Performance Analysis for Crowdsourcing Context Submission using Hierarchical Clustering Algorithm and Classification

S.P.Jadhav
PG Student, Department of Computer Engineering
Smt. Kashibai Navale College of Engineering, Pune 411041
Maharashtra, India

M.R.Patil
Assistant Professor, Department of Computer Engineering
Smt.Kashibai Navale College of Engineering, Pune 411041
Maharashtra, India

ABSTRACT
Very well know that the complexity and volume of the data is increasing rapidly in some Crowdsourcing websites. The term Crowdsourcing means the action of outsourcing tasks, traditionally performed by an employee or contractor, which are now performed by a large group of people. It is more expensive and more time consuming process because of increase in rate of submission and so short listing the winners. Data submitted by crowdsourcing websites can be noisy, inconsistent. To overcome this problems related to data one of the method was proposed which named as text mining method; this method performs the number of operations like extraction of data, preprocessing process, tf-idf calculation and calculation of similarity. Results obtained by existing system shows that k-means algorithm with text mining methods do not do the entire trick of evaluating submissions. Hence proposed system uses hierarchical clustering algorithm with text mining methods and classification for relation submission to overcome the problems present in the existing system.

General Terms
Crowdsourcing, Clustering, Information Retrieval (IR), Classification, Pre-processing.

Keywords
Apriori Algorithm, Clustering, Crowdsourcing, Hierarchical clustering, TDM (Term Document Matrix), IR (Information Retrieval).

1. INTRODUCTION
Crowdsourcing, text mining and information retrieval are closely related with each other. Crowdsourcing is widely used for information retrieval and information retrieval is distributed form of text mining.

1.1 Text Mining
Text mining is nothing but using data mining techniques for using useful patterns from unstructured data. In the data mining concept one phrase is very much applicable, which is nothing but "Garbage in, Garbage out". The data gathering method had some disadvantage like output data is out of range values, sometimes impossible data combination occurred, missing values and etc. By analyzing the data properly that has not been carefully screened can reduce the misleading results.

1.2 Crowdsourcing
The method of Crowdsourcing was introduced by Jeff Howe. Crowdsourcing can be defined as "an act of outsourcing task, traditionally performed by employee or contractor which are now performed by large group of people". The research of Crowdsourcing comes from variety of fields such as computer science, management, and many other domains that have discovered Crowdsourcing as a useful approach. There are some problems occurred during managing and maintaining the Crowdsourcing websites context.

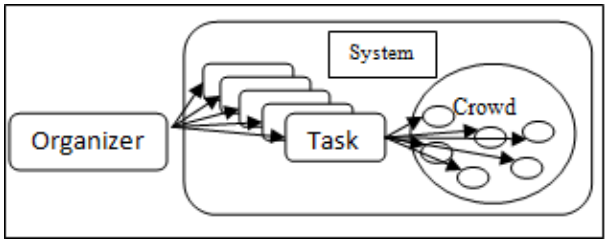

Fig 1: General Architecture of Crowdsourcing

Looking at above general architecture, it can be said that task is the work assigned to crowd. Crowd is the group of people who performs the task. Organizer can be either organization or individual who assigns the task to crowd and evaluates the results. System facilitates organizer to assign the task to crowd and to review the results.

1.3 Clustering
For identifying the structure in a set of unlabeled data the method of clustering is used. To identify the essential group in a set of unlabeled data is the main aim of the clustering. Proposed system uses text mining process with clustering to form the clusters of large amount of unstructured data. Clustering technique works for transposing words and phrases in zigzag data structure, such as submissions to crowd sourcing websites, into numerical values which can connect with the structured and labeled data in a database and analyze this data with the data mining techniques. Hence the proposed system overcomes the limitations or problem of submission by using the method of clustering with the text mining approach.

1.4 Information Retrieval

Recalling the selective and systematic, logically stored information defined as the information retrieval (IR). In the section of IR we are studied about representation, storage and processing of data. Gathering and organizing the information in one or more area is the main goal of this method. This systems found in many areas like in search engine, library catalogue, shopping store catalogue, etc. There are some applications of information retrieval, for which this method is used.

1.4.1 Voice Extraction

In this application non-regular queries are taken as input and output was produces by searching the best match and by using the automated speech retrieval, the speech is recorded as per the relevant query.

1.4.2 Image Retrieval

Image retrieval system and images, based on text or images that contain a given shape or color from large set of digital information

1.4.3 Music Retrieval

Music retrieval systems and a piece of melody or enter the notes of a musical theme.

1.4.4 Document Retrieval

Document retrieval is designed as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual.

2. RELATED WORK

Existing work has shown that the cluster formation of crowdsourcing website context need further enhancement [1], [2], [3], [4], [5], [6], [7]. In [1], behavioral model is implemented which leads into dynamically growing online data which is sometimes noisy, missing values etc. resulting into difficulty into submission of crowdsourcing website context. To overcome this problem, text mining methodology is used in [2]. Methodology consist of data extraction followed by stemming, stop word removal and tokenization, tfidf calculations and finally clustering using k-means algorithm. In [2], [3], system uses k-means algorithm for clustering. But this approach also has some shortcomings as k-means algorithm is a static algorithm. In [4], algorithm calculates probability value based on which probabilistic classifier indexes the document to the concern group of cluster using three steps preprocessing, rule generation and probability calculation. [5], [6], [7], summarizes and contextualizes relation between IR and crowdsourcing, overview of what crowdsourcing means, why it is important, core points while designing crowdsourcing mechanism and different aspects of crowdsourcing such as computational techniques and performance analysis.

Table 1. 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	Partially 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

Proposed system consists of five major modules as shown in following figure. Those five major modules are pre-processing, Term Document Matrix (TDM), Hierarchical Clustering, Classification and Relation Submission.

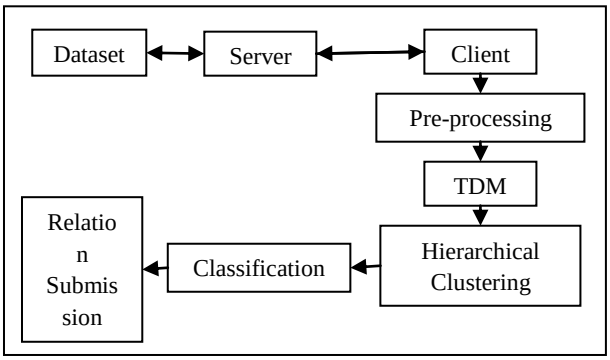

Fig 2: General Architecture of Propose System

Looking at figure of general architecture of proposed system it can very well say that client will extract dataset from server. Extracted dataset will go through five modules shown in architecture of proposed system

3.1 Preprocessing

In the process of data mining the data preprocessing is one of the important steps. The raw data occurred in the data is highly susceptible to noise, missing values and inconsistency. With all this raw data the quality of data as well as result is affected. Hence the data preprocessing is one of the method for maintaining and improving the quality of data. Preprocessing of data consists following three steps:

3.1.1 Stemming

Stemming process is used to measure the root of the words. By applying the semantic knowledge words are converted into stems. This process improves the effectiveness and reduces the index size.

3.1.2 Stop Word Removal

Main goal behind using stop word removal algorithm is to reduce the indexing file and improve efficiency. In this process file which is output of the stemming process (stemmed file) and file containing list of stop words are taken as an input. Words of stemmed file which are similar to stop word list get removed using this algorithm so that resulted file will be without stop words.

3.1.3 Tokenization

This process separates the words if any of the character comes in between them.

3.2 Term Document Matrix (TDM)

3.2.1 TF-IDF

TF-IDF calculation uses two weighting algorithms called tf and tf-idf algorithm.

Algorithm takes number of documents as input which calculates tf value of every word in every document by using following formula:

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d): w \in d}$$

…… (1)

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

$$idf(t,d) = log\frac{|D|}{|\{d \in D: t \in d\}|}$$

…… (2)

Where,

- $|D|$: Cardinality of D, or total number of document in the corpus.
- $|\{d \in D: t \in d\}|$: Number of documents where the term t appears i.e., tf (t, d) ≠0. If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the formula to appears 1+|{d ∈ D : t ∈ d}|.

Finally it calculates *TF*IDF* value by using following formula:

$$tfidf(t,d,D) = tf(t,d) * idf(t,d)$$

…… (3)

In this way *TF*IDF* value for every term in all documents is generates.

3.2.2 Cosine Similarity

$$\cos\sin(x,y) = \frac{\sum_i xiyi}{\sqrt{\sum_i xi^2}\sqrt{\sum_i yi^2}} = \frac{(x,y)}{||x||||y||}$$

…… (4)

- Where, Cos sin (x, y) = cosine Similarity between documents x and y.
- Where x = no. of documents $\{x_1,x_2,\ldots..,x_n\}$
- y = no. of documents $\{y_1,y_2,\ldots..,y_n\}$
- ||x|| = norm of matrix of document x.
- ||y|| = norm of matrix of document y.

3.3 Hierarchical Clustering

Hierarchical clustering algorithm, takes *TF-IDF* calculated terms as input which find outs similarity between terms. Algorithm puts similar terms which are found using average linkage criteria and average of similar values between terms to form the clusters in one cluster and others in different clusters.

Average linkage clustering can be done using following formula:

$$D(X,Y) = \frac{1}{N_X \times N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} d(x_i, y_j)$$

$$x_i \in X, y_j \in Y$$

…… (5)

Where,

$d(x,y)$ *is the distance between objects* $x \in X$ *and* $y \in Y$;
X *and* Y *are two sets of objects*(*clusters*);
N_X *and* N_Y *are the numbers of objects in clusters* X *and* Y *respectively.*

3.4 Apriori Algorithm

Clustered terms are given as input to Apriori algorithm. The terms are checked in input documents and it forms the binary matrix.

3.5 Relation Submission

Using Apriori algorithm resulted matrix generates rules which are relations.

3.6 Case Study

Let's consider three data sets each containing single text file for the calculation of *tf*idf* and implementation of Apriori algorithm for relation submission.

Dataset 1: the game of life is a game of everlasting learning

Dataset 2: the unexamined life is not worth living

Dataset 3: never stop learning

TF FOR DATASET 1

Words	Term frequency (*tf*)
The	0.1
Game	0.2
Of	**0.2**
Life	**0.1**
Is	**0.1**
A	**0.1**
Learning	**0.1**

TF FOR DATASET2

Words	Term frequency (*tf*)
The	0.142857
Unexamined	0.142857
Life	0.142857
Is	0.142857
Not	0.142857
Worth	0.142857
Living	0.142857

TF FOR DATASET 3

Words	Term frequency (*tf*)
Never	0.333333
Stop	0.333333
Learning	0.333333
Terms	**Idf**
Game	1.098726209
Of	1.098726209
A	1.098726209
The	0.405507153
Unexamined	1.098726209
Life	0.405507153
Is	0.405507153
Not	1.098726209
Worth	1.098726209
Living	1.098726209
Never	1.098726209
Stop	1.098726209
learning	0.405507153

Above table does not contain few words of dataset as during experiment few words removed in stemming, stop word removal and tokenization phase.

TF*IDF VALUES FOR EVERY TERM FOR EACH DATA SET

Sr. No.	Sr. No	Sr. No	Words	Dataset1	Dataset1	Dataset1
1	14	27	The	0.04055	0.0579	0
2	15	28	unexamined	0	0.1569	0
3	16	29	Life	0.04055	0.0579	0
4	17	30	Is	0.04055	0.0579	0
5	18	31	Not	0	0.1571	0
6	19	32	Worth	0	0.1569	0
7	20	33	Living	0	0.1569	0
8	21	34	Never	0	0	0.3662
9	22	35	Stop	0	0	0.3662
10	23	36	Learning	0.0405	0	0.1351
11	24	37	Game	0.2197	0	0
12	25	38	Of	0.2197	0	0
13	26	39	a	0.1098	0	0

APRIORI ALGORITHM AND ASSOCIATION RULE

Support count= occurrence of term/total transactions

Sr. no	Sr. no	Sr. no.	Words	occurrence	Support count
1	15	29	The	2	0.66
2	16	30	Unexamined	1	0.33
3	17	31	Life	2	0.66
4	18	32	Is	2	0.66
5	19	33	Not	1	0.33
6	20	34	Worth	1	0.33
7	21	35	Living	1	0.33
8	22	36	Never	2	0.66

9	23	37	Stop	1	0.33
10	24	38	Learning	2	0.66
11	25	39	Game	2	0.66
12	26	40	Of	2	0.66
13	27	41	A	1	0.33

Here minimum support count is considered as 0.6 so the words having support count less than 0.6 are removed in the process of Apriori algorithm which are indicated by highlighted rows.

Terms having minimum support count ≥ 0.6 are used as input for Apriori algorithm to find out most frequent item set followed by relation submission i.e. rule generation.

All the calculations are performed using formulae mention in methodology section.

4. RESULTS

System evaluates the performance of clustering approach by measuring the accuracy in time. Results are obtained by comparing two algorithms hierarchical algorithm which is used in proposed system and k-means algorithm used in existing system. Following table helps to analyze the result. Following table represents number of documents which are filtered and time required to form clusters using both algorithms k-means and hierarchical respectively.

Table 2: Comparison between existing and proposed system

		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

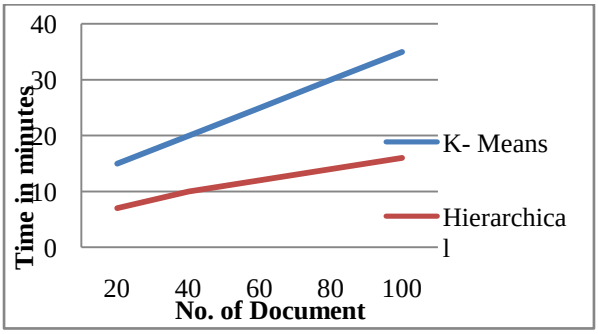

Fig 3: Comparison of K-means and hierarchical clustering algorithm

5. CONCLUSION AND FUTURE WORK

In proposed scheme, hierarchical clustering algorithm and text mining techniques are applied on modern research field of Crowdsourcing. Main motto is to detect outstanding innovative ideas which are submitted by crowd due to their likelihood of using the unique set of words and separating these words from the noise. 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. Clustering could facilitate the current situation of which expert committees commonly are unable to cope with. On further work clustering output can be used as a search engine. Proposed work can use different clustering algorithms.

6. ACKNOWLEDGMENT

For giving the helpful comments the author of the paper owe the thanks and sense of gratitude to reviewers.

7. 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 Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[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", 2010.

[7] 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.