Chapter 14
Text Semantic Mining Model Based on the Algebra of Human Concept Learning

Jun Zhang
Shanghai University, China

Xiangfeng Luo
Shanghai University, China

Xiang He
Shanghai University, China

Chuanliang Cai
Shanghai University, China

ABSTRACT

Dealing with the large-scale text knowledge on the Web has become increasingly important with the development of the Web, yet it confronts with several challenges, one of which is to find out as much semantics as possible to represent text knowledge. As the text semantic mining process is also the knowledge representation process of text, this paper proposes a text knowledge representation model called text semantic mining model (TSMM) based on the algebra of human concept learning, which both carries rich semantics and is constructed automatically with a lower complexity. Herein, the algebra of human concept learning is introduced, which enables TSMM containing rich semantics. Then the formalization and the construction process of TSMM are discussed. Moreover, three types of reasoning rules based on TSMM are proposed. Lastly, experiments and the comparison with current text representation models show that the given model performs better than others.

DOI: 10.4018/978-1-4666-2476-4.ch014

INTRODUCTION

With the rapid growth of the Web, how to represent and organize the large-scale texts have drawn a lot of attentions. One of the most important works on text knowledge representation is to find out the semantics in texts. Plenty of scholars focus on many kinds of models that are used to represent text knowledge through various text analyzing methods. Such models are always expected to contain rich semantics, to obtain a robust reasoning ability and to be automatically constructed.

Currently, models referring to represent text knowledge can be mainly divided into four types. (1) Statistics models, which are generated by statistical methods. The typical ones include vector space model (VSM) (Salton & Wong, 1975) and latent semantic analysis (LSA) (Landauer & Foltz, 1998). VSM uses some words extracted from a text and their weights to represent the text semantics, but it doesn't take the relations between the words into account. Thus VSM is only able to express a little semantics in the text while much more semantics has been lost. On the contrary, the LSA model carries more semantics than the former one but its complexity is high because the construction of LSA is involved with the operation of singular value decomposition, whose complexity goes very high. (2) Cognition based models, whose basic idea is inspired by cognitive theories. Element fuzzy cognitive map (EFCM) (Luo & Xu, 2008) is one of the typical models. It obtains more semantics than VSM and a lower computation complexity than LSA. Meanwhile, it can be applied to large-scale text collections since it is constructed automatically. (3) Probability topic models, such as author-topic model (ATM) (Michal & Thomas, 2004), author-recipient-topic model (ART) (McCallum & Corrada-Emmanuel, 2004) and correlated topic models (CTM) (Blei & Lafferty, 2006). These models always need a lot of complex computations, which make probability topic models unsuitable to be used in large-scale text collections. (4) Ontology based models, which are based on ontology languages and most of them are semi-automatically constructed. Ontology inference layer (OIL) (Horrocks & Fensel, 2000), web ontology language (OWL) (McGuinness & Harmelen, 2004) and simple html ontology extensions (SHOE) (Heflin & Hendler, 1999) are typical ontology based models. Since possessing a lot of semantics, ontology based models attracts plenty of researches on them. However, they can only be applied to special areas that contain a lot of human experiential knowledge, as the generation of ontology based models needs a mass of manual work. Thus, up to now, ontology based models still cannot be applied to automatically process large-scale text collections.

Consequently, according to the discussions above, we can see that some models are carrying abundant semantics but cannot be constructed automatically (e.g. OWL); some ones are both allowed to be automatically established and carrying a lot of semantics but still can't be applied to large-scale collections for their high complexities (e.g. CTM and ATM); some ones can be set up automatically with a lower complexity but carry little semantics (e.g. VSM). As a result, through the analysis of those models, we consider that a good text knowledge representation model should satisfy the two conditions listed below.

1. Contain rich text semantics;
2. Construct automatically with a lower complexity;

According to the two conditions, this paper proposes a text knowledge representation model called text semantic mining model (TSMM) based on the algebra of human concept learning. According to Cognitive Informatics (Wang, 2002, 2007a), a concept is defined as a cognitive unit to identify and/or model a real-world concrete entity and a perceived-world abstract object whereas the formal treatment of concepts and a new mathematical structure known are defined as Concept Algebra (Wang, 2006). Moreover, in