

From Information Retrieval to Information Interaction

Gary Marchionini

University of North Carolina at Chapel Hill, School of Information and Library Science
100 Manning Hall
Chapel Hill, NC 27599, USA
march@ils.unc.edu

Abstract. This paper argues that a new paradigm for information retrieval has evolved that incorporates human attention and mental effort and takes advantage of new types of information objects and relationships that have emerged in the WWW environment. One aspect of this new model is attention to highly interactive user interfaces that engage people directly and actively in information seeking. Two examples of these kinds of interfaces are described.

1 Introduction

Information retrieval (IR) is hot. After 40 years of systematic research and development, often ignored by the public, technology and a global information economy have conspired to make IR a crucial element of the emerging cyberinfrastructure and a field of interest for the best and brightest students. The new exciting employers are Google, Amazon, and eBay and the extant giants like IBM and Microsoft have active IR research and development groups. In many ways, research in IR had plateaued until the WWW breathed new life into it by supporting a global marketplace of electronic information exchange. In fact, I argue that the IR problem itself has fundamentally changed and a new paradigm of information interaction has emerged. This argument is made in two parts: first, the evolution of IR will be considered by a broad look at today's information environment and trends in IR research and development and second, examples of attempts to address IR as an interactive process that engages human attention and mental effort will be given.

2 Information Objects and People

As a scientific area, IR uses analysis to break down the whole problem into components and first focus on the components that promise to yield to our techniques. IR has always been fundamentally concerned with information objects and with the people who create, find, and use those objects; however, because people are less predictable and more difficult and expensive to manipulate experimentally, IR research logically focused on the information objects first. Traditionally, information objects have been taken to be documents and queries and research has centered on

two basic issues: representation of those objects and definition of the relationships among them. Representation is a classical issue in philosophy, information science (e.g., Heilprin argued that compression was the central representation problem [9]), and artificial intelligence. The IR community has demonstrated a variety of effective representations for documents and queries, including linguistic (e.g., controlled vocabulary) assignments and a large variety of mathematical assignments (e.g., vectors) based on term-occurrence, relevance probability estimates, and more recently hyper-link graphs. IR research has mainly focused on equality (e.g., of index terms) and similarity relationships—similarity between/among objects—and developed a large variety of matching algorithms that are exploited in today’s retrieval systems. A schematic for the traditional IR problem is depicted in Figure 1.

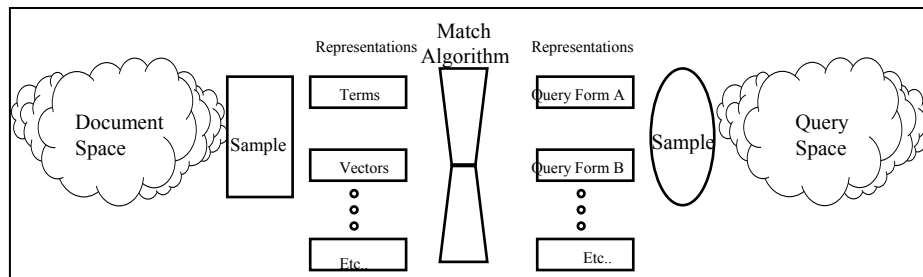


Fig. 1. Content-Centered Retrieval as Matching Document Representations to Query Representations

The figure shows that samples of document and query objects from the respective universe of all objects are each represented in some fashion, most often using the same representation form. For example, a simple approach used in early commercial retrieval systems was to represent documents and queries with terms assigned from a controlled vocabulary and simply match overlaps. A more contemporary example returns ranked sets of similarities by representing documents and queries as vectors of inverse document frequency values for a specific set of terms in the sample ordered by cosine similarity. In cases where the document and query representations are in different forms (e.g., different metadata schemes or human languages), crosswalks, translations, or interlingua must also be added to the process. This content-centered paradigm has driven creative work and led to mainly effective retrieval systems (e.g., SMART, Okapi, Iquery), however, progress toward improving both recall and precision seems to have reached a diminishing return state.

Two important changes have been taking place in the electronic information environment that expand this schema and stimulate new kinds of IR research and development. These changes are due to new types and properties of information objects and to increasing attention to human participation in the IR process. The IR community has begun to recognize these changes as illustrated by the two grand research and development challenges identified for IR research at a recent strategic workshop [1]: global information access (“Satisfy human information needs through natural, effi-

cient interaction with an automated system that leverages world-wide structured and unstructured data in any language.”), and contextual retrieval (“Combine search technologies and knowledge about query and user context into a single framework in order to provide the most ‘appropriate’ answer for a user’s information needs.” P.330).

The very nature of information objects of interest to IR has both broadened and qualitatively morphed. On one hand, IR has broadened its interest in objects that are not strictly text to include statistics, scientific data sets and sequences, images, sounds, video, animations and other multimedia. In many cases, the same retrieval paradigms are applied with these kinds of objects (e.g., color histograms rather than term occurrences). Additionally, new kinds of objects are emerging such as executable code modules, transaction protocols and forms, and advisory agents and processes, each offering new kinds of feature sets that may be leveraged for retrieval. What is more significant than new types of objects is the trend toward all objects becoming more dynamic and less static and dependable for IR purposes. For example, an active blog is an object that is continually changing and its representations must likewise be continually updated as well. This change emanates from new capabilities within objects and from new capabilities in the external environment that contains them. Internally, electronic objects increasingly are designed to exhibit behavior—to ‘act’ according to external conditions. Hypertext is of course the classic example; recommender systems illustrate more contemporary examples; and context-aware sensor-program devices illustrate the latest trend. In addition to the increasingly sophisticated behavior inherent to information objects is the trend toward the global information space (cyberinfrastructure) to store and use context. For example, a retrieval system may not only represent webpage content but continuously update access times and referral pages. Additionally, the system may save increasingly detailed traces of fleeting ephemeral states arising in online transactions—perhaps as extreme as client-side mouse movements as well as clicks. Thus, our objects acquire histories, annotations, and linkages that may strongly influence retrieval and use. It is important to keep in mind that this applies to query objects as much as document objects. For example, consider the implications for retrieval of a query on the World Trade Center before and after 9/11.

These changes in the very nature of information objects offer new challenges and opportunities for IR. The IR community has moved to accept these challenges on multiple fronts—consider for example, the evolution of the TREC tracks over time. Clearly, entirely new kinds of features are available to use in our object representations. Likewise, object contexts will help enormously in representation and revealing relationships. What seem particularly promising are opportunities to discover new kinds of features within objects, and new kinds of relationships among objects that can be leveraged for retrieval purposes. Hyperlinks and citations are literal relationships formed by object creators and these relationships have been creatively used as features for representing those objects in page rank and hub-authority algorithms. Explicit recommendations are relationships formed by third parties between objects and opinion and can be used to cluster objects of similar opinion. Implicit recommendations are relationships formed by social behavior—the traces of many people acting with objects—and are also leveraged for retrieval purposes. I suspect there are

scores of features natively associated with electronic objects (e.g., trustworthiness, various costs, perceptibility) and even more relationships among electronic objects (e.g., counterargument, derived from, alternative for) that have yet to be tapped for retrieval purposes. The truly exciting thing about IR today is that there is so much new ground to plow that even relative novices can make important discoveries.

Taken alone, this basic change in the nature of information would bloat the basic IR paradigm with a large array of alternative representation options and matching algorithms. A second trend has been in play that combines to require a new paradigm for IR. This trend is an increasing consideration of the people using an information retrieval system. Although there have always been voices representing people in IR research (e.g., advocates of subjective relevance such as Saracevic [17], Schamber [18], and Harter [7]; those who focused on the cognitive processes in retrieval such as Belkin [2], Ingwersen [10], and Marchionini [12]), there are increasing efforts in the IR research community to incorporate people into the retrieval problem. This represents maturation in our approach to IR research and development as we aim to expand our problem-definition to include major facets that have long been set aside in order to focus on the content facets of the overall IR problem.

Figure 2 depicts a different paradigm for the retrieval problem than the classic matching paradigm in Figure 1. The information sample here is shown as a cloud rather than a fixed database since it is dynamic. In this figure, the emphasis is on the flow of representations and actions rather than discrete matches. The indexes are multiple and dynamic. The classical techniques for representing information objects remain useful but may be controlled/selected by users rather than fixed by the system. The relationships of similarity, however, may be determined by the human information seeker on the fly according to their needs and capabilities. Thus, the problem shifts from the system optimizing matching to put burden on the human information seeker to engage in an ongoing process. In such a user-centered paradigm, people have responsibilities and capabilities. Expecting a two-word query to Google to solve every information need is both lazy and naïve and people must go beyond this to be successful. One challenge is that people tend to want to be lazy and naïve (this is sometimes a good cost-benefit tradeoff decision) when doing complex and tedious tasks, especially in the many cases when retrieval tasks are embedded in larger activities. Our approach to this challenge is to imagine information seeking as a core life process where people are constantly connected to information sources just as our bodies are connected to the environment through filters and selectors that are highly tuned to the environment. In such a paradigm, the crucial system design challenges become the control mechanisms for interacting with representations in agile and engaging ways. Note that some of these interactions incorporate the existing query/results patterns so ubiquitous today.

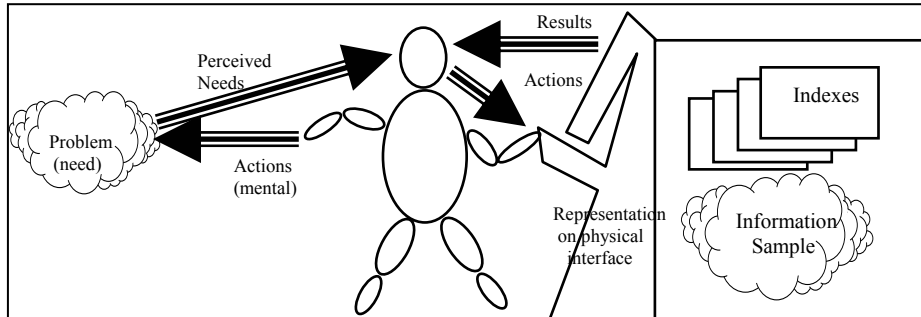


Fig. 2. User-centered information interaction

IR research brings users to bear in various ways. There have been long-standing efforts to provide automatic or user-controlled query expansion [e.g., 14], systems that strongly benefit from user relevance feedback [e.g., 15, 16], and efforts to provide user assistance [e.g., 13]. In addition to leveraging human-generated metadata, researchers are looking for ways to use user behavior or conscious annotations to add additional metadata/features to objects. Some have worked toward modeling users with profiles that may be explicitly completed by users or automatically generated by monitoring actions. More recently, attention to recommender systems both explicitly and automatically capitalizes on user actions with information systems [e.g., 11]. These efforts are even being subsequently leveraged for opinion mining that generates new inferred relationships that may in turn be used as features for retrieval. Other efforts aim toward providing memory aids for users (e.g., [5]) and in extreme examples, of gathering more or less complete interaction histories [e.g., 3]. Other approaches may leverage caching to preserve user-system interaction states over long periods of time (the Internet Archive preserves web page states, but imagine resources for preserving all interactions with specific resources for very long periods—something telephone companies have had to do routinely). Still others aim to create integrated environments that use data mining rules and contemporaneous activity to contextual information retrieval [e.g., 4]. There are likewise aims to create anticipatory information systems that go well beyond the selective dissemination systems of yore to leverage context and user profiles. All of these efforts will enrich a human-centered IR paradigm and advance the field toward more complete consideration of all information seeking factors. I estimate that the greatest progress come from methods that actively include human capabilities in the IR process. To this end, a number of researchers aim to focus on the crucial human-system interaction mechanism that serves as the linchpin of this paradigm.

3 Highly Interactive Interfaces

The concept of direct manipulation was introduced into interface design by Shneiderman [20] and has been applied to interfaces specific to information retrieval by a number of researchers. Shneiderman and his colleagues applied direct manipulation techniques in retrieval environments called dynamic query systems [19]. The key elements of these highly interactive interfaces are active engagement of the user with visual representations that update immediately upon input actions and allow immediate reversal of actions. In the case of dynamic queries, there is close coupling of results display and mouse or keyboard actions. Other researchers have created highly interactive environments for database and information retrieval, most notably, the Xerox PARC group's series of systems (e.g., the Hyperbolic Tree, Perspective Wall, Web Forager). See Hearst [8] for a review of interfaces for IR. Two examples from our work at Carolina on agile view interfaces for information interaction follow.

2.1 Digital Video Retrieval

As part of our efforts to develop an open source video digital library (www.open-video.org), we have created and systematically evaluated a series of visual surrogates (representations) for video content. These surrogates include keyframe-based storyboards and slide shows, fast forwards, and excerpts. These surrogates are considered to be alternative view options for users who have identified a partition of the corpus (either through a text query or a set of interactions/selections with corpus overview representations). Figure 3 shows an Open Video Overview for a set of videos. The views are agile in that simple mouse actions are used to change views. Users can choose which representation best meets their needs and quickly get overviews of the result set with different visual and textual cues emphasized. Clicking on a specific surrogate for a segment (see Figure 4), brings up a full metadata record with three different previews specific to that video's content: a 7 second excerpt, a storyboard (containing up to 36 keyframes), and a fast forward (at 64X speed). Clicking one of

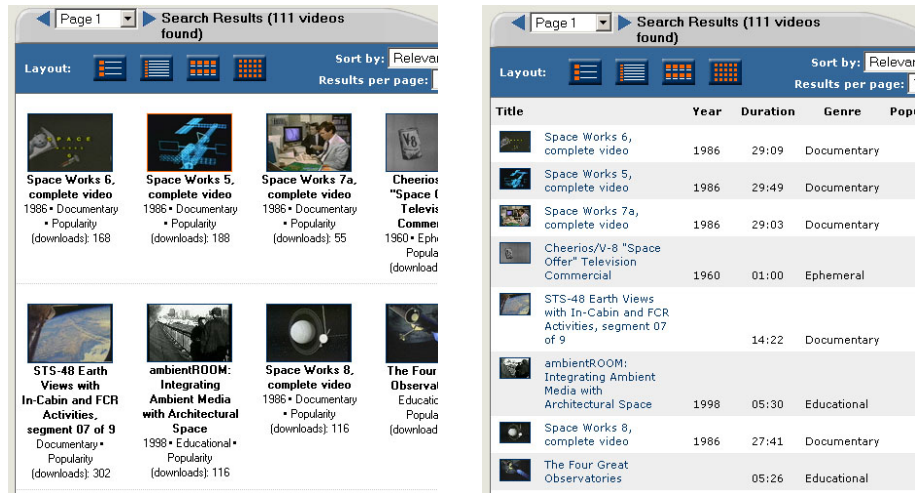


Fig. 3. Alternative Overviews of a Search Result Set

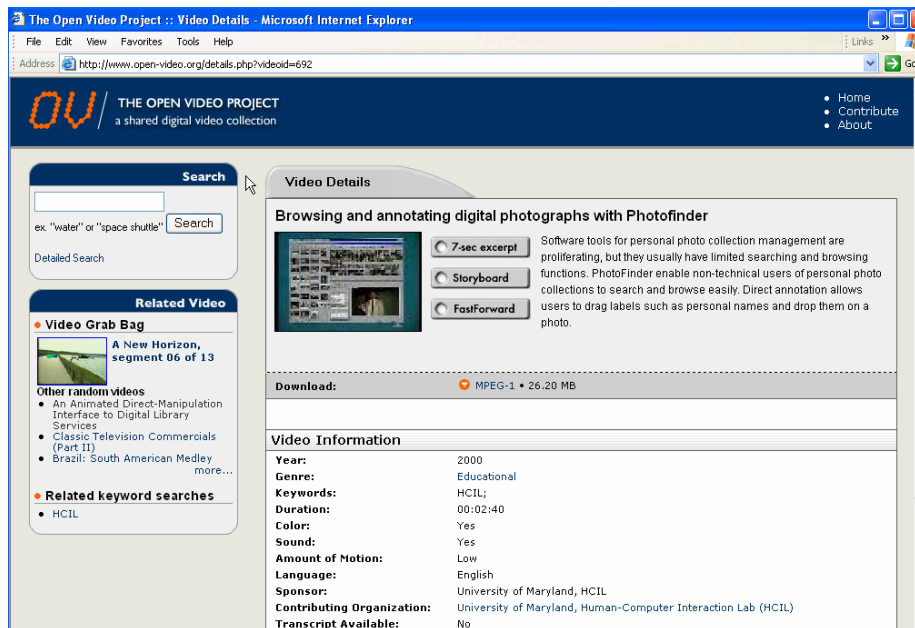


Fig. 4. Alternative Previews for a Specific Video Segment

the radio buttons immediately displays the preview in the preview panel without opening new windows or changing the user's context. Textual metadata is also displayed. The retrieval paradigm is to allow the user to move quickly to different levels of video granularity, most of which have alternative representations that emphasize different video features, in order to both determine whether it is worth loading and

viewing the video and to understand the overall context within which the specific video sits. The actual system also leverages other relationships such as popularity of download, similarity based on usage within a session (implicit recommendations), and various standard bibliographic indexes with hyperlinks as appropriate. Note that textual queries are also supported—the idea is not to replace a demonstrably useful capability but to augment it and put information seekers in control of strategic search decisions.

What I consider most significant about this project is that the design decisions are based on an interaction framework (agile views) and the object representations are empirically validated with extensive user testing (see the project website for a series of papers). This principled and systematic approach to system development is rooted in the strong evaluation cultures of the HCI and IR communities. User feedback and strong growth in usage of the Open Video digital library demonstrates the efficacy of this empirical evaluation tradition to theory and development.

2.1 Interacting with Databases of Webpages

Another example of a highly interactive interface that couples query and results seamlessly is the Relation Browser++ (<http://idl.ils.unc.edu/rave>). The idea is to present users with a complete information space and allow them to interact with a variety of partitions based upon several attribute sets. Figures 5-7 show a sequence of explorations in the US Energy Information Administration website with more than 10000 webpages represented in the underlying database. The webpages have been classified according to their pertinence to four main facets (these facets and the classifications shown here are based on the EIA website structure): fuel types, geography, sector, and process, each of which has a small number of attribute values (sub-facets). The opening screen (not shown) provides the number of pages and relative-length bars for each attribute value in the entire corpus. Note that pages can be indexed by several attribute values within and across facets. As the user moves over an attribute value, the number of pages and bars in other facet attribute values are immediately updated. This allows people to explore relationships across facets. Clicking on an attribute value partitions the database to include only those webpages meeting that condition. Browsing/mousing can continue dynamically, or the user can click the search button to retrieve the results. In Figure 5, the user has clicked on the attribute value 'natural gas' and the search button. We see that 2916 pages are related to natural gas, 128 of these are also related to alternative fuels, 576 are related to the commercial sector, and 403 are related to imports/exports. The search button changes to 'restart' after clicking and the results are displayed in a panel in the same window. This is an important element of interactive interfaces---maintaining user context so that interactive flow is not lost. New window displays should be avoided until discrete transition points. In the RB++, all browse and search activity takes place in the same window with updates optimized to avoid cognitive interrupts. New windows are only used when the user clicks on a record to jump to that particular webpage. Note that the number of

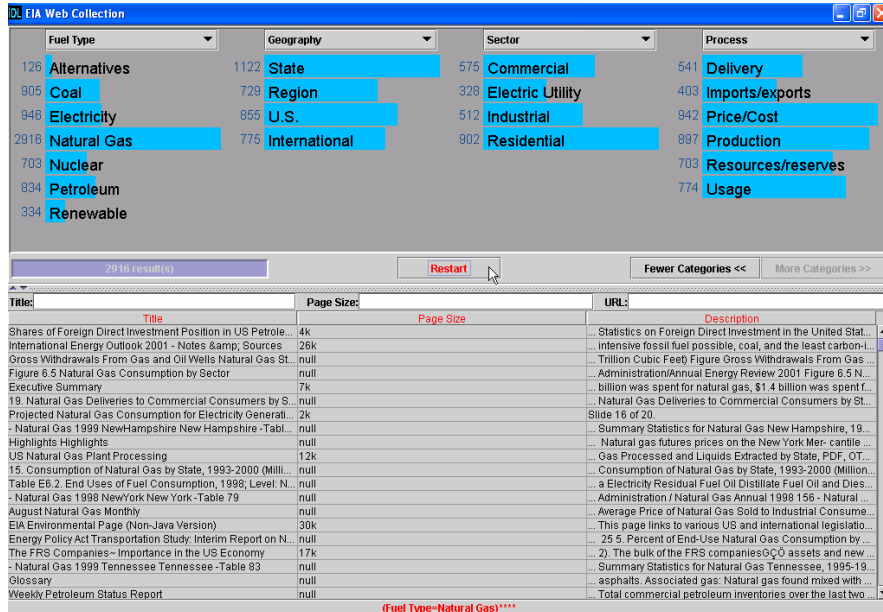


Fig. 5. Relation Browser++ display for Energy Information Administration websites after clicking on natural gas.

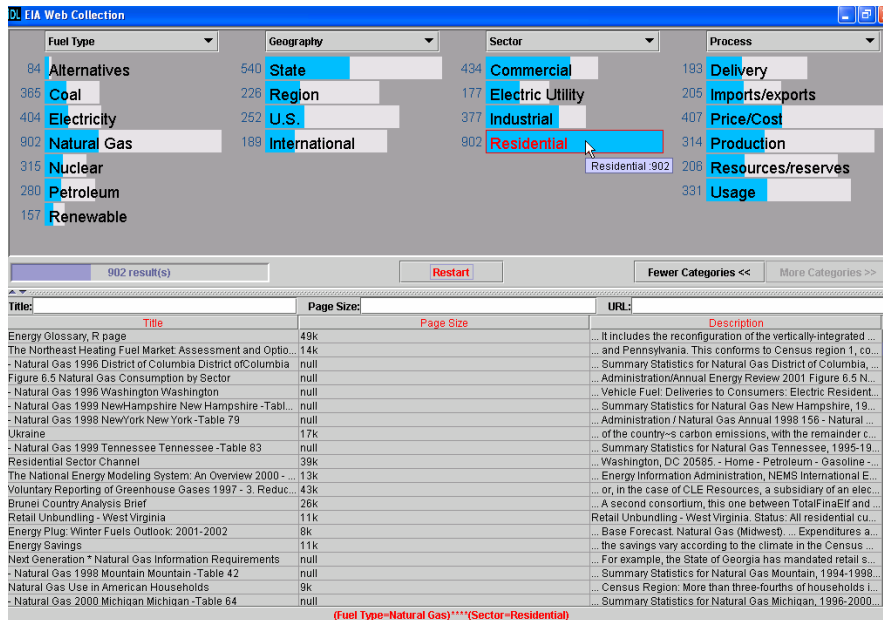


Fig. 6. Relation Browser++ display after moving mouse over residential attribute

results appears to the left of the search button and the SQL-like query statement that produced the results appears at the bottom of the window.

At this point, the user can continue browsing, execute a string search within the results, retrieve a specific page, or start over. Figure 6 shows an example where the user simply moved the mouse to the residential attribute value. All the attribute value numbers and bars are immediately updated; thus 902 webpages are related to natural gas and the residential sector, and 540 webpages are related to residential sector natural gas at the state level. Moreover, note that the posting results have also been updated to show 902 results are available. Each mouse move is in fact a new query. The user can continue to narrow down the results by mousing or can enter text queries in any of the three result set fields (title, page size, URL). The string search immediately returns matches anywhere in the field with the matching characters highlighted. Figure 7 shows the results at the instant that the 's' in the string 'house' is typed from the state shown in Figure 5---yielding 50 results from the 2916 natural gas pages.

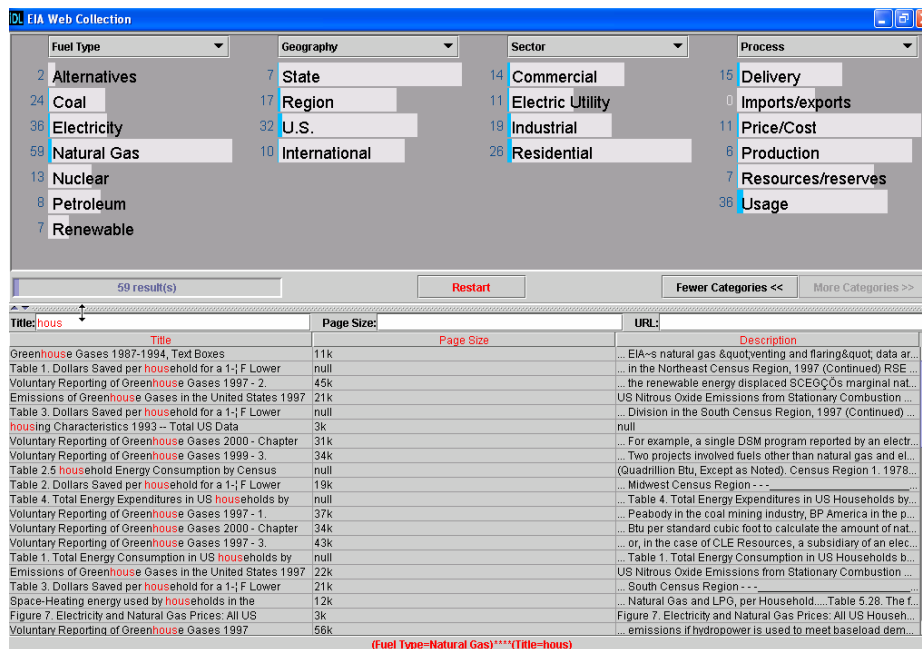


Fig. 7. Relation Browser++ display after clicking on residential and entering first four characters of word 'house' in the title field.

The Relation Browser++ is the third iteration in an ongoing effort to develop a general purpose interface that couples browsing and searching in databases of web content. The current version is a Java applet linked to a MySQL database. The Relation Browser++ and its predecessors (RAVE-the relation attribute viewer, and original relation browser) have been applied to dozens of applications (see examples at <http://idl.ils.unc.edu/rave>) where the number of facets and the number of attribute values are small (to preserve screen real estate to ensure no scrolling in the browsing panel). The interface works with medium size databases of tens of thousands of records. We have an implementation with almost 3 million transaction log records but getting the metadata to the client-side applet is very slow. For very large databases like a WWW search, the Relation Browser+ would be most appropriate as a way to interact with a set of search results or a specific category of pages. Our current work aims to develop automatic ways to find good slicing facets and then populate the underlying databases with appropriate pages. The first problem is the category discovery problem and the second is the text classification problem. We are investigating clustering techniques with some customized heuristics [6] with several of our statistical agency research partners.

3 Conclusion

The examples given here are meant to illustrate beginning steps toward engaging people in a continuous way during information seeking. The video retrieval and database browsing examples aim to harvest the fruits of IR progress so that machines can do what they do best under the control of humans with capabilities to recognize patterns and draw inferences from a wide array of possible relationships. In this way, information is created on the fly as well as retrieved (e.g., as users see new patterns or relationships among objects). A person with an information problem is best able to meet that need through action, perception, and reflection rather than through query statements alone. Thus, the notion of information interaction rather than information retrieval to better reflect the active roles of people and the dynamic nature of information objects in the electronic environment.

Acknowledgements

This work has been partially supported by National Science Foundation Grants # IIS 0099638 and EIA 0131824. Thanks are also due to Sheila Denn and Miles Efron who provided thoughtful comments and suggestions on an earlier draft of this paper.

References

1. Allen et al. (2003). Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. ACM SIGIR Forum 37(1), 31 – 47.

2. Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, vol. 5: 133-143.
3. Bell, G. & Gemmell, J. (2003). Storage and media in the future when you store everything. Microsoft Research <http://research.microsoft.com/~gbell/storeall.htm> accessed on January 14, 2004.
4. Budzik, J., Hammond, K. & Birnbaum, L. (2001). Information access in context. *Knowledge Based Systems*, 14(1-2), 37-53.
5. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. (2003). Stuff I've seen : A system for personal information retrieval and re-use. *Proceedings of SIGIR 2003 The 26th Annual International Conference on Retrieval and Development in Information Retrieval* (Toronto, Canada July 28-Aug 1, 2003). NY: ACM Press, 72-79.
6. Efron, M. Marchionini, G., & Zhang, J. (2003). Implications of the recursive representation problem for automatic concept identification in on-line governmental information. Proceedings of the ASIST Special Interest Group on Classification Research. (Los Angeles: Oct. 19-22, 2003).
7. Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9), 602-615.
8. Hearst, M. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.) *Modern information retrieval*. pp. 257-323. NY : ACM Press and Addison-Wesley.
9. Heilprin, L. (1985). Paramorphism versus homomorphism in information science. In Laurence Heilprin (Ed), *Toward foundations of information science*. White Plains, NY: Knowledge Industry Publications.
10. Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham
11. Konstan, J. (2004). Introduction to recommender systems: Algorithms and evaluation. *ACM Transactions on Information Systems*, 22(1), 1-4.
12. Marchionini, G. An invitation to browse: Designing full-text systems for casual users. *Canadian Journal of Information Science*, 12(3/4), p. 69-79, 1987.
13. Meadow, C., Cerny, B., Borgman, C., & Case, D. (1989). Online access to knowledge: System design. *Journal of the American Society for Information Science*, 40(2), 86-98.
14. Mitra, M., & Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. *Proceedings of SIGIR 1998 The 21st Annual International Conference on Retrieval and Development in Information Retrieval* (Melbourne, Australia Aug 24-28, 1998). NY: ACM Press, 206-214.
15. Robertson, S. (1997). Overview of the okapi projects. *Journal of Documentation*, 53(1), 3-7.
16. Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science*, 41(4), 288-297.
17. Saracevic, T. (1975). Relevance: A review of and a framework for thinking on the notion in information science, *Journal of the American Society for Information Science*, 26 321-343.
18. Schamber, L., Eisenberg, M. & Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition, *Information Processing & Management*, 26(6), 755-776.
19. Shneiderman, B. (1994), Dynamic queries for visual information seeking, *IEEE Software* 11(6), 70-77.
20. Shneiderman, B., (1983). Direct manipulation: A step beyond programming languages, *IEEE Computer* 16(8), 57-69.