A CROWDSOURCING WORKER QUALITY EVALUATION ALGORITHM ON MAPREDUCE FOR BIG DATA APPLICATIONS

S.GAYATHRI[1], Mrs.N.VASUNTHIRA DEVI [2]
1 (Mphil (Research Scholar), Annai Vailankanni Arts and Science college, Thanjavur, Tamilnadu, India)
2(Assistant Professor, Department of Computer Science, Annai Vailankanni Arts and Science college, Thanjavur, Tamilnadu)

Abstract:

Crowd sourcing is a up-to-the-minute rising scattered computing and big business model on the surroundings of Internet embryonic. With the enlargement of crowd sourcing systems, the information magnitude of crowdsourcers, contractors and errands grows hastily. The worker quality appraisal based on big data analysis technology has become a critical challenge. This paper first proposes a general worker quality evaluation algorithm with the purpose of is useful to any significant errands such as cataloging, toning, filtering, tagging and many other up-and-coming applications, exclusive of assassination resources. Second, we become conscious the appraisal algorithm in the Hadoop platform using the Map Reduce parallel programming model. Finally, to efficiently verify the accurateness and the effectiveness of the algorithm in a wide multiplicity of big data scenarios, we manner a series of experiments. The trial results exhibit that the proposed algorithm is precise and helpful. It has high computing recital and flat scalability and it is appropriate for significant member of staff value evaluations in a big data atmosphere.

Keywords — **Map Reduce, big data scenarios, crowdsourcers, Hadoop platform.**

I. INTRODUCTION

Crowd sourcing is a distributed problem-solving and production model. In this distributed computing model, enterprises distribute tasks through the Inter-net and recruit more suitable workers to involve in the task to solve technical difficulties. Nowadays, more and more businesses and enterprises have begun to use the crowd sourcing model. For enterprises, the crowd sourcing model can reduce production cost, and promote their technology and creativity. The crowd sourcing model is oriented to the public, and every Internet user can choose to participate in the crowd sourcing tasks that they are interested in to provide solutions for enterprises. However, for one task, there may be a large number of workers involved in it and provide solutions. The crowdsourcers will be confused when they faced with such a huge number of solutions and it is difficult for them to make a final choice.

Moreover, not every person is qualified to serve enterprises because of their different backgrounds and different personal qualities. There may even be Malicious workers in crowd sourcing platform. Therefore, worker quality control has gradually Become an important challenge for the crowd sourcing model. It is of great importance to mine the information about the worker's self quality from a large number of worker data to provide the crowdsourcers some reference.

This paper mainly studies the core problem of worker quality control: worker quality evaluation. The worker quality evaluation will help enterprises recruit high-quality workers who can provide them high-quality solutions. It is of great significance to both the quality of the tasks and the environment of the crowd sourcing platform.

II. LITERATURE REVIEW

A literature review is an account of what has been published on a topic by accredited scholars and researchers. Occasionally you will be asked to write one as a separate assignment, but more often it is part of the introduction to an essay, research report, or thesis. In writing the literature review, your purpose is to convey to your reader what

knowledge and ideas have been established on a topic, and what their strengths and weaknesses are. As a piece of writing, the literature review must be defined by a guiding concept (e.g., your research objective, the problem or issue you are discussing or your argumentative thesis). It is not just a descriptive list of the material available, or a set of summaries

Besides enlarging your knowledge about the topic, writing a literature review lets you gain and demonstrate skills in two areas

Information seeking: the ability to scan the literature efficiently, using manual or computerized methods, to identify a set of useful articles and books

Critical appraisal: the ability to apply principles of analysis to identify unbiased and valid studies.

Using Crowd sourcing and Active Learning to Track Sentiment in Online Media

Tracking sentiment in the popular media has long been of interest to media analysts and pundits. With the availability of news content via online syndicated feeds, it is now possible to automate some aspects of this process. There is also great potential to crowd source Crowd sourcing is a term, sometimes associated with Web 2.0 technologies that describes outsourcing of tasks to a large often anonymous community. Much of the annotation work that is required to train a machine learning system to perform sentiment scoring. We describe such a system for tracking economic sentiment in online media that has been deployed since August 2009. It uses annotations provided by a cohort of non-expert annotators to train a learning system to classify a large body of news items. We report on the design challenges addressed in managing the effort of the annotators and in making annotation an interesting experience.

Parallel rough set based knowledge acquisition using Map Reduce from big data

Nowadays, with the volume of data growing at an unprecedented rate, big data mining and knowledge discovery have become a new challenge. Rough set theory for knowledge acquisition has been successfully applied in data mining. The recently introduced Map Reduce technique has received much attention from both scientific community and industry for its applicability in big data analysis. To mine knowledge from big data, we present parallel rough set based methods for knowledge acquisition using Map Reduce in this paper. Comprehensive experimental evaluation on large data sets shows that the proposed parallel methods can effectively process big data.

Dryad: distributed data-parallel programs from sequential building blocks

Dryad is a general-purpose distributed execution engine for coarse-grain data-parallel applications. A Dryad application combines computational "vertices" with communication "channels" to form a dataflow graph. Dryad runs the application by executing the vertices of this graph on a set of available computers, communicating as appropriate through flies, TCP pipes, and shared-memory FIFOs. The vertices provided by the application developer are quite simple and are usually written as sequential programs with no thread creation or locking. Concurrency arises from Dryad scheduling vertices to run simultaneously on multiple computers, or on multiple CPU cores within a computer. The application can discover the size and placement of data at run time, and modify the graph as the computation progresses to make efficient use of the available resources.

Dryad is designed to scale from powerful multi-core single computers, through small clusters of computers, to data centers with thousands of computers. The Dryad execution engine handles all the difficult problems of creating a large distributed, concurrent application: scheduling the use of computers and their CPUs, recovering from communication or computer failures, and transporting data between vertices.

CrowdER: crowd sourcing entity resolution

Entity resolution is central to data integration and data cleaning. Algorithmic approaches have been improving in quality, but remain far from perfect. Crowd sourcing platforms offer a more accurate but expensive (and slow) way to bring human insight into the process. Previous work has proposed batching verification tasks for presentation to human workers but even with batching, a human-only approach is infeasible for data sets of even moderate size, due to the large numbers of matches to be tested. Instead, we propose a hybrid human-machine approach in which machines are used to do

an initial, coarse pass over all the data, and people are used to verify only the most likely matching pairs. We show that for such a hybrid system, generating the minimum number of verification tasks of a given size is NP-Hard, but we develop a novel two-tiered heuristic approach for creating batched tasks. We describe this method, and present the results of extensive experiments on real data sets using a popular crowd sourcing platform. The experiments show that our hybrid approach achieves both good efficiency and high accuracy compared to machine-only or human-only alternatives.

III. IMPLEMENTATION

Current system:

Crowdsourcers almost release tasks at all times due to the large-scale crowd sourcing platform. Additionally, a large number of workers participate in these tasks. Therefore, the crowd sourcing platform will generate a large amount of data every moment, including crowd sourcing tasks, worker behaviors, and the solutions of tasks. The large amount of data put forward new demands to the calculated performance of crowd sourcing platform. The use of big data technology to specially process these massive data is a key issue that the crowd sourcing platform needs to consider.

Shortcomings of the current system:

1. Most of these crowd sourcing systems rely on offline or artificial worker quality control and evaluation or simply ignore the quality control issues.
2. High computational cost.
3. Performance accuracy is less.

Proposed system:

Therefore, to evaluate the quality of the workers in the crowd sourcing platform accurately, we first propose a general worker quality evaluation algorithm. This algorithm achieves the worker quality evaluation for multiple workers and multiple problem types with no pre-developed answer, and the algorithm has a stronger scalability and practicality compared with the algorithm. Second, we propose to use the Map Reduce programming model to realize large-scale parallel computing for worker quality and implement the proposed algorithm in the Hadoop platform. Finally, we conduct a series of experiments to analyze and evaluate the performance of worker quality evaluation algorithm.

Advantage of proposed system:

1. The proposed algorithm is effective and has a high performance.
2. It can meet the needs of parallel evaluation of the large-scale workers in a crowd sourcing platform.

IV. CONCLUSIONS

In this paper, we first proposed a general worker quality evaluation algorithm, which is applied to any critical crowd sourcing tasks without pre-developed answers. Then, to satisfy the demand of parallel evaluation for a multitude of workers in a big data environment, we implement the proposed algorithm in the Hadoop platform using the Map Reduce programming model. The experimental results show that the algorithm is accurate and has high efficiency and performance in a big data environment.

REFERENCES

1. *D.C. Brabham, "Crowdsourcing as a Model for Problem Solv-ing: An Introduction and Cases," Convergence the Internation-al Journal of Research Into New Media Technologies, vol. 14, no. 1, pp. 75-90, 2008.*
2. *M. Allahbakhsh, B. Benatallah, A. Ignjatovic, et al, "Quality Control in Crowdsourcing Systems: Issues and Directions," IEEE Internet Computing, vol. 17, no. 2, pp. 76-81, 2013.*
3. *A. Doan, R. Ramakrishnan, and A.Y. Halevy, "Crowd sourcing Systems on the World-Wide Web," Communications of the ACM, vol. 54, no. 4, pp. 86-96, 2011.*
4. *P. Clough, M. Sanderson, J. Tang, et al, "Examining the Limits of Crowdsourcing for Relevance Assessment," IEEE Internet Computing, vol. 17, no. 4, pp. 32-38, 2013.*
5. *B. Carpenter, "Multilevel Bayesian Models of Categorical Data Annotation," unpublished, 2008.*
6. *A. Brew, D. Greene and P. Cunningham, "Using crowdsourc-ing and active learning to track sentiment in online media," In Proceedings of the 6th Conference on Prestigious Applications of Intelligent Systems, 2010.*

7. J. Howe, "The Rise of Crowdsourcing," Wired Magazine, vol. 14, no.14, pp. 176-183, 2006.
8. V. C. Raykar, S. Yu, L. H. Zhao, et al, "Learning From Crowds,"Journal of Machine Learning Research, vol. 11, no. 2, pp. 1297-1322, 2010.
9. J. Manyika, M. Chui, B. Brown, et al, "Big Data: The next frontier for innovation, competition, and productivity," 2011.
10. S. C.H. Hoi, J. Wang, P. Zhao, et al, "Online feature selection for mining big data," BigMine, pp. 93-100, 2012.
11. K. Michael, K.W. Miller, "Big Data: New Opportunities and New Challenges," Computer, vol. 46, no. 6, pp. 22-24, 2013.
12. C. Lynch, "Big Data: How do your data grow?," Nature, Vol.455, No. 7209, pp. 28-29, 2008.
13. F. Chang, J. Dean, S. Ghemawat, et al, "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems, Vol. 26, No.4, 2008.
14. M. Joglekar, H. Garcia-Molina, and A. Parameswaran, "Evalu-ating the crowd with confidence," Proceedings of the 19th ACM SIGKDD international conference on Knowledge discov-ery and data mining, ACM, pp. 686-694, 2013.
15. J. Zhang, T. Li, and Y. Pan, "Parallel rough set based knowledge acquisition using MapReduce from big data," Big-Mine, pp. 20-27, 2012.
16. J. Dean, and S. Ghemawat, "MapReduce: Simplified data pro-cessing on large clusters," Communications of the ACM, vol. 51, no.1, pp. 107-113, 2005.
17. D. Hastorun, M. Jampani, G. Kakulapati, et al, "Dynamo: Ama-zons highly available key-value store," In: Proceedings of the 21st ACM Symposium on Operating Systems Principles, pp. 205-220, 2007.
18. M. Isard, M. Budiu, Y. Yu, et al, "Dryad: Distributed data-parallel programs from sequential build-ing blocks," European Conference on Computer Systems, pp. 59-72, 2007.
19. J. Wang, T. Kraska, M. J. Franklin, et al, "CrowdER: crowdsourcing entity resolution," Proceedings of the VLDB Endowment, vol. 5, no. 11, pp. 1483-1494, 2012.
20. N. Maisonneuve, and B. Chopard, "Crowdsourcing Satellite Imagery Analysis: Study of Parallel and Iterative Models," GIScience, pp. 116-131, 2012.