Information mining: use of the document dimensions to analyse interactively a document set

J. Mothe [(1),(2)], C. Chrisment[(1)], T. Dkaki[(3)], B. Dousset[(1)] D. Egret[(4)]

(1) Institut de Recherche en Informatique de Toulouse
(2) Institut Universitaire de Formation des Maîtres de l'académie de Toulouse
(3) Institut Universitaire Technologique de Strasbourg
(4) Centre de Données astronomiques de Strasbourg

(1) Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex
Fax : 33 5 61 55 68 52 Tel : 33 5 61 55 63 22
e-mail : mothe@irit.fr

Abstract

This paper introduces a new interface that integrates document mining. A key point is the co-operation of different modules with which the user interacts in order to visualise in a global way a document set. This visualisation is done according to different point of views, given by the different document dimensions. The graphical visualisations are animated and 4-Dimensional. The general mining process follows the framework of a Knowledge Discovery in Databases process. We present the several phases implied by document mining and show how this technique can be used based on a document set.

1. Introduction

Most of the Information Retrieval Systems (IRS) retrieve a list of document references and titles in response to a user's query. Then the user has to sequentially determine the relevance of each returned reference and to access the corresponding document. Finally she/he reads the document if considered of interest.

With the growth in the amount of electronic information, more and more documents are available on a given subject. As a result, more and more documents match the user's queries and the lists of retrieved documents become more and more time consuming to be processed by the users. Some techniques have been proposed to ease the user in evaluating the interest of a document or of a group of documents. One of these techniques is document summarisation. In that case, the system provides the user with a summary of the retrieved documents in addition to their references. A summary is a brief account giving the main points of the whole document. Different approaches have been used to summarise the documents. One kind of approach is to create natural language summaries. In that case, the summaries can be generated by extracting the most important sentences from the initial document [31]. Another approach is to extract pre-defined elements of information that give the main points of the document. Information extraction systems [32] can be used to achieve this goal. They analyse texts in order to extract pre-defined elements e.g. person or organisation names, dates. To some extend the indexing task in IRS can be viewed as a summarisation task as the goal of this process is to make a characterisation of the document content through a reduction of the initial document length. Summaries can be generated combining these two kinds of techniques (extraction of pre-defined elements, creation of natural language summaries). In that case, the summaries include relevant sentences and factual information as person names, dates [15]. Alternatively, new sentences can be generated using important extracted items and natural language processing [27].

In the end, the approaches used in IR share several main features. First, the summaries are mainly computed according to a single document dimension that is its content. Second, these summaries are used to help the user to determine the document relevance before the decision of accessing the document is made. Summaries are not computed in order to help the user extracting the knowledge from the returned information. Finally, the summarisation is based on the reduction of a single document. There had been little work aimed at providing the user global summaries or global views of a document set. However, offering the user tools that provide an initial meaningful structure of the returned documents could be of a great help [7].

Classification techniques aim at providing global document views. [16] presents a mediated classification where a large document set is classified according to a given classification structure (derived from the classification of a specific collection). [17] introduces an interactive interface allowing the document collection consultation through the search and browsing of large category hierarchies to which the document are associated. Classification can provide global document visualisation; generally in IR the classifications are limited to document classification based on "bag of words" document characterisation.

In this paper, we present an approach to mine a retrieved document set in order to help the user analysing the documents. The general framework is based on the knowledge discovery from database (KDD) process. In section 2, we parallel our framework to the KDD framework. The whole process includes the creation of an

information warehouse. The information warehouse structure is explained section 3. Section 4 is devoted to the basic document mining functions used in order to extract some useful knowledge. Section 5 relates the experiments. In addition, it gives a short example with some results and explains how the graphical views help the user in discovering the document set. The last section concludes the paper.

2. From Knowledge Discovery in Databases to Knowledge Discovery in Texts

2.1. General framework of KDD and decision support

Database technology allows efficient storage and access to more and more huge quantities of data. This is no more sufficient and the Database as a discipline has to deal with knowledge induction and decision support starting from these data. This is the purpose of the growing research fields of KDD and data mining. Techniques are provided to extract or infer useful and previously unknown knowledge from growing volumes of data. This goal is reached by discovering global patterns and relationships among the data. As examples, this technology can be used to analyse customer behaviours or to predict market evolutions, as long as data are connected to purchasing records or credit card operation records.

A KDD process can be divided into three stages [4, 13] (see Figure 1):

Data selection and pre-treatment: This consists in collecting, homogenising, cleaning and reducing the data. This phase consists in improving the space of data to explore in order to provide the only useful pieces of information to mine according to the users' objectives. This is achieved by filtering and summarising the data from on-line databases -named the data sources- that traditionally support the OLTP (On-Line Transaction Processing). The filtering or cleaning is based on the data sources querying. Summarising is used to turn the detailed data into more global data. The resulting data is generally stored in a data warehouse. Indeed, a data warehouse is defined as a repository of data designed to support management decision making and provides integrated and historical data from which data analysis can be done [20, 38]. A data warehouse is designed in order to support the OLAP (On-Line Analytical Processing). In addition, data marts can be created which are repositories designed to serve a specific group of users who share the same kind of knowledge needs. Generally, a data mart is a relevant subset of the data warehouse.

Data analysis and mining: The main objective is to mine the cleaned information in order to discover existing but previously unknown relationships among the data [13]. A distinction can be done between data analysis and data mining. One of the most used analysis methods is the multidimensional analysis, which is typically interactive and uses rollup (increasing the level of aggregation, looking at more global information) and drill-down (decreasing the level of aggregation, looking at the detailed data) operations. The knowledge extracted from the data is either visualised by the user through an adapted interface or automatically written down via reporting tools. On the other hand, data mining refers more to the automatic discovering of rules that could then be used in a system expert for example. It is directly related to the machine learning and KDD (the rules modelling the discovered knowledge).

Multidimensional analysis is used in OLAP mechanisms and a data warehouse is typically modeled multidimentionally in order to ease this type of analysis [4]. The dimensions are defined according to the type of data manipulated by the operational sources. As examples, the dimensions can be the product, the period of the sale, the seller or the shop, etc. Generally, the dimensions are hierarchical so that rollup and drill-down operations can be performed. To multidimensionality analysis is associated a graphical interface that allows to visualised the figures of 2 or 3 dimensions simultaneously.

With regard to *data mining*, several classes of mining functions have been defined [1]:

- *Classification and clustering:* the objective is to find a partition of the data (mapping them into predefined classes or into clusters constructed according to the data similarities). Classification and clustering methods use a similarity measure to compare the different elements to classify.

- *Dependencies*: it includes the discovering of (weighted) dependencies between variables, relations between fields, temporal dependencies or sequences, regressions. One tries to find data correlation e.g. under the ('antecedent / consequence') form.

Interpretation: This step objective is to fulfil the user's needs in term of knowledge and to allow him or her to take the relevant decisions. The information produced by the analysis or mining process has to be presented to the user in the most synthetic and expressive way. This is typically the case with multidimensional analysis.

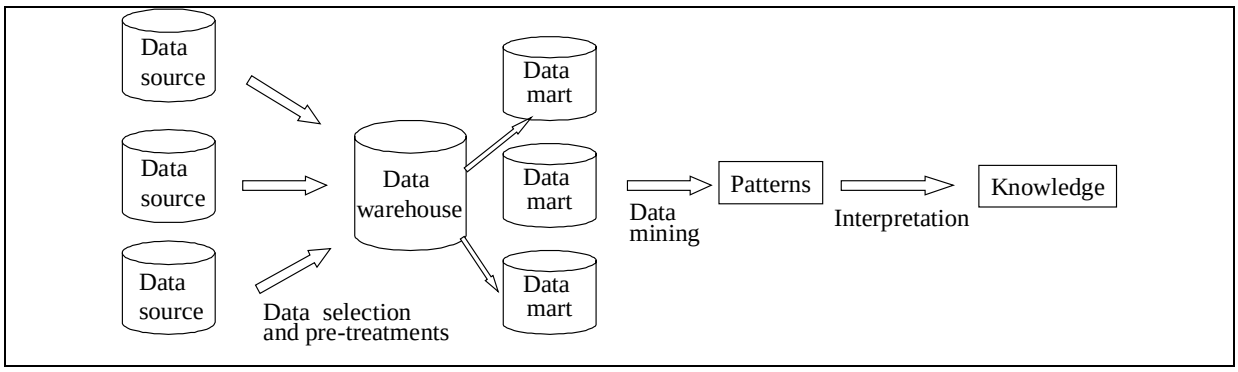


Figure 1: Overview of a global decision support process.

2.2. Knowledge discovery from text and document mining

To parallel database technology with information retrieval technology, database technology allows efficient storage and access to large data sets whereas IR technology does the same with documents. Whereas querying databases generally fulfil well the on-line applications, it is not enough when decision making or prediction is involved. Data warehouse has been suggested as a way to supply selected and summarised data in a format that fits data mining processes. In the same way, querying document sources fulfils only raw information needs; more advanced information can be provided to the users. Classification methods [21] is one of the way to provide more advanced information. The IR community has been active in that domain for a long time [19]. Classification can be a way to provide information more advanced than a simple list of documents, as most of the IRS does. On the Web for example, different researches are carried on to attempt to classify Web documents based on the words they contain [5, 11, 25]. Generally, the clustering is based on the similarity of the document content. It is computed using a similarity function on the index-based representation of the documents [34]. Automatic creation of ontologies, which define the terms and relations comprising the vocabulary of a domain, is another application of text mining [14]. But textual documents contain a rich range of information and not only keywords. On the Web, some researches have made use of the hypertext nature of the Web. Mining the hypertext references has been used in Information Retrieval [22] in order to determine the authority of the pages or for visualisation purposes; [6] analyses the relationships between the hyperlinks and the concepts used in the Web pages.

Documents have many dimensions that could be used in order to extract some knowledge from them. Meta information such as the author names, organisation, date of publication or last update is under-used in IR. We argue that these kinds of information can be of a great help when exploring a set of documents that has been retrieved according to a given information need. It is possible to extract some knowledge or relationships between the different document dimensions as relationships between the authors and the topics of the documents or the links between the authors themselves; the strength of these links and their evolution. In the same way, it is possible to extract which are the organisations involved in a given topic and how all those relationships have evolved over time [28].

One of the preconditions to be able to mine documents is to have information from these documents available in an appropriate format. In the case of databases, that is precisely the goal of data warehouses. In the same way, document warehouses can play this role when dealing with documents. Figure 2 gives an overview of the steps needed to create a document warehouse. The phases are the same than in the case of a data warehouse building; but the way to achieve them differs.

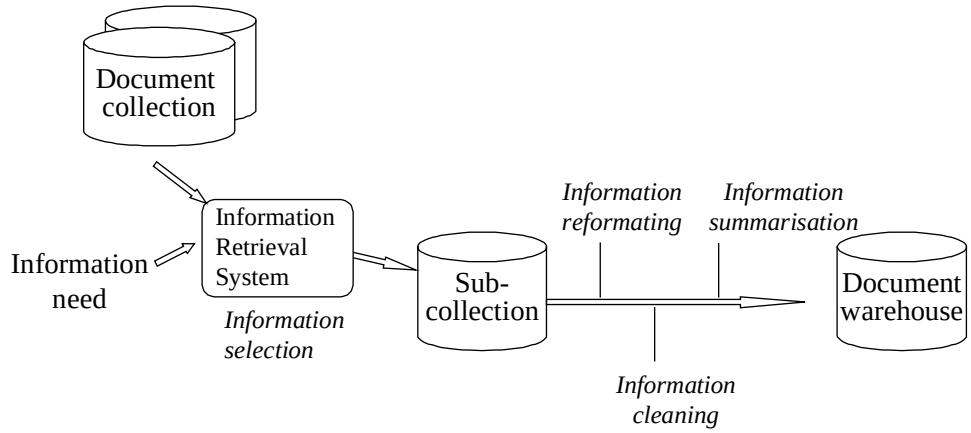


Figure 2: Overview of the document warehouse creation

The *Information selection*: has in charge to gather the information related to the domain of interest, which is described through an information need. It can correspond to the result of an IRS (or several IRS queried in parallel).

The *Information reformating*: unlike data from databases, documents are not necessarily structured. Moreover, whatever the type of document considered, a document contains many different kinds of information. One difficulty when dealing with text remains to be able to attach a "role" or a semantic to the elements that constitute a document. This role corresponds to what is called a dimension in data mining applications. It should not be confused with the linguistic role of the terms: meta-data (such as the author names or the document date) correspond to document dimensions, the document content description corresponds to another dimension. The text analysis is necessary in order to make the values of the different document dimensions being clearly identified. It can be done through the marking-up of the different elements of information. This step corresponds to the writing of the documents under the form of templates.

The *Information cleaning*: the cleaning is used to decide what is the information that has to be kept. That includes the dimensions to take into account and eventually the values to be considered (e.g. if the journal where the documents were published is a document dimension, one can be interested on the European journals only) as well as solving some syntax and semantic problems (e.g. synonyms). Notice that even if the users –of the same organisation- share the same domain of interest, they do not necessarily share the same knowledge needs and are not necessarily interested in the same parts of the information. Customisation of the document warehouse is necessary to fulfil the need of personalised views on the documents. It can be done during the mining itself.

The information *summarisation*: in the case of data warehouse, the summarisation corresponds mainly to aggregation functions done on numerical data according to the value of some of the attributes (e.g. adding the purchases of a given product over each month). In the case of texts, the main important point is that the summarisation format must fit the information mining that will be performed.

At the end, whatever the information sources (structured data or documents), the objectives are identical: to provide an homogeneous warehouse that can be mined. The information as to be synthesised so that the mining could be more efficient.

The information selection and reformatting are out of the scope of this paper. The detailed method we use to rewrite documents under the form of relevant templates can be found in [10]. The next section focuses on information summarisation.

3. Multidimensional contingency tables: a warehouse for document miming

A document warehouse answers the need of summarised information. The format of the summarised information has to be adapted to apply data mining functions, which basically uses numerical data as input. Contingency tables are a way to transform non-numerical information into numerical information that is widely used in statistical applications. They have been shown to be efficient to represent summarised information in the case of databases [13] and they are the starting point of many mining functions. Classification, clustering, factorial analysis (principal component analysis, correspondence analysis) are easily performed on contingency tables. Multidimensional analysis can be performed on these structures. Moreover, a contingency table is a basic representation from which many other representations can be derived.

A contingency table is obtained by dividing up a population according to two variables, I and J (using the OLAP vocabulary, the variables correspond to the dimensions). The columns of the table correspond to the modalities (or values) of the variable J, whereas the lines of the table correspond to the modalities of I. The table could be viewed as the characterisation of the lines (objects) according to the columns (the characteristics). In fact, the two variables play symmetric roles and can be treated the same way. In statistical applications, the intersection Tij of a row i and a column j corresponds to the number of objects in the population for which the variable I has the value i and the variable J has the value j simultaneously. Usually, the two variables are of the same nature (e.g. if the population comprises individuals, the two variables can be the colour of the eyes and the colour of the hair). This principle can be extended to different document representations taking into account the document dimensions.

A quite basic representation can be obtained using the document references and the indexing terms as the two variables. The contingency table expresses then the term frequency for each document. A vector-based representation [36] or a language model based [33] can be easily deduced from this representation; traditional term weights can be computed as all the commonly measures can be deduced from the contingency table (document size, collection size, term frequency, inverse document frequency). Doing so, the documents are represented as the contingency table lines and terms are represented as columns. In other words the objects are the documents whereas the terms are the document characteristics.

More sophisticated and synthetic representations of a document set can be obtained taking into account the other document dimensions. For example, the objects can be the document author names and the characteristics can be the document indexing terms. That means that the documents are viewed according to their

authors. In addition, what make authors being similar or different is the terms they use in the documents they write.

If the lines of the contingency table are the terms used in the documents and the columns are the dates, we obtain a characterisation of the terms according to time from which can be deduce some term evolution. In fact the crossing can involve any kind of information (dimension), as soon as this dimension exist in the documents. Such a crossing corresponds to a 2-D synthetic view of the document set.

The two-dimensional contingency tables can easily be extended to three-dimensional contingency tables (see Figure 3). In that case Tijk corresponds to the number of documents (or document units, depending on the granularity of the analysis) for which I=i and J=j and K=k. Again I, J, K can be any document dimension. When K corresponds to time (e.g. date of publication), the resulting contingency table enable the analysis of the relationships that exist between two variables and to analyse these relationships according to time.

The contingency tables can be generalised to n-dimensional tables.

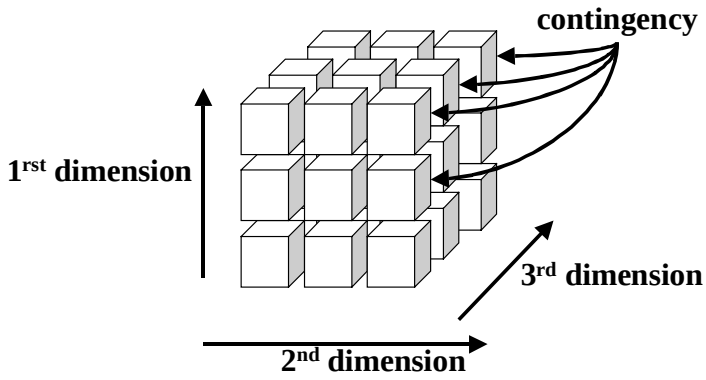


Figure 3: 3-D contingency tables

According to our framework, the summarised information has to be user-oriented. Indeed, some users can be interested in analysing only the documents that have been published in some journals or to consider only the documents that contain at least one of the concepts of a given list. This is achieved by building contingency tables using some parameters; this is done through the COOCCUR module.

This module is so called because is based on the calculation of the co-occurrence of chosen elements. The module COOCCUR computes contingency tables from "templated" documents. But, because of the textual nature of the initial information, COOCCUR does not simply count the occurrences of the values of the chosen document dimension as it would have done on data (from databases) but it combines that with the use of synonymy dictionaries. In order to personalise the repositories, lists of modalities that have to be considered for each chosen variables (i.e. the list of values for each chosen dimension) can be specified. They correspond to profiles that are used to filter the information. In the same way, list of modalities that should not be considered can also be specified. The meta-data associated with the document warehouse corresponds to the description of the dimensions that are used to build each contingency table (with their corresponding parameters). Examples are shown Figure 4.

	List +	List -	Syn		TITLE	AUTHOR	JRN	TEXT	DATE
TITLE									
AUTHOR									
JRN									
TEXT		Stoplist	TEXT.Syn			DATE			
DATE	DT								

Figure 4: Example of meta data associated with a contingency table

In Figure 4, the document dimensions are shown as well as the description on how the contingency tables are built. In that example, there are two contingency tables. One 2-dimensional table crosses the *TEXT* (document content) and the *JRN* (journal of publication). One 3-dimensional table crosses the *AUTHOR* names with the *TEXT* over time. With regard to the *TEXT*, a stop word list is used (the words contained in that list will be ignored during the documents parsing) and a dictionary of synonyms will be used (*TEXT.Syn*). The dates considered have to be in the *DT* list. DT and Stoplist correspond to positive and negative profiles.

4. Mining methods

Contingency tables correspond to a first level of knowledge modelling as they memorise the element relationships. In addition, they fit more advanced mining techniques. In the data mining field, different mining functions have been defined. Among them, the classification and the discovering of dependencies are the most important.

4.1. Classification and clustering

The objective of the classification and clustering is to find a partition of the data (mapping the data into predefined classes or into clusters constructed according to the data similarities). Classification and clustering methods use a similarity measure to compare the different objects to classify according to their characteristics. Thanks to the multi-dimensional document warehouse we defined, it is possible to easily expend the traditional document classification. Usually, documents are classified according to their content (i.e. according to the terms they are indexed with). We generalise the classification principal to the classification of any object defined from the document set. For example, the authors of the documents from the set can be classified according to the terms they use in their publications or according to the journals where they publish in. That leads to a huge flexibility of the classification in addition to a vast range of knowledge extracted from the document set.

Whatever the objects to classify, different similarity functions can be used and we implemented the most common ones. With regard to clustering, we implemented single, complete linkage clustering as well as coocurrence based clustering. A dendrogram (cf. Figure 6) results from clustering methods. It can be cut at any level. If cut at a top level, the classes are less numerous -but the objects from a class are less close each other (according to the dimensions and the similarity function chosen)- than if cut at a lower level. The content of a class can be edited. With regard to supervised classification, a *colour* is associated to each class (i.e. all the objects that belong to a group are associated with a colour).

4.2. Factorial analysis

Factorial analysis methods have been little used in IR. They are based on the mathematical properties of the Singular Value Decomposition (SVD) of matrices. One of the main feature is that is possible, given a set of points described in a N-dimension space (the matrix describes the set of points) to find the K-dimension space (K<N) that keep best the distance between the points. Latent Semantic Indexing [8] is based on SVD but takes advantage only of the dimension reduction that is associated with the SVD [12]. We better use the Correspondence Factorial Analysis (CFA) [2]. This method is based on SVD, but instead of using the absolute object representation (the contingencies) it is based on the object profiles (similar to probabilities or % of contribution). Another specificity of CFA is the distance measure used to compare profiles. The measure (χ_2) aims at favouring the specificities instead of the too recurrent phenomena. In addition, χ_2 makes it possible to represent the objects and the characteristics in the same reduced space.

It is possible to represent the documents and the indexing terms according to a CFA [24] in order to discover the correlations that may exist between the documents and explain them based on the term distribution. Furthermore, it is possible to analyse any kind of contingency tables (i.e. whatever the document dimensions used are) through a CFA.

The classification and factorial methods are not new and have largely been used in IRS and in many other fields that implies the object analysis. The main original point is that these methods are components of an interactive interface and that they can be combined in order to explore a document set.

5. interactive document mining

This section shows the principle of a document set mining based on the interactive interface we introduced. The user interactively analyses the several document dimensions in order to extract advanced information from the texts.

5.1. The document set

The document set used to illustrate the mining methodology is a subset of the Astrophysics Data System (ADS) database of astronomy abstracts, restricted to 6190 articles for which at least on of the authors is affiliated to a French institution. The document dimensions used are the terms corresponding to the document content, the author names and the affiliations. After pre-treatment, the document set is described by 5229 terms, 6455 author names and 71 affiliations.

When provided with such a document set, it is quite obvious that the user cannot read all the documents even if s/he needs to have an overview and some key points hidden in the mass of information. The interactive mining interface we present aims at helping the user to explore efficiently a document set.

5.2. Thematic maps

Thematic maps can be induced from the document set. Each one giving different kinds of advanced information. The details concerning the thematic map of the affiliations can be derived from the analysis of the correlation of the affiliation names and the document concepts dimensions. Different kinds of information can be extracted from such relationships; one of the most interesting is the thematic specificities of the institutes. The interface allows a direct visualisation of these specificities through the CFA and the graphical representation using the first factorial axes (cf. Figure 5).

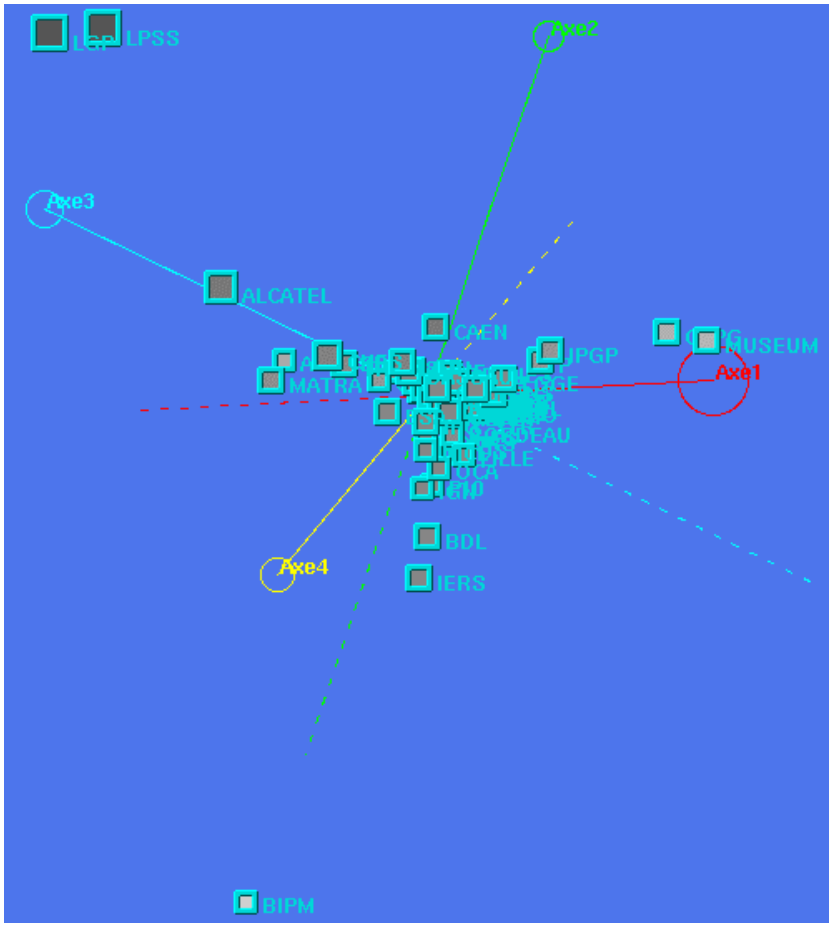


Figure 5: Institute correlations according to the thematic

These axes are the ones that best keep the distance between the objects. 4 axes are displayed simultaneously: the two first axes correspond to the plan, the 3-D effect is given by the size of the squares (each one representing an institute in Figure 5) whereas the 4-D is represented by the inner colour of the squares, from black to white. The centre (origin of the axes) corresponds to the medium profile. Thus, the institutes around the centre, are non-specific, they refer to common topics (according to the document set studied). On the contrary, the institutes far away from the centre have some specificities related to their characterisation (i.e. the concepts used in the publications); the favoured direction give some additional information (cf. Figure 6). As examples, the LGP and LPSS labs share the same specificites as they are not centred but are close each other with regard to the 4 first dimensions. Figure 6 corresponds to the visualisation the user obtains when s/he asks to visualise simultaneously the institutes and the concepts. If selected (via the mouse), the detail of the elements are displayed. The distance between the institute names (in yellow) and the topics (in blue) are representatives of their probable correlations that exist.

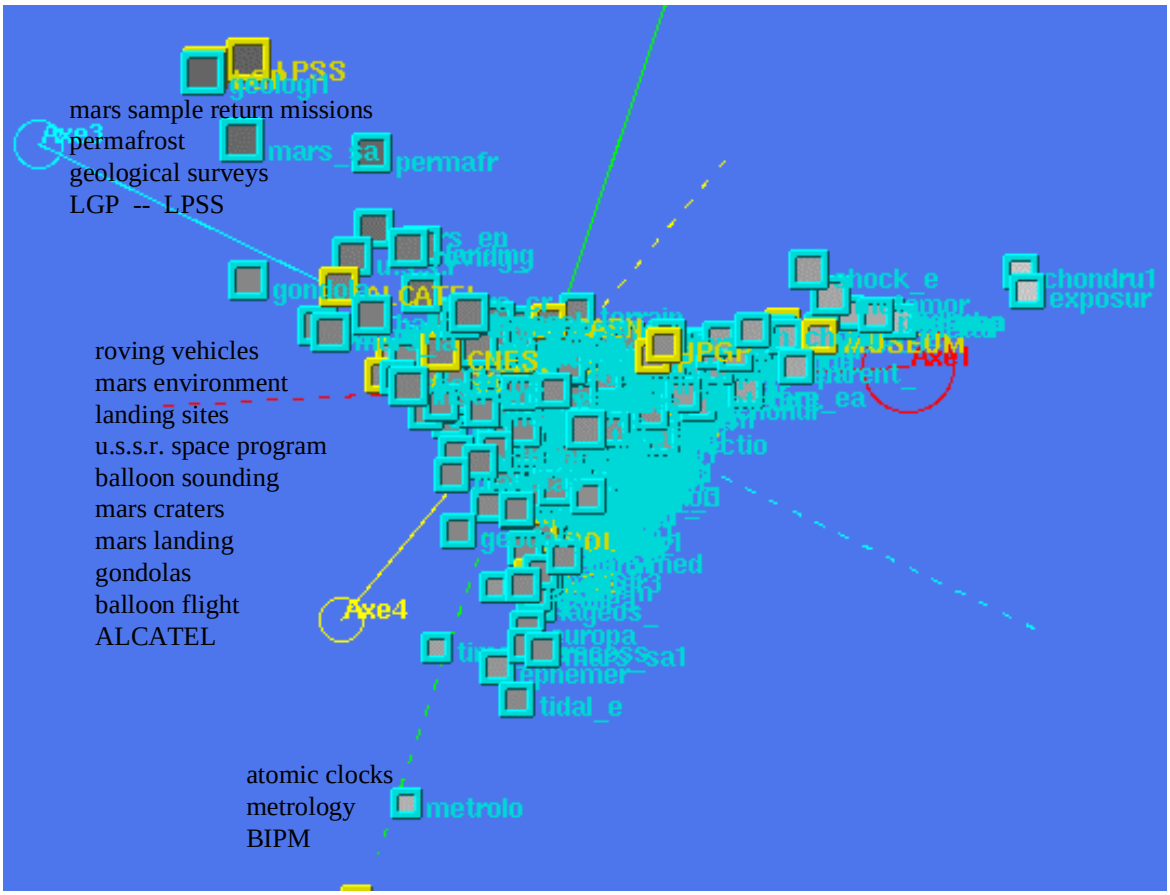


Figure 6: Thematic map of the institutes

From this visualisation, it is then possible for the user to focus on one (or several) institute(s) by an interactive filtering of the data based on the selection. For example, it is possible to automatically deduce what are all the authors belonging to a selected institute (e.g. the French "Centre de recherche en physique de l'environement terrestre et planétaire - CRPETP") and then to obtain a map of these authors. Because they are several authors and affiliations in a single article and because we did not link each author to her/his real affiliation, we get the persons who are co-authors of a person belonging to the selected institute.

Figure 7 shows the results of a dendrogram from the clustering of the authors associated to the CRPETP, according to the topics. These clusters are exported to a CFA based on the same data. The distance between the authors reflects their distance regarding their characterisation via the terms. In addition, the terms that are specific to the authors could have been shown as done in Figure 5 and the same kind of analysis could have been made.

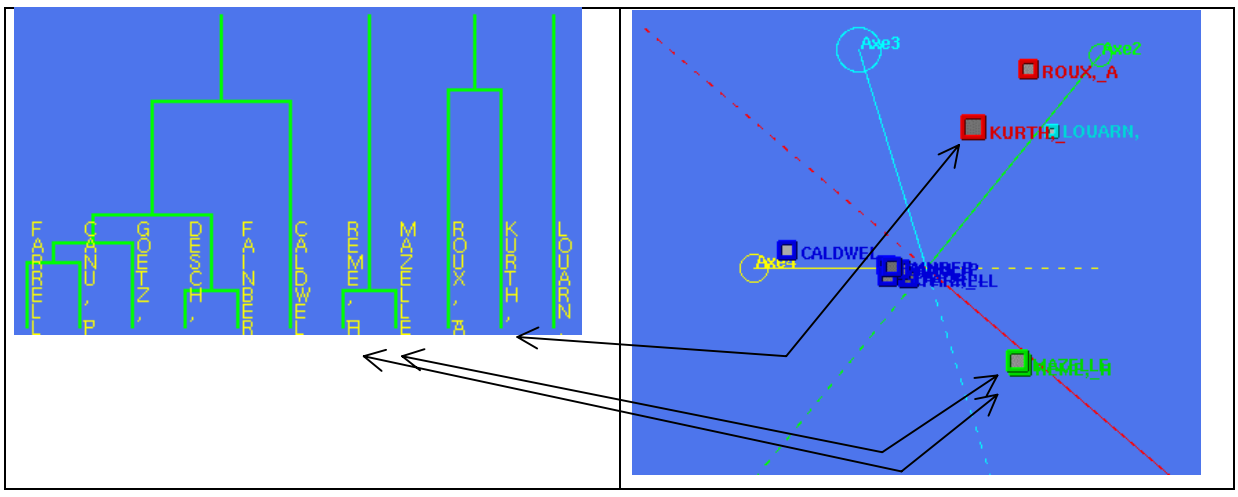
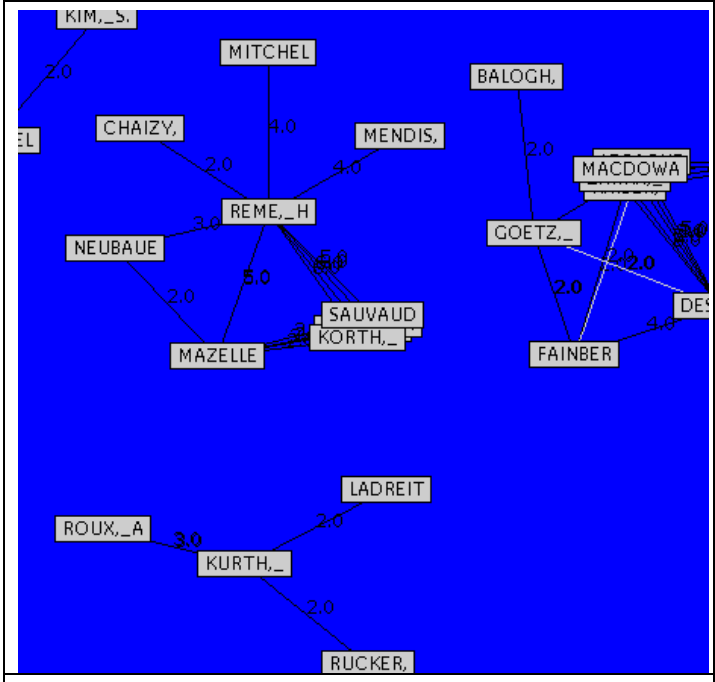
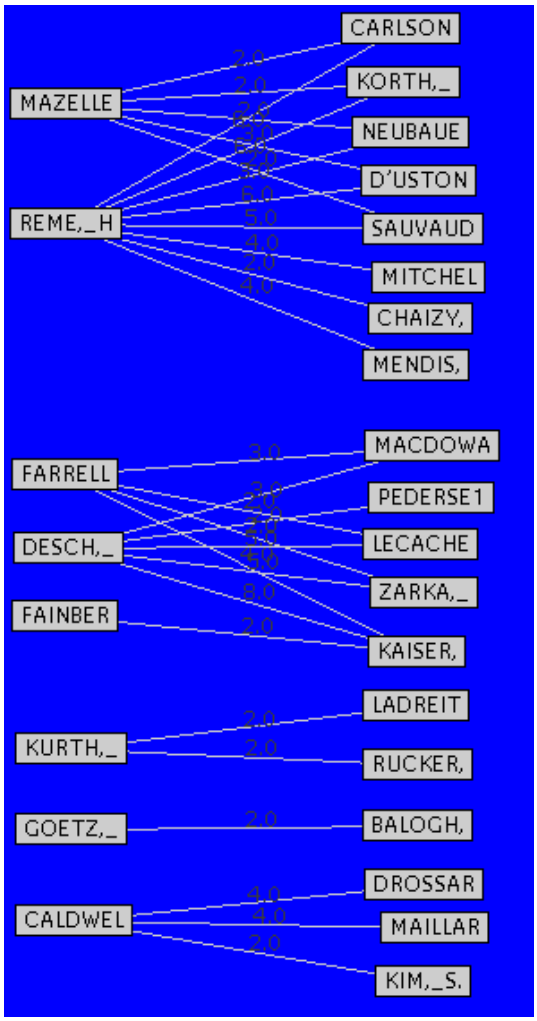


Figure 7: Thematic maps of the authors associated to a selected affiliation

These authors can then be analysed deeper in order to see for example with whom they collaborate (Figure 8).



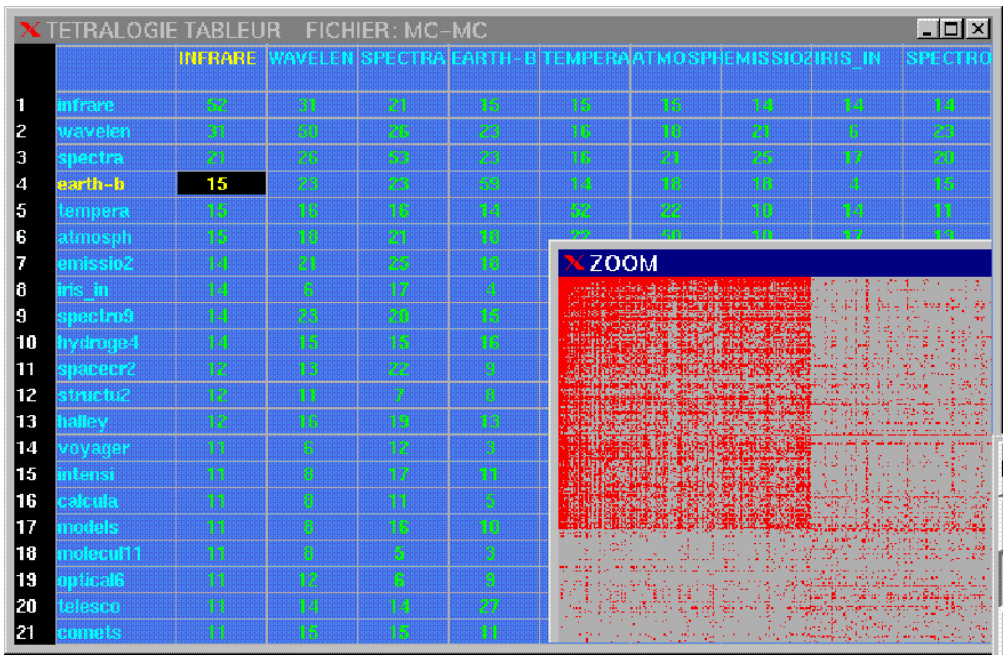
The links between the previously selected authors themselves and with the other authors are displayed. The weights of the links correspond to the co-authoring value.



The previously selected authors are displayed on the left part of the figure and their links with the other authors are weighted according to the co-authoring value. The intra CRPETP author links are no more displayed.

Figure 8: Relationships between chosen authors and all the authors

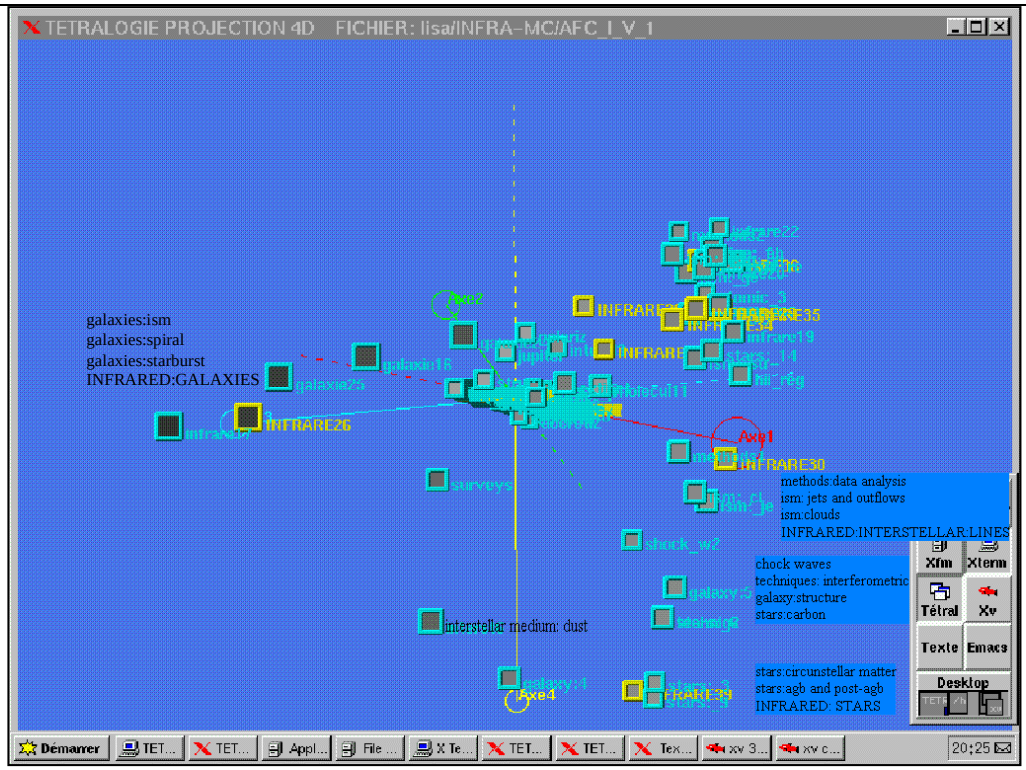
In parallel, it is possible to focus on a domain. To proceed, it is possible to automatically detect the terms that are strongly related to a given topic. As an example, in the next figures, the focus area is the Infrared field. First, based on the co-occurrences of the terms, the terms strongly related to terms that contain the word Infrared are selected. They constitute a new profile that is be used to filter all the information (See Figure 9).



The contingencies correspond to term coocurrences. They are computed on document as units. They could have been based on smaller units as paragraphs. A red point in the "Zoom" frame corresponds to a non null co-occurrence value.

Figure 9: Interactive selection of terms linked to the Infrared area.

From all these terms, it is possible to detect some sub-domains. It is done thanks to the analysis of the correlation of the co-occurrences of the terms using a CFA. Figure 10 gives the results obtained.



Several groups of terms appear. They underline sub-topics of the Infrared domain. As examples : **Galaxies:ism ; Galaxies:spiral ; Galaxies : starburst and INFRARED : GALAXIES** correspond to a sub-field. Another domain is **stars:circumstellar matter ; infrared : stras ; stars : agb and post-agb.** These two sub-domains are related thanks to the key-words (**Galaxy stellar content and Interstellar medium :dust**). These words are shared by two subdomains. Other sub-domains can be view on this graphical representation with the links they have each other.

Figure 10: Term correlation

6. conclusion

In this paper we presented a new way of browsing a set of documents. It is based on knowledge discovery techniques. The document set is first summarised or synthesised under the form of multi-dimensional contingency tables. The document dimensions used for the summarisation correspond to either content description or meta-information. The summarised information corresponds to a document warehouse one can mine in order to extract advanced information such as correlation and relationships between the document dimensions. The mining is an interactive process where the user plays a major role. The results are displayed through dynamic graphical views. Thank to the system, the user can discover useful relationships and associations between the information, which avoid the reading of the raw information. In addition, the user can discover what are the main features of the documents and their links. A next step will be to formally evaluate the interface and the all process as presented. One of the difficulties is that, because of the interactive functionalities, the traditional measures used in IR are insufficient and have to be complemented by other measures [23, 35, 37]. [3] presents an interesting framework to evaluate interactive information retrieval systems ; however this framework is not fully transposable to be applied to the evaluation of knowledge discovery systems and further research has to be carried on that direction.

7. References

1. R. Agrawal, T. Imielinski, A. Swami, *Database Mining: A Performance Perspective*, IEEE transactions on knowledge and data engineering, pp 914-925, Vol.5, N.6, 1993.
2. J.P. Benzécri, *L'analyse de données*, Tome 1 et 2, Dunod Edition, 1973.
3. P. Borlund, Experimental components for the evaluation of interactive information retrieval systems, Journal of Documentation, V. 56, N.1, pp 71-91, 2000.
4. S. Chaudhuri, U. Dayal, An overview of data warehousing and OLAP technology, ACM SIGMOD Record, Vol. 26, N.1, pp 65-74, 1997.
5. C. Chekuri, M.H. Goldwasser, P. Raghavan, E. Upfal, Web search using automatic classification, 6th International Conference on the World Wide Web, 1997.
6. F. Crimmins, T. Dkaki, J. Mothe, A. Smeaton, "Tétrafusion: Information Discovery on the Internet," *Intelligent Information Retrieval*, pp 55-63, July-August 1999.
7. M. Czerwinski, K Larson, Trends in future Web designs: What's next for the HCI professional?, In Interactions – New visions of human-computer interaction, pp 9-14, Vol 6, 1998.
8. S. Deerwester, S. Dumais, G.W. Furnas, T.K. Landawer, R. Harshman, Indexing by latent semantic indexing analysis, journal of the american society for information science, Vol. 41, N.6, pp 391-407, 1990.
9. T. Dkaki, B. Dousset, J. Mothe, Mining information in order to extract hidden and strategical information, 5th international conference on computer assisted information retrieval, RIAO 97, pp 32-51, 1997.
10. T. Dkaki, B. Dousset, J. Mothe, Information mining in order to graphically summarise semi-structured documents, CODATA, 2000.
11. S. Dumais, H. Chen, Hierarchical classification of Web documents, 23nd International Conference on Research and Development in Information Retrieval, Athenes, 2000.
12. C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika, Vol.1, pp 211-218, 1936.
13. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, ISBN 0-262-56097-6, 1996.
14. R. Fikes, A. Farquhar, distributed Repositories of highly Expressive Reusable Ontologies,, IEEE Intelligent systems, Vol. 14, N. 2, pp 73-79, 1999.
15. R. Gaizauskas, A. Robertson, Coupling information retrieval and information extraction: a new text technology for gathering information from the Web, pp 356-370, Proceedings of Computer-Assisted Information Searching on Internet Conference, RIAO, Montréal, Canada, 1997.
16. D.J. Harper, M. Mechkour, G. Muresan, Document clustering for mediated information access, 21st symposium on Information Retrieval, pp 92-107, 1999.
17. M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hiearchy.
18. M.A. Hearst, Untangling Text Data Mining, 37th Annual Meeting of the Association for Computational Linguistics, June 1999. http://www.sims.berkeley.edu/~hearst
19. M.A. Hearst, The use of categories and clusters for organizing retrieval results, in Natural Language Information Retrieval, T. Stralkowski (Ed.), Kluwer Academic Press, 2000. http://www.sims.berkeley.edu/~hearst
20. W.H. Inmon, What is a data warehouse?, PRISM, Vol.1, N.1, 1995.

21. A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Surveys, Vol. 31, N.3, pp 264-323, 1999.
22. J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, Vol.46, N.2, pp 212-235, 1999.
23. J. Koenemann, NJ. Belkin, A case for interaction: A study of interactive information retrieval behavior and effectiveness, CHI'96, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 205-212, New York, 1996.
24. L. Lebart, A. Salem, L. Berry, Exploring textual data, Kluwer Academic Publishers, ISBN 0-7923-4840, 1998.
25. Y. Maarek, I.Z. Ben Shaul, Automatic organizing bookmarks per contents, 5th International Conference on the World Wide Web, 1996.
26. P. Martin, P. Eklund, Embedding Knowledge in Web Documents, 8th International World Wide Web Conference, Toronto, 1999.
27. K. McKeown, D. R. Radev, Generating summaries of multiple news articles, pp 74-82, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, 1995.
28. J. Mothe, "Internet-Based Information Discovery: Application to Monitoring of Science and Technology," *J. Research in Official Statistics*, No. 1, pp. 17–30, 1998.
29. J. Mothe, Correspondence analysis method applied to document re-ranking, Internal report, IRIT/00-21-R.
30. J. Mothe et al., WebSmart: data marts from Web pages, submitted to Decision Support System Journal, 2000.
31. C.D. Paice, Constructing literature abstracts by computer: techniques and prospects, N°26, pp 171-186, Information Processing and Management, 1990.
32. M.T. Pazienza (Ed.) Information extraction – A multidisciplinary approach to an emerging information technology, ISBN 3-540-63438-X, 1997.
33. J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, 21nd International Conference on Research an Development in Information Retrieval, pp 275-281, Melbourne, 1998.
34. K. van Rijsbergen, Information Retrieval, Butterworths, London, Second Edition, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html
35. S.E Robertson, N.M. Hancock-Beaulieu, On the evaluation of IR systems, Information Processing & Management, Vol. 28, N.4, pp 457-466, 1992
36. G. Salton, The SMART Retrieval System – Experiments in automatic document processing, Prentice Hall Inc., Englewood Cliffs, NL, 1971.
37. L.T. Su, Evaluation measures for interactive information retrieval, Information Processing & Management, Vol. 28, N.4, pp 503-516, 1992
38. J. Widom, Research problems in data warehousing, International Conference on Information and Knowledge Management, 1995.