PhD Thesis Review: Knowledge Acquisition through Text Mining

by Dragoş Marcel VESPAN

The evolution of internet as a mean for sending information led to the growth of on-line knowledge resources to the diversification of forms and formats used for their storage and transmission: text, data, video and audio. Although hardware restrictions of storage space and data transmission speed is no longer a problem, the text still remains the most efficient form for presenting knowledge over the internet, compared to different audio, video and multimedia formats.

The web was designed as an informational space with the purpose of getting over the limitations of classical communication means and of allowing the communication between humans and machines. The most important barrier in accomplishing this objective is represented by the fact that information and knowledge on the internet is exclusively designed for human consumption. This barrier can be overpassed through using methods and techniques for representing the knowledge from text documents so that it can be automatically acquired and processed by the machines.

The objective of the thesis is represented by the identification and the application of methods and techniques used for knowledge extraction from organizational web resources.

The necessity results from the development of the internet as a mean of knowledge transmission in all society domains, from the huge number of internet users and from the perspective of efficiency increase of the web based communication process required by the development and use of applications which can access knowledge.

The chapter “Text mining – characteristics and application domains” presents the evolution of the internet as a mean of information transmission and focuses on the heterogeneity and diversity of web text documents. The state of the art in text mining is presented with focus on its multidisciplinary character, existing approaches and applicability domains of text mining. Several functional architectures of text mining systems are also described.

The chapter "Document representation" analyses the document as the fundamental element of knowledge acquisition through text mining. Several descriptive representation techniques of documents are presented, focusing on the use of tables as a mean for organizing information. Documents features are defined and document representation techniques in vector space are analyzed.

The chapter “Automated classification of documents” approaches the problem of knowledge acquisition by using categorization algorithms. Main algorithms used in text categorization are identified and analyzed. Measures for the evaluation of text categorization algorithms are defined and compared.

The chapter "Semantic Web – Characteristics and languages" approaches the problem of knowledge representation in the context of Semantic Web. Models of information distribution over the internet are compared and the Resource Description Framework examples are used to represent data from AES web documents. Process models for services of AES are described and ontology development methods through text mining are presented.

The chapter "Web document representation – OntoDev system" presents the personal contributions of the author. An algorithm for web document representation which stores information related their table structure is defined. The representation of ontological concepts on the web resources is analyzed. Log analysis techniques used for the identification of internet user behavior are presented.

The final section of the thesis, “Conclusions”, underlines the importance of web documents based ontology development in knowledge representation. This chapter presents also the original contributions of the author and future research directions of presented theoretical aspects.

Ontologies play a fundamental role in the process, distribution and reuse of web based knowledge applications. Ontologies are used in electronic commerce in order to activate machine based communications between sellers and buyers. Search engines also use ontologies for retrieving web pages that contain words semantically similar but syntactically different. In knowledge acquisition systems, the main issue is the determination of document specific content.

The thesis proposes the use of table structure for web documents and of hyperlink structure of websites in order to develop a domain ontology. Relation between concepts which are described through tables and the hierarchical structure of websites constitute the basis for domain ontology

development. The concepts of such domain ontology can be used for web document categorization. Document classes are then used in log analysis in order to identify internet user behavior.

In order to validate the proposed approaches, the OntoDev system was developed by using Visual Basic 2005 and SQL. For this, the functional architecture and the logical structure of the database were defined. Data sources used were represented by the logs recorded by AES proxy servers and the websites on the following subdomains of *ase.ro*.

The original contributions of the author, presented in the thesis, are:

1. The definition of a theoretical framework for the approach of the theme proposed. The research elaborated in this thesis begins with the study and the analysis of the text mining domain, the definition of txt mining and of the concepts that belong to knowledge acquisition through text mining, text mining systems architectures and its applicability domains.
2. The analysis of main document representation methods – describes the way information is presented in documents both from the point of view of the form and of the content. Document representation is the most important aspect of ontology development. This way, in the thesis there were analyzed different models of document representation and feature (words, terms, concepts) extraction.
3. Definition and implementation of a web document representation algorithm which preserves the table structure characteristic to web documents and which does not alter the information and the knowledge these documents contain. This algorithm has been developed on the basis of the DOM structure of web documents and has been tested on the web pages of the Academy of Economic Studies sites.
4. Analysis of document categorization algorithms identifies common and specific aspects, advantages and disadvantages of applying a certain algorithm and defines their evaluation measures. In order for webpage categorization based on ontology concepts, there were analyzed several document categorization algorithms: decision tree based algorithms, probabilistic algorithms, algorithms based on support vector machines, Rocchio algorithm, perceptron and kNN algorithm. In order to be compared, there were identified several performance measures, like precision and recall, efficiency and utility.
5. Study of ontologies and ontology description languages as a mean for knowledge representation kn Semantic Web. The use of Semantic Web is not restricted only to the access of static documents which provide information about a certain domain, but is extended also to services, which offer useful behaviors. An ontology can be used for website indexing. Each ontology concept is associated with certain web pages where it can be found on. The analysis of AES web resources revealed that there is a connection between the level of concept generality and its representation in document text or in hyperlink text: more general concepts are better represented in hyperlinks and more specific concepts are better represented in textual content of web documents.
6. Identification of models of internet user behavior on the basis of access log analysis and of the association between ontological knowledge and visited websites. On the Internet, the user is the key element and its behavior prediction becomes, this way, fundamental for web mining. Associations between concepts, web documents and paths followed by visitors inside websites represent the basis of internet user behavior models. By processing the web resources of the Academy of Economic Studies, there were identified seven internet user behaviors models.

Future research directions of proposed theme are the definition of automatically web based ontology development, elimination of linguistic boundaries in ontology representation and development of ontology based service retrieval.

Prof. dr. Ion Ivan, PhD