Information Processing and Management 44 (2008) 1771–1783

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Navigating information spaces: A case study of related article search in PubMed

Jimmy Lin [a,b,*], Michael DiCuccio [b], Vahan Grigoryan [b], W. John Wilbur [b]

[a] The iSchool, University of Maryland, College Park, MD, USA
[b] National Center for Biotechnology Information National Library of Medicine Bethesda, MD, USA

ARTICLE INFO

*Article history:*
Received 20 November 2007
Received in revised form 11 April 2008
Accepted 14 April 2008
Available online 12 June 2008

*Keywords:*
MEDLINE
TREC genomics
Browsing
Interactive IR

ABSTRACT

The concept of an "information space" provides a powerful metaphor for guiding the design of interactive retrieval systems. We present a case study of related article search, a browsing tool designed to help users navigate the information space defined by results of the PubMed® search engine. This feature leverages content-similarity links that tie MEDLINE® citations together in a vast document network. We examine the effectiveness of related article search from two perspectives: a topological analysis of networks generated from information needs represented in the TREC 2005 genomics track and a query log analysis of real PubMed users. Together, data suggest that related article search is a useful feature and that browsing related articles has become an integral part of how users interact with PubMed.

Published by Elsevier Ltd.

## 1. Introduction

The idea of an "information space" provides a powerful metaphor for designers of information retrieval systems. It reminds us that documents[1] do not exist in isolation, but are embedded in an interconnected tapestry spanned by semantic relationships. To give two simple examples: Web pages exist in a rich environment criss-crossed by hyperlinks; scholarly articles can be viewed as nodes in a vast citation network that comprise the body of knowledge of a particular discipline. The information space metaphor allows us to invoke physical imagery to describe topological features of virtual spaces: densely-connected "patches", isolated "islands", etc. More importantly, the metaphor allows us to think of information seeking as the process of navigating unfamiliar territory, thereby providing a model to guide the design of information retrieval systems. Indeed, a previous study has confirmed that users often invoke physical navigation metaphors when using the World Wide Web (Maglio & Matlock, 2003). Thus, it behooves researchers and developers to take this idea seriously.

For the most part, retrieval systems provide few aids to help users navigate the information space defined by their search results. The ubiquitous ranked list, the primary presentation device used by nearly all search engines today, treats each "hit" as if it were an isolated and independent entity. Since ranked list results are sorted based on the likelihood of relevance to the user's query (Robertson, 1977), important document relationships are often hidden, even if such features were used in the ranking algorithm itself (e.g., PageRank, Page, Brin, Motwani, & Terry, 1999). Metaphorically, the ranked list provides a "soda-straw" view onto different points in an information space – users see individual search results, but very little context in terms of the local neighborhoods in which the hits reside. For example, the user may not be immediately aware that result

* Corresponding author. Address: The iSchool, University of Maryland, College Park, MD 20742, USA.
*E-mail addresses:* jimmylin@umd.edu (J. Lin), dicuccio@ncbi.nlm.nih.gov (M. DiCuccio), grigoryv@ncbi.nlm.nih.gov (V. Grigoryan), wilbur@ncbi.nlm.nih.gov (W.J. Wilbur).

[1] Or passages, records, audio clips, pictures, etc. – whatever unit of retrieval a system operates on.

0306-4573/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.ipm.2008.04.002

two and result seven provide almost the same information, and hence are redundant. The user may not realize that result four and result five are both connected to the same document (for example, via a hyperlink or a citation), which turns out to be an authority on the particular topic of interest.

There is a substantial body of work on organizing and presenting search results. Clustering and classification techniques have been used to group together similar objects in many research prototypes (e.g., Hearst & Pedersen, 1996; Leuski & Allan, 2000; Pratt & Fagan, 2000; Dumais, Cutrell, & Chen, 2001, just to name a few). These methods provide valuable clues about the structure of a particular information space (see Section 6 for an overview of related work). Clusty[2] is an example of statistical clustering technology deployed at Web scale. However, we are unaware of any published evidence supporting its effectiveness, and indeed, Hearst has pointed out problems associated with such methods (Hearst, 2006). Elsewhere, we are encouraged by the development of other navigational aids in the Web environment, for example, the query suggestion and refinement feature implemented in many search engines today (Cui, Wen, Nie, & Ma, 2003). These and other techniques provide the user with a richer set of tools for navigating information spaces to solve complex problems.

This work explores one conception of an information space in the biomedical domain – as a large network of documents connected by content-similarity links (what we refer to as related document networks).[3] In this view, the "local neighborhood" around each document consists of other documents that are similar in content. We specifically focus on the following questions:

- What does the information space defined by content-similarity links look like? That is, what are the topological features of related document networks?
- How are topological features of related document networks relevant to information seeking?
- Is the ability to navigate the information space via these content-similarity links useful for information seeking?

This article is organized as follows: We first begin with an overview of the related article search feature in PubMed. Section 3 describes the TREC 2005 genomics test collection used in our experiments and the methods for constructing the related article networks. Section 4 presents results from our analysis of these networks, characterizing both their topology and the accessibility of relevant documents via browsing. To complement this analysis, we examine user transaction logs from the PubMed search engine to understand how real users behave – these results are presented in Section 5. Before concluding, we discuss related work in Section 6.

## 2. Background

The context for this work is MEDLINE, the authoritative repository of abstracts from the medical and biomedical primary literature maintained by the US National Library of Medicine (NLM). As of 2007, MEDLINE contains over 17 million records (called citations), each of which includes bibliographic information, abstract text, as well as links to the full text (if available).

PubMed, NLMs public gateway to MEDLINE, provides a related article search feature to help users browse the literature. We have previously developed a probabilistic content-similarity algorithm (Wilbur, 2005; Lin & Wilbur, 2007) that, given a MEDLINE citation as a "query", retrieves a ranked list of other MEDLINE citations that might also be of interest to the user.[4] This functionality is invoked whenever the user examines a MEDLINE abstract: the right panel of the browser is automatically populated with results of this related article search (see Fig. 1). The goal is to unobtrusively suggest other interesting items (currently, MEDLINE articles, 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. In most retrieval systems, users' interactions center around issuing queries and examining search results. The ability to interleave exploration-focused browsing of related articles with traditional query-focused access enhances a user's ability to glean useful information from free-text collections.

The current PubMed interface displays links to five related articles. These articles are in turn connected to others via links of the same type. Together, these connections define a vast document network in which the nodes represent MEDLINE citations and the links reflect content-similarity. Each implicit invocation of the related article search provides the user with a view of the local neighborhood in the information space. The user is presented with articles that are "nearby", i.e., similar in content, and clicking on the related links moves the user through this environment, i.e., traversing the information space defined by these networks.

The theory of effective view navigation (Furnas, 1997) provides a framework for formally articulating the functionality provided by related article search. Furnas details his theory in terms of two types of graphs: a logical graph, which represents how information objects are truly connected, and a view graph, which represents local structure visible to the user at a particular point in time. The set of all content-similarity links between abstracts defines the logical graph of MEDLINE. The local view graph consists of the current abstract linked to five other abstracts that are most similar in terms of content. We begin with theoretical evidence that supports the usefulness of this browsing tool.

[2] http://clusty.com/.

[3] In this article, we use "document" and "article" interchangeably.

[4] Although MEDLINE citations contain only abstract text and associated bibliographic information, PubMed provides access to the full text articles (if available). Thus, it is not inaccurate to speak of searching for articles, even though the search itself is only performed on information in MEDLINE.


**Fig. 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.

The cluster hypothesis (van Rijsbergen, 1979), an idea in information retrieval that dates back several decades, is the simple observation that closely associated documents tend to be relevant to the same requests. That is, relevant documents tend to be more similar to other relevant documents than they are to non-relevant documents. Voorhees (1985) demonstrated that different document collections exhibit this property to varying degrees. Historically, IR researchers have attempted to exploit this observation by indexing cluster representatives (e.g., centroids) instead of complete documents (Jardine & van Rijsbergen, 1971; Salton, 1971; Voorhees, 1985). The primary justification for adopting such approaches – to reduce computational and storage requirements – are for the most part no longer relevant today for everything other than Web-scale collections. Nevertheless, the cluster hypothesis remains a powerful generalization and an active area of research (Diaz, 2007; Liu & Croft, 2004). In addition, there has been work on cluster-based retrieval interfaces (Hearst & Pedersen, 1996); see (Hearst, 2006) for a recent discussion. Related article search represents another attempt to exploit this characteristic of document collections. Examining a MEDLINE abstract (as in Fig. 1) represents a conscious and deliberate action on the user's part (e.g., selection from a list of results), which suggests some level of interest in the citation. In these situations, the cluster hypothesis predicts that similar documents are more likely to be relevant (i.e., "nearby", in our conception of information space).

Similar theoretical support comes from information foraging theory (Pirolli & Card, 1999), which hypothesizes that, when feasible, natural information systems evolve toward states that maximize gains of information per unit cost. Furthermore, the theory claims that information seekers behave in a manner that is not unlike our hunter–gatherer ancestors foraging in physical space. One basic assumption in information foraging theory is the existence of information patches (note once again the physical metaphor) – the tendency for relevant information to cluster together. An information seeker's activities are divided between those that involve exploiting the current patch and those that involve searching for the next patch – thus, the user is constantly faced with the decision to pursue one or the other activity. Pirolli and Card introduce the concept of an information scent, or the perception of value, cost, or access of an information source based on proximal cues. Viewed within this theory, PubMed's related article search provides exactly these scent cues to help users make decisions about their foraging behavior. Users might decide to follow related article links (i.e., remain in the same information patch), or seek out entirely different locations (e.g., by issuing a new query). Information scent is similar to Furnas' notion of a "residue" that indicates the content of objects reachable by following links (Furnas, 1997).

This discussion helps to motivate a number of specific questions about the topology of related document networks. For example:

- How densely or sparsely connected are related document networks? Do they contain clusters or "information patches", as assumed by the cluster hypothesis and information foraging theory?

- What is the prevalence of disconnected components, representing isolated information patches? Are relevant documents "reachable" from each other?
- What is the distribution of relevant documents within this network? Are the findings consistent with the cluster hypothesis and the theory of information foraging?

These questions focus our exploration of related article search as deployed in PubMed and its underlying conception of an information space. We would like to quantify the effectiveness of this approach, but the natural question is: compared to what? This work is not a comparative evaluation of different content-similarity algorithms – see (Lin & Wilbur, 2007) for such an analysis. Neither is this work a comparative study between alternative conceptions of information spaces: PubMed's current implementation focuses only on content similarity as the underlying model for browsing the document collection. This work is best viewed as a case study focusing on one particular approach. Experimental evidence suggests that related article search is intrinsically effective, and that it has become an integral component of users' interactions with PubMed.

## 3. Methodology

Our experiments employed the test collection developed in the genomics track (Hersh et al., 2005) at the 2005 Text Retrieval Conference (TREC). TREC is a yearly evaluation forum, organized by the US National Institute of Standards and Technology (NIST), that brings together dozens of research groups from around the world to work on shared IR tasks. Different "tracks" at TREC focus on different retrieval problems, ranging from spam detection to question answering. For several years, text retrieval in the biomedical domain occupied one such track, the genomics track.

An IR test collection consists of three major components: a collection 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. The genomics track employed a ten-year subset of MEDLINE (1994–2003), which totals 4.6 million citations, or approximately a third of all citations at the time it was collected in 2004 (commonly called the MEDLINE04 collection). 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 biologists. In total, five templates were developed, with ten fully-instantiated topics for each; examples are shown in Table 1. In some cases, the actual topics deviated slightly from the template structure (in order to accommodate real requests). Note that GTTs differ from the topic structure in the standard ranked retrieval task, the so-called *ad hoc* retrieval task. Topics in the *ad hoc* task consist entirely of free text: a short "title", a sentence-long "description", and a paragraph-long "narrative".

Relevance judgments for the topics were gathered using the pooling methodology, the standard for many retrieval tasks (Harman, 2005). In total, 32 groups submitted 59 runs to NIST (both automatic runs and those that involved human intervention), which insured a rich, diverse pool of results. Relevance judgments were provided by an undergraduate student and a Ph.D. researcher in biology.

We used a simple approach to construct a related document network for each topic in the TREC 2005 genomics track test collection. For each relevant PMID, we retrieved its top five related PMIDs; together, these results define a directed graph where the nodes represent MEDLINE citations and the edges represent content-similarity links.[5] Retrieval of related articles was accomplished through the eutils API[6] provided by the National Center for Biotechnology Information (NCBI). To ensure that all nodes in the network were found within the MEDLINE04 collection, we placed date restrictions on the eutils query and discarded all PMIDs that did not appear in the collection.

In constructing the document networks, we specifically focused on the structure of the information space that the user is likely to traverse. Thus, we expanded related links from relevant documents only, since users are unlikely to pay attention to related links if the current abstract is not relevant. Another design choice merits discussion: we only performed one round of related link expansion from known relevant documents. Alternatively, one might perform the expansion multiple iterations, e.g., expand related articles of related articles, and so forth. There are several reasons why this possibility was rejected. First, as already discussed, users are unlikely to follow links that appear irrelevant (even if they lead to relevant documents later on, since the current PubMed interface does not provide "lookahead"). Metaphorically, users are generally unwilling to wander away from a relevant document unless the path leads to another relevant document. Second, multiple iterations of expansion from relevant documents significantly increase the size of the networks – which may result in unwieldy structures since we are interested in both statistical and visual analysis.

We used two separate toolkits in this study:

- *SocialAction* (Perer & Shneiderman, 2006),[7] a network visualization tool developed at the University of Maryland. Although the tool was originally developed for analyzing social networks, we 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 structure. Second, the tool is able to compute some statistics to support formal characterization.

[5] Note that edges are annotated with both a similarity score and a retrieved rank; these features are not used in our current experiments.
[6] http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.
[7] http://www.cs.umd.edu/hcil/socialaction/.

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

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

- *JUNG (Java Universal Network/Graph Framework)*,[8] an extensible open source toolkit for analyzing graphs.

We found the combination of visual and statistical analysis particularly powerful. SocialAction was useful as an exploratory tool and helped to generate a number of hypotheses about the data. JUNG supported processing of the related document networks in "batch mode", allowing us to rapidly extract interesting features that provide supporting evidence for our hypotheses.

## 4. Analysis of related article networks

Using the methods described in the previous section, we built a related document network for each of the topics in the TREC 2005 genomics track test collection. Topic 135, which contained no relevant documents, was discarded from the topic set – this yielded a total of 49 networks. This section begins with a general characterization of these networks and then focuses on the potential effectiveness of related article search.

### 4.1. Network characteristics

The distribution of topics by number of relevant documents is shown in Fig. 2. We see that most topics have fewer than forty relevant documents, but approximately a fifth of the topics have more than one hundred relevant documents. The tail of the distribution contains many outliers – the two largest topics contain 709 and 437 relevant documents, respectively.

We introduce the notion of "information density" to quantify the extent to which relevant documents cluster together. Networks with high information density are those that contain few, highly-connected clusters. Within a cluster, related article search can serve as an effective tool for gathering relevant information, since a user browsing the network is more likely to encounter relevant documents. Thus, the overall effectiveness of related article search depends on the information density of the underlying network. First, we present an informal analysis, followed by more rigorous statistical characterizations.

Based on visual inspection of all 49 related document networks using the SocialAction tool, we were able to identify three prototypes:

- A network with high information density, exemplified by topic 131, "Provide the genes L1 and L2 in the HPV11 virus in the role of L2 in the viral capsid".
- A network with moderate information density, exemplified by topic 121, "Provide information on the role of the gene BARD1 in the process of BRCA1 regulation".
- A network with low information density, exemplified by topic 129, "Provide information on the role of the gene Interferon-beta in the process of viral entry into host cell".

Visualizations of these networks are shown in Fig. 3. In all three, nodes are labeled with PMIDs – they are not readable, but the specific labels are irrelevant since our focus is on the overall topology of the network. Relevant PMIDs are colored in green, while irrelevant PMIDs are colored in tan. Arrows show the directionality of related article links. That is, possible user navigation paths flow along the arrows. In some cases, the arrows are bidirectional, indicating that the two PMIDs are on each other's top five related result list.

Table 2 provides key statistics for the three representative networks. All three topics visualized in Fig. 3 have roughly the same number of relevant documents (42, 42, and 38, respectively – shown in the column labeled #Rel). Thus, topological differences between the networks reflect the structure of the information space defined by the relevant documents. The final

[8] http://jung.sourceforge.net/.

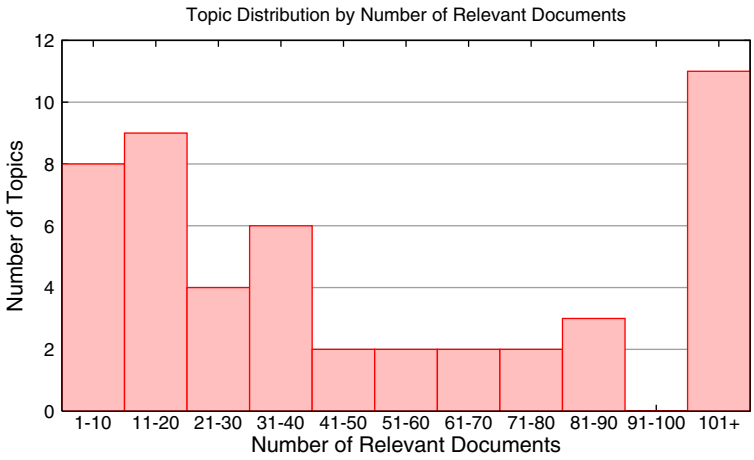


**Fig. 2.** Histogram showing the distribution of topics by number of relevant documents.

three columns of Table 2 show the total number of nodes in the network (#N), the total number of disconnected components (#C), and the percentage of nodes in the largest component (%ILC). As expected, networks with higher information density have fewer total nodes and a greater fraction of nodes in the largest component.

By construction, the information density of a document network reflects the extent to which the cluster hypothesis holds, i.e., how similar relevant documents are to each other. At one extreme, if relevant documents are randomly scattered in the document space, each relevant document is likely to form its own component, and the network should consist entirely of "stars" (a relevant article linked to its top five related articles). The bottom graph in Fig. 3 is an example of a network with low information density (although not as extreme as the case just described). We can invoke the metaphor of isolated "information islands", where users may encounter significant difficulty in moving from islands of relevant information to other islands. Paths between relevant documents must extend through irrelevant documents (which represent paths that users are unlikely to navigate through).

At the other end of the spectrum, if relevant documents are indeed clustered together, we should see a small number of large components. The top graph in Fig. 3 is an example of a network with high information density. The middle graph in Fig. 3 shows a network with moderate information density – a large, densely-connected component and a number of small, isolated components. We see that the cluster hypothesis holds to varying degrees with different topics, which is consistent with previous work. These results are also consistent with the assumptions made by information foraging theory: relevant documents are indeed distributed in patches, through which users can productively wander to gather relevant documents. However, it is apparent that the patch structure varies from topic to topic – in some cases, relevant documents are concentrated in a small number of "superpatches". The existence of these structures would compel the user to browse the neighborhood of relevant documents, since the expected marginal benefit may be higher than searching for the next patch.

Let us try to formalize this notion of information density. The clustering coefficient (Watts & Strogatz, 1998) is one common way to characterize the density of a network. The clustering coefficient quantifies the extent to which neighbors of a node are also neighbors of each other. This value is typically compared to the clustering coefficient of a comparable random graph (Erdös & Rényi, 1959), and a network is said to exhibit small-world characteristics if its clustering coefficient is much higher; cf. (Albert & Barabási, 2002). Table 3 shows this analysis applied to the three representative networks under consideration. Although the clustering coefficient is higher than one would expect from a comparable random graph, the small-world characteristic is far less pronounced than other networks that have been analyzed (for example, real-world social networks have clustering coefficients that are several orders of magnitude higher than random).

There are, however, two drawbacks to the clustering coefficient for measuring information density. First, it assumes an undirected graph, which is not true in our case (we return to this issue in Section 4.2). Second, it does not directly quantify the most important characteristic for information seeking: the presence of large connected components. The document network most ideal from an information seeking point of view would contain one single large connected component from which any relevant document could be reached from every other relevant document. It is possible for graphs with high clustering coefficients to have multiple disconnected components (if, for example, the individual components are themselves highly connected).

We propose that our notion of information density can be quantified by measuring the percentage of nodes in the largest component (%ILC). This is preferable to directly counting the number of components since that value is dependent on the number of relevant documents, which exhibits large variance (as can be seen from Fig. 2); see also (Smucker & Allan, 2007). Of course, the three graphs in Fig. 3 are merely representative samples from a continuous spectrum. Where does each of the 49 networks we analyzed fall along this continuum of information density? We plotted the distribution of topics according to this measure in Fig. 4. The %ILC metric was divided in ten equal-sized bins, and we show the number of

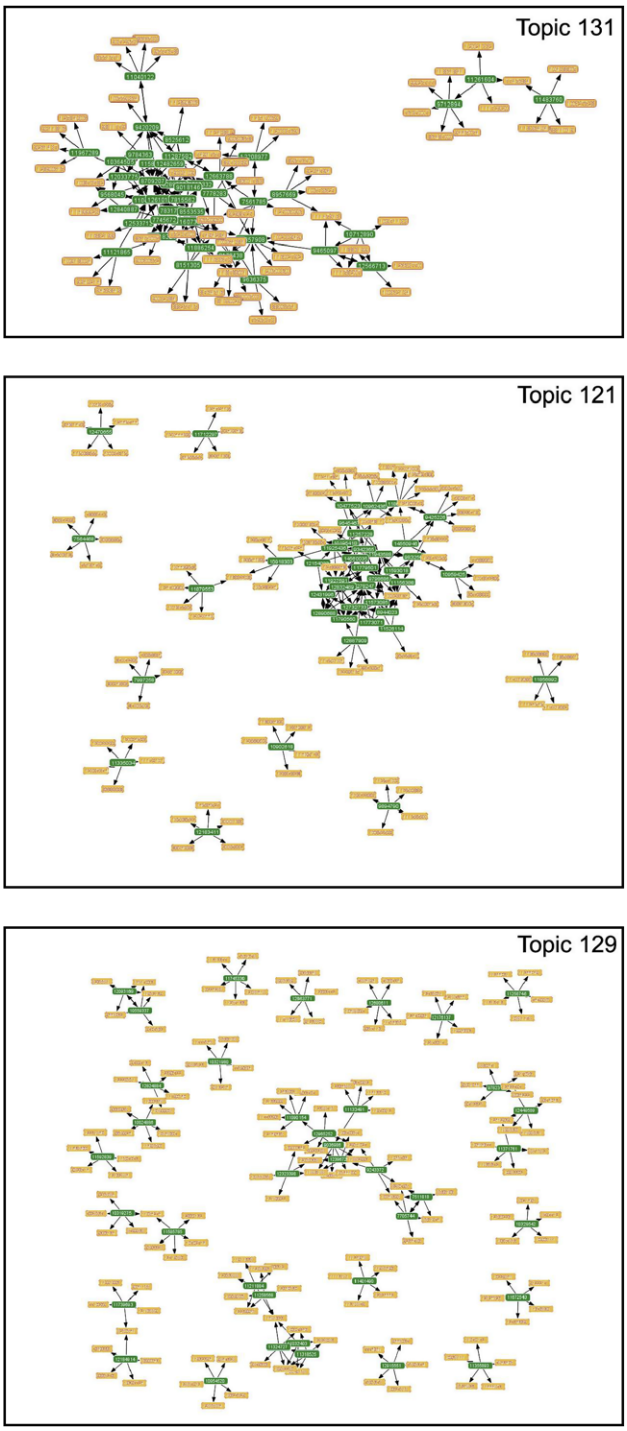


**Fig. 3.** Visualizations of three representative document networks; see Table 2 for key statistics.

document networks that fall into each bin. Although the thresholds are somewhat arbitrary, we considered networks >70%ILC to be high in information density, networks <30%ILC to be low in information density, and all other networks to have moderate information density.

The histogram suggests that the majority of the document networks are moderate to high in information density. This means that, at least with respect to the types of information needs represented in the TREC 2005 genomics track, related article search is potentially an effective tool for helping users navigate information spaces. That is, there are relatively few topics in which relevant documents are scattered in information islands.

**Table 2**
Key statistics for three representative document networks

| Density | Topic | #Rel | #N | #C | %ILC |
|---|---|---|---|---|---|
| High | 131 | 42 | 108 | 2 | 86 |
| Moderate | 121 | 42 | 129 | 10 | 58 |
| Low | 129 | 38 | 190 | 22 | 19 |

The last four columns list the number of known relevant documents (#Rel), total number of nodes in the entire network (#N), the total number of components (#C), and percentage of nodes in the largest component (%ILC).

**Table 3**
Clustering coefficient for the three representative document networks, compared to the clustering coefficient of a comparable random graph

| Density | Topic | CC | $CC_{random}$ |
|---|---|---|---|
| High | 131 | 0.1466 | 0.0363 |
| Moderate | 121 | 0.0766 | 0.0254 |
| Low | 129 | 0.0513 | 0.0106 |

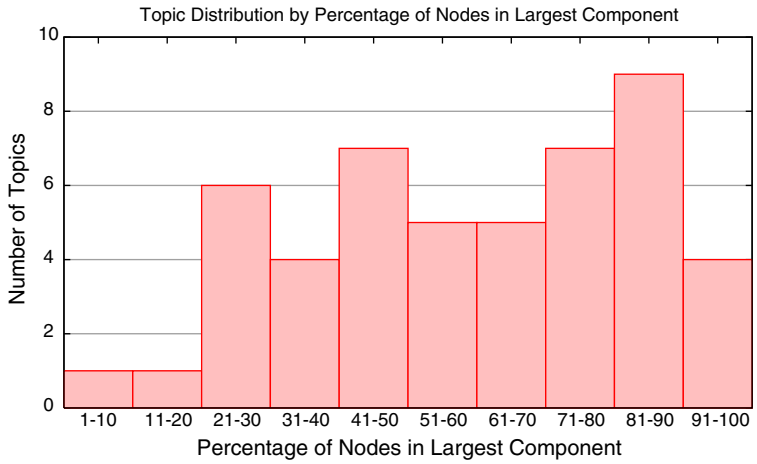


**Fig. 4.** Distribution of topics by percentage of nodes in the largest component. Percentages have been divided into ten bins, and the bar graph shows the number of networks that falls into each bin.

### 4.2. Access to relevant articles

Although information density characterizes the space defined by various information needs, we do not yet have a complete story. In particular, the existence of large clusters alone does not necessarily imply easy access to relevant articles via browsing related article links since there may not be navigable paths between nodes in the same component. Because related article links are directional, the components are *weakly connected*, in the graph-theoretic sense. Consider the simple case of an irrelevant article that is related to two relevant articles: although the three would form a connected component, there would be no navigable path from one relevant article to the other in our document network. Of course, there may in reality be paths between the two relevant articles, but they must go through irrelevant articles.

Results presented in this section attempt to directly characterize the effectiveness of related article search. One obvious method 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 is the same as computing precision at five (P5) on the related document search results, averaged over all known relevant documents for a particular topic.

The scatterplot in Fig. 5 shows this analysis. Each point represents a topic. The *y*-axis shows the mean precision at five for related article search, computed from the known relevant documents for a particular topic. The *x*-axis characterizes the related document network in terms of fraction of nodes in the largest component. Fitting a regression line to the plot yields an $R^2$ value of 0.509 (additional analysis reveals that the association between the two variables is significant, $p \ll 0.01$). As expected, the effectiveness of related article search increases with information density, which confirms the results of the previous section. Averaged across all topics, P5 is 22.4% – this means that for a randomly selected information need from the TREC 2005 genomics test collection, whenever the user encounters a relevant MEDLINE citation, PubMed related article search will suggest one other relevant article. Since related links are displayed unobtrusively in the PubMed interface, the user is essentially able to gather twice as many relevant documents from the same interface screen.

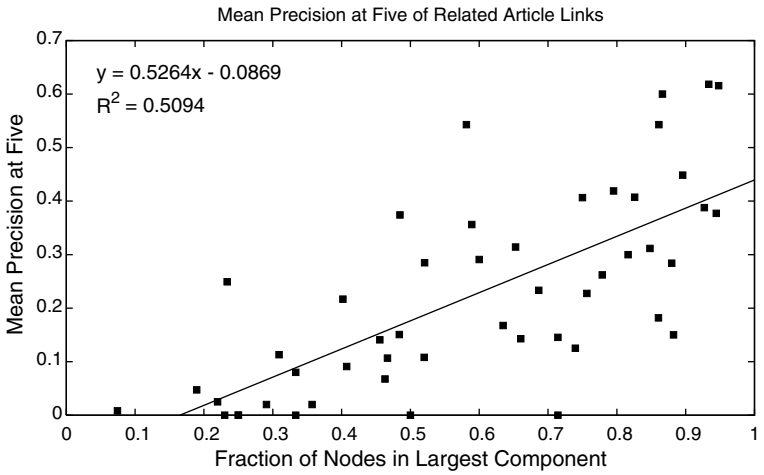


**Fig. 5.** Scatterplot relating the mean precision at five of related article links to the fraction of nodes in the largest component of that document network. Positive correlation suggests that as the information density increases, the effectiveness of related article search also increases.

It is perhaps worth stressing that these precision figures represent a conservative estimate, since relevance is but one factor that would compel a user to click on a related article link. An article not directly relevant to the user's present information need might nevertheless be interesting and worth pursuing, for example, because it touches on a connected topic or prompts the user to consider some new idea. Browsing related articles may be a useful tool for serendipitous knowledge discovery, but this is obviously very difficult to evaluate. Therefore, our study focused only on relevance, and thus is rather conservative in the more general problem of quantifying the degree to which a related link is worth clicking (for whatever reason).

This analysis, however, does not quantify the recall of related article search, which is another important measure of effectiveness, especially for scientists who are conducting in-depth research about a particular topic. In a separate experiment, we attempted to quantify residual recall in the following manner: starting from a relevant document, we compute the recall of *additional* relevant documents reachable via links in the document network. This gives us the *residual* recall (since we are explicitly discounting the initial relevant document), which quantifies the amount of relevant information accessible by browsing related article links (without encountering an irrelevant article). We can average this recall measure across all known relevant documents for a topic, and then compute the mean across all topics to arrive at the per-topic mean achievable residual recall (essentially, a macro-average). In Fig. 6, we plot this measure against the fraction of nodes in the largest component. A linear regression yields an $R^2$ value of 0.29; additional analysis shows that the association between the two variables is significant, $p \ll 0.01$. One downside of mean achievable residual recall is its sensitivity to the number of total relevant documents; topics with few relevant documents are greatly affected by idiosyncrasies in the document network. Nevertheless, we see a weak positive correlation between fraction of nodes in the largest component (and by extension, information density) and mean achievable residual recall.

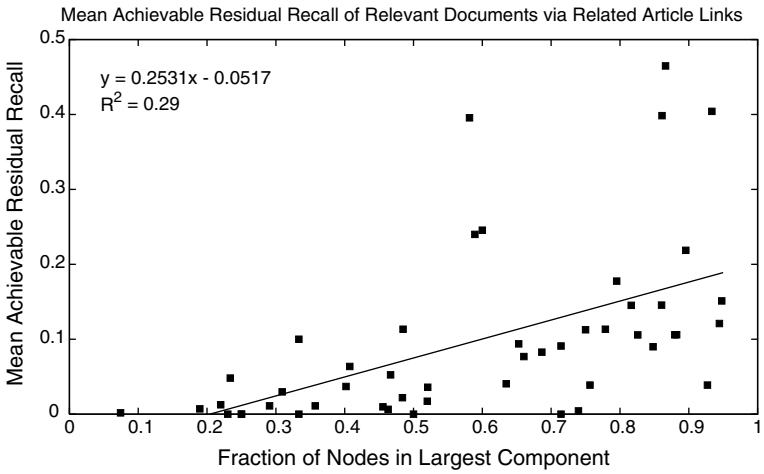


**Fig. 6.** Mean achievable residual recall by browsing the document network starting from a known relevant document. Each point represents an average across all relevant documents in a topic.

The mean achievable residual recall, averaged across all topics in the TREC 2005 genomics track, is 9.8%. That is, given a randomly-selected information need represented by the test collection, and starting from a known relevant document, it is expected that an *additional* 9.8% of relevant documents are reachable via traversals of related article links that do not go through any irrelevant documents. This translates into a mean of 14 documents; a median 2.0 documents. Thus, every time the user finds a relevant article, these numbers of additional relevant articles are easily accessible by browsing related links.

We conclude from this analysis that related article search as deployed in PubMed represents an effective information-seeking tool, at least for the types of information needs captured in the TREC 2005 genomics track test collection.

## 5. Analysis of real user behavior

To complement our laboratory experiments with the TREC collection, we analyzed the behavior of real PubMed users by examining transaction logs collected by NCBI, the unit within the US National Library of Medicine that is responsible for administering PubMed. Our goal was to understand how often and under what circumstances the related article search feature is invoked by real searchers. This work represents the first log analysis of PubMed related article search ever conducted.

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 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.

In one week, we observed 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 disregard these sessions since they do not represent interactive information-seeking behavior with PubMed. The distribution of sessions by length (after eliminating those of length one) is shown in Fig. 7. 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 provides an upper bound on actual attempts at addressing information 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 generated by clicks on related article links is shown on top in Fig. 8, sorted in terms of session length. As a point of comparison, the same histogram for PubMed queries is shown on the bottom of Fig. 8. Figures for both histograms become noisier with increasing session length due to the paucity of data for longer sessions, but the trends appear clear. As session lengths grow, users become more likely to follow related article links. The figure levels out at approximately five percent of

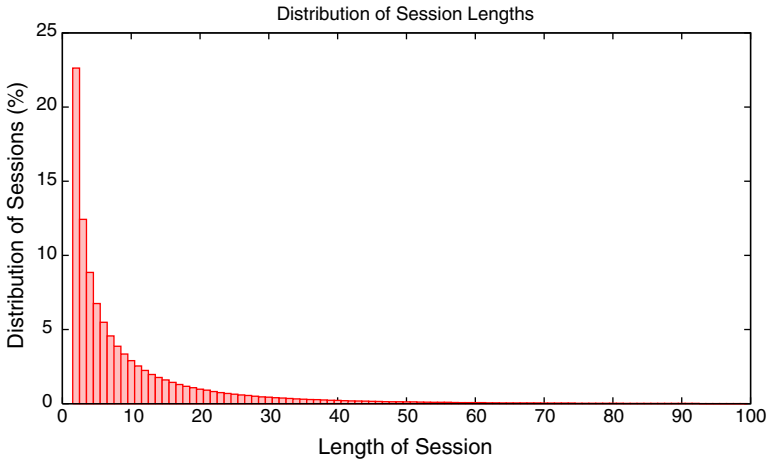


**Fig. 7.** Distribution of sessions by length (ignoring those of length one).

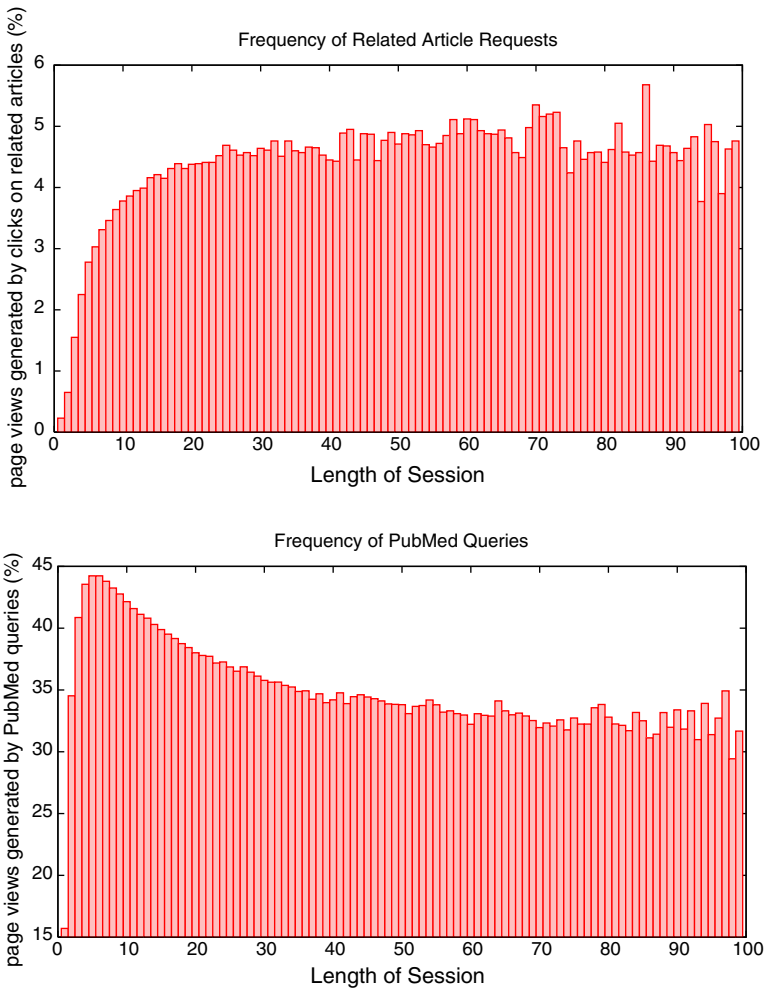


**Fig. 8.** Percentage of page views generated by clicks on related article links (top) and PubMed queries (bottom), binned by session length.

all page views for longer sessions. The frequency of search queries follows a different pattern – short sessions are dominated by queries (over 40%), but the fraction of page views generated by queries shrinks as sessions become longer, settling at a frequency of approximately one query for every three page views. These two distributions fit our intuitions: since issuing queries serves as the entry point to MEDLINE from the PubMed interface, it is no surprise that short sessions are dominated by querying. Longer sessions represent more involved search sessions, where users invoke a broader repertoire of information-seeking skills, e.g., perusing search results to better understand the topic, exploring related articles, etc. Therefore, as session length increases, the frequency of issuing queries naturally decreases. Based on this evidence, it is clear that related article search has become an integral part of searchers' interaction with PubMed. Combined with the results from analyzing related document networks, we suggest that content-similarity browsing is indeed an effective tool for information seeking.

Finally, we present evidence that users navigate through related document networks via multiple traversals of content-similarity links. We show this by analyzing subsequent actions of users *after* they have clicked on a related article link. This distribution is shown in Fig. 9; see caption for more detailed description of what the bars represent. We see that once users begin browsing related 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.

## 6. Related work

Despite superficial similarities between browsing related document networks and approaches to interactive retrieval based clustering and classification, we believe there are important differences. Clustering refers to the automatic grouping of documents by some measure of content similarity – Scatter/Gather is a classic example (Hearst & Pedersen, 1996). Although clustering does not by itself specify a visualization, it is possible to render the document groupings graphically

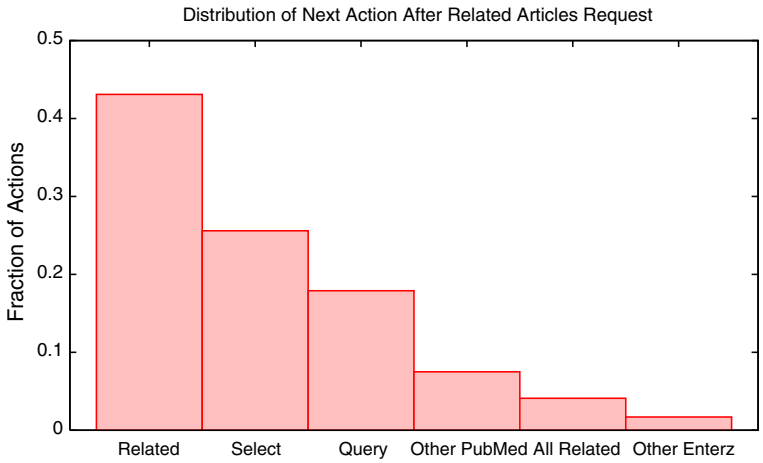


**Fig. 9.** Distribution of next action after a user has click on a related article link: clicking on another related article link ("Related"), selecting another abstract to view via the browser "back" button ("Select"), issuing a new PubMed query ("Query"), performing other actions with PubMed ("Other PubMed"), examining all related articles ("All Related"), performing actions with other Entrez databases ("Other Entrez").

in a search interface (e.g., Leuski & Allan, 2000). One downside of clustering approaches is that they often create groupings that are not semantically coherent; they often conflate many dimensions of document similarity since the algorithms typically operate in a high-dimensional feature space difficult for users to understand. Thus, users often find it difficult to see what the cluster is "about". This issue is exacerbated by the challenge of finding descriptive cluster labels, which is itself a difficult problem (Hearst, 2006). Put differently, the primary relationship conveyed by a cluster-based interface is that of group membership (i.e., a document is a member of the group defined by all documents in the cluster). In contrast, the primary relationship conveyed in related document networks is pairwise content similarity, which we believe is easier for users to comprehend, since they do not need to understand the entire contents of a cluster.

In contrast to clustering, classification approaches place search results into a finite set of pre-defined categories (e.g., Dumais, Cutrell, & Chen, 2001). Although categorized results are generally more meaningful, developing the category structure in advance requires knowledge engineering – although in the case of MEDLINE the existence of controlled vocabulary MeSH® terms provides a pre-existing set of categories that can be readily exploited (Demner-Fushman & Lin, 2006; Pratt & Fagan, 2000). Another downside of classification approaches is that a static category structure cannot adapt to different queries – it is difficult to convey differences between documents along dimensions in which there are no pre-existing categories. Once again, the focus on pairwise content similarity distinguishes our approach from previous work on classification, although one can imagine hybrid methods that take advantage of existing categories when available.

The idea of browsing related documents has previously been explored (Wilbur & Coffee, 1994; Smucker & Allan, 2006). However, these cited articles focus primarily on user simulations to explore the effectiveness of this browsing technique, whereas we specifically examine the topological characteristics of the document networks. Furthermore, to our knowledge, this work represents the first detailed query log analysis of related article search deployed in a large operational environment.

## 7. Conclusion

Within the domain of the medical and biomedical primary literature, this work presents a case study examining one conception of an information space – that which is defined by content-similarity links between documents in the collection. We focus on the effectiveness of related article search, a feature deployed in the PubMed search engine that allows users to easily traverse content-similarity links.

This study takes a two-pronged approach. Laboratory experiments with the TREC 2005 genomics track test collection allow us to characterize the topological structure of the networks through which users are likely to navigate. In most cases, we find dense information patches, representing regions in which many relevant documents are clustered closely together. Further analysis suggests that users are indeed able to gather many relevant documents via browsing. These results are confirmed by an empirical study of real PubMed users. Transaction logs reveal that searchers do often take advantage of related articles links, suggesting that this mode of browsing has become an integral part of users' interactions with PubMed.

## Acknowledgements

We are grateful to Adam Perer for support with the *SocialAction* software and Ben Shneiderman for valuable discussions. 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


Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47–97.

Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 829–839.

Demner-Fushman, D., & Lin, J. (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (COLING/ACL 2006)* (pp. 841–848). Sydney, Australia.

Diaz, F. (2007). Regularizing query-based retrieval scores. *Information Retrieval, 10*(6), 531–562.

Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of SIGCHI 2001 conference on human factors in computing systems (CHI 2001)* (pp. 277–284). Seattle, Washington.

Ellen, V. (1985). The cluster hypothesis revisited. In *Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1985)* (pp. 188–196). Montreal, Canada.

Erdös, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae-Debrecen, 6*, 290–297.

Furnas, G. W. (1997). Effective view navigation. In *Proceedings of SIGCHI 1997 conference on human factors in computing systems (CHI 1997)* (pp. 367–374). Atlanta, Georgia.

Harman, D. K. (2005). The TREC test collections. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 21–52). Cambridge, Massachusetts: MIT Press.

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49*(4), 59–61.

Hearst, M. A., & Pedersen, J. O. (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)* (pp. 76–84). Zürich, Switzerland.

Hersh, W., Cohen, A., Yang, J., Bhupatiraju, R., Roberts, P., & Hearst, M. (2005). TREC 2005 genomics track overview. In *Proceedings of the fourteenth text retrieval conference (TREC 2005)*, Gaithersburg, Maryland.

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*(5), 217–240.

Leuski, A., & Allan, J. (2000). Strategy-based interactive cluster visualization for information retrieval. *International Journal on Digital Libraries, 3*(2), 170–184.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics, 8*, 423.

Liu, X., & Croft, B. W. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004)* (pp. 186–193). Sheffield, United Kingdom.

Maglio, P. P., & Matlock, T. (2003). The conceptual structure of information space. In K. Höök, D. Benyon, & A. J. Munro (Eds.), *Designing information spaces: The social navigation approach* (pp. 385–403). London, the United Kingdom: Springer-Verlag.

Page, L., Brin, S., Motwani, R., & Terry, W. (1999). The PageRank citation ranking: Bringing order to the Web. Stanford Digital Library Working Paper SIDL-WP-1999-0120, Stanford University.

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

Pirolli, P., & Card, S. (1999). Information foraging. *Psychology Review, 106*(4), 643–675.

Pratt, W., & Fagan, L. (2000). The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association, 7*(6), 605–617.

Robertson, S. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294–304.

Salton, G. (1971). *The SMART retrieval system – experiments in automatic document processing*. Englewood Cliffs, New Jersey: Prentice-Hall.

Smucker, M. D., & Allan, J. (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)* (pp. 461–468). Seattle, Washington.

Smucker, M. D., & Allan, J. (2007). Measuring the navigability of document networks. In *Proceedings of the SIGIR 2007 web information-seeking and interaction workshop* (pp. 37–40). Amsterdam, The Netherlands.

van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworth.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.

Wilbur, W. J. (2005). Modeling text retrieval in biomedicine. In H. Chen, S. S. Fuller, C. Friedman, & W. Hersh (Eds.), *Medical informatics: Knowledge management and data mining in biomedicine* (pp. 277–297). New York: Springer.

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