A MapReduce Implementation of the Spreading Activation Algorithm for Processing Large Knowledge Bases Based on Semantic Networks

Jorge González Lorenzo, Department of Computer Science, University of Oviedo, Oviedo, Asturias, Spain
José Emilio Labra Gayo, Department of Computer Science, University of Oviedo, Oviedo, Asturias, Spain
José María Álvarez Rodríguez, Department of Computer Science, University of Oviedo, Oviedo, Asturias, Spain

ABSTRACT

The emerging Web of Data as part of the Semantic Web initiative and the sheer mass of information now available make it possible the deployment of new services and applications based on the reuse of existing vocabularies and datasets. A huge amount of this information is published by governments and organizations using semantic web languages and formats such as RDF, implicit graph structures developed using W3C standard languages: RDF-Schema or OWL, but new flexible programming models to process and exploit this data are required. In that sense the use of algorithms such as Spreading Activation is growing in order to find relevant and related information in this new data realm. Nevertheless the efficient exploration of the large knowledge bases has not yet been resolved and that is why new paradigms are emerging to boost the definitive deployment of the Web of Data. This cornerstone is being addressed applying new programming models such as MapReduce in combination with old-fashioned techniques of Document and Information Retrieval. In this paper an implementation of the Spreading Activation technique based on the MapReduce programming model and the problems of applying this paradigm to graph-based structures are introduced. Finally, a concrete experiment with real data is presented to illustrate the algorithm performance and scalability.

Keywords: Algorithm, MapReduce, OWL Web Ontology Language (OWL), Resource Description Framework (RDF), Semantic Networks, Semantic Web Initiative, Web of Data

DOI: 10.4018/jksr.2012100105
INTRODUCTION

The Semantic Web (Berners-Lee, Hendler, & Lassila, 2001) is considered as an extension of the World Wide Web, adding metadata understandable by machines; and the same issues about size addressed in the traditional web are present. The Semantic Web data is expressed using RDF triples, each consisting of a subject, a predicate and an object. A set of such triples is called an RDF graph. If data is modeled through a graph, then we can use graph algorithms to explore this data. One of these algorithms is the spreading activation algorithm (Todorova et al., 2009). But, if we take into account that the whole Semantic Web has billions of triples, it is obvious that the generated graph would be very large to fit into one single machine. Fortunately, this kind of size issues has been solved before in the traditional web by using parallel and distributed computing approaches. One of the most successfully applied framework for parallel and distributed processing is MapReduce (Dean & Ghemawat, 2004). MapReduce has proven to be efficient and is used by Google internally for processing petabyte order datasets. This success has motivated the apparition of the open source initiative Hadoop (http://hadoop.apache.org), which is an Apache project mainly developed and supported by Yahoo.

The main motivation of our research is providing a way of reducing the amount of time needed for processing large RDF datasets. In this article, an implementation of the spreading activation algorithm using the MapReduce programming model is presented. The spreading activation algorithm is used to find out related concepts starting from a set of activated nodes. The algorithm propagates this activation through the graph vertices, and, at the end, the related concepts are the ones with highest activation level. First, we describe the main problems of solving graph problems with MapReduce, and then the decisions and details about the final implementation of the algorithm are shown. Finally, a result section is presented to demonstrate the scalability of the implementation.

PREVIOUS WORK

MapReduce

MapReduce is a framework introduced by Google in 2004 for processing huge datasets using a large number of machines in a parallel and distributed way (Dean & Ghemawat, 2004). MapReduce framework transparently handles system-level details, such as scheduling, fault tolerance or synchronization. The main advantages of the framework is the simplicity of the map and reduce operations, that allow a high degree of parallelism with little overhead, at the cost of writing programs in a way that fits this programming model. MapReduce has proven to be efficient and is used by Google internally for processing petabyte order datasets. This success has motivated the apparition of the open source initiative Hadoop (http://hadoop.apache.org), which is an Apache project mainly developed and supported by Yahoo.

MapReduce handles all the information using tuples of the form <key, value>. Every job consists of two phases: a map phase and a reduce phase. The map phase process the input tuples and produce some others intermediate tuples. Input tuples are divided in groups, each of them processed by a map function running in a single machine. Then, these intermediate tuples are grouped together according to their key value forming a group. Finally, each group is processed by the reduce function, producing a set of output tuples.

Spreading Activation

The spreading activation algorithm (hereafter SA) was introduced as an approach for modeling the human memory and its cognitive processes by following its low-level structure (Todorova et al., 2009). It takes advantage of the similarity between neural networks and graph models, so the same ideas behind the spreading activations mechanisms can be applied to graphs modeled...
Adventure Game Learning Platform
www.igi-global.com/chapter/adventure-game-learning-platform/70094?camid=4v1a