FACILITATING E-BUSINESS BY RETRIEVING RELEVANT BUSINESS OPPORTUNITIES ON THE INTERNET

Jing Bai, Jian-Yun Nie and François Paradis
DIRO, Université de Montraél, Québec, Canada

Keywords: e-Business, call for tenders, classification, retrieval.

Abstract: The Web is a useful medium that contains more and more business opportunities. However, it is often difficult to identify relevant ones using generic search engines. In this paper, we describe a system MBOI dedicated to the matching of business opportunities on the Web. It collects automatically calls for tenders, analyzes and classifies them. User profiles are automatically constructed to help document retrieval. Query translation is also provided in order to allow users to find calls for tenders written in a different language.

1 INTRODUCTION

Business opportunities become valuable only if they are offered to the right enterprises at the right time. As there are more and more business opportunities published on the Web, finding and selecting relevant ones is a crucial activity for businesses, as evidenced by the recent studies in Business Intelligence (Betts, 2003). A basic form of business opportunity announcement is call for tenders (CFT). A CFT announces the interest of a company or organization to purchase a good or a service. Large organizations can have staff dedicated to the purpose of finding CFTs, but smaller companies cannot afford it.

There are a large number of electronic tendering sites available on the Web, usually covering an economic zone (e.g. TED for the European Union, SourceCan for Canada and FedBizOpps for USA) or sector of activity. While these sites do increase the accessibility to the CFTs, their scope is limited in several respects: First, tendering sites are usually specific to some countries, specialization domains and languages. There is no site that federates all the tendering sites. Second, different sites use different standards to organize the data. For example, different classification schemas are used on TED, FedBizOpps and SourceCan. This can confuse the users when browsing through the hierarchy. Third, the sites do not provide information about the companies involved in CFTs, which is useful additional information for interested users. Finally, the browsing and search on these sites are user-independent. For users with different background, the same query words would result in the same answers. For example, a query on “tank” would strongly depend on the domain of interest to have a correct interpretation. It should retrieve different CFTs in military (a vehicle) and forestry (water tank) domains. A better system should take into account the user’s background in selecting CFTs.

This paper describes a system developed to facilitate the search of relevant CFTs on the Web. This system is capable of collecting CFTs from different sites, classifying them according to a given class schema, translating user’s query to a different language automatically, and determine the relevant CFTs according to user’s profile.

In the following sections, we will first describe the general architecture of the system called MBOI. Then we will describe the main functionalities implemented in it. An important technical contribution of our system lies in the utilization of user background. We will show that the CFTs retrieved can be more relevant when user background is considered.

2 THE MBOI SYSTEM

The MBOI system (Matching Business Opportunities on the Internet) is built from a research project supported jointly by the Natural Science and Engineering Council of Canada and Nstein technologies. Its aim is to facilitate the search for relevant business opportunities on the Web. It has been used by an organization of tendering

Bai J., Nie J. and Paradis F. (2007).
FACILITATING E-BUSINESS BY RETRIEVING RELEVANT BUSINESS OPPORTUNITIES ON THE INTERNET.
In *Proceedings of the Second International Conference on e-Business*, pages 174-179
DOI: 10.5220/0002115101740179

intermediate. Its basic functionalities are as follows:

- Automatic collecting of CFTs from other tendering sites;
- Information extraction and filtering from the collected CFTs to identify the key information;
- Automatic classification of CFTs according to a class schema;
- Automatic query translation into another language;
- Query refinement by considering user profile;
- Synthetic analysis of companies.

Figure 1 shows the system architecture of MBOI. The different processes will be discussed below.

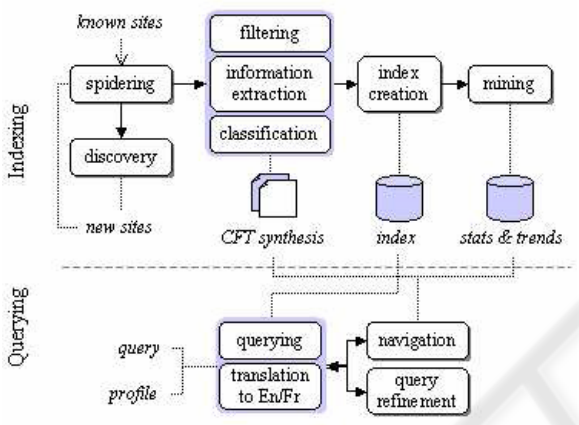

Figure 1: System Architecture.

2.1 Collecting Documents

The first challenge of such a system is to collect new CFTs from other sites. This is a problem similar to, however different from, Web crawler. There is much work lately on Intelligent Web Robots (Chau, 2003) and topic-focused Web crawling (Aggarwal, 2001). In our case, however, we are not just crawling for a *topic*, but rather for certain types of documents. Moreover, the typical crawl strategy might not be adequate here, since the information is often not linked directly (for example, a company site will not offer links to its competitors). To extract information, we use wrappers (Soderland, 1999), i.e. tools that can recognize textual and/or structural patterns, to collect new CFTs from 40 sites, including SourceCan, which is tendering site in Canada, FedBizOpps (Federal Business Opportunities), which maintains a central database of solicitations in US government, and regional or organization sites, which publish the tenders informally.

Below is a simple example of CFT for office supplies of the Saskatchewan government.

Reference Number: CFAB4
Source ID: PV.MN.SA.213412
Published: 2003/10/08
Closing: 2003/10/28 02:00PM
Organization Name: Saskatchewan Government
Title (English): Office Supplies
Title (French): Fournitures de Bureau
Description: The Government of Saskatchewan invites tenders to provide office supplies to its offices in Regina. The supplier is expected to start delivery on December 5, 2003, and enter an agreement of at least 2 years. Contact: Bernie Juneau, (306) 321-1542

The most important information in this CFT is "office supplies". This is the *subject* of the call, according to which retrieval and classification operations will carry out.

2.2 Information Extraction

CFTs are often semi-structured or non-structured. In many cases, the key information is hidden in the Description as a free text. In addition, the description usually contains only one or a few sentences related to the subject of the call, but a long description about the procedure, such as the submission deadline, the contact person, etc., which is not useful for identifying relevant CFTs.

To deal with this problem, a module of information extraction is used to recognize different types of information from the description part. These types include named entities (Maynard, 2001) (e.g. place, date, name of person or organization, etc.) and concepts. Named entities and concepts are extracted by a tool from our industrial partner of the project – Nstein technologies. This tool uses techniques similar to GATE (Cunningham, 2002) based on rules for the recognition of named entities, but the rules have been adapted to CFTs. In addition, terms (nouns or nouns phrases) that fit in some syntactic structures and appear quite frequently in the collection of CFTs are considered to encode important concepts for CFTs. The above process is aided by a dictionary of concepts. Figure 2 shows an example of CFT with concepts and named entities (goegraphical locations) tagged.

We have considered the following generic entities: geographical location, organization, date, time, money, URL, person, email and phone number. In addition, we have also considered the following specific entities to our collection:

- FAR (Federal Acquisition Rules). These are tendering rules for U.S. government agencies. A CFT may refer to an applicable paragraph in the FAR (e.g. FAR Subpart 13.5.).

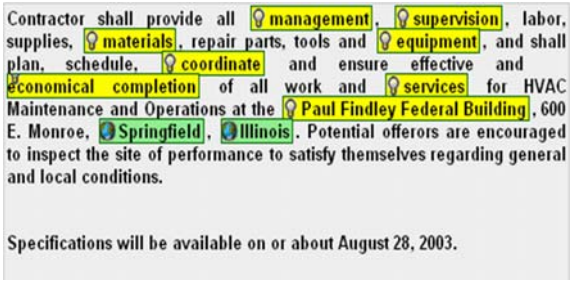

Figure 2: Result of information extraction.

- CLIN (Contract Line Item Number). The line item defines a part or sub-contract of the tender.
- Dimension. A dimension almost always refers to the physical characteristics of a product to deliver (e.g. .240MM x 120MM.).

The result of the information extraction process is useful for two purposes: 1). it can help determine which part of a CFT is related to the subject; 2). a user can choose a related concept or named entity to modify her query, for example, to narrow down a geographical location. We will explain these two utilizations in the following sections.

2.3 Filtering

As a CFT contains a large proportion of sentences (that we call procedural sentences) specifying the submission procedure, which are not useful for retrieval and classification, it is necessary and useful to remove them.

We observe that there are many named entities of certain types in procedural sentences, for example, person names, phone number, and so on. An intuitive approach is to use named entities to determine if a sentence is useful or not. In order to determine the capability of each type of named entity to predict the usefulness of a sentence, we have randomly picked 1 000 sentences, and manually classified them into important and non-importance sentences. The following table shows how each type of named entity can predict the useful sentences (+) or useless sentences (-). For example, phone number (a negative indicator) appeared in 40 sentences, 39 of which were labeled negative. Dimensions (a positive indicator) appeared in 8 sentences, all of which were recognized positive.

Locations and organizations are the most problematic entities, with very low accuracy. That is partly because they often appear along with the subject in an introductory sentence. For example the first sentence in our example CFT contains an organization (Government of Saskatchewan), the subject (office supplies) and a location (Regina). Therefore, we only use the other types of named entity to select or remove sentences. The grey zones shown in Figure 2 are the sentences that have been filtered out using this approach.

Table 1: Named entities in FBO documents.

type	Freq. in FBO	Accuracy
Location (-)	123344	50% (66/132)
Person (-)	48469	N/A
date & time (-)	170525	96% (101/105)
Money (-)	30606	100% (18/18)
URL & email (-)	29177	100% (38/38)
phone number (-)	25938	98% (39/40)
FAR (-)	142762	100% (56/56)
CLIN (-)	10364	80% (4/5)
Dimensions (+)	5290	100% (8/8)

2.4 Querying with User Profile

We implement the retrieval operation using statistical language modeling (LM) for information retrieval (IR) (Ponte 1998; Zhai 2001). LM approach has been shown to be very effective. In this approach, for each document D and query Q, we build a language model, $P(t|D)$ and $P(t|Q)$, reflecting the probability of each term t being generated from them. Then the ranking of the document is determined according to the following score, which is based on negative KL-divergence:

$$Score(D,Q) = \sum_{i \in V} P(t \mid Q) \log P(t \mid D)$$

It has been found that smoothing is important for document model to deal with the zero-probability problem. Several smoothing methods have been studied in IR (Zhai 2001). Here, we use the following Jelinek-Mercer smoothing:

$$P(t \mid D) = \alpha P_{ML}(t \mid D) + (1-\alpha) P_{ML}(t \mid C)$$

where P_{ML} is the Maximum Likelihood (ML) estimation of probability and α a smoothing parameter (set at 0.5 empirically).

On the query side, the traditional approach is to estimate it by ML. This results in a retrieval process independent from user profile. As we stated earlier, it is important to take into account the user background so that the retrieved CFTs can correspond better to the user's interests. In addition, user queries are usually very short. Users tend to omit some terms that seem obvious to them, but turn out to be important for IR systems. The addition of a user profile model allows us to recover part of the interesting terms omitted by the user.

To consider a user profile, we also use a language modeling approach. This choice is motivated by the ability of LM to extract important elements from a highly noisy context. This choice is

also preferred to a user profile defined manually by the user because users are often unable to define a correct profile themselves. A language model for a user profile, $P(t|U)$ is interpreted as the probability that the term t corresponds to a topic of interest of the user. In our case, we use the documents that a user has read, or the set of documents from the Web site of a company, to create a user profile.

For each query submitted by the user (or the company), the query is complemented by the user profile through the following smoothing process:

$$P(t \mid Q) = \lambda P_{ML}(t \mid Q) + (1-\lambda)P(t \mid U)$$

where λ is another smoothing parameter, which is also set at 0.5 empirically.

It would be interesting to test the impact of using user profile with real queries and relevance judgments. Unfortunately, there is not such a test collection for CFTs. However, we have simulated several profiles for companies such as Lockheed Martin, Canadian Coast Guard, etc., by collecting documents from the companies' Web sites. We also created profiles for different specialization domains: Military, Forestry, etc. These latter are created from the documents in the corresponding directory in ODP (Open Directory Project: http://dmoz.org). We have been able to see the difference of search results with and without user/domain profile. For example, with the query "tank", without user/domain profile, most of CFTs retrieved concern "gas tank". However, when the "Military" domain profile is turned on, the top CFTs concern military tanks. When the "Forestry" profile is turned on, we retrieve more CFTs concerning water tanks. For the query "rescue", with "Military" profile, we can retrieve CFTs concerning rescue operations of Canadian Coast Guard, while when "Forestry" profile is turned on, more CFTs are about forest fires.

Although these examples are insufficient to provide a strict measure of the impact of the user/domain profiles, we can get an intuition about the usefulness of user and domain profiles. A more formal evaluation is performed on TREC collections and it is described in (Bai et al. 2007). We have observed that once the domain model is integrated, the retrieval effectiveness (average precision) is increased on the order of 10%. This result shows that domain models can indeed improve the retrieval effectiveness. We can expect a similar (or even larger) effect on matching CFTs.

2.5 Querying and Query Refinement

Our index supports both simple keyword queries and precise queries over named entities. For example, the simple keyword query *bush* will return all documents where the word occurs, including documents about bush trimming and President Bush. In contrast, the precise query *person:Bush* will only return documents about (president) Bush.

We provide an interface for query refinement, where extracted information is shown and can be added to the query. This can be used to disambiguate terms (e.g. starting with keyword *bush* and then adding *person: Bush*) or to add criteria such as location or classification code.

Figure 3 shows a query and its results in our system. Here the query *snow removal* was entered. The bottom right part of the screen displays the results, in the usual manner, i.e. call for tenders are listed by order of relevance. Each document has a small excerpt (from its filtered contents) where the query keywords are highlighted, as well as some extracted information (here, the classification codes). The boxes on the left represent information extracted from the top 100 result documents. Concepts are phrases extracted by the concept extraction tool, which represent the salient ideas of the document. Organizations, locations and categories are the named entities discussed above.

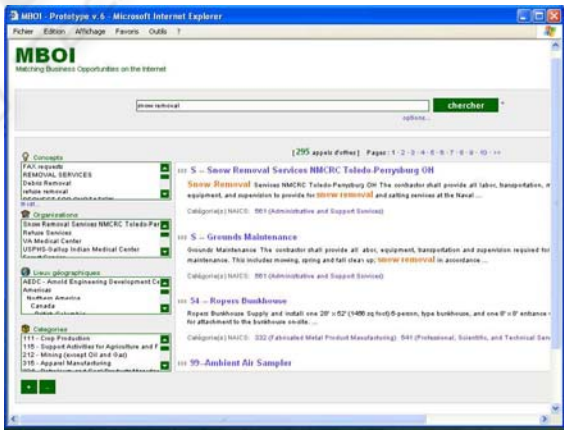

Figure 3: Querying in MBOI.

2.6 Query Translation

Query translation is an important aspect in international tendering. A user in a different country, who does not speak the language of another country, can still be interested in business opportunities in the latter country. However, most users cannot afford hiring a professional translator to translate their

queries. An automatic query translation can suffice for this step.

There have been many studies on query translation in IR (Peters, 2003). In our case, we use an approach based on parallel texts – texts with their translations in a different language. A statistical translation model is trained from a set of parallel texts, which tells the probability to translate a source word into a target word. We use IBM model 1 (Brown et al. 1991) as the translation model because this model does not consider the sentence structure and word order in queries. This corresponds to the situation of IR where queries often do not follow strict syntactic rules.

We use English-French as our test languages. Ideally, the translation model should be trained on a set of in-domain parallel texts. In our case, we have collected 100,000 pairs of documents from the TED site (Tenders Electronic Daily - for the European community) in both English and French. These texts are used to train two translation models in both directions.

As can be seen in Figure 4, the English query "snow blower" has been translated by French words "neige" (snow), "neigeux" (snowy), "soufleuse" (blower), etc. This translation, combined with the original query, allows retrieving relevant French CFTs.

Figure 4: Query translation.

2.7 Classification

There are many classification algorithms proposed in literature, such as SVM, Naïve Bayes, etc. (Yang, 2001). Naïve Bayes (NB) (Jason, 2003) is a simple approach, which has been shown to be reasonably effective, and it is very efficient in time. The classification task in NB is formulated as follows:

$$c(D) = \arg\max_{c} \prod_{t \in D} P(t \mid c) P(c)$$

where $P(t|c)$ is the probability of the term t in the class c, and $P(c)$ is the prior probability of class c. We use NB in our case. It has been found that feature selection is useful for NB (Yang, 2001). Following this study we select 8000 strongest features according to Information Gain.

2.8 Synthetic Analysis

Another aspect of our system is to present the user with useful synthetic information. This information includes:

- (Figure 5) The *hot list* for a given period: the top categories of CFTs and the top contracting authorities, together with the total amount of the contracts.
- (Figure 6) Company profile. This information is entirely extracted from the CFT and awards documents. It includes the known addresses for this organization, the categories of awarded contracts and business relationships.

Figure 5: Most active entities.

Figure 6: Company profile.

3 TESTS ON CLASSIFICATION

On FedBizOpps (FBO), calls for tenders have been manually classified according to two classification schemas, FCS (Federal Supply Code) and NAICS (North American Industry Classification System, http:// www.census.gov/naics). So we can use them to test the accuracy of classification. In our test, we only consider the first three digits of NAICS, i.e. the corresponding sector. There are 92 such categories.

We collected 21,945 CFTs from FBO, covering the period of September 2000 to October 2003. This collection is split into two parts: 60% for training, and 40% for testing. We used Rainbow package (McCallum, 1996) to perform NB classification.

Table 2 shows a comparison of the classification results with and without sentence filtering. What is the most interesting to observe is Micro-F1.

The sentence filtering reduces the size of the whole collection from around 600,000 sentences to 96,811. The results, identified in the table as *sent.filt*, show a strong increase in the micro-F1 measure (+7.6%). This shows that sentence filtering can be highly useful for the classification of CFTs. This allows removing many procedural sentences that are not directly related to the subject of the CFT.

Table 2: Classification on FBO.

method	macro-F1	micro-F1
baseline	.3297	.5498
sent.filt.	.3223	.5918 (+7.6%)

4 CONCLUSION

The system we described in this paper has been in use by our commercial partners, and deployed in several applications: as an aid for business opportunities watch, as a CFT search facility for the Canada's metal industry portal, and as an thematic watch for the travel industry. The system has been found very useful in all these applications, which shows that such a system would be of great help to facilitate the distribution of business information. We believe that the finding of relevant business opportunities is the first step to a business success. This is part of e-Business.

From a technical point of view, our study shows that sentence filtering brings a strong increase to classification accuracy (Micro-F1). User/domain profiles seem to be useful. Their usefulness has been formally tested in another study (Bai et al. 2007). All our results indicate that both information extraction and filtering, and user/domain profiles are highly useful for retrieving CFTs.

The system can be improved on several aspects: the translation module can be more precise; we can use a more effective classification approach such as SVM. However, the general approach presented here seems promising for business intelligence.

REFERENCES

Aggarwal, C.C., Al-Garawi, F. and Yu. P.S. Intelligent crawling on the world wide web with arbitrary predicates. *WWW Conference*, 2001.

Bai, J., Nie, J., Cao, G., Using query contexts in information retrieval, *Proc. SIGIR*, 2007, to appear.

Betts, M. The future of business intelligence. *Computer World*, 14 April 2003.

Brown, P.F., Pietra, S.A.D., Pietra, V.D.J. and Mercer, R.L. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19: 263-312, 1992.

Cai, L. and Hofmann, T. Text categorization by boosting automatically extracted concepts. *Proceedings of SIGIR*, pp.182.189, 2003.

Chau, M., Zeng, D., Chen, H., Huang, M. and Hendriawan, D. Design and evaluation of a multi agent collaborative web mining system. *Decision Support Systems*, 35(1):167.183, 2003.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *ACL*, 2002.

Jason, D. M., Rennie, Lawrence, Shih, J. T., & Karger, D. R. Tackling the poor assumptions of Naive Bayes text classifiers. *ICML*, 2003.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. Named entity recognition from diverse text types. *Recent Advances in Natural Language Processing*, pages 257.274, 2001.

McCallum, A. K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996. http://www.cs.cmu.edu/.mccallum/bow,

Peters, C.. Braschler. M., Gonzalo, J. and Kluck. M., editors. *Advances in Cross-Language Information Retrieval Systems*. Springer, 2003.

Ponte, J. and Croft, W.B., A language modeling approach to information retrieval. *Proceedings of SIGIR*, pp. 275-281, 1998.

Soderland. S. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 44(1), 1999.

Yang, Y. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67.88, 1999.

Zhai, C, and Lafferty, J., A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *Proc. SIGIR*, pp. 334-342, 2001.