Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Patterns of Document Access in Searching and Browsing

David Bodoff
Hong Kong University of Science and Technology, dbodoff@ust.hk

Lydia Zhang
Hong Kong University of Science and Technology, zhangjin@ust.hk

Follow this and additional works at: http://aisel.aisnet.org/amcis2000

Recommended Citation
Bodoff, David and Zhang, Lydia, "Patterns of Document Access in Searching and Browsing" (2000). *AMCIS 2000 Proceedings*. 289.
http://aisel.aisnet.org/amcis2000/289

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2000 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Patterns of Document Access in Searching and Browsing

David Bodoff, Lydia Zhang
{dbodoff, zhangjin}@ust.hk
Department of Information and Systems Management
Hong Kong University of Science and Technology
Clear Water Bay
Hong Kong

Abstract: In the current work, we investigate the feasibility of using past experience to predict which documents will be accessed by users. Document access may be viewed as a surrogate measure of relevance, in which case the discussion here regards a method to improve retrieval effectiveness. But our main concern here is with users' access patterns per se. The prediction of future document accesses based on the past, is applied to two different IR services, (a) browsing and (b) keyword search. A straightforward method using conditional probabilities shows promise in both cases, while very different access patterns are observed for users of the two different IR services. These results have potential technical uses in improving document retrieval, and also shed light on the very significant differences between users of different IR-related services.

Introduction

Information Retrieval (IR) is by now a well researched field, and sophisticated IR systems are deployed and in widespread use. Since the late 1990's, Internet and Intranet search engines are very visible implementations of these IR systems. Yet in spite of all that is known about the technical and cognitive aspects of text searching, textual information is not fully and easily exploited as part of standard business routines (Gordon, 1997), or even as part of exceptional strategic research. With so much information and so many knowledge workers, the unanswered question remains how to fully exploit the available information when doing our productive jobs.

One example of this challenge, relevant to the current research, was reported by (Vandenbosch, 1997). They studied executives' use of Executive Information Systems (EIS), and particularly watched whether the executives used the EIS for unfocused scanning of the environment or for focused question-answering. This was an important element to track, since scanning had previously been linked with increased organizational effectiveness, while question-answering was considered as related to increased efficiency. The authors found that executives were more likely to use the EIS for narrow question-answering, and this was considered as a lost opportunity.

In the current work, we investigate the feasibility of using past experience to predict which documents will be accessed by users. Document access may be viewed as a surrogate measure of relevance, in which case the discussion here regards a method to improve retrieval effectiveness. But our main concern here is with users' access patterns per se. This prediction of future document accessess based on the past, is applied to two different IR services, (a) unfocused scanning or browsing, and (b) keyword search. A straightforward method using conditional probabilities shows promise in both cases, while very different access patterns are observed for users of the two different IR services. These results have potential technical uses in improving document retrieval, and also shed light on the very significant differences between users of different IR-related services. As the users in this study were institutional, these results may contribute to a better understanding of how corporate users search and browse documents.

1 Browsing versus Searching

Browsing and searching are two different ways to access documents. Numerous authors have compared these two processes in the physical and digital worlds. Users' goals are different for the two processes, and the available technologies are different as well. Regarding users' goals, searching versus browsing indicate two ends of a continuum. With browsing, the user looks "through information without a particular problem to solve or question to answer, while focused search (Huber 1991) occurs when people are looking for something specific" (Vandenbosch, 1997).

The browsing activity can be further broken down into different types. One type of browsing involves "exploring topic areas" (Gutwin, 1999). In this type of browsing, the user wants "to find documents in a general area, but without knowing exactly what they are looking for" (ibid. p. 82). Through browsing, "users may try to gain an understanding of the topics that are part of the area, may wish to gather contextual information for directing and focusing their search, or may simply hope to come across useful material" (ibid. p. 83). This particular type of browsing is related to a service known as "current

awareness" in the library sciences literature. "The term 'current awareness' was coined to describe the state of keeping up with new developments" (Marchionini, 1997).

Regarding technologies, keyword search is the most common technology used to support search, while for browsing, the most popular technologies are hypertext and directories (Chen, 1998). In the special case of current awareness browsing, there exists a special case of a directory, called a clipping or tracking service. A clipping service gathers together into one place all documents related to one area. In essence, it is a directory with one entry. To provide this kind of clipping service, the service provider surveys numerous document sources, and identifies -- either manually or automatically -- all the documents that pertain to the given topic. Once these documents have been identified and co-located, the user can easily browse them. In this way, the user's current awareness is supported.

2 Favorite Documents

In this study we are interested in finding which documents, if any, appear to be more frequently retrieved and selected by searchers. We do not address the question of matching a document to a particular query. Rather, we address the question of identifying the a priori probability that each document will be selected and considered as relevant to an arbitrary query. These two concepts -- a priori relevance versus conditional relevance -- are reviewed in the following paragraph as they pertain to Information Retrieval.

Let the probability of a document's relevance to a query be denoted as a conditional probability P(R | D,Q). IR research commonly invokes Bayes theorem by which

$$P(R \mid D,Q) = P(R \mid D) * \frac{P(Q \mid R,D)}{P(Q \mid D)} \quad (1)$$

(see v. Rijsbergen, 1977; Turtle 1991, Fuhr 1990).

The focus of most research in IR is on the term P(Q|R,D) which is the probability that document D will be relevant to query Q. This term is implemented in search engines by a matching function between a query and a document. On the other hand, the first term -- P(R|D) -- represents the a priori probability that document D is relevant to an arbitrary query. In other words, this term represents a sort of a priori popularity of a document. This a priori probability is conceptually important as a component of (1), but is not often used in practice. In the current work we investigate the possibility of estimating this often-neglected term of a priori relevance. A good estimate may be used in expression (1) to improve retrieval performance. In addition, this investigation sheds light on user behavior, since the term represents users general level of interest in different documents across time.

We expect to find that users' interest in documents is not randomly distributed, but that some documents do attract more attention than others. If this is so, and not all documents have the same a priori probability of relevance, then we will want to investigate possible methods for predicting each document's prior probability of relevance P(R|D).

3 The Prediction Model

To use an analogy from equity markets, we may adopt a fundamental or a technical model. A fundamental analysis would predict a document's probability of access based on its underlying features, e.g. its publisher, topic, length, etc. A technical analysis would predict the document's future accesses on the basis of its past accesses. In the current work we adopt a technical analysis of this sort, as this is the data that was most readily available. We aim to predict which documents will be accessed at a time *t* on the basis of the document's access history from times 1,2,...*t-1*. We use a simple conditional probability as a prediction model: What is the probability of a document's being accessed in time t after its publication, given that it was accessed (at least once) in time(s) 1,2,…*t*-1 after its publication.

An additional aspect of the predictive model is whether it treats individual documents or classes of documents. In the simplest case, we can treat each document separately and indivisibly. In this model which we call Model 1, the retrieval history of an individual document is modeled to predict that document's probability of relevance. The main idea of Model 1 is that some individual documents are more likely to be helpful to an arbitrary user than other documents. This is the approach we take here.

A second approach which we call Model 2, is to identify *characteristics* of documents that seem to predict their probability of relevance --e.g. their length, their author, their URL, etc. This approach has been effectively used (Fuhr, 1991; Gey, 1995) to estimate probabilities P(D|Q), and it may also be used to predict P(D). In general, this is an interesting question for IR research. The main idea of Model 2 is that some document *types* - -i.e. documents possessing certain identifiable characteristics - - are more likely to be helpful to an arbitrary user as compared with other document types. In the current study we adopt Model 1. An investigation of Model 2 is left for future work.

Combining the elements discussed in this section, the final functional form of the model is thus specified as $P(R_t \mid D_i, R_{t-1}, R_{t-2}, \ldots)$ i.e. the a priori probability of relevance

for document D_i in time *t*, given that document D_i was relevant in times *1, 2,…t-1*. We do not have any theoretical reason for supposing that any particular time period t should be predicted by any particular previous periods 1..*t*-1. As a result, in this exploratory study, we investigate different independent and dependent time periods.

4 Experimental Setting and Available Data

We obtained data from an online data aggregator called Wisers. This company makes available over the Internet electronic copies of all major Chinese language newspapers in Hong Kong, as well as one English-language newspaper. Wisers indexes the documents using a proprietary bi-lingual Chinese and English language search engine. The service is subscription-based. Wisers provides a number of different services through their World Wide Web interface. The two that are relevant to the current research are the full-text keyword search and "tracking". Full-text search is used to search the archives using keyword queries. The tracking service supports current awareness by collecting all the day's stories pertaining to a topical area (e.g. "The Environment", "Law and Judiciary", etc.) into one folder that can be easily browsed. Not every document is included in a tracking folder, while some documents are included in more than one folder. All documents, including those that are included in a tracking folder, are indexed and accessible through the full-text search.

Whether the user clicks on a pre-designated folder, or submits a keyword query, Wiser displays to the user a list of document titles with the first few lines of text from each. As with the familiar form of search engine interface, the user may then click on a document title to view it. If the user does this, then Wisers records this as a document access.

We may regard these document accesses as a surrogate measure for a document's relevance to the user. We believe this is a reasonable surrogate because the Wisers interface included a substantial excerpt from the beginning of each document, so users had a reasonably good idea of the article's contents (this interface was subsequently modified). To the extent that document accesses is considered as a surrogate for document relevance, then results reported here suggest methods for increasing the accuracy of search engine relevance predictions as discussed in previous sections. To the extent that one is less confident in this surrogate as a measure for relevance, then our results merely allow us to predict probability of future *accesses*, but not necessarily of relevance. These predictions are interesting for what they say about user access patterns per se, with technical implications limited to questions of caching and efficiency.

There are two relational tables that record document accesses. The first table, "search_access", records document accesses that are made when the user clicks on a document from the results screen of a keyword search. The second table, "browse_access", records accesses that are made when the user clicks on a document listed in the contents of a tracking folder. Each table has essentially the same four fields: Username, DocID, Query_Date/Acc_Date, and Query/Acc_Point. Username is a foreign key of a User Data table[1]. DocID contains substrings that denote the article's publisher, the date of publication, and the article number (e.g. article 17 of Ming Pao Newspaper from August 28, 1999). The third field is called Query_Date in the search_access table and Acc-Date in the browse_access table. This field denotes the date of document access. The last field is called Query in the search_access table and Acc-Point in the browse_access table. This field contains the user's keyword query in the search_access table, and the name of the tracking folder (e.g. "Environment") in the browse_access table.

An example record from the search_access table is:

User	Docid	query_date	Query
User0002	l999071702 80012	8/7/99	young,migrants need,health,at-tention

An example record from the browse_access table is:

User	Docid	query_date	Query
User0004	l999073103 00100	8/1/99	Information Technology

where "Information Technology" is the name of the Wiser's pre-defined tracking folder through which this user was browsing when he clicked on this document.

5 Data Sampling

For the current exploratory study, Wisers selected a subset of their very large dataset. The data was limited according to user and access date. The user field was limited to four financial services firms in Hong Kong. The access dates were limited to a three-month period, August-October 1999, that was believed to be typical. In terms of the attribute names of our two tables, the data set was limited by selecting on the Query_Date field for the query table, and the Acc_Date field for the tracking table, for that three-month period. Thus, the data represents all

[1] The user data is not used in the current work for the following reason: Each "user" identifies only the financial industry company whose subscription is being used. The transaction logs do not attempt to identify individual users within the corporate subscriber.

the user accesses that were made during that period. The resulting search_access and browse_access tables have 72,415 and 30,871 records respectively.

In the current investigation we were exclusively focused on whether accesses in the early days of a document's existence are good predictors of its being accessed later. A problem we faced is that these periods are relative to each document's creation date, so no single window of logged accesses can give us this information. To arrange this sort of data, we therefore did the following: From the sample of document accesses we were given, we considered only documents that were created from August 1 to August 28 1999. Then for each of these documents, we considered only its access patterns for its first 9 weeks of creation. In this way, all documents under consideration were alike, in that for each one, we had available the first 9 weeks of that document's access history. After limiting the sample in this way, we were left with 64,831 *documents*. The access patterns for these documents are discussed in the following sections.

6 Preliminary Data Analysis

If we do not limit the sample to the first nine weeks after each document's publication, then we find 72,415 accesses via keyword queries and 30,871 via browsing. On the other hand, during the first nine weeks after a document's publication, we found 6871 accesses through keyword search and 6867 accesses through browsing. This indicates that as time elapses following its publication, a document is more likely to be accessed via keyword search and less via tracking. This stands to reason, since the users of a tracking service are primarily interested in what's new, while focused searches are often archival.

Next, we wanted to ascertain whether the distribution of relevancy data across documents shows any structure. Some documents are positively assessed more frequently than others, but this would also occur if these positive assessments were randomly assigned to documents. Let d denote the number of documents in the data set ; n document accesses; q queries; $m=n/q$ number of document accesses per query. Then the distribution of randomly assigned relevancy assessments could be modeled as d independent Binomial variables, each with $p = m/d$ and q trials. That is, for this model each document faces an equal m/d chance of being accessed in response to each of q queries. If the data can fit such a distribution, then we might believe that the total number of document accesses is randomly distributed among the documents. This, in turn, would indicate that all documents have the same random probability of "attracting" accesses, and there is no further point in trying to predict a unique probability of access for each document.

With the current data set, it was not possible to identify q, and it was therefore not possible to identify m. That is, with 6871 accesses, do these represent 6871 queries (trials) where each query chose one document to access (m=1)? Or does the data represent (say) 687 queries (trials) where each query chose ten documents to access (m=10)? The Binomial distribution differs slightly under these different scenarios. Because the software did not maintain any session awareness or attempt to identify an indivisible query session, it is not possible to answer this question. We therefore compare the data to many different possible Binomial distributions. Three are shown below. We try to fit the data to these Binomials.

Table 1 below shows the numbers of documents with 0,1,2, etc. accesses during the 9 weeks after that document's publication . The last two columns show the observed data for keyword search (6871 accesses) and for browsing (6867 accesses). The first three columns show three different Binomial distributions assuming 6871 accesses and 64,831 documents. In the first Binomial, q is just the total number of document accesses -- 6871; in the second, q is equal to half the number of accesses -- 3436; in the third, q=687. The table shows, for the hypothetical distributions and for the observed data, how many documents (frequency) had the different numbers of accesses (# accesses).

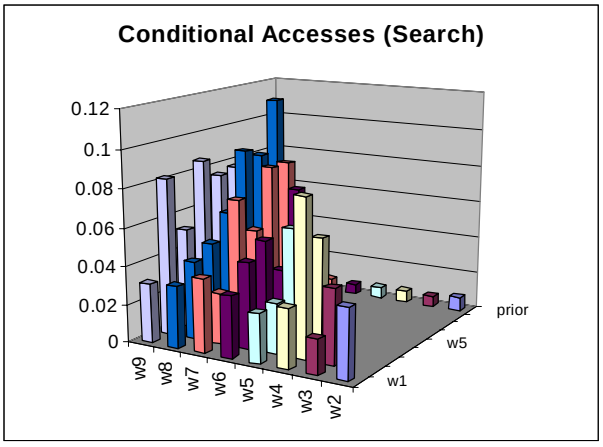
Comparing the observed data with the Binomial process, it is apparent that some documents are attracting more than their random (Binomial) share of attention, for both search and browsing services. If we can predict which documents these are, retrieval effectiveness may be improved. In addition, this predictability will inform us about user behavior in the two settings.

Convinced that the data does have some structure, and some documents have higher probabilities of access than others, we turn to the goal of predicting an a priori probability of access for each document. As described above in section 3, we chose a simple conditional probability as a prediction model to predict an individual document's relevance in time t from relevance in times $t-1$, $t-2$, etc.

Table 1	Binomial #1	Binomial #2	Binomial #3	search data	browse data
# Accesses	*Frequency p=1/64831, 6871 trials*	*Frequency p=2/64831, 6871/2 trials*	*Frequency p=10/64831 6871/10 trials*	*Observed frequency (6871 total)*	*observed frequency (6867 total)*
0	52492	52645	57321	61148	61316
1	11071	10977	7071	2398	2279
2	1178	1123	420	607	596
3	86	79	18	303	282
4	4	6	1	123	92
5	0	1	0	90	52
6	0	0	0	40	30
7	0	0	0	33	28
8	0	0	0	40	53
9	0	0	0	15	25
10	0	0	0	4	18
More	0	0	0	30	60

This raises the question of how to divide the data into "periods". It is possible to search for those "bins" that result in the strongest predictive ability. This could easily be achieved with data mining or other search tools. But we might then find meaningless patterns such as "if a document is retrieved in during days 12-18 and again during days 28-34 following publication, it has a .2 probability of being retrieved again during days 45-60". We did not want to unleash a blind search on this data, and felt that more could be accomplished by a more holistic approach at this preliminary stage. We simply divided accesses into week-long periods following publication of a document. The question then was whether accesses in week *n* could be predicted on the basis of accesses in weeks *n-1, n-2, …1*.

Figure 2

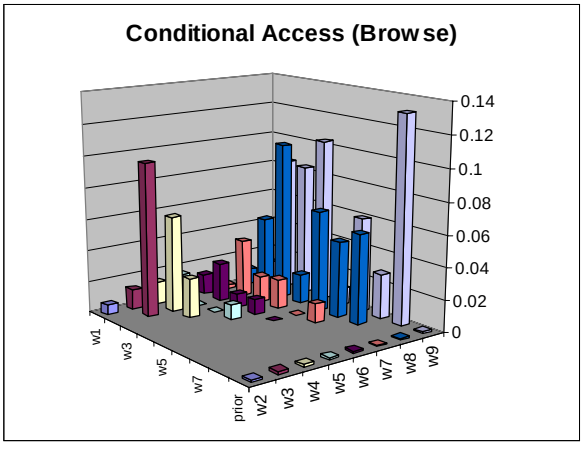


7 Results

Our results show that in general, the conditional probabilities are very much (2-20 *times*) stronger than the prior probabilities, so accesses in earlier weeks can predict accesses in later weeks. Since each of these probabilities is derived from a sum of many binary random variables (one for each document), we were able to assume a normal approximation, and we tested for the difference of means between the prior and conditional probabilities. Almost every conditional probability was significantly different from the prior at alpha=.01.

A more interesting question regards the varying strength of these conditional probabilities over time. We calculated the full (lower) matrix that represents the probability of $week_m$ conditional on $week_n$, with $n<m$. We did this for both the keyword search accesses and the browsing accesses. The results are strikingly different and are shown below in figures 2,3:

Figure 3

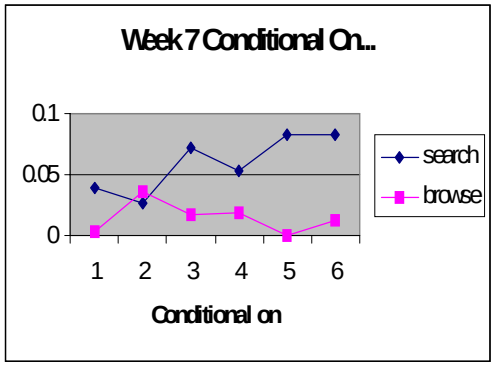


Each figure shows the probability of access in weeks 2-9, conditional on access in weeks 1-8, as well as an unconditional priori probability. For keyword search, retrievals in week *n* are most highly predicted by retrievals in immediately preceding weeks *n-1*, *n*-2, and less highly predicted by periods that are farther back in time. Also, it is found that in general, this prediction from week *n-1* to week *n* is stronger for high *n*. The overall sloping of this graph is clear: stronger conditional predictions are made along the diagonal, and with older documents.

For browsing, the picture is nearly opposite! This graph is visually clear only if the axes are reversed and the viewpoint is rotated as in figure 3. Now, with the exception of week 8 predicting week 9, in all other cases the strongest predictions of week *n* are the document's accesses in relatively *earlier* weeks, not its accesses in nearby weeks *n-1*, *n-2*. For every one of weeks 3-9, the strongest predictors are weeks 2 or 3, with the predictive ability diminished after that. It appears that the predictive power decreases after weeks 2-3, then rise again in weeks n-1, n-2; but the conditional probability does not rise again to the predictive levels of the early weeks (for week 9, week 8 does emerge as a better predictor than the early weeks. This 8-9 pair is the only exception to the pattern). This pattern is in direct contrast to keyword search. This graph is also different from the keyword case in the second respect, i.e. for browsing, the predictions are not stronger for older documents, as they were in the keyword search case. For the browsing data, this pattern is not clear.

A slice of the 3-dimensional graph for week 7, conditional on weeks 1-6, shows the difference in the trend over time, for searching as opposed to browsing.

Figure 4



Three points should be made that add further support to these initial findings. First, the numbers of data points are not small. Each of these conditional probabilities is a fraction with a numerator of about 30 and a denominator of about 500. So the "coverage" of these rules is not insignificant. Second, it should be reiterated that calendar dates and related environmental conditions have no bearing whatever on these trends, since the "week" numbers are relative to the publication date of each individual document. All these documents became 2 or 3 or *n* weeks old at very different times, so that peculiarities in the environment at any point in absolute time do not have any impact on this data. Lastly, we note that the search interface defaults to one month's history for keyword searches and the current (1) day for tracking/browsing service. Data analysis showed that the conditional probabilities showed no patterns that we could sensibly attribute to these artifacts.

Regarding statistical tests, more work is needed to identify a test that can measure the 3-dimensional trends that are captured in the graphs. We have 3-dimensional surfaces under two (keyword versus browsing) conditions. More work is required to identify a statistical test that can measure the difference in these shapes. As an intermediate goal we would like to identify a test for the difference between two corresponding rows of data -- for example, the two lines in figure 4.

8 Discussion

What is the meaning of the opposing trends in keyword versus browsing data? We believe that the results can be understood in hindsight, and hope that useful hypotheses can be garnered for further testing. The apparent explanation for the search data, is that a newspaper document that is downloaded even many weeks after publication is apparently one of some lasting value, a "good" document, while an article that is read in the early days of its publication may be read just because it's "news". Thus, it seems plausible (and in fact we informally expected as much) that later accesses would be able to predict still-later accesses, while early accesses would not be terribly meaningful.

The tracking service results were not expected, but an explanation is offered here for future testing. The tracking service is used for current awareness browsing. In general, this use is highly time-sensitive and older documents would not likely be accessed as often as new ones. Indeed, as indicated in the beginning of section 6, the data supports this trend. The question is whether accesses of an older document (late accesses) indicate a "good" document that will be re-accessed, as in the keyword case. The data shows that the strongest effect is quite the opposite. Early accesses (i.e. of young documents) are the better predictors of later accesses. In hindsight, we understand from the data that a "good" or favorite document for keyword search is not necessarily a good document that will be re-accessed for current

awareness browsing. The way to tell whether a document is "good" for users of keyword search is whether it has been accessed recently, after it was no longer "news". On the other hand, the way to tell if a document is good for current awareness users, is whether the document was frequently accessed when it was *new*. That indicates that the document may continue to be a useful one for those trying to maintain current awareness. Accesses at later dates may indicate something positive about the document, but not that it is highly useful in the context of current awareness services. This explanation is perhaps not entirely convincing; the very different graphs reported here require further investigation.

Regarding the generalizability of these results, we note that the database we studied was for newspaper articles. The half-life of these articles is much shorter than the half-life of (say) scholarly works. It is conceivable that some of the particular trends reported here are relevant only to a database of "newsy" documents. Nevertheless, we note that the result under discussion shows a *difference* (from keyword to browsing) in the *differences* (from conditioning a probability on $week_i$ as compared to $week_j$) of *conditional* probabilities. While it is true that newsy documents have a short half-life, this does not explain why the *conditional* probability on $week_1$ should be *higher or lower* than the conditional on $week_{t-1}$; and it certainly does not explain why this difference over time should be opposite in the browsing versus keywords cases. Still, a deeper understanding is necessary of the interactions between the type of document -- from news to eternal works -- and the type of user -- those interested primarily in news versus those interested in eternal works.

Conclusion

In this study we investigate the feasibility of predicting which documents will be accessed, based on each individual document's previous history of accesses. We find that this is indeed feasible, as there is structure in the distribution of access across time and across documents. We used a simple model of conditional probability for individual documents, and found encouraging results. Most interestingly, we found that the best predictors are essentially opposite in the case of keyword search versus tracking/browsing. In the case of keyword search, recent accesses are more accurate predictors, while for browsing, early (2-3 weeks after publication) accesses are more highly predictive of later access. These results shed light on the significant differences in users of different IR-related services. As well, these results may indicate a method for improving retrieval effectiveness by considering a document's probability of access P(D), in addition to its probability of relevance to a given query P(D|Q).

Acknowledgement

The authors would like to express their gratitude to Gabrielle Wong, Reference Librarian, Hong Kong University of Science and Technology

References

Chen, Hsinchun, A. Houston, R. Sewell, B. Schatz, "Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques", *Journal of the American Society for Information Science* (49:7), 1998, pp. 582-603

Fuhr, N. and Buckley, C. "A Probabilistic Learning Approach for Document Indexing," *ACM Transactions on Information Systems* (9:3), July 1991, pp. 223-248.

Gey, F.C., Chen, A., He, J. and Meggs, J. "Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections," *Proceedings of TREC-4*, Gaithersburg, MD, Department of Commerce, National Institute of Standards and Technology

Gordon, Michael D. "It's 10 A.M. Do You Know Where Your Documents are? The Nature and Scope of Information Retrival Problems in Business," *Information Processing & Management* (33:1), 1997, pp. 107-121.

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., and Frank, E. "Improving Browsing in Digital Libraries with Keyphrase Indexes," *Decision Support Systems*, 27, 1999, pp. 81-104.

Marchionini, G. "Research and Development in Digital Libraries," see *ils.unc.edu/~march/gigital_library-R-and_D_html*, 1997.

Turtle, H. and Croft, W.B. "Evaluation of an Inference Network-Based Retrieval Model," *ACM Transactions on Information Systems* (9:3), July 1991, pp. 187-222.

Vandenbosch, B. and Huff, S.L. "Searching and Scanning: How Executives Obtain Information from Executive Information Systems," *MIS Quarterly*, March 1997, pp. 81-105.

Van Rijsbergen, C.J. "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval," *Journal of Documentation* (33:2), June 1997, pp. 106-119