Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 490 – 495

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keyword Prediction with ARM on Bibliographic RDF Data

Nidhi Kushwaha, Bharat Singh, Rajesh Mahule, O P Vyas

Indian Institute of Information Technology, Allhabad, India

Abstract

Web-3.0 provides an easy way to utilize the in-depth knowledge of the huge data that grows day-by-day in the internet Our aim with this paper is to work with the Linked Open Data Cloud data, where the main problem with the dataset is inconsistencies, bulkiness. We are exploring bibliographic data which is one of the cloud data. The authors found some useful information in the dataset that should be explored for judging the improvement of the search query's result. After analysis we came to know that many of the papers residing in RKBExplorer did not have keyword information. Because of that the search engine based on the RKBExplorer only able to use the information in this database going to retrieve the papers, authors of that paper and their related cited papers with given paper author or title. But assume the situation where the user wants to enter the search string, then what would be the result? Would it retrieve all the related paper even if their keywords are not assigned? In this paper we are trying to answer this question, with the help of data mining algorithm ARM on the features retrieved from the RDF data. We have developed a novel approach through which we can answer the user's query which is mixture of important the strings, we called them tags of the papers

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Linked Open Data Cloud; Data Mining; Query-Answering System; RDF; Assoication Rule Mining.

1. Introduction

Linked Open Data (LOD) Cloud Projects [1] had started in the year 2008 by Tim Berners-Lee. It comes with the idea of open source sharing of information on the web, which globally connects the data using unique URIs. Publishing the data gives an opportunity to the universities and researchers to load the data for the internet users in various application domains. The Google Rich snippet and Yahoo Search monkey show the good example of embedding information of RDF data into the less informative XML documents. Today's various applications and browsers support RDF data. This shows its continuous growth and usefulness with the current working environment. Traditional web consists millions of pages connected with each other. But the logic behind the connectivity was missing. This causes the problem to connect the future relatedness of the documents. Ontology development needs a specialized person who has good knowledge of that field. Because of domain

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.019

dependent these developed Ontologies had different connections and concept names [10]. Now the problem was how to combine them, it has been said that use the well known predicates for newly developed Ontology can reduce this problem. One of the way found by Ontology engineers was the use of ontology development tools and the important predicate link known as "owl:sameAS". Linked Open Data Cloud, a continuously growing cloud done this work under some norms defined in [1][10]. The Cloud has various domain information including a cross domain giant "DBpedia" [1]. Many universities and organizations come forward to success the dream of Sir Tim Berners-Lee (in 2006). The cloud itself presents an example of the diverse information sources. Various works have been going on to connect this diverse information. Consider an example, the person is related to FOAF ontology may also connect with the DBLP paper (DBLP Ontology) with the has-author relationship. The co-author of the paper might be taught in the same university (University Ontology) who's one of the student presents this paper in the conference (Event Ontology) held in Japan (Geoname Ontology). Former example explained the connectivity of a person to another, to a paper, to an organization, and to the country itself. Accessing all this information automatically without moving through hyperlinks of the pages was the initial idea of the LOD Cloud. Linked Open Data Cloud, a phenomena of Tim Berners Lee already taken a well established space in the current applications [1][2][3]. Many organizations have come forward in the last decades to provide successful open source Knowledge Base [1]. This knowledge has been utilized in many applications [4] to give the meaningful results by combining different data sources. This was only possible because of their same structured format. The triplet of RDF contains information about the concepts in the form of their relationship among all other but related concepts. These links also have some special characteristics. Different links can attach with different objects or value. Like an actor of the movie can be a director also. But this is not the case when we are taking about bibliographic database. RKBExplorer [5] contain information about bibliographic data, we explore the data for generating tags from it. RKBExplorer provides the unified view of the heterogeneous data sources. ReSIST project [5] proposed a semantically enabled knowledge structure. The aim of the project was to provide services from different but related data sources. In the next sections we describe about methodology, implementation and the results of the framework, finally concludes with future work in the last section.

2. Triplet Extraction through multiple RDF datasets

We have studied 3-Bibliographic datasets named as: DBLP, IEEE and ACM. These are the very basic and most utilizes datasets in the bibliographic searching. Users in this search are not ordinary users they are trained enough to utilize the result of the searching queries [9]. But the time consumption for utilization of these searches is the main problem in this. Linked Open Data Cloud is a good example of semantic connectivity among the huge knowledge source. The information provided by the datasets in this cloud is semantically linked with each other. We can consider the datasets as a huge graph in which the vertices are the subjects and objects.

Linkage information is consumed as a predicate unique. between the subject and object. So in-conjunction, the information is called as a triplet. The linkage information gives us an opportunity to specifically utilize the objects or value. Here the authors have listed some interested information about the three RDF datasets. Thanks to the bibliographic RDF converter organization that provide a common ontology for all the three datasets. Observation told that some common information like "sub-area-of", "has-author", "fullname", "has-title", "has-date", "year-of" presents in the three data sets(IEEE,ACM,DBLP).

Some information has similar meaning but different predicates are used to them like "cites-publication-reference" in DBLP, ACM and "is-very-strongly-related-to", "is-strongly-related-to", "is-related-to" information in IEEE. Another example of this is "has-ieee-keyword" of IEEE and "address-generic-area-of-interest" in ACM dataset.

In our discussion we called these similar terms as the complementary terms (purple color). Our aim to utilize the complimentary terms as well as direct information of the 3 datasets with the least preprocessing steps for the ease of the use it.