Fuzzy Learning of Co-Similarities from Large-Scale Documents

Sonia Alouane-Ksouri, Ecole Nationale d’Ingénieurs de Tunis, Université de Tunis El Manar, Tunis, Tunisia

Minyar Sassi Hidri, Ecole Nationale d’Ingénieurs de Tunis, Université de Tunis El Manar, Tunis, Tunisia

ABSTRACT

To analyze and explore large textual corpus, we are generally limited by the available main memory. This may lead to a proliferation of processor load due to greedy computing. The authors propose to deal with this problem to compute co-similarities from large-scale documents. The authors propose to enhance co-similarity learning by upstream and downstream parallel computing. The first deploys the fuzzy linear model in a Grid environment. The second deals with multi-view datasets while introducing different architectures by using several instances of a fuzzy triadic similarity algorithm.

Keywords: Data Grid, Document Co-similarity, Fuzzy Linear Model, Large-scale Environments

INTRODUCTION

With the increasing number of available documents, the processing efficiency and scalability of the systems and their underlying computations becomes a major concern. For economic and operational reasons it is often preferable not to execute the computations on a single machine. One of the major problems is the similarity computing due to this huge data. Several methods dealing with this task are referred to as co-clustering approaches and have been extensively studied. In (Hussein et al., 2010), a co-similarity measure has been proposed, called X-Sim (Hussein et al., 2010, Bisson & Hussain, 2008) which builds on the idea of iteratively generating the similarity matrices between documents and words. This measure works well for unsupervised document clustering.

However, in recent researches, the sentence has been considered as a more informative feature term for improving the effectiveness of document clustering (Chim & Deng, 2008; Torres-Moreno & Meunier, 2001). While considering three levels documents, sentences and words to represent the data set, we are able to deal with a dependency between them. It is based on weights computing

DOI: 10.4018/ijfsa.2015100104
based on statistical models (Salton & Buckley, 1988). But it has spawned the view that classical probability theory is unable to deal with uncertainties in natural language and machine learning.

We proceed to a fuzzification control process which converts crisp similarities to fuzzy ones (Zadeh 1965). The conversion to fuzzy values is represented by the membership functions (Kundue et al., 1997; Deng-Feng & Chun-Tian, 2002; Deng-Feng, 2004; Deng-Feng, 2014). These fuzzy similarity matrices are used to compute fuzzy co-similarity between documents, sentences, and words in a triadic computing called FT-Sim (Fuzzy Triadic Similarity) (Alouane-Ksouri et al., 2013a). Several extensions to the co-clustering methods have been proposed to deal with such multi-view data. Some works aim at combining multiple similarity matrices to perform a given learning task (Bisson & Grimal, 2012; Tang et al., 2009; de Carvalho & Melo, 2012), the idea being to build clusters from multiple similarity matrices computed along different views.

Following the development of the Web and the availability of storage spaces, documents become increasingly accessible. Data can be provided from several sources (multi-source) and can be seen as a set of matrices (multi-view). By treating separately these views, we risk a huge loss of information. When analyzing large text corpora, the amount of documents to be processed on a single machine is generally limited by the available main memory. The increase of this latter leads to a proliferation of processor load.

To overcome these limitations, a parallel solution adapted to FT-Sim model, called PFT-Sim (Parallel FT-Sim), have been proposed (Alouane-Ksouri et al., 2013b). It offers a parallel strategy for data analysis and divides the computing task and data processing into sub-tasks. We use here the term computing topology to refer to the architecture of parallel and distributed computing.

The rest of the paper is organized as follows: in section 2 we present how to learn fuzzy triadic co-similarity for document co-clustering. In section 3 we present the different solutions to deal with computing co-similarity from large-scale documents. Section 4 concludes the paper and highlights future work.

**TRIADIC DOCUMENT CO-SIMILARITY COMPUTING**

In this work, we focus on the preprocessing step which including data representation and similarity computing. It’s proved that the sentence has been considered as a more informative feature term for improving the effectiveness of document clustering. While considering three levels (documents-sentences-words) to represent the data set, we are able to deal with a dependency between documents-sentences, as also between sentences-words and, by deduction, between documents-words.

The sentence-word-based document similarity (or triadic similarity) considers weighting scheme in computing the document similarity with sentences and the sentence similarity with words. A weighted value may be assigned as a link from a document to a word (or sentence) indicating the presence of the word(sentence) in that document.

Different weighting schemes such as the tf-idf (Salton et al., 1975) may be incorporated to better represent the importance of words in the corpus, but it has spawned the view that classical probability theory is unable to deal with uncertainties in natural language and machine learning (Shu-Ping & Deng-Feng, 2013; Shu-Ping & Deng-Feng, 2014; Deng-Feng & Shu-Ping, 2014). One such alternative is what is known as the fuzzy sets theory (Zadeh 1965). It has proved useful in the context of control theory, pattern recognition, and medical diagnosis. So, we proceed to a fuzzification control process which converts crisp similarities to fuzzy ones. The conversion to fuzzy values is represented by the membership functions. They allow a graphical representa-
Related Content

Soft Computing Paradigms and Regression Trees in Decision Support Systems
www.igi-global.com/chapter/soft-computing-paradigms-regression-trees/24328?camid=4v1a

A Fuzzy Expert System for Star Classification Based on Photometry
www.igi-global.com/chapter/a-fuzzy-expert-system-for-star-classification-based-on-photometry/178411?camid=4v1a

Integrated Network Topological Control and Key Management for Securing Wireless Sensor Networks
www.igi-global.com/article/integrated-network-topological-control-and-key-management-for-securing-wireless-sensor-networks/104158?camid=4v1a
Malicious Application Detection and Classification System for Android Mobiles
www.igi-global.com/article/malicious-application-detection-and-classification-system-for-android-mobiles/190635?camid=4v1a