Chapter 85
Document Clustering Using an Ontology-Based Vector Space Model

Ruben Costa
Universidade Nova de Lisboa, UNINOVA, Portugal

Celson Lima
Federal University of Western Pará, Brazil

ABSTRACT

This paper introduces a novel conceptual framework to support the creation of knowledge representations based on enriched Semantic Vectors, using the classical vector space model approach extended with ontological support. One of the primary research challenges addressed here relates to the process of formalization and representation of document contents, where most existing approaches are limited and only take into account the explicit, word-based information in the document. This research explores how traditional knowledge representations can be enriched through incorporation of implicit information derived from the complex relationships (semantic associations) modelled by domain ontologies with the addition of information presented in documents. The relevant achievements pursued by this work are the following: (i) conceptualization of a model that enables the semantic enrichment of knowledge sources supported by domain experts; and (ii) implementation of a proof-of-concept, named SENSE (Semantic Enrichment knowledge Sources).

1 INTRODUCTION

The subject of knowledge representation gained a new dimension with the advent of the computer age. Particularly, with the creation of the World Wide Web, new forms of knowledge representation were needed in order to transmit data from source to recipient in common data formats, and to aid humans to find the information they want in an easily understandable manner. With the evolution of the Semantic Web, knowledge representation techniques moved into the spotlight, aiming at bringing human understanding of the meaning of data to the world of machines. Such techniques create knowledge representa-

DOI: 10.4018/978-1-5225-5191-1.ch085

tions of Knowledge Sources (KSs), whether they are web pages or documents (Figueiras, Costa, Paiva, Jardim-Gonçalves, & Lima, 2012).

Information retrieval (IR) techniques were primarily designed for the access and retrieval of library documents, and more recently web pages. IR is often regarded as synonymous with document retrieval and text retrieval, though many IR systems also retrieve pictures, audio, and other types of non-textual information. The word "document" is used herein to include not just text documents, but any "clump" of information.

People have the ability to understand abstract meanings that are conveyed by natural language. This is why intermediary reference librarians are useful; they can talk to a librarian about his/her information needs and then find the documents that are relevant. The challenge of information retrieval is to mimic this interaction, replacing the librarian with an automated system. This task is difficult because machine comprehension of natural language is generally still an open research problem.

Ontologies are the foundation of both content-based information access and semantic interoperability over the web. Various definitions of what constitutes an ontology have been formulated and have evolved over time. A good description of these can be found in (Corcho, Fernandez-Lopez, & Gomez-Perez, 2003). From the authors' perspective, the best definition that captures the essence of an ontology is the one given by Gruber (Gruber, 1993): "an ontology is a formal, explicit specification of a shared conceptualization". As elaborated in (Studer, Benjamins, & Fensel, 1998): Conceptualization refers to an abstract model of some phenomenon in the world which identifies the relevant concepts of that phenomenon. Explicit means that the types of concepts used and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine processable.

Typically, in human endeavour, shared conceptualizations are defined over a lengthy period of time, based on the shared experience of a group of people, sometimes referred to as a community of practice (Wenger & Snyder, 2000). They will involve the definition and use of abstractions that are designed to capture the important aspects of some practical context in order to support a particular activity or type of activity. As such, a shared conceptualization is a socially constructed model or reality that is distinct from reality and is optimized to support the goals and activities of the community of practice in which it was defined. Communities engaged in different activities are likely to form shared conceptualizations that are quite different views of reality, and make up shared ''world-views'' (Checkland & Scholes, 2000) that provide a basis for highly effective and efficient communications within the respective communities.

In order to understand and formalize the shared worldviews of such communities in the form of ontologies to support the integration of diverse human activities, it is important to consider approaches that derive from an interpretive philosophical standpoint rather than from a positivist, scientific/engineering one (Fitzgerald & Howcroft, 1998). In such an approach, it is important to interpret, accommodate, and model what is, rather than trying to change reality to fit a single model. This inevitably results in different ontologies for different communities, but the challenge then is to find ways to allow those communities to collaborate effectively with one another whilst maintaining their existing, efficient, effective separate worldviews. The implication is that the emphasis must be shifted from developing a standard representation of a single ''reality'', towards providing mechanisms for supporting communication between differing perceptions of reality, focusing our attention on the overlaps at the boundaries and the specific conceptualizations that are required for such communication to happen.

With respect to the work reported here, it is proposed to use knowledge available in domain ontologies in order to support the process of representing knowledge sources (e.g. project reports, meeting minutes, descriptions of problems/solutions) thus improving the classification of such knowledge sources. A