Grid-Based Fuzzy Processing for Parallel Learning the Document Similarities

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

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

Kamel Barkaoui, CEDRIC-CNAM Paris, Paris, France

ABSTRACT

Document co-clustering methods allow to efficiently capture high-order similarities between objects described by rows and columns of a data matrix. In Alouane et al. (2013), a method for simultaneous computation of similarity matrices between objects (documents or sentences) and between descriptors (sentences or words), each one being built on the other one, according to a fuzzy triadic model based on the three-partite graph. Because of the development of the Web and the high availability of storage spaces, documents become more accessible. This makes the fuzzy computing very expensive. In the present case, the development of fuzzification algorithms of fuzzification requires the integration of a deployment platform with the required processing power. The choice of a grid architecture seems to be an appropriate answer to our needs since it allows us to distribute the processing over all the machines of the platform, thus creating the illusion of a virtual computer able to solve important computing problems which require very long run times in a single machine environment. The authors propose to enhance similarity by upstream and downstream parallel processing. 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: Document Co-Clustering, Downstream Processing, Fuzzy Linear Model, Grid Computing, Three-Partite Graph, Upstream Processing

INTRODUCTION

The joint classification of objects and their descriptors – for example documents (or sentences) with the words that constitute them – also called “co-classification”, have been widely studied these last years. In past work (Alouane et al., 2013), a method for simultaneous calculation of matrices of the similarities between objects and between descriptors, each

DOI: 10.4018/ijssmet.2014010104
one being built on the other one, according to a fuzzy triadic model based on the three-partite graph (Long et al., 2006).

Thus, the fuzzy sets theory, which was demonstrated in several applications (Azar, 2010a, b; Azar, 2012), is used to deal with uncertainty and it has been used to convert crisp similarities to fuzzy ones has been introduced.

The conversion to fuzzy values is represented by the membership functions (Kundu, 1997). They allow a graphical representation of a fuzzy set (Zadeh, 1965). These fuzzy similarity matrices are used to calculate fuzzy similarity between documents, sentences and words in a triadic computing called FT-Sim (Fuzzy Triadic Similarity).

The development of the algorithms of fuzzification which we propose requires the integration of a deployment platform with the required processing power. The choice of the type of platform (grid, cloud) essentially depends on the nature and load of the processing to be realized (Foster et al., 2008). In our case, an architecture of the computing grid - type perfectly answers our needs, as it allows us to distribute the processing over all the machines of the platform, thus creating the illusion of a virtual computer perfectly capable of solving important computing problems which require very long run times in a classic environment with one single machine (Caromel et al., 2008).

Moreover, with the development of the Web and the high availability of the storage spaces, more and more documents become accessible. Data can be provided from multiple sites and can be seen as a collection of matrices. By separately processing these matrices, we get a huge loss of information.

Several extensions to the co-clustering methods have been proposed to deal with such multi-view and distributed datasets. Some works aim at combining multiple similarity matrices to perform a given learning task (Tang et al., 2009; De Carvalho et al., 2012; Bisson et al., 2012a; Bisson et al., 2012b). The idea being to build clusters from multiple similarity matrices computed along different views.

We propose to enhance the proposed similarity by upstream and downstream processing. The first processing deploys the fuzzy linear model in a Grid environment. The second one deals with multi-views datasets while introducing different architectures by using several instances of a fuzzy triadic similarity algorithm. Thus, we consider a model in which datasets are distributed into \( N \) sites (or relation matrices). They describe the connections between documents for each local dataset.

Our goal is then to compute a Documents\( \times \)Documents fuzzy similarity matrix for each site trying to take into account all the representative information expressed in the relations.

To combine multiple occurrences of FT-Sim, we propose waterfall, ring, merging and splitting based parallel architectures.

The rest of the paper is organized as follows: in section 2 we highlight related work in document co-clustering. In section 3 we provide an overview of the fuzzy triadic similarity computation. In section 4, we detail the downstream parallelism based on grid computing. In section 5, we present architectures to deal with upstream parallel computing. Section 6 presents the Middleware layer of our model. Section 7 concludes the paper and gives indications of some future work.

**RELATED WORK**

Most of the existing clustering methods focus on datasets described by a unique data matrix, which can either, be a matrix which describes objects by their characteristics, or a relation matrix that describes the intensity of the relation between instances of two types of objects, such as a Documents \( \times \) Words matrix. In the latter case, both types of objects can be clustered; methods dealing with this task are referred to as co-clustering approaches and have been extensively studied.

However, in many applications, datasets involving more than two types of interacting...
Web Service Enabled Online Laboratory
www.igi-global.com/chapter/web-service-enabled-online-laboratory/44011?camid=4v1a

Approaches for Semantic Web Service: An Overview
www.igi-global.com/article/approaches-for-semantic-web-service/101789?camid=4v1a