Interactive multidimensional document visualization

Josiane Mothe
Inst. de Recherche en Informatique de Toulouse
www.irit.fr/~Josiane.Mothe/

Taoufiq Dkaki
Inst. de Recherche en Informatique de Toulouse
www.atlas.fr

1 Introduction

Complementary approaches are used to improve information retrieval efficiency. One approach consists in focusing on the system inner processes used (e.g. document indexing, query-document similarity computing). Another approach can be to improve the user-system interaction [7]. Positive and negative relevance feedback type mechanism is among the most used and lots of studies have shown its efficiency. Classification is also used: grouping together documents according to their similarity and using those classes when consulting or querying the collection. When proposing an interactive process, powerful interfaces are needed. INQUERY [1] proposes a graphical view of the retrieved documents where similar documents are graphically close; the retrieved documents are listed to the user according to the class they belong to. VIBE [6] represents the information in a two dimensional space (document vs term) so that it is possible to graphically see which are the possible relevant documents according to some selected terms. Some complementary studies have shown that it is efficient to use the document structure or the factual information extracted from a document. ENVISION [5] represents the retrieved documents into two-dimensional tables. Rows and columns represent either factual information such as authors, date, or estimated relevance and cells are filled with document references, navigation through those tables is possible.

In this poster we propose an interactive document visualization tool. It takes into account the fact that document relevance depends on the document features which are considered (information content, publication date, author affiliation, ...). Those elements are crossed in order to discover the strong links that exist between them. Discovered relationships are then displayed on a 4-dimensional graphical view. This interface allows the user to graphically select the elements he is most interested in and to automatically construct new filters that will change interactively the set of the displayed documents.

2 Document representation

- Information extraction

We propose to extract different elements that characterize the document and that may help the user in the information seeking task. We use: *Phrase extraction* based on a statistical approach and *Factual extraction* based on document tags in order to extract (attribute: value) features such as author names, publication date, author affiliations, references...[3].

- Information filtering

Once those element values have been extracted from the raw information, some of them can be easily used to filter the information. As an example, a phrase subset can be used to filter the documents in which those phrases occur or on the opposite the documents in which those phrases do not occur (positive filtering and negative filtering). In fact, any of the extracted elements (references, authors, dates,...) can be used for filtering purpose.

- Information summarization

We summarize extracted elements under the form of contingency tables which is a powerful basic 2D-knowledge representation (Zembowicz et Al., pp 329-349 in [4]). The crossed features can be any of the extracted elements. In addition, the element values can be either all the values that have been automatically extracted from the whole set of studied documents or some values that are stored as a filter (filters are interactively build -see 3).

- Information analysis

To analyse the information we use the Correspondence Factorial Analysis (CFA) method [2]. It is used to represent variables and characters in a same space. Within the context of the data analysis, variables correspond to the contingency table columns and characters corresponds to the rows; here, variables and characters correspond to extracted elements of information. The space axes are computed as a variable combination so that the information explained in that space is the most important in term of inertia. Then, a graphical representation in a space using the first principal axes gives an interesting view of the set of variable and character values. The distance between a variable and a character (or between two characters or two variables) corresponds to a dependence relationship.

3 Interactive document visualization

- Multidimensional document representation

We provide graphical representations in a 4 dimensional space. The displayed axes are the 4th first principal axes resulting from a CFA. The user can query and visualize any extracted element summarization or crossing. Some are more relevant. Among them, the result of a (authors vs phrases) crossing can be quoted: it provides the "experts" of a domain (Fig.2). In the same way,

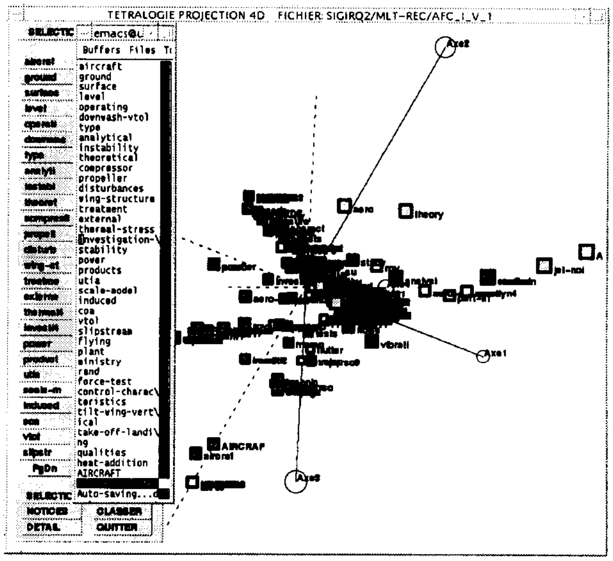

Figure 1: Authors+Document references vs Phrases correlations
Phrases associated with the term AIRCRAFT are displayed.The term typology according to the document set is given.

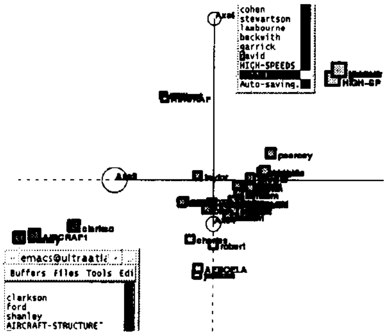

Figure 2: Authors/Phrases correlations
Two groups have been displayed, their content can then be used as filters.

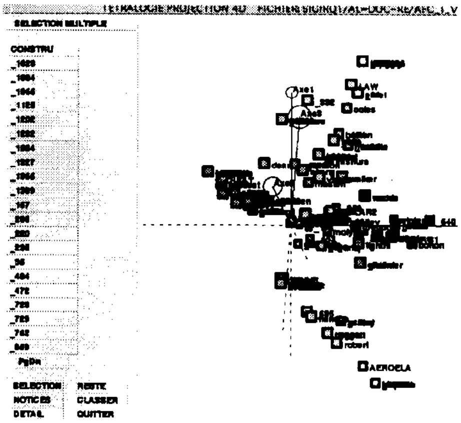

Figure 3: Authors+Document references vs Phrases correlations

the visualization of a (document references + authors vs phrases) crossing will provide the user with relevant information simultaneously about both documents and authors according to their similarities with the selected phrases (Fig.3). On the other hand, the visualization of the phrases extracted from the whole set of filtered documents in the query term space is relevant to characterize more accurately what phrases subsets are related to each query terms according to the document collection content (Fig. 1). It can help the user to assess the relevance of his query formulation and helps him reformulating it.

- Interactive filter creation

Another key point of our approach is the interactive filter creation, providing the user a powerful tool to analyze and visualize the document set. At any time he can graphically select some element values and use them to re-filter the document set. It is used to analyze deeper some document subsets (Fig. 1 and 2).

4 Experimentation

We applied this approach to the Cranfield collection using our system TETRALOGIE. We are now experimenting this approach to a bigger data collection from a 200 000 documents sample from the Adhoc TREC6.

References

[1] J. Allan et Al. *INQUERY does battle with TREC-6.* 1997.

[2] J.P. Benzecri. *L'analyse de donnees.* Dunod Edition, 1973.

[3] D. Dousset, T. Dkaki, and J. Mothe. Mining information in order to extract hidden and strategic information. In *Computer-Assisted Information Searching on Internet, RIAO97*, pages 32–51, 1997.

[4] U.M. Fayyad et Al. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, 1996.

[5] L.T. Nowell et Al. Visualizing search results: Some alternatives to query-document similarity. In *SIGIR*, pages 67–75. ACM Press, 1997.

[6] K.A. Olsen et Al. Visualization of a document collection: the vibe system. *Information Processing and Management*, 29(1):69–81, 1993.

[7] TREC-6. Text retrieval conference. 1997.