Knowledge Extraction by Internet Monitoring to Enhance Crisis Management

El Hamali Samiha
National School for Computer Science ESI, Algiers, Algeria
s_el_hamali@esi.dz

Nadia Nouali-TAboudjnent, Omar Nouali
Research Center of Scientific and Technique Information CERIST, Algeria
{nnouali, onouali}@cerist.dz

ABSTRACT

This paper presents our work on developing a system for Internet monitoring and *knowledge extraction* from different web documents which contain information about disasters. This system is based on *ontology of the disasters domain* for the knowledge extraction and it presents all the information extracted according to the kind of the disaster defined in the ontology. The system disseminates the information extracted (as a *synthesis of the web documents*) to the users after a *filtering* based on their profiles. The profile of a user is updated automatically by interactively taking into account his feedback.

Keywords

Crisis management, internet monitoring, knowledge extraction, ontology, information filtering.

INTRODUCTION

Several crises and disasters happened in the last decades. After the southern Asian tsunami, the Katrina hurricane in the United States, and the Kashmir earthquake, people throughout the world used the Web as a source of information and a means of communication. In the hours and days immediately following a disaster, a lot of information becomes available on the web. With a single Google search, 17 millions "emergency management" sites and documents have been listed (Min & Peishih, 2008). The user finds a lot of information which comes from different sources, for instance : during the first 3 days of tsunami disaster in south-east Asia, 4000 reports were gathered by the Google News service, and within a month of the incident, 200000 distinct news reports from several online sources (Yiming, et al., 2007 IEEE). The information can be documents of organizations sites, blogs, and wikis, etc. This huge mass of information can create a cognitive overload for the user which disturbs him because the information obtained is not always relevant, and do not have all the same degree of credibility. For a better use of this information in crisis management (by decision makers, organizations, simple citizen), information systems for Internet monitoring can be a helpful solution. The proposal of this paper deals with the design of a system for Internet monitoring to improve disaster management process. The system is based on ontology of disasters domain (Provitolo, Müller, & Dubos-Paillard, 2009). Knowledge is extracted and disseminated for a user profile-based filtering. A brief description of related work is presented, followed by the system architecture and the modules description. A discussion of the system validation, direction of future work and fruitful extensions conclude this paper.

RELATED WORK

A management system of disasters aims to minimize the negative impact of the disaster, so it must take into

Reviewing Statement: This short paper has been fully double-blind peer reviewed for clarity, relevance and significance.

account a lot of challenges like urgency, and quality of information (Yujia, 2008). To deal with these challenges, information must be synthesized in order to not create an overload of knowledge to the user; it must be provided in real or quasi-real time so that it will be useful. The delivered information must be filtered according to a user needs in order to be relevant to its duties. Much of information on the Web is in the form of natural language documents. A promising approach to accessing the knowledge in such documents is centered on IE (Information Extraction) that reduces the documents to tabular structures from which the fragments of documents can be retrieved as answers to queries. However, the time and effort needed for manually annotating a large number of texts and the prerequisite of templates that stipulates which types of information are extractable are major challenges of exploiting such extraction techniques for practical purposes (Yangarber & Grishman, 2001). Many IE systems rely on predefined templates and pattern-based extraction rules or machine learning techniques in order to identify certain entities within text documents, the XAR framework (Ashish & Mehrota, 2008) is an example of these systems which use also rules based on NLP techniques. This framework has been used for disasters management in the RESCUE disaster portal[1] (Yiming, et al., 2007 IEEE) to extract information about a disaster. Other works (Tanev, Piskorski, & Atkinson, 2008) are based on machine learning techniques; (Chapman & Ciravegna, 2006) is based on NLP techniques to extract information from textual documents. Documents on the Web use limitless vocabularies, structures and composition styles for defining approximately the same content. This makes it hard for any IE technique to cover all variations of writing patterns. For example, although content similarity between two disaster documents might be expected, expressions used for both sources may vary significantly. More importantly, traditional IE systems lack the domain knowledge required to pick out relationships between the extracted entities. Ontology is a conceptualization of a domain into a machine readable format; it can represent the entire domain by concepts and relations between them. These observations led us to the use of an ontology of disasters proposed in (Provitolo, Müller, & Dubos-Ribald, 2009) coupled with a general-purpose lexical database (WordNet[2]) and an entity-recognizer (GATE[3]) as guidance tools for identifying knowledge fragments consisting of not just entities, but also the *relations* between them. The other part of this work is the filtering information. The main objective for this component is to support personalization and ranking of information through creation and use of profiles and contexts. In crisis situations, overflow of information can be as bad as no information at all, so filtering and personalization is required. The user profile presents the preferences of the user. In this paper, we use a ranking model based on similarity between the web document and the user profile.

THE SYSTEM ARCHITECTURE

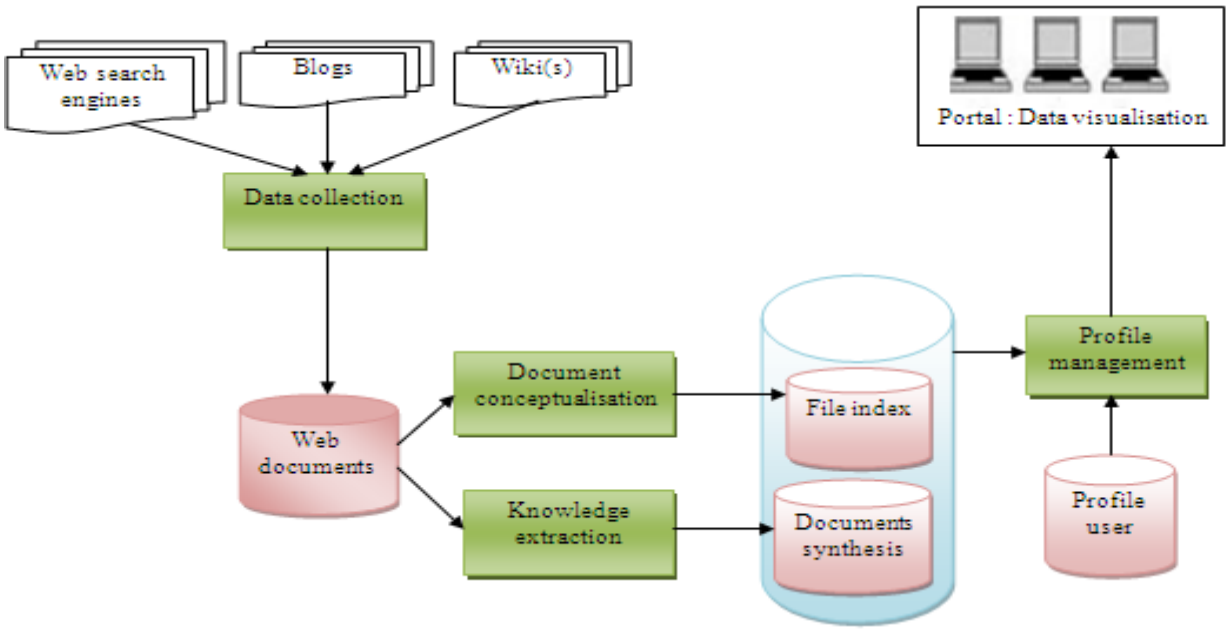

Figure 1 : system architecture

[1] http://www.disasterportal.org/ontario
[2] Available at http://www.cogsci.princeton.edu/~wn
[3] Available at http://gate.ac.uk

Our system is based on 4 modules (figure 1) which are described below.

Data collection from the Web

This module uses Web services to collect information from the various sources of the Web such as search engines, blogs, wikis, etc. The result of this module is a list of URL (S) which will be used thereafter to rebuild the web documents, and to record them in a database of documents for treating the documents without using Internet access, but they will be locally accessible. In the first phase, only the textual content will be taken into account.

Indexing of the collected documents

This module provides a file which is an index of collected documents. This index is a conceptual representation of the database of collected documents where each document is a set of concepts in the ontology of crisis (Provitolo, Müller, & Dubos-Paillard, 2009). This presentation is used to disseminate the results to the user based on similarities between the concepts of the document and the concepts which represent the user profile.

The indexing module is organized according to the following steps :

1) the extraction of terms of a document to build the free language,
2) the use of a stop list to eliminate the words which do not have a semantic,
3) use of lemmatization and stemming methods to present terms with their stem and lemma,
4) use the TF/IDF method to associate a weight for each stem,
5) construct a file of simple index (document is a set of stems with its weight),
6) the ontology of disasters (Provitolo, Müller, & Dubos-Ribald, 2009), is used to find the concepts relative to each term.
7) Use the method of CF.IDF to associate a weight for each concept (resulting from the ontology) in the document. Each document D is represented by a list of concepts, example document D is represented by the concepts earthquake, magnitude, injured.

Knowledge extraction

The extraction of knowledge aims to find all the concepts relative to the disaster (in the ontology) from a document, and for each concept find the properties, attributes and relations with the other concepts in the ontology. For example, an earthquake is a disaster (disaster is a concept in the ontology); which has the following attributes: date, place, magnitude, etc. The concept earthquake has a relationship to the concept damage; damage can concern materials or humans being, etc. We will use the method of (Alani, et al., 2002) to knowledge extraction using the ontology of disasters. Each web document is divided into paragraphs, which are in turn broken down into sentences. Each paragraph is analyzed syntactically and semantically to identify any relevant knowledge to extract. The Apple Pie Parser[4] is used for grouping grammatically related phrases as the result of syntactical analysis. Semantic examination then locates the main components of a given sentence (i.e. 'subject', 'verb', 'object'), and identifies named entities (e.g. 'Haiti' is a 'Place', '12 January 2010' is a date) using GATE and WordNet. GATE is also be used to resolve anaphoric references (personal pronouns).

The following is an example paragraph: "The Haiti earthquake was a catastrophic magnitude 7.0 M_w earthquake. The earthquake occurred on 12 January 2010, and causes 230,000 died and 250,000 damaged residences." The challenge now is to extract binary relationships between any identified pair of entities. Knowledge about the domain specific semantics is now required which can be inferred from the ontology and used to decide which relations are required and expected between the entities in hand. At this stage, our system submits a query to the ontology to obtain such knowledge. In addition, three lexical chains (synonyms, hypernyms, and hyponyms) from WordNet are used in order to reduce the problem of linguistic variation between relations defined in the ontology and the extracted text. Since a relation may have multiple entries in WordNet (polysemous words), the mapping between a relation and an entry in WordNet takes into account syntactic and semantic clues present in a sentence. For example, the relation of *human damage* is mapped into the concept of 'causes' which, according to

[4] http://www.cs.nyu.edu/cs/projects/proteus/app/

WordNet, has five noun senses and two verb senses. The verb sense which is associated to make consequence of a happened event is selected since this term is associated to the concept of earthquake which is a disaster in the ontology. By providing the IE process with direct access to the concepts and relations in the ontology (Provitolo, Müller, & Dubos-Paillard, 2009), our approach bypasses the need for predefining external templates. Annotations provided by GATE and WordNet highlight that "Haiti" is a place, and "12 January 2010" is a date. Relation extraction is determined by the relations defined in the disaster ontology relative to 'earthquake' concept which matches with three attributes (date, place, and magnitude), and the relation 'human damage' and 'material damage'. This paragraph generates three knowledge triples about the earthquake:

- Attributes of the earthquake (concept) : date(12 January 2010), place (Haiti), and magnitude (7.0),
- Relation human damage : 230,000 died,
- Relation material damage : 250,000 residences,

The result of the knowledge extraction is an XML file as shown in figure 2, which will be recorded in a database as a synthesis of the web document.

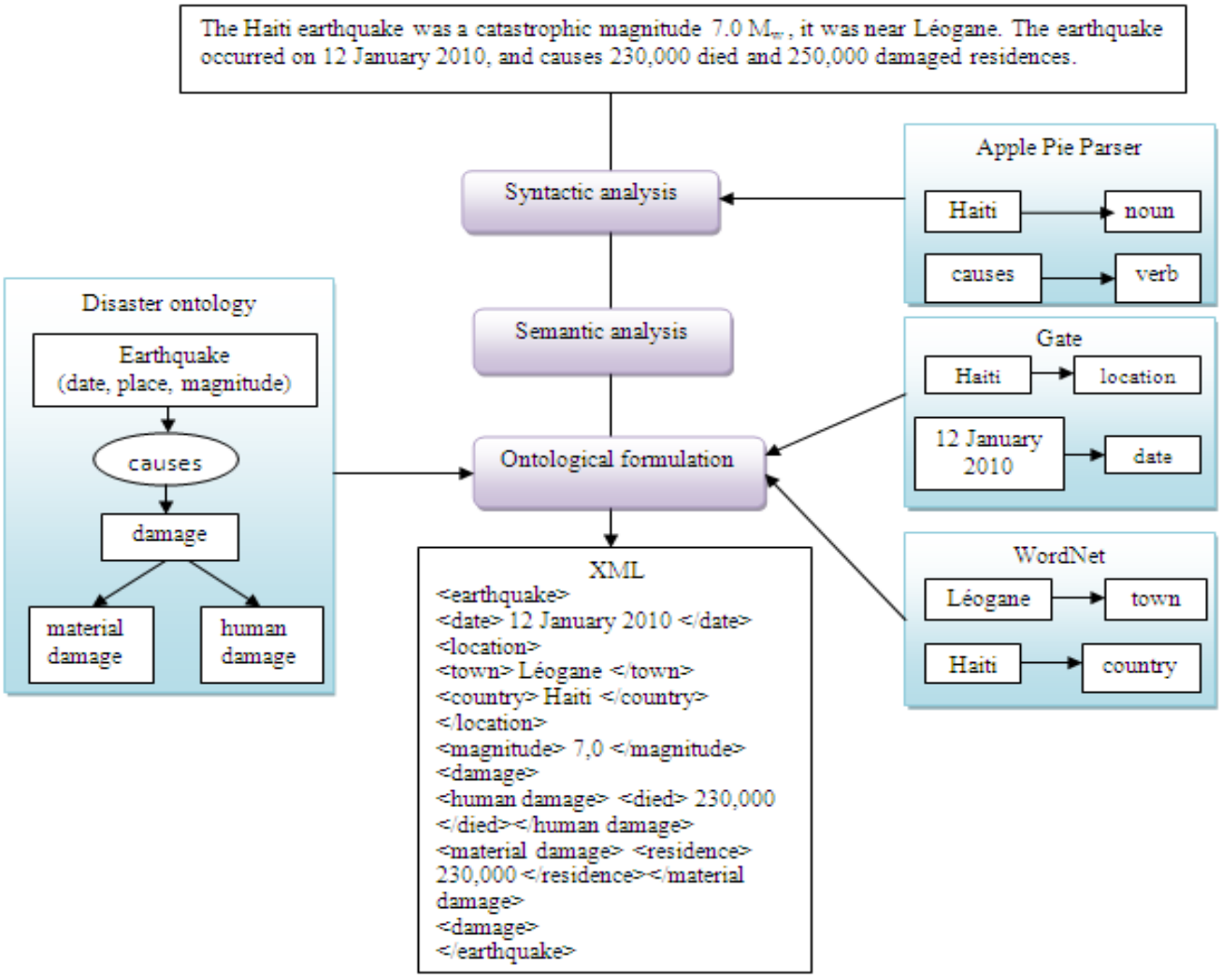

Figure 2: knowledge extraction module

Personalized retrieval using profiles

To achieve personalized information dissemination the system must have flexible and accurate mechanisms to represent and update the information needs from different users. In our system, user characteristics are captured as a user profile. When a new document is available, it will be scored by its similarity to the user profile. If the relevance score is beyond a certain threshold, it will be disseminated to the user, with an order of preference. The user at the first login must specify his preferences; the system uses this method for a first description of the profile using the concepts of the ontology of disasters. The user profile will be updated by analyzing his feedback. A first filtering is done by measuring the similarity between the document and the profile using the

conceptual representation in the index file (similarity measure between the concepts describing the profile, and the concepts in the index file to represent the document). This first filtering chooses potentially relevant documents to be disseminated. A second filtering is based on the similarity between the synthesis of the document (given by the knowledge extraction in the XML file) and the user profile.

CONCLUSION

Our system includes many standard useful capabilities as extracting knowledge from unstructured web documents, personalized retrieval, and user profile updates. The knowledge extraction in our system retrieves metadata triplets (subject – relation – object) from the ontology automatically to overcome the limitations of predefined fixed templates. This knowledge will be disseminated to the user according to his profile to gain time especially in the disasters management. We aim to validate our approach by implementing these modules, and collaborating with crisis management organizations for a practical validation. As perspectives, we plan to use the ontology of disasters with the XAR framework (Ashish & Mehrota, 2008) as a basis of knowledge extraction rules and take into account multimedia documents.

REFERENCES

1. Ashish, N., & Mehrota, S. (2008). *XAR: An Integrated Framework for Semantic Extraction and annotation.*

2. Chapman, S., & Ciravegna, F. (2006). *Focused Data Mining for decision support in Emergency responce Scenarios.* in proceedings ISWC 2006.

3. Goh, O. S., & Fung, C. C. (2005). *Automated Knowledge Extraction from Internet for a Crisis Communication Portal.* Springer-Verlag Berlin Heidelberg 2005.

4. Min, S., & Peishih, C. (2008). *Automatic Extraction of Abbreviation for Emergency Management Websites,Iscram Conference-Washington,USA,May 2008.*

5. Nieuwenhuis, K. (2007). *Information Systems for Crisis Response and management.* DECIS Lab,Delftechpark 24, Delft, The Netherlands.

6. Provitolo, D., Müller, J. P., & Dubos-Paillard, E. (2009). *Vers une ontologie des risques et des catastrophes : le modèle conceptuel.*

7. Provitolo, D., Müller, J.-P., & Dubos-Paillard. (2009). *Validation of an ontology of risk and disaster through a case study of the 1923 Great Kanto Eartquake.* 3rd International Conference on Complex System and Applications, Le Havre, Juillet 2009.

8. Tanev, H., Piskorski, J., & Atkinson, M. (2008). *Real-Time News Event Extraction for Global Crisis Monitoring.* E. Kapetanios, V. Sugumaran, M. Spiliopoulou (Eds.): NLDB 2008, LNCS 5039, pp. 207–218, Springer-Verlag Berlin Heidelberg 2008.

9. Yangarber, R., & Grishman, R. (2001). *Machine Learning of Extraction Patterns from Unannotated Corpora: Position Statement," Proc.* Workshop on Machine Learning for Information Extraction, Berlin, 2001, pp. 76-83.

10. Yiming, M., Dimitri V, K., Ram, H., Sharad, M., Nalini, V., Naween, A., et al. (2007) – IEEE, *On-Demand Informatio Portals for Disaster Situations.*

11. Yujia, C. (2008). *An information assisant system for the prevention of tunnel vision in crisis management,Human media interaction(HMI), University of Twente, thre Netherlands, April 2008.*

12. Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., et al. (2002). *Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from theWeb.* University of Southampton.