

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, tokenizers, 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

1. INTRODUCTION	11
1.1 PROBLEM DEFINITION	13
1.2 THESIS RESEARCH OBJECTIVES AND APPROACHES	14
1.3 THESIS CONTRIBUTIONS	15
1.4 THESIS STRUCTURE	15
2. BACKGROUND ON KNOWLEDGE DISCOVERY AND TEXT MINING	18
2.1 KNOWLEDGE DISCOVERY	18
2.2 THE KNOWLEDGE DISCOVERY PROCESS	20
2.3 DATA MINING TASKS	23
2.4 DATA REPOSITORY	25
2.5 KNOWLEDGE DISCOVERY IN TEXT	27
2.6 PREVIOUS RESEARCH ON KDT AND IE	29
2.7 KNOWLEDGE DISCOVERY IN TEXT USING PATTERNS	33
2.8 CONCLUSION	35
3. A REVIEW OF CONCEPT EXTRACTION METHODS	38
3.1 APPROACHES FOR CONCEPT EXTRACTION	39
3.1.1 STATISTICAL WITH NLP APPROACH	39
3.1.2 NLP APPROACHES.....	40
3.1.3 MACHINE LEARNING APPROACHES.....	43
3.2 CONCLUSION	44
4. THE PROPOSED PATTERN BASED ARABIC CONCEPT EXTRACTION METHOD	46
4.1 THE PROPOSED APPROACH	47
4.1.1 FORMAL DESCRIPTION OF THE PROPOSED APPROACH	48
4.2 ARABIC SEMANTIC CONCEPT EXTRACTOR FRAMEWORK	58
4.2.1 FEATURE EXTRACTION PHASE	59

4.2.2	CONCEPTS EXTRACTION PHASE	63
5.	EVALUATION FOR THE ARABIC SEMANTIC CONCEPT EXTRACTOR METHOD	66
5.1	THE EXPERIMENT'S ELEMENTS:	67
5.2	THE EXPERIMENTS STEPS:.....	71
5.3	THE EXPERIMENTS RESULTS:.....	71
5.3.1	FIRST EXPERIMENT	72
5.3.2	SECOND EXPERIMENT	78
5.3.3	THIRD EXPERIMENT	80
5.3.4	FOURTH EXPERIMENT	81
5.4	EVALUATION SUMMARY.....	83
6.	CONCLUSION AND FUTURE WORK.....	86
6.1	CONCLUSION.....	86
6.2	FUTURE RESEARCH.....	87
	REFERENCES.....	89