Knowledge Extraction from Question and Answer Platforms on the Semantic Web

A systematic review of technologies available for information extraction.

Shayne Weerakoon
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
e-mail: shayneweerakoon@gmail.com

Guhanathan Poravi
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
e-mail: guhanathan.p@iit.ac.lk

***Abstract* — Knowledge Extraction is the process of getting structured data from unstructured or semi-structured sources. Much research has been conducted in this field and applying these technologies to the web has become a key effort in the past few years. This is due to changes from web 1.0 where the web was simply a set of static pages where user interaction was minimal. With the rise of web 2.0, the internet is no longer a medium to access static information. Users can now share their own thoughts easily thus increasing the amount of user generated content. This has made the web ripe with knowledge, however not all information can be easily accessed. This paper aims to bridge the gap between knowledge available and the knowledge accessed using knowledge extraction.**

Keywords - knowledge extraction; information extraction; social knowledge; web 2.0;OBIE;

I. Introduction

Social Knowledge is the collective body of knowledge produced by your immediate community or social circle [15]. This can be from a collective knowledge base as small as a family or it could be a massive body of knowledge such as Wikipedia. One of the best examples of a social knowledge base is Wikipedia, where various members of different societies and cultures share their knowledge.

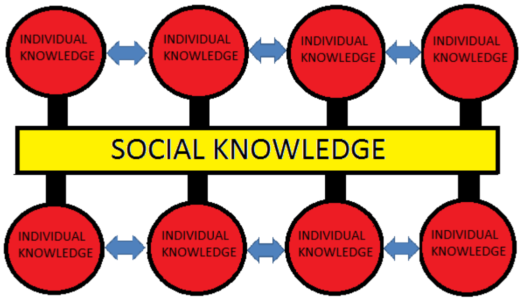

Figure 1 - Social Knowledge[15]

One of the defining characteristics of social knowledge is that it is the result of a group sharing and collaborating knowledge.

The era of Web 2.0 has completely changed how the internet is used. The internet is no longer a place to consume static content. Users cannot create their own content, thus giving rise to a larger number of user generated content(UGC) [3]. Considering the rising use of social media and platforms such as Stack Exchange, the volume of UGC available has never been higher, and is constantly rising, therefore it is possible to state that the volume of social knowledge being shared is high as well.

This can be clearly seen on a platform such as Stack Exchange. Stack Exchange allows specific communities to create separate sites. For examples, electronics and software engineering will each have their own respective sites. This means that users can collaborate knowledge within their own communities. This is one of the defining characteristics of social knowledge, as mentioned previously.

This paper focuses on the extraction of social knowledge from popular Question & Answer(Q&A) platforms and begins by introducing the reader to the domain of knowledge extraction. It then goes on to systematically review the research and technologies available for knowledge extraction tasks before finally proposing a high-level solution.

II. Knowledge Extraction

Before going into knowledge extraction, we must first understand information extraction. Information Extraction (IE), according to Russell and Norvig[14], could be defined as processing of natural language text and to retrieve occurrences of a particular class of objects or events and occurrences of relationships among them. IE automatically extracts structured information from unstructured or structured documents. Most frequently, IE is done by means of natural language processing(NLP) on human readable text. What differs knowledge extraction (KE) from IE is that the information extracted is then stored in a schema, based on the requirements and the type of information being extracted. The two main types of information extraction are as follows:

978-1-5386-6539-8/18/$31.00 ©2018 IEEE
DOI 10.1109/ISMS.2018.00029