Information Extraction for Prediction: Application for web service for conference alerts

Vandana korde
Department of Computer Engineering Sardar Vallabhbhai National Institute of Technology, Surat

Gite Hanumant R.
Department of Computer Science&IT,
Dr. B.A.M.University, Aurangabad

C. Namrata Mahender
Department of Computer Science&IT,
Dr. B.A.M.University, Aurangabad

ABSTRACT
In the general framework of Knowledge discovery ,data mining techniques are usually dedicated to information extraction from structured database .Text mining techniques ,on other hand are dedicated to information extraction(IE) from unstructured textual data and Natural language Process(NLP) can then see as helpful tool for text mining procedure. In this paper we discussed about our work related to IE and proper structuring of the web news related to conference like name of conference, date, location and area of interest etc. Here we have also emphasised on the major issues while extracting and correlating those information for further processing.

Keyword: Data mining, Text mining, Information Extraction

1. INTRODUCTION
Today's world is full of information. Abundant of information is available just even for a relatively small word search. Examples of such data include email, text, Web pages, newsgroup postings, news articles, call-centre text records, business reports, research papers, and so on. In its raw form, the data has limited value since we can do little with it beyond keyword search. Consequently, over the past two decades, significant efforts have focused on the problem of extracting structured information (e.g., researchers, publications, co-author and advising relation-ships, etc.) from such data. The extracted information is then exploited in search, browsing, querying, and mining.

In recent years, the explosion of unstructured data on the World-Wide Web has generated significant further interests in the above extraction problem, and helped position it as a central research goal in the database, AI, data mining, IR, NLP, and Web communities. [1] An illustrative (but far from exhaustive) list of current projects that address this research goal include: (1) entity matching and approximate joins at AT&T Research, MSR and Stanford, (2) answering structured queries over text at Columbia and UCLA, (3) intelligent email and personal information management (PIM) at CMU, Massachusetts, MIT and Washington, (4) extracting and querying semantic entities/relations at IIT Bombay, CMU, MSR and Washington, (5) data cleaning at MSR, (6) doing OLAP-style analysis using extracted information at IBM Almaden and Wisconsin, (7) standardization efforts at IBM Watson on interfaces for NLP extraction tools, (8) managing unstructured data in bioinformatics at Illinois and Michigan, and (9) Web-based community information management (CIM) at Illinois and Wisconsin.

In our work we concentrate on the extraction of related information for an individual from the conference alerts because every information is not needed by every individual which can further be useful for personalised web services.

2. TEXT MINING AND IE
Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) [2**]** is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. This paper suggests a new framework for text mining based on the integration of *Information Extraction* (IE), and Knowledge Discovery from Databases (KDD), *data mining*.

Text prediction is one of the most widely used techniques to enhance the communication rate in augmentative and alternative communication, as like text mining approach is used to the Prediction of Disease Status from Clinical Discharge Summaries[3],[4]. Prediction from text can be just as ambitious as prediction for numerical data mining. In statistical terms, prediction has a very specific characterization, and it need not deal with just topic assignment to documents. Prediction for text follows the classical lines of all numerical classification problems.

As Information Extraction is one of important application of NLP, its impact while mining may provide with good performance. Because the traditional data mining assumes that the information to be "mined" is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge.

While constructing an IE system is a difficult task, there has been major recent progress in using machine learning methods to help automate the construction of IE systems [5, 6, 7 and 8]. However, the accuracy of current IE systems is limited and therefore an automatically extracted database will inevitably contain significant numbers of errors. So in our work cleaning and structuring the information is our key objective and then correlate the information.

3. PROPOSED APPROACH
Our approach to text mining is motivated by practical applications. However, the design and development of prediction methods, by considering the need of text predictors we are trying to develop a module which gives the some prediction form text document. There are many sources of news on the Web, often taken from newswire services such as Reuters or Associated Press, news of some event according to

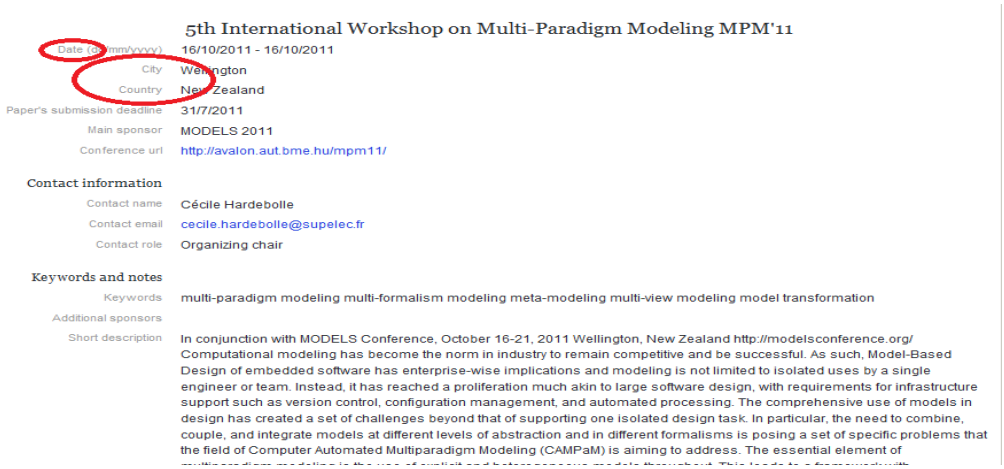
field , that contain valuable information Such information may be used, for instance, to analyse . We considered the web news of conference event; we generally have following details like name of conference, date, location and area of interest etc from any conference alert site. By collecting the old conference data and the current information, we extract certain relation, to find that following predictions can be performed which will be of a great importance in personalized web structures

1. Prediction on the buzz area of research.
2. Provide personalized conference alerts depending on subject interest.
3. To know the proper zone, areas where the research is growing or blooming in full fledge.

The stages we are following for extraction and correlating the unstructured data .

In the first stage we are collecting the conference data from different real-time Web crawler that monitors news and alerts about conference for e.g http://confrancealret.com,http://ourglocal.com,http://infomatrics.in, http://wikicfp.com our work we have considered such for major websites

Second thing general information like subject, name of conference, important dates, country, contact information, organizing committee is there in all such sites but the pattern or the web structure is always varying. So the most important part of our work is that standardized the required data after extracting them from this websites. Figure 1 and 2 shows the variation of formats and structure of data on a conference alert website


5th International Workshop on Multi-Paradigm Modeling MPM'11

Date (dd/mm/yyyy)	16/10/2011 - 16/10/2011
City	Wellington
Country	New Zealand
Paper's submission deadline	31/7/2011
Main sponsor	MODELS 2011
Conference url	http://avalon.aut.bme.hu/mpm11/

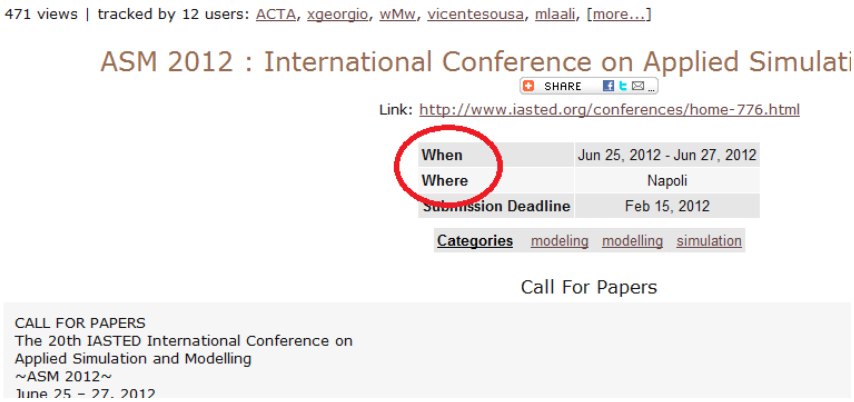
Contact information

Contact name	Cécile Hardebolle
Contact email	cecile.hardebolle@supelec.fr
Contact role	Organizing chair

Keywords and notes

Keywords	multi-paradigm modeling multi-formalism modeling meta-modeling multi-view modeling model transformation
Additional sponsors	
Short description	In conjunction with MODELS Conference, October 16-21, 2011 Wellington, New Zealand http://modelsconference.org/ Computational modeling has become the norm in industry to remain competitive and be successful. As such, Model-Based Design of embedded software has enterprise-wise implications and modeling is not limited to isolated uses by a single engineer or team. Instead, it has reached a proliferation much akin to large software design, with requirements for infrastructure support such as version control, configuration management, and automated processing. The comprehensive use of models in design has created a set of challenges beyond that of supporting one isolated design task. In particular, the need to combine, couple, and integrate models at different levels of abstraction and in different formalisms is posing a set of specific problems that the field of Computer Automated Multiparadigm Modeling (CAMPaM) is aiming to address. The essential element of

Figure 1: http://infomatrics.in


471 views | tracked by 12 users: ACTA, xgeorgio, wMw, vicentesousa, mlaali, [more...]

ASM 2012 : International Conference on Applied Simulati

SHARE

Link: http://www.iasted.org/conferences/home-776.html

When	Jun 25, 2012 - Jun 27, 2012
Where	Napoli
Submission Deadline	Feb 15, 2012

Categories modeling modelling simulation

Call For Papers

CALL FOR PAPERS
The 20th IASTED International Conference on
Applied Simulation and Modelling
~ASM 2012~
June 25 - 27, 2012

Figure 2: http://wikicfp.com

And after extracting make a proper normalized database for further processing.

In our work we are using Stanford open source NLP tool for information extraction , which extracts Entity as name of conference, date of conference, location, website url, subject of conference mainly. Output is in the form tag. Further the required relative information is selected from this document and database is constructed. Figure 3 shows the schema of our database.

this database with existing KDD tools. So we discussed about our work related to IE and proper structuring of the web news related to conference like name of conference, date, location and area of interest.

The database which we have developed from the extracted information will be most useful for predictions and alerts for a personalized web services

Conference Table={Con_name,Date,Venue,Website,organization ,Con_topic_key}

Subject Table= {Subject_id , ,subject_name,subject_keyword}

Topic subject Table= {Topic_id, subject_id, Topic_name}

Topic Combination Table= {T1,T2..Tn, Con_topic_key,}

Figure 3 shows the schema of our database

4. CONCLUSION:

Text mining is a relatively new research area at the intersection of natural-language processing, machine learning, data mining, and information retrieval. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text corpora can be developed. In this paper, we have presented an approach that uses an automatically learned IE system to extract structured databases from a text corpus, and then mines**.**

5. REFERENCES

[1] AnHai Doan, Raghu Ramakrishnan, Shivakumar Vaithyanathan: Managing Information Extraction SIGMOD 2006, Chicago, Illinois, USA. June 27–29, 2006

[2] Haralampos Karanikas and Babis Theodoulidis Manchester, "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK 2001.

[3] Hui Yang, Phd, Irena Spasic, Phd, John A. Keane, Goran Nenadic, Phd" A Text Mining Approach to the Prediction of Disease Status fromClinical Discharge Summaries", Journal of the American Medical Informatics Association Volume 16 Number 4 July / August 2009.

[4] Azadeh Nikfarjam, Ehsan Emadzadeh "Text mining approaches for stock market prediction", IEEE 2010.

[5] M. E. Califf, editor. Papers from the Sixteenth National Conference on Artificial Intelligence Workshop on Machine Learning for Information Extraction, Orlando, FL, 1999.

[6] C. Cardie. Empirical methods in information extraction. AI Magazine, 18(4):65–79, 1997.

[7] F.Ciravegna and N. Kushmerick, editors. Papers from the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, Sept. 2003.

[8] N. Kushmerick, editor. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence Workshop on Adaptive Text Extraction and Mining, Seattle,WA, Aug. 2001