multilingual and multwriting access to information on the Internet", Actes de RIAO 2000, coord. CID, collège de France, Paris, 12-14 avril 2000

Textual mapping for multilingual and multiwriting access to information on the Internet

A. Lelu*, M. Hallab*, F. Papy, S. Bouyahi**, H. Rhissassi*, N. Bouhaï*, F. Tang*****
* Université Paris 8 / Département Hypermédias.
** Université Paris 8 / Département Documentation.
2 rue de la Liberté, 93200 Saint Denis
*** Enseignant de chinois, Paris.
lelu@cnam.fr

Abstract

We present here a prototype Web search engine with 1) a fuzzy lexical access to the list of simple and compound indexing terms, 2) a simplified, symmetric and powerful statistical expansion dialogue, 3) a cartographic applet for analyzing and browsing the results, 4) a multilingual and multiwritings extension, based on a N-grams coding scheme. The neural mapping algorithm, as well as the statistical expansion process, lean on a mathematically clean and consistant background (Hellinger metrics and i.d.f. weighting).

Introduction

In this work, embedded in a prototype search engine co-funded by the French Ministry of Education, Research and Technology, we have tried to deal with three basic needs, not really fulfilled at the present time on the World Wide Web :1) The need for overviews and synthetic information, 2) The need for a simplified and unified mechanism for selecting the documents superset which presumably "hides" the relevant information, and thus has to be mapped, 3) The need for taking into account the diversity of languages and writings.

1 - The need for overviews and synthetic information.

When faced to a large amount of Web pages drawn from a previous request, or iterately selected, one must have a global idea, an overall view of their content. This is the reason why we have designed and implemented :

- a "fuzzy" clustering algorithm, the Axial K-Means (Lelu, 1994), which makes such a representation emerge from the data, as a set of unsupervised topics, without any need for a human classification of the texts neither any semantic organization of the descriptive terms. This neural algorithm is based upon a "Winner Takes All" network of variants of Oja neurons (Oja, 1991), resulting in an oblique factors representation: each term or document is characterized by its projections, i.e. centrality degrees, upon the axes abstracting the main topics in the dataspace. Each of these axes may be considered as the first factor of a minimal Spherical Factor Analysis (Domengès & Volle, 1979) applied to a thematically homogeneous subset of documents. Due to the Hellinger metrics used in this method, and other specific features, the semantic quality of our analysis is generally compared favorably to other algorithms, for example:

  Leximappe http://www.cisi.fr/Vfr/Produits/sampler.htm, U-Map www.umap.com, SemioMap www.semio.com.

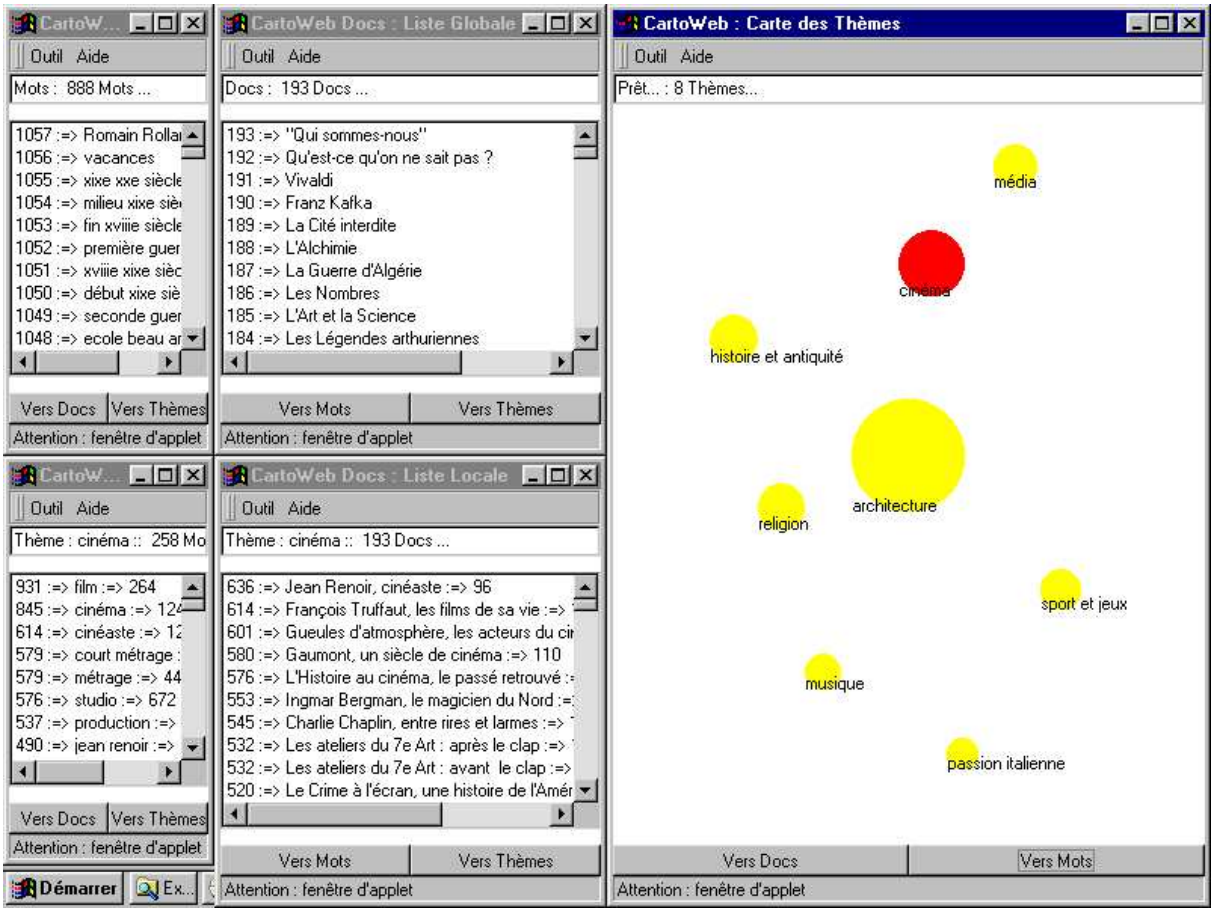
- an original navigation interface : our topics are automatically displayed over a 2D global map, which is 1) much more legible than when all documents and/or all terms have to be displayed, 2)

semantically consistant by itself, in that semantically related topics are close to each other in the map space.

These principles are demonstrated by a Java applet (see Figure 1) we designed for one to explore and make his interest bounce when set to three types of entities :

. the documents,
. their descriptors,
. the topics issued from the "m-n" relationship between documents and descriptors: each topic is embedding a data-pole, a high density area in the data space, which gathers both homogeneous documents and the homogeneous descriptors they share.

Browsing is quite simple, due to a symmetric organization of the interface (Lelu, Tisseau-Pirot & Adnani 1997), with a set of 2-buttons windows:



.

Figure 1: Screenshot of our cartographic browser, applied to the collection of abstracts of the books "Gallimard Jeunesse"

. in the documents list window: "To the terms" and "To the topics" buttons,
. in the terms list window: "To the documents" and "To the topics" buttons,
. in the topics map window: "To the documents" and "To the terms" buttons,

A given document or term may “enlighten” one or several topics in the map, respecting the context effects: the well-known polysemy of many terms, and the multiplicity of topics embedded in a single document.

2 - The need for a simplified and unified mechanism for selecting the documents superset which presumably "hides" the relevant information, and thus has to be mapped.

We have set up two principles, issued from our pratice and design experience of textual Human Machine Interfaces :

- Principle 1: we think it better and fair with regard to the user not to answer him directly a list of documents when he types his query. Any retrieval system indeed stores all what it considers as document descriptors, and the typed character string may - or may not - be present in the list. We try to help the user making his mind by himself : we have decided 1) to dig out the compound phrases from the texts, which are known to be the real semantic elementary units, 2) to display the phrases which are lexically neighboring to the query string, and "fuzzily" match this string - because of typing errors or mismatches of both the user or the authors -, whatever the position of this approximate string in the phrase.

We have designed and implemented in our prototype search engine such an approximate string matching algorithm, which can be favorably compared to those based on dynamic programming

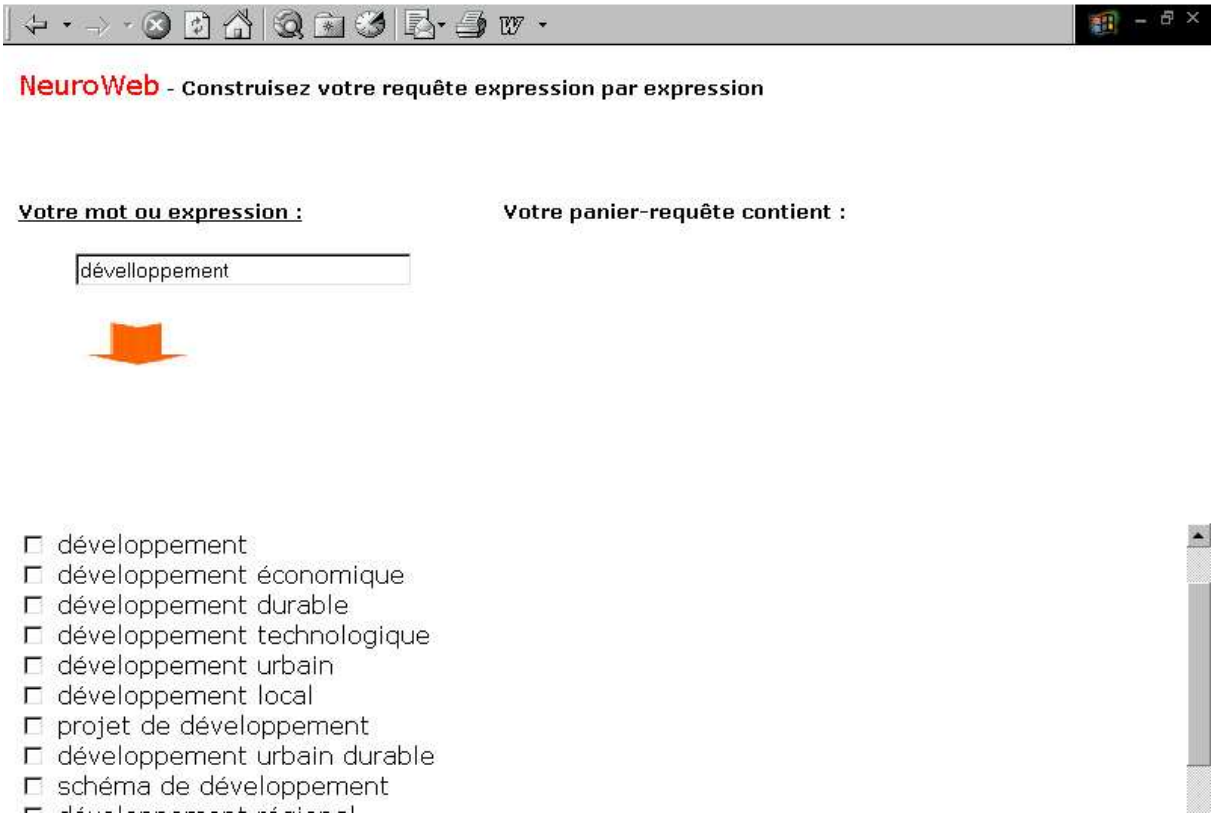


Figure 2: The user types a query (with a misspelling), and the system answers the list of the lexically closest single and compound terms.

(Lelu & Hallab, 2000).

- Principle 2: reducing the diversity of query types to one common kernel.

The Web search engines implement a wealth of query mechanisms, which may be reduced to four general schemes:

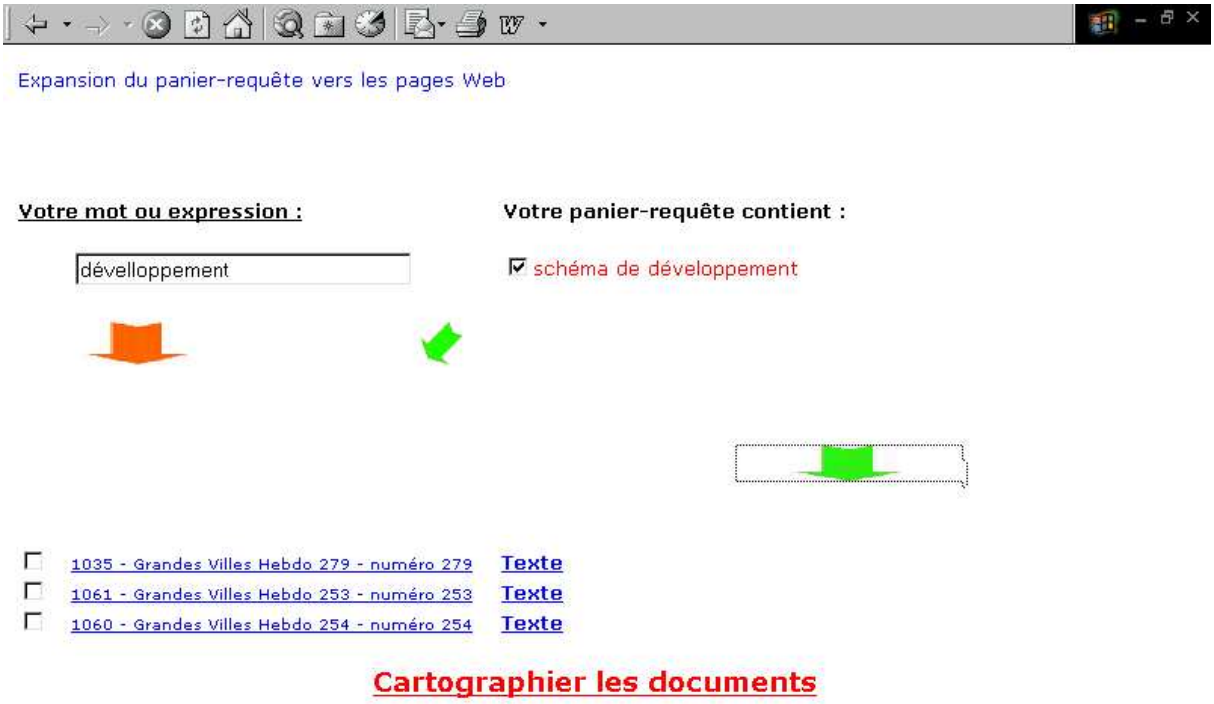


Figure 3: The user chooses the term "schéma de développement" (i.e. puts it in his query-bag) and clicks the "To the documents" arrow.

. 1(or n) term(s) -> documents : boolean query, often implicit and weighted, and giving rise to a ranked document list.

. 1(or n) document(s) -> terms : provides the list of the most relevant terms in a document or a set of documents.

. 1 document -> documents : global similarity query, i.e. "statistical expansion" of a document, or a part of a document.

. 1 term -> terms : request for semantically neighboring terms, based either on manually edited thesaurus relations, or "statistical zooming".

We have designed and implemented a unique query scheme, which gathers and generalizes the above ones:

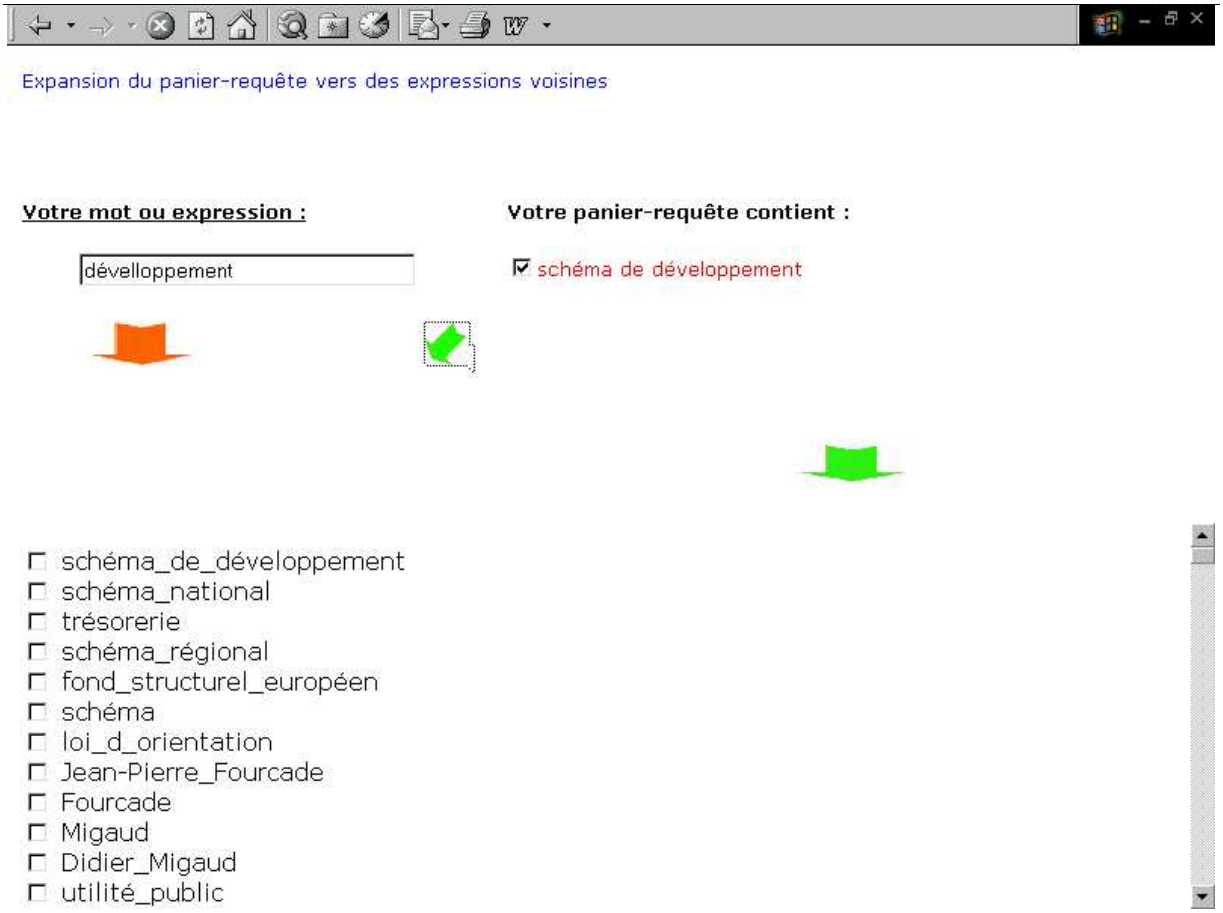


Figure 4: The user clicks the "To the terms" arrow, and the system answers the list of the semantically closest terms (linked by statistical co-occurrence)

. the user is provided with a "query basket" he may fill indifferently with term(s) and/or document(s) best characterizing his interest field.

. this unique query area gives rise to two different types of answers, depending on his clicking the "To the terms" button or the "To the documents" one, which enable him to iteratively modify his query basket, and progressively focus on his subject of interest. A unique and symmetric mathematical scheme, using the Hellinger metrics and i.d.f. weighting, underlies these different operations – considered as to be expanded to the terms or to the documents, a query-basket is considered as a “bag of terms” or as a “bag of documents”, as long as a term may be treated as a document indexed with one word, or a document as a word indexing one document.

3 - The need for taking into account the diversity of languages and writings.

Our propositions, in the above two sections, are laying on the concept of vector representation of documents, the components of which correspond to the chosen descriptors. The problem is to find out, whatever the language and the writing, the right descriptors which have to be 1) the most revevant as possible for a given language and a given writing system, 2) identifiable and extractable in the present state of the art.

- For French and English languages, we are relying on the Nomino morpho-syntactic analyzer (P. Plante, UQAM,<www.ling.uqam.ca/nomino>), in order to draw out lemmas and candidate compound phrases, next to be filtered and controlled in our Hypermap assisted indexing environment (Rhissassi

& Lelu, 1998).

-For languages in which the process of word separation is not trivial and depends on the semantic context, such as agglutinative languages or asian writings, we have developed a technique for coding the documents by their N-gram frequencies, which compares favorably with similar approaches (Lelu, Hallab & Delprat, 1998) - the result being 1) the mapping of a collection of texts, i.e. the extraction and graphic display of thematic clusters, 2) the computation of proximity indices between documents, based on their oblique coordinates in the space spanned by our thematic axes, which may be considered as some kind of a "patent semantic indexing", and not a latent one, as semantic interpretations can be drawn from each thematic axis. The figure 6 gives an idea of the man-machine dialogue, and the figure 7 shows how “highlighting” chinese bigrams suggests relevant words and phrases, without any word isolation process.

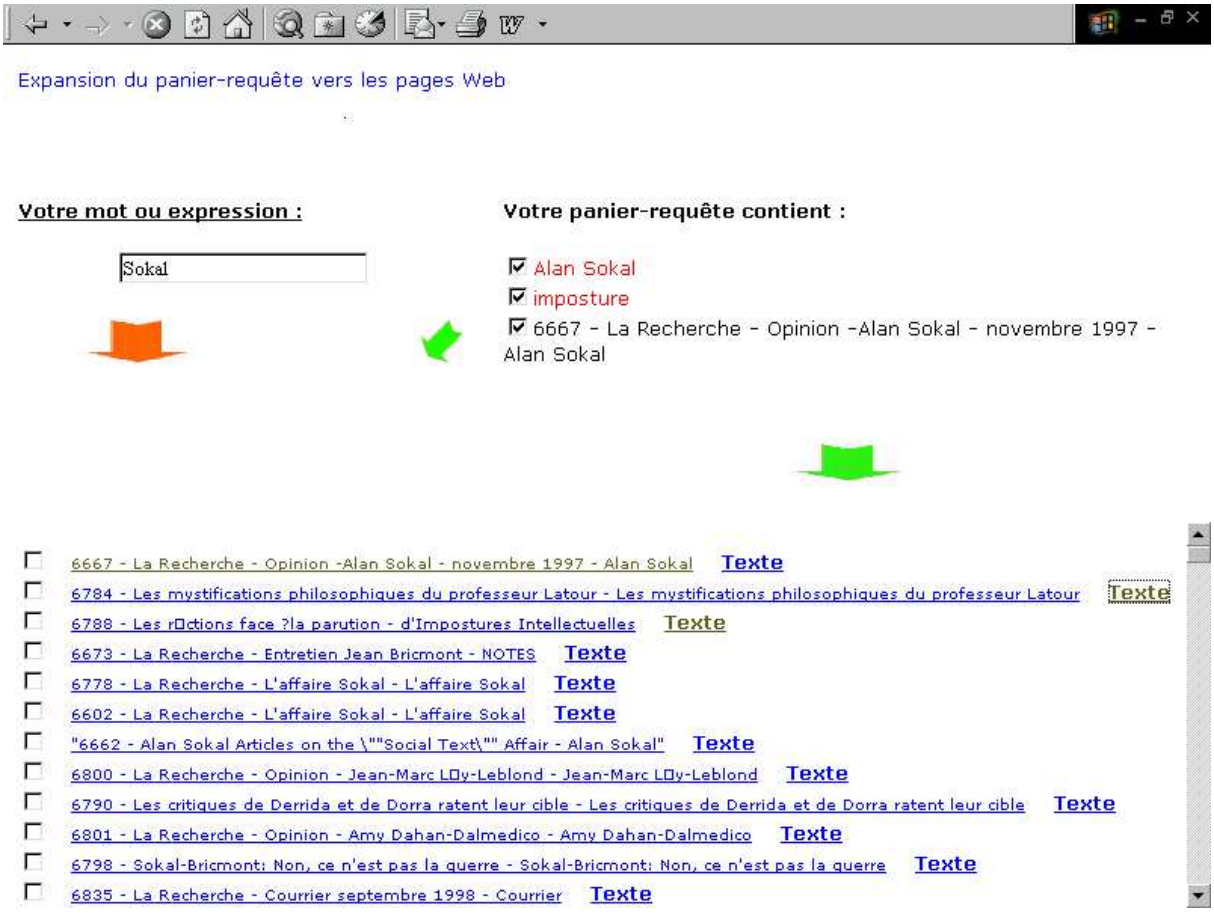


Figure 5: The user puts in his query-bag two terms and one document particularly specific of the "Sokal affair", and clicks the "To the documents" arrow.

Figure 6: The user types a detailed query sentence; the system then answers the list of the most relevant statistically extracted semantic topics; finally the user chooses one of these topics in order to get the list of the topic's documents closest to his query.

Conclusion and evaluation stakes

The core of our search engine can be accessed at http://hdyn.hymedia.univ-paris8.fr/neuroweb. Our CartoWeb applet and our chinese version ought to be fully integrated by the time of the conference. A sample of 8000 french-language Web pages issued from randomly chosen sites can be accessed, and 7000 chinese pages issued from Takungpao, a major Hong-Kong press site will be on-line in the sequel.This work is in progress: it follows that a full evaluation has not been set up yet.

However, a partial evaluation of our N-gram coding and "patent indexing" took place in the frame of the french Amaryllis evaluation program, applied to the problem of multilingual querying: it appeared that documents in english answered to french queries had high recall and precision rates when the query was correlated with one or a few extracted topics, and very low ones otherwise… Hence our results are quite sensitive to how fine-grained is the analysis, suggesting parametrization elements for our Web engine, and further research tracks.

Acknowledgements

We are indebted to the French Ministry of Education, Research and Technology, for co-funding the prototype Web search engine within the frame of which the present work takes place. Special thanks to Qi Chong and Bruno Delprat for their contributions to the processing of Chinese Language.

.

Figure 7: Example of the highlighted Web page issued from the “Sports” cluster (see figure 6)

Bibliographical References

Domengès D., Volle M. – Analyse factorielle sphérique : une exploration – Annales de l'INSEE, N°**, vol.**, 1979

A. Lelu - "Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets" - New Approaches in Classification and Data Analysis - E. Diday, Y. Lechevallier & al. eds., pp.241-248, Springer-Verlag, Berlin, 1994

A. Lelu, A.G. Tisseau-Pirot, A. Adnani - " Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation. " - Hypertextes et Hypermédias, vol.1, N°1, éditions Hermès, Paris, 1997

A. Lelu, M. Hallab, B. Delprat - "Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes" - Actes des 4emes Journées Internationales d'Analyse Statistique des données Textuelles, coord. S. Mellet, UPRESA " Bases, Corpus et Langage ", Université de Nice, 1998

Lelu A., M. Hallab - Consultation " floue " de grandes listes de formes lexicales simples et composées : un outil préparatoire pour l'analyse de grands corpus textuels. - to be published in : Actes des 5emes Journées Internationales d'Analyse Statistique des données Textuelles, EPFL, Lausanne, Mars 2000

Oja E. – Data compression, feature extraction and autoassociation in Feedback Neural Networks – Artificial Neural Networks, T. Kohonren et al. Eds., Elsevier, 1991

Rhissassi H., Lelu A., " Indexation assistée et cartographie sémantique pour la génération automatique d'hypertexte ", Actes de la conférence CIDE'98, coord. M. Mojahid, INPT, Rabat, Maroc, 15-17 Avril 1998, Europia Productions, pp.131-139.