Search Result Visualisation with xFIND

Keith Andrews
IICM, Graz University of Technology and
Know-Center Graz
Graz, Austria
kandrews@iicm.edu

Christian Gütl
GUETL IT Research & Consulting and
Infodelio Information Systems
Graz, Austria

Josef Moser
Hyperwave Research and Development
Graz, Austria

Vedran Sabol
IICM, Graz University of Technology
Graz, Austria

Wilfried Lackner
IICM, Graz University of Technology
Graz, Austria

Abstract

The xFIND gatherer-broker architecture provides a wealth of metadata, which can be used to provide sophisticated search functionality. Local or remote documents are indexed and summaries and metadata are stored on an xFIND broker (server). An xFIND client can search a particular broker and access rich metadata for search result presentation, without having to fetch the original documents themselves. Search result sets are not only presented as a traditional ranked list, but also in an interactive scatterplot (Search Result Explorer) and using dynamic thematic clustering (VisIslands).

1. Introduction

The amount of information readily accessible to individuals at their workplace and at home is rapidly increasing. There are now more than one billion unique, indexable web pages[11]. Monolithic, centralised search engines are increasingly unable to cope with the exponential growth of the web. Since individual pages are visited perhaps only once or twice a year, the centralised index is inherently out of date. More general search queries often return many hundreds or thousands of matching documents. Hence the motivation for both a scalable resource discovery framework and for visualisation tools to help end users explore search result sets.

This paper briefly describes the xFIND gatherer-broker system, before presenting in detail two interactive visualisa-

tion techniques which benefit from the xFIND architecture: the Search Result Explorer and VisIslands.

2. Related Work

The first widespread client-server indexing and retrieval system overvthe internet was WAIS[13], using a protocol derived from Z39.50[2]. Harvest[3] was the first distributed, gatherer-broker information discovery and access system for the web, an integrated set of tools, written largely in C, for gathering information from diverse repositories, building topic-specific content indices, and searching the indices. Combine[16] is a gatherer-broker system, originally developed as part of the European DESIRE project. The SOIF format for document summaries introduced by Harvest is widely used, but is in the process of being superseded by RDF[14], a format for describing resources in XML. However, gatherer-broker systems supporting RDF are only beginning to emerge[12]. The xFIND system described later in this paper, builds upon the tradition of Harvest, but is written mainly in Java and has its own set of protocols.

In terms of visualisation of search result sets, two approaches are generally taken: visualisation of metadata attributes and visualisation via derived document similarities. Envision[15] plots the results of bibliographic searches in a two-dimensional scatterplot based on their metadata attributes. The mapping of particular attributes to visual representations such as the x-axis, y-axis, icon size, and icon shape, is controlled by drop-down menus. The Search Result Explorer described later takes the same scatterplot approach as Envision, but uses xFIND's extended set of meta-

data.

Bead [5, 6] uses a force-directed placement technique to lay out documents in a corpus as a landscape. More similar documents lie closer together in the landscape. Searches can be made and the results highlighted. ThemeScape (part of the SPIRE text visualisation suite)[17] examines a corpus of text documents and extracts a set of discriminating terms (words), usually nouns, which characterise topics in the corpus. A combination of stoplists, synonym substitution, and statistical analysis is used to select 200 to 300 "good" discriminating terms. A landscape of topical terms is generated and documents contribute to the height field at each point, based on their contribution to each topic. The VisIslands visualisation described in this paper builds on the work of Bead and ThemeScape, but uses a pre-clustering technique to improve run-time performance.

3. The xFIND System

xFIND[8, 9, 10], the extended Framework for Information Discovery, is a is a framework for distributed information discovery and knowledge management. For scalability, xFIND uses a gatherer-indexer-broker architecture, similar to that pioneered by the Harvest[3] system. As shown in Figure 1, xFIND consists of *gatherers*, *indexers* and *brokers*. For reasons of portability and platform independence, xFIND is implemented almost entirely in Java. In addition to traditional metadata like title, keywords, and description, xFIND also gathers and indexes document headings, and information about embedded links and images. Metadata regarding aspects of information quality, such as authority, diction, and target audience can be manually or semi-manually added.

The xFIND gatherer gathers information about documents and resources, both local and remote, and pre-processes this information. The gathering process is adaptive. At configurable intervals, entire servers, particular sub-sites, or individual documents can be gathered. To reduce network load, a local gatherer can be configured to gather information from a locally mounted file system. The gatherer pre-processes HTML and plain text documents (soon also PDF) and generates a pre-defined set of metadata (a document descriptor or summary) for each document. Thumbnails of embedded images are also generated and stored in the document descriptor. Only the document descriptors are passed on to one or more xFIND indexers.

The document descriptors harvested by a gatherer can be fetched by one or more authorised xFIND indexers. The xFIND indexer indexes a set of document descriptors and renders them searchable. An indexer may specialise, for example, in a particular topic or geographical location. Statistical information such as term frequency and discriminating terms (discriminators) are generated. Furthermore, trusted external systems (rating systems, ACF systems, etc.) are allowed to contribute additional metadata fields. The xFIND architecture also provides for the contents of a particular index to be replicated in whole or in part, in order to minimise network loads.

An xFIND broker is the starting point for user interactions. A broker can distribute its search queries to a particular set of indexers. The broker is also able to expand queries using a thesaurus. The results of a distributed search are collated and compiled into a uniform search result set. Brokers can be individually tailored for a division, a department, a group of employees or even for a single user as well as supporting particular topics. The xFIND broker provides standard search functionality (simple, extended, and expert search) through HTML forms. Search queries can combine both full-text and descriptive and evaluative metadata. The standard result set is a linear list ranked by relevance.

The enhanced metadata provided by xFIND derives from two sources: metadata automatically extracted from a document and quality rating metadata. Automatically extracted metadata are generated by the gatherer, which identifies the URL, mime type, file size, and creation or modification time of each document. Further information extraction depends on the type of the object, as well as the proper filter for processing the object. At the time of writing, plain text and HTML filters are available for textual information, and an image filter (supporting gif and jpg file format) for multimedia documents. Support for PDF and audio and video formats is planned. For HTML documents, common meta-attributes such as title, keywords, description, and language are parsed and processed by the gatherer. The full-text content is retained for indexing. Headings, links, and (thumbnails of) embedded images are extracted and retained as metadata. The gatherer also creates an electronic fingerprint of each information object, which allows a determination of the trustworthiness of information in case of replication.

Quality ratings can either be entered manually or imported from an external quality rating system. The integration of external metadata is also supported for non-textual objects such as images or video clips. To simplify the manual or semi-automatic provision of metadata, xFIND supports the definition of metadata for an entire document structure, a directory, or a particular document by inserting additional metadata files. More specific metadata overrides more general metadata.

4. The Standard xFIND Search Interface (Ranked List)

For the visualisation examples presented here and in the following sections, an xFIND broker specialising in the topic of Knowledge Management will be used. At the time of writing, the broker has access to an index of some 44,878
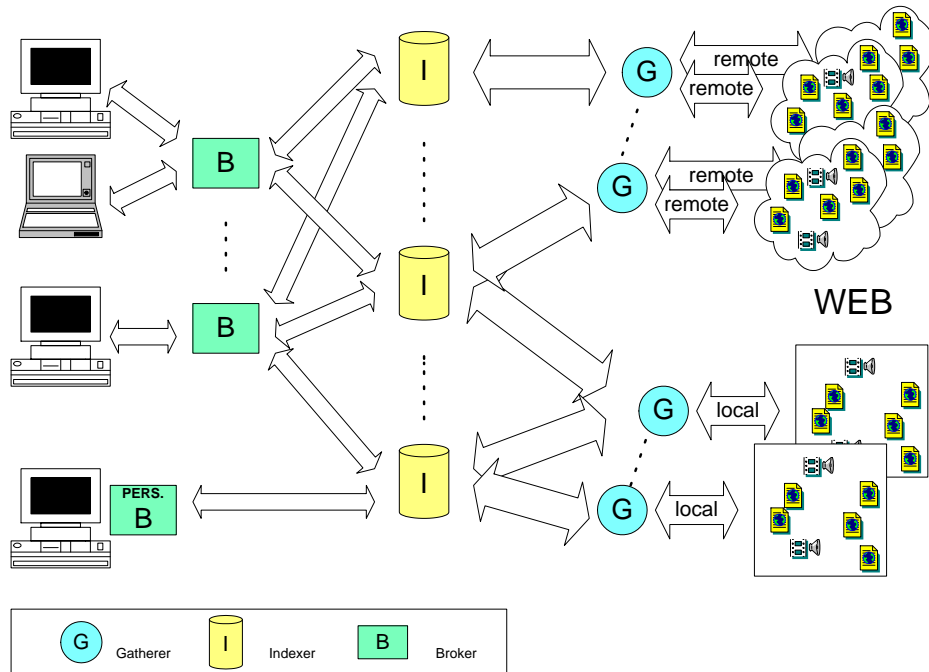
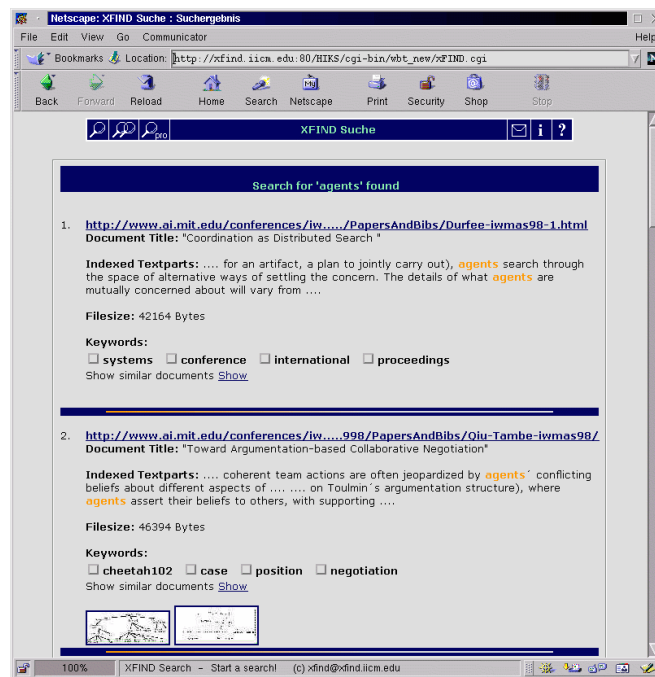**Figure 1.   The distributed architecture of the xFIND system.**

Figure 2.   The standard ranked list returned by an xFIND broker. Note the context of the query term "agents", the most important terms (discriminators) contained in each document, and the thumbnail images.

Daily
http://agents.www.media.mit.edu/groups/agents
http://bots.internet.com
http://agents.umbc.edu
Once a Week
http://www.uibk.ac.at/sci-org/voeb
http://www.ai.mit.edu
http://xfind.iicm.edu

Table 1. These sites are indexed for the broker on Knowledge Management, as well as a number of other sites at less frequent intervals.

documents, gathered from the sites shown in Table 1, as well as a number of other sites less frequently. The query used in each case is the single word "agents", for which there are 314 matching documents.

The default presentation of search results by an xFIND broker takes the form of a traditional ranked list. Figure 2 shows the first two matching documents. The most relevant document to the query "agents" is entitled "Coordination as Distributed Search". Note the context of the query term "agents" is shown for each document, the most important terms (discriminators) contained in each document are listed, and thumbnail images are shown for documents containing embedded images.

5. The Search Result Explorer

The Search Result Explorer uses a scatterplot (starfield display) to allow interactive exploration of the search result set based on the rich metadata associated with each object, in a manner similar to Envision[15]. Documents are plotted according to two of their metadata attributes (corresponding to the x and y axes). Further metadata attributes can be mapped to icon size and icon colour, allowing four dimensions of metadata to be visualised and explored simultaneously. If too many documents would be mapped to the same proximity, a group icon is used to represent that subset of documents. For group icons, the size and colour of the group icon is determined (under user control) by the maximum, minimum, median, or average value of the group's members. Since it is possible to zoom in on specific areas of the display, an overview window is provided in the lower left corner to help maintain context and orientation.

Figure 3 shows the first 210 (a user-configurable limit) of the 314 matching documents plotted by relevance on the y axis and document size on the x axis. The most relevant document is shown at the top of the plot. Here, the colour of each document icon is determined by the document's age,

from yellow older documents to white recent documents. Relevance is mapped to icon size, providing a redundant encoding. More relevant documents are both larger and towards the top of the plot. The most relevant document has been selected and its metadata displayed. It is the same document from 10th October 1998 entitled "Coordination as Distributed Search".

Figure 4 illustrates the interactive nature of the plot. Most of the rich metadata attributes provided by xFIND can be mapped to either axis or to icon size or colour. Figure 5 shows the result of the change. The y axis now corresponds to the modification date of the document, and document relevance is mapped to both icon size and icon colour (more relevant are orange, less relevant are white). It can be seen at a glance, that the most relevant documents are about a year old and reasonably small.

6. Dynamic Thematic Clustering with VisIslands

The VisIslands interface supports dynamic thematic clustering of search result sets, in a manner similar to SPIRE's Themescape[17] and its commercial successor Cartia[4].

The search result set is first pre-clustered using hierarchical agglomerative clustering (or optionally k-means clustering)[1]. The cluster centroids are then distributed randomly in the viewing rectangle. The documents belonging to each cluster, as determined by the initial pre-clustering, are then placed in a ring around each centroid. This arrangement is fine-tuned using a linear iteration force-directed placement algorithm derived from [7]. Documents similar to one another are attracted towards each other. After a certain cut-off point, the arrangement has stabilised, and each document contributes its weight to the height field of the grids within which it lies. Dense areas of many documents have corresponding peaks. The overall result is like a contour map of islands. A more three-dimensional visualisation of the islands is currently being implemented.

Figure 6 shows the islands visualisation for the first 200 documents matching our example query. Pre-clustering has identified 22 clusters. Cluster 17 concerning "mind, human, language" has been selected. Note the corresponding visual cluster of red documents in the islands display. Figure 7 shows the metadata associated with Cluster 17.

Figure 8 focuses on Cluster 22, containing 108 documents. Note that fine tuning with force-directed placement has attracted one document which pre-clustering assigned to Cluster 22 over towards the "webmaster" and "architectures" clusters. Zooming in on Cluster 22, Figure 9 shows that, in fact, many of the documents assigned to Cluster 22 on pre-clustering, should perhaps have been assigned to the cluster called "erläuterungen".
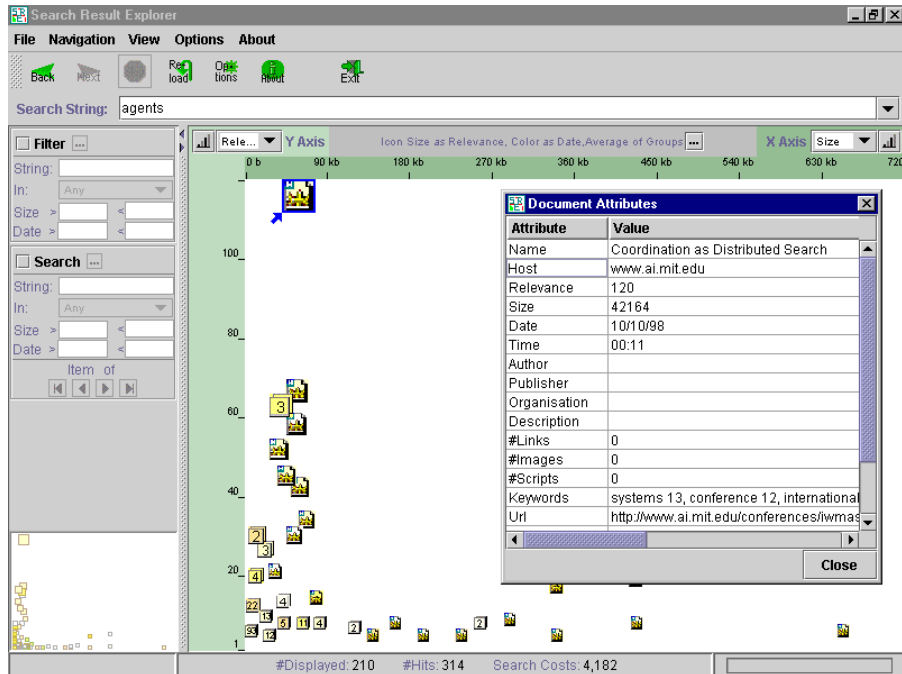
Figure 3. The Search Result Explorer plots search results along two axes. Here, document relevance is mapped to the y axis and document size to the x axis.
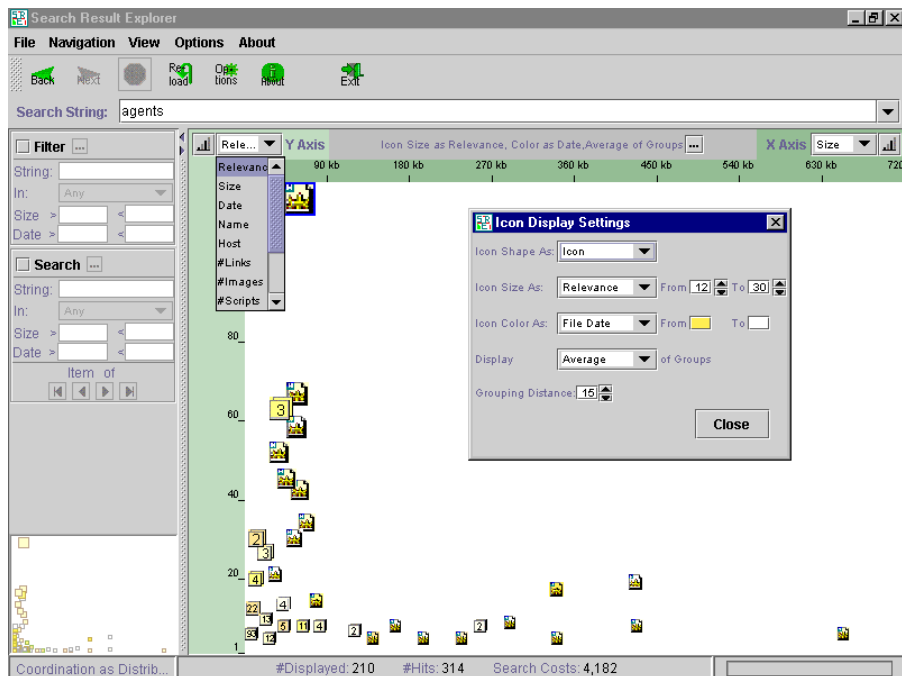

Figure 4. Search Result Explorer: Most of the rich metadata attributes provided by xFIND can be mapped to either axis, or to icon size or icon colour.
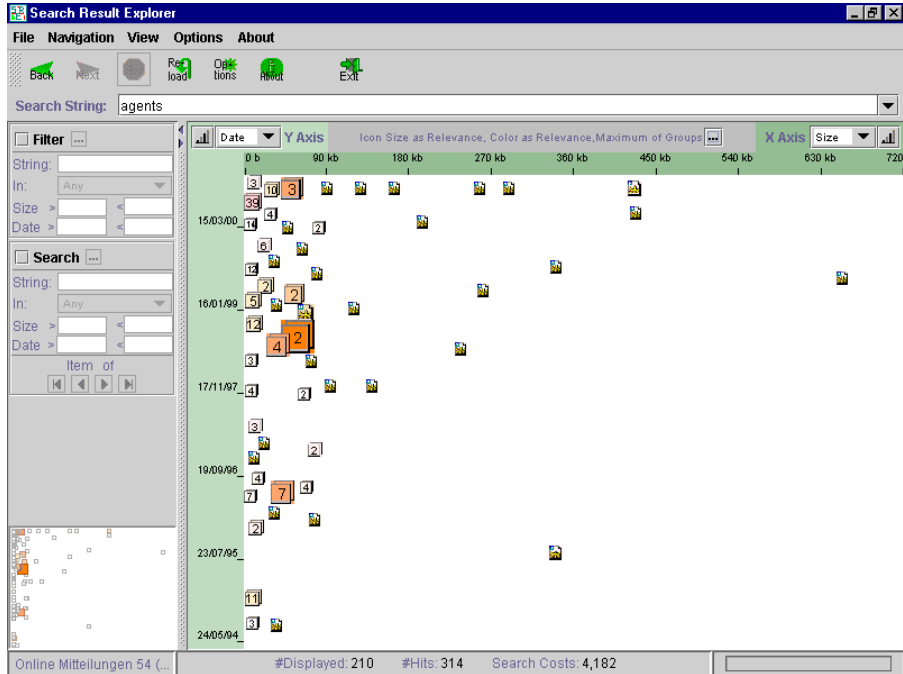
Figure 5. Search Result Explorer: An alternative view. Modification date has been mapped to the y axis. More relevant documents are now both larger and more orange.
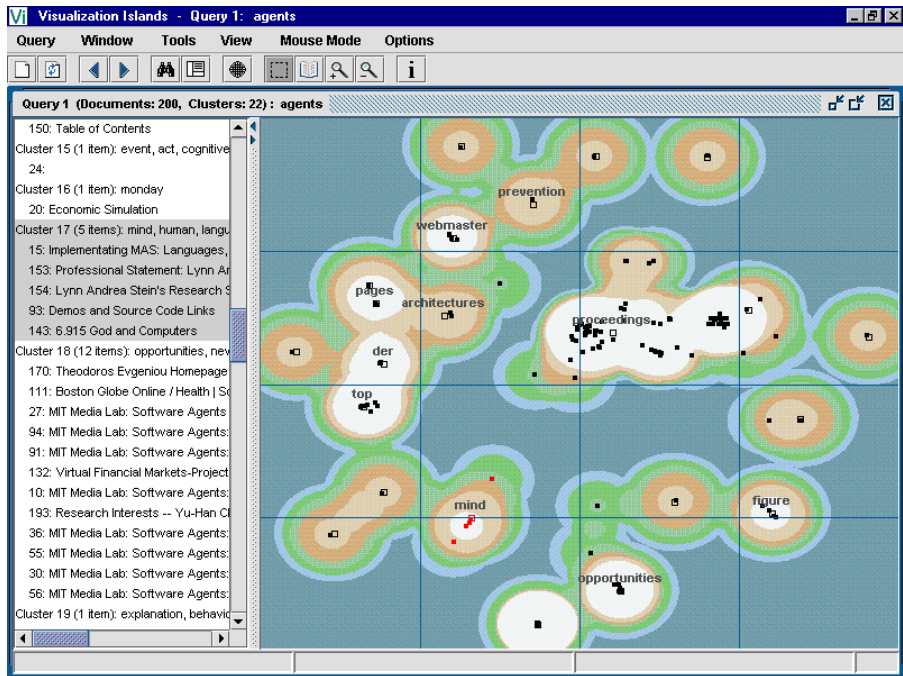

Figure 6. VisIslands: Pre-clustering using hierarchical agglomerative clustering has identified the 22 clusters shown in the left-hand pane. Cluster 17 has been selected.
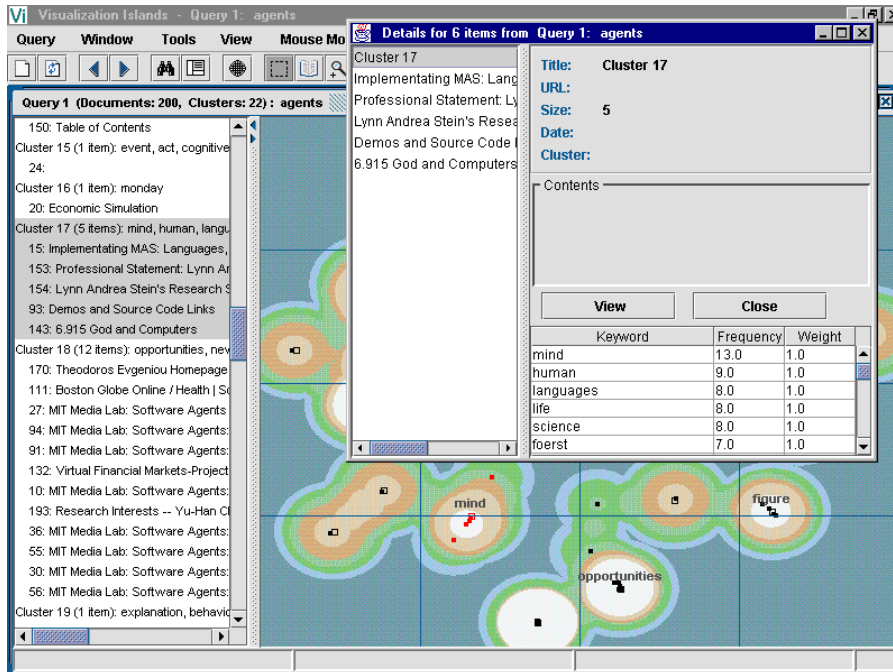
Figure 7. VisIslands: The metadata associated with Cluster 17 is displayed. Its most frequent terms include "mind", "human", and "language".
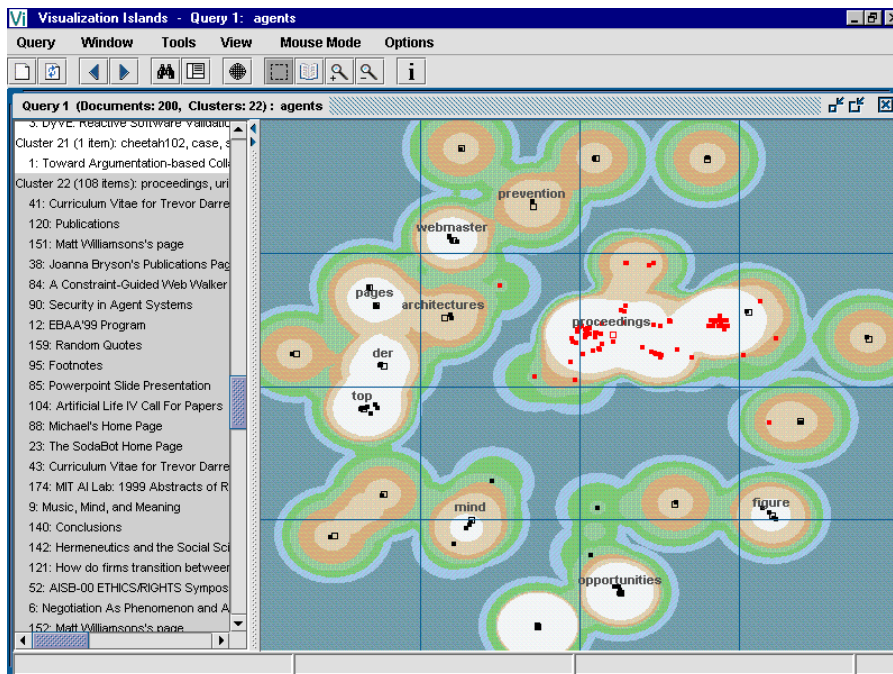

Figure 8. VisIslands: Cluster 22 deals with a variety of topics including "proceedings" and "conference".
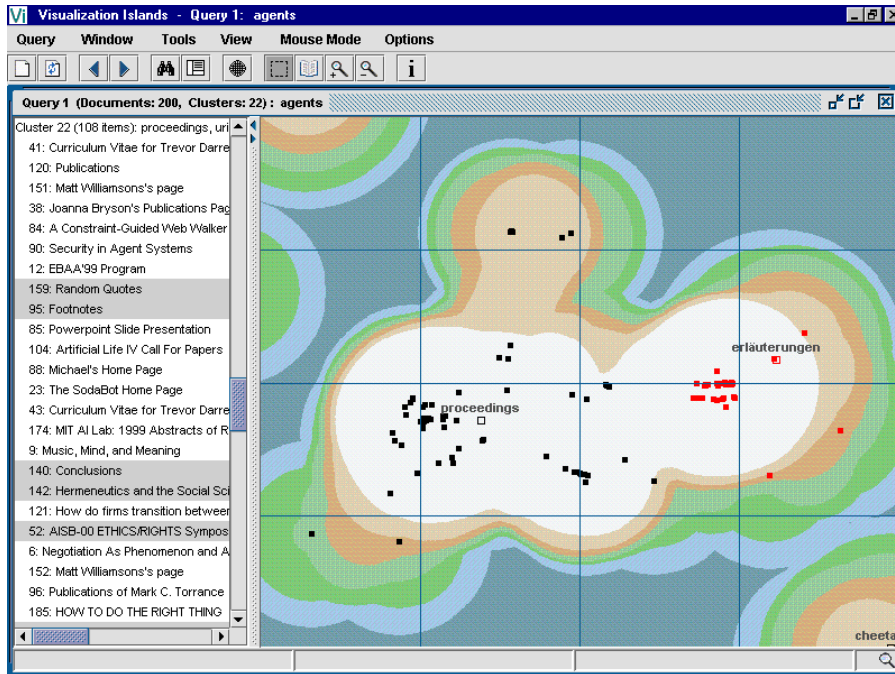
Figure 9. VisIslands: After zooming in on Cluster 22, the group of documents on the right-hand peak has been manually selected.

7. Future Work

The xFIND project is ongoing. Current work includes development of a utility to automatically detect the language of a document, so that stop list and stemming algorithms can be applied automatically. Also, a filter for indexing PDF documents is planned.

On the visualisation side, the Search Result Explorer and VisIslands are currently implemented as independent xFIND clients. A Master's thesis is just commencing to build an integrated search client incorporating query formulation and history, ranked result lists, and a framework for plugging in one or more synchronised visualisations, embracing both the Search Result Explorer and VisIslands.

8. Concluding Remarks

The xFIND infrastructure builds on the original work of Harvest and provides a rich, highly configurable knowledge management framework. The Search Result Explorer and VisIslands visualisations demonstrate the potential of information visualisation techniques applied to the exploration of search result sets. The electronic version of the paper includes full colour illustrations.

9. Acknowledgements

References

[1] M. N. M. Anil K. Jain and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, Sept. 1999. http://www.acm.org/pubs/citations/journals/surveys/1999-31-3/p264-jain/.

[2] ANSI/NISO. Z39.50-1995. Library of Congress, 1995. http://lcweb.loc.gov/z3950/agency/document.html.

[3] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, M. F. Schwartz, and D. P. Wessels. Harvest: A scalable, customizable discovery and access system. Technical Report CU-CS-732-94, University of Colerado,

Mar. 1995. `ftp://ftp.cs.colorado.edu/pub/cs/techreports/schwartz/Harvest.ps`.

[4] Cartia. Mapping the information landscape, 2000. `http://www.cartia.com/`.

[5] M. Chalmers. Using a landscape metaphor to represent a corpus of documents. In *Spatial Information Theory, Proc. COSIT'93*, pages 377–390, Boston, Massachusetts, Sept. 1993. Springer LNCS 716. `http://www.dcs.gla.ac.uk/~matthew/papers/ecsit93.pdf`.

[6] M. Chalmers. Adding imageability features to information displays. In *Proc. UIST'96*, Seattle, Washington, Nov. 1996. ACM. `http://www.dcs.gla.ac.uk/~matthew/papers/uist96.pdf`.

[7] M. Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proc. Visualization'96*, pages 127–132, San Francisco, California, Oct. 1996. IEEE Computer Society. `http://www.dcs.gla.ac.uk/~matthew/papers/vis96.pdf`.

[8] C. Gütl. Extended framework for information discovery: Programmers documentation. IICM, Graz University of Technology, 2000. `ftp://xfind.iicm.edu/pub/Archive/xFIND-0.95/Documentation/xFIND_ProDoc_V095.pdf`.

[9] C. Gütl. xFIND: Extended framework for information discovery. IICM, Graz University of Technology, 2000. `http://xfind.iicm.edu/`.

[10] C. Gütl, K. Andrews, and H. Maurer. Future information harvesting and processing on the web. In *Proc. European Telematics: Advancing the Information Society*, Barcelona, Spain, Feb. 1998. `http://www2.iicm.edu/~cguetl/papers/fihap/fihap_en.html`.

[11] Inktomi and NEC Research Institute. Web surpasses one billion documents. Press Release, Jan. 2000. `http://www.inktomi.com/new/press/billion.html`.

[12] C. Jenkins, M. Jackson, P. Burden, and J. Wallis. Automatic rdf metadata generation for resource discovery. In *Proc. WWW8*, Tononto, Canada, May 1999. IW3C2. `http://www8.org/w8-papers/2c-search-discover/automatic/automatic.html`.

[13] B. Kahle, H. Morris, F. Davis, K. Tiene, C. Hart, and R. Palmer. Wide area information servers: An executive information system for unstructured files. *Electronic Networking: Research, Applications and Policy*, 2(1):59–68, Spring 1992.

[14] O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. W3C Recommendation, 1999. `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222`.

[15] L. T. Nowell, R. K. France, D. Hix, L. S. Heath, and E. A. Fox. Visualizing search results: Some alternatives to query-document similarity. In *Proc. SIGIR'96*, pages 67–75, Zurich, Switzerland, Aug. 1996. ACM.

[16] B. Roslund and S. Lundberg. Combine harvesting robot: User's guide, 2001. `http://www.lub.lu.se/combine/docs/uguide.html`.

[17] J. A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(9):814–835, July 1999. `http://www.vistg.net/hat/Wise_draft/Ch5.Wise.html`.