THE REGNET PROJECT

A Review of Academic Research on Information Retrieval

By Charles H. Heenan

Engineering Informatics Group
Department of Civil and Environmental Engineering
Stanford University
Stanford, California 94305

Email: heenan@stanford.edu

August 6, 2002

Acknowledgement and Disclaimer

This report is intended to review current academic research into information retrieval for unstructured multimedia content. This review has been performed as part of the Regnet Project, which is funded by the National Science Foundation under Grant No. EIA-0085998.

Any opinions, findings, and conclusions or recommendations expressed in this report are those of the author and do not necessarily reflect the views of the National Science Foundation.

Introduction

The purpose of this document is to give an overview of academic research into information retrieval (IR) of unstructured content. Unstructured content typically includes text, speech, music, video, or still images. This evaluation focuses on the retrieval of unstructured text. The scope of the evaluation includes papers published no earlier than 1990 for conferences sponsored by the Association for Computing Machinery (ACM). The research for this document was conducted as part of the Regnet Project at Stanford University. The Regnet Project is funded by the National Science Foundation and is focused on the application of information technology to regulation management and regulatory compliance.

Methodology and Organization

The research for this document was limited to academic papers available through the Association for Computing Machinery's online digital library. The initial queries used to gather documents for review were:

- text mining
- text discovery
- text retrieval
- information mining
- information discovery
- information retrieval
- data mining
- data discovery
- data retrieval
- information management
- knowledge management
- text classification
- information classification
- text categorization
- information categorization

Of the thousands of papers that matched one or more of these queries, more than 500 were selected for an initial review. Of those, 60 papers were selected for the more detailed reviews that form the basis of this document. These 60 papers were chosen either because they represent areas of active research or because they are particularly creative or cutting-edge. In the former case, there are often other similar papers in the ACM portal, any one of which reasonably could have been chosen. In the latter case, there may be no other papers that approach the given research question in the same way. No doubt, there are relevant papers in the ACM portal that did not match any of the starting queries. Likewise, there are relevant papers that were not reviewed in detail. This review is not meant to be exhaustive but is intended to focus on the issues and approaches that may be of relevance to the Regnet Project. Those wishing to pursue further research in this area are encouraged to visit the ACM digital library at http://www.acm.org.

The field of text-based information retrieval is hardly new. In the ACM archive, there exists a mountain of published technical papers on various aspects of the text IR problem. A major topic addressed by information retrieval research is the dual problem of synonymy and polysemy. This problem stems from the fact that, in response to a given query, any retrieval engine must strike a balance between the conflicting demands of precision and recall.[1] With available techniques, an

[1] Precision is the ratio of relevant retrieved documents to retrieved documents.
Recall is the ratio of relevant retrieved documents to relevant documents.

increase in the precision of a retrieval engine tends to result in a concomitant decrease in the recall performance of that engine. That is to say, if you tweak an IR system so that a very high percentage of the result set is relevant to the query, you increase the risk that many other relevant documents will be excluded from the result set. Conversely, if you optimize an IR system so that a very high percentage of all documents that are relevant to the query are included in the result set, you increase the risk that many irrelevant documents will also be included.

The first section of this paper addresses the use of query expansion in solving the problem of synonymy. The second section addresses the use of context vectors in solving the polysemy problem. The third section discusses new developments in search interfaces.

Section 1

An Approach to the Synonymy Problem: Query Expansion

One recurring problem in text IR is how to deal with multiple terms that refer to the same concept. For example, if a query interface does not take this into account when processing search terms, then its search results will be incomplete. Although this is an important problem, it is a relatively simple one to address, however, and developers of text IR systems have tended to solve it with query expansion enabled by controlled vocabularies containing synonym lists or classification hierarchies.

A query expansion-enabled interface will take as input a given search term, look for synonyms in the controlled vocabulary, and return documents that match either the search term or any of its synonyms. More sophisticated query expansion-enabled interfaces use controlled vocabularies that incorporate classification hierarchies in addition to synonym lists. This type of interface uses a hierarchy of superordinate and subordinate relationships to conduct more thorough query expansion operations. For example, if a user enters a search on the term "dog," such an interface might not only return documents that match the term "dog" but also documents that match terms subordinate to "dog" in a classification hierarchy, such as "golden retriever" or "border collie." If the text collection contains documents in multiple languages, the controlled vocabulary can allow query expansion to include an international element by allowing for multilingual synonym lists and classification hierarchies. Overall, the impact of simple as well as more sophisticated query expansion-enabled search tends to be more complete search results and a better search experience for the user.

Despite the relative theoretic ease with which one can use query expansion to address the problem of multiple terms referring to the same concept, the fact remains that constructing synonym lists and classification hierarchies is an onerous, manual task. However, recent work out of Northwestern University in Illinois and out of Monash University in Australia reveals creative ways to conduct query expansion without first having to construct controlled vocabularies.

Reference-Based Query Expansion

Bradshaw, Scheinkman, and Hammond of Northwestern University's Intelligent Information Lab point out that people do not always submit unambiguous search queries to information retrieval

systems. Citing studies on the searching behavior of digital library users, Bradshaw, et al. note that "people rarely use features of [a] query interface such as the Boolean operator "and" or phrase delimiters such as quotation marks to indicate how they intend query words to be grouped together." Moreover, they note that people "rarely form queries of longer than three words" even though more detailed queries are often necessary to get highly-specific search results. Consequently, in their view it is short-sighted for existing indexing systems to assume that searchers will submit accurate, unambiguous queries when the evidence indicates that they will not.

In response to this problem, Bradshaw, et al., have come up with a creative way to index documents so that a query will yield high-quality search results even if the query terminology is imprecise: research documents are indexed according to how they have been referenced in other articles. This approach is based on the observation that, in research papers, the text "surrounding a citation (the reference) is usually a concise description of the information the cited document provides." Using references in this way is a powerful approach to indexing documents because "references pair concise, on-point descriptions of information with the documents that contain that information." Consequently, an information system that enables query expansion by incorporating document-reference information "is much better equipped to deal with the brief and often incomplete way people typically describe an information need." Such a system can provide more accurate, relevant search results even to short queries "because a few words is often enough to eliminate from consideration many irrelevant documents that would be retrieved by standard retrieval techniques based on content." Such a system also has the advantage that generating the reference-based indexes it requires for query expansion is a process that can be automated.

Despite the virtues of reference-based query expansion, there remain a number of limitations to the idea. First, it seems that a system like the one Bradshaw, et al., propose will be limited to conceptually homogenous text archives. A system whose approach to indexing depends upon the way authors of documents cite other documents requires as much; otherwise, there is likely to be insufficient cross-referencing of documents for the citation index to be of benefit. In this light, it makes sense that Bradshaw and her colleagues chose an archive containing only computer science research articles as the underlying text for their system. Second, even if a reference-based approach to query expansion could work for conceptually heterogeneous document collections, the fact remains that more recent articles will tend to be under-indexed as compared to older articles. For a period of time, any newly-published article will not have been cited by any other authors, although the article itself will contain citations to earlier work. Presumably, if a newly-published article addresses a topic on which others have published before, then the existing index of cross-references may succeed in returning the new article in response to queries for which it is relevant. However, authors of newly-published articles that also break new conceptual ground may have to wait until their papers are cited by others before their work is fully incorporated into a cross-reference index. Nevertheless, reference-based query expansion represents an important research contribution to the field of text-based information retrieval.

Multilingual Query Expansion

Chau and Yeh of Monash University's School of Business Systems also look at the information retrieval problem that is created when more than one term refers to the same (or similar) conceptual content. In this case, Chau and Yeh are interested in the problem as it applies to multilingual

heterogeneous document collections as opposed to highly specific text archives like the one used in the Bradshaw study. Yet, like Bradshaw et al., Chau and Yeh explore query expansion as a possible solution.

The application of query expansion to a multilingual corpus is appropriate due to the problems searchers tend to face when looking for resources that are not in their own language. Unless a searcher is bi-lingual, it can be difficult "to formulate [a] query specifying an information need by producing appropriate keywords" in another language. Chau and Yeh add that native users of Asian languages face additional difficulties even when they are able conceptually to specify their information need because "most Asian characters, such as Chinese, cannot be composed easily and directly from the computer keyboard." To deal with both the difficulty in choosing search terms and the difficulty of entering eastern ideograms on western keyboards, Chau and Yeh propose an explorative approach to searching in which an information-seeker will browse through a map, directory, or hierarchy of concepts that are normalized to the information-seeker's native language. The user of such an interface submits a query by clicking on a concept of interest and the system returns results by showing the documents that populate that concept category. Of course, the documents that are returned may be in any number of other languages besides the searcher's native tongue.

While the formulation and submission of a query in this system occurs at the moment a user clicks into a concept category, the groundwork for multilingual expansion of that query occurs well in advance. Chau and Yeh's approach to multilingual query expansion requires preprocessing the document collection so that multilingual content can be grouped into appropriate concept categories ahead of time. This preprocessing uses "the co-occurrence statistics of a set of multilingual keywords extracted from a parallel corpus. " (A parallel corpus is a collection of documents containing identical text written in multiple languages.) The reason for calculating these co-occurrence statistics is that "semantically related multilingual keywords representing similar concepts tend to co-occur in similar patterns (i.e. similar inter- and intra-document frequency) within a parallel corpus." By analyzing these statistics, "multilingual keywords extracted from a parallel corpus [can be] sorted into keyword clusters (concept classes)." Once these keyword clusters have been identified, "each…cluster is given a concept label in each language involved."

On balance, the approach to multilingual query expansion outlined by Chau and Yeh is compelling. The authors make a point of addressing the fact that there is an "inexact correspondence between keywords across languages" due to cultural or linguistic differences. They acknowledge that, as a result, a "one-to-one mapping of a keyword and its foreign counterparts may not always be possible." So rather than attempting to make perfect matches between terms in one language and those in another, Chau and Yeh focus on clusters of relationships in the expectation that those clusters will be of value to the information seeker. To the extent that this approach makes it easier for speakers of Asian languages to formulate queries and to pose them to the system, Chau and Yeh's expectation appears to be appropriate.

However, the authors have not addressed the problem of how concept clusters can be labeled accurately and efficiently. Even within one language, the question of assigning concepts to categories can be a drawn-out manual process full of subjectivity. If an automated concept-to-category assignment tool is used, then the process for creating assignment rules can itself become drawn-out and subjective. When dealing with multilingual text collections, the difficulty of placing concepts in categories and of labeling those categories becomes even greater. At the same time, the

likelihood that a monolingual end-user will be able to distinguish between high- and low-quality concept-labels decreases precisely because a typical user knows only one of the languages being used. One avenue for future research on multilingual text retrieval could be to explore how to develop high-quality concept names for multilingual concept clusters in an efficient manner. Another avenue for future research could be how to strengthen the monolingual end user's relative inability to assess the quality/accuracy of concept labels.

Section 2

An Approach to the Polysemy Problem: Context Vectors and Context Distance

Query expansion is a simple and productive approach to the problem created when multiple terms refer to the same concept. Unfortunately, an equivalently simple approach does not exist for the opposite case that arises when morphologically identical terms refer to separate concepts.

To solve the polysemy problem in information retrieval requires the disambiguation of word meanings when separate ideas are expressed by the same term. A common example of this circumstance is the word "bank," which can refer to a river bank, a bank of public telephones, and a place that stores money. If an information seeker submits a query of "bank," the difficulty is how to enable the search system to determine what type of "bank" is meant. This sort of ambiguity has direct implications for query expansion as well, because in one case the query should expand to include synonyms such as "shore" or "edge" while in another case the synonym list should include "financial institution" or "investment house."

Developers of some early text information retrieval systems chose simply to ignore the polysemy problem. These early systems would return all documents deemed "relevant" to the query, where relevance is based upon strict word similarity. While this may yield many relevant documents, they are likely to be buried among other documents that do contain the search term but that are irrelevant on a semantic level. More importantly, defining relevance according to strict word similarity means some documents that are relevant will not be returned because they do not contain the specific search term. The result of ignoring the polysemy problem in this way is both low precision and low recall. This is problematic on both counts: low precision creates difficulties in separating the wheat from the chaff, so to speak, in the list of returned documents, while low recall is precisely the sort of problem that query expansion is meant to offset.

Query expansion does not hold the answer: although recall would improve, precision would go through the floor if a system were to expand a query on "bank" to include all synonyms of all the various senses of that word (ie, synonyms of "bank" as in "river bank" *and* synonyms of "bank" as in "financial institution," *and*…*and*…). Recent research on the use of context distance for word sense disambiguation holds great promise.

The work of Jing and Tzoukerman of Columbia University and Bell Labs, respectively, suggests one solution for the issue of polysemous terms. Starting from the assumption that a given word or phrase has a dominant meaning in a given document, they then "represent this meaning in the form of a context vector." These context vectors are based on "all occurrences of the same word in [a] document" and are derived from the terms that occur within a window of 10 words surrounding the target word. The more frequently a given word or phrase appears within the window of a given

target word, the stronger a signifier that word is when it comes to sense disambiguation. For example, if the term "savings and loan" always occurs within the 10-word context window for "bank," there is a strong likelihood that the bank in question is the financial institution type and not shore-of-a-river type. For each term within these context windows, a weight is assigned based on the frequency of occurrence.

For example, Figure 1 shows an example of the target word "bank" and its corresponding context vector. Note that the words "savings," "million," and "loan," etc, help "to disambiguate the target word 'bank' as the money bank rather than the river bank."

Target word : bank
Context vector :
{ savings(0.44) federal(0.44) million(0.44)
loan(0.33) company(0.22) farmington(0.22)
board(0.22) agreed(0.22) billion(0.22)
nationwide(0.22) }

Figure 1: Context vector for the target word *bank*. The weight following each term in the context vector indicates the importance of that term in the vector.

On the basis of the context vector in Figure 1, an information retrieval engine that is responding to a query on the term "river bank" can adjust so as to return only exact boolean matches and to exclude matches on the more general term "bank."

However, context vectors alone may not be sufficient to determine whether two morphologically related (or even identical) target words are semantically related. First, an approach is needed for evaluating how closely related a given pair of context vectors may be. Since context vectors are composed of individual terms, Jing and Tzoukerman achieve this by focusing on the "level of mutual information between words in context vectors." For Jing and Tzoukerman, this mutual information level is signified by term co-occurrence frequency. If the terms in one context vector have strong co-occurrence relationships with the terms in another context vector, then the respective target words (regardless of morphology) are more likely to be semantically related. Figure 2 shows a table of word pairs with varying levels of co-occurrence strength (corpus relevance).

Word Pairs	Corpus Relevance
gaza, palestinians	0.600
nyse, dow	0.571
composite, dow	0.537
wheat, grain	0.443
south, year	0.117
food, told	0.052
miles, people	0.051

Figure 2: Word pairs and co-occurrence strength for each pair (corpus relevance).

The calculation of word pair co-occurrence strength (corpus relevance) makes it possible to calculate the distance between context vectors even when the terms in those vectors are not the same. For example, "the word 'bank' may occur with the word 'money' in one context, and with the word 'loan' in [another.] If [one] can capture the close relatedness of 'money' and 'loan', [one] can deduce that 'bank' probably has similar meanings in the two occurrences." Jing and Tzoukerman observe that "a model which relies on exact word repetition will fail in this case since it will miss the relations between 'money' and 'loan.'" Figure 3 shows an example of just such a case. Note that the only shared term in the context vectors is "loan." Despite this, the strong corpus relevance between the terms in each context vector is sufficient to indicate that the two target words concern the same topic.

bank : { savings(0.44) federal(0.44) million(0.44)
loan(0.33) company(0.22) farmington(0.22)
board(0.22) agreed(0.22) billion(0.22)
nationwide(0.22) }
banks { fdic(0.56) depression(0.42) number(0.28)
failed(0.28) post(0.28) fund(0.28) year(0.28)
fslic(0.28) loan(0.14) deposits(0.14) }

Figure 3: "Bank" and "Banks" - Morphologically distinct, but semantically linked.

Figure 4, on the other hand, shows an archetypal polysemy problem: two morphologically identical instances of the word "bank." Are they conceptually identical? Or are they semantically as different as if they were morphologically unrelated? The Jing – Tzoukerman approach allows us to say that, although the two terms are morphologically identical, they are conceptually distinct.

bank : { savings(0.44) federal(0.44) million(0.44)
loan(0.33) company(0.22) farmington(0.22)
board(0.22) agreed(0.22) billion(0.22)
nationwide(0.22) }
bank : { west(0.45) gaza(0.45) strip(0.45)
state(0.22) boasted(0.22) called(0.22)
hailed(0.22) israeli(0.22) plo(0.22)
occupied(0.22) }

Figure 4: A bank is not always a bank.

Jing and Tzoukerman's work is an important contribution to a central problem in information retrieval: how to find an optimal balance between precision and recall. On the one hand, one could maximize recall for a given query simply by returning the set of all records in the document repository. The fact that this results in abysmal precision makes it non-sensical. On the other hand, attempts at maximizing precision must have some way of dealing with polysemy. Otherwise, either those attempts will fail or recall will suffer. In short, precision and recall are two sides of the same coin. The goal is to find an optimal balance between them. Jing and Tzoukerman show us that query expansion and sense disambiguation can take us a long way towards this goal.

The remainder of this paper discusses new developments in search interfaces, including the categorization of search results and the categorization of databases as opposed to text content.

Section 3

Search Interfaces: The Categorization of Search Results

The synonymy and polysemy problems pertain to the task of query fulfillment in that, to be good, a search engine must respond to a query by returning a list of documents with the maximum quantity of relevant records and the minimum quantity of irrelevant records. Yet, there exists a separate set of problems that pertain to the user interface for viewing these search results. Typically, search results are presented in the form of a ranked list, broken down so that only 10 or 20 are viewable on a given web page. Even if precision and recall are optimized, a list of search results will contain some documents that are not useful for the searcher and others that are useful. The list of search results is likely to contain subsets of documents that are similar, or that are related to the search query in a similar way. If precision and recall are not optimized (as is more commonly the case), then the list of search results will also contain irrelevant documents scattered among the relevant ones.

Susan Dumais of Microsoft Research and Hao Chen of UC Berkeley have conducted research into alternatives to the traditional ranked list display of search results. They have found that users are able to find documents more efficiently when search results are organized into topical categories than when they are presented with a standard ranked list.

Dumais and Chen tasked the study participants with finding documents via a traditional list interface as seen in Figure 5, and then by means of category-style interfaces, one of which is shown in Figure 6.

Figure 5: A ranked list interface for search results.

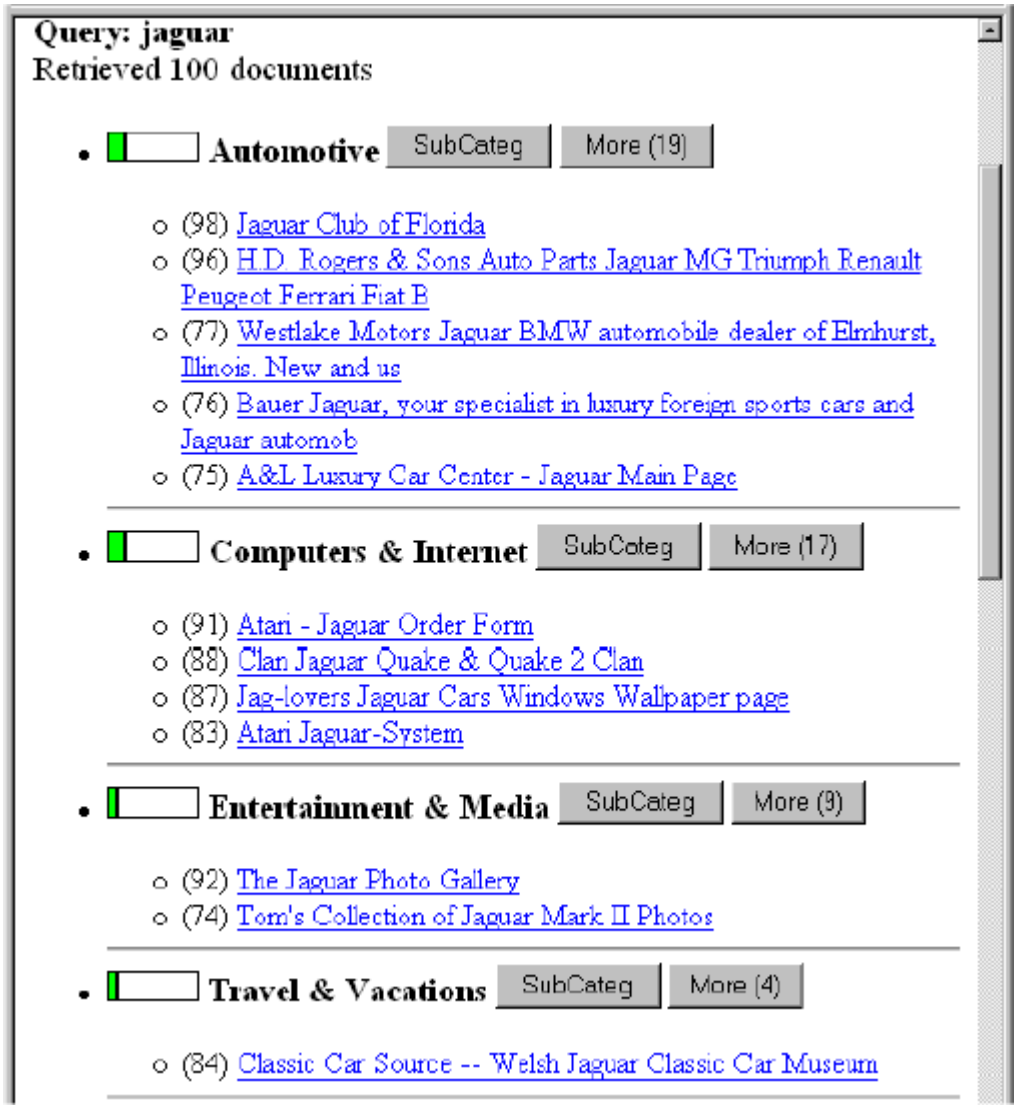

Figure 6: A category-based interface for the same search results as shown in Figure 5.

Dumais and Chen used four variations on category interfaces like the one shown in Figure 6 and three variations on list interfaces like the one shown in Figure 5. In every case, users were more efficient at locating information through a category interface than through a list interface. The relative advantage of a category-based interface was even greater for "difficult" searches as opposed to "easy" ones (See Figure 7).

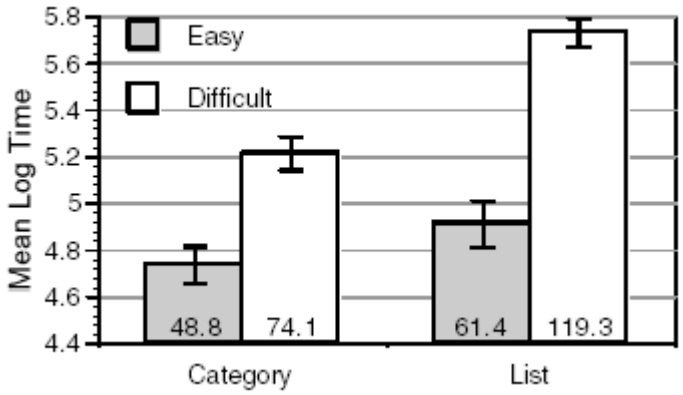

Figure 7: Mean log time to complete tasks for easy and difficult queries for each interface type.

Interestingly, as of April 2002, only one high-profile commercial search engine appears to have incorporated categorization of search results into its user interface. Teoma.com is a web search engine that went live to the public in early 2002. Figure 8 shows the Teoma interface after it has completed a search on the term "Knowledge Management." At the bottom-left of the screen is the standard ranked list of search results. However, at the top right of the screen is a section labeled "Refine – Suggestions to narrow your search." Although Teoma packages the links in this section as suggestions for query refinement, they function as subcategories within the domain of knowledge management.

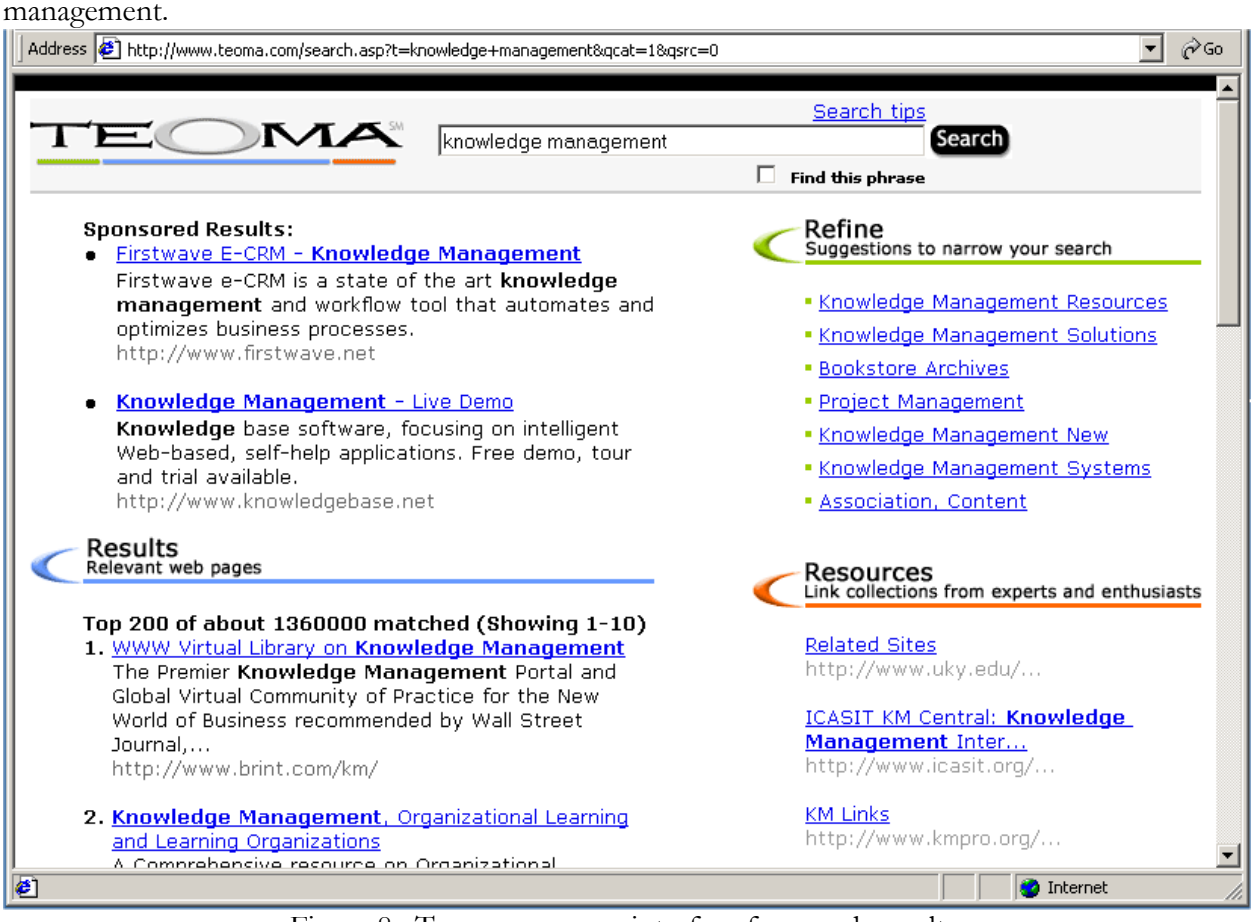

Figure 8: Teoma.com user interface for search results.

During Teoma's beta release, the user interface even used the Windows Explorer "folder" iconography to represent these links explicitly as categories and subcategories within the realm of "Knowledge Management." It is unclear why they switched metaphors, but the fact remains that clicking on links in the "Refine" section of the Teoma interface will yield a subset of documents from the primary search as well as a new list of links (sub-subcategories) for further refinement. Regardless of what metaphor is used to represent the idea of categorization of search results, in the future it is likely that other search websites will follow Teoma's lead and incorporate categorization of search results into the user interface.

Search Interfaces: The Incorporation of Subjective "Expert" Opinion

Besides research on using categorization for making search results more accessible, there is research from Intel Corporation on how the use of "expert" opinion can facilitate interdisciplinary search. John Light, of Intel, published a paper in 1997 in which he discusses search technology and some of the assumptions underlying the then-state-of-the-art search systems. One of his observations is that "text retrieval is currently very Aristotelian. That is, answers are judged as either right or wrong." This raises problems when there is a high degree of speciation within general fields of inquiry, because the same terms can convey radically different meanings between disciplines. The consequence of this for searchers is greater difficulty in finding material outside of one's own domain of specialization. This, Light adds, is problematic because

> "some of the most interesting and important searching being done today is across disciplines. Whether it is done by someone who is a novice or expert in his own discipline, these searches are in a space where the searcher doesn't really know or understand the vocabulary. Historically, some of our greatest inventions have resulted from connecting disparate disciplines, so supporting searches in foreign domains is critically important. Our current search methods, which rely heavily on the user's ability to pick individual words, make that hard."

One of the solutions to this problem is for search systems to turn the binary, Aristotelian right-or-wrong approach to search on its head by incorporating the knowledge of subjective domain experts. Light proposes "a largely automated system that uses expert information that is provably and intentionally *subjective*." He adds that "the application of a human viewpoint is an additional advantage to the system, not a drawback" and that "one way to look at the expert contribution is as that of an *editor* of a publication."

Light envisions experts as fulfilling a number of roles. Two such roles are topic identification and vocabulary definition. According to Light, experts would need to identify a "large list of narrow topics within [a given] document set." These topics could then be used by non-experts to construct queries themselves. In addition, Light argues that experts would need to be responsible for the creation of "a description of the vocabulary used to discuss each topic," where the topic is described "by a list of words or phrases that are specific to the topic."

One question that arises from Light's idea that expert knowledge could be used to improve search is that of labor: Who is going to spend the time necessary to create these lists of topics and domain specific vocabulary definitions? As it turns out, countless individual weblog developers have been

doing just that on a voluntary basis for some time. In May, 1999, the online news site, Salon.com, described weblogs as "personal personal web sites operated by individuals who compile chronological lists of links to stuff that interests them, interspersed with information, editorializing and personal asides. A good weblog is updated often, in a kind of real-time improvisation, with pointers to interesting events, pages, stories and happenings elsewhere on the Web. New stuff piles on top of the page; older stuff sinks to the bottom." Although there is little standardization from one weblog to the next and there is no guarantee that some set of weblogs has rigorously defined specific vocabularies, weblogs do represent a tremendous amount of quasi-expert information on increasingly narrow topical niches.

Since weblogs tend to have a common format, it should be possible for search engines to harvest this information. The result would be that weblog developers will have unknowingly filled-in for the role of "editor" that John Light argues can improve the quality of web search. Again, as with the idea of using categorization for organizing search results, few commercial search engines are taking advantage of weblog information in a way that would fulfill Light's vision. Yet, again, it is Teoma.com that is leading the way.

When a user submits a query through Teoma's search interface, Teoma looks for weblogs and other pages that contain lists of links that deal with the user's query. Links to these list-of-links pages are shown at the bottom right of the Teoma search results page, under the heading "Resources – Link collections from experts and enthusiasts" (See Figure 8). Figure 9 shows the page that is listed first under the Resources heading in Figure 8. It is a list of links to pages dealing with the original search term, "Knowledge Management." Although it is impossible to verify the qualifications of any given "expert" or "enthusiast" who has created a list of links page, the idea of incorporating such pages into a search interface is a good one and—in at least some cases—it does add value.

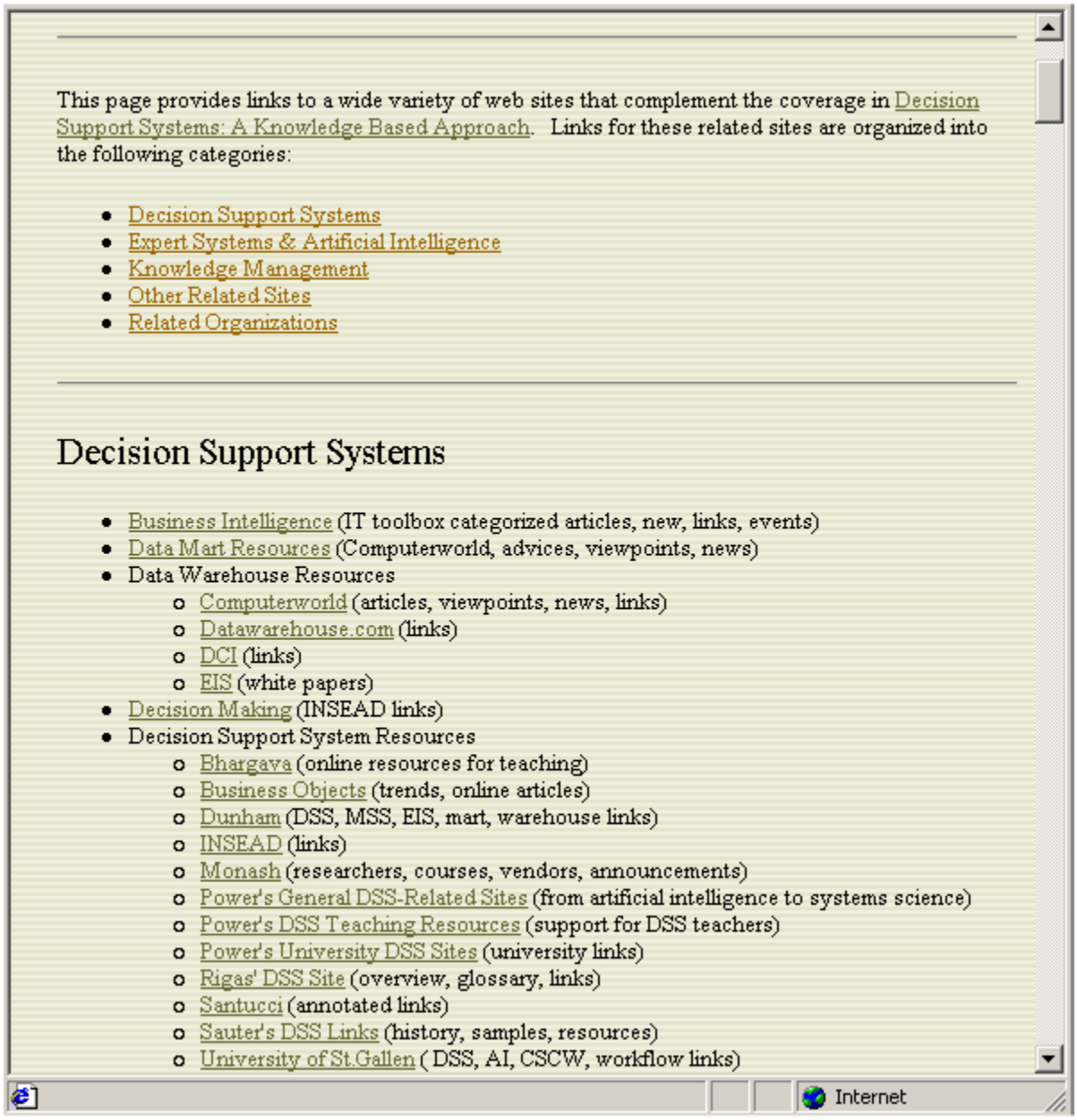

Figure 9: A list-of-links page that was included by Teoma.com in response to a search on the term "Knowledge Management."

Search Interfaces: The Deep Web

While the categorization of search results and the incorporation of "expert" information does add value to search interfaces, the fact remains that traditional web content (content directly accessible through links) represents only a fraction of the information on the Internet. Recent studies indicate that traditional, static web makes up two billion pages of the Internet. While that is a sizable figure, it pales in comparison to the 500 billion pages that are estimated to exist on the "hidden," or "deep" web. Deep web pages reside in web-connected databases and are only accessible through the

mediation of a query interface. These web-based interfaces to databases dynamically generate a list of links in response to searches entered by users. The problem is that "traditional search engines cannot handle such interfaces…." As a result, they "ignore the content of these resources, since [the search engines only work by taking] advantage of the static link structure of the web to "crawl" and index web pages."

There do exist a number of sites that are focused on addressing the problem that is presented by the deep web. For example, Invisibleweb.com (Figure 10) and Searchengineguide.com are two manual categorization efforts in which databases are grouped under topical headings. A click into a category such as "education" will yield a list of sites through which one can access database search interfaces. Through these interfaces, one can "Find a Teacher," "Find a College," or even "Find a School District."

Figure 10: Home page for InvisibleWeb.com, "The Search Engine of Search Engines." The main portion of the page contains a manually populated classification of online databases.

Unfortunately, categorizing online databases manually can be just as time consuming—if not more so—than categorizing online documents, particularly because online databases do not offer

unmediated access to their content. In response to this difficulty, one research effort out of Columbia University points the way towards a more efficient method of categorizing online databases based upon their content.

Ipeirotis, Gravano, and Sahami frame the problem by relating their experiences searching for documents with the keyword "cancer" on the PubMed medical database from the National Library of Medicine. A manual query of the PubMed database for "cancer" yielded "1,301,269 matches, corresponding to high-quality citations to medical articles." However, since these documents are dynamically generated in response to a query, they are not "'crawlable' by traditional search engines." For example, using the same query of "cancer" via websearch engine Alta Vista to find pages in the PubMed site "returns only 19,893 matches. This number not only is much lower than the number of PubMed matches reported above, but…the pages returned by AltaVista are links to other pages on the PubMed site, not to *articles* in the PubMed database." In short, traditional web queries will not work for accessing information in deep web repositories such as the PubMed database.

Ipeirotis, et al., have developed a creative, automated approach for approximating what a database is "about" through the use of query probes and the evaluation of the results from each probe. If a database returns many documents in response to a query about "cancer," but returns zero documents in response to a query about "NHL hockey," that information can be used to help decide whether to classify the database as being about healthcare/medicine or about sports. The more query probes that are submitted, the more refined and accurate will be the ultimate classification of the database itself.

On balance, the approach of using query probes seems to be an effective innovation for categorizing databases without manual intervention. It is likely that this approach will be highly effective when applied to narrow, topically focused databases. The only drawback to this approach is that heterogeneous databases (ones that contain roughly equal numbers of documents about a range of topics) may pose a greater categorization challenge because of lower variation in the database's response to different query probes.

Conclusion

Academic research into information retrieval systems is proceeding apace. New user interfaces for effectively conveying search results have moved from the research lab to the "live" web, and this flow of innovation seems unlikely to fade. New approaches to the problems of synonymy and polysemy are pushing the frontiers of retrieval algorithms to new levels of effectiveness.

However, at the same time there exist a number of fundamental questions about how the "effectiveness" of a retrieval algorithm should be defined, and therefore evaluated. The traditional criteria for a retrieval system have been precision and recall. A system that is precise will return a very low percentage of irrelevant documents given a specified query or classification rule. Yet, while the documents that are returned by such a system will tend to be on-topic, there is no guarantee that those documents represent anything more than a small percentage of all the on-topic documents in the search database. On the other side of the coin, a system that has high levels of recall can be expected to return a significant percentage of all the documents in the database that are on-topic to a given search query. Yet this increase in recall almost always comes at the expense of precision.

Ultimately, the optimal relationship between an information retrieval system's precision and recall is likely to vary depending upon the application domain and upon the needs of the system's users.

If different search and categorization algorithms set the balance between precision and recall differently, clearly some algorithms will not be appropriate for some information seekers' needs. It is important for information seekers to be aware of the variation that exists among search and categorization approaches, and to understand which approach is right for a given information need. In some cases, an information seeker may need a recall-oriented tool. In others, exhaustiveness is less important and a precision-oriented algorithm may be more appropriate. In the end, users of search services should keep in mind that what goes on behind the query submission box varies widely from site to site and that this variation has an impact upon search results. One must not be lulled into an Internet-enabled laziness with respect to information retrieval. Information seekers wishing to be thorough should employ a range of search tools rather than one favorite engine. When viewing search results (or categorization results), they should be just as mindful of what is not returned as they are of what is. And in some cases, they should even consider making a trip to the library of a local research university or other institution. After all, not everything is digital or available electronically. Not everything has been indexed by search engines or categorization schemes. At least, not yet.

References:

Information Visualization

Information visualization is a broad research area. In this paper, only some of the visualization research has been discussed – namely the use of categorization for optimizing the usability of search interfaces. That research is cited under a separate heading, below. Nonetheless, the following papers are noteworthy and readers should consider consulting them to gain a wider context on the field of information visualization. Of particular note is the work of Peter Pirolli, et al., on information scent.

Au, Peter; Carey, Matthew; Sewraz, Shalini; Guo, Yike; Ruger, Stefan. "New Paradigms in Information Visualization" *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 307 – 309. July 2000.

Chi, Ed H.; Pitkow, James; Mackinlay, Jock; Pirolli, Peter; Gossweiler, Rich; Card, Stuart. "Visualizing the Evolution of Web Ecologies" *Proceedings of the Conference on Human Factors in Computing Systems,* Pages 400 – 407. 1998.

Graham, Martin; Kennedy, Jessie B.; Hand, Chris. "A Comparison of Set-Based and Graph-Based Visualisations of Overlapping Classification Hierarchies" *Proceedings of the Working Conference on Advanced Visual Interfaces,* Pages 41 – 50. 2000.

Kreuseler, Matthias; Schumann, Heidrun. "Information Visualization Using a New Focus+Context Technique in Combination with Dynamic Clustering of Information Space" *Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with the Eighth ACM International Conference on Information and Knowledge Management*, Pages 1 – 5. 1999.

Light, John. "A Distributed, Graphical, Topic-Oriented Search System" *Proceedings of the Sixth International Conference on Information and Knowledge Management*, Pages 285 – 292. 1997.

Miller, Nancy E.; Wong, Pak Chung; Brewster, Mary; Foote, Harlan. "TOPIC ISLANDS – A Wavelet-Based Text Visualization System" *Proceedings of the Conference on Visualization*, Pages 189-196. 1998.

Pirolli, Peter; Card, Stuart; Van Der Wege, Mija. "The Effect of Information Scent on Searching Information: Visualizations of Large Tree Structures" *Proceedings of the Working Conference on Advanced Visual Interfaces*, Pages 161 – 172. 2000.

Shneiderman, Ben; Feldman, David; Rose, Ann; Ferre Grau, Xavier. "Visualizing Digital Library Search Results with Categorical and Hierarchical Axes" *Proceedings of the Fifth ACM Conference on Digital Libraries*, Pages 57 – 66. 2000.

Categorization of Search Results

Recent work out of Microsoft Research indicates that categorization of search results, as opposed to simple ranked results, facilitates information retrieval. The following papers discuss this topic in detail.

Borner, Katy. "Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing" *Proceedings of the Fifth ACM Conference on Digital Libraries*, Pages 234 – 235. 2000.

Chen, Hao; Dumais, Susan. "Bringing Order to the Web: Automatically Categorizing Search Results" *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems*, Pages 145 – 152. 2000.

Dumais, Susan; Chen, Hao. "Hierarchical Classification of Web Content" *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 256 – 263. July 2000.

Categorization of an Information Space for Browsing

Chaffee, Jason; Gauch, Susan. "Personal Ontologies for Web Navigation" *Proceedings of the Ninth International Conference on Information and Knowledge Management*, Pages 227 – 234. 2000.

Geffner, S; Agrawal, D; El Abbadi, A; Smith, T. "Browsing Large Digital Library Collections Using Classification Hierarchies" *Proceedings of the Eighth International Conference on Information and Knowledge Management*, Pages 195 – 201. 1999.

Graham, Martin; Kennedy, Jessie B.; Hand, Chris. "A Comparison of Set-Based and Graph-Based Visualisations of Overlapping Classification Hierarchies" *Proceedings of the Working Conference on Advanced Visual Interfaces,* Pages 41 – 50. 2000 (cross-referenced with above).

Approaches to Information Retrieval

The following papers discuss interesting avenues of research in information retrieval as a whole. Several of these papers are discussed in greater detail in the body of this essay. Of the ones that were not discussed, the a primary theme is the use of software agents as retrieval facilitators. The news article "Use the Blog, Luke" is also of particular interest.

Belkin, Nicholas J.; Croft, Bruce W. "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communications of the* ACM, Volume 35, Issue 12, Pages 29 – 38. December 1992.

Chau, Michael; Zeng, Daniel; Chen, Hinchun. "Personalized Spiders for Web Search and Analysis" *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Pages 79 – 87. June 2001.

Dorre, Jochen; Gerstl, Peter; Seiffert, Roland. "Text Mining: Finding Nuggets in Mountains of Textual Data" *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pages 398 – 401. 1999.

Ipeirotis, Panagiotis; Gravano, Luis; Sahami, Mehran. "Probe, Count, and Classify: Categorizing Hidden-Web Databases" *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, Pages 67 – 78. May 2001.

Jing, Hongyan; Tzoukerman, Evelyne. "Information Retrieval Based on Context Distance and Morphology" *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 90 – 96. August 1999.

Johnson, Steven, Use the blog, Luke (http://www.salon.com/tech/feature/2002/05/10/blogbrain/print.html), May 2002.

Lam, Wai; Lai, Kwok-Yin. "A Meta-Learning Approach for Text Categorization" *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 303 – 309. September 2001.

Menczer, Filippo; Belew, Richard K. "Adaptive Information Agents in Distributed Textual Environments" *Proceedings of the Second International Conference on Autonomous Agents*, Pages 157 – 164. 1998.

Plaisant, Catherine; Shneiderman, Ben; Doan, Khoa; Bruns, Tom. "Interface and Data Architecture for Query Preview in Networked Information Systems" *ACM Transactions on Information Systems*, Volume 17, Issue 3, Pages 320 – 341. July 1999.

Singh, Lisa; Scheuermann, Peter; Chen, Bin. "Generating Association Rules from Semi-Structured Documents Using an Extended Concept Hierarchy" *Proceedings of the Sixth International Conference on Information and Knowledge Management*, Pages 193 – 200. 1997.

Stuckenschmidt, Heiner; van Harmelen, Frank. "Ontology-Based Metadata Generation from Semi-Structured Information" *Proceedings of the International Conference on Knowledge Capture*, Pages 163 – 170. 2001.

Tansley, Robert; Bird, Colin; Hall, Wendy; Lewis, Paul; Weal, Mark. "Automating the Linking of Content and Concept" *Proceedings of the Eighth ACM International Conference on Multimedia*, Pages 445 – 447. 2000.

Voss, Angi; Nakata, Keiichi; Juhnke, Marcus. "Concept Indexing" *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, Pages 1 – 10. 1999.

Wong, Kam-Fai; Song, Dawei; Bruza, Peter; Cheng, Chun-Hung. "Application of Aboutness to Functional Benchmarking in Information Retrieval" *ACM Transactions on Information Systems*, Volume 19, Issue 4, Pages 337 – 370. October 2001.