IMPROVING ACCESS TO MULTILINGUAL ENTERPRISE INFORMATION SYSTEMS WITH USER MODELLING

Context enriched cross-language IR

Alberto Díaz
Ingeniería T. Informática de Sistemas, Centro de Estudios Superiores Felipe II, Aranjuez, Madrid,Spain
Email: adiaz@cesfelipesegundo.com

Pablo Gervás
Departamento de Sistemas Informáticos y Programación, Universidad Complutense,Madrid, Spain
Email: pgervas@sip.ucm.es

Antonio García
Departamento de Comunicación, Universidad Rey Juan Carlos, Madrid, Spain
Email: antonio.garcia@cccom.urjc.es

Keywords: Multilingual, user modeling, enterprise information systems, cross-language information retrieval

Abstract: The enterprise systems for the processing and retrieval of textual information are usually based on the techniques used for the Internet search engines, incorporating natural language techniques, graph theory, as well as traditional information retrieval instruments. This paper presents a simultaneous integration of two additional processes into this task in order to improve the performance of this type of systems: user modeling and cross language information retrieval. The different situations that can appear in a multilingual framework are presented, as well as the techniques and resources to be applied to them. We describe the user modeling task and we defend its usefulness to a multilingual information system through the presentation of a prototype that has been applied to the electronic newspaper domain.

1. INTRODUCTION

Modern day systems for processing and retrieval of textual information rely on a number of technical tools that allow a semblance of intelligent access which enables us to cope with the information overload that prevails in most professional settings at the current time. Many different such tools have been developed in the last few years, generally in response to specific problems. Some of these tools are gradually becoming standard in information processing circles, and they are incorporated into any information management tool that aims to remain competitive. A reasonable benchmark for the success of everyday information access instruments is the basic Internet search engine, already incorporating natural language techniques (AskJeeves), natural language and graph theory (Google), as well as traditional information retrieval instruments. Two arrivals in the field that have yet to be incorporated into your run-of-the-mill search engine are user modelling (UM) and cross-language information retrieval (CLIR).

The traditional approach to solving the question of information access in the field of information retrieval (Salton and McGill, 1983) is: the user wants documents about a certain topic, he produces a free form textual description of the topic as a query, the information retrieval engine generates from this query a set of index terms, those index terms are matched against index terms derived form previously treated documents, the documents which match best are returned to the user in ranked order.

This traditional approach is based on two basic assumptions that need not necessarily apply in the context of information access within multinational corporations.

On one hand traditional IR assumes that both documents and queries will always be in the same language. This is more and more often not the case neither within the intranets of big corporations nor even in the more general sphere of information available on the Internet. The typical framework for information management becomes slightly more complicated when multilingual aspects are taken into account, depending on whether the query and the documents are in one or more languages, and various combinations thereof (query in several languages and documents in several languages, different parts of a document in different languages...). Even a more fundamental underlying assumption, that the user who writes a query in one language is not interested in documents in different languages however relevant, is progressively undermined by the existence of operative translation systems[1] or rapid glossing systems (Bauer et al, 1995). A user who receives documents in a language he does not understand can resort to these automatic tools to get some idea of the content. Cross language information techniques may provide a solution to this problem.

On the other hand traditional IR requires that the sort of query that a typical user writes when faced with an information need contain all the relevant information that may be useful for the system to find the most appropriate information. This is not true in most cases, a fact made especially apparent in our everyday experience with search engines on the Internet. Even the most concise query an experienced user can think of has a big potential for retrieving documents outside the topic the user had in mind. This is because a user who is searching for information does so in a specific context, and, when composing his query, tends to think in terms of relevant words within that context. However relevant those words may be in the context, the chances are that they also appear in many others which are not even remotely related. The search engine cannot tell the difference and will return them nonetheless, thereby producing a page of results with a lot of noise. The application of a customised profile of the user as background definition of the context in which the search is taking place can solve this problem.

2. MULTILINGUAL INFORMATION ACCESS AND USER MODELING

As a general rule, a context of documents and queries with several languages always requires a translation process. For example, a query in a specific language can be formulated by means of a translation of a query (vectors of terms of the query), translation of documents, or by a translation system, with all the problems that derive from it.

Using a information retrieval point of view there are three main possibilities: to translate the query into the target languages of the documents, to translate the documents into the source language of the query, or to map queries and documents into an intermediate (language-independent) indexing space.

2.1 Cross Language information retrieval

Electronically accessible information exists in many different languages, especially in the case of the intranets of multinational corporations. Providing the means for retrieving relevant information across language boundaries is an important challenge that has lead to the development of Cross Lingual Information Retrieval (Grefenstette, 1998a), a specific area of research devoted to the task of filtering, selecting and ranking documents that might be relevant to a query expressed in a different language.

Traditional IR works with the words of the original user query, and most of the effectiveness during retrieval comes from the retrieval engine finding those words (or related words with slight morphological alterations) in the relevant documents. Because different languages are involved in one and the other, in CLIR this is no longer possible. The main problem in a CLIR system is to know how a term expressed in one language may be written in another. Some means of obtaining the set of translations of the original term into the new language must be available. Different techniques have been employed in the past to solve this problem.

One favourite approach is to use machine readable bilingual dictionaries to obtain translations, either directly, by looking up words in the source query obtain words for the target query, or indirectly, by using a source language corpus to obtain context vectors for the words in the query, translating the words in the source vector to obtain a target vector, and using a target language corpus to

[1] http://www.babelfish.com

obtain an appropriate word for that context vector (Grefenstette, 1998a).

Another favourite approach is to use a parallel corpus, that is, the same text written in different languages. This can be done either by applying statistical techniques this can be used to produce bilingual term equivalents by comparing which strings co-occur in the same sentences over the whole corpus (Sheridan, 1996), or by means of Latent Semantic Indexing (Littman, 98; Oard, 1998; Evans, 98).

2.2 User modeling

The internal and communicative behaviour of an interactive computer system is affected by its knowledge about the user, which may be implicit in the design of the system, or explicitly available. User-adapted interaction, however, requires the use of an explicit model of the user, i.e., the knowledge about the user must be explicitly represented and modifiable, and the system has to contain mechanisms to exploit this explicit information to adapt its behaviour to specific users dynamically.

A user model (Herzog, 1999) is a knowledge source which contains explicit assumptions on all aspects of the user that may be relevant for the behaviour of the system. A user modeling component is that part of an interactive system whose function is to

- Incrementally build up a user model,
- To store, update and delete entries in it,
- To maintain the consistency of the model,
- To supply other components of the system with assumptions about the user.

Extending an information retrieval system with a user model provides enough additional information to allow an important reduction of the generic ambiguity problems inherent to the paradigm: in general it is difficult, perhaps impossible, for a user to define exactly his interests in the brief manner expected of an information retrieval system without involuntarily introducing a certain ambiguity. As a result of the extension with a user model, the amount of irrelevant information received by the user is greatly reduced. If a user has previously defined his general interests by means of a user model as well as specifically presenting a query of several words, and if the information access system can manage this information in a useful intelligent way, the information received by the user will be that much more relevant.

3. INTEGRATION OF USER MODELING WITH CROSS LANGUAGE INFORMATION RETRIEVAL

The main contribution of this paper lies in integrating cross language information retrieval techniques - which enable the system to handle queries and documents in two different languages - with a complex user-modelling component - specifically designed to allow cross language modelling and translation of language specific models.

3.1 Definition of a user model

The user model has to adapt to the different aspects of each domain, in order to allow a better definition of the user interests. To propose a specific user model we have to decide what is our domain. Different domains suggest different models for the users (Amato & Straccia, 1999).

We work in the electronic newspaper domain. Nowadays there are more than 6000 digital newspapers in Internet. Most of them offer basic search and delivery engines, as well as different personalization options. However, the opportunities of introducing improvements in that context are important, much more if newspapers in different languages are used. Our sources of information are two daily newspaper in two different languages, one Spanish and one English. The output is an email with the relevant news of the day per user.

Two models per user are constructed, one in each language, and each model is applied to the news in the corresponding language.

The model stores personal information (login, email, maximum number of news per message, etc) and information about the user interests in each language. These preferences are represented both in terms of structural and content-based information.

The sections of the newspapers acts as structural information. The user can select the sections of the Spanish newspaper and the sections of the English newspaper in which he is interested. The sections of a newspaper are not language independent, and they can not be translated directly.

An initial content-based information consists of a set of keywords (fine-grained interest), introduced by the user as interesting. This set of keywords is stored in the same language of the model.

As an additional content-based selection feature, an alternative classification of the news items, obtained by means of automatic categorisation of the documents against a different set of categories, is provided. Internet users are already familiar with the categories systems employed in search engine directories. The category system of Yahoo! has been chosen as an alternative way of representing interests. Since it is designed for a wider purpose and a wider domain than newspaper section headings, it constitutes a good second opinion. The first level of 14 categories from Yahoo! is presented as a choice (coarse-grained interest). These categories are language independent.

Users can show their preferences giving their degrees of importance on each element: without interest, of some interest, interesting, very interesting. The personalization architecture allows an extra level of user specification: each of the three features (sections, categories and keywords) has a weight that represents its importance for the user interests.

Our model approach allows users to define his long-term interests, but the model is not complete if it does not take into account the short-term interests of a user (Billsus & Pazzani, 2000). This kind of information is modeled by means of a keyword representation of those news items for which the user provides feedback, an issue described below.

3.2 User models and languages

To obtain one model per language we can proceed in two different ways:

- The user defines as many models as languages can be managed the system.
- The user only defines the model in his preferred language

It is very important not to overload the users. If we rely on a model of interaction in which the user is expected to define and update not just a single personal profile but one for each of the languages over which the system operates, users will soon give up hope of using the system in the real world. Statistics say that the typical user only introduces one or two words when he searches the web, so such a user is unlikely to devote time to filling in several models of his interests in different languages. Given these circumstances, we have opted for an approach where the user defines a single model in his preferred language, which is then processed by the system to obtain information to be applied in dealing with either language. This is more comfortable for the user and there are more possibilities that the user will not feel overloaded by the input requirements of the system.

After having decided that it is better that the user defines one model in his language, we are in the same situation that we described in section 2, that is, we are in a Cross Language Information Retrieval framework but with user models instead of queries. As we argued we have several possibilities to get multilingual access to all the documents.

User models usually contain less information than documents, and they can be processed in the same usual way as queries: we can translate the models instead of the documents. Moreover, we are then in a situation to achieve a much better translation. That is because a user model is not, in general, just a set of keywords, as queries usually are. A user model consists, in general, of some aspects that are relevant to the user and all the different aspects which constitute a context that provides more information to guide the translation.

3.3 The translation process

With our user model we only have to translate the set of keywords. In a first step we have used a combination of two of the techniques presented in the section 2. We used domain dependent machine translation with a restricted dictionary to reduce translation alternatives and when we have more than one translation we have chosen the first translation alternative appearing in the dictionary, which relies on the fact that in our dictionary the first translation alternative given is the most common. Moreover, when there are no translations we have used the same word in the two languages, for example, proper names.

3.4 The process of retrieval

The system applied the user models to the daily news, using text classification techniques. For each user, one relevance value per news item is computed and a ranking is constructed. The news items with more relevance are sent to the user in a email message.

The message that is received by a user contains: the name of the user, the date and for each news item, its title, its relevance, the section that it belongs to, a user-adapted summary (Acero et al., 2001) and a hyperlink to the news item in the digital newspaper. Moreover, the interests of the user are shown in the message by means of a synopsis of his user model.

The representation of the documents is obtained applying the Vector Space Model (VSM) to their texts (Salton, 1983). A representation for each category can be obtained by applying text categorisation techniques (Sebastiani, 1999) and using a set of training documents, in this case, the web pages indexed under the corresponding category by the version of Yahoo! for each language. The keywords are also represented with VSM, using the weight assigned for each word in the model. To perform the selection we applied cross language category-pivoted categorisation (Sebastiani, 1999) with the categories and cross language information retrieval with all the keywords. Also all the news items are processed to check if they belong to one of the sections selected in the user model.

3.5 Applying relevance feedback

The user has the possibility of providing positive or negative feedback on the items that he has been sent by voting for or against each specific one. Specific feedback keywords are obtained for the news items for which the user has voted, and they can have positive or negative weights, depending on the vote of the user (Nakashima & Nakamura, 1997). The additional information obtained in this way reflects a short-term interest of the user, as opposed to the a long-term interest that is reflected in the sections, categories and keywords.

When a user introduces feedback it is necessary to apply the translation process again. The vote of the user provides relevance information for a document in a given language, and this can be applied directly to the user model in that language. The system takes care - by means of a process of domain dependent machine translation with a restricted dictionary - of the additional processing required to apply this additional information to the model in the other language.

4. CRITICAL DISCUSSION AND EVALUATION

An important point to consider is that simple user models usually result in the introduction of irrelevant information. The integration of textual content analysis tasks can be used to achieve a more elaborated user model, to obtain a suitable representation of document contents, and to evaluate the similarity between user's interests and information. Representative examples of information access systems that integrate this kind of techniques are WebMate (Chen and Sycara, 1998), News Dude (Billsus and Pazzani, 1999) and SIFT (Yan and Garcia-Molina, 1995). WebMate is a tool that compiles information from a list of URLs that the user wants to monitor (e.g., newspaper home pages) or from the search results using popular engines. The information is selected by its accordance with a user profile, which represents these multiple interests using vectors of terms and its weights. News Dude is a system that compiles a personalized news program. Besides representing users short-term and long-term interests, it takes into account the news previously heard by the user to avoid presenting the same information twice. Finally, SIFT is an information filtering system that also models user's interest topics using keyword vectors provided by the user which are updated automatically by relevance feedback.

The user model proposed represents separately short-term needs and long-term multiple interests. Users can express their preferences both in terms of newspaper sections and news stories content. This representation works like a stereotypical definition that avoids starting with an empty user model that is trained by user feedback - as done in WebMate and NewsDude. In those cases, the initial training phase may become frustrating for users if many irrelevant news items are selected. This new approach solves, also, the difficulties presented for some users, beginners mainly, by the method of providing a set of keywords and their associate weights used by some filtering systems - like SIFT. Lastly, application of implicit feedback allows these initial definitions to be enhanced and to evolve together with user's interest.

An initial evaluation of a prototype of our system has given good feelings about the performance. This evaluation has been developed using a working pattern adapted to a monolingual version of the system used in previous experiments. This pattern includes several aspects as interface evaluation, newspaper sections, categories, summaries, bilingual capacity and user estimated recall and precision.

In general, users found the system suitable. They are satisfied with the different aspects of the user model, they estimate that the translation of the keywords is sometimes less than adequate but they value in a positive way the possibility to receive news in different languages.

We have yet to perform a more complete evaluation with a larger number of users and the relations between the different features that appear in our system must be studied in greater detail. For instance, how the multilinguality and the user modeling affect the traditional way of evaluating information retrieval systems, i.e. recall and precision measures.

5. CONCLUSIONS

The manner in which the user modeling and cross language information retrieval processes may find their way into an integrated information handling set up has been considered.

First, the different tendencies and possibilities to manage multilingual information have been presented, that is, the cross language information retrieval techniques and resources.

Then, we have presented the user modeling process and how it can improve access to multilingual enterprise information systems.

We have presented a set of methods to achieve an advanced user model that offers intelligent personalised access services to news. The proposed user model takes into account long-term and short-term multiple interests of the users and the changing character of these interests. Although the work presented is focused on journalistic field, the techniques introduced in the paper are applicable to various other domains.

The initial evaluation of a prototype has shown promising performance of the framework and more refined ways of evaluation are to be studied in order to obtain more definite conclusions.

REFERENCES

Acero, I., Alcojor, M., Díaz, A., Gómez, J.M., Maña, M. Generación automática de resúmenes personalizados. Procesamiento del Lenguaje Natural, 27 (2001), pp. 281-290

Amato, G. and Straccia, U., 1999. User Profile Modeling and Applications to Digital Libraries. In S. Abiteboul and A.M. Vercoustre (eds.), *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Lecture Notes in Computer Science, Vol. 1696, 184-197, Springer-Verlag.

Bauer, D., Segond, F., Zaenen, A., 1995. Locolex: The translation rolls off your tongue. In *Proceedings of the ACH/ALLC '95*, Santa Bárbara, California, July 11-15.

Billsus, D., Pazzani, M.J., 1999. A Hybrid User Model for News Story Classification. In: Proceedings of the Seventh International Conference on User Modeling, Banff, Canada

Billsus, D., Pazzani, M.J., 2000. User Modeling for Adaptive News Access. In: *User Modeling and User-Adapted Interaction* 10: 147-180. Kluwer Academic Publishers.

Chen, L., Sycara, K.P., 1998. WebMate: A Personal Agent for Browsing and Searching. In: Proceedings of the Second International Conference on Autonomous Agents, Minneapolis

Evans, D.A, Handerson, S.K., Monarch, I.A., Pereiro, J,. Delon, L. and Hersh, W.R., 1998. Mapping vocabularies using latent semantics. In Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Grefenstette, G., 1998. The Problem of Cross-Language Information Retrieval. In Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Herzog, G., 1999. User Modeling. In: http://www.dfki.uni-sb.de/fluids/User_Modeling.html. FLUIDS: Future Lines of User Interface Decision Support. Project in the *Telematics Engineering* sector of the Telematics Applications Programme managed by the European Commission (DG XIII).

Littman, M.L., Dumais, S.T. and Landauer, T.K., 1998. Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Nakashima, T., and Nakamura, R., 1997. Information filtering for the Newspaper. 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. August 20-22, 1997, Victoria, B.C., Canadá.

Oard, D., and Dorr, B.J., 1998. Evaluating cross-language text filtering effectiveness. In Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Picchi, E., and Peters, C., 1998. Cross-language information retrieval: a system for comparable corpus querying. In Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

Salton, G. and Mc Gill, M., 1983. *An Introduction to Modern Information Retrieval*, Mc Graw-Hill, New York.

Sebastiani, F., 1999. A Tutorial on Automated Text Categorization. In *Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99).*

Sheridan, P. and Ballerini, J.P., 1996. Experiments in multilingual information retrieval using the SPIDER systeem. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 58-65, August 1996, http://www-ir.inf.ethz.ch/Pulbic-Web/sheridan/papers/SIGIR96.ps.

Yan, T.W., Garcia-Molina, H., 1995. SIFT – A Tool for Wide-Area Information Dissemination. In: Proceedings of the USENIX Technical Conference