Vocabulary Switching and Automatic Metadata Extraction or How to Get Useful Information from a Digital Library

Jürgen Krause
Jutta Marx
German Social Science Information Centre (IZ)
Lennéstr. 30
D-53113 Bonn
krause|marx@bonn.iz-soz.de

1 Introduction

Nowadays, users of information services are faced with highly decentralised, heterogeneous document sources with different content analysis. To reduce this problem a great deal of work is carried out all over the world in the context of so-called virtual or digital libraries. The idea behind these efforts is to combine different information sources to solve the problem of finding useful information in the „borderless world of the internet“. In the virtual library projects of the Deutsche Forschungsgemeinschaft (DFG) and in a broader sense in the Global Information Initiative of the German Federal Ministry of Education and Research (bmb+f), the first objective is to combine all existing library catalogues. This bibliographic information will then be linked to grey literature of the WWW and of course with literature databases of information centers such as the German Social Sciences Information Centre (IZ), Bonn.

Unfortunately, most of these approaches are technologically oriented, with an emphasis on simultaneous access to distributed document pools with different structures (syntactic heterogeneity). The semantic heterogeneity, e.g. the fact that there are different systems for content analysis used with the resources tied together in a virtual library, is much harder to deal with. Standardization efforts such as the Dublin Core Initiative (DC) are a useful precondition for comprehensive search processes but they assume a hierarchical model of cooperation, accepted by all players.

Because of the diverse interests of those partners, such a strict model can hardly be realised. Projects should consider even stronger differencies in document creation, indexing and distribution with increasing „anarchic tendencies“. To solve this problem, or at least to moderate it, we suggest a system consisting of automatic transformation modules which map the users query to the underlying syntactic and semantic meta data of each database in question.

The IZ research group is trying to establish such a system by combining different approaches in the field of automatic extraction of metadata and manipulation of query terms by using cross-concordances and statistical transfer modules (vocabulary switching). First results have been acheaved in the context of the projects CARMEN[1] (Content Analysis, Retrieval and MetaData: Effective Networking), ViBSoz[2] (Social Science Virtual Library), ETB[3] (The European Schools Treasury Browser), and the already finished project ELVIRA[4] (Electronic Retrieval and Analysis System for Industrial Associations).

2 Metadata Extraction

The first step to handle semantic heterogeneity should be the attempt to enrich the semantic information about documents, i.e. to automatically fill up the gaps in the documents metadata. Especially in the domain of Social Sciences, metadata is not always consistent nor complete. To handle this problem a set of cascadeing deductive and heuristic extraction rules are applied to the documents.

The following example shows the extraction of a title tag from a htlm-document. The numbers in brackets indicate the significance (1 = lowest significance; 3 = highest significance). The remaining heterogeneity will be handled by the term manipulation, described below.

```
If (<title>-tag existing && <H?>-tag existing)
  If (<title>-tag==<H?>-tag highest order) {
    Titel[3]=<title>-tag
  } elsif (<title>-tag contains <H?>-tag highest order) {
```

[1] For more information see http://www.mathematik.uni-osnabrueck.de/projects/carmen/

[2] Meier, W., Müller, M.N.O. und Winkler, S. (2000). 'Virtuelle Fachbibliothek Sozialwissenschaften: Problembereich und Konzeption'. In: Bibliotheksdienst 34: 1236-1244.

[3] For more information see http://www.en.eun.org/etb/

[4] Krause, Jürgen; Stempfhuber, Maximilian (2000): Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA (Forschungsberichte 4).

```
    Titel[2]=<title>-tag
  } elsif (<H?>-tag highest order contains <title>-tag) {
    Titel[2]=<H?>-tag highest order
  } else {
    Titel[2]=<title>-tag + <Hx>-tag
  }
} elsif (<title>-tag existing) {
    Titel[2]=<title>-tag
} elsif (<H?>-tag existing) {
    Titel[1]=<H?>-tag highest order
} elsif (paragraph existing enclosed in <strong>|<b>) {
    Titel[0]= last paragraph enclosed in <strong>|<b>
  } elsif (paragraph existing enclosed in <em>|<i>) {
    Titel[0]= last paragraph enclosed in <em>|<i>
  }
}
```

The development of such rules has been made on the basis of a representative test corpus consisting of electronic documents from universities, domain specific information centers and libraries, primarily from the fields of mathematics, physics and social sciences.

3 Vocabulary Switching with Cross-Concordances

Cross-concordances are lists of intellectually acquired links between terms of different classification schemes or thesauri with similar meaning. In the above mentioned projects cross-concordances are created between universal and special classifications in the domain of mathematics, physics, education and social sciences (e.g. PICA Basic Classification vs. IZ Social Science Classification).

They are used to transform terms of the original query into terms from the classification or thesaurus specific to the target database. For example a query can be formulated by using the SWD, which is well-known to the users. The target database may be SOLIS which is indexed with the IZ Thesaurus. The term „Soviet Union“ has to be translated into the search term „USSR“ by a synonym relation for the query to be successful.

Problems my arise if there are no sufficient resources (time, money, domain experts) to create such cross-concordances or if not all documents within a virtual library are indexed with classifications or controlled vocabulary.

4 Vocabulary Switching with Statistical Transfer Modules

Statistical transfer modules - which need much less resources to be created - can be used to supplement or replace cross-concordances. They allow a statistical crosswalk between two different thesauri or even between a thesaurus and the terms of automatically indexed documents. The algorithm is based on the analysis of co-occurrence of terms within two sets of comparable documents.

The main problem of this approach is to find /to generate such documents of similar content. Unlike cross-concordances, the transformation is not based on intellectually acquired term-to-term links but on a weighted vector of terms.

Here are some examples of weighted relations between SWD (left hand side) and IZ Thesaurus (right hand side), translated into English for better understanding.

History 1933 - 1945 vs. Drittes Reich (weight 0.8571)
University vs. College (weight 0.8363)
USA vs. North America (weight 0.8802)
Knowledge Based System vs. Information System (weight 1.0) + Artificial Intelligence (weight 1.0)

Other interesting examples are compound terms, which may have to be split up into single words e.g. 'Jugendarbeitslosigkeit' - 'Jugendlicher / Arbeitslosigkeit'.

5 Further Research

Metadata extraction, cross-concordances and statistical transfer modules are evaluated in the context of ViBSoz and CARMEN. A report on current results will be given together with the paper presentation.