# Cross-Language Information Retrieval using Algorithms for Mining Association Rules

Mestrando: **André Pinto Geraldo**
Orientadora: **Viviane Pereira Moreira**

Programa de Pós-Graduação em Computação
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{andre.geraldo,viviane}@inf.ufrgs.br

***Abstract.*** *This work proposes the use of algorithms for mining association rules as an approach for Cross-Language Information Retrieval. These algorithms have been widely used to analyze market basket data. The idea is to map the problem of finding associations between sales items to the problem of finding term translations over a parallel corpus. The proposal was validated by means of experiments using different languages, queries and corpora. The results show that the performance of our proposed approach is comparable to the performance of the monolingual baseline and to query translation via machine translation, even though these systems employ more complex Natural Language Processing techniques. A prototype for cross-language web querying was implemented to test the proposed method. The system accepts keywords in Portuguese, translates them into English and submits the query to several web-sites that provide search functionalities.*

## 1. Introduction

Cross-Language Information Retrieval (CLIR) is the retrieval of documents in one natural language, based on a query formulated in another natural language, e.g. retrieval of documents written in English based on a set of keywords in Portuguese. The main motivation for CLIR is the growing need for exploring documents in foreign languages. This need has increased dramatically with the explosive growth of the Internet. The Web has content in many languages and the distribution of this content by language is very different from the distribution of Internet access. While English is still dominant in terms of content (~66%) [GlobalReach], the percentage of users that access the Internet in English is less than 30% [World_Internet_Statistics]. Despite being one of the languages with the largest number of native speakers, Portuguese is an extremely underrepresented language on the Web. It is estimated that only 1.4% of the Web's content is in Portuguese [GlobalReach]. This means that the 58 million Portuguese-speaking Web users are limited to a very small subset of the available information. Thus, a system which accepts search terms in Portuguese and retrieves documents in English is valid contribution to this community. We have built a prototype, based on the proposed approach for CLIR, which does that. This prototype is presented in Section 5.

CLIR systems can be used to break the language barrier and can be of use to people who are able to read in a foreign language but are not proficient enough to write a query in that language; e.g. a Portuguese speaker can be able to read documents in Spanish. CLIR can also aid people who are not able to read in a foreign language but have access to translations resources. Moreover, CLIR enables search for images and videos, which typically have metadata available in English only. Related work on CLIR is discussed in Section 2.

This work summarises the MSc dissertation which can be found at http://www.inf.ufrgs.br/~apgeraldo/dissertacao.pdf. As by-product of the dissertation, three papers were published:

(i) the proposed approach for CLIR using association rules was presented at SPIRE 2009 [Geraldo, Moreira et al. 2009]. and is summarized here in Section 3.

(ii) an extension of the BM25 ranking algorithm to further emphasize rare terms was published as a poster at SBBD 2007 [Geraldo and Moreira Orengo 2008].

(iii) experiments with the proposed approach for using association rules for CLIR were sub-mitted to the evaluation campaign CLEF 2008 [Geraldo and Orengo 2008] and are revised in Section 4.

This work has also been accepted for the MSc Dissertation contest at PROPOR 2010[1]. The three selected dissertations will be presented in April at the conference.

## 2. Related Work

The first research on CLIR was done by Salton, who showed that CLIR systems could perform nearly as well as monolingual systems, using a good quality thesaurus [Salton 1970] . According to [Grefenstette 1998], CLIR involves basically three problems: (i) knowing how a term expressed in one language might be written in another, i.e., crossing the language barrier; (ii) deciding which of the possible translations should be retained. Retaining more than one translation is useful in promoting recall. However, using wrong translations will reduce precision; and (iii) deciding how to properly weight the importance of translation alternatives when more than one is retained.

Many approaches have been proposed to solve these problems. These solutions typically use resources such as Machine Translate (MT) systems, Machine-readable dictionary (MRD), thesauri or multilingual corpora.

The approach for CLIR we propose is statistical. Other statistical approaches have been previously presented. [Nie, Simard et al. 1999] and [Kraaij, Nie et al. 2003] propose a probabilistic translation model which extracts translation probabilities from parallel corpora mined from the web. Their results are comparable to query translation using Systran [SYSTRAN].

CLIR systems that achieve the best results do so by combining several techniques, such as good quality translation resources, stemming (or decompounding), more elaborate weighting schemes, query expansion and relevance feedback [Savoy 2004; Agirre and Lacalle 2007].

## 3. CLIR using Association Rules

An association rule (AR) is an implication of the form $X \Rightarrow Y$, where $X = \{x_1, x_2, \ldots, x_n\}$, and $Y = \{y_1, y_2, \ldots, y_m\}$ are sets of items. The problem of mining ARs in market-basket

[1] http://www.inf.pucrs.br/~propor2010/

data was firstly investigated by [Agrawal, Imielinski et al. 1993]. In the rule "90% of customers that purchase bread also purchase milk", the antecedent is bread and the consequent is milk. The number 90% is the confidence factor (*conf*) of the rule, which is calculated according to equation 1. The confidence of the rule can be interpreted as the probability that the items in the consequent will be purchased given that the items in the antecedent are purchased. An AR also has a support level associated to it. The support (*sup*) of a rule refers to how frequently the sets of items $X \cup Y$ occur in the database. Equation 2 shows how the support of an AR is calculated.

$$conf(X \Rightarrow Y) = \frac{n(X \bigcup Y)}{n(X)} \quad (1) \qquad \sup(X \Rightarrow Y) = \frac{n(X \bigcup Y)}{N} \quad (2)$$

where *n* is the number of transactions and *N* is the total number of transactions.

Mining ARs means to generate all rules that have support and confidence greater than predefined thresholds.

Our proposal is to map the problem of finding ARs between items in a market-basket scenario to the problem of finding cross-linguistic equivalents on a parallel corpus. A parallel corpus provides documents in a language *B* that are exact translations of documents in a language *A*. The approach is based on co-occurrences and works under the assumption that cross-linguistic equivalents would co-occur a significant number of times over a parallel corpus. In this work, the transaction database is replaced by a text collection; the items that the customer buys correspond to the terms in the text; and the shopping transactions are represented by documents.

The proposed approach to use algorithms for mining ARs for CLIR is divided into the following five phases depicted in Figure 1:

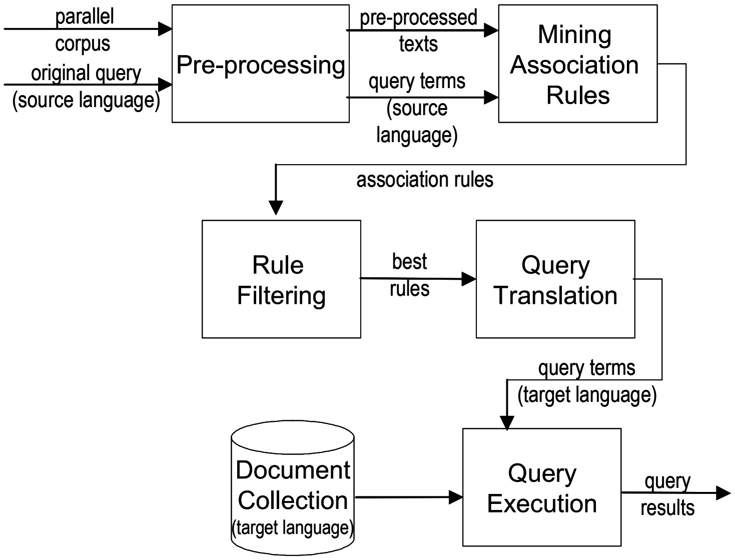

**Figure 1 – Phases of the proposed approach**

(*i*) **Pre-processing**: The inputs for this phase are a collection of parallel documents and the original query in the source language. During this phase, the original text and its equivalent in the other language are initially treated separately. We remove stop-words, apply stemming, and break the documents into sentences. The output of this phase is a set of pre-processed parallel sentences. During this phase, an inverted index containing all stems in the document collection and the list of sentences in which they appear is also built. The inverted index will be used in the next phase to enable selection of the sentences over which the algorithm for mining ARs is run.

(*ii*) **Mining ARs:** This step consists in generating ARs for the terms in the query. In order to speed up rule generation, only sentences that contain the query terms are considered. The output of this phase is a set of ARs for each query term.

(*iii*) **Rule Filtering:** The aim of this step is to keep the rules that most likely map a term to its translation. Rule filtering is based on the following heuristics:

- Select the AR with the highest confidence. Such a rule will be called *M* and it has the greatest chance of being the correct mapping.
- Select the ARs that have confidence of at least 80% of *M*.
- Select ARs with confidence equal to (100 – *M* ± 0.1). The rationale is that when word in language A is translated to two words in language B, the confidences of the ARs tend to be complementary to 100%.

*(iv)* **Query Translation:** Each term in the original query is replaced by all possible translations that remain after the filtering process.

*(v)* **Query Execution:** The last step is to execute the queries in a search engine. At this stage, the CLIR problem has been reduced to monolingual retrieval.

Some preliminary experiments we did for CLEF [Geraldo and Orengo 2008] showed encouraging results. Our approach was ranked amongst the top scoring methods. The test was done using collection of library catalogues in English and the query topics were in Spanish. The experiments described here use a different test collection and different languages for the query topics. In addition, we compare our proposed approach to MT and test the combination of the two. We also provide a deeper analysis of the results.

There are two basic strategies for generating the ARs to create a bilingual lexicon: (i) *eager* – mining rules for all terms in the collection a priori; and (ii) *lazy* – mining rules on demand for query terms only prior to query processing. Our approach mines the ARs on demand, according to a lazy strategy as advocated by [Veloso, Almeida et al. 2008]. In their work, the lazy strategy brings improvements in terms of the quality of the rules that are generated, because with this strategy, you can upgrade or replace the parallel corpus at any time. However, in our work the gain is in the number of ARs that are generated, as we only mine rules for the terms in the query. On the other hand, this strategy slightly delays querying. To speed up this process, we could build a cache of ARs. Only words that were not in the lexicon would need mining at query time.

The main advantage of our approach is that it is simpler than other co-occurrence based methods [Yang, Carbonell et al. 1997; Nie, Simard et al. 1999; Kraaij, Nie et al. 2003; Orengo and Huyck 2003] and yet the results are comparable or superior. The method does not require the generation of a term by document matrix, which is costly. The pruning of the itemsets that are below the thresholds for support and confidence allows for efficiency in terms of memory management. It is also simpler than MT systems, which typically need more complex Natural Language Processing (NLP) capabilities [Geraldo, Moreira et al. 2009].

## 4. Experiments

In order to validate the proposal, we carried out four sets of experiments in which we vary: (i) the language of the queries; (ii) the parallel corpus used to mine the ARs; (iii) the document collection used for searching; and (iv) the set of queries. The experiments aim at demonstrating the independence of the proposal in relation to these factors.

[Zettair] was the monolingual IR system used in all experiments in conjunction with BM25+ [Geraldo and Orengo 2008], our proposed ranking function. We used the Apriori Algorithm [Agrawal and Srikant 1994] to mine the ARs, . The document collections used in this study differ in two groups, the first consist of complete editions of the newspaper Los Angeles Times in 1994 (113,005 documents), and Glasgow Herald 1995 (88,874 documents). Each news document contains on average 569 terms. The second type of collection consists of meta-data from the British Library (1,000,101 documents). Each document has, on average, just 19 terms. The queries we used were from the CLEF Campaigns [CLEF].

In the first experiment we aimed at assessing whether the proposed approach would work for queries in different source languages. Queries in Finnish and Portuguese were used to retrieve documents from the LA Times collection, which is in English. The parallel corpus used to mine the ARs was a synthetic bilingual collection created by automatically translating 20% of the LA Times collection. We did not consider any relevance information in order to choose which documents to translate, thus not including any bias. The results have shown that our CLIR system achieves 88% of the monolingual performance in terms of mean average precision (MAP). A t-test has shown that this difference is not statistically significant at a 95% confidence level. In addition, the method performed consistently for both source languages.

For the second experiment, the goal was to evaluate different alternatives of parallel corpora to serve as basis for the mining process. Three alternatives were tested: (i) automatically translating a sample of the collection used for querying (AR-LATimes); (ii) automatically translating a sample of a collection from the same domain as the one used for querying (AR-GH); and (iii) using EuroParl (AR-EuroParl), a manually translated corpus. The results have shown that best performance was achieved by AR-LATimes, while the worst performance was achieved by AR-EuroParl. Recall-Precision curves are presented in Figure 2. We conclude that the domain of the documents used for mining the ARs plays an important role as when documents from a different domain are used, performance suffered. This experiment also tested automatically translating queries using [Google_Translator 2010] (MT-Google) and [LEC_Power_Translator] (MT-LEC). Our approach outperforms LEC and its result is comparable to GoogleTranslator. Intuitively, machine translation systems were expected to perform better than ARs as they employ much more sophisticated NLP methods. This lack of significant difference favours our simpler proposal. In order to test whether ARs and machine translation could be used in conjunction to improve results, we have also tested the combination of the best AR strategy with the best automatic translator

(MT+AR). The combination was done by performing a set union of the query terms generated from both strategies. For Portuguese, this run was significantly better than all other bilingual runs and even outperformed the monolingual baseline. This gain can be attributed to the query expansion effect brought by our approach. For more details on this experiment, please refer to [Geraldo, Moreira et al. 2009].

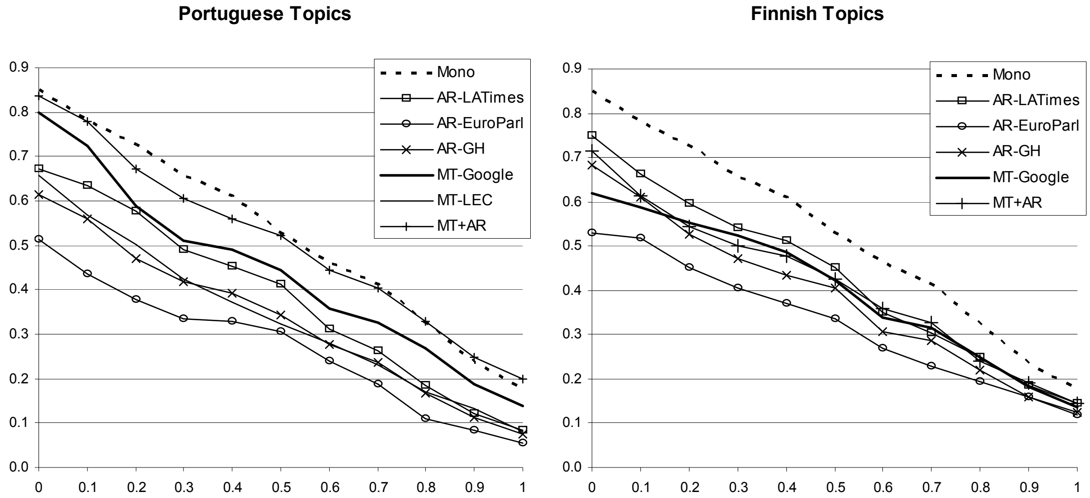

**Figure 2 Recall-Precision curves for the Portuguese and Finnish queries applied over an English collection.**

The third experiment compared the quality of the system in response to different sets of queries. The results have shown that the quality for both sets of queries are comparable. This experiment was submitted to CLEF 2008 Robust-WSD [CLEF] task and it was ranked in first place at the evaluation campaign

In the fourth experiment, we varied the collection used for querying. This time data was much more sparse. The results show that MAP is 86% of the monolingual result. This experiment was ranked by [CLEF] in third place amongst all participating groups. Note that this good result was obtained even without the use of techniques for enhancing the results such as relevance feedback which were employed by other participating groups. For more details on this experiment, please refer to [Geraldo and Orengo 2008].

## 5. Prototype for Cross-Language Web Querying

Despite being one of the languages with the largest number of native speakers, Portuguese is an extremely underrepresented language on the Web. It is estimated that only 1.4% of the Web's content is in Portuguese . This means that the 58 million Portuguese-speaking Web users are limited to a very small subset of the available information. Thus, a system which accepts search terms in Portuguese and retrieves documents in English is valid contribution to this community. We have built a prototype system which does that.

Our prototype was implemented in PHP and JavaScript. It uses asynchronous calls due to the large number of requests required. It is accessible from http://www.inf.ufrgs.br/~apgeraldo/busca/. Currently, it takes the query terms in Portuguese, translates them using our proposed approach and sends the corresponding English keywords simultaneously to four search engines (Google, Yahoo!, Bing and AOL Search), two image search engines (Google Images and Yahoo Images), two video libraries (YouTube and Yahoo Videos), one academic search (Scholar Google), two digital libraries (Wikipedia and ACM) and two e-commerce sites (Amazon and Ebay). The user will have the option to translate the search engines' results into Portuguese using a MT system.

The prototype can be easily extended to perform searches on image or video databases using their captions. Search on digital libraries or any web site that has a search box can also be added without difficulty.

## 6. Conclusions

This work proposed a new method for CLIR based on association rules to identify the co-occurrence of terms over parallel corpora. According to experiments performed, it was observed that the results obtained by the proposed method are comparable to those of a monolingual system and to the results of the state-of-the-art approaches.

Even the good results obtained by experiments, there are still many opportunities for improvements. In this study we only considered cases where a term in the source language is translated into one or more simple terms in the target language. Further work will include the treatment of cases in which one term translates to two or more terms.

## Acknowledgements

This work was partially supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brazil) and INCT-Web.

## References

Global Reach. http://global-reach.biz/globstats/refs.php3 accessed on 19-Oct-2007

Agirre, E. and O. L. Lacalle (2007). UBC-ALM: Combining k-NN with SVD for WSD. SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations. Prague**:** 342-345.

Agrawal, R., T. Imielinski, et al. (1993). Mining Association Rules between Sets of Items in Large Databases. Proc. of the ACM SIGMOD Conference on Management of Data. Washington, D.C.

Agrawal, R. and R. Srikant (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference. Santiago, Chile**:** 487-499.

Cross-Language Evaluation Forum. http://www.clef-campaign.org accessed on 17-May-2010

Geraldo, A. P. and V. Moreira Orengo (2008). Ajustando a importância dos termos: uma extensão à BM25. Anais da Sessão de Pôsteres do Simpósio Brasileiro de Bancos de Dados.

Geraldo, A. P., V. P. Moreira, et al. (2009). On-Demand Associative Cross-Language Information Retrieval. String Processing and Information Retrieval, 16th International Symposium (LNCS 5721). J. Karlgren, T. J. and H. Hyyro. Saariselkä, Springer**:** 165-173.

Geraldo, A. P. and V. M. Orengo (2008). Ajustando a importância dos termos: uma extensão à BM25. XXIII Simpósio Brasileiro de Banco de dados. Campinas, BR, SBC.

Geraldo, A. P. and V. M. Orengo (2008). UFRGS@CLEF2008: Using Association rules for Cross-Language Information Retrieval. Evaluating Systems for Multilingual and Multimodal Information Access (LNCS 5706). F. Borri, A. Nardi and C. Peters. Aarhus, Denmark, Springer**:** 66-74.

Global Reach. http://global-reach.biz/globstats/refs.php3 accessed on 19-Oct-2007

Google Translator 2010. http://www.google.com/translate_t accessed on 17-May-2010

Grefenstette, G. (1998). Cross-Language Information Retrieval. Boston, Kluwer Academic Publishers.

Kraaij, W., J. Nie, et al. (2003). "Embedding web-based statistical translation models in cross-language information retrieval." Computational Linguistics **29**(3): 381-419.

LEC Power Translator. http://www.lec.com/power-translator-software.asp accessed on 17-May-2010

Nie, J., M. Simard, et al. (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. SIGIR**:** 74-81.

Orengo, V. M. and C. R. Huyck (2003). Portuguese-English Cross-Language Information Retrieval Using Latent Semantic Indexing. Advances in Cross-Language Information Retrieval - Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 (LNCS 2785). C. Peters, M. Braschler, J. Gonzalo and M. Kluck. Rome, Italy, Springer.

Salton, G. (1970). "Automatic Processing of Foreign Language Documents." Journal of the American Society for Information Science **21**(3): 187-194.

Savoy, J. (2004). "Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval." Information Retrieval **7**(1-2): 121-148.

Systran. http://www.systransoft.com/ accessed on 22/01/2009

Veloso, A., H. Almeida, et al. (2008). Learning to Rank at Query-Time using Association Rules. SIGIR-08. Singapore**:** 267-274.

World Internet Statistics. http://www.internetworldstats.com/stats7.htm accessed on 17-May-2010

Yang, Y., J. Carbonell, et al. (1997). Translingual Information Retrieval. 15th International Joint Conference on Artificial Inteligence (IJCAI), Nagoya, Japan.

Zettair. www.seg.rmit.edu.au/zettair/ accessed on 17-May-2010