Discovery of Business Opportunities on the Internet with Information Extraction

François Paradis and **Jian-Yun Nie**
Université de Montréal
Québec, Canada
{paradif,nie}@iro.umontreal.ca

Arman Tajarobi
Nstein Technologies
Québec, Canada
arman.tajarobi@nstein.com

Abstract

In this paper we describe a tool for the discovery of business opportunities on the Web. The aim of our system is to help a user decide which call for tenders should be examined further. The project and its goals will be presented. We then focus on one aspect, *named entity* extraction, and its use to improve classification and user navigation. We show that filtering contents improves the classification results, but so far we did not get a significant improvement by combining named entities and extraction. However there are benefits to the interface to facilitate query refinement and collection browsing.

1 Introduction

Finding and selecting business opportunities is a crucial activity for businesses, as evidenced by the recent interest in Business Intelligence [2]. A key functionality is the ability for suppliers to find Call for Tenders (CFT) relevant to their businesses, based on their domain of expertise, the geographical location, the submission/execution dates, the contract size, etc. Sometimes the only information freely available is the solicitation notice: a fee must be paid to obtain the full documentation. Thus it is important to select the right candidate CFTs.

Large organisations can have staff dedicated to the purpose of finding CFT. For smaller organisations, or for secondary suppliers that might not respond directly to CFT but want to monitor the activity in their sector, tendering services are offered. Typically a professional will have a portfolio of enterprises and send them digests of candidate CFTs on a routine basis. Many electronic tendering sites are also available, usually covering an economic zone (e.g. "TED" for the European Union, "SourceCan" for Canada) or sector of activity (e.g. "NetMetal" for the metal industry in Canada). While these sites do increase the accessibility to the data, an organisation trying to go outside is usual sphere of activity will still have difficulties. Firstly it has to deal with country/domain specific standards, such as regulations and classification codes. Secondly it might be missing some contextual information, such as the recurring nature of the tender, the tendancy of the contracting authority to select local suppliers, etc.

Our solution is to augment the original data with extracted information from CFT and other supporting documents, and to cross-reference this information in such a way that the user can discover the patterns of a company or an industry, or relate to information she is familiar with. For example, multi-classification can provide more than one angle to a CFT (for instance, a code for a product-oriented schema, and a code for a service-oriented schema), or an alternative schema when the user is not familiar with the schema provided in the CFT. By combining information from CFTs and awards, a pattern of the activity in a sector, or the business relationships between organisations, can be presented to the user.

Two challenges are: how to discover new CFT and related documents on the Web, and how to extract information from these documents, knowing that the Web offers no guarantee on the structure and stability of those documents. There is a much work lately on Intelligent Web Robots [4] and topic-focused Web crawling [1]. 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, *wrappers* [13], i.e. tools that can recognise textual and/or structural patterns, will have limited success because of the diversity and volatility of Web documents. Free-text techniques such as named entity extraction [12], do not assume a particular document model or structure, but have lower accuracy.

In this paper we describe a tool to help the discovery of business opportunities on the Web. In particular we focus on Information Extraction, and its use to improve classification and user navigation. We first present the general framework of the project and the test collection we built to evaluate our results. We then discuss the application of Information Extraction in the project. We present some results of filtering contents based on named entities. We also show how these named entities are used to aid the user refining her queries and browse the information space.

2 The MBOI Project

The MBOI project (Matching Business Opportunities on the Internet) deals with the discovery of business opportunities on the Internet. In the first phase of the project we have implemented a tool to aid a user in this process. This first

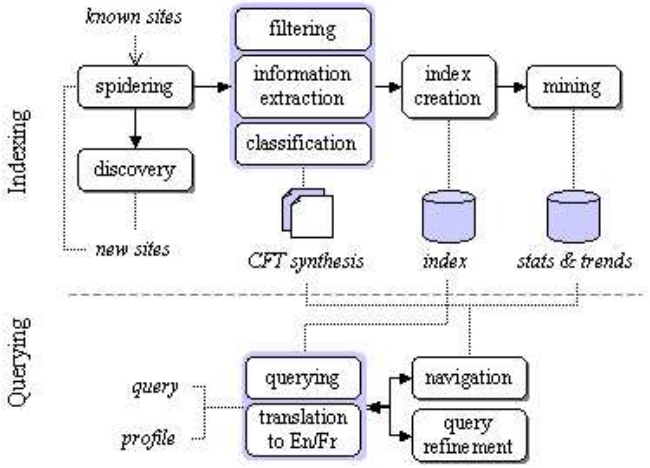

Figure 1: System Architecture

system, called CERVO was developed by CIRANO (Interuniversity Centre of Research on the ANalysis of Organizations), and was used by SDTI (Information Technology Development Service) and its professionals to help businesses located in the Ste-Hyacinte area (Quebec) in the search for CFTs. Other applications followed, notably a business portal for the metal industry in Canada, NetMetal. In the latter, rather than defining profiles for each organisation, some key profiles were defined corresponding to simple themes; users could then express complex queries by combining these profiles. Finally, it should be noted that CERVO was also applied to another domain, namely the tourism industry, in a project with UQAM (University of Quebec in Montreal). There again, although the information needs are typically less precise and more changing, the basic process is very similar to a business watch.

Figure 1 shows the system architecture for the second phase of the project. The various processes for collecting the information, extracting information, index and mining, querying and navigation, are discussed below.

2.1 Collecting documents

The first step of indexing is *spidering*, i.e. to collect documents from Web sites. In the earlier work with CERVO, a list of 40 sites was defined. These include *aggregator* sites, such as SourceCan, which collect information from other Web sites, *government* sites, such as FedBizzOpps (Federal Business Opportunities – i.e. U.S. agencies), which maintain a central database of solicitations sent to them, or finally regional or organisation sites, which publish the tenders informally. At the moment our robot needs to be told where to find the documents: it is supplied with a URL seed or pattern, and a username/password if access is restricted. It can also fill out forms if CFTs can only be accessed through search forms.

The discovery of new sites has not been implemented yet. It will consist of following promising links from the known sites, using a model for business relationships, and the extracted information.

Eventually we would like to collect a wide range of documents: press releases, solicitation notices, awards, quarterly reports, etc. At the moment however we collect only solicitation notices and *awards* i.e. details about the winning bid to the tender. Although the solicitation notice and its award should share a common identifier, it is not always straightforward to pair them up, because they can exist on different sites, and the publisher can substitute their own identifier to the contracting authority. Different versions of the solicitation can exist (e.g. French and English). Moreover, amendments can be published.

Figure 2 (top) shows a simplified pre-solicitation notice for the office supplies of the Saskatchewan government. The notice was published on the Merx site, along with an amendment to change the date of the delivery from December 5, 2003 to January 5, 2004. The two documents share a common identifier (`CFAB4`).

2.2 Information extraction

The next step is the extraction of information and classification. The extracted information is put in a *CFT synthesis*, which is an XML document inspired from xCBL (Common Business Language) and OASIS UBL (Universal Business Language) [5].

Business-related documents, in particular CFTs, are typically classified according to an industry standard, for example, NAICS (North American Industry Classification System) or CPV (Common Procurement Vocabulary, for the European Union). Some CFTs are manually classified with these codes, whereas some others are not. However we note that some of the contents of the CFT is not relevant for classification because it is not indicative of the topic. For example, in figure 2, only the first sentence is deemed relevant. The second sentence states the delivery date and contractual obligation, and the last sentence provides a contact for more information. We consider this *procedural language* as noise, and try to remove it before classification by *filtering*. At a more technical level, filtering could also be simply removing HTML layout tags.

Figure 2 shows an example of a CFT synthesis, which combines information from the presolicitation notice and its amendment (represented as two `published-document` elements). It is identified by the Saskatchewan Government with number `031021-5`, and by Merx with number `CFAB4`. The synthesis includes the free text description of the CFT (`description` element) and the various extracted elements: for example the contracting authority (Saskatchewan government), contact person (Bernie Juneau), classification code (NAICS `418210`), etc.

Elements in the synthesis can be repeated if they apply more than once to the CFT. In our example there is a French and an English title because the notice on Merx was bilingual. Similarly there could be multiple classification codes because the CFT was classified under several schemas. Elements can also have an associated confidence measure, to represent the accuracy of information extraction. In our example there are two execution dates, but the amended date is given more confidence.

Presolicitation (on Merx):

Reference Number: CFAB4
Source ID: PV.MN.SA.213412
Published: 2003/10/08
Closing: 2003/10/28 02:00PM
Organisation 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

Amendment (on Merx):

Reference Number: CFAB4
Description: The start delivery date has been revised to January 5, 2004.

Synthesis (inferred in MBOI):

```
<call-for-tender>
<contracting-authority-solicitation-id>
    031021-5</...>
<title xml:lang="en">Office Supplies</title>
<title xml:lang="fr">Fournitures de Bureau</...>
<description xml:lang="en">The Government of
Saskatchewan
    invites tenders...</description>
<date-closing>2003-10-28</date-closing>
<execution-date-start confidence="0.1">
    2003-12-05</...>
<execution-date-start confidence="0.9">
    2004-01-05</...>

<classification>
    <classification-system>NAICS</...>
    <code>418210</code>
</classification>

<published-document>
    <publisher-id>Merx</publisher-id>
    <publisher-solicitation-id>CFAB4</...>
    <original-url>...</original-url>
    <cached-url>...</cached-url>
    <date-published>2003-10-08</...>
    <date-cached>2003-10-09</date-cached>
    <language>eng</language>
    <format>text/html</format>
    <document-type>presol</document-type>
</published-document>
<published-document>...
    document-type>amendment</document-type>
</published-document>
<contracting-authority>
    <name>Saskatchewan Government</name>
    <authority-type>provincial government</...>
</contracting-authority>
<contact><name>Bernie</name>
    <surname>Juneau</surname>
    <phone>(306) 321-1542</phone>
</contact>
</call-for-tender>
```

Figure 2: A Call for Tender

2.3 Indexing and mining

Indexing is the process of building an *inverted file* or index to support querying. In our case the index is partioned into *fields*, corresponding to elements in the CFT synthesis, so that we can combine free text search over all contents, with search for specific dates, locations, classification codes, etc. The retrieval engine we are using is *lucene*. Dates and money amounts are represented in a special format to accomodate the range queries supported in lucene, e.g. to search for money amounts between $50,000 and $100,000.

Other information is also pre-computed and stored in a database (data *mining*). This information is inferred from CFT synthesis; it is used to provide statistics about CFTs or information about organisations. At the moment the inference is quite simple, but in the future it is possible to envisage technique akin to business intelligence.

2.4 Querying and navigation

There are two possible uses of *profile* in our system. One is to have the user define a recurring need (query), so that the system can send results automatically whenever CFT are found. Another is to save contextual or background information and use it transparently to affect querying. The first was implemented in CERVO, and we are currently working on the second.

Our system includes CFTs in French and English. In order to allow a searcher to find documents in another language, we have implemented translation of queries using CLIR (Cross-Language Information Retrieval) techniques [11]. To this end, we have build a statistical translation dictionary from a collection of 100,000 pairs of documents of the TED site (Tenders Electronic Daily - for the European community).

3 Test collection

It is well-known that the results of information extraction and classification techniques vary according to the domain and collection. Since no CFT collection was available, we defined our own, using pseudo-XML documents from FedBizOpps (FBO). The appeal of this collection is that some information is already tagged, and therefore can ease the evaluation.

To eliminate duplicates, we kept only one document per CFT, i.e. chose a document amongst and pre-solicitations and amendments. We only considered documents with two classification codes, FCS and NAICS[1], so that we can later measure the effectiveness of conversion. We obtained 21945 documents, covering the period from September 2000 to October 2003. Finally, we splitted this collection in two: 60% for training, and 40% for testing.

The NAICS codes are hierarchical: every digit of a six-digit code corresponds to a level of the hierarchy. For example, in figure 2, the industry code is `418120` (Recyclable Paper and Paperboard Wholesaler-Distributors) and the sector code, `418` (Miscellaneous Wholesaler-Distributors). Each of

[1]Since the NAICS codes were not tagged in XML at the time (as they now are), they were extracted from the free text description. To ensure 100% precision, very strict patterns were used: e.g. "The NAICS code is XXXXXX".

the three participating countries, the U.S., Canada and Mexico, have their own version of the standard, which mostly differ at the level of industry codes (5th or 6th digit). However, there are exceptions, as demonstrated by our example: the equivalent code in the American version would be `424120` (Stationery and Office Supplies Merchant Wholesalers) and the sector code, `424` (Merchant Wholesalers, Nondurable Goods).

We reduced the category space by considering only the first three digits, i.e. the corresponding sector. This resulted in 92 categories (vs. 101 for FCS). We did not normalise for the uneven distribution of categories: for NAICS, 34% of documents are in the top two categories, and for FCS, 33% are in the top five.

As mentioned above, filtering can play an important role in the classification of CFTs. We thus built a subset of the collection in order to experiment with filtering techniques. We manually labeled 1000 sentences from 41 documents. The label was "positive" if the sentence was indicative of the tender's subject, or "negative" if not. Sentences with descriptive contents were labeled positive, while sentences about submission procedure, rules to follow, delivery dates, etc. were labeled negative. In the example of figure 2, only the first sentence would be labeled positive. Overall, almost a quarter of the sentences (243) were judged positive.

4 Information Extraction

Although CFTs do not follow a standard format, they usually contain the following information: contracting authority, openning and closing date, legal notices on the conditions of submission, etc. These pieces of information are often present but mixed up in a semi-structured text. For example the location can be tagged as "`<td>Location</td><td>Montreal</td>`", identified with "`Location: Montreal`", or simply appear in the description body: "`This contract is to be performed in Montreal.`".

This information is called *named entities* [7] in the litterature, and usually refers to names of people, organisations, locations, time or quantities. Their extraction proceeds from a syntactic or lexical analysis, possibly with the help of a dictionary. In our experiment we use NStein *NFinder*, which uses a combination of lexical rules and dictionary.

4.1 Filtering

In a previous study [9] we have examined the performance of named entities extraction on a small sample of 40 documents from our collection. Note surprisingly we found that money extraction, a relatively simple task, obtained the highest score, while organisation and person scored the lowest. Some of the problems we identified were: four-digit SIC (Standard Industrial Classification) codes incorrectly identified as years, acronyms eroneously idenfied as organisations (e.g. "frequency domain helicopter electromagnetic (HEM)"), two capitalised words identified as a person (e.g. "Space Flight", "Will Result"), missing US state abbreviations (e.g. TX, FL), etc.

This time we are interested to see if those entities can be used predict the relevance of a sentence in a call for tender. We have considered the following entities:

- geographical location. In a call for tender, this can be an execution or delivery location. A location can also be part of an address for a point of contact or the contracting authority (although these are preferably tagged as metadata in FBO, they often appear in the text body).
- organisation. Most often the organisation will be the contracting authority or one its affiliates. For predetermined contracts it can be the contractor.
- date. This can be a delivery date or execution date (opening and closing dates are preferably tagged as metadata).
- time. A time limit on the delivery date, or business hours for a point of contact.
- money. The minimum/maximum contract value, or the business size of the contractor.
- URL. The Web site of the contracting authority or a regulatory site (e.g. CCR).
- person. A point of contact in the contracting agency.
- email, phone number. The details of a point of contact.

The entities above have a particular use in our collection. However they are generic in the sense that they also apply to many other domains. We have also considered the following entities, specific to our collection:

- FAR (Federal Acquisition Rules). These are tendering rules for U.S. government agencies. A call for tender may refer to an applicable paragraph in the FAR (e.g. "FAR Subpart 13.5").
- CLIN (Contract Line Item Number). The line item define a part or sub-contract of the tender. Line items usually appear as a list (e.g. "CLIN 0001: ...").
- dimensions. In the context of a tender, a dimension almost always refers to the physical characteristics of a product to deliver (e.g. "240MM x 120MM").

All entities except CLIN and dimensions are negative indicators: their presence is an indication of a negative passage or sentence, i.e. not relevant to the subject of the tender. CLIN and dimensions on the other hand are positive indicators, since they introduce details about the contract or product.

We have extracted entities using NStein *NFinder* for the top four entities, and simple regular expressions for the others. Table 1 summarises the results by entity type. The first column gives the total frequency of the entity in the FBO collection. The second column shows the accuracy of the entities as positive/negative indicators on the 1000 sentences subcollection of FBO. 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 labeled positive.

Locations and organisations are the most problematic entities, with very low accuracy. That is partly because they often

Table 1: Named entities in FBO documents

type	Freq. in FBO	Accuracy as a predictor
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)

appear along with the subject in an introductory sentence. For example the first sentence in our example CFT contains an organisation (Government of Saskatchewan), the subject (office supplies) and a location (Regina).

4.2 Classification

CFTs are often classified by industry type, according to the several existing standards: SIC (Standard Industrial Classification), NAICS (North American Industry Classification System), FCS (Federal Supply Codes), CPV (Common Procurement Vocabulary), etc. The classification codes are not always included in documents, and even when they are, it is interesting to classify the same CFT according to other standards. For instance, an American user will probably be familiar with NAICS, but maybe not with CPV (the European Union standards). Furthermore, these standardss are regularly updated, and the different codes between two versions can be the source of errors. We can make these conversions explicit by classifying CFTs according to several standards: e.g. CPV, NAICS version 1997/Canada, NAICS version 2002, etc.

Table 2 shows results of the classification of NAICS codes on our FBO collection, using the title and description fields. The first result is our baseline: a Naive Bayes classifier [6] where the 8000 top terms were selected according to their InfoGain score. The following thresholds were applied: a rank cut of 1 (rcut), a fixed weight cut of 0.001 (wcut), and a category cut learnt after cross-sampling 50test set over 10 iterations (scut). More details about these thresholding techniques can be found in [14; 15]. Our results were obtained with *rainbow* [8].

The second line, "patterns", is the classification result after replacing named entities in the collection with generic names. For example, "$42.00" with "CU" (for currency). This will force the classifier to consider some terms which would normally not be indexed (such as numbers), and diminish the impact of some other terms (such as locations). Not surprisingly it does not affect results much, because these entities are pretty much uniformely distributed over the classes. We tried different combinations of entities, with little difference in the results. A micro-F1 score of .5466 was obtained with the most accurate entities: date, time, money, URL, email, phone number, FAR and dimensions.

We trained a Naive Bayes classifier on the 1000 sentences subcollection of FBO, for the positive and negative classes. The task seems to be relatively simple, since when we tested

Table 2: Classification on FBO

method	macro-F1	micro-F1
baseline	.3297	.5498
patterns	.3199	.5466
sent.filt.	.3223	.5918 (+7.6%)
sent.filt.patt.	.3497 (+6.1%)	.5939 (+8%)

the classifier on a 40/60 split we obtained a micro-F1 measure of 85%. We thus filtered the whole collection with this classifier, keeping only the positive sentences. The collection size went from around 600,000 sentences to 96,811. The new, filtered documents were then classified with another Naive Bayes classifier. The results, reported in the table as "sent.filt." show a strong increase of the micro-F1 measure (+7.6%).

Finally the last line in table 2 combines sentence filtering with patterns, i.e. the sentence filtering classifier was retrained with the patterns in the sentences. We expected a good increase since the classifier could now generalise better from its small training set, but the results were somewhat disapointing. The macro-F1 shows the strongest increase, +6.1% over the baseline, or +8% over the non-pattern filtering. This tend to indicate that the use of patterns did not change the overall measure (micro-F1), but provided more stability in the classes results (macro-F1).

4.3 Query Refinement

Our index supports 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. However the 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 Nstein *Nconcept* tool, which represent the salient ideas of the document. Organisations, locations (*lieux géographiques*) and categories are the named entities discussed above.

The user can refine her query by selecting an entity and adding it to the query with "+". She can also require the absence of an entity (i.e. the NOT operator) with "-". Figure 4 shows a "refined" query, where the location "Illinois" has been added to the keywords "snow removal". More constraints can be added on the dates, awarded contract amount (*montant du contrat*), or awardee (*fournisseur*). Finally the

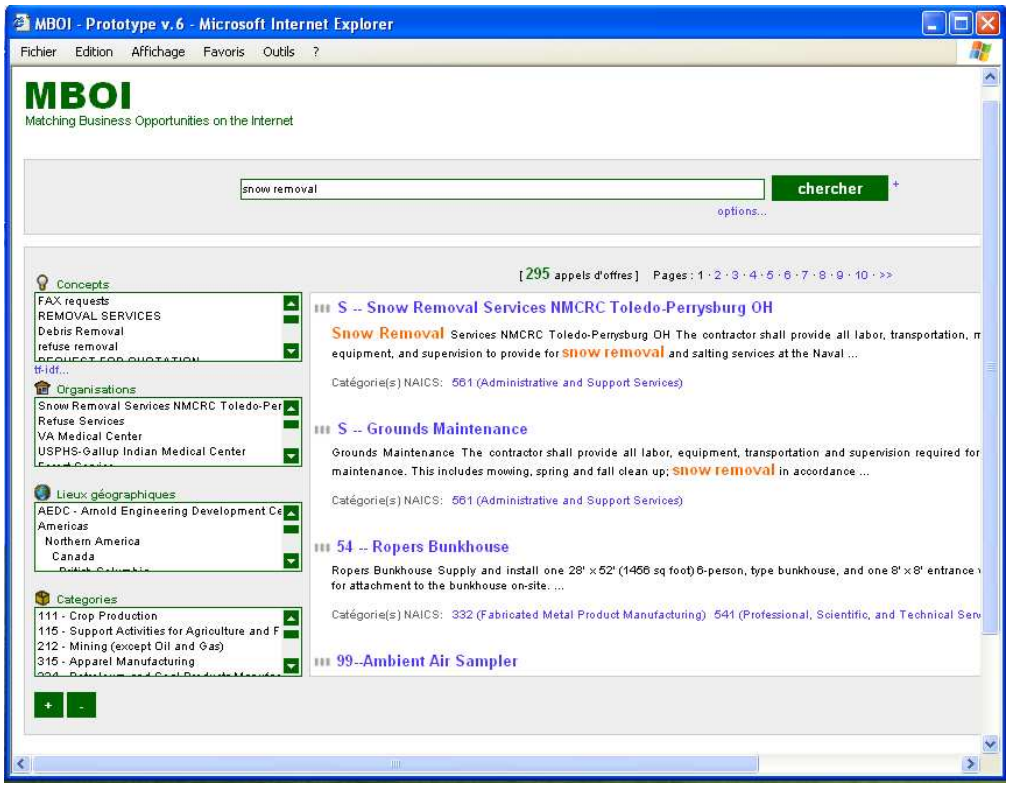

Figure 3: Querying in MBOI

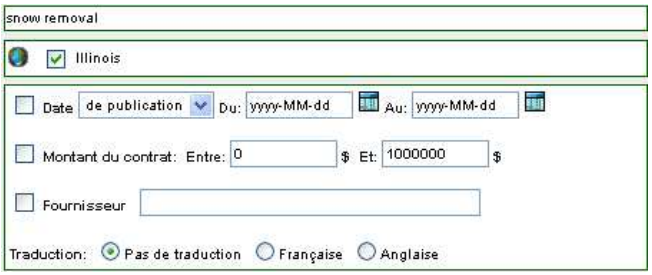

Figure 4: Query Refinement

query can be translated (*traduction*) to French (*Française*) or English (*Anglaise*).

4.4 Navigation

Another aspect of our use of information extraction is to let the user browse the information space starting from the named entities.

For example figure 6 shows the top entities or "hot list" (*palmarès*) for a given period.

The first block shows the top categories (*activités*), i.e. categories that had the most number of CFTs (*appels d'offres*). Selecting a category sends a query to retrieve the associated CFTs. Similarly the classification tree can be navigated, and the relevant CFTs displayed at each node.

The second block shows the top contracting authorities (*organismes adjudicateurs*), this time based on the amount of the

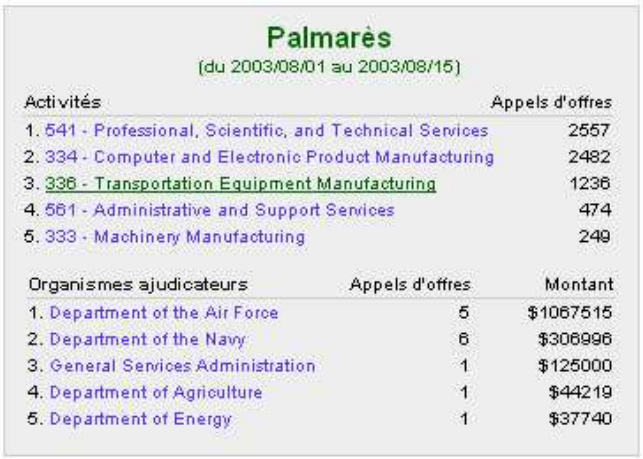

Figure 6: Most active entities

contracts (*montant*). Clicking on a link this time brings up a company profile page, as shown in figure 5. This information is entirely built from the CFT and awards documents. The first box shows the known addresses for this organisation (here, its different branches), the categories of awarded contracts (*activités des contrats octroyés*), and business relationships (*relations d'affaires*), which in this case are the awardees to its tenders. Finally the CFTs for this organisation are listed.

Figure 5: Company profile

5 Conclusion

Our tool has been in use by our commercial partners, and deployed in several aplications: as an aid for business opportunies watch for the St-Hyacinthe (Québec) region, as a CFT search facility for the Canada's metal industry portal (NetMetal[2]), and as an "issue" or "thematic" watch for the Quebec travel industry.

Our study of information extraction and classification techniques show results comparable to those reported in the literature. We have shown that sentence filtering brings a strong increase to classification. However the combination of filtering and named entities seemed to bring only a small increase in the macro-F1 measure. In another work we tried using controlled vocabulary derived from the classification schema, but could not demonstrate significant gain [10] either. We believe the advantages on the interface to query refinement and collection browsing are obvious; however we would like to be able to evaluate the querying the same way we evaluated classification. To that end, we are currently working with SDTI to build a query collection.

Another way to improve classification is to combine it with information extraction. Our first experiments with conditional classification, i.e. using the conditional probabilities observed on the training set between two classification standards, have not been up to our expectations. A possible way to improve these results is to combine it with fixed conversion rules that are available between some standards. We are also trying to use *concepts*, as extracted by Nstein *Nconcept* to inmprove classification. Preliminary results are promising. This is supported by other works [3], which have shown that term indexing could be improved with latent semantic indexing and boosting strategies.

Other directions for future work incude Intelligent Web Robots [4], Multilingual Information Retrieval [11], and Filtering. Multilingual Information Retrieval can help international tenders be aware and submit for calls published in foreign languages. This is especially important considering that contracting authorities tend to favor local tenders, and therefore publish only in the local language.

[2] `http://www.netmetal.net/`

Acknowledgments

This project was financed jointly by Nstein Technologies and NSERC. We would also like to thank the following people who have made this project possible: Claude Martineau, from SDTI and Robert-Gérin Lajoie from Cirano.

References

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

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

[3] L. Cai and T. Hofmann. Text categorization by boosting automatically extracted concepts. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 182–189, 2003.

[4] M. Chau, D. Zeng, H. Chen, M. Huang, and D. Hendriawan. Design and evaluation of a multi-agent collaborative web mining system. *Decision Support Systems*, 35(1):167–183, 2003.

[5] E. Dumbill. High hopes for the universal business language. *XML.com, O'Reilly*, November 7 2001.

[6] J. T. Jason D. M. Rennie, Lawrence Shih and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[7] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named entity recognition from diverse text types. In *Recent Advances in Natural Language Processing*, pages 257–274, 2001.

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

[9] F. Paradis, Q. Ma, J.-Y. Nie, S. Vaucher, J.-F. Garneau, R. Gérin-Lajoie, and A. Tajarobi. Mboi: Un outil pour la veille d'opportunités sur l'internet. In *Colloque sur la Veille Strategique Scientifique et Technologique*, 25–29 October 2004.

[10] F. Paradis and J.-Y. Nie. Étude sur l'impact du sous-langage dans la classification automatique d'appels d'offres. In *CORIA*, 9–11 mars 2005.

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

[12] E. F. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147, 2003.

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

[14] Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67–88, 1999. An excellent reference paper for comparisons of classification algorithms on the Reuters collection.

[15] Y. Yang. A study on thresholding strategies for text categorization. In *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval*, 2001.