LAMP-TR-145 July 2007
CS-TR-4877
UMIACS-TR-2007-36
HCIL-2007-10

Exploring the Effectiveness of Related Article Search in PubMed

Jimmy Lin[†‡], Michael DiCuccio[‡],
Vahan Grigoryan[‡], and W. John Wilbur[‡]

[†]College of Information Studies
University of Maryland
College Park, Maryland, USA
E-mail: jimmylin@umd.edu

[‡]National Center for Biotechnology Information
National Library of Medicine
Bethesda, Maryland, USA
E-mail: {dicuccio,grigoryv,wilbur}@ncbi.nlm.nih.gov

Abstract

We describe two complementary studies that explore the effectiveness of related article search in PubMed. The first attempts to characterize the topological properties of document networks that are implicitly defined by this capability. The second focuses on analysis of PubMed query logs to gain an understanding of real user behavior. Combined evidence suggests that related article search is both a useful and often exploited feature in PubMed.

Publication Date: July 13, 2007
Keywords: biomedical domain, TREC genomics track, cluster hypothesis, visualization, log analysis

Please cite as: Jimmy Lin, Michael DiCuccio, Vahan Grigoryan, and W. John Wilbur. Exploring the Effectiveness of Related Article Search in PubMed. Technical Report LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10, University of Maryland, College Park, July 2007.

1 Introduction

Text mining in the biomedical domain is seen as an important component of modern genomics research. Alongside curated databases such as Entrez Gene and Online Mendelian Inheritance in Man (OMIM), collections of scientific articles such as MEDLINE serve as indispensable tools for discovery. It is one major goal of bioinformatics researchers to develop techniques, algorithms, and systems that support such breakthroughs.

Hearst defines text mining as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." (Hearst, 2003) A key capability is the linking of extracted elements to form new facts or new hypotheses. Our work in the context of the PubMed search engine is certainly in this spirit.

A recently-revised functionality in PubMed is the related article search feature (see Figure 1). When the user examines a MEDLINE abstract, the right panel of the browser is automatically populated with titles of articles that may also be of interest, as determined by a probabilistic content similarity algorithm (Wilbur, 2005) based primarily on abstract text—these are the results of an implicit related article search. Our goal is to unobtrusively suggest other interesting items (articles initially, but pointers to genes, proteins, sequences, etc. are also possibilities) to facilitate knowledge discovery and the linking together of otherwise unrelated facts. This feature supports a qualitatively different approach to exploring large document collections. The ability to interleave exploration-focused browsing with traditional query-focused access enhances a scientist's ability to glean useful information from free-text databases.

This paper explores the following question: to what extent is the related article search feature in PubMed useful for scientists? As a first step in answering this complex question, we present results from two complementary studies. Section 2 focuses on topological characteristics of the document networks that are implicitly defined by these links. Section 3 analyzes the behavior of real PubMed users with respect to this functionality.

2 Networks of Related Articles

The current PubMed interface for displaying MEDLINE abstracts includes links to five related articles. These articles are in turn connected to others via similar links. Together, these connections define a document network in which the nodes represent MEDLINE articles and the links reflect content similarity (computed primarily from abstract text). A user browsing related articles is implicitly traversing the information space defined by these networks; cf. (Wilbur and Coffee, 1994; Hearst and Pedersen, 1996; Smucker and Allan, 2006). We can better understand the usefulness of this capability by studying their topological characteristics. For example:

- How densely or sparsely connected are these networks?
- Do they contain disconnected components, which would represent potentially unrelated bodies of literature?

In this section, we present an initial exploration of these questions using an existing network visualization tool.

2.1 Test Collection

Ultimately, we would like to characterize the distribution of "interesting" articles in related document networks. However, we face the dual challenge of more concretely defining what "interesting" means and finding corresponding human judgments. As an alternative, we employed the test collection developed

Figure 1: Typical screenshot of PubMed as the user examines an abstract. The "Related Links" panel on the right is populated with titles that may also be of interest.

in the genomics track (Hersh et al., 2005) at the 2005 Text Retrieval Conference (TREC), a yearly evaluation forum that brings together dozens of research groups from around the world to work on shared IR tasks.

An IR test collection consists of three major components: a corpus of documents, a number of information needs (called "topics", in TREC parlance), and relevance judgments, which specify the relevance of documents with respect to information needs. By studying the distribution of relevant documents in the related documents networks, we hope to infer the distribution of interesting documents, since relevance certainly defines one aspect of what it means to be interesting.

The genomics track employed a ten-year subset of the MEDLINE database (1994–2003), which totals 4.6 million citations. Each citation is identified by a globally unique pmid. Information needs were captured by generic topic templates (GTTs), consisting of semantic types (e.g., genes) embedded in prototypical questions, as determined from interviews with real biologists. In total, five templates were developed, with ten fully-instantiated topics for each—examples are shown in Table 1.

2.2 Methods

For convenience, our experiments employed the *bm25* ranking algorithm as implemented in the Lemur toolkit[1] instead of the actual algorithm deployed in PubMed. Since both retrieval models basically treat the distribution of elite terms in documents as Poissons, and the performance of both methods are shown to be comparable in separate experiments (Lin and Wilbur, 2007 in preparation), the use of *bm25* served as a expedient.

We started by processing the list of relevant documents (reldocs) from the TREC 2005 genomics track. For each relevant document, we retrieved the top five related pmids using the entire abstract text

[1] http://www.lemurproject.org/

#1	**Information describing standard [methods or protocols] for doing some sort of experiment or procedure.** *methods or protocols:* purification of rat IgM
#2	**Information describing the role(s) of a [gene] involved in a [disease].** *gene:* PRNP *disease:* Mad Cow Disease
#3	**Information describing the role of a [gene] in a specific [biological process].** *gene:* casein kinase II *biological process:* ribosome assembly
#4	**Information describing interactions between two or more [genes] in the [function of an organ] or in a [disease].** *genes:* Ret and GDNF *function of an organ:* kidney development
#5	**Information describing one or more [mutations] of a given [gene] and its [biological impact or role].** *gene with mutation:* hypocretin receptor 2 *biological impact:* narcolepsy

Table 1: Templates and sample instantiations from the TREC 2005 genomics track.

as the query. Together, the results define the related document network for a particular topic: the pmids are the nodes and a directional link connects each "query" pmid to one result.[2] Each related document network defines the information space for a particular information need, showing connections from relevant documents. By examining the topology of these networks, we can characterize the effectiveness of related article search in PubMed.

Analysis was conducted using *SocialAction* (Perer and Shneiderman, 2006),[3] a network visualization tool developed at the University of Maryland. The tool was originally developed for analyzing social networks, although we have adapted it to our task. *SocialAction* provides two major capabilities: First, a graphical component allows us to visually inspect the document networks and form qualitative impressions about their structures. Second, the tool is able to compute statistics to support formal characterization.

2.3 Results

Of the 50 topics from the TREC 2005 genomics test collection, we eliminated four—one because it had no known relevant documents and three due to glitches with the analysis tool. For all results reported in this section, we focus on the 46 remaining topics. The distribution of topics with different numbers of relevant documents is shown in Table 2. As can be seen, a large number of topics have few relevant documents, suggesting that they represent difficult information needs.

One obvious way to characterize the effectiveness of related article search is by the expected number of relevant links that appear to the user when examining a relevant abstract. That is, on average, how many of the five suggested articles are relevant? This analysis is shown as a scatterplot in Figure 2. Each point represents a topic; the x-axis shows the number of relevant documents for that topic and the y-axis shows the average number of relevant articles suggested (out of five). Fitting a regression line to the plot yields an R^2 value of 0.47.

[2]Each link is associated with both a score and a rank (not presently used).

[3]http://www.cs.umd.edu/hcil/socialaction/

Bin	Number of Topics
< 25	17
$[25, 50)$	12
$[50, 75)$	6
$[75, 100)$	3
≥ 100	8

Table 2: Distribution of topics in terms of relevant document counts.

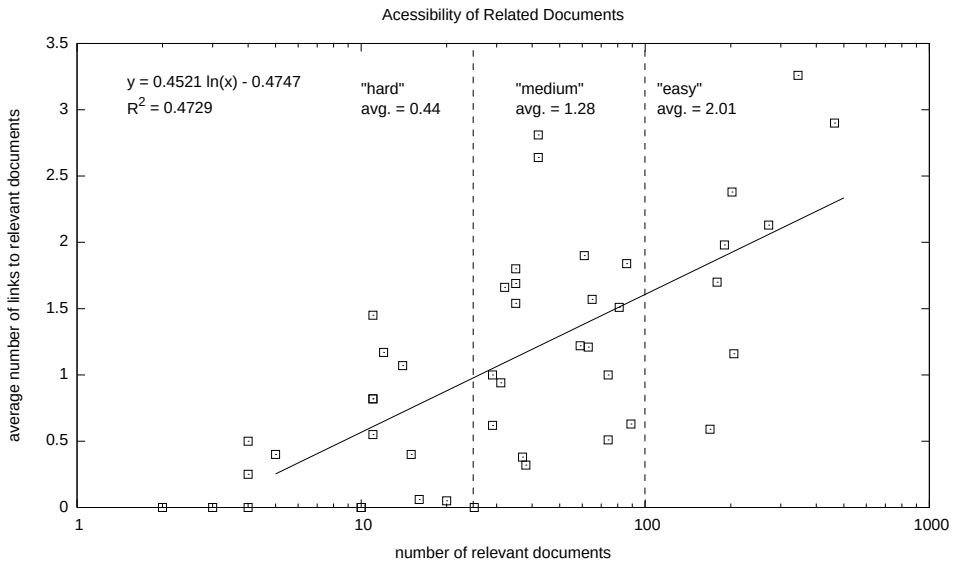

Figure 2: Scatterplot relating topics in terms of number of relevant documents to the average number of links to relevant related articles.

We arbitrarily divided topics into "hard", "medium", and "easy" based on the number of relevant documents (less than 25, between 25 and 100, and more than 100, respectively). The average number of relevant links for each of the categories is also shown in Figure 2. For medium topics, the user can expect to see 1.28 relevant titles in the related article panel while browsing; 0.44 for difficult topics and 2.01 for easy topics. These statistics demonstrate the potential usefulness of the related articles, since access to additional relevant MEDLINE articles comes at little cost and requires no explicit actions from the user.

A visualization of the document network for topic 132 "provide information about the genes APC (adenomatous polyposis coli) and wnt in colon cancer", which has 31 relevant documents, is shown in Figure 3. Relevant documents are shown in black and non-relevant documents are shown in light gray (the nodes are labeled with pmids, which are not readable but also not essential to our analysis)—the links represent the top five related pmids for each relevant pmid (ignoring both retrieval score and rank). This is a graphical representation of the information space for this information need—the links represent possible navigation paths for the user. This network structure is typical of topics with moderate numbers of relevant documents (the medium topics, as previously defined).

In Figure 3, we see a large, tightly connected cluster in which relevant documents share many links. This indicates that if a user were in this part of the information space, a large number of relevant abstracts would be potentially accessible via a single click. Since PubMed shows these titles in its interface (Figure 1), a user is directly alerted to additional relevant articles without requiring any

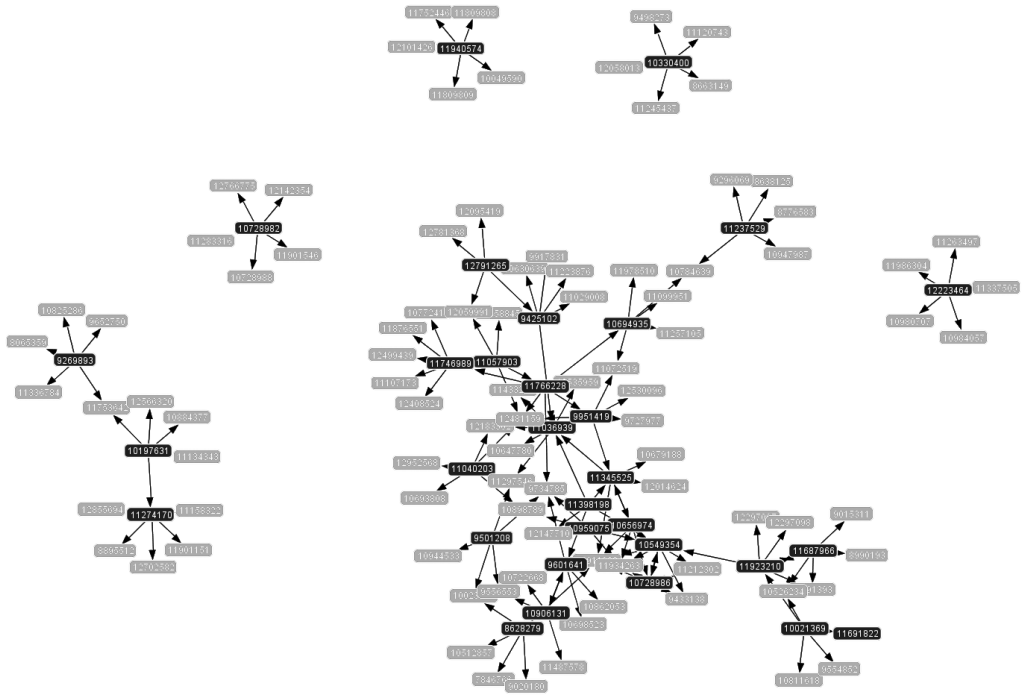

Figure 3: The document network for topic 132, which has 31 relevant documents (shown in black). The network has six components.

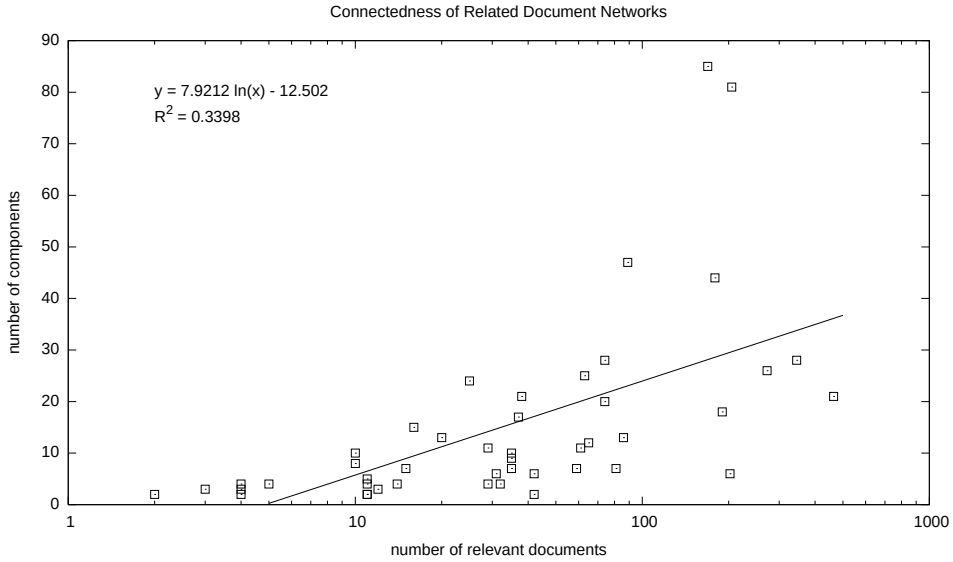

Figure 4: Scatterplot relating topics in terms of number of relevant documents to the number of components in its related document network.

explicit action. The related document network, in essence, allows us to visualize the effectiveness of single traversals of related article search.

In addition to the central cluster of documents, we note five additional components not connected to the large cluster.[4] Four of these have a distinctive star-like shape, with a relevant document in the center and non-relevant documents radiating outwards. This pattern indicates that a relevant document is isolated in the information space (at least via one-click traversals). In practical terms, these represent relevant documents that are more difficult to arrive at through browsing related articles alone.

One coarse-grained method for quantifying the effectiveness of related article search is by counting the total number of components in the network associated with each topic. A large number of components, relative to the number of relevant articles, is suggestive of "information islands", where relevant abstracts are rather dissimilar in content. The upshot is that a user browsing related links in this situation is less likely to encounter other relevant material.

Naturally, we would expect topics with fewer relevant documents to be more fragmented and topics with more relevant documents to be more densely connected, but what exactly is the nature of this relationship? An analysis is provided in Figure 4. Each point in the scatterplot represents a topic: x-axis shows the number of relevant documents in log scale, and the y-axis shows the number of components in its related document network. Fitting a regression line to the plot yields an R^2 value of 0.34. Very roughly, the number of components grows linearly with respect to the log of the number of relevant documents. The connectedness of the related document networks suggests that related article links can provide a useful tool for exploring document collections for information needs with sufficiently large numbers of relevant documents. For topics with few relevant documents, users may find themselves in isolated "information islands"; however, for such needs, formulating a good query may likewise be difficult.

[4] A caveat: there are cases in which two relevant documents are connected only through directional links into a non-relevant document. For simplicity, we still consider the entire structure "connected".

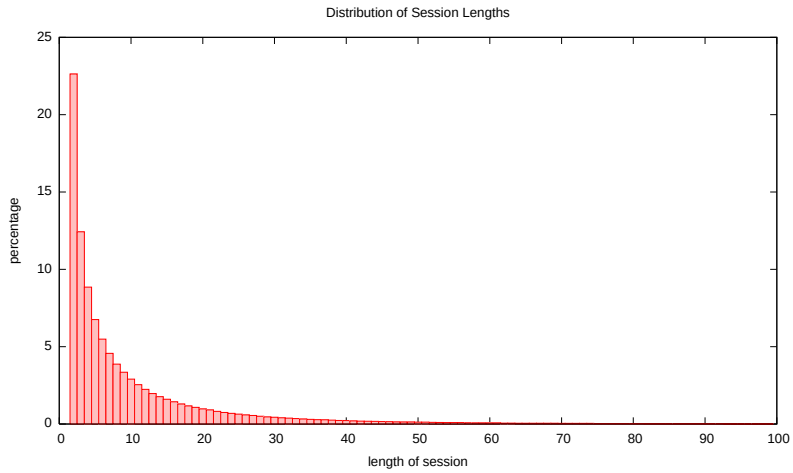

Figure 5: Distribution of sessions by length (ignoring those of length one).

3 Analysis of Real User Behavior

To complement our laboratory experiments with the TREC collection, we analyzed the behavior of real PubMed users by examining query logs collected by NCBI. Our goal was to understand how often and under what circumstances the related article search feature is invoked by real searchers.

3.1 Methods

We focused on a set of logs gathered between June 20 and June 27, 2007, which represents a typical week in terms of usage patterns. The basic unit of analysis is the session, which is accurately tracked through a browser cookie. Sessions are comprised of page views (CGI invocations). Note that our definition of a session is coarse-grained and may contain long periods of user inactivity (we currently do not perform any temporal segmentation). In addition, a user who engages PubMed with multiple browser windows (or tabs) will show up in our logs as a single session, since there is no effective way to separate the source of the CGI requests. In this data set, we discarded all sessions longer than 100 page views. This eliminates an insignificant fraction of sessions and partially addresses the skew in certain statistics caused by public computer installations.

The logs contain a wealth of information, including timestamp and details of the CGI invocation, which allows us to reconstruct with reasonable accuracy the actions of a particular user. Certain client-side actions, such as use of the browser "back" button, are not captured, although it is possible to infer some of these behaviors.

3.2 Results

In one week, we observe 35,136,632 page views across 7,964,643 sessions. Of those sessions, 62.8% consisted of a single page view—most of these represent bots and direct access into MEDLINE (e.g., from an embedded link or another search engine). Although this accounts for a large portion of all traffic, we mostly disregard these sessions since they do not represent interactive information-seeking behavior (with the system). The distribution of sessions by length (after eliminating those of length one) is shown in Figure 5. Note the distribution is still heavily dominated by short sessions.

Of all sessions in our data set, 1,941,329 (24.4%) include at least one PubMed search query and view of an abstract—we believe that this figure roughly quantifies actual attempts at addressing information

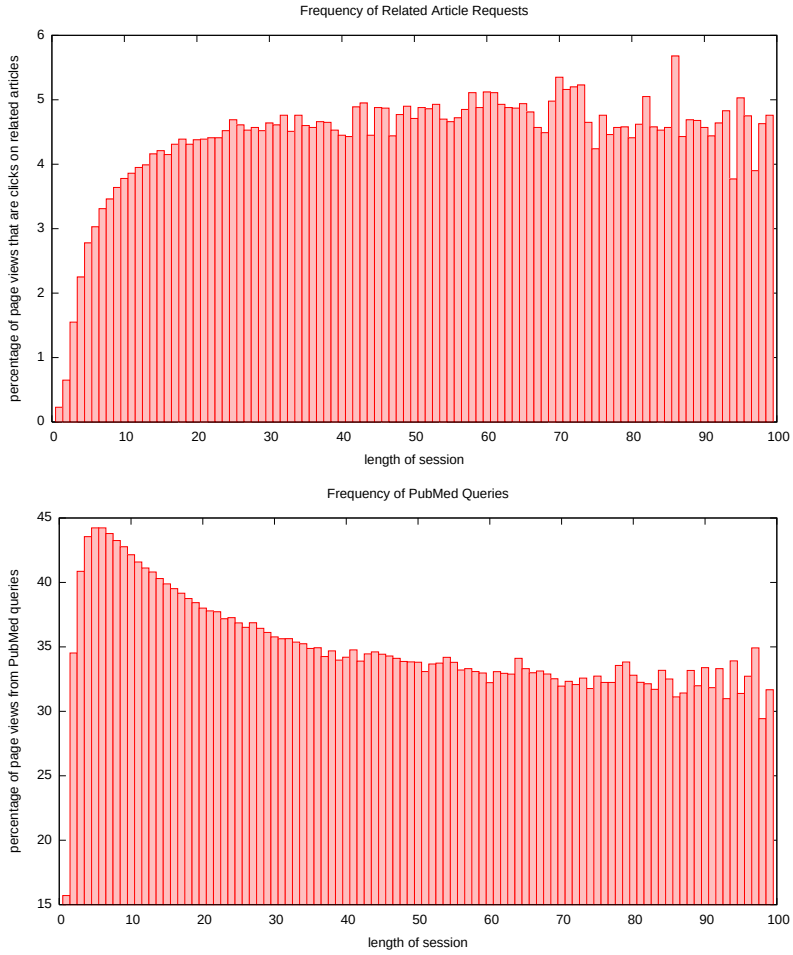

Figure 6: Fraction of page views that are clicks on related articles (top) and PubMed queries (bottom), binned by session length.

needs with PubMed. Of these sessions, 359,542 (18.5%) also include a click on a suggested related article. In other words, roughly a fifth of all non-trivial sessions involve examination of related articles. This figure provides a lower bound on usage, since session counts are dominated by short sessions and many of those represent situations where related article search is less applicable (e.g., known-item retrieval, which typically takes two page views).

Separately analyzing sessions of different lengths provides a more nuanced view of the log data. The percentage of all page views that are clicks on related articles (on the right panel in Figure 1) is shown on top in Figure 6. Disregarding shorter sessions, we see that they represent approximately five percent of page views. As a point of comparison, the same histogram for PubMed queries is shown on the bottom of Figure 6. Between a third and a half of page views are actual search queries. These statistics suggest that related articles are frequently found to be useful—or at least interesting enough to attract the user's attention.

Thus far, log data establishes usage statistics for the related article search feature in PubMed. However, is there evidence for users navigating the information space through multiple traversals of the related article links? We can begin to answer this question by analyzing subsequent actions of users *after* they have clicked on a related article link. This distribution is shown in Figure 7; see caption for more detailed description of what the bars represent. We see that once users begin browsing related

Figure 7: Distribution of next action after a user has click on a related article: clicking on another related article ("Related"), selecting another abstract to view via the browser "back" button ("Select"), issuing a new PubMed query ("Query"), performing other actions with PubMed ("Other PubMed"), examining all related articles ("All Related"), performing actions with other Entrez databases ("Other Entrez")

articles, they are likely to continue doing so (more than forty percent of the time)—more so than selecting another abstract to view (via the browser "back" button) or issuing a new query. These results lend credence to the network analysis performed in the previous section.

In summary, analysis of PubMed logs confirms that related article search receives high sustained usage. Our results provide at least indirect evidence that this capability is effective for information-seeking tasks.

4 Conclusion

We have taken a two-pronged approach to exploring the effectiveness of related article search in the context of the PubMed search engine. Laboratory experiments with TREC test collections examine more abstract properties of the between-article links and the document network they define. In general, we see that related article links create dense clusters of relevance that make for a potentially fruitful browsing experience. These results are confirmed by a empirical study of real PubMed users. Query logs tell us that searchers do often click on related articles suggested by the system, and that it forms an integral part of their information-seeking experience.

With theoretical implications for information retrieval, results of our work appear to support the Cluster Hypothesis (van Rijsbergen, 1979), the simple idea that closely associated documents tend to be relevant to the same requests. In fact, this observation forms the underlying basis of *why* browsing related articles can be useful in the first place. Although previous attempts to exploit the cluster hypothesis for document indexing have largely been unsuccessfully; see, for example, the classic work of Voorhees (1985), our work suggests that interactive browsing of related documents on the retrieval end is a powerful capability, at least in the biomedical domain.

5 Acknowledgements

We are grateful to Adam Perer for support with the *SocialAction* software and Ben Shneiderman for valuable discussion. This work is supported by the Intramural Research Program of the NIH, National Library of Medicine. The first author would like to thank Esther and Kiri for their kind support.

References

Marti A. Hearst and Jan O. Pedersen. 1996. Reexaming the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*, pages 76–84, Zürich, Switzerland.

Marti Hearst. 2003. What is text mining?
available at http://www.ischool.berkeley.edu/~hearst/text-mining.html, accessed July 10, 2007.

William Hersh, Aaron Cohen, Jianji Yang, Ravi Bhupatiraju, Phoebe Roberts, and Marti Hearst. 2005. TREC 2005 genomics track overview. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, Maryland.

Jimmy Lin and W. John Wilbur. 2007, in preparation. A probabilistic topic-based model for related document search: Applications to the biomedical domain.

Adam Perer and Ben Shneiderman. 2006. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700.

Mark Smucker and James Allan. 2006. Find-Similar: Similarity browsing as a search tool. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 461–468, Seattle, Washington.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

Ellen Voorhees. 1985. The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1985)*, Montreal, Canada.

W. John Wilbur and Leona Coffee. 1994. The effectiveness of document neighboring in search enhancement. *Information Processing and Management*, 30(2):253–266.

W. John Wilbur. 2005. Modeling text retrieval in biomedicine. In Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, editors, *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*, pages 277–297. Springer, New York.