ABSTRACT

KNOWLEDGE EXTRACTION USING DISTRIBUTED WEB CRAWLING WITH APPLICATION OF DATA MINING TECHNIQUES

PROBLEM DESCRIPTION

The web is an information repository that contains unstructured data that cannot be used directly by machines. For example, Wikipedia is a huge source of human knowledge, which is often not immediately accessible to computers. Web Knowledge Extraction provides methods and techniques for retrieving knowledge on the Web expressed in natural language.

The aim of the project is to extract knowledge from unstructured texts available on the web and build a knowledge base. The latter requires a complex structuring phase to standardize relations and make the content machine-readable. The text resources are the web pages of the Wikipedia domain, because it includes numerous topics and it is a reliable source.

Systematic analysis of the entire web is called crawling, on the contrary analysing a specific topic is called focused crawling. The project develops a focused crawler, to ensure an accurate analysis of the resources in the domain of interest.

Analysing a domain like the web and ensuring efficient domain coverage is a complex and time-consuming task. This activity can be redistributed on a computer cluster with Multi-agent architecture. The project analyses the procedures for the distribution of the knowledge extraction with particular attention to fault tolerance.

CONTEMPORARY TECHNICAL SCENARIO

In the literature, there is not a complete application that allows the crawling of the web and to extract the information from web pages in a distributed system. The thesis is composed by three main activities:

- Knowledge extraction
- Web Crawler
- Multi-agent protocol.

In the knowledge extraction activity it is necessary to define a procedure to extract the information from a text. The study of a text in natural language is called Natural Language Processing (NLP). It is composed of procedures that allow computers to understand the entities involved in a sentence or the meaning of a text.

Natural language understanding is an AI-complete problem.

The studies of focused web crawlers are many and the two approaches used are: analysis based on link web structure and partial PageRank.

For the communication in distributed mode the literature suggests use of an architecture P2P with a Multi-agent protocol. For communication can be used the gossip protocol.

PERSONAL CONTRIBUTION TO PROBLEM SOLUTION

The goal of the work thesis is to build a mix of techniques from different sectors of computer science. In the project there are many small and significant improvements to the state of art.

The goal is achieved through the following innovations:

▪ A new way to extract information from a text and build a knowledge base using ontology and semantic similarity. The information extracted from the text in the form of triples (subject, relationship and complement) is saved in the Neo4j database. Triples provide unstructured information. Further actions have been planned to make them machine-processable. For this, the most relevant ontology for the analysed text is selected and the properties are used to extract information based on meaning.

▪ Focused web crawler based on semantic features and web structure. The link structure of the web pages is saved in a graph, on which a PageRank is run with a Personalization Vector proportional to the similarity value of the page with the searched topic. Pages with higher scores have a higher priority.

▪ A protocol to coordinate agents independently avoiding silence and delay failure. The system is P2P based on the gossip protocol. There is not a step in which the agents make joint planning. The pages to be analysed are chosen in a probabilistic way.

EXPERIMENTAL AND APPLICATIVE CONTENTS

The solution is validated through different tests.

The extraction of knowledge is tested by comparing the results of the thesis's NER with the results of Dandelion's NER on the same text. In this case the Dandelion NER is a test oracle used to evaluate performance.

To evaluate the approach used in the web crawler: the program runs on a pages domain, where there are a subset of topic pages. The test evaluates how many topic pages are found in relation to the pages visited.

In particular, three indices evaluated are:

- The percentage of topic pages found compared to the topic pages that there are in the dataset.
- Topic pages rate compared to the pages visited.
- Percentage of pages visited to find all topic pages.

Finally, the last test evaluates if there is an effective convenience in the Multi-agent approach, because management work of the Multi-agent protocol can require too much computing power and be inconvenient compared to a configuration with a single agent. The system is tested in two configurations with single agent and Multi-agent and it verifies if an increasing number of agents corresponds to an increase of the visited pages.