Knowledge Mining Wikipedia: An Ontological Approach

Herbert Lee, Marvel Digital Ltd., Hong Kong
Keith C. C. Chan, The Hong Kong Polytechnic University, Hong Kong
Eric Tsui, The Hong Kong Polytechnic University, Hong Kong

ABSTRACT
The organization of information in the knowledge economy has become a primary business process in many enterprises. The better information is organized and stored, the easier it can be retrieved, so that the most relevant information will always be available. Ontology is a versatile technology for organizing information; however, the main obstacle that prevents ontology prevailing is the difficulty in its management, that is, the building and maintaining of ontology. To solve this problem, a structured knowledge source is needed. In this paper, a thorough review on the feasibility of using Wikipedia as the source for ontology management is presented. Different approaches that are based on the use of Wikipedia for such purposes are also discussed. Finally, an ontology management system, based on the use of Wikipedia as the knowledge source, is proposed.

Keywords: Data Mining, Information Organization, Information Retrieval, Ontology, Semantic Search, Wikipedia

THE PROBLEM
We are living in a world where information is critical for better daily decision making. Information is available everywhere. The paramount issue concerning information is therefore not about availability but about findability, and that information has to be made available in the right context. Information is available in abundance and if we can turn information into knowledge, we can increase our competitive advantage. The question we are faced with is how we can turn information into knowledge. One way to tackle the problem is to get information better organized and stored. The easier information can be retrieved, so that the most relevant information will always be available when needed, the more we can exploit the power provided through such information. It is only by properly organizing information that it will be much easier to turn information into knowledge assets. Organizing information in the IT sense is through information categorization. There are a few widely used information categorization schemes: taxonomy, folksonomy and ontology. Each of these schemes has its own advantages and limitations. However, among these schemes, ontology is the most versatile and powerful of all. Its drawback is that it is difficult to build and maintain.

Ontology can be used, among other purposes, to automate the categorization of infor-
mation. Enterprises have started to look into this technology to solve their ever increasing problems associated with the information explosion. However, building enterprise ontology is not a simple task and there is also the subsequent problem of having its maintenance. While larger organizations may allocate resources for the development of ontology, small and medium sized enterprises are definitely not in this position. On the other hand, for enterprises to migrate from small-scale ontology to large scale multiple ontologies, consistency will always be a big issue. Problems such as duplicating data elements, untested ontology elements, and ambiguous definitions will surface (McCready, 2008). There have been many attempts to develop enterprise ontology in the past two decades, however, only a few have been successful (Uschold et al., 1998). The main obstacle to success seems to lie in the issues concerning ontological engineering.

Ontological Engineering refers to the set of activities that is concerned with the ontology development process, the ontology life cycle, the methods and methodologies for building ontologies, and the tool suites and languages that support them (Perez et al., 2004). Many studies concerning ontological engineering emphasize that it is a difficult process (Warren et al., 2006; Fensel, 2007). A typical example that is often cited is the case of Cyc, that was one of the largest ontology projects in the world (Lenat & Guha, 1999). It was to be an upper-level ontology that would model world common sense knowledge. The entire Cyc ontology contained hundreds of thousands of terms, along with millions of assertions relating the terms to each other, forming an upper ontology whose domain is the whole of human consensus reality. The mega project was funded by the E.U. and was initiated in 1983, with the development planned to span over a decade.

Ontology is mostly built by knowledge experts and computer engineers, and it is basically a manual or, at most, a semi-automated process (Cristani & Cuel, 2004; Kovacs et al., 2006). Two paramount issues need to be tackled in order to make ontology pervasive. Firstly, the ontological process must be made simple and inexpensive such that SMEs (small and medium enterprises) can build their ontologies using existing IT staff. Secondly, there has to be a simple way of building universal ontology that can be shared among all users and can be used to power semantic technology (Richards, 2006; Mikroyannidis et al., 2010).

WIKIPEDIA AS A KNOWLEDGE SOURCE

Undoubtedly there is a great barrier to making ontology pervasive. Automated methods in ontological engineering have to be sought, and the technique that is associated with the automated construction of ontology is referred to as ontological learning. It is a sub-domain under information retrieval (IR). Traditional approaches to building ontology fall into two basic categories. One is to rely on carefully hand-crafted human resources, and ontology so built is generally of high quality, but restricted in size and coverage. On the other hand, automated methods using text mining techniques can produce ontology with incredible size and coverage, but at the expense of quality. This is because human language is inconsistent and ambiguous. Approaches based on statistical techniques (data mining) might emulate human intelligence for specific tasks only, and in specific contexts. It might be inappropriate when there is a need to generalize or move into new domains and tasks. Another approach which seems to be quite successful in recent years is to apply text mining techniques on a semi-structured knowledge source. One of the best candidates for this knowledge source is Wikipedia.

Wikipedia, as a knowledge source, provides a middle ground between the two approaches by offering both quality and quantity. It consists of more than 14 million articles from thousands of contributors, is carefully crafted and it keeps on increasing, in terms of both quality and quantity every day. The Wikipedia articles are constantly monitored for quality and accuracy by the mass community of users so as to arrive
Calibration of the Parameters of a Model of an Engineering System Using the Global Optimization Method

Improvement of 2-Partition Entropy Approach Using Type-2 Fuzzy Sets for Image Thresholding
[www.igi-global.com/article/improvement-of-2-partition-entropy-approach-using-type-2-fuzzy-sets-for-image-thresholding/136068?camid=4v1a](www.igi-global.com/article/improvement-of-2-partition-entropy-approach-using-type-2-fuzzy-sets-for-image-thresholding/136068?camid=4v1a)