Journal of Engineering and Applied Sciences 13 (23): 10037-10038, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Performance Evaluation of RM-MapReducer in Web Page Categorization

[1]P. Malarvizhi and [2]N. Radhika
[1]Department of CSE, Karpagam University, Coimbatore, Tamil Nadu, India
[2]Department of Computer Science and Engineering, Amrita School of Engineering,
Amrita University, Coimbatore, Amrita Vishwa Vidyapeetham, India

Abstract: This study presents the performance evaluation of Relevancy Measure based MapReducer (RM-MapReducer) in supporting web page categorization with different datasets between 200 and 400 web pages as test data. Experiments were performed using datasets of WebKB and ODP with the real time crawler dataset. It was observed that the real time crawler dataset provides better results compared to the other dataset of WebKB and ODP using MapReduce programming model.

Key words: Web page categorization, RM-MapReducer, dataset, web pages, programming model

INTRODUCTION

The dimension and the dynamic nature of the web creates a need for classification of web pages (Malarvizhi and Radhika, 2016) to manage the information retrieval process. The rapid growth of web increases the need to have automated assistance to organize the huge amount of information provided by keyword-based search engines. Categorization of web pages is a process that assigns the labels of the category of predefined to a web page (Malarvizhi and Radhika, 2016). The explosive growth of web introduces a need for web page classification (Uysal, 2016) to manage the information retrieval tasks and to improve the performance of the search (Qi and Davision, 2009). In this research, MapReduce parallel programming model with functions map and reduce is used for categorizing the web pages based on the relevancy measure. MapReduce, the Googles popular framework is an attractive parallel model with functions map and reduce suitable for parallel processing of arbitrary data (Malarvizhi and Pujeri, 2012).

MATERIALS AND METHODS

Relevancy Measure based MapReducer (RM-MapReducer): MapReduce, the parallel programming model was used to categorize the web pages based on the relevancy measure. The Relevancy Measure RM is computed by comparing the similarities between the two vectors v_1 of frequent keyword and v_2 of keyword with the predefined category vector C is given below. For all $w \in W$ do; for all $c \in C$ do; c←(RM)max.

RM is the sum of the two similarities of v_1 and v_2 (Malarvizhi and Pujeri, 2012). MapReduce is a popular programming model with functions map and reduce and meets a number of varieties of applications (Chen and Schlosser, 2008). The MapReduce programming model assigns a predefined category label to a web page based on the (RM) max value. The dataset of web pages are given as input to the map function of the MapReducer and the reduce function of the MapReducer computes the relevancy measure for each category and assigns the (RM)max value category label to the web pages.

RESULTS AND DISCUSSION

This comparative study was performed on dataset of 400 web pages and was implemented using java. The performance of the study was evaluated on test data with the evaluation metrics of precision and recall.

Dataset: Two datasets, dataset 1 and 2 are created for test data. Dataset 1 contains the web pages related to the category of C1-4 of course, student, department and conference. Web pages of category course, student and department were collected from WebKB dataset and category conference was collected from computer science conferences of the ODP website. Dataset 2 contains the web pages of category C1-4 of course, student, department and conference collected from web by using the web crawler websphinx. Websphinx is a customized web crawler used to collect the web pages from web. The 200 web pages were collected for each dataset of C1 of 62 web pages, C2 of 62 web pages, C3 of 35 web pages, C4 of 41 web pages and totally 400 web pages were collected for dataset.

Corresponding Author: P. Malarvizhi, Department of CSE, Karpagam University, Coimbatore, Tamil Nadu, India

Table 1: Precision and recall of dataset 1

Category	Precision	Recall
C1	0.94	0.96
C2	0.95	0.92
C3	0.71	0.94
C4	0.93	0.72

Table 2: Precision and recall of dataset 2

Category	Precision	Recall
C1	0.94	0.97
C2	0.97	0.94
C3	0.73	0.94
C4	0.94	0.71

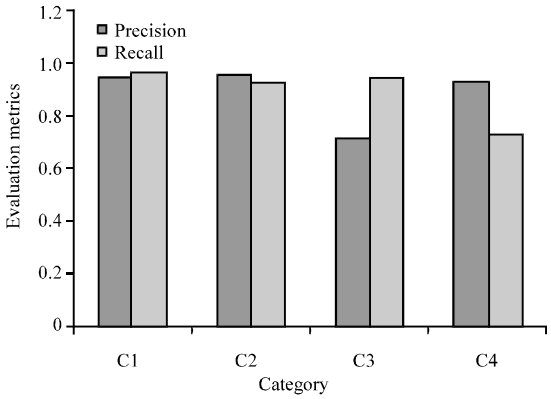

Fig. 1: Performance of the RM-MapReducer on dataset 1 of WebKB and ODP

Performance measures: The performance of the study was evaluated on test data with the evaluation metrics of precision and recall. The evaluation metrics used are given in Eq. 1 and 2:

$$\text{Precision} = \frac{\text{Relevant pages} \bigcap \text{retrieved pages}}{\text{Retrieved pages}} \quad (1)$$

$$\text{Recall} = \frac{\text{Relevant pages} \bigcap \text{retrieved pages}}{\text{Relevant pages}} \quad (2)$$

Comparison of performance measures: The study was analyzed with two datasets in terms of precision and recall measures and the results showed that the classification done was highly accurate with real time crawler dataset. Table 1 and 2 gives the performance measures of the resultant categories of the dataset1 and dataset 2. The results are plotted as graph in Fig. 1 for dataset 1 and in Fig. 2 for dataset 2. From the table and graphs, it is

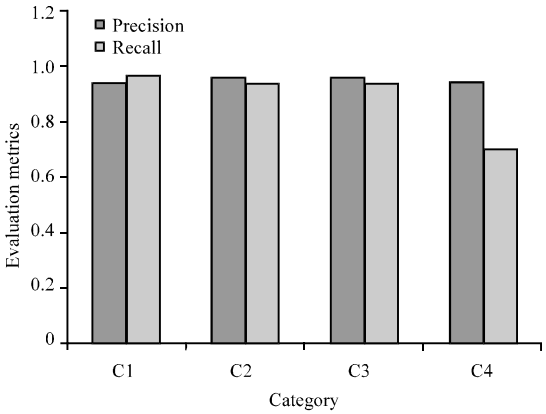

Fig. 2: Performance of the RM-MapReducer on dataset 2 of real time crawler

evident that the precision and recall obtained by the MapReducer for dataset 2 performs better compared to dataset 1.

CONCLUSION

In this research, the behavioural comparison of the MapReduce parallel programming model based on relevancy measure in web page categorization is performed for different datasets and the results demonstrates that the performance of the relevancy measure based MapReduce programming model in real crawler is accurate than the benchmark dataset of WebKB and ODP.

REFERENCES

Chen, S. and S.W. Schlosser, 2008. Map-reduce meets wider varieties of applications. Intel. Res. Pittsburgh, Tech. Rep., 5: 1-8.

Malarvizhi, P. and N. Radhika, 2016. Web page categorization: An overview. Intl. J. Adv. Res. Comput. Eng. Technol., 5: 1370-1373.

Malarvizhi, P. and R.V. Pujeri, 2012. Distributed approach to web page categorization using map-reduce programming model. Intl. J. Eng. Technol., 3: 373-386.

Qi, X. and B.D. Davison, 2009. Web page classification: Features and algorithms. ACM. Comput. Surv., 41: 12-31.

Uysal, A.K., 2016. An improved global feature selection scheme for text classification. Expert Syst. Appl., 43: 82-92.