IEEE --- 2005 International Conference on Emerging Technologies
September 17-18, Islamabad

Information Extraction from Heterogeneous Sources Using Domain Ontologies

Waqas-ur-Rehman Chaudhry, Farid Meziane
School of Computing Science and Engineering
University of Salford
Greater Manchester, M5 4WT
United Kingdom
chwaqas@hotmail.com, F.Meziane@salford.ac.uk

Abstract

The main objective of this paper is to describe the KAARE (Knowledge Availability, Access, Retrieval and Extraction) system, a generic business model for knowledge extraction of semi structured and unstructured data from web pages. The system is ontology driven and provides a set of generic tools that will enable an effective access, retrieval and filtering of information available on the World Wide Web. The interactive model is composed of five managers namely the Query Manager, the Ontology Manager, the Search Manager, the Information Manager, and the Presentation Manager. Each manager is responsible for carrying out the delegated tasks from which valid inferences can be made.

Index Terms: Information Extraction, Ontologies, Semantic web.

1. Introduction

The WWW serves a huge, widely distributed and diverse community of users. Currently, there are more than 16 millions web sites hosting more than 3 billions web pages [1] covering nearly each and every topic of life with a constant and rapid increase. The WWW and its associated information services are now undeniably the richest source of information.

The diversity and density of the WWW has created a significant data extraction problem as its present structure makes it difficult to make use of that information in a systematic way. The WWW is managed by many individuals, companies, and adhere to various standards and formats, mainly by using HTML, which with no doubts provides a convenient and nice way to present information to humans, but imposing a real challenge for automatic extraction of relevant knowledge and information with respect to a service or application. These and many other such related factors contribute to the difficulty of extracting demanded Web data and hence limiting its success factor. Hence, Knowledge discovery from web resources is becoming a priority for many researchers and industries. The most common way currently in use for searching and retrieving information from the WWW is based on keywords search or similarity based search using one or more search engines, and then in order to extract relevant pieces the user has to browse the large number of returned URLs. Moreover, these approaches can encounter many major difficulties including synonymy and polysemy problems.

The remaining of the paper is organised as follows. Section 2 summarises the benefits and vision of the semantic web. Section 3 presents our approach and in section 4 we present the generic algorithm used for the implementation of our system and in section 5 we present a case study. Section 6 looks at some related work and in section 7 we evaluate our system and draw some conclusions on the first prototype.

2. The semantic web

The WWW contains a huge collection of documents, which are read, understood, and processed mainly by humans and its current structure is not machine friendly. The amount of electronic information keeps on growing and the internet users

are facing the information overload paradox and existing tools and techniques do not provide adequate relief from this problem. Moreover, they are not able to exploit the semantic content of these information sources, so it can be hard at times to find out the meaningful relationships between different pieces of information.

These and many other such similar problems are the bottlenecks for the future growth and utilisation of the web, and in order to overcome them, web contents should be processed by computers if we want to achieve the vision of the semantic web which aims at providing an information enriched with machine processable semantics. This will allow various intelligent services to understand the information and to perform knowledge level information transformation, search, retrieval and extraction [5].

Ontologies are in no doubt the most important form of knowledge representation currently in use for the Semantic Web. In order to overcome the problems caused by present search and retrieval techniques to access information, ontologies are providing the ways to retrieve and extract information based on the actual content of a page and help navigate the information space based on semantic concepts [9, 10]. Tools like ontologies facilitate access to and description of the content of documents and are an important step towards offering efficient resource discovery [6 ,13] on the Web. They can be generic like WordNet[1] or can be domain dependant covering the concepts related to a particular domain e.g. academic ontology covering the concepts related to academia. The proposed knowledge extraction tool (KAARE), will be populated with a domain ontology, which will provide concepts related to a domain of interest, in order to disambiguate word sense, automatic query expansion and for efficient information retrieval and extraction.

3. Our approach

In order to overcome the shortcomings of keyword based search in response to information selection requests, we have proposed a generic ontology based information extraction tool. The model presented here constitutes a suitable basis for building an effective solution to extracting unstructured information from the WWW by providing an extensible architecture and will provide fast and accurate selective access to this information; performing selective dissemination of relevant documents depending on filtering criteria.

The KAARE project aims at providing a set of integrated software components for accessing heterogeneous data sources and extracts the required information based on domain dependent ontologies. However, ontology selection will be based on the search criteria and the information users are requesting.

3.1. The proposed system

A scenario for a practical implementation of the KAARE system is to allow end users easy access to specific knowledge bases. The system is composed of the following steps:

Step 1: Capturing requirements
Step2: Targeting external resources of information corresponding to the requirements.
Step 3: Collecting information
Step 4: Structuring information
Step 5: Presenting the information to the end user
Step 6: Getting feedback from end user to define the whole process as iterative

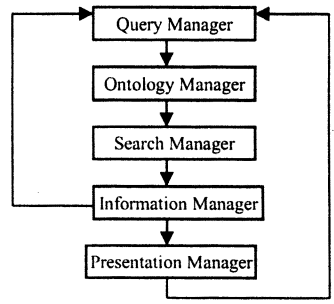

Figure 1. The process/business model.

To validate this model, we used the domain of information extraction for the educational sector, a wide and disparate unstructured or semi structured information source, which is having a wide variety of worldwide end users. The overall iterative model is designed as a set of five managers and is summarised in Figure 1. The detailed architecture of the KAARE system is described in Figure 2.

3.1.1. Query manager

The query manager handles the user's query. This is used as an input to the system and can be based on keywords, index terms and/or natural language. The query manager performs stop word removal,

[1] http://wordnet.princeton.edu/

stemming, and query expansion using domain ontology.

3.1.2. Ontology manager

Based on the keywords present in the user's query the ontology manager will select, load, and process the appropriate ontology to expand the query. Later the ontology will be used by the search and presentation managers to carry out further tasks.

3.1.3. Search manager

Through a series of queries to search engine(s) the search manager attempts to locate target resources, which corresponds to the user requirements i.e. identification and retrieval of related documents/links possibly containing the information to be extracted.

3.1.4. Information manager

The information manager extracts relevant information from trusted target resources obtained through the search manager. Rules have been devised for each group of queries.

3.1.5. Presentation manager

The presentation manager applies different information presentation techniques e.g. filtering, sorting, and cleaning the extracted information obtained through information manager before presenting it to the potential end user.

4. KAARE algorithm

The detailed algorithm and the different steps involved in the KAARE system are as follows:

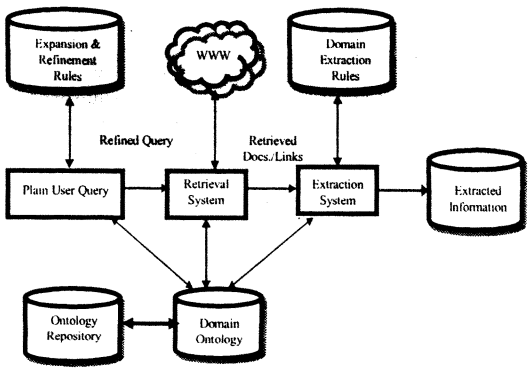

Figure 2. Overall architecture of the KAARE system

4.1. Domain ontology selection and its manipulation tools

This phase involves the development and/or selection of the ontology for a specific application domain. In our case new ontologies can be deployed as well as those already published provided they comply with the OWL language syntax, which is the W3C recommendation for describing ontologies and the one adopted for our system.

A domain ontology is automatically selected based on the structure of the query, but users have the choice to select a different one if they wish to. In order to process the selected ontology from within the system, Jena 2.1[2] API is being used because of its comprehensiveness for the manipulation of ontologies, especially those expressed in OWL.

4.2. The query refinement process

This phase is concerned with the processing of the user's query using domain ontology along with the formulation, expansion and refinement rules. The query can be expressed as a natural language sentence or as a list of keywords. The refined query will be passed to the search engine to retrieve the relevant documents possibly containing the required information. Formulation and expansion of the query is not a trivial task in information retrieval [3], as both factors contribute well in recall and precision. Moreover, the shorter the query is the poorer the recall and precision are.

In addition, a little syntactic variation in the keywords like plural, gerund forms and suffixes can cause problems in retrieving the correct documents. Stemming is a technique which can be used, in order to partially overcome these syntactic variations in the keywords by replacing keywords with their stems after removal of the affixes (i.e. suffix and prefix), for example, connect is the stem for the variants: connected, connecting, connection, and connections, and Author for author's, authors, authorship etc. In our proposed model we used the "porter algorithm" [14] for stemming. Similarly in order to remove the stop words from within the query SMART English stop list[3] is used. We then check for the synonyms and the similarity factor of the terms present in the query with

[2] Jena – A Semantic Web Framework for Java, Available: http://jena.sourceforge.net/

[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

the domain ontology being loaded/selected using equivalent classes, class hierarchy and relationships in the ontology using super classes and sub classes etc., so that the user's query can be refined and expanded in order to pass the refined query to the search engine interface.

4.3. Search engine interface

This phase involves the development of a module to be used as an interface to the search engine (Google[4] in this case). The user will type a query, after applying query formulation, expansion and refinement rules based upon the domain ontology, the resulted refined query is then passed to the search engine, in order to obtain the documents related to the query with respect to the domain ontology. The documents/links that are most related to the query are then downloaded and ranked accordingly.

The results obtained will be passed to the extraction system and used as the basis for the information extraction process.

4.4. Domain extraction rules development

This phase concerns the definition of the extraction rules to be used when analysing the returned websites. This will mainly concern the structure of the web pages and the information they contain. The detail of the work performed by this module is as follows:

4.4.1. Page segmentation and tokenization

Once the documents/links are obtained, the contents of the documents are downloaded and read character by character. Furthermore tokens are being created based on variable criteria for further processing depending on the IE task. It then checks for the existence of the information to be extracted, if the information is present then it will return the extracted information otherwise the system will download all the links in the current page and then after performing filtration e.g. removal of duplicate links etc. it checks for the existence of the information to be extracted.

4.4.2. Knowledge extraction

Since the system selects the ontology to be used, the user has to select the information he wants to extract or define his own. Once the information is defined, each "text line" is parsed character by character to extract the required units of knowledge using domain ontology and extraction rules. Due to the variety of information to be extracted and type of the results to be obtained we have to create different extraction rules for different information extraction related tasks to perform the extraction task successfully.

Similarly when trying to extract the contact information for example which may include telephone number, fax number, email address, etc., typically the required information can be identified and located by the presence of a keyword followed, preceded or sometimes both by a value, where for instance in case of telephone number each value may be followed or preceded by the words like Telephone Number, Telephone Number:, Telephone, Telephone: etc. and value in its simplest form is a string of digits with or without country code, sometimes enclosed within "( )" if present, with varying length depending upon the particular format of a country. The overall extraction algorithm [12] is given in Figure 3

4.4.3. Information presentation

Upon the basis of the results obtained from the search engine interface, in order to extract the required values the system navigate through the sites and their web pages are being checked/matched for presence of the required values to be extracted. Depending upon the user's query, and hence the values to be extracted, the extraction method can be very simple or complicated. Moreover the extraction system, which consists of extraction rules and domain ontology will be recursively called to find the required values from both within the sites and from all other sites returned by the search engine. Once the information manager extracts the required information, the extracted components will then be passed to the presentation manager for presentation in a format suitable to the user.

```
extract(page, existence-item)
 {
 apply extraction rules to page
 if existence item found
   return item
 else
   {
   extract all links in the current
page
   index all links in current page
   let S be the set semantic links
   for each link ks in S
      extract(ks, existence-item)
   }
   return "not found"
 }
```

Figure 3. Extraction algorithm

[4] www.google.com

5. Case study

After the removal of stop words, syntactic variations, it is possible that a particular concept is having 1:m (one to many) association/relationship as compared to 1:1 (one to one) not only in a particular section of the ontology but also in other sections as well e.g. The keyword "thesis" is associated with two concepts in the university ontology, firstly with master level students and secondly with PhD not only in one department but it may relate to other departments of the university in an academic domain.

So there is a further need to disambiguate the concept. In our case we disambiguate by using the interrelationship between the set of keywords occurring together in a particular query to determine the most appropriate sense/context based upon matching terms from their list of synonyms in ontology, and then once a particular region of interest is determined, the closeness of a concept associated with a particular region of interest, is determined by using either depth first search or breadth first search.
We are currently working on the prototype, and in order to demonstrate the effectiveness of the proposed model, we have selected the academic domain for the prototype model. Different sample queries in English are being tested upon selected academic sites. We show here in table 1 and table 2 the performance of the system in terms of precision with regards to the information actually extracted by the system, based on a set of sample queries tested on a set of 17 universities from the UK, US and Australia. In the first query, we have tried to extract general information from the top level of the sites, and the second sample query presented here will extract the detailed level information from within deeper levels.

Query 1: Extract university address, telephone number, and fax number.

Table 1.. System precision for query 1.

	Actual	Extracted	Precision
University Address	13	8	61%
University Phone No.	17	14	82%
University Fax No.	12	8	66%

Query 2: Extract all faculties/schools and schools/departments in a university alongwith their addresses, telephone numbers, and fax numbers.

Table 2.. System precision for query 2

	Actual	Extracted	Precision
Name of Faculties in University	17	13	76%
Name of Schools in University	17	12	70%
Faculties / Schools Address	13	7	54%
Faculties / Schools Phone No.	17	14	82%
Faculties / Schools Fax No.	12	8	66%

We have chosen academic domain because of its broadness and the heterogeneity in the style with different versions of information, usually spread over several domain names, containing unstructured text with a variety of data present in all possible data types and international formats. Moreover there is no knowledge of search path and depth in advance.

6. Related work

Over the past decade due to the growth of electronic data, cost-effective ways of finding the relevant information and extracting useful information from them are increasingly important, and hence several techniques are currently being used to alleviate this problem. Khan [12] proposed the ideas of conceptual distance to retrieve audio data using a query expansion mechanism in the domain of sports. In terms of techniques used to access the information this work is most related to ours but in a different domain.

OntoSeek [13] proposed the use of ontologies for better recall and precision in narrow domain like product catalogues. Similarly MnM [7] is a Semantic Annotation Tool for extraction of knowledge structures from web pages by integrating (Marmot, Badger and Crystal, from the University of Massachusetts at Amherst), This system incorporates the functionalities of other systems like WebOnto to provide browser interface for manipulating ontologies, Marmot (a NLP tool for text analysis), Badger (to analyze text and to produce case frame instantiations depending upon domain specific guidelines), and Crystal (for learning rules). The main purpose is to integrate a template-driven IE engine using a domain specific ontology to

supply the necessary semantic content, to disambiguate extracted information.

7. Conclusion

The overall aim is to specify, and deploy an open framework based on ontologies i.e. capturing information from different sources across all the different knowledge layers, structuring, filtering, and finally providing users with access to specific knowledge. Moreover, it proposes integration of technologies allowing development of an innovative solution as well as exploring and validating enabling technologies for the emergence of the Semantic Web.

The proposed model effectively enables users to access, retrieve and filter information from the web, relevant to their interests and needs, matching their quality expectations with less human intervention. The process is interactive in order to integrate feedback and fulfil specific user requirements. The model will take into account different types of semi-structured, and unstructured documents from a wide variety of sources, provided these documents correspond to the given application domain.

In this paper, we have presented the results of the first prototype using some simple queries, although the precision is not high at this stage because of the heterogeneity and varying style of the contents presented, but we are working on it and we believe it can be improved with a larger sample size as more rules will be included. In future the system will not only extract general information from the WWW, but will also be able to extract information in response to more complex queries e.g. extract detailed information like name, email address, phone number, address of professor(s) of a particular subject in a particular university and/or in different universities, list publication(s) of a person in a particular area etc.

Moreover, in future implementation our extraction approach will be resilient to changes in source document formats. For example, changes in HTML formatting codes do not affect our ability to extract and structure information from a given Web page. Finally the model will contribute effectively to the emergence of semantic web, by providing methodology, tools and both global and generic solutions.

References:

[1] K.Thirunarayan, On Embedding Machine-Processable Semantics into Documents, Proceedings of the 9th International Conference on Application of Natural Language to Information Systems, pp. 368-373, Salford, 2004.

[2] X. Gao and L. Sterling, Semi Structured Data Extraction from Heterogeneous sources, Internet-based Knowledge Management and Organizational Memories, pp. 83-102. Idea Group Publishing, 2000.

[3] C. H. Wen, J. Nie, J. Ma W. Probabilistic Query expansion using query logs, WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. pp. 325-332 ACM 1-58113-449-5/02/0005.

[4] V. C. Story, V. Sugmaran, A. Burton-Jones, The Role of User Profiles in Context-aware Query Processing for the Semantic Web, Proceedings of the 9th International Conference on the Application of Natural Language to Information Systems, pp. 51-63, Salford, 2004.

[5] R.Ba˜nares-Alcantara et al., An Ontology-Based Knowledge Management Platform, Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03) pp. 177-182, Mexico, 2003.

[6] OntoWeb. IST project IST-2000-29243 OntoWeb, 2002. (http://www.ontoweb.org.)

[7] M. Vargas-Vera, J. Domingue, E. Motta, S. Buckingham Shum and M. Lanzoni, Knowledge Extraction by using an Ontology based annotation tool, in proceedings of the Workshop Knowledge Markup & Semantic Annotation, K-CAP'01, Victoria Canada, October 2001, ACM 089791886/97/05.

[8] M. Gómez, C. Abasolo, and E. Plaza. Domain-independent ontologies for cooperative information agents. Lecture Notes in Artificial Intelligence, 2128:118–129, 2001.

[9] Y. Sure, J. Angele and S. Staab, Guiding Ontology Development by Methodology and Inferencing, Volume 31, Issue 4, pp. 18-23, December 2002, ISSN: 0163-5808.

[10] Xiaohua Hu, et al., Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents. International Conference on Web Intelligence (WI 2004), 20-24 September 2004, Beijing, China, pp. 77-83, ISBN 0-7695-2100-2.

[11] F. Meziane, M.K. Kasiran, Extracting unstructured information from the WWW to support merchant existence in ecommerce, Proceedings of the 8th International Conference on the Application of Natural Language to Information Systems, pp. 175-185, Berg, 2003.

[12] L. Khan, Ontology-based Information Selection, PhD thesis, Department of Computer Science, University of Southern California, 2000.

[13] N. Guariono, C. Masalo and G. Vetere, Ontoseek: Content based access to the web. IEEE Intelligent Systems, 14(3):70-80, May 1999.

[14] M. F. Porter, An Algorithm for Suffix Stripping, Program, 14 (3), pp. 130-137; 1980.