ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 7, July 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Text Mining Approach for Evaluating Crowdsourcing Contests using Hierarchical Clustering

R. Jayanthi[1]
Assistant Proffessor
PG and Research Department of Computer Science
Quaid-e-Millath Govt. College for Women (Autonomous)
Chennai, India

J. Josphin Mary[2]
M. Phil Research Scholar
PG and Research Department of Computer Science
Quaid-e-Millath Govt. College for Women (Autonomous)
Chennai, India

Abstract: ***The word crowdsourcing is a combination of word crowd + outsourcing. It means the tasks are distributed to a large group of people especially online users and collect ideas from those people. It is also defined as a contribution of large number of people's work on a particular task. It combines the efforts of numerous people. It is also called distributed problem solving technique. Due to numerous people share their ideas, it is very difficult to evaluate the submissions. And also it is very expensive and time consuming task. This paper gives an approach for overcome the decision making problem with the help of keyword based analysis and Text mining techniques.***

Keywords: Text mining, Clustering, Crowdsourcing, Unstructured, Text, Information Extraction.

I. INTRODUCTION

The word Crowdsourcing is a combination of the word crowd and outsourcing. The term crowdsourcing was coined by Merriam Webster in 2006. There are so many crowdsourcing websites and one of the popular is Amazon's Mechanical Turk. Here the crowdsourcer post their task, time limit and award for that task. The users who are interested in crowdsourcing contest they can share their ideas through crowdsourcing websites.

Crowd voting, crowd funding, crowd contest, micro tasks and macro tasks are some types of crowdsourcing. Crowdsourcing system was proposed in order to reduce the company's production cost. The key assumption in the "wisdom of crowds" is that the opinions, insights, ideas, and knowledge of the "many" can be better than that of a given expert. To have a "wise" crowd, in Surowiecki's (2004) framework, there are four prerequisites: 1) cognitive diversity, by which each individual involved has some private information; 2) independence, wherein each person's opinion or decision is not influenced by those around them; 3) decentralization, through which individuals can specialize and tap into local fonts of knowledge; and 4) aggregation, which stresses the importance of structural mechanisms for translating many private opinions or decisions into a collective decision. If we compare these four prerequisites with the three defining elements of crowdsourcing.

Due to numerous peoples sharing their own ideas. It is very difficult to identify which ideas or work is best among all submissions. To overcome this problem we have to move on text mining techniques. Text mining is used to process structured and unstructured data also. Unstructured text contains very large amount of information which is not easily used for further processing. So that we need some techniques to overcome this problem for extracting required patterns. Text mining plays an important role of extracting useful patterns from unstructured text. It is one of the emerging technologies for Knowledge Discovery Process. Document organization and pattern discovery becomes the main task in data mining [1]. This paper explains how text mining techniques are used to evaluate crowdsourcing submission and also explain the working, merits and demerits of crowdsourcing systems.

The rest of the paper is organized as follows: Section II describes crowdsourcing. Section III describes crowdsourcing websites. Section IV describes the role of text mining in crowdsourcing system. Section V describes proposed methodology Section VI describes the implementation details. Section VII describes the conclusion. Section VIII future work.

II. CROWDSOURCING

A. *Working of Crowdsourcing System*

Crowdsourcing system works based on the following criteria.

1. Crowdsourcing workers are trained and tested before working on tasks.
2. Qualified workers choose tasks from the catalog of work.
3. Each task has an efficient worker interface.
4. Smart quality control ensure work is correct.
5. Worker promotions and rewards.
6. Elasticity and output based pricing.

Figure 1. Working of crowdsourcing system

B. *Advantages of Crowdsourcing*

- ✓ Significant cost saving.
- ✓ Free marketing
- ✓ Customer loyalty
- ✓ Talent seeks you
- ✓ Instant hiring/outsourcing potential
- ✓ No or very little overhead

C. *Disadvantages of Crowdsourcing*

- ✓ Questionable ownership
- ✓ Deadweight members
- ✓ Paying more does not improve quality
- ✓ Waste of time
- ✓ Marketing risk

III. CROWDSOURCING WEBSITES

Table I The following table shows some examples of crowdsourcing websites

S.no	Websites	Descriptions
1	ClickWorker.com	It is a small NASA experimental project that uses public volunteers for scientific task
2	MicroWorkers.com	Microtasking is the process of splitting a job into its component microwork and distributing this work over the Internet.
3	CloudCrowd.com	*CloudCrowd* is an online work platform that provides writing, editing, and other work.
4	Crowdflower.com	Once data is uploaded to the platform, the system automatically allocates the work to contributors and tests them against known answers hidden within the task
5	CrowdSource.com	Crowdsourcing is the process of getting work or funding, usually online, from a crowd of people.
6	MTurk.com	It is a crowdsourcing Internet marketplace enabling individuals and businesses (known as Requesters) to coordinate the use of human intelligence to perform tasks that computers are currently unable to do.
7	DoMyWork.net	DoMyWork allows people to communicate and work together. You can either sign up as a user or as a member of staff. Users can submit their work requests, of whatever kind, and members of staff can then opt to work on that project at the price specified by the user.
8	InfiniteWorkers	InfiniteWorkers is an innovative online platform that connect employers and workers from all around the world. This platform helps online freelancers to make money by completing a variety of tasks and employers can find the best workers to get work done.
9	RapidWorkers.com	We are combining Workers and Employers to minimize publicity and marketing costs while at the same time providing money to everyday users.
10	ShortTask.com	It connect online jobseekers with providers

IV. ROLE OF TEXT MINING IN CROWDSOURCING SYSTEM

Text mining is the process of extracting patterns and useful information from large amount of unstructured data sets. Text mining is also called as Knowledge discovery process and Text data mining. The main aim of text mining is that the linking of extracted information to form new facts or new hypothesis to be explored further more conventional means of experimentation [3]. The crowdsourcing system includes organizers who assigns task to the crowd. Crowd is nothing but a group of people who performs tasks on the crowdsourcing systems [4]. In Crowdsourcing system the crowdsourcer post their task, time period and reward for that particular task. The online users who are interested on crowdsourcing system, they can log on to particular crowdsourcing websites, select the particular task and share their ideas to those organizations. Due to this crowd, so many ideas come to the organizations and also it is very difficult to select which is best which ideas or task satisfied the required criteria. So crowdsourcing system moves on to text mining techniques.

Text mining performs operations in the following order. The first step is pre-processing technique like stopword removal and stemming, pos tagging, tokenization. And the result of the pre-processing goes as an input to the next process is clustering. And finally the tasks are clustered as which satisfies the required criteria and which are not satisfying the criteria. Text mining applies clustering based on the similarity measurements. Similarity measurements are calculated using TF-IDF values and cosine similarities. In this paper data are clustered using TF-IDF values. TF-IDF is an abbreviated form of Term Frequency – Inverse Document Frequency.

V. PROPOSED METHODOLOGY

The proposed methodology considered data from crowdsourcing slogans contest. Proposed methodology consists of five phases and the System Architecture is shown in fig.1.

Table II

Phase	Process	Description
I	Keyword base slogan extraction	Slogans are extracted using keywords
II	Stopword Removal	Stopwords like articles, prepositions, delimiters are removed from extracted slogans
III	TF – IDF Calculations	TF – IDF values are calculated for all extracted slogans
IV	Rating the slogans	All the slogans are rated using the key features of keywords.
V	Clustering	Clustering the higher rating slogans

Table 1 shows the steps included in proposed methodology. After this process are completed. The resulted slogans are given to the expert committee. The expert committee will judge which is best and which slogan will getting the reward.

This proposed methodology just gives a decision support to the expert committees. And also it saves time and reducing the work of expert committees

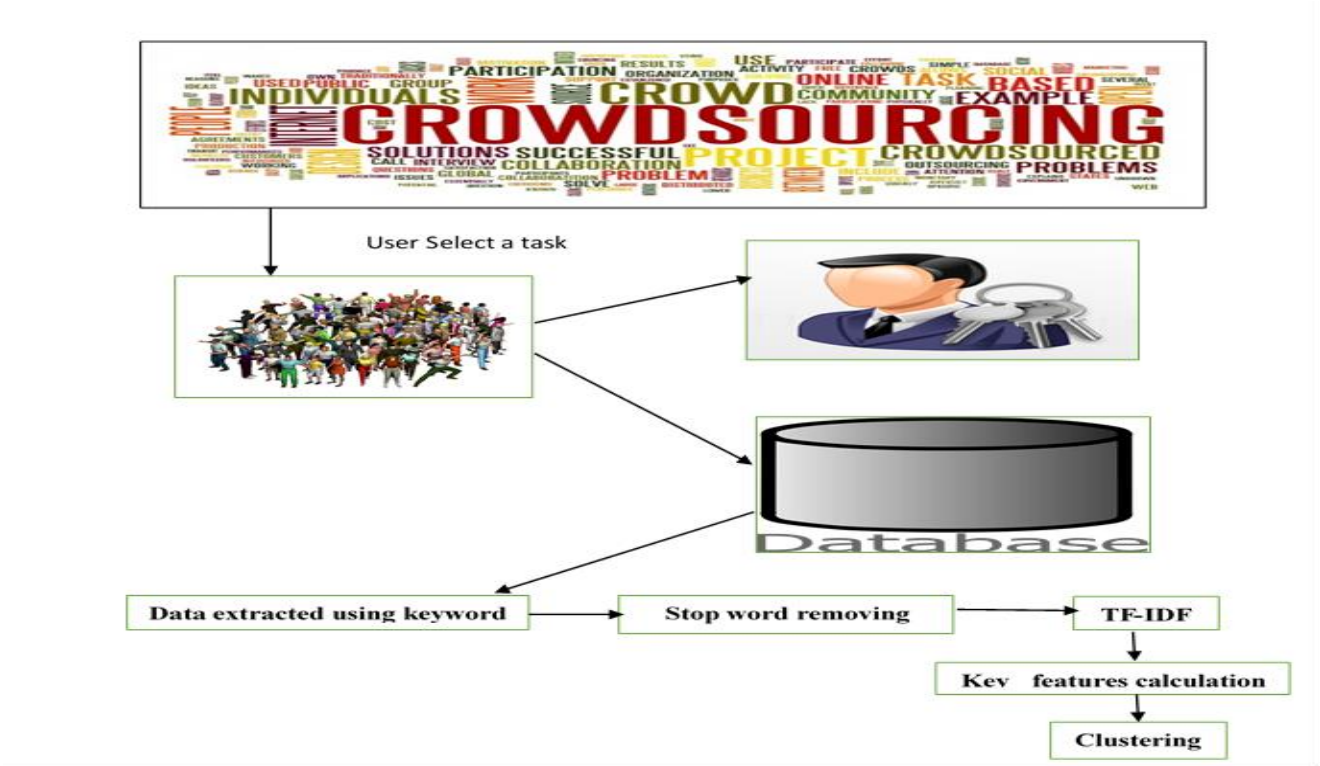

Figure. 2 System Architecture

VI. IMPLEMENTATION OF THE PROPOSED METHODOLOGY

In this paper, the dataset is taken from crowdsourcing slogan contest. The sample dataset which is considered for implementation is first pre-processed and shows how these data are further processed using Text Mining technique.

A. Slogan Table

Table III

C. No	C. Name	Slogans
1	Ashok	Save paper, Save trees, Save plant
2	Jack	Don't waste time. Reduce, reuse, recycle.
3	Jas	Millions of plastic materials are thrown away. Please recycle it
4	Alwin	Purchase items from recycled products.

B. Slogan Extraction

The above table shows the slogan details. Here the keyword is Recycle and trees. So the sentence containing the keyword are extracted. So out of four only three slogans are extracted[6].

1. Save paper, save trees, save plant.

2. Don't waste time. Reduce, reuse and recycle.

3. Millions of plastic materials are thrown away. Please recycle it.

C. Stop word removing

This step removes all stop-words and delimiters from the extracted above slogans.

1. Save paper save trees save plant.
2. Waste time Reduce reuse recycle.
3. Millions plastic materials thrown away Please recycle.

D. TF-IDF calculations

TF-IDF is an abbreviated form of Term Frequency and Inverse Document Frequency. TF can be calculated by using the below formula

TF= No. of times t appears in document / Total no. of words in a document.

IDF = log_e (Total no of document / no of documents with term in it)

Table IV

Terms	TF	IDF	TF-IDF
Recycle	0.5	0.69314718056	0.34655
Trees	0.17	1.0986122887	0.18676

Table V

Terms	TF	IDF	TF-IDF	Label
Save	0.5	1.0986122887	0.54930614435	Other
Paper	0.17	1.0986122887	0.1867640890790	Trees
Trees	0.17	1.0986122887	0.1867640890790	Trees
Plant	0.17	1.0986122887	0.1867640890790	Trees

Table VI

Terms	TF	IDF	TF-IDF	Label
Waste	0.2	1.0986122887	0.219722457	Recycle
Time	0.2	1.0986122887	0.219722457	Recycle
Reduce	0.2	1.0986122887	0.219722457	Recycle
Reuse	0.2	1.0986122887	0.219722457	Recycle
Recycle	0.2	0.40546510811	0.081093021	Other

Table VII

Terms	TF	IDF	TF-IDF	Label
Millions	0.14285714	1.0986122887	0.156944609	Recycle
Plastic	0.14285714	1.0986122887	0.156944609	Recycle
Materials	0.14285714	1.0986122887	0.156944609	Recycle
Thrown	0.14285714	1.0986122887	0.156944609	Recycle
Away	0.14285714	1.0986122887	0.156944609	Recycle
Please	0.14285714	1.0986122887	0.156944609	Recycle
Recycle	0.14285714	0.40546510811	0.064950153	Other

Table 4 shows the TF-IDF values of the keywords Re-cycle and Trees. Table 5, 6, 7 shows the TF-IDF values of the slogans. Using this TF-IDF values slogans are clustered. Slogans are labelled in manner, which slogans TF-IDF values are related to Re-cycle's values that slogans are labelled as Re-cycle. Like that slogans are rated.

E. Rating the slogans

In this step key features are assigned to each keywords.

Key features of recycle

1. Reduce
2. Reuse
3. Restore

Key features of trees

1. Save
2. Prevent
3. plant

Slogan 2 contains two key features of recycle. So it will be rated as 2 star rating. And Slogan1 contains 2 key features of trees. So this will also be rated as 2 star rating. So out of three slogans Slogan 1 and Slogan 2 gets higher rating. These two will be clustered and giving to the expert committees. The decision making is done by the expert committee and they will judge which Slogan is best and that will be awarded.

F. Distribution of slogans

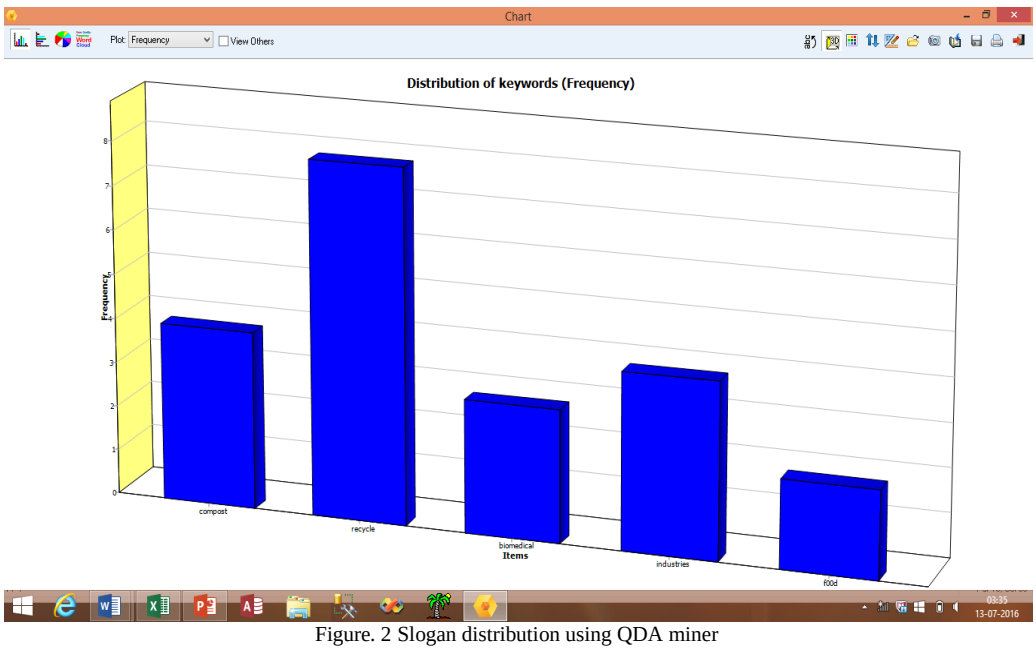

Figure. 2 Slogan distribution using QDA miner

Figure. 2 shows the slogan distribution using TF-IDF values in QDA miner tool.

Table VIII Distribution of slogans

S.NO	TERMS	FREQUENCY
1	Recycle	36.1
2	Compost	19.0%
3	Food	9.5%
4	Biomedical	19.0%
5	Industries	24.3%

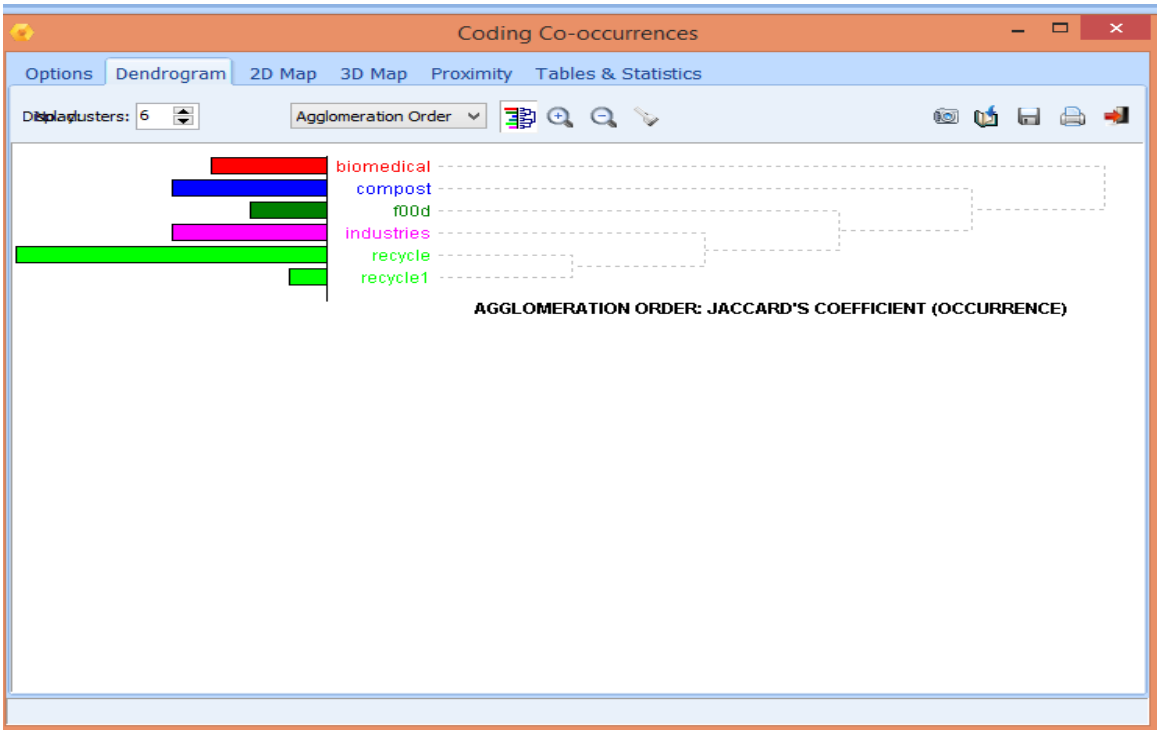

Figure 4 Data in agglomerative clustering

The above diagram shows the clustering of slogans using hierarchical clustering. The clusters are formed using jaccard's coefficient.

VII. Conclusion

Slogans are classified based on the similarity measures. Term frequency and Inverse document frequency are one of the best way of finding similarity between documents. In this paper, we have considered two types of slogans "recycle and trees". TF-IDF values are calculated for keywords recycle and trees and also calculated for slogans also and then slogans are assigned to particular keyword. Hierarchical clustering algorithms are used for clustering purpose. By using this approach result is very effective and efficient. This kind of text mining approach reduces the time and cost for crowdsourcing organizations.

VIII. Future Work

In this paper, slogans are labelled and clustered based on TF/IDF values. Hierarchical clustering algorithms are used for slogans clustering. In future other algorithms are used for clustering like k-means, birch rock etc. Results of these algorithms are compared with hierarchical clustering in future. TF-IDF based similarities are proposed in this paper. In future it can combined with cosine similarity or any other vector space model and the results can be compared.

References

1. Amrut M. Jadhav1, Devendra P. Gadekar2, A Survey on Text Mining and Its Techniques. International Journal of Science and Research (IJSR)
2. Thomas P. Walter Institute of Information Management, University of St. Gallen thomas.walter@unisg.ch Andrea Back Institute of Information Management, University of St. Gallen andrea.back@unisg.ch ,A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests
3. Vishal Gupta Lecturer Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India Email: vishal@pu.ac.in, Gurpreet S. Lehal Professor & Head, Department of Computer Science, Punjabi University Patiala, India. "A Survey of Text Mining Techniques and Applications"
4. Ms. P. V.Wazalwar#1, Ms. M. A. Potey ,D. Y. Patil College of Engg, Akurdi, University of Pune, India. "Approaches to Improve Student Feedback with Wisdom of Crowd"
5. Corney, J. R., Torres-Sanchez, C., Jagadeesan, A. P., and Regli, W. C. (2009) "Outsourcing labour to the cloud", International Journal of Innovation and Sustainable Development, 4, 4, 294-313.
6. Anandhi Sheshasaayee and R.Jayanthi, "A Text Mining Approach For Extract Opinion from Unstructured Text". Indian journal of science and technolog. Vol-8(36) December 2015.
7. Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011) Crowdsourcing systems on the World-Wide Web, Communications of the ACM, 54, 4, 86.
8. . Kleemann, F., Voß, G. G., and Rieder, K. (2008) Un (der) paid innovators: "The commercial utilization of consumer work through crowdsourcing, Science", Technology & Innovation Studies, 4, 1, 5-26.
9. Kozinets, R. V., Hemetsberger, A., and Schau, H. J. (2008) "The Wisdom of Consumer Crowds: Collective Innovation in the Age of Networked Marketing", Journal of Macromarketing, 28, 4, 339-354.
10. Leimeister, J. M., Huber, M., Bretschneider, U., and Krcmar, H. (2009) "Leveraging Crowdsourcing: Activation- Supporting Components for IT-Based Ideas Competition", Journal of Management Information Systems, 26, 1, 197-224.
11. Malone, T. W., Laubacher, R., and Dellarocas, C. N. (2010) "The collective intelligence genome, MIT Sloan Management Review," Spring, 51, 3, 21-31.

AUTHOR(S) PROFILE

R. Jayanthi, MCA., M.Phil., working as an Assistant Professor in PG & Research Department of Computer Science at Quaid-E-Millath Govt. College for Women(Autonomous), Chennai. She is pursuing her Ph.D in the University of Madras. Her areas of interests are Data Mining, Text Mining, Natural Language Processing, Information Extraction and Business Intelligence. She has published articles in more than 10 International Journals.

J. Josphin Mary, received her M.sc. Computer science in 2013 from Valliammal College for Women. She is pursuing her M.Phil. Computer Science under the supervision of Mrs. R. Jayanthi, MCA., M.Phil., Assistant Professor in PG & Research Department of Computer Science at Quaid-E-Millath government College for Women, Affiliated to university of Madras. She has presented papers in International conferences and published papers in International Journals. Her area of interest is Text Mining, Cryptography and Network Security.

