Analysis of Anchor Text for Web Search

Nadav Eiron and Kevin S. McCurley
IBM Almaden Research Center *

ABSTRACT

It has been observed that anchor text in web documents is very useful in improving the quality of web text search for some classes of queries. By examining properties of anchor text in a large intranet, we hope to shed light on why this is the case. Our main premise is that anchor text behaves very much like real user queries and consensus titles. Thus an understanding of how anchor text is related to a document will likely lead to better understanding of how to translate a user's query into high quality search results. Our approach is experimental, based on a study of a large corporate intranet, including the content as well as a large stream of queries against that content. We conduct experiments to investigate several aspects of anchor text, including their relationship to titles, the frequency of queries that can be satisfied by anchortext alone, and the homogeneity of results fetched by anchor text.

Keywords

Intranet web search, text indexing, anchor text

1. INTRODUCTION

One significant difference between the problems of web search and traditional text search is the availability of link structure between documents. Among other things, this can be used to effectively rank hypertext documents [7, 11], and it was known already in 1994 that anchor text is useful for web search [10]. For the purposes of this paper, we define anchor text to be the "highlighted clickable text" that is displayed for a hyperlink in an HTML page, which is to say the text that appears within the bounds of an `<A>` tag. Thus for a tag of the form

```
<A HREF="http://foo.com/">buy furniture</A>
```

we would say that the text "buy furniture" was associated with the document located at the URL `http://foo.com/`. We also include `ALT` text for image hyperlinks, when such text is available.

In spite of the fact that many commercial search engines rely heavily on anchor text, there have been several studies in which anchor text was observed to provide little or no improvement for a search task. This dichotomy was observed by Craswell, Hawking, and Robertson [2], as well as Westerveld, Kraaij, and Hiemstra [18]. In both cases they pointed out that anchor text provides a significant boost to the quality of results for site finding or home page finding tasks, whereas previous research had found little impact for the TREC subject finding (or ad hoc query) tasks. We believe that this dichotomy is related to the fact that the predominant use of web search engines is for the entry page search task, but that ad hoc queries appear in a very heavy tail of distribution of queries. Thus commercial search engines achieve their success by serving most of the people most of the time.

The question of whether anchor text is helpful for a search task appears to depend crucially on the way the task is defined, and it raises the question of determining precisely when anchor text is helpful for web search, and why? Our goal in this research is to address this question. It is our hope that these results and observations will provide future insight into the best use of this unique feature of the web in designing web search tools.

There is at least one clear reason why anchor text is helpful for search. It has often been observed that users of web search engines tend to submit very short queries, consisting of very few terms on average. In many ways, anchor text shares this characteristic, since anchor text is typically very short, and provides a summarization of the target document within the context of the source document being viewed. Thus the process of creating anchor text for a document is a close approximation of the type of summarization presented by users to a search system in most queries. When anchor text terms match a query, the target document will usually be very relevant to the query (and anchor) terms.

Our main premise is that, on a statistical basis at least, *anchor text behave very much like real user queries.* For this reason, a better understanding of the relationship between anchor text and their target documents will likely lead to more effective results for a majority of user queries. Our methodology for this investigation is based on experience with a large intranet corpus, combined with knowledge of how people express their information needs through queries to a search engine on that corpus.

The structure of the paper is as follows. In the following

*650 Harry Road, San Jose, CA 95120

SIGIR 2003 Toronto, Canada

section, we discuss the nature of the search problem, and some ways in which web search differs from traditional text search. In section 4 we describe the large corporate intranet that formed the corpus for our investigations, and why it was chosen over other more standard corpuses. In section 5 we present evidence for the hypothesis that in a statistical sense, anchortext and titles look more like queries than content. In section 10 we show that documents retrieved by anchor text techniques are in some sense "more cohesive" than documents retrieved by content indexing.

2. THE WEB SEARCH PROBLEM

As part of our understanding of the role of anchor text in web search, we should return to first principles. At a high level, we might typically think of a search problem as an attempt to satisfy a user's informational goal or need. A user's informational need is translated into a query that is fed to a text search system. The text search system uses this query to locate documents that are *relevant* to the information task as specified by the query.

This definition is extremely broad, and in practice there are a number of factors that make it difficult to construct uniform evaluation techniques for the effectiveness of a search tool for this task.

Definitions of "relevance" The concept of relevance has multiple definitions, and is often a source of ambiguity in evaluating a system. See [15, Chapter 5] or [17]. In particular, the relevance of a result may depend on the state of the user.

Use of prior knowledge The user may know precisely what they are looking for, perhaps by having seen it before, or by having been told that it exists. A user may also know precisely the concept they are searching for, or they may not know exactly how to express the terminology of the concept they are searching for.

Authoritativeness One critical feature that has fueled the rapid growth of the web is the fact that the barrier to publication is extremely low, and everyone can have a voice. This is unfortunately also a curse for information retrieval, because it results in a huge document collection of wildly varying degrees of credibility and authority. Typically, users value a few authoritative pages more than they would thousands of non-authoritative pages on the subject, and system design revolves around the need to prune the list of potentially relevant documents.

Goal of a system In web search it is often the case that queries match many documents, many more than the user could possibly ever read. However, because of the diversity of the sources and the sheer size of the corpus, many of these pages are either irrelevant in the context the user intended, or they are not authoritative enough to be useful for the user. It is typically more useful to the user for the search engine to return a smaller result set that contains pages that are either highly authoritative on the subject of the query (analogous to the home page finding task of TREC), or that are well linked to pages with additional information on the subject. For this reason, the TREC entry page search task provides a good model for the task faced by many web search systems.

An additional problem that arises on the World Wide Web is that authors often clamor for the attention of readers, which prompts the need for adversarial analysis into the design of web search. We have specifically chosen to avoid these issues by concentrating on the problem of intranet search. There is independent interest in focusing on intranets in particular, because the social forces surrounding their creation are different from the world wide web, and their usage is also different.

3. EVALUATION TECHNIQUES

Much of the purpose of the TREC conference has been to develop a standard methodology by which the effectiveness of text search tools can be judged. Unfortunately, test methodology for search tends to incorporate assumptions about the nature of the search task. In cases when these assumptions are a close match to how users interact with search tools on a particular corpus, this methodology can be very effective. Unfortunately, there are many pitfalls in extrapolating from the results of these tests outside the scope of their implementation and definition. For example, the quality and homogeneity of the underlying corpus can directly affect the quality of search results for a particular approach, because some methods may be more robust to vagaries of a particular corpus.

The methodology used in TREC has been to use standard corpuses by which methods can be tested and judged against each other, using queries and tasks that are defined ahead of time. In this paper we pursue a somewhat different approach to understanding the nature of web search. We believe that in the evaluation of the success of a system, statistics about queries can be as significant as statistics about the content. For this reason, our investigation focuses on the structure of a large corporate intranet, and we use the knowledge of a set of queries by people searching for information within that corpus. One advantage of our approach is that we are able to see how real users express their information needs, by examination of the query logs on a search engine designed to search the underlying corpus.

We believe that such examination is complementary to the TREC approach, and that both have some value. In some ways this is similar to the evaluation of algorithms and complexity in computer science. The performance of an algorithm may be evaluated for its worst case performance, or it can be evaluated for its average case performance. It can also be evaluated as it performs on a probability distribution of inputs that mimics a real world data stream. The approaches are somewhat different, but complementary.

One disadvantage of our approach is that we are unable to make sweeping claims about precision and recall of a particular system, because the actual need is unknown to us. This is an area in which TREC excels, because a great deal of effort is expended to construct queries for which the underlying "right answer" is known. One danger in the TREC approach is the methodology only allows a small number of queries, and may not be representative of an actual workload. Moreover, these queries are designed to match the definition of the task, without examining the suitability of this definition against actual usage. By contrast, our approach of studying a relatively encapsulated corpus with a large set of users searching for a large number of concepts offers some hope for better modeling of the process by which users translate their information need into queries.

Our approach does not eliminate the danger that the evaluation does not match the actual usage, because there is also a danger that the observed query statistics will reflect the state of the system as it already exists, but if the system is changed, the query statistics may change as people adjust their usage to the behaviour of the system.

4. OUR EXPERIMENTAL DATA

We chose to run most of our experiments on a combination of documents on the internal and external sites of a large corporate intranet [1] (we refer to this collectively as the IBM Intranet), from which we crawled approximately 20 million URLs. The IBM intranet consists of approximately 7,000 servers worldwide, with documents in many different languages, and produced by a wide variety of content generation methods. Aside from the obvious content differences, this large intranet appears to mirror the commercial part of the web in many ways. Due to a variety of factors, the data in this crawl has a high degree of duplication, and once we removed all duplicate and near duplicate documents, we were left with 2,952,344 documents.

Tokenization and parsing of these documents producing 2,569,880 anchor text records, containing 57,144,748 tokens, of which 762,965 are distinct. Among these distinct tokens, only 448,534 are purely alphabetic, which reflects the fact that many of the anchor texts are machine-generated from databases or contain filenames. This is a fact of life among people working to index intranets or the World Wide Web – a large amount of the content that is to be indexed is not "pure text", but instead contains a mixture of text and other data produced by databases.

4.1 The Query Load

One of the advantages of our corpus over other web data sets (e.g., the TREC data sets) is that we have access to the complete logs of queries made against this corpus by real-world users. In other words, not only do we have a large and fairly characteristic web data set, but we know how IBM employees characterize what they are looking for in this data set.

The queries we used in our experiments are based on a "live" log of the queries submitted to the IBM intranet search engine over a period of about five months in 2002. We have cleaned up the queries to eliminate queries that we believe originated with automated tools, and not actual human users (such as excessively long queries, queries with many "or" operators, etc.). We also eliminated all queries that used operators limiting the query to a specific host or a group of hosts, removed queries that contained phrases, and down-cased all terms. The final set of queries we use included 448,460 distinct queries, representing a total of 1,265,395 queries executed against the search engine. Figure 1 shows the resulting distribution of queries and query terms. As expected, the query distribution is a tail-heavy power law distribution.

It has often been observed that users of web search tend to submit very short queries [4, 16]. Our query logs exhibit a very similar behaviour (a more detailed analysis of IBM intranet query logs appeared in [19]). In particular, more than 50% of all queries consist of a single term.

[1]The identity of the corporation is withheld in order to comply with the anonymity of the submission.

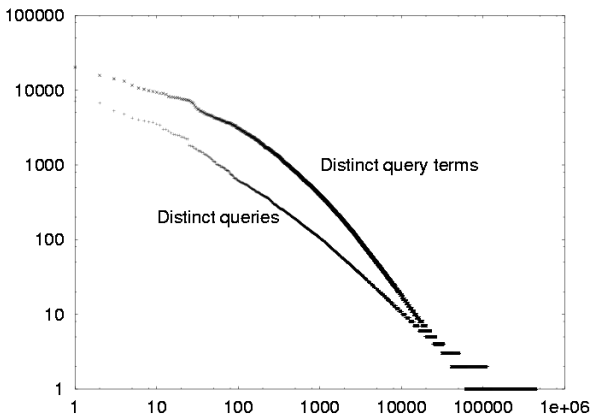

Figure 1: The distribution of distinct queries and distinct query terms

5. FEATURES OF ANCHOR TEXT

We believe that, in large part, the suitability of anchor text as features for hypertext search stems from the inherent similarity between anchor text and the queries user typically submit to search engines. Essentially, anchor text is typically a short summary of a document, rarely more than a few words long. Search queries are, many times, similar in nature: They are a few words long (see Figure 2), and express a summary of a subject that the user is interested in. Even when a search engine uses anchor text from many source pages as a single bag of words associated with the target page, the relative succintness of the anchor text remains unchanged. In many cases, multiple anchor texts referring to the same target page will be identical, but there are a significant fraction of pages that have multiple distinct anchor texts associated with them. In the corpus we studied there are more than 508,000 documents that have at least five distinct anchortexts associated with them.

Both anchor text and queries tend not to be complete sentences. There are a few very common colloquialisms such as "next" and "click here", but beyond that anchor text is usually just a noun phrase, providing a description of the target phrase. Similarly, most queries are either a noun phrase (such as "DB2 performance") or contain a collection of nouns and adjectives. Rarely are verbs key terms in a query. In that sense, the vocabulary of queries, as well as their grammatical form, is similar to that of anchor text, and vastly different from that of the full page content. The query logs that we studied show a heavy use of jargon and acronyms of the institution, and the same is true of anchor text in our corpus.

The distribution of the number of terms in queries and anchor text for the IBM intranet is shown in Figure 2. We have also included the distribution of the number of terms in titles from the corpus, as well as the number of distinct terms from all anchortext to a page. The number of distinct terms has a somewhat more heavy tail distribution than the others. It appears that titles and collected anchortext are very similar in their distribution of terms, and in the sections that follow we will address the degree to which they complement each other.

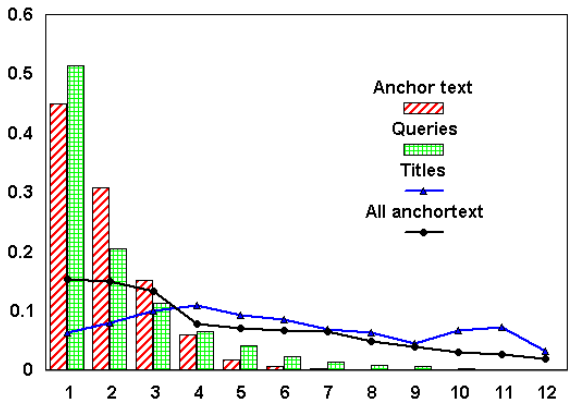

Figure 2: Number of distinct terms in titles, queries, individual anchor text, and collected anchortext. Collected anchortext aggregates all the anchor text from the links to the page.

6. ANCHOR TEXT, TITLES, AND QUERIES

We have argued that one reason that anchor text is so useful for web search is because most users use short queries, and by so doing they tend to choose a very small number of terms that precisely summarize the pages they are looking for. We can model this by hypothesizing the existence of a space of "concepts" that users typically search for. When a user wants to search for this concept, they select a set of terms that describe the concept, and submit this as a query. Similarly, when an author wishes to insert a hyperlink to a page, they need to select anchor text that will provide a short summary for the page that is linked to. There are differences in the constraints placed on the selection of anchor text by an author and the selection of query terms by a search engine user, but as a first order approximation we could imagine them as being both selected from the same "concept space".

It was observed by Jin, Hauptmann, and Zhai [5] that document titles also bear a close resemblance to queries, and that they are produced by a similar mental process. Thus it is natural to expect that both titles and anchor text capture some notion of what a document is *about*, though they are linguistically dissimilar. One advantage that anchortext has over titles is that while there is typically only one title authored by the same author as the document itself, there can be many anchor texts, authored by the same author as the document or by other authors. Thus anchor text provided collective summary information. While it is natural to expect that titles should reflect a short summary of the document, there are instances in which the language of the document itself does not represent the collective wisdom of what the document is about. A good example is provided by the query "layoffs" which are often known in the corporate world by other language. Thus the use of anchor text can help with the problem of synonomy.

We performed an experiment to test this hypothesize that anchortext behaves like titles in capturing "aboutness" of documents. In order to do this, we compared the distribution of documents retrieved by matches in their titles, anchortext, and content. The result sets by content tended to be much larger (as one would expect), but we would expect the documents fetched by title to be statistically more relevant to the queries. The question is whether the documents fetched by anchortext share this feature, and for this we used methodology to measure similarity of document collections.

6.1 Term Frequency Similarity

One common way to measure the similarity of two document collections is through similarity of their term frequency distributions see [6]. This reduces the problem to one of comparing two probability distributions, and for this there are numerous techniques (see [9]). One well known measure from information retrieval is the cosine distance. Given two probability distributions p and q on a set of terms T, we define the cosine distance as

$$\cos(p,q) = \frac{\sum_{t\in T} p(t)q(t)}{\sqrt{\sum_{t\in T} p(t)^2}\sqrt{\sum_{t\in T} q(t)^2}}$$

Another measure that is commonly used to measure distance between probability distributions is the Kullback-Leibler divergence, defined as

$$D_{\mathrm{KL}}(p,q) = \sum_{t\in T} p(t)\log(p(t)/q(t))$$

This measure has several drawbacks however, including the fact that it is asymmetric, and is only defined in the case where p is absolutely continuous with respect to q (because of the fact that q may vanish while p does not). Hence as an empirical measure, the Kullback-Leibler divergence is extremely sensitive to sparse data in the observations. An alternative measure is the (balanced) Jensen-Shannon divergence, which we define as:

$$\mathrm{JS}(p,q) = \frac{1}{2}(KL(p,(p+q)/2) + KL(q,(p+q)/2))$$

Note that the Jensen-Shannon divergence is symmetric, and measures the distance from the two distributions to their mean. Because of our normalization, it is bounded between 0 and 1, and is 0 precisely when p and q are identical, and equal to 1 when p and q have disjoint support. The Jensen-Shannon divergence may be used in a test of the hypothesis that two samples are taken from the same distribution. Under this hypothesis, we would expect that $\mathrm{JS}(p,q) = 0$. The larger $\mathrm{JS}(p,q)$ is, the more evidence that p and q are different, and in fact $\mathrm{JS}(p,q)$ is proportional to the minus logarithm of the probability that the two distributions are identical.

6.2 Similarity of Titles and Anchor Text

In order to test our hypothesis that documents returned by matches on anchortext are statistically more relevant to queries, we assume that documents fetched by title matches satisfy this and compare the results from using anchor text to the results from using titles. We picked 102 distinct random queries from the intranet query logs (representing 125 total queries from the logs) and fetched the result sets for them when querying by anchor text, titles, and content[2]. For each query, we then compared the term frequency distributions from the documents fetched by the three methods.

[2]We limited the number of result set documents we fetch to 1000 for practical reasons. If a query's result set was larger, we sampled uniformly at random 1000 documents from the result set.

After discarding 9 queries for which we had empty result sets for one of the methods, there were 93 remaining queries, and for 59 of those the JS divergences showed the documents fetched by anchor text and titles to be more similar than the documents fetched by anchor text and content or the documents fetched by titles and content. Similar results were achieved using the cosine distance. This suggests that fetching documents by anchor text matches seems to produce documents that are very similar to documents that are fetched by title.

7. INCORPORATION INTO MODELS

The main purpose of this work is to examine properties of anchor text that makes it particularly attractive for use in web search, but due to lack of space, we do not directly address the issue of how to best incorporate anchor text into models of information retrieval. In [2] it was observed that using anchor text alone in a BM25 ranking scheme was particularly good at the site finding task. In the final version of this paper we will include a reference to recent work [3] on TREC-style experiments on a different approach to mixing the effects of anchor text, content, titles, and other ranking schemes. We believe that this remains a fruitful area of research for the future.

In [8], they suggested a method by which it is possible to combine a language model for anchor text with a language model for content in a system based on statistical language models. The reasoning is that "anchor texts and the body texts ('content only') provide two very different textual representations of the documents." They followed a probabilistic model for information retrieval, in which documents are ranked by their probability of relevance to a given query. Under fairly standard assumptions, we have

$$P(D|T_1,\ldots,T_n) \approx P(D)\prod_{i=1}^{n}\{(1-\lambda)P(T_i|C)+\lambda P(T_i|D)\}$$

The quantity $P(T_i|C)$ is estimated using the distribution of terms in the collection and the observation of queries made on the collection. The crucial difficulty in this approach is to estimate $P(T_i|D)$, namely the relevance of a term to a given document. In [8] it was suggested to mix two models of content for anchortext and anchor text, using

$$P(T_i|D) \approx \mu P_{\text{content}}(T_i|D) + (1-\mu)P_{\text{anchor}}(T_i|D)$$

If a term appears in the anchortext of a document, then this term may be a likely candidate for inclusion in the model of similar documents, and this framework provides a way to incorporate this. Along similar lines, the incorporation of a language model for titles was addressed in [5]. The fact that anchor text functions in much the same way as title in summarizing a document suggests that it might make sense to unify these, particularly since they have significant correlation to each other. In section 8 we present evidence that these are strongly correlated to each other.

In the rest of this paper, our goal is not to measure the results of a particular search engine on a particular corpus, but rather to examine characteristics of anchor text that are relevant to different approaches. We make no attempt to address the ranking of documents according to their relevance to a query, although this is clearly the major problem in an environment in which there are simply too many potential results to present to a user. Our goal is only to make statistical comparisons between result sets fetched using different features of web documents, and for this reason we use only simplistic models of information retrieval hereafter. Specifically, we use Salton's Vector Model [14, 13] and the classic probabilistic Binary Independence Retrieval model [12]. The inclusion of stemming, case preservation, multilinqual methods, or other sophisticated techniques are useful for achieving better retrieval and ranking of results, but they would unnecessarily complicate our statistical observations so we omitted them. We refer the reader to [1] for a survey of information retrieval models.

8. QUERYING BY ANCHOR TEXT

As a first test, we wished to judge how effective anchor text alone would be in text indexing. We therefore constructed an inverted index of the anchor text for the document set, by concatenating together all of the anchor text of links pointing to a document, and using that as a virtual document instead of the actual content. We created similar indices to allow us to perform queries on the content and the titles of documents. We then selected a set of 10,000 random queries from the query log, resulting in a total of 7474 distinct queries (so that they represent a representative sample of the distribution of actual queries).

Table 1 shows the number of queries that had non-empty result sets using each of the three indices. We break down the results by the size of the query.

Terms in query	1	2	3	4	5	all
% satisfied by anchor text	72	62	42	40	25	60
% satisfied by titles	70	55	29	32	14	55
% satisfied by content	74	82	85	87	78	79

Table 1: Number of queries of various sizes for which results were found using each of the three indices we built

One interesting observation from this is that the use of multiple terms in queries has a different effect on titles than it does on anchor text. While on short queries titles and anchor text provide similar performance (at least by this crude measure), the advantage anchor text has over titles grows significantly as the size of the query grows. This suggests that the events of individual query terms appearing in anchor text for a document are not independent. Hence we might expect that a large fraction of multi-term queries are not ad hoc queries, but are instead entry page queries using a multi-term name for the information need, or terms that may be used to defined the same concept in different situation and by different people.

We also examined the following related question: how often does searching anchor text find documents that do not contain the search terms themselves? Much to our surprise, out of 1000 pages picked at random from the corpus we found only 664 for which all terms that appear in anchor text also appear in the content. 130 of these 1000 pages had *none* of their anchor text contained in the document itself. However, when looking at the results of our 7474 randomly chosen queries, we find that a full 86.5% of the pages that were found by querying on anchor text were also found by querying on content. We therefore conclude that the terms that typically occur in queries are more likely to be repeat-

ed in the content than the average anchor text term. This phenomenon can be explained by the prevalance of anchor text that serves navigational purposes (such as "next", "up", "prev", "click here") which typically is not repeated in the text, but is not useful for querying either. These results confirm the intuition that anchor text querying typically finds smaller result sets that are almost always entirely contained in the result sets found by querying on content.

8.1 Overlap of Titles and Anchor Text

Since titles and anchortext have been observed to fulfill a very similar function, it is natural to ask if they are essentially identical. For the IBM intranet, we found 320,826 distinct alphabetic terms in anchor text, whereas the titles contained only 139,617 distinct alphabetic terms. In order to further address question this, we investigated how often anchortext for a document contained terms in the document that were *not* in the title of the document. In order to concentrate our attention on "important" terms, we confined our counting to terms that had actually appeared in the search logs. Among the 2,395,766 documents for which we had anchor text, content, *and* title information, 60.6% contained terms that people had searched for, were contained in their body and their anchor text, but not in their title. Fully 13.5% had 6 or more additional such query terms in their anchortext. Thus it appears that anchor text provides a potentially important enrichment of the information supplied by authors in their titles.

9. ANCHOR TEXT TERM DISTRIBUTION

Just as ordinary text has certain words that appear frequently such as "the" and "if", so does anchortext. In particular, there are many commonly used phrases or words in anchortext such as "click here" or "home", or "next", and these query terms should receive relatively little weight in evaluating the relevance of a document to a query. One notable difference is that the most common terms in anchortext are generally *not* the same terms that appear frequently in text. In particular, Table 2 shows the sixteen mostly commonly occuring words in both anchortext and content of the IBM intranet. There is some overlap to be sure, but each has their own common vocabulary and terms that often appear in anchortext can be somewhat rare in content. This is not altogether surprising, but it suggest that any indexing method that depends on term frequencies should keep track of these statistics separately in order to fully exploit the semantic meaning of anchor text. A similar suggestion was made in [8].

Most common alphabetic terms		
Anchor text	Content	Titles
next	the	(omitted)IBM
topic	to	calendar
domain	a	bookserver
(omitted)IBM	of	bookmanager
prev	and	for
previous	in	of
page	for	via
to	(omitted)IBM	and
index	is	news
the	this	by
for	b	index
and	on	linux
search	that	software
of	by	guide
linux	or	java
contents	you	channel

Table 2: Most frequent terms in anchor text, content, and titles of the IBM intranet. Note that only 5 of the top 16 anchor text terms appear as most frequent terms in content.

10. HOMOGENEITY AND ANCHOR TEXT

The predominant use of short queries in web search suggests that the most common goal of users is to find an "entry page" for a given topic that they specify, and inspection of the most common queries from the intranet search log that we studied is consistent with this. Some examples of very popular queries from the intranet search log include "vacation", "benefits", "travel", or "reserve". Each of these terms has broad usage, but it is natural to expect that users are searching for the entry page of a specific business function known under this name. One problem of searching content for these terms is that it tends to turn up pages that represent every possible use of the term. Our observation from using anchor text in search is that it tends to concentrate on pages that would naturally be summarized by that single term, and are therefore fairly narrow in their scope.

In order to test this hypothesis, we need to measure the *homogeneity* of the result sets using anchor text and content to perform our lookups. The precise definition of homogeneity of a corpus has several definitions, but we chose to use a methodology similar to that of Kilgariff and Rose [6]. They proposed several quantitative measures of homogeneity, based on the principle that you should randomly divide the corpus into two pieces and measure the similarity of the two pieces to each other. In [6] it was suggested to use a Spearman or Chi square statistic, or else use a cross-entropy measure. Instead we chose to use the Jensen-Shannon divergence and cosine distances as measures of dissimilarity for the two halves of the corpus.

Our experiment was as follows. We used the same 102 distinct random queries that were described in section 5. We then computed the homogeneity, in terms of term distribution, for each result set.

We discarded queries for which result sets were small, since homogeneity of small document sets is very sensitive and cannot be regarded as a reliable indication of the utility of the documents to a human user. Figure 3 shows the results for the 14 queries out of the 102 for which both anchor text and content result sets contained more than 800 documents. The results indeed indicate that, except for one outlier query, the homogeneity of the anchor text result sets is higher than that of the result sets obtained by querying on content.

The improved homogeneity of results returned by anchor text suggests that documents returned by matches on anchor text will tend to be focused on just one meaning of the terms queried for, and that this meaning will be the most common meaning. One of the reasons for this phenomenon is the brevity of anchor text when compared to documents. When matching a multi-word query against a long document, the

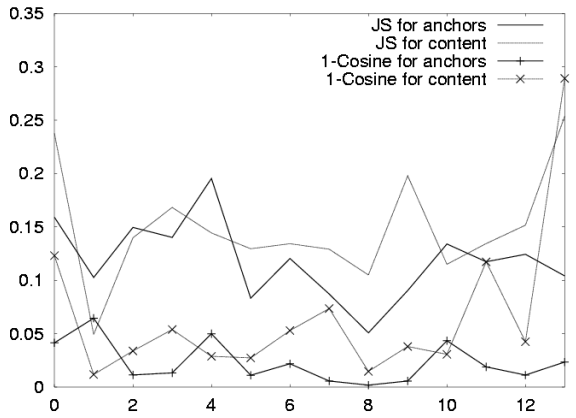

Figure 3: Homogeneity of result sets by anchor and by content for queries with large result sets.

various words of the query may match in different parts of the document. These parts may talk of only distantly related subjects, and therefore, while each one of the words queries on indeed appears, they never appear within the same context. Hence, the document is not a good match for the concept represented by the query. Anchor text, on the other hand, is short enough that if multiple words from the query appear in it, they always appear in great proximity, and will therefore tend to have the same meaning as they have in the query itself. One might expect the same to be true of titles, however, as we mentioned in Section 8, titles are not rich enough to be useful for multi-word queries.

11. CONCLUSIONS

We have presented a statistical study of the nature of anchor text and real user queries on a large corpus of corporate intranet documents. We have found significant evidence that supports our hypothesis that anchor text resembles real-world queries in terms of its term distribution and length. We have also found that anchor text is typically less ambiguous than other types of texts, resulting in a more coherent and focused result set for queries based on anchor text than for those based on other features of the corpus. In addition to providing a better match than titles or content to the language people use for queries, anchor text also holds the promise of providing more authoritative results to queries.

We have also studied the nature of titles of documents in our corpus. While traditional text search engines have found titles to be a very useful feature of a document, we have found titles to be far less useful than anchor text. While titles are typically longer than individual anchor text, many pages (especially highly relevant pages) have many individual anchor texts pointing to them. This provides for a much better indication of the summarization of the page in different contexts, by different people, than that afforded by a single title which is authored by one author.

We believe that the study of real user queries on real corpora is essential if an improvement of the average "user experience" with hypertext search is to be gained. Current typical search engine users, as well as hypertext authors, are far from being the professional information retrieval experts who were the typical users of earlier search engines.

We hope our results, and other similar results that may follow, will lead to a better understanding of when and why different features of a hypertext corpus contribute to the performance of search engines. Such understanding should ultimately lead to better search engines that apply the most appropriate search techniques depending on query and corpus characteristics.

12. REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press and Addison Wesley, New York, 1999.

[2] Nick Craswell, David Hawking, and Stephen E. Robertson. Effective site finding using link anchor information. In *Proc. of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, pages 250–257, New Orleans, Louisiana, USA, September 2001. Association for Computing Machinery.

[3] Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin, and David P. Williamson. Searching the workplace web. In *Proceedings of the Twelfth International World Wide Web Conference*, Budapest, 2003.

[4] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society of Information Science and Technology*, 53(3):235–246, 2000.

[5] Rong Jin, Alex G. Hauptmann, and ChengXiang Zhai. Title language model for information retrieval. In *Proc. of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, pages 42–48, Tampere, Finland, August 2002. Association for Computing Machinery.

[6] Adam Kilgariff and Tony Rose. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, May 1998.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proc. of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, pages 27–34. Association for Computing Machinery, 2002.

[9] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Association for Computationl Linguistics*, pages 25–32, 1999.

[10] Oliver A. McBryan. GENVL and WWWW: Tools for taming the Web. In *Proceedings of the First International Conference on the World Wide Web*, Geneva, Switzerland, May 1994. CERN.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[12] S. E. Robertson and K. Sparck Jones. Relevance weighting of search items. *Journal of the American Society for Information Sciences*, 27(3):129–146, 1976.

[13] G. Salton. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[14] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, January 1968.

[15] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[16] C. Silverstein, M. Henzinger, H. Marais, , and M. Moricz. Analysis of a very large altavista query log. Technical Report SRC 1998-014, Digital Systems Research Center, 1998. See also SIGIR Forum 33(1), 6-12.

[17] C. J. van Rijsbergen. *Information Retrieval*. Butterworth's, London, 1979.

[18] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, URLs and anchors. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC-10)*, pages 663–672, 2002.

[19] Jason Zien, Jörg Meyer, John Tomlin, and Joy Liu. Web query characteristics and their implications on search engines. In *Poster Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, 2001.