVista	0.39681553	0.25289029	7 0.01847379
	9 AltaVista	0.39240971	0.26747379	8 0.01858058
	10 AltaVista	0.38779126	0.27955146	9 0.01915146
	11 AltaVista	0.38452524	0.28889903	10 0.01974369
	12 AltaVista	0.38860388	0.3058	11 0.02033689
	13 AltaVista	0.39100971	0.31944078	12 0.02069223
	14 AltaVista	0.38733301	0.32510097	13 0.02079903
	15 AltaVista	0.39813981	0.33929223	14 0.02079903
	16 AltaVista	0.40179612	0.3564233	15 0.02184757
	17 AltaVista	0.4037165	0.36613689	16 0.02220291
	18 AltaVista	0.41517282	0.39442136	17 0.02230971
	19 AltaVista	0.41274272	0.40489515	18 0.0224165
	20 AltaVista	0.41033495	0.41033495	19 0.02277184
	1 Del.icio.us	0.03721553	0.00675922	20 0.02277184
	2 Del.icio.us	0.03549029	0.00947379	
	3 Del.icio.us	0.03608447	0.01312816	
	4 Del.icio.us	0.03703883	0.01550097	
	5 Del.icio.us	0.04216408	0.02052524	
	6 Del.icio.us	0.0392466	0.02094757	
	7 Del.icio.us	0.04108835	0.02281068	
	8 Del.icio.us	0.03986311	0.02325825	
	9 Del.icio.us	0.03868447	0.02389126	
	10 Del.icio.us	0.03784854	0.0243767	
	11 Del.icio.us	0.03952718	0.02658155	
	12 Del.icio.us	0.03787864	0.02756311	
	13 Del.icio.us	0.0382466	0.02900485	
	14 Del.icio.us	0.03998835	0.03159417	
	15 Del.icio.us	0.03880388	0.03183107	
	16 Del.icio.us	0.03845922	0.03343204	
	17 Del.icio.us	0.03844078	0.03487379	
	18 Del.icio.us	0.03723981	0.03577864	
	19 Del.icio.us	0.03686796	0.03598932	
	20 Del.icio.us	0.03923786	0.03923786	
	1 Furl	0.03155146	0.00611068	
	2 Furl	0.05129029	0.01158058	
	3 Furl	0.04928447	0.01600388	
	4 Furl	0.05528641	0.0197068	
	5 Furl	0.05250874	0.02140388	
	6 Furl	0.05173592	0.02299612	
	7 Furl	0.05077087	0.02566408	
	8 Furl	0.05299223	0.02891553	
	9 Furl	0.05217282	0.0300699	
	10 Furl	0.05128835	0.0319767	

data

11 Furl	0.05167282	0.03362427
12 Furl	0.0509	0.03623981
13 Furl	0.05087379	0.03798447
14 Furl	0.04985728	0.03831748
15 Furl	0.0474165	0.0385534
16 Furl	0.04610097	0.03913398
17 Furl	0.0453835	0.04018447
18 Furl	0.0431068	0.04054078
19 Furl	0.04327184	0.04198932
20 Furl	0.0427932	0.0427932
1 Google	0.24093398	0.04289126
2 Google	0.23116602	0.06956019
3 Google	0.26804272	0.1059835
4 Google	0.30990583	0.13132524
5 Google	0.32568058	0.15449515
6 Google	0.3232466	0.16802621
7 Google	0.32025922	0.18013495
8 Google	0.32001068	0.19552913
9 Google	0.3201534	0.20934272
10 Google	0.33213786	0.22771942
11 Google	0.33117476	0.23752718
12 Google	0.32934854	0.24824369
13 Google	0.3282	0.25638155
14 Google	0.33223592	0.2672767
15 Google	0.3416068	0.28086214
16 Google	0.33944175	0.29419806
17 Google	0.34246699	0.30383786
18 Google	0.33226602	0.31214563
19 Google	0.33333592	0.32505825
20 Google	0.33466019	0.33466019
1 Live	0.27006019	0.06081942
2 Live	0.28706214	0.09896408
3 Live	0.30705049	0.14062913
4 Live	0.29097184	0.14869612
5 Live	0.30598447	0.17185728
6 Live	0.3185	0.18952233
7 Live	0.32498835	0.20121553
8 Live	0.32430291	0.21227379
9 Live	0.32487476	0.22521845
10 Live	0.32612816	0.23927573
11 Live	0.32575437	0.24946699
12 Live	0.32556893	0.25862427
13 Live	0.3285301	0.27027184
14 Live	0.32709515	0.27691456
15 Live	0.32810583	0.28647767
16 Live	0.32786796	0.29331456
17 Live	0.32479126	0.29885825
18 Live	0.3294068	0.31285922
19 Live	0.32345243	0.31703301
20 Live	0.32472427	0.32472427
1 Open Director	0.03462718	0.00461456

data

2	Open Director	0.03822621	0.00983495
3	Open Director	0.0324301	0.01117573
4	Open Director	0.03322621	0.01423786
5	Open Director	0.03337864	0.0172165
6	Open Director	0.03237864	0.01813107
7	Open Director	0.03119806	0.01847379
8	Open Director	0.02967864	0.01858058
9	Open Director	0.02945728	0.01915146
10	Open Director	0.02925825	0.01974369
11	Open Director	0.02864563	0.02033689
12	Open Director	0.02790291	0.02069223
13	Open Director	0.02734369	0.02079903
14	Open Director	0.02686408	0.02079903
15	Open Director	0.02753592	0.02184757
16	Open Director	0.02716117	0.02220291
17	Open Director	0.02655825	0.02230971
18	Open Director	0.02430485	0.0224165
19	Open Director	0.02413883	0.02277184
20	Open Director	0.02277184	0.02277184
1	Reddit	0.0226534	0.00212621
2	Reddit	0.03497767	0.00464951
3	Reddit	0.02783883	0.00565922
4	Reddit	0.02858447	0.00796796
5	Reddit	0.02851262	0.00905243
6	Reddit	0.02424272	0.0108165
7	Reddit	0.03424272	0.0162068
8	Reddit	0.0334767	0.01660291
9	Reddit	0.03233107	0.01675049
10	Reddit	0.03082816	0.01848835
11	Reddit	0.03164563	0.02012427
12	Reddit	0.03059515	0.02012427
13	Reddit	0.03023689	0.02573398
14	Reddit	0.03103689	0.02707573
15	Reddit	0.03070971	0.0273534
16	Reddit	0.03090874	0.02794175
17	Reddit	0.0403165	0.03809223
18	Reddit	0.04166699	0.0398165
19	Reddit	0.0408699	0.03996408
20	Reddit	0.03996408	0.03996408
1	Yahoo Directo	0.0864068	0.0142466
2	Yahoo Directo	0.07407864	0.0216068
3	Yahoo Directo	0.06859223	0.02546311
4	Yahoo Directo	0.06942621	0.02722039
5	Yahoo Directo	0.06591359	0.02814757
6	Yahoo Directo	0.07121456	0.03325922
7	Yahoo Directo	0.06988738	0.03515631
8	Yahoo Directo	0.07507573	0.04634078
9	Yahoo Directo	0.07291942	0.04678252
10	Yahoo Directo	0.0725068	0.04960874
11	Yahoo Directo	0.07387476	0.05284757
12	Yahoo Directo	0.07266699	0.05465631

data

13	Yahoo Directo	0.07224757	0.0562301
14	Yahoo Directo	0.07177476	0.05658544
15	Yahoo Directo	0.07147282	0.05757476
16	Yahoo Directo	0.07033883	0.05768155
17	Yahoo Directo	0.0691767	0.05826602
18	Yahoo Directo	0.06718641	0.05987379
19	Yahoo Directo	0.06164854	0.05998058
20	Yahoo Directo	0.06067087	0.06067087

data

Yahoo Directo	Del.icio.us	Furl	Reddit	Google	Live	AltaVista
0.0142466	0.00675922	0.00611068	0.00212621	0.04289126	0.06081942	0.07396019
0.0216068	0.00947379	0.01158058	0.00464951	0.06956019	0.09896408	0.1223466
0.02546311	0.01312816	0.01600388	0.00565922	0.1059835	0.14062913	0.14440777
0.02722039	0.01550097	0.0197068	0.00796796	0.13132524	0.14869612	0.17582524
0.02814757	0.02052524	0.02140388	0.00905243	0.15449515	0.17185728	0.2052932
0.03325922	0.02094757	0.02299612	0.0108165	0.16802621	0.18952233	0.22600388
0.03515631	0.02281068	0.02566408	0.0162068	0.18013495	0.20121553	0.23867087
0.04634078	0.02325825	0.02891553	0.01660291	0.19552913	0.21227379	0.25289029
0.04678252	0.02389126	0.0300699	0.01675049	0.20934272	0.22521845	0.26747379
0.04960874	0.0243767	0.0319767	0.01848835	0.22771942	0.23927573	0.27955146
0.05284757	0.02658155	0.03362427	0.02012427	0.23752718	0.24946699	0.28889903
0.05465631	0.02756311	0.03623981	0.02012427	0.24824369	0.25862427	0.3058
0.0562301	0.02900485	0.03798447	0.02573398	0.25638155	0.27027184	0.31944078
0.05658544	0.03159417	0.03831748	0.02707573	0.2672767	0.27691456	0.32510097
0.05757476	0.03183107	0.0385534	0.0273534	0.28086214	0.28647767	0.33929223
0.05768155	0.03343204	0.03913398	0.02794175	0.29419806	0.29331456	0.3564233
0.05826602	0.03487379	0.04018447	0.03809223	0.30383786	0.29885825	0.36613689
0.05987379	0.03577864	0.04054078	0.0398165	0.31214563	0.31285922	0.39442136
0.05998058	0.03598932	0.04198932	0.03996408	0.32505825	0.31703301	0.40489515
0.06067087	0.03923786	0.0427932	0.03996408	0.33466019	0.32472427	0.41033495



data

data

data