MBOI: Discovery of Business Opportunities on the Internet
Extended Abstract

Arman Tajarobi, Jean-François Garneau
Nstein Technologies
Québec, Canada
{arman.tajarobi,jf.garneau}
@nstein.com

François Paradis
Université de Montréal
Québec, Canada
paradifr@iro.umontreal.ca

We propose a tool for the discovery of business opportunities on the Web, more specifically to help a user find relevant *call for tenders* (CFT), i.e. invitations to contractors to submit a tender for their products/services. Simple keyword-based Information Retrieval do not capture the relationships in the data, which are needed to answer the complex needs of the users. We therefore augment keywords with information extracted through natural language processing and business intelligence tools. As opposed to most systems, this information is used at all stages in the back-end and interface. The benefits are twofold: first we obtain higher precision of search and classification, and second the user gains access to a deeper level of information.

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. A major hurdle to the discovery of new documents is the poor degree of "linkedness" between businesses, and the open topic area, which makes topic-focused Web crawling (Aggarwal et al., 2001) unapplicable. To extract information, *wrappers* (Soderland, 1999), i.e. tools that can recognise textual and/or structural patterns, have limited success because of the diversity and volatility of Web documents.

Since we cannot assume a structure for documents, we exploit information usually contained in CFTs: contracting authority, opening/closing date, location, legal notices, conditions of submission, classification, etc. These can appear marked up with tags or as free-text.

A first type of information to extract are the so-called *named entities* (Maynard et al., 2001), i.e. names of people, organisations, locations, time or quantities. To these standard entities we add some application-specific entities such as FAR (regulation number), product dimensions, etc. To extract named entities we use Nstein NFinder™, which uses a combination of lexical rules and a dictionary. More details about the entities, statistics and results can be found in (Paradis and Nie, 2005a).

We use another tool, Nstein Nconcept™, to extract *concepts*, which capture the "themes" or "relevant phrases" in a document. NConcept uses a combination of statistics and linguistic rules.

As mentioned above, CFTs not only contains information about the subject of the tender, but also procedural and regulation information. We tag passages in the document as "subject" or "non-subject", according to the presence or absence of the most discriminant bigrams. Some heuristics are also applied to use the "good predictors" such as URL and money, or to further refine the non-subject passages into "regulation". More details can be found in (Paradis and Nie, 2005b).

Another information to extract is the industry or service, according to a classification schema such as NAICS (North American Industry Classification System) or CPV (Common Procurement Vocabulary). We perform multi-schema, multi-label classification, which facilitates use across economic zones (for instance, an American user may not be familiar with CPV, a European standard) and confusion over schemas versions (NAICS version 1997/Canada vs. NAICS version 2002). Our classifier is a simple Naive Bayes, trained over 20,000 documents gathered from an American Government tendering site, FBO (Federal Business Opportunities). Since we have found classification to be sensitive to the pres-

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 30–31,
Vancouver, October 2005.

ence of procedural contents, we remove non-subject passages, as tagged above. The resulting performance is 61% micro-F1 (Paradis and Nie, 2005b).

Finally, a second level of extraction is performed to infer information about organisations: their contacts, business relationships, spheres of activities, average size of contract, etc. This is refered to as *business intelligence* (Betts, 2003). For this extraction we not only use CFTs, but also awards (i.e. past information about successful bids) and news (i.e. articles published about an organisation). For news, we collect co-occurences of entities and classify them using a semantic network. For example, the passage "Sun vs. Microsoft" is evidence towards the two companies being competitors.

The extracted information is indexed and queried using *Apache Lucene.*, with a Web front-end served by *Jakarta Turbine*. The interface was designed to help the user make the most of the extracted information, whether in query formulation, document perusing, or navigation.

Our system supports precise queries by indexing free-text and extracted information separately. For example, the simple keyword query "bush" returns all documents where the word occurs, including documents about bush trimming and president Bush, while the query "person:Bush" only returns documents about President Bush. However such queries are not very user-friendly. We thus provide an interface for advanced queries and query refinement.

The extracted information from the 100 top query results is gathered and presented in small scrollable lists, one for each entity type. For example, starting with keyword "bush", the user sees a list of people in the "person" box, and could choose "Bush" to refine her query. The list is also used to expand the query with a related concept (for example, "removal services" is suggested for "snow"), the expansion of an acronym, etc.

Queries can be automatically translated using Cross-Language Information Retrieval techniques (Peters et al., 2003). To this end we have built a statistical translation model trained from a collection of 100,000 French-English pair documents from a European tendering site, TED (Tenders Electronic Daily). Two dictionaries were built: one with simple terms, and one with "concepts", extracted as above. The intuition is that simple terms will offer better recall while concepts will give better precision.

The interface shows and allows navigation to the extracted information. When viewing a CFT, the user can highlight the entities, as well as the subject and regulation passages. She can also click on an organisation to get a company profile, which shows the business intelligence attributes as well as related documents such as past awards or news.

We are currently expanding the business intelligence functionalities, and implementing user "profiles", which will save contextual or background information and use it transparently to affect querying.

Acknowledgments

This project was financed jointly by Nstein Technologies and NSERC.

References

Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. 2001. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings International WWW Conference*.

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

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

François Paradis and Jian-Yun Nie. 2005a. Discovery of business opportunities on the internet with information extraction. In *IJCAI-05 Workshop on Multi-Agent Information Retrieval and Recommender Systems*, 31 July.

François Paradis and Jian-Yun Nie. 2005b. Filtering contents with bigrams and named entities to improve text classification. In *Asia Information Retrieval Symposium*, 13–15 October.

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

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