Abstract

The explosive growth of information in textual documents creates great need of techniques for novel, potentially useful and ultimately understandable patterns in data from large text collections. Recently, there are great efforts that are being done for automated discovery of useful or interesting knowledge from unstructured text. Consequently, Information Extraction from Text (IE) is now one of the most promising areas of Knowledge Management (KM) including Knowledge Identification and Extraction. It can – for example - provide support in protocol analysis either in an automatic way (unsupervised extraction of information) or semi-automatic way (e.g. helping knowledge engineers locating the important facts in protocols, via information highlighting).

The knowledge discovery (KD) task addressed by this work is that of extracting concepts from a set of domain specific documents, which is considered an essential step for ontology building.

In this thesis we propose a novel concept extraction method based on the phenomena that "most of class siblings usually have the same features or have similar semantic patterns". Considering that Arabic is lacking the existence of tools as well as necessary resources that could be used in IE, the proposed method succeeds to extract concepts from a raw Arabic text with no use of predefined resources such as lexicon, tokenizes, part of speech taggers, parsers, phrase boundary identifiers, and semantic role labelers.

The main contributions of the thesis include:

- An approach for using learned patterns to automatically extract domain concepts from Arabic free text documents as an essential step to facilitate building domain ontology.

- The proposed approach depends on the input text data without a need for more ANLP resources. The approach makes use of the surrounding text environment of concepts for extracting more sibling concepts in the same domain

To test its generality, the proposed algorithm has been implemented and evaluated, by applying it for extracting Arabic concepts related to two agricultural classes: “agricultural operation”, and “material”. The results demonstrated that the proposed approach is capable of automatically extracting concepts that belong to a defined class with an acceptable degree of accuracy (F-score result was 79% (precision = 73% & recall = 85%).

Contents