Contrastive Indexing of Full Text Documents

Laurent Romary, Patrice Bonhomme

Contrastive Indexing of Full Text Documents

Laurent Romary
Patrice Bonhomme
UMR LORIA, B.P. 239, F-54506 Vandoeuvre Les Nancy
{romary,bonhomme}@loria.fr

In the context of the general Aquarelle scenario (see A. Michard's introduction in this issue), the creation of folders allows a user to put together pieces of information which he considers useful for his own purpose. In particular, he may include textual fields which in turn have to be made accessible for further retrieval. To this end, we designed a full text indexing method which, rather than providing an absolute set of indexes for each textual field, aims at contrasting each of them to the other fields the user might point to either in the same folder or within other folders he has created or extracted from an Aquarelle server.

The basic idea behind the contrastive indexing method is to consider a given document or rather the set of tokens it contains as a sample taken from the set of all tokens belonging to the reference corpus of documents it belongs to (see [1]). The frequency of the token within the document can then be compared to the expected distribution computed from the reference corpus, in order to evaluate whether it is inkeeping with it, or on the contrary too far from it not to be interpreted as indicating a particular relevance for the document. For each document, we thus compute a set of so-called contrasting tokens which is a good indication of its informational content *relatively* to the contents of the documents it is compared to. As a consequence, this method has different interesting properties which both from a linguistic and information retrieval point of view makes it a good option for an optimal full text indexing mechanism (detailed in [2]):

- there is no need for a specific list of grammatical words (i.e. *stop-list*) which have to be avoided during the indexing process, since these are generally subject to a uniform distribution among a set of documents taken from the same field or belonging to the same textual genre (e.g. historical descriptions, newspaper articles etc.);
- furthermore, not only grammatical words are being dropped from the candidate list of indexing terms, but also those words which, although being meaningful from an absolute point of view, are uniformly represented within the reference set of documents and thus are not relevant for the description of any specific one. For instance, within a set of documents describing historical buildings words like 'architecture' or 'architectural' are not likely to appear as contrasting tokens;
- as a consequence, the method can be seen - at least to a large extent - as being language independent, as it relies on a local model of linguistic distribution. Still, even if this has not been specifically tested, we might expect that for highly inflectional languages the result might not be as good as those observed for French and English, unless a lemmatizing phase is considered;
- finally, it can be observed that a given document can be indexed differently according to the set of documents which contextualizes it, thus providing a way to account for the different viewpoints users might project onto it when building up folders.

The full text indexing module has been considered as a semi-automatic process provided to the user during the folder editing stage. As a matter of fact, the user always has the possibility to edit and validate the set of candidate terms before these are actually inserted within the folder itself.

Given the robustness of the method as we have observed it in our first trials within the *Aquarelle* project (described in [3]), we have thought of extending it towards a general mechanism of content identification within a set of more less homogeneous documents. Indeed, what results from the contrastive indexing process is a kind of thematic description of the document in comparison with a given reference which acts as a background, hence the possibility to iteratively group together documents with similar descriptions and further to build up a thematic map of the whole reference database. This concept has been recently applied within a project funded by the DGLF (*Délégation Générale à la Langue Française*) aiming at automatically producing thematic descriptions of a given web site. The contrastive indexing method, combined with a hierarchical clustering algorithm has allowed us to produce topic maps of a given web site independently of its actual language or content domain.

References

[1] P. Bonhomme, E. Bourion, B. Gaiffe, F. Rastier, L. Romary, and N. Valceschini. Accès sémantique aux banques textuelles. rapport intermédiaire, CRIN-INALF, mars 1996. Programme Cognition, Communication intelligente et ingénierie des langues.

[2] P. Bonhomme and L. Romary. Apport de la Statistique Lexicale dans la Recherche d'Information. Premières Journées du Chapitre Français de l'ISKO, Lille, France. International Society for Knowledge Organization. Sept. 1997.

[3] P. Bonhomme and L. Romary. A contrastive indexing method and its integration in aquarelle folders. In ERCIM, editor, Proceedings of the third DELOS Workshop, 1997.