Crowdsourcing Based on Clustering

Sheetal Jadhav and M.R.Patil

Smt.Kashibai Navale College of Engineering, Vadgaon ,Pune -411041

Summary
Crowdsourcing is an act of outsourcing tasks, traditionally performed by an employee or contractor, which are now performed by a large group of people. Recent survey deals with the problem of evaluating the submissions to crowdsourcing websites on which data is increasing rapidly in both volume and complexity. Thus, with an increasing number of submissions, the process of rate submissions, select winners and adjust monetary rewards is getting more complex, time consuming and hence more expensive. To overcome this problem text mining methodology can be used which consist of series of operations like Data extraction, pre-processing tf-idf calculation, calculating similarity and clustering and finally cluster submission. But results using these operations in existing methodology also show that this aspect does not do the entire trick of evaluating submissions. Hence propose methodology uses classification using Apriori algorithm which will find relations of clustered terms.

Key words:
Crowdsourcing, Clustering, Information Retrieval, Classification, Pre-processing.

1. Introduction

the phrase "Garbage in, Garbage out", is particularly applicable to data mining. Data gathering method are often loosely controlled, resulting in out of range values (e.g. income-100), impossible data combination (e.g. Gender: male, Pregnant: Yes), Missing values etc. Analyzing data that has not been carefully screened for such problems can reduce misleading results. Thus, presentation and quality of data is first and foremost before running an analysis.

1.1 Crowdsourcing

Crowdsourcing was introduced by Jeff Howe is a form of human computation that is the method of having people to do things that we might consider assigning the potential of large crowds of people by issuing open calls for contribution to particular tasks. A crowdsourcing activity contains a number of activities that involve resources within and beyond organizational boundaries, mostly human participants and information technology. The research for crowdsourcing comes from variety of fields such as computer science, management, and many other domains that have discovered crowdsourcing as an useful approach. But there is a problem in evaluating the submissions of crowdsourcing websites.

1.2 Clustering

Clustering can be examined the most significant unsupervised learning problem which deals with identifying a structure in a set of unlabeled data. So, the goal of clustering is to identify the essential grouping in a set of unlabeled data. With clustering we are using text mining process which is semi-automated process of extracting patterns from large amounts of unstructured data sources. Clustering technique works by transposing words and phrases in zigzag data structure, such as submissions to crowd-sourcing websites, into numerical values which can then be connected with structured and labeled data in a database and analyzed with data mining techniques. In this way, proposed system will overcome the problem of submission using clustering with text mining approach.

1.3 Information Retrieval

Information retrieval may be defined as selective, systematic recall of logically stored information. The main goal is to collect and organize information in one or more domain areas. The area of classic IR studies representation, storage and processing of text documents. Information retrieval (IR) is the study of finding information that matches their information needs. IR systems can be found everywhere such as search engines, library catalogue, shopping store catalogue etc.the information retrieval and text ming are closely related with each other as IR is distributed form of text mining. The main areas of applications in Information retrieval are voice extraction, image retrieval, music retrieval and document retrieval.

2. Related Work

Researchers have proposed related to the Crowdsourcing based on clustering. Thomas Walter and Andera Back presented a solution deals with problem of evaluating submissions to crowdsourcing websites ,text mining methodology for data extraction, Preprocessing, tf*idf calculation, Clustering. S.Subbaiah promoted probabilistic Classifier for Text mining using Preprocessing, rule

Manuscript received February 5, 2016
Manuscript revised February 20, 2016

generation, and Probability calculation. The proposed algorithm calculates the positive probability value and negative probability value for each term set or pattern identified from the document. Matthew Lease and Emine Yilmaz summarize and contextualize six novel research contributions at the intersection of information retrieval (IR) and crowdsourcing.

Table1: Evaluation of Related Work

Sr.No	Algorithm	Description
1	Stemming Algorithm	used in preprocessing step to delete suffixes , reduces inflected words e.g. computer, computing and compute to compute.
2	Stop Word Cleaning Algorithm	partly manual process, searches text by a predefined list of stop words(e.g. the , is, at ,but) and deletes them from text.
3	Tokenization Algorithm	process of breaking a stream of text up into words, phrases or symbols
4	Tf and tf*idf algorithm	calculates frequencies of terms in all contest
5	Clustering Algorithm	clustered submissions per contest
6	Classification Algorithm	indexes the document to the concern group of cluster.

3. Methodology

Our proposed system consists of application which has been divided into five major modules; clustering preprocessing, Term Document matrix, Text mining, Submission selection

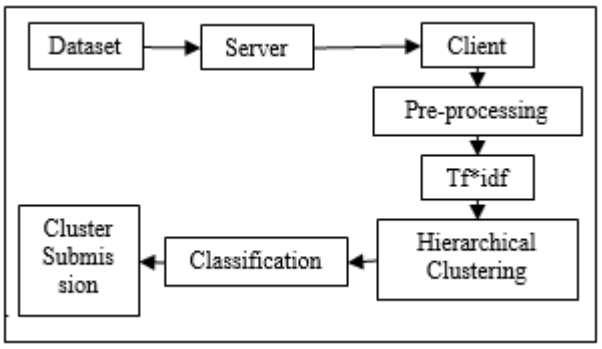

Fig1: General Architecture of Propose System

3.1 Preprocessing

Data pre-processing is often neglected but important step in data mining. Raw data is highly susceptible to noise, missing values and inconsistency. The quality of data affects the result. In order to improve the quality of data and mining results raw data is preprocessed. It is one of the most critical steps.

3.1.1 Stemming

Stemming is the process used to measure root or stem of the word. The process converts words to their stems that include language based semantic knowledge. It improves effectiveness and reduces the index size.

3.1.2 Stop Word Removal

This process is performed to reduce the indexing file size and to improve efficiency. Here stem file and stop word list file is given as an input. Those words are removed which are found in our original file from stop word list. The resultant file is obtained in which stop words are removed.

3.1.3 Tokenization

This process is performed to separate the words if any character comes in between them.

3.2 Term Document Matrix

It uses two weighting algorithms called tf and tf-idf. Algorithm takes number of documents as input which calculates tf value of every word in every document by using following formula:

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 * f(t,d)}{\mathrm{maxf(w,d):w \to d}} \qquad (1)$$

Algorithm also calculates idf value of every word in every document by using following formula:

$$\mathrm{idf}(t,d) = log + \frac{|D|}{|d:t \to d|} \qquad (2)$$

3.3 Hierarchical Clustering

In this algorithm, TF-IDF calculated Terms for different documents given as input which find outs similarity between terms. Algorithm puts similar terms which are found using average linkage criteria and average of similarity values between terms to form the clusters in one cluster and others in different clusters

3.4Apriori Algorithm

Clustered terms are given as input to Apriori algorithm. The terms are checked in our input documents and it forms the binary matrix. From this matrix Rule that is Relations are generated using Apriori algorithm.

3.5 Relation Submission

Output of Apriori algorithm will be relations.

4. Results

To evaluate the clustering approach, we measure its timing accuracy. Therefore we compare the two kind of clustering algorithms; k-means algorithm which is previously used and hierarchical algorithm which is proposed one. The intention is to prove efficiency of hierarchical clustering algorithm. We have analyzed the experimental results using following table 1 which shows the number of documents to be filtered by k-means algorithm and hierarchical algorithm respectively.

Table2: Experimental Results

		Clustering Algorithm	
		k-means(min)	Hierarchical(min)
Number of documents	20	15	7
	40	20	10
	60	25	12
	80	30	14
	100	35	16

If there are 20 documents in the input, results has been shown in above table that it requires 15 minutes to process using k-means and 7 minutes using hierarchical clustering. Hence, hierarchical clustering over comes the traditional k-means techniques.

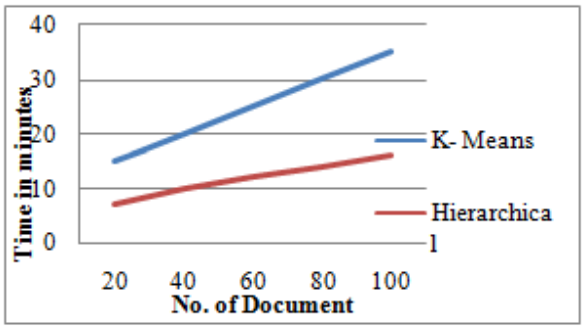

Fig2: Comparison of K-means and hierarchical clustering algorithm

5. Conclusion And Future Work

We used long existing Clustering algorithms and applied them on the modern research field of crowdsourcing contests. Our intention is to detect outstanding innovative ideas submitted by crowds due to their likelihood of using unique sets of words and hence separating them from a mass of so called noise. These show that Clustering can serve as an approach to detect outstanding ideas. In this way Data Mining Approach help to evaluate submission of crowdsourcing web contents and their quality using Clustering. Clustering could be used as decision support of expert committees as it provides fast and direct entrance to unique ideas. Concerning the rising problem of an increasing number of ideas, concepts or solutions being submitted by the crowd, Clustering could facilitate the current situation of which expert committees commonly are unable to cope with. On further work our clustering output can be used as a search engine. We can use different clustering algorithms in our application.

Acknowledgments

The author owes sincere thanks and deep sense of gratitude to reviewers for their helpful comments.

References

[1] J. C. Bongard, Member, IEEE, Paul D. H. Hines, Member, IEEE, Dylan Conger, Peter Hurd, and Zhenyu Lu," Crowdsourcing Predictors of Behavioral Outcomes", IEEETRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEM, VOL 43, NO. 1, JAN2013.

[2] Thomas Walter, Andera Back," A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests", 2013, 46th Hawaii International Conference on System Sciences.

[3] E. A.Calvillo1, A. Padilla, J. Munoz, J. Ponce, J. T. Fernandez," Searching Research Papers Using Clustering and Text Mining",2013, IEEE.

[4] S.Subbaiah," Extracting Knowledge using Probabilistic Classifier for Text Mining", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22.

[5] Matthew Lease ,Emine Yilmaz," Crowdsourcing for information retrieval: introduction to the special issue",Springer Science Business Media New York, March 2013.

[6] Kai Kuikkaniemi," White paper: Crowdsourcing in Media Industry".

[7] Ellen M. Voorhees," Variations in relevance judgments and the measurement of retrieval effectiveness", Information Processing and Management 36 (2000) 697-716, December 1999.

[8] Man-Ching Yuen, Irwin King and Kwong-Sak Leung," A Survey of Crowdsourcing Systems", 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social ComputinG.

Sheetal Prabhakar Jadhav received the B.E. degree in Computer Science and Engineering from Tatyasaheb Kore Institute of Engineering and Technology in 2011. During 2011-12, she was a lecturer with Computer Engineering Department at Polytechnic College in Mumbai University. Her research interests include big data, text mining and data mining.