

## Document Clustering, classification and Data Mining

Some solutions suggested for the disambiguation of query terms include automatic query enrichment by using local feedback and local context analysis (Ballesteros and Croft, 1997) and allowing for simple machine translation geared toward the semantics of titles and other metadata used on small portions of text to allow the user to see results and then choose the correct term meanings (Hayashi, Kikui, and Susaki, 1997).

Less of a problem and more of a question is that of where in the system to have translation occur. The majority of work done on CLIR involves translation of the query, either manually, or sometimes, automatically. But as Hull and Grefenstette (1996, p.485) note, there is no reason in principle why the problem could not be approached using document translation. However, due to difficulties with automatic translation and the labor intensiveness of manual translation, this is likely to be less efficient for most systems at this point.

### Basic Approaches to CLIR

With a general understanding of what CLIR is and general problems it brings to the IR field, I will explore several important approaches to CLIR systems. Machine translation, controlled vocabulary and dictionary-based approaches will be discussed as well as latent semantic indexing and corpora-based approaches.

### Machine Translation

One approach to CLIR is to use machine translation (MT), which can automatically translate queries or documents. This can be beneficial because the query can be translated from the language of the user to another language for search, and the results can be translated back into the user's language for viewing (Fluhr, 1996). One of the few available examples of research done on MT with specific regard to CLIR comes from Fluhr and Radwan (1993) (as cited in Oard and Dorr, 1996).

Unfortunately, MT systems often make translation errors because of missing information in the term index or ambiguous definitions. Oard and Dorr (1996) note that MT only produces high quality translations for specific domains, such as those containing specific technical terminology, possibly because semantic accuracy suffers when insufficient domain knowledge is incorporated into a translation system.

### Controlled Vocabulary

The controlled vocabulary approach has long been the most dominant and effective for CLIR. One of the first CLIR system experiments (Salton, 1970) involved the use of controlled vocabulary, with surprising results for its relative simplicity. Usually with CLIR a multilingual thesaurus of some sort is created to hold a list of descriptors for each document in a collection and the semantic relations between them, and each term in the thesaurus must be translated for each language involved (Fluhr, 1996). The descriptors can be added to the thesaurus manually or automatically (if the system can learn from previous indexing which terms are likely to be important) (Fluhr). Sheridan and Ballerini (1996) used a multilingual thesaurus and query expansion with the SPIDER system to achieve results with Italian and German which were significant, although 32% less precise than those from using only a monolingual group of documents with monolingual queries.

[next](#) [previous](#)

---

©2005 Jatit