TétraFusion: Information Discovery on the Internet

Francis Crimmins and Alan F. Smeaton, Dublin City University
Taoufiq Dkaki and Josiane Mothe, Institut de Recherche en Informatique de Toulouse

THE TÉTRAFUSION SYSTEM DESCRIBED HERE SUPPORTS KNOWLEDGE DISCOVERY FROM THE WEB BY HELPING USERS PERFORM DATA-MINING OPERATIONS ON SETS OF HARVESTED URLS. POTENTIAL APPLICATIONS RANGE FROM DOMAIN OVERVIEWING TO SCIENCE MONITORING TO COMPETITIVE INTELLIGENCE.

INFORMATION SEEKING CAN TAKE many forms, from a search to find an item known to exist to an explorative investigation into a new domain. When the collection of objects being searched is textual in nature, *information-retrieval* techniques can serve to find known items or items potentially relevant to a user's known information need. When a user explores a new domain, attempting to summarize the essence of an area previously unknown to the user, we call this *knowledge discovery*. For this task, where the user does not know the area, let alone know what to look for, IR techniques are of only limited value.

Knowledge discovery from text collections involves a user navigating through a text collection, reading, summarizing, browsing, and generally surfing through the collection to form an abstraction of what that new domain is all about. The task might involve conventional IR functions to direct the user's browsing and restrict the set of text objects being browsed. However, IR only partly supports the knowledge-discovery task, as all of the navigation, abstraction, and assimilation is left to the user, in addition to the responsibility of having to formulate an IR query that encapsulates the domain in which the knowledge discovery is to be performed. Furthermore, because we are usually operating within limiting time constraints, the knowledge-discovery task cannot always be based on an exhaustive analysis of many individual text documents. If the IR operation presents many documents to a user that are not relevant to the information-discovery task or fails to locate and present documents that are relevant to the task, the user-formulated domain abstraction will be inaccurate.

One of the most useful sources of information to arrive in recent times is the Web. What makes this resource particularly important for information discovery is that the range of domains it covers is huge, which means that no matter what subject you're researching, there is bound to be something on that subject on the Web. While the size and breadth of the Web's contents are a bonus for those performing information discovery, there is a price to pay—the difficulty of locating all and only those Web pages relevant to some topic. In essence, it is impossible to achieve both high precision and high recall in a Web search because of the size and diversity of the Web's contents and the fact that no single index of the Web's contents can claim to be exhaustive. Even the major Web search engines have limited coverage, so in using them we are always faced with the problem of choosing appropriate search terms.

As a support to the knowledge-discovery task, IR systems such as Web search engines have another important weakness in that they do not provide efficient aggregation of results. With an IR system, a user must read a full document to determine its contents, whereas in a knowledge-discovery task a user might prefer to have a time-saving summary or an overview of the document contents.

Together, these points contrive to illustrate that knowledge discovery is an important kind of information seeking, especially when applied to text documents on the Web. Rely-

1094-7167/99/$10.00 © 1999 IEEE

ing solely on IR systems to help users perform this task, however, is ineffective and an inefficient use of resources. As this article will show, our TétraFusion system addressed many of these weaknesses by building a number of layers of information harvesting and analysis on top of the simple IR functions.

Approach

TétraFusion combined two existing systems, Tétralogie and Fusion2, operating at IRIT in Toulouse, France, and Dublin City University in Dublin. Each system has been operational on reasonably large scales for some time. Like its predecessor Fusion, Fusion2 is a Web metasearch engine that has been serving user Web queries for more than two years.[1] Tétralogie is an information-mining tool that was initially developed for knowledge discovery from well-structured documents such as bibliographic records.[2] More recently, Tétralogie has evolved to take into account sources that are not well-structured, such as the Internet.[3] The two individual systems were first combined during 1997, producing the system we now call TétraFusion. Since then, each of the two components has been especially revised to yield a more integrated and combined tool.

Fusion2 is a metasearch engine that operates in client-server mode. A text-query input by the user goes to our Fusion2 server, which broadcasts it in parallel to six popular Web search engines. The top-ranked URLs from the six Web searches are combined using data-fusion techniques into a consolidated ranked list that Fusion2 presents to the user. As users view URLs, they are invited to mark some as being relevant to their query. After some have been so marked, the user can invoke a query-expansion process whereby the known relevant URLs are analyzed, word stems are identified and ranked in order of their potential usefulness as search terms, and the top-ranked such terms are presented back to the users. This procedure lets users manually select new search terms to add to the original query, which is then rebroadcast to the underlying Web search engines to generate a second or subsequent set of URLs for the users to view.

Uniquely among metasearch engines, Fusion2 incorporates the concept of relevance feedback, which allows a user's query, as represented by the set of search terms, to be refined in light of identified relevant URLs—a technique known to improve effectiveness in IR on closed collections of documents.

Tétralogie is a knowledge-discovery system that can use any structured textual information as an information source. Organizations such as Cedocar, Inria, Nestlé, Toulouse University Library, and others currently use the system, mostly for bibliographic and science-monitoring studies. Tétralogie has been supported by the French Secrétariat Général de la Défense Nationale and Conseil Régional de la Haute Garonne.

The Tétralogie system takes as input any raw structured information that has been retrieved or harvested using either a specific server or an IR engine. The system first pre-

FUSION2 LETS USERS OR OTHER AGENTS QUERY SEVERAL DIFFERENT SEARCH ENGINES IN PARALLEL AND USES DATA-FUSION CONCEPTS TO FUSE THE RESULTS RETURNED.

treats this raw information to reduce its volume and to homogenize it when necessary.[4] This pretreatment requires an information-extraction process to extract predefined elements such as author names, author affiliations, date and place of publication of the paper, and a set of representative phrases from the document text. Tétralogie then reduces documents using co-occurrence functions on these extracted elements and stores them as 2D and 3D contingency tables. These matrices in turn serve as the input to the mining process, which is used to discover global patterns and hidden information from the document set and to provide an overview of this discovered information. For example, the mining process lets users discover the main subdomains from within a field, collaborative work as evidenced by collaborations between authors and affiliations, a thematic map for the harvested documents, and so on. The mining process relies on classification methods and factorial-analysis functions. Finally, the system presents the data-mining results to the user in graphical form. This graphical representation does not consist solely of static objects but has associated with it facilities for users to focus on specific elements, to change the view they have on the results, or to perform drill-down or roll-up operations to and from documents. The Tétralogie software interface can incorporate the knowledge-discovery process into the framework of a loop including the four main tasks (information filtering, information reducing, information mining, and knowledge visualization)[5] to achieve useful user-directed knowledge discovery.

Compared to other knowledge-discovery systems, Tétralogie allows interactive, as opposed to static, knowledge discovery with the system's modules collaborating for improved efficiency. The information sources can be either factual data or free text. The pretreatment of the free text includes techniques commonly used in IR (stemming, term conflation, and so on), while the pretreatment of other structured elements is also performed. In addition, Tétralogie lets users analyze temporal information and have 4D visualization, a powerful representation that lets users visualize more subtle and hidden types of information.

A natural evolution of Tétralogie is to apply its information mining to URLs harvested from the Web. We would like to ensure that these URLs are diverse—not just the ones that contain the most number of occurrences of the user's initial search terms. This criterion would not be achieved by using a conventional Web search engine, but is achieved by integrating Fusion2 (with automatic rather than manual relevance feedback) and Tétralogie into the system we call TétraFusion. Figure 1 shows a functional overview of TétraFusion.

Harvesting URLs from the Web with Fusion2

Fusion2 lets users or other agents query several different search engines in parallel and uses data-fusion concepts to fuse the results returned, thus improving the search's overall quality. Since the original version of the system was launched in August 1996, it has served thousands of Internet searches.

Metasearch engines accept queries just like any other search engine, but instead of maintaining their own database of indexed documents, these services rely on other external search services to provide the information necessary to fulfil user queries. They all

operate in essentially the same way, querying some underlying Web search engines in parallel to answer user queries, but differing in the processing they perform on the results returned by the Web search engines before presenting them to the user.

As we mentioned, a user interacts with the Fusion2 system by entering a query, which is sent to the Fusion server, where it is broadcast in parallel to six Web search engines. The results returned from each engine are parsed to extract document titles, URLs, and rank positions.

After a time-out period, the system performs a data-fusion operation on the URL lists returned from search engines at that point. This data fusion performed on our server machine relies on rank position rather than retrieval status value (RSV) or URL score, because not all search engines return scores. Duplicate objects have their ranks summed, and objects are penalized if they have not been retrieved by a particular search engine. The system then takes the fused ranking of URLs and sends it back to the client for display to the user.[1]

Fusion2's relevance feedback and query expansion let users modify their search by marking selected documents as being relevant to their information need. These URLs then return to the Fusion server. Such pages are retrieved from the Web, after which their text is analyzed by removing HTML tags and stopwords and the remaining text is stemmed. From this, the system extracts a list of candidate search terms. Once these have been ranked, the top-scored return to the user for display. When the user receives the list of candidate extra search terms, he or she can select some, or none, of these. They are added to the original query as a form of query expansion, and the user is asked to rerun the expanded query.

In the URL harvesting approach used in TétraFusion, a sequence of searches using the query are sent to the retrieval engines to retrieve the top 10 documents, numbers 11 to 20, 21 to 30, and so on. This initial querying process is complete when the top-ranked 100 URLs have been returned from each engine. This approach is necessary because the underlying engines return 10 documents at a time in response to incoming queries. Figure 2 shows the flow of control of the URL-harvesting system.

TétraFusion sends the initial set of returned documents for term analysis as described previously and automatically adds the top 20 search terms from these URLs to the original query in a process called *pseudo relevance feedback*. This technique causes query latency, which probably explains why it is not used in conventional Web searching. Nonetheless, this technique has proven be useful in improving IR precision—in TREC, for example— which is precisely why we used it in TétraFusion (see http://trec.nist.gov). The expanded query then goes out, and 100 URLs are once again retrieved from each engine, as shown in Figure 2, this time based on the expanded query. The final set of URLs returned from these searches go back to the client.

This approach leverages existing systems to make efficient use of network bandwidth and resources. This was the idea behind the Harvest system, where *brokers* provided the indexing and query interface to the gathered information.[6] These brokers achieve this step by requesting information from information *gatherers* and other brokers. In our system, we use the existing Web search engines, which have the resources at their disposal to attempt to index the Web. By querying them, we get a pool of URLs for the Tétralogie system to work with.

Analysis of URLs in TétraFusion relies on the same pretreatment of the pages, data mining, and finally visualization for the user via the same visual representation that the Tétralogie system uses—except that in TétraFusion, we use different predefined elements to extract from Web pages.

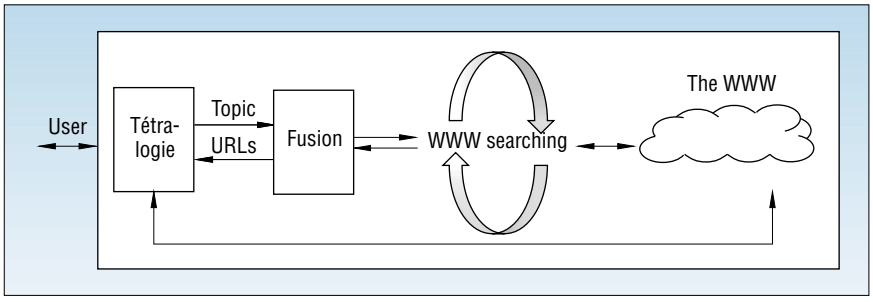


Figure 1. Functional overview of TétraFusion.

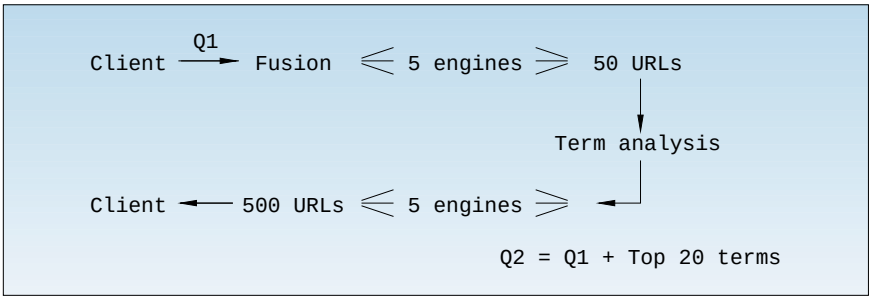


Figure 2. Flow of control for URL harvesting.

Analyzing pages with Tétralogie

Tétralogie provides different complementary tools for mining Web pages and extracting information from them. We'll now take a step-by-step look at the way Tétralogie analyzes Web pages.

Information pretreatment. Once URLs have been identified and the pages downloaded directly to Tétralogie, they must be pretreated to turn raw information into a more structured form on which mining functions can be applied. This pretreatment includes information extraction, filtering, and summarization tasks.

Information extraction. Information extraction consists of extracting predetermined elements of information. In previous work with the Tétralogie system, we have extracted information from semistructured bibliographic databases where the structure of the documents is marked up using a set of predefined tags.[4] In the context of the Web, most of the documents are written in HTML, which provides tags to mark up structural elements

of the documents (such as metatags and reference tags). We use these tags to extract some elements and also extract representative phrases from the document content.

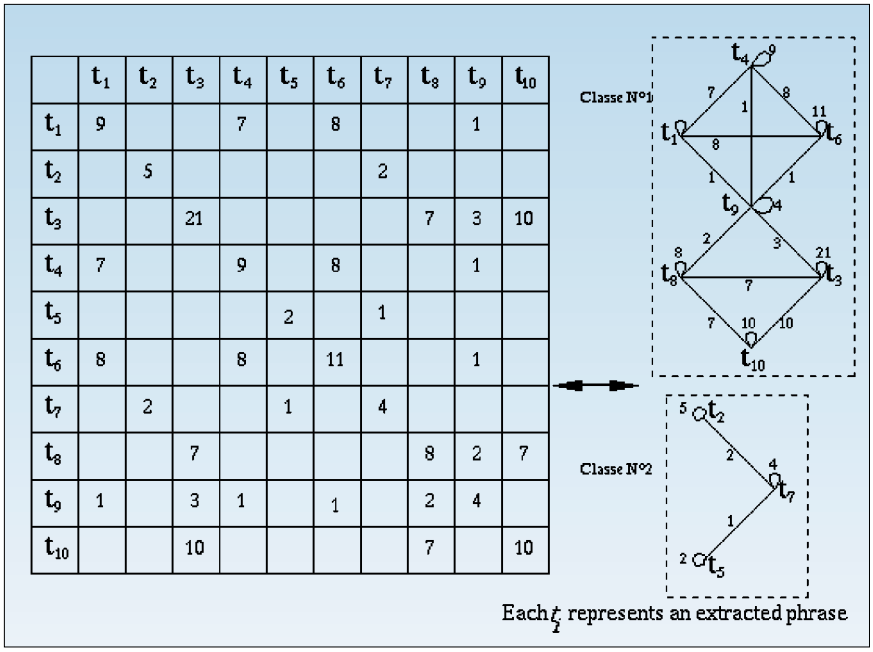
- *Phrase extraction*. In IR, the use of phrases as indexing items can improve retrieval effectiveness. In a study of the usefulness of phrases identified syntactically versus those identified by simple word co-occurrence, Mandar Mitra and his colleagues found statistical and syntactic phrases had about the same level of improvement.[7] In our approach, we base phrase detection on a statistical technique. When the same combination of terms occurs several times in documents, it is considered to be a phrase because it is most likely to correspond to a real concept. Phrases can be composed of more than two words, depending on their word-combination frequency. In addition, we use a stop-list to delete nonuseful terms and Porter's stemming algorithm to conflate different variants of a word to a single form.
- *Citation and site extraction*. We use the tags provided by HTML to mark up the hyperlinks to extract either the full URLs cited (for example, http://www.irit.fr/SSI/ACTIVITES/EQ_SIG/Welcome_a.html) or the server cited (http://www.irit.fr) in the harvested documents. The latter element gives direct information on the organization cited (IRIT, France).
- *Metainformation extraction.* When available, metatags such as the author tag (`<META NAME=Authors CONTENT="Dupont">`) are used to extract predefined elements from the documents. Unfortunately, such tags are still not widely used in Web documents, although as this changes, there is great potential for using this information.

Information filtering. While values are extracted for predefined elements from the harvested documents, they can also be filtered positively or negatively. For example, it is possible to give a stop-list of terms or phrases that should not be considered during the extraction phase (negative filter), while it is also possible to enlist only phrases that must be considered as a positive filter.

Information summarization. After filtering, our system summarizes remaining information into the form of a contingency table that is a powerful 2D-knowledge representation.[5] A 2D contingency table summarizes a weighted relationship between two kinds of extracted elements (the relationship `Is_Treated_By` between phrases and authors). Let $T = [t_{i,j}]$ be such a contingency table; $t_{i,j}$ corresponds to the number of documents where the element value *i* of the first element (the *i*th phrase) and the element value *j* of the second element (the *j*th author) co-occur. In fact, the two elements involved in a relationship can be of different types or of the same type (for instance, the copublishing relationship among authors).

Mining functions. Tétralogie integrates different mining functions to accomplish classification and correlation-detection tasks. For classification, it uses three kinds of classification, both supervised and nonsupervised. For each of these, the elements to classify correspond to the rows of the contingency tables and the distance measure used is based on variables represented as the columns. On the other hand, the correlation discovery relies on factorial analysis. These are the methods used:

- *Hierarchical ascendant classification* is a nonsupervised classification method. At the first step, Tétralogie reduces classes to singletons, where each class contains a unique extracted element value. It then aggregates these classes recursively in pairs according to their distance with the closest two classes merged. This process repeats until all the elements belong to a single class. The result of this is a tree that can be cut at any level according to the number or the size one wants for the classes. Any kind of distance function can serve to evaluate how close two classes are, such as Euclidean or χ^2.
- *Classification by partition* (CP) is a supervised classification where the user gives the number of classes expected as well as a representative element of each class. CP is an iterative process, and at each iteration, each item to be classified is associated with the closest class representative. Then, for each class, the representative element is recomputed (for instance, the center of gravity). The iterations stop when the classes are stable—when there are no more element shifts from one class to another.
- In *graph-based classification*, the contingency tables of co-elements (such as phrase-to-phrase tables) are turned into graphs where the vertices correspond to the extracted values. Two vertices are linked with an edge if and only if their cell value is larger than a threshold, which the user can select (see Figure 3). With regard to factorial-analysis methods, we can view the contingency tables as items



	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
t_1	9			7		8			1	
t_2		5					2			
t_3			21					7	3	10
t_4	7			9		8			1	
t_5					2		1			
t_6	8			8		11			1	
t_7		2			1		4			
t_8			7					8	2	7
t_9	1		3	1		1		2	4	
t_{10}			10					7		10

Figure 3. Example graph-based classification applied to a co-phrase contingency table.

(rows) described in an *n*-dimensional space with each dimension corresponding to a column.[8] Factorial-analysis methods reduce the number of these dimensions to something more manageable without loss of information.

- *Principal component analysis* (PCA) reduces data dimensionality into spaces that are the most important as determined by the eigenvalues of the variance/covariance matrix of the table using Euclidean distance.
- *Correspondence factorial analysis* (CFA) is performed the same way as PCA is, except that the distance measure used is the X^2 distance. This has the mathematical property of allowing representation of both rows and columns of the tables in the same space and thus can show the associations between rows and columns.

Graphical representation and manipulation. A graphical representation of complex information is intuitively easier for a user to comprehend, so all results obtained from the document analysis are displayed in graphical views. This graphical approach is similar to that taken by the Document Explorer system, which is used to visualize Web document structure based on an analysis of the document content.[9] The summarized information stored in contingency tables and the graph-based classification is displayed in a spreadsheet tool. The tree resulting from a hierarchical ascendant classification can also be displayed, while the factorial-analysis functions yield a set of points in a space.

We developed a visualization tool that can visualize up to four dimensions simultaneously. The first two are the *x* and *y* dimensions on screen. The third is obtained using a perspective view that magnifies close objects or points and reduces remote objects, while also modifying distances between objects depending on the closeness of the objects to the viewer. The fourth dimension is obtained using several levels of gray to represent objects, each level of gray corresponding to a value on the fourth axis. The disposition of the axes—that is to say, the azimuth used to observe the data—is chosen to visualize as clearly as possible the partition of the set of points.[2]

To fully exploit the advantages of a 4D visualization and let users really explore information relationships, we have added dynamics to the display that let users modify their point of view on the results set using several kinds of methods or actions. These actions induce transformations on the representation of the set of points and include the following:

- *Zooming around a given point*. The user can focus on some of the points that appear most interesting.
- *Scanning the set of points using several angles*. Two points can seem to be close to each other according to a given point of view, whereas in fact they might be far away. A scanning by rotation can make the correlations between the points visual. The user can chose a continuous and animated rotation in the space.
- *Modifying the visualization space*. To increase the amount of visualized information, it might be useful to incorporate other axes into the visualized space. We provide a function that produces a sliding of the axes with the first axis no longer represented, the second axis becoming axis number 1, and so on. This axis sliding can repeat until the whole information content is visualized and the user can select any combination of axes.

Another aspect of visualization is the selection of subsets of elements for deeper analysis. A user can graphically isolate some data subset and apply the data-mining functions only on the documents filtered according to this selection of information, thus generating a new graphical representation for user-controlled visualization. This last point illustrates the whole approach we have taken to mining and analyzing harvested information. It is the user who chooses the combination of mining and visualization processes (filtering-mining-visualization), making the Tétralogie's knowledge-discovery process flexible and adaptable. Figure 4 shows an overview of page analysis in Tétralogie.

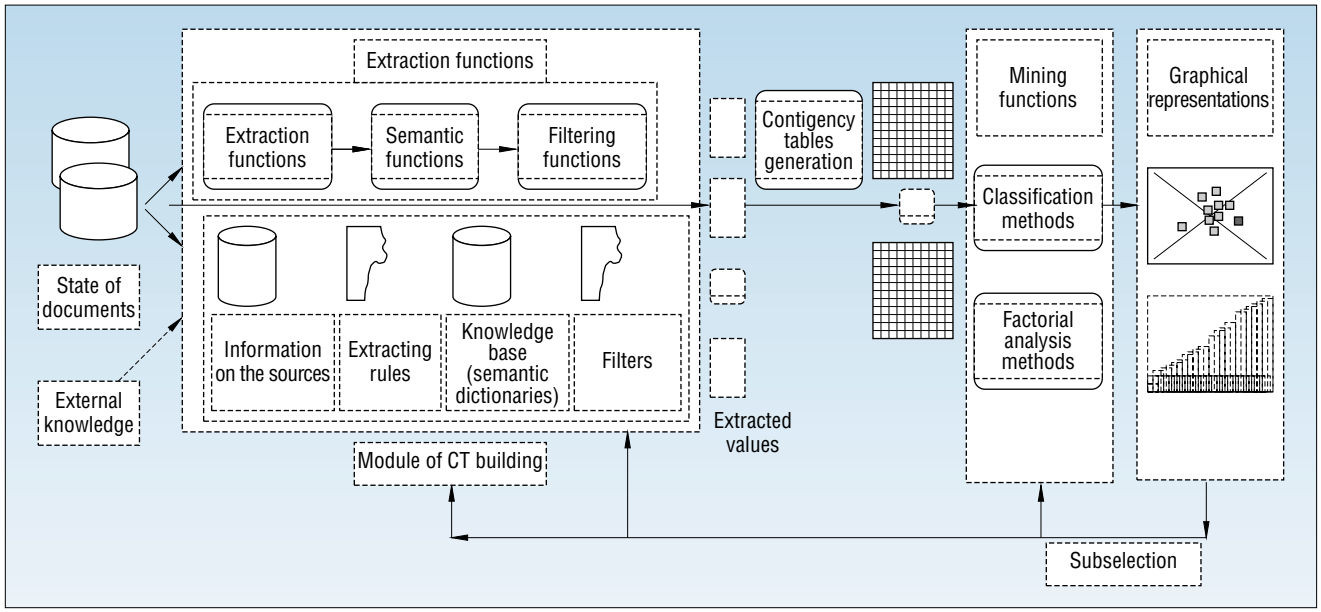


Figure 4. Page analysis in Tétralogie.

See the "Sample session" sidebar for an example of Tétralogie's operations.

BECAUSE TÉTRAFUSION OPERates on the Web, it is not constrained by subject domain or even by the language in which the Web pages are written. The TétraFusion system has managed to leverage two existing systems, Tétralogie for data mining and Fusion2 for metasearching, into one integrated tool that is particularly well-suited to the knowledge-discovery task. Operationally, TétraFusion would be scalable to really large usage because the component that determines URLs for harvesting, Fusion2, already supports many thousands of Web searches and can easily support many more, while the actual URL harvesting, interactive data mining, and knowledge-discovery component, derived from Tétralogie, is a personalized tool running on a user's client machine. At present, the TétraFusion system is not publicly available, although we are planning to include a version for public use in future developments. ■

Acknowledgments

This work was facilitated by a joint CNRS-Forbairt Exchange program.

References

1. A.F Smeaton and F. Crimmins, "Relevance Feedback and Query Expansion for Searching the Web: A Model for Searching a Digital Library," *Proc. First European Conf. Digital Libraries*, Springer-Verlag, Berlin, 1997, pp. 99–112.

2. J. Mothe, T. Dkaki, and B. Dousset, "Mining Information to Extract Hidden and Strategic

Sample session

To illustrate TétraFusion in operation, let's look at the processing for a sample query. The area we explore is represented with the initial query *"data mining" and "information mining,"* which returned a set of 500 URLs after iterative Web searching with Fusion2. These were downloaded with only three documents not harvested due to connection failures. Before starting the information-pretreatment process, the 495 documents (two further documents were eliminated due to HTML syntax errors) were filtered to keep only the URLs containing at least one of the two phrases "data mining" or "information mining."

Information extraction. We performed the pretreatment on the 245 remaining URLs, extracting different predefined elements. Table A displays the results.

Summarization (contingency tables). To discover the topics of interest for each site and identify the important sites, we crossed extracted phrases and document servers. The crossing table lets us list the most strongly related topics for each site, and the results obtained give the phrases that characterize each source. (See Table B for an example.)

We can assess the summarized information by accessing the corresponding Web sites. As an example, when connecting to http://www.is1.co.uk, we see that this site offers software and consulting services for data mining.

We then applied several mining functions to this crossing to provide complementary knowledge from the analyzed documents.

Correspondence factorial analysis. Figure A displays the results of the CFA. This screenshot only partly conveys the detailed information shown in a live interaction. It shows the distribution of both the topics (extracted phrases) and the Web site in a space computed according to the topics. The distance between items is representative of probable item correlations. At least five distinct groups appear in this screenshot. Two distinct group contents are displayed in Figure A. Taking the top right group, we see several phrases and two Web sites. That means that these two sites are mostly characterized by the linked phrases. Both sites are calls for participation on intelligent agents, knowledge management, and so on.

Hierarchical ascendant classification. We can display another complementary view of the data aggregated from downloaded Web pages consisting of clusters of sites crossed with phrases. The result of the classification is a tree (Figure B1) that the user can navigate by cutting at any level to visualize the content of nodes (Figure B2). This classification tree can help the user visit the minimum number of sites and still have the largest overview. Suppose that we want to have the largest

Table A. Pretreatment results.

Attribute	Number of different values	Meaning
URLs	245	Document URL
Site	183 (36 sites are referenced more than twice)	Document site
Title	1,943 phrases and single terms	Phrases extracted from the document titles (tag `<Title>`)
SubTitles	3,244 phrases and single terms	Phrases extracted from the document subtitles (tags `<H1>` … `<H7>`)
Links	3,800 (1910 external http references)	URLs referenced in the document set
Links-site	624 occur more than twice	Site of URLs referenced in the document set
Text	12,524 phrases and single terms	Phrases extracted from the rest of the textual information

Table B. Crossing table results.

URL	Topics
www.data-miners.com	Technology, data-mining, data-warehousing, business, product service, company, marketing, order-book
www.isl.co.uk	Data-mining-consultancy
www.datamining.com	Discovery, decision-support
www.santafe.edu	Technology, data-mining, kurt-thearling
www.europe.digital.com	Technology, information, solution, digital-data-mining, digital, burgandy-bar

Information," *Proc. Computer-Assisted Information Searching on Internet (RIAO97)*, C.I.D., Paris, 1997, pp. 32–51.

3. J. Mothe, "Internet-Based Information Discovery: Application to Monitoring of Science and Technology," *J. Research in Official Statistics*, C.I.D., Paris, No. 1, 1998, pp. 17–30.
4. C. Chrisment et al., "Extraction et Synthèse de Connaissances à Partir de Données Hétérogènes," *Ingénierie des Systèmes d'Information* (Knowledge Extraction and Synthesis from Heterogenous Collections), Vol. 5, No. 3, Aug. 1997, pp. 367–400.
5. U.M. Fayyad et al., *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, Calif., 1996, pp. 329–349.
6. C. Bowman et al., *Harvest: A Scalable, Customizable Discovery and Access System*, Tech. Report CU-CS-732-94, Dept. of Computer Science, Univ. of Colorado, Boulder, 1995.
7. M. Mitra et al., "An Analysis of Statistical and Syntactic Phrases," *Proc. Computer-Assisted Information Searching on Internet (RIAO97)*, C.I.D., Paris, 1997, pp. 200–214.
8. J.P. Benzecri, L'Analyse de Données (Data Analysis), Vols. 1–2, Dunod Edition, 1973.
9. R.H. Fowler, W.A.L. Fowler, and J.L. Williams, "3D Visualization of WWW Semantic Content for Browsing and Query Formulation," *Proc. WebNet 96*, Assoc. Advancement of Computing in Education, Charlottesville, Va., 1996.

Francis Crimmins is a researcher in the School of Computer Applications at Dublin City University. His interests include filtering, resource discovery, summarization, and networked IR. His previous experience includes working as a networking and

overview of information mining and data mining but still cannot afford to visit any more than, say, 18 sites. If the sites can be divided into 18 classes, we can visit one site representing an element from each class. In the example, one of the documents from the set we have highlighted can give a quick idea of the group content.

Graph-based classification. To discover subtopics within the document set (that are all on the same general topics) and their specific terminology, we studied cophrase crossing. Figure C displays the co-occurrence matrix as well as its global view (in black and gray) where each point corresponds to a non-null value in the matrix. Clusters have been highlighted, and a cluster's content can be listed by graphically selecting it. Notice that the selected elements can then in turn be used to filter the document set. Thus, the user can drill down to visualize the underlying information filtered by the selected elements.

These analyses open up the area of the initial query to let the user explore terminology, search among Web sites, and view documents in a knowledge-discovery task in what might be a completely new domain for the user.

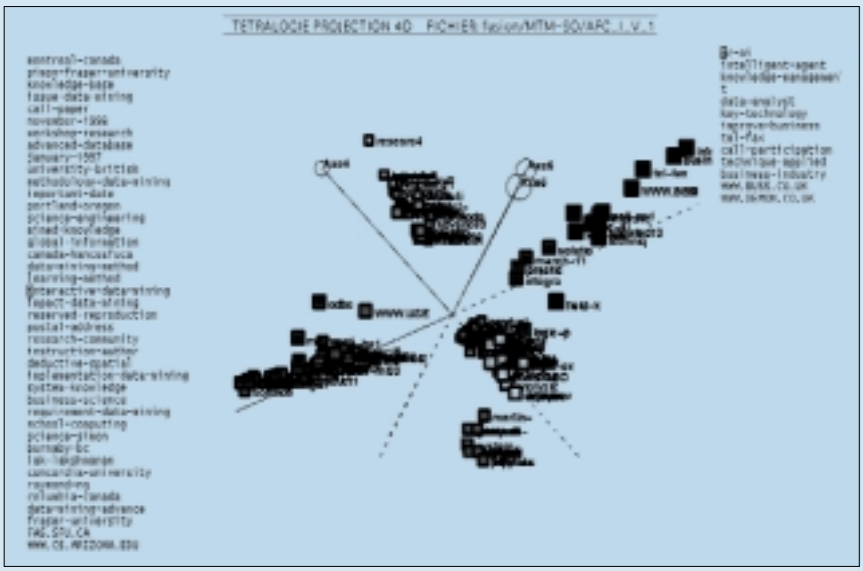


Figure A. Results of a CFA on phrases and sites crossing displaying site specifics. This figure displays some of the specific clusters that have been found.

This example illustrates the kind of results a user gets when using TétraFusion, but static screenshots cannot capture the true interactive nature of the interface. The tool we present meets the criteria that have been defined for visualization interfaces, such as the possibility for the user to focus attention on certain elements of the visualized information and to have several levels of abstraction in the information being displayed. On the other hand, the information-harvesting and pretreatment phases can be time-consuming, depending on the response time of the queried servers. Thus, the overall system has a query latency that is traded against a high degree of document analysis. The selection of URLs to be analyzed, including the process of pseudo relevance feedback, could be speeded up only by using a central index that stores word occurrence and other information for each indexed document. This information is generally unavailable except by harvesting the URLs directly, which is what causes the query latency in the first place. Operating on such a central index, as the Tétralogie system does, TétraFusion would have a very much reduced execution time.

Different domains have been mined using the Tétralogie system. Before an organization analyzes a new domain, it evaluates the system's performance on a known corpus to examine the results obtained. Of course, this assessment is difficult to quantify and remains a qualitative criteria. For example, Tétralogie has been successfully used to evaluate the public image of a well-known international corporation by mining all articles posted in newsgroups about it. The articles were filtered according to the corporation's name, acronyms, and distributed products. Among the patterns that emerged was the detection of attacks from competing organizations. Another corpus mining has shown the influence of the use of omega3 (a fish oil) on the arteries and heart, which is a research issue discussed in many publications. In mining astronomical publications, we have highlighted evolutionary trends in astronomical literature—both copmpleted and new research programs, new terminology, observations, and so forth. In general, returned experiences from Tétralogie end-users show that classifications and clustering methods are easy to use and

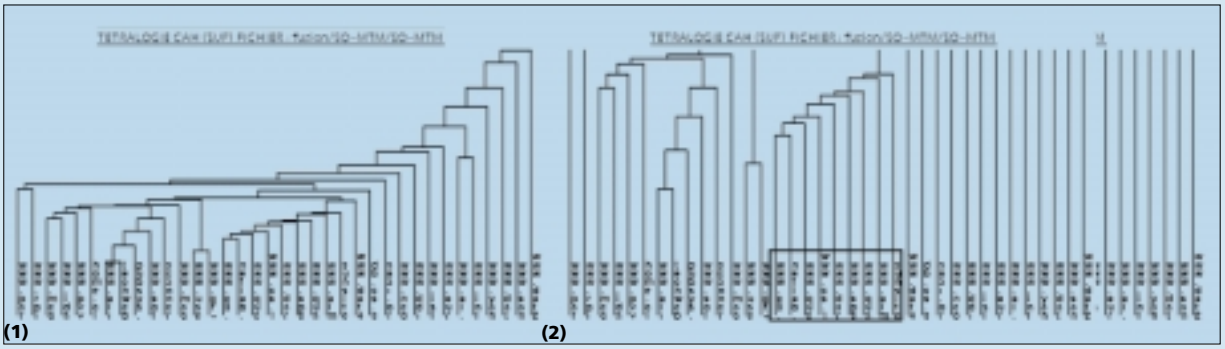


Figure B. Sites and phrases classification tree: (1) before any cutting, (2) after cutting.

understand. The factorial-analysis methods and visualization seem to need a little more training to be easily and fully interpreted by users, but they seem to be ready to train because of the increase in value obtained using this technique.

As the discontent grows with navigating and finding information on an increasingly complex and growing Web, the need for tools such as ours and the tolerance of increased execution time will also increase.

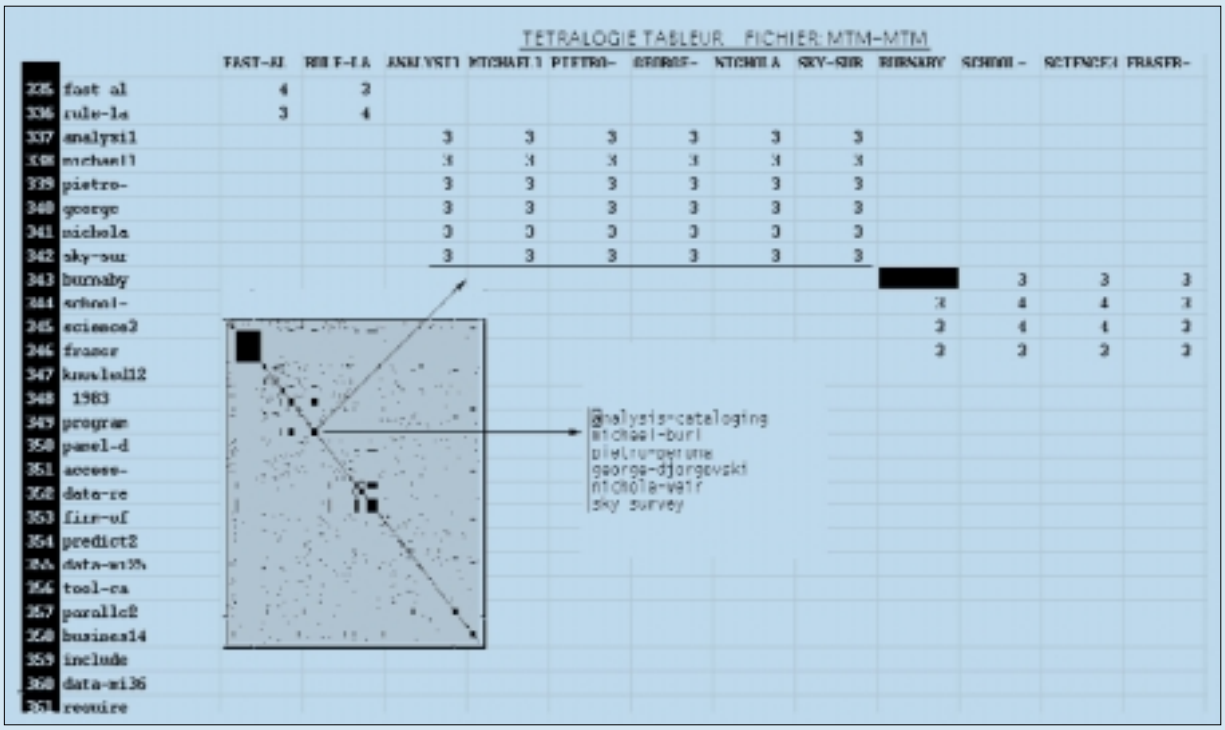


Figure C. Result from crossing phrases with phrases.

security specialist with Sun Microsystems. He holds a BSc and MSc in computer applications from Dublin City University and is pursuing research as a graduate student. Contact him at the School of Computer Applications, Dublin City Univ., Glasnevin, Dublin 9, Ireland; francis@compapp.dcu.ie.

Taoufiq Dkaki is an assistant professor at Robert Schuman University of Strasbourg, where he teaches IR and competitive intelligence. His research interests include knowledge discovery and competitive intelligence. He graduated from the Engineer National Superior School of Computing of Toulouse and holds a PhD in computer science from Paul Sabatier University of Toulouse. Contact him at the Institut de Recherche en Informatique de Toulouse, 118, Route de Narbonne, 31062, Toulouse Cedex, France; dkaki@irit.fr.

Josiane Mothe is an assistant professor at the Toulouse I University and Directeur d'étude at the Institut Universitaire de Formation des Maîtres de Toulouse. She carries out her research at the Institut de Recherche en Informatique de Toulouse. Her interests cover IR and text mining. She holds a PhD in computer science from Paul Sabatier Toulouse University. Contact her at the Institut de Recherche en Informatique de Toulouse, 118, Route de Narbonne, 31062, Toulouse Cedex, France; mothe@irit.fr.

Alan F. Smeaton is a professor of computing and head of the School of Computer Applications at Dublin City University, where he also leads the Multimedia Information Retrieval research group. He is an associate editor of the *ACM Transactions on Information Systems* and *Information Retrieval*. His interests cover IR from text, image, audio and digital video. He holds a BSc, MSc, and PhD from the National University of Ireland and is a member of the ACM, IEEE Computer Society, and the Irish Computer Society. Contact him at the School of Computer Applications, Dublin City Univ., Glasnevin, Dublin 9, Ireland; asmeaton@compapp.dcu.ie.