IJSTE - International Journal of Science Technology & Engineering | Volume 2 | Issue 05 | November 2015
ISSN (online): 2349-784X

A Contrast Framework for Similarity Detection

Divya B Nair
PG Student
Department of Computer Science and Engineering
Mount Zion College of Engineering, Kadammanitta, Pathanamthitta

Smita T Thomas
Research Scholar
Department of Computer Science and Engineering
Vels University, Chennai

Abstract

The research works play an important role in promoting the science and technology competitiveness of a country. Due to the limitation of information open and sharing. It is possible to approve similar projects by different government departments. These similar projects are a waste of both scientific resources and money. To overcome this, here propose a contrast framework for detecting similar projects based on big data mining technology, providing evidence-based decision making for government departments during the project approval process. Firstly constructed a big data file associated with government approved projects including all details about the project: titles, principal investigators, research organizations, keywords, and bibliographies. Secondly, a contrast framework is proposed to detect similar projects by mining information from the above big data file. Finally, implement the Hadoop architecture to speed up the data mining algorithm.

Keywords: Big date mining; contrast framework; Distributed computing; Hadoop architecture; similar project detection

I. INTRODUCTION

To support the science and technology most countries invest a large amount of money for many research fields, are usually established or approved by different government sectors or departments. If there is no information open and sharing, it is possible to apply for more number of projects from different government sectors with one research work. This will result in waste of scientific resources and money. Therefore, it is a challenge to find a way of efficiently detecting similar projects.

Document retrieval techniques have been used for detecting similar projects in past years. Many researchers used document retrieval techniques to detect similar projects. Mr. Jiang invented a prototype system of scientific research project management based on text mining technology. It focused on text segmentation and text modelling. Mr. Zuo invented a non- segmentation to find similar projects but it uses frequently closed suffix-tree modest vector to find the similarity of different projects. Mr. Fang proposed a method to find duplicate projects by using TF-IDF method.

Big data gathered a huge attention worldwide. The logic is; - Capture vast content of data related with research objects using Data Mining techniques. It has been already proved successfully for several areas. Example: - Election result analysis prior to the election campaign in US president election, Election advertisement etc.

This paper made a contrast framework to detect similar project based on big data mining technology. Here we create a big data file organising a wide variety of scientific research projects including project titles principle investigators, research organizations, key word sets and bibliographies. The contracts framework mines the information from the organised big data file for finding similarities. Efficiency of Data mining algorithm can be stimulated by implementing Hadoop architecture.

II. CONTRAST FRAMEWORK

The following diagram represents the framework. It consists of 5 main parts. The big data file, 2.Task Parser, 3. Match finder, 4. Hadoop computing, 5. Result viewer

Big data file consists of a huge collection of information about research works. Task parser performs translation of users request to machine executable instructions and passes them to Hadoop computing. Match finder finds the similarity of projects by data mining and information integration. The Hadoop architecture is used for managing distributed computing; it manages the computer cluster system and detects similar projects. The final result is visualised through the result viewer to the requester.

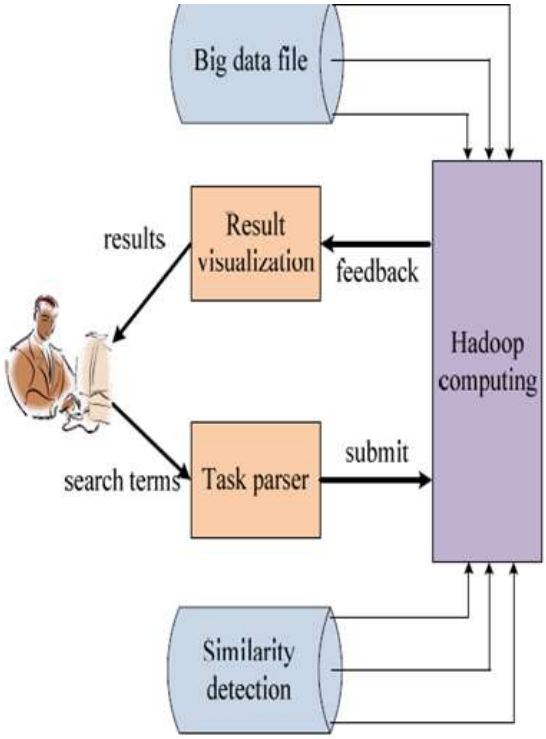

Fig. 1: Contrast Framework

Project title, keyword sets, research organizations, principle investigators and bibliography of scholar paper have an association relationship among each other. A big data file is created from a wide range of network due to huge collection of projects and bibliographies. Its metadata includes a project identifier, keyword, research organization and principle investigators. The association relationship among five elements in the big data file can be obtained by project identifier. The group of terms describing the research work is referred as keyword set. It is obtained from project title with text segmentation technique and keyword from the scholar paper. The relation among project identifier and scholar paper is gained from acknowledgement of a paper. Due to the massive collection of scholar paper the keyword sets are automatically constructed by text mining and text extraction techniques. To improve the accuracy the final results are manually verified. The details about principle investigator and the research organization are gathered from the database of approved research projects. It is difficult to identify the data if it is not complete.

A. Match Finder

Using the keyword set, research organizations and principle investigators the match finder module calculates the similarity of projects. Then a weighted average method is used to obtain the similarity of projects. The keyword set is used to detect the similarity of research projects as key word set defines the main contents of research projects. If two projects guided by same principle investigator may cause being similar that two different principle investigators. However the match finder checks whether the principal investigators are same or not to find the similarity. Same procedure is being done for research organization to detect the similarity. Even though it calculated the similarity of organization and principal investigator the main similarity is obtained from calculating the similarity among the key words. In order to increase the accuracy, it used the principal investigators and research organisation.

B. Hadoop Computing

Distributed computing is used to speed up these algorithms. And a Hadoop architecture is used to implement the distributed computing for similarity detection is shown in figure. It includes a job tracker who plays the role of a manager dividing the task into multiple subtasks and organizes them to be executed. The Tasktrakers play the role of workers who executes the sub tasks and reports the results to the Job traker, the Tasktrakers download the data from the big data file and generates an intermediate result. The Job traker communicate with the Tasktrakers and record all its status. The distributed computing include two types of subtasks, one is the map task and the other is the reduce task. The map task generates the intermediate result with the split data for similarity detection. The reduce task finally organize all the intermediate result and calculate the similarity.

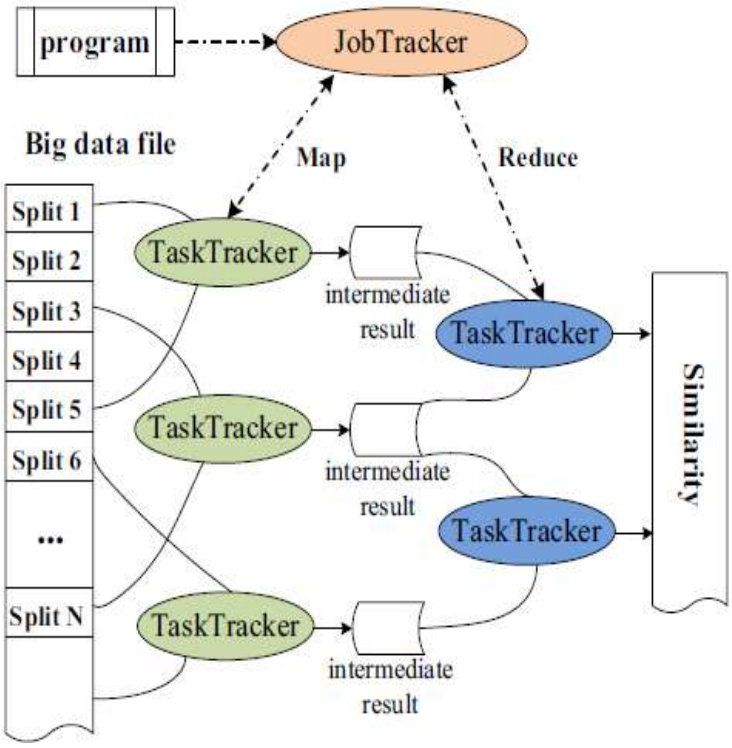

Fig. 2: Hadoop Architecture

III. CONCLUSION

This paper proposed a contrast frame work of detecting similar projects by integrating data from a huge database that merges massive information of several sources with big data mining techniques. The framework can help the different government sectors to avoid same projects and hereby allocate the scientific resources optimally.

REFERENCES

[1] Li, Lirong Song, Hui Zhao ,A Discriminate Framework For detecting scientific research projects based on Big data mining Shanqing Institute of Scientific and Technical Information of China
[2] H. Zhang, T.Chow, A multi-level matching method with hybrid similarity for document retrieval [j], Expert Systems with Applications, 2012, 39 (3):2710, 2719.
[3] J.Reid, M. Lalmas, K. Finesilver, M. Hertzum, Best entry points for structured document retrieval- part II: types, usage and effectiveness [J], information processing and management, 2006, 42(1):89-105
[4] J.Reid, M.Lalmas, K.Finsilver, M. Hertzum, Best entry point for structured document retrieval-part I: Characteristics [J], information processing and management. 2006, 42(1):74-88.
[5] P. Kalczynki, A.Chou, Temporal document retrieval model for business news archives [J] Information Processing and Management, 2005, 41(3):635-650
[6] S.Jing. Research and application of text mining methods for scientific management [D]. Dalian: Dalian University of Technology 2006.
[7] C.Zuo. Research and implementation of finding duplicate science projects with the non-segmentation technology [D].Chongqing: Chongqing University, 2010.
[8] Y.Fang. Research on duplicated science projects detection by improving the TF-IDF method [J]. Information Research, 2012(1):1-3.
[9] Y.WU. Research on classification and duplicated project detection for science projects based on hierarchical clustering [D]. Tianjin: Tianjin University of Finance and Economics, 2008.
[10] M.Lin, Y.Kang, C. Zang. Research on fuzzy C mean algorithm based on research project management application [J]. Computer Engineering and Design, 2010, 31(7):1570-1572.
[11] Y. Liu, F. Zhang, Q. Niu. Study on how to avoid the repeated projects of scientific research management [J]. Science and Technology Management Research, 2010(21): 198-200.
[12] Google flu trends. [EB/OL]. http://www.google.org/flutrends.