Semantic Approach to Web-Based Discovery of Unknowns to Enhance Intelligence Gathering

Natalia Danilova, School of Engineering and Mathematical Sciences, City University London, London, UK

David Stupples, School of Engineering and Mathematical Sciences, City University London, London, UK

ABSTRACT

A semantic Web-based search method is introduced that automates the correlation of topic-related content for discovery of hitherto unknown intelligence from disparate and widely diverse Web-sources. This method is in contrast to traditional search methods that are constrained to specific or narrowly defined topics. The method is based on algorithms from Natural Language Processing combined with techniques adapted from grounded theory and Dempster-Shafer theory to significantly enhance the discovery of related Web-sourced intelligence. This paper describes the development of the method by showing the integration of the mathematical models used. Real-world worked examples demonstrate the effectiveness of the method with supporting performance analysis, showing that the quality of the extracted content is significantly enhanced comparing to the traditional Web-search approaches.

Keywords: Grounded Theory, Information Quality, Information Retrieval, Natural Language Processing, Semantic Similarity

INTRODUCTION

The quality of decisions made in business and government correlates directly to the quality of the information used to support these decisions. Most of the information used for intelligence analysis may, in the future, be harvested from the Web as this is fast becoming the richest source. An organisation’s Intranet knowledge base can be efficiently searched using enterprise search systems based upon either semantics such as ontologies, or meaning-based computing. These technologies require comprehensive (and often automatic) clustering, indexing and tagging of the Intranet knowledge base information. The existing Web, as originally envisaged by Berners-Lee (2001), was expected to form into a Semantic Web, that encourages simply the inclusion of semantic content in Web-pages, making it both human readable and machine readable. However, most of the current Web remains poorly semantically tagged, making it impossible to apply enterprise search methods for Web-based information extraction.

DOI: 10.4018/ijirr.2013010102
Traditional information retrieval often violates one of the fundamental laws associated with handling complexity. One of them is the “Law of Requisite Variety” (Ashby, 1956), which states that only variety can master variety, reducing disturbances and promoting harmonious order. When applied to the task of Web-based discovery of unknowns, if the retrieval algorithm does not involve enough variety to deal with the scope and complexity of the Web, its results appear to be attenuated, thus losing information in the process. In effect, we need to identify a search solution for the Web that can handle the vast variety and quantity of information involved and then to filter relevant Web-pages of high quality and discarding information that is either too topically remote or irrelevant information noise.

The main contribution of this paper is that it addresses this variety search issue and presents a novel semantic approach method to Web-based discovery of previously unknown intelligence. It provides a comprehensive theoretical background to the proposed solution together with demonstration of the effectiveness of the method from results of the experiments, showing how the quality of collected information can be significantly enhanced.

The rest of the paper contains research background on Web-based search of unknowns, detailed discussion on the semantic approach, results of real-world experiments, and post-analysis and conclusions.

BACKGROUND ON WEB-BASED SEARCH OF unknowns

If the Web is to be used for improving decision-making, then a more effective search method is needed to collect and correlate the best information.

Government and business decisions are made with varying degrees of certainty. Donald Rumsfeld (2002) stated: “There are ‘known knowns’ – that is things we know we know; there are ‘known unknowns’ – that is some things we know we do not know; but there are also ‘unknown unknowns’ – that is things we don’t know we don’t know.” Effective decision-making requires trusted, focused and relevant information. We should be comfortable with both ‘known knowns’ (KKs) and ‘known unknowns’ (KUs), as these are straightforward to find. The problem being that much of the rich information required for good decisions may be in the category of ‘unknown unknowns’ (UUs). So an important question to ask is how can we find the relevant UUs to enrich the knowledge on the topic, reduce the associated uncertainty and improve decision-making (Figure 1)?

Recent research projects aimed at discovery of UUs focus primarily on ontology-based knowledge acquisition techniques. Lehmann et al. (2007) presented a new user interface ‘DBpedia’ to explore a large ontology-based data set by finding connections between different objects, thus, discovering UUs. The

Figure 1. Discovery of unknowns reduces uncertainty and enhances decision quality
Related Content

Semantic Web, Ontology, and Linked Data
[www.igi-global.com/chapter/semantic-web-ontology-and-linked-data/198543?camid=4v1a](www.igi-global.com/chapter/semantic-web-ontology-and-linked-data/198543?camid=4v1a)

Ontological Random Forests for Image Classification
[www.igi-global.com/article/ontological-random-forests-for-image-classification/132502?camid=4v1a](www.igi-global.com/article/ontological-random-forests-for-image-classification/132502?camid=4v1a)

Analysis of Success Factors of Introducing SAP System for ERP Implementation in Small and Midsize Enterprises in Taiwan
[www.igi-global.com/chapter/analysis-success-factors-introducing-sap/73781?camid=4v1a](www.igi-global.com/chapter/analysis-success-factors-introducing-sap/73781?camid=4v1a)
Nature-Inspired-Based Modified Multi-Objective BB-BC Algorithm to Find Near-OGRs for Optical WDM Systems and Its Performance Comparison