Ontology Driven Web Extraction from Semi-structured and Unstructured Data for B2B Market Analysis

S.M. Hazzaz Imtiaz
School of Electronics & Computer Science
University of Southampton, UK.
+44 (0)23 8059 5415
hsmi@ecs.soton.ac.uk

John Darlington
School of Electronics & Computer Science
University of Southampton, UK.
+44 (0)23 8059 9045
jd@@ecs.soton.ac.uk

Landong Zuo
School of Electronics & Computer Science
University of Southampton, UK.
+44 (0)23 8059 5415
lz@ecs.soton.ac.uk

ABSTRACT
The Market Blended Insight project[1] has the objective of improving the UK business to business marketing performance using the semantic web technologies. In this project, we are implementing an ontology driven web extraction and translation framework to supplement our backend triple store of UK companies, people and geographical information. It deals with both the semi-structured data and the unstructured text on the web, to annotate and then translate the extracted data according to the backend schema.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods - *Semantic networks*.

I.2.6 [**Artificial Intelligence**]: Learning - *Knowledge acquisition*.

I.2.7 [**Artificial Intelligence**]: Natural Language Processing - *Language parsing and understanding*, *Text analysis*.

General Terms
Algorithms, Languages.

Keywords
Semantic web, Ontology Driven Web Extraction, Market analysis.

1. BACKGROUND AND RELATED WORK
In today's competitive economic environment, it is important for a business to understand emerging trends, segment its market based on the functions, products and services of its prospective clients, scan beyond typical structured business data sources and interpret a market opportunity. There is a huge amount of semi and unstructured text relating to organizations available in public domain on the web. The organizations typically describe themselves in their websites using various attributes, behaviors, relationships with clients, suppliers etc. The opinions of consumers and between organizations or the products that it wants to market are stated in discussion forums, reviews, media or news sites, people blogs etc. Understanding these information about prospective clients as well as competitors, effectively and on a timely basis is important for an organization to formulate their marketing strategy. The project has concentrated on developing an ontology driven web information extraction system which generates semantic content for the topic of interest to the marketing user. It consists of domain specific wrappers generated by hand, named entity recognition with relevant grammars and gazetteers, relation extraction from both semi and unstructured data etc. The ontology driven web extraction provides obvious advantages such as exploiting the ontological data in class labels, synonyms for wider search, restrictions for verifying instances etc. The extraction is focused as we are only looking for interesting classes and relations existing in that ontology and not other content. Finally the system interface to the search engines opens it to a much wider range of documents.

Several approaches have been proposed for extracting semantically annotated data from the web. McDowell [1] proposed an ontology driven , domain independent, web extraction system named 'OntoSyphon'. It takes any ontology as input, uses that to specify web searches that identify possible semantic instances, relations, and taxonomic information. Yildiz [2] used ontology contents and predefined OWL[2] semantics for the automatic extraction rule generation process. It uses words in the concept names and properties and populates their values by extracting closely preceding or following values matching that particular datatype. Schutz [3] proposed an ontology extension mechanism by relation extraction from text in the Football domain using the RelExt tool. It identifies triples which are pairs of concepts connected through a relation or verb and measure their relevance in terms of a highly ranked subject and a highly ranked direct or indirect object, which is then integrated into an existing ontology. Our approach is similar to that of McDowell in that we are ontology driven, automatic and only extract the instances of the ontology concepts the user is interested in and similar to Schutz in that we also extract relation triples. However we deal with both semi and un-structured text for knowledge extraction with some site specific wrappers and crawler configurations. We also concentrate on a set of organization websites to mine for relevant facts and have the option to use a crawler, count relevance and focus crawl the user's pages of interest.

2. FOCUSED WEB CRAWLING
We have used WebSphinx [4]web crawler and re-implemented the crawler to run in an ontology focused crawling mode similar to that of Ehrig [5]. The use of topical or focused crawling mode can

Conference'04, Month 1–2, 2004, City, State, Country.

[1] The Market Blended Insight project (DTI Project No: TP/5/DAT/6/I/H0410D) is a three year applied research project funded under the UK Governments Technology Programme.

[2] Web Ontology Language (OWL) http://www.w3.org/2004/OWL/

keep it limited to relevant topics of marketing user's interest. It takes in the basic crawling parameters being specified as a root webpage or several root web pages, depth level, regular expression link visit pattern, maximum number of pages, breadth or depth first crawl etc. Additionally for running in an ontology focused mode and computing the relevance based on it, it also takes in the background ontology and the entities we are interested in that ontology. The documents are preprocessed with a GATE[1] pipeline containing a morphological analyzer to get to the word roots, relevant gazetteers and JAPE[2] grammars to annotate semantic entities contained in the ontology. The page relevance is scored by counting the number of entities of interest, entities which are linked to these by the taxonomy or by relations in the ontology graph multiplied by different weight measures as described in [5].By limiting the crawl only to a specific website we can discover all the pages of interest in that website. From the relevant pages, for text mining, we can extend the ontology with instances as described in section 4.

3. SEMI STRUCTURED DATA EXTRACTION WITH WRAPPERS

The JAPE[2] grammar is written as a set of phases where each phase may consist of one or more pattern/action rules with a priority ordering. Having a sequential phasing means, a progressively complex annotation can be built at later stages, where annotations generated by the rules of previous stages are used in the rules at the later stages. For example: before we annotate a whole address, we annotate the postcode, county, city, street name and house number, using a combination of gazetteers and jape grammars. Similarly, extraction rule for a contact person may consist of his name and optionally his address, email, phone number etc.

The CETF keeps a set of jape scripts as wrappers which covers standard HTML page structures such as table or list wrappers. The table header cells are mapped to that concept's attributes in the backbone ontology using a string distance measure such as the Levenshtein Distance algorithm [6].If its below a threshold, a WordNet[3] based semantic similarity metric of Lin [7] is employed. In case of nested tables, a concept may be linked to many other concepts through object properties and a column may contain concepts with attributes instead of just attributes of the main concept. For example: a table may describe the instances of the person concept with the columns: first name, surname, age, email, phone number etc. Therefore if the person concept in the backbone ontology has similar attributes, they will be mapped to it. When it contains address of the person in a column, the extractor needs to call the appropriate wrapper to annotate address and get different address attributes such as postcode, city, street etc. and map it to the corresponding object property in the generated dataset. Figure 1 shows the inferred company roles and relationships, extracted from Architect Journal Specification[4] and The Barbour ABI[5] websites.

There are plenty of free web directory services for organizations. Some list companies of a particular business such as Applegate directory for manufacturers, some are trade associations such as glass and glazing federation, some are local council company directories, some list companies by their roles such as UK wholesalers etc. They usually list companies by some hierarchy

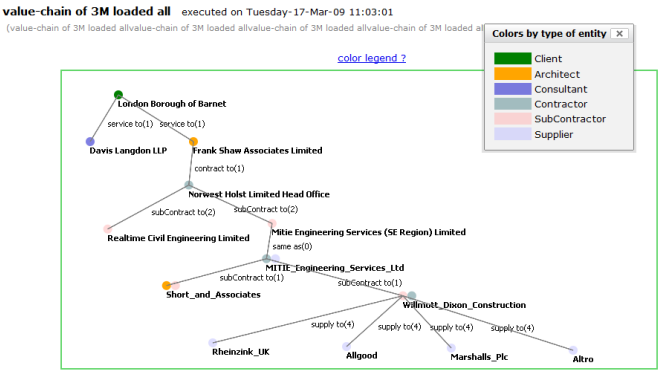

Figure 1. Company roles and relationships extracted from Ajspec[2] and Barbour ABI[3] website

based on business activity, alphabetical list and often contain company pages reflecting that business details. Thus a crawler can exploit these rules to get to a company page to extract the company details such as address, important contact persons, a textual description and also its business activity hierarchy. We can map this hierarchy to ours and create a finer level micro-segmentation of companies. For example a company listed just as a restaurant in our backbone can be discovered as an Italian restaurant in a directory website hierarchy or in the textual description. Currently the CETF stores the crawler configuration for several directory websites and wrappers in the form of jape scripts to extract company information such as contact data, business activity and contact person details. The user can select the key concepts i.e. company, address, person etc. to extract from any of them.

4. UNSTRUCTURED FREE TEXT EXTRACTION

To extract relational triples, the system takes as input the relations that we want to extract, their argument types and some of its seed instances to boot strap the process. We consider the binary relationships of two entities. The well known 'hyponymy' or is-a and the 'meronymy' or part-whole binary relations can be expressed using a small number of lexico-syntactic Hearst patterns [8]. Patterns like "NP_0 such as {NP_1, NP_2 . . . . (and | or)} NP_n", where NP stands for noun phrase i.e. "Software vendors such as Microsoft or Oracle", indicate hyponymy relation. We have exploited these natural language patterns to build up our gazetteer of instances of desired types using the Google search API. The figure 2 shows some of the instances extracted for the search engine query "high value electronic items|goods|objects such as" which may be useful for a postal service provider building up an ontology of high value goods with instances, to extract from its prospective customers websites.

Banko [9] has shown that by using a few syntactic patterns, most of the other binary relation patterns can be grouped into categories such as E_1 verb E_2, E_1 verb preposition E_2, E_1 NP preposition E_2

[1] GATE, A General Architecture For Text Engineering http://gate.ac.uk/

[2] JAPE: Regular Expressions Over Annotations, http://gate.ac.uk/sale/tao/index.html#x1-1790007

[3] WordNet, http://wordnet.princeton.edu/

[4] Architect Journal Specification, http://www.ajspecification.com/

[5] Barbour ABI, http://www.barbour-abi.com/xml_feed/rss_project_news_clickthrough_barbour-abi.htm

etc. where E_1, E_2 indicate semantic categories i.e. PERSON established COMPANY. Important business facts such as company A acquiring company B, or company A merging with company B or Person X becoming CEO of company A etc. can be extracted mostly using the above patterns. For annotating these relations and be able to run inference on them, argument phrases

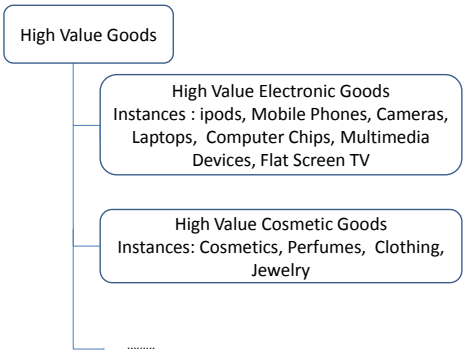

Figure 2. Instances of classes generated using Hearst patterns with Google search API

are normalized which involves mapping them as objects to some known concepts. If the term in question doesn't belong to a semantic category i.e. hasn't matched with the gazetteer instances or with the grammar pattern, another mapping is attempted using a string distance measure Levenshtein [6] with a threshold value. If it is a multi-word expression such as a noun phrase, the syntactic head of the noun phrase is taken into account and previous steps repeated. Next, relation phrases are normalized, ignoring differences in verb root forms and trying to match it with the verbal expression of any existing relation between the classes of instances under consideration. If the relation doesn't match, the string distance matching as in the previous step is attempted. If it fails, its synonyms are checked against all possible relations between the classes. If it is a multi-word expression, the most representative word or the verb of that expression is taken into account and previous steps repeated. To identify new relations between entities as well as new entities, the process starts with a set of seed patterns [10] and each candidate pattern in the corpus is scored using some function and against these seed patterns. The high scoring patterns, above a threshold, are included in the set of seed patterns. However, Stevenson [10] considered pattern elements to belong to either a semantic category or a lexical item. But, the class of entities the project has considered so far i.e. organization, person etc., it would not be accurate to judge similarity for lexical items. Thus only relations between classes of entities are considered here. The WordNet based semantic similarity metric of Lin [7] is used for our purpose.

The project is currently working to use dependency analysis, where a sentence can be represented using a set of directed binary links between a head and its modifiers where the links are the relations such as subject, object etc. Using the dependency tree, one can use different information extraction models. Greenwood [11] has provided a comparison of these models i.e. predicate-argument structure, chain, linked chain, subtree and showed linked chains to be the most suitable for information extraction tasks with relatively high precision and recall. We can use Stanford[1] or Minipar [12] parsers to obtain the dependency or parse trees more readily. However, it would require more expensive pipeline processing and annotating these different patterns correctly would require additional work. Thus, this can be attempted only when we rank the document highly.

5. CONCLUSION

The project has implemented a wide range of B2B market search scenarios. To cover the client bases from insurance companies, banks, construction material supplier to postal delivery services, it has looked at heterogeneous data sources of semi or unstructured data, covering different domains. With a relatively simple user interface, it has exposed high level parameters such as concepts to be extracted, extraction schedule, start URL, search engine queries, while abstracting the lower level parameter such as crawler configuration for directory websites. The project has brought the latest advances in web extraction and text mining technology to timely supplement the semantic content already held in backend triple stores and enabled the user to query and inference this data for his marketing campaign. It enables him to identify and target the desired market segment, consisting of organizations who have the required need and behavior for the propensity to buy his products.

6. REFERENCES

[1] McDowell, L. K. and Cafarella, M. 2006 Ontology-driven Information Extraction with OntoSyphon.

[2] Burcu, M. S. Y. 2007. ontoX - A Method for Ontology-Driven Information Extraction. In Computational Science and Its Applications (ICCSA 2007), LNCS 4707, Springer-Verlag, 2007, S. 660 - 673.

[3] Schutz, A. and Buitelaar, P. 2005. RelExt: A Tool for Relation Extraction from Text in Ontology Extension.

[4] Miller, R. C. and Bharat, K. 1998. Sphinx: A framework for creating personal,site-specic web crawlers. Computer Networks, 30(1-7):119-130.

[5] Ehrig, M. and Maedche, A. Ontology-focused Crawling of Web Documents.

[6] Levenshtein, V. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, Translated from Doklady Akademii Nauk SSSR.

[7] Lin, D. 1998. An Information-Theoretic Definition of Similarity. In Proceedings of the 15th International Conference on Machine Learning.

[8] Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In Procs. of the 14th International Conference on Computational Linguistics, pages 539–545.

[9] Banko, M. and Etzioni, O. The Tradeoffs Between Open and Traditional Relation Extraction.

[10] Stevenson, M. and Greenwood, M. A. 2006. Learning Information Extraction Patterns using WordNet. In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006 22 - 28 May 2006, Vol. 2006 (2006), pp. 95-102.

[11] Stevenson, M. and Greenwood, M. A. 2007. Comparing Information Extraction Pattern Models. In Proceedings of the Information Extraction Beyond The DocumentWorkshop (COLING/ACL 2006), pages 12-19, Sydney, Australia.

[12] Latat, D. L. 2001. Language and text analysis tools.

[1]The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml