Analysis Of Crowdsourced Data Using Data Mining Approach For Decision Support

Miss. Mayura Nalawade
Department of Information Technology
MIT College of Engineering
Pune, India
mayura.nalawade@gmail.com

Prof. B.A.Dixit
Department of Information Technology
MIT College of Engineering
Pune, India
dixit.bharati@gmail.com

Abstract— **Over the previous decade, crowdsourcing has developed as a significant problem solving and information gathering standard on the Internet. Crowdsourced data from blogs, forums and social media have become important source of information and people's feedback and opinion about almost every daily topic. It is well accepted that when information from different sources is processed and analyzed together, increases the potential of gaining meaningful insights to a large extent. This paper focuses on gathering the reviews of different e-commerce websites based on different performance parameters. This unstructured crowdsourced data is processed and analyzed in order to remove redundancy and make it more representative to help the analyst in decision making.**

Keywords-***Analytics; Crowdsourcing; Multiviewpoint- Based Similarity; TF-IDF.***

I. INTRODUCTION

The term crowdsourcing was introduced in 2006 by Jeff Howe. It is a combination of two words "crowd" and "sourcing". It can be defined as online production model that has emerged in recent years. Crowdsourcing depends on web 2.0 technologies[1]. It works across the boundaries of the organization and companies by outsourcing tasks that are traditionally performed by an employee or contractor.

Crowdsourcing is basically about sharing of knowledge[4]. The reason for increased popularity of crowdsourcing is its three features namely low cost, increased efficiency and globalization of labor[14]. Wikipedia, YahooAnswers, Blogs, Forums, Social networking sites etc that contain user generated content are good examples of productive crowdsourcing [8].

However, crowdsourced data can hardly be used directly to yield usable information. Crowdsourced data requires preprocessing and intelligent analysis so that useful information can be yielded from it [8]. Researchers and businesses are trying hard to increase the efficiencies and to maximize gains from crowdsourced data. Blogs, Forums, Social networking sites etc have become the most important source of news and people feedback and opinion about almost every daily topic [2]. This data is now being hugely used for business analytics.

With this massive amount of information over the web from different social networks and Blogs, etc, there has to be an automatic tool that can determine what people are talking about and how it has an impact on the business and performance of different companies [2]. Firms may commonly apply analytics to business data, to describe, predict and improve business performance. Analytics is discovery and communication of meaningful patterns in data.

The goal of this paper is to pre-process and analyze crowdsourced data using data mining techniques to improve its usability as information, to find patterns in data and to help analyst for decision making and planning. The crowdsourced data is collected in form of reviews from different sources about different e-commerce websites like Flipkart, Amazon, Myntra and Snapdeal. The reviews are collected based on different parameters like delivery, returns, customer care and refund. Fake reviews are identified and eliminated. Pre-processing is performed and using cosine similarity similarities are calculated. Multiviewpoint clustering is applied to remove the redundancy in the reviews. A main method to the clustering problem is to treat it as associate degree optimization method. An optimal partition is found by optimizing a particular function of similarity (or distance) among data. Efficiency of clustering algorithms for this approach depends on the appropriateness of the similarity measure to the data at hand [7].

The results in form of graphs are provided which help in business analysis and decision making.

II. RELATED WORK

The crowdsourcing process is now being used in many different fields across the world. Most of the work is now being focused to improve the quality of crowdsourced data and analyze it in a way so that it can be more beneficial in the applied context.

Authors have used social networks as source of news and people's feedback. The Arabic hot topics were collected from twitter. Clustering was based on unigram words which occurred more than 20 times. These words were used as features for clustering using bisecting K means clustering algorithm. Entropy and purity were calculated. 72.5% was the score recorded for the quality of the generated topic [2].

This work deals with problem of evaluating the submissions to crowdsourcing website, using text mining, similarity measures and k-means clustering algorithm. The goal was to provide decision support to the expert committees' process of analyzing and evaluation submissions to crowdsourcing websites [3].

This approach used crowd assisted strategies instead of automated data mining for analyzing crowdsourced data. Analyst took help of people called crowd workers who were asked to examine the views or explanations from the crowd. To detect redundancy in explanations clustering with representative selection was used. To check the provenance of explanations highlighting tasks were introduced. Embedded web browser was used to capture workers browsing history. The analyst was provided with an explanation management interface to help him with sorting and filtering of responses [12].

The author contributed seven different methods for improving the standard and variety of worker-generated explanations. The experiments show that using (S1) feature-oriented prompts, providing (S2) better illustrations, and together with (S3) reference gathering, (S4) chart reading, and (S5) annotation subtasks will increase the standard of replies by 28 % for US staff and 916% for non- US workers. Feature-oriented prompts increase explanation quality by 69 % to 236% dependent on the prompt. Author also shows that (S6) pre-annotating charts will focus employee's attention on appropriate details, and determine that (S7) generating explanations iteratively will increase description diversity without increasing employee attrition. Author utilized proposed techniques to generate 910 explanations for sixteen datasets, and located that 63% were of high quality. These results demonstrated that paid crowd staff will dependably generate numerous, high-quality explanations that support the analysis of particular datasets [11].

III. IMPLEMENTATION DETAILS

A. System Overview

In the proposed system the reviews for e-commerce websites like Flipkart.com, Amazon.com etc will be extracted. These reviews will be extracted based on different parameters like delivery, refund, customer care and exchange. The reviews will be filtered for spam detection based on presence of bold words, presence of question in the review, product comparison, rating deviation. The filtered reviews are then pre processed. These processes improve the quality of data initially. TF-IDF calculation was used for calculating number of times a word appeared in document. We calculate cosine similarity for finding similar reviews. The Multiviewpoint Based similarity algorithm is used for clustering similar reviews [7]. We used J48 classifier for identifying real and fake reviews.

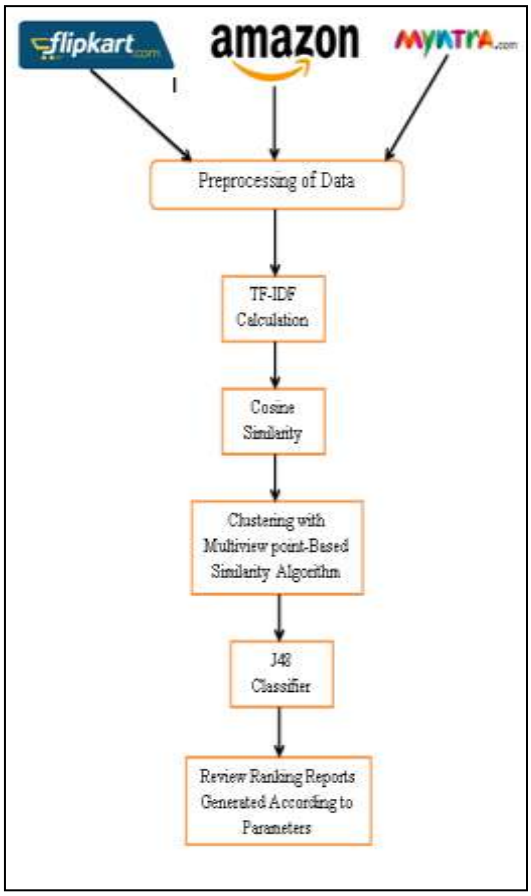

Figure1: System Architecture

B. Mathematical Model for Proposed Work

The system is represented as:

$$S = \{D,P,T,X,C\} \quad (1)$$

Let, S is a set of all variables used in system development.

A. Data extraction process

$D = \{d_1, d_2, d_3\ldots\}$;
where, D is the main set of documents like d_1; d_2; d_3:::

B. Preprocessing

$P = \{p_1, p_2, p_3, p_4\}$; where, p is the main set of steps like p_1, p_2, p_3, p_4
p_1 = Stemming of documents
p_2 = Stop word Removal Tokenization
p_3 = Tf-idf calculation
p_4 = hierarchical clustering

C. tf-idf
$T = \{T_1, T_2, T_3\ldots\}$;

Where, T is file of tf-idf calculated terms for the document like T_1, T_2, T_3…

i. It first calculates tf value of every word in every document by following formula:

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

tf(t,d) : tf value of term in a documents (d_1, d_2, d_3, d_n).

ii. Then it calculates idf value of every word in every document by following formula:

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

Idf(t,d) : idf value of term in a document d.

|D|: cardinality of D, or the total number of documents (d_1, d_2, d_3, d_n)
|d : t, d|: No. of documents (d_1, d_2, d_3, d_n) where term t is present.

Finally, it calculates tf×idf value by following formula:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

tf(t,d) = term frequency for no. of terms t = {t_1, t_2, t_n} from no. of documents d = {d_1, d_2, d_3, d_n}.
idf(t,d) = inverse document frequency for no. of terms t = {t_1, t_2, t_n} from no. of documents d = {d_1, d_2, d_3, d_n}.

D. Cosine similarity
X = {x_1, x_2,…x_n}; where, X is main set of documents containing cosine values.

$$CosSim(x,y) = \frac{\sum XiYi}{\sqrt{\sum X_i^2}\sqrt{\sum Y_I^2}}$$

Where, x = no. of documents {x_1, x_2,…x_n} and y = no. of documents {y_1, y_2, …y_n}

C. Algorithm

Multiviewpoint-Based Similarity[7]
1.procedure BUILDMVSMATRIX
2. for r←1:c do
3.Ds\Sr←$\sum_{di \epsilon Sr}$di
4.n_S\Sr ←|S\Sr|
5. end for
6. for i←1:n do
7. r←class of d_i
8. for j ←1:n do
9. if dj ε S_r then
10. a_{ij} ←$d_i^t d_j$-$d_{i\ ns\backslash Sr}^{t\ Ds\backslash Sr}$-$d_{i\ ns\backslash Sr}^{t\ Ds\backslash Sr}$+1
11. else
12. a_{ij} ←$d_i^t d_j$-$d_{i\ ns\backslash Sr-1}^{t\ Ds\backslash Sr-dj}$-$d_{i\ ns\backslash Sr-1}^{t\ Ds\backslash Sr-dj}$+1
13.end if
14. end for
15. end for
16. return A= {a_{ij}} n_xn
17.end procedure.

D. Experimental Setup

The system is built using Java framework (version jdk 6) on Windows platform. The Netbeans (version 7.4) is used as a development tool. The system doesn't require any specific hardware to run, any standard machine is capable of running the application.

IV. RESULTS AND DISCUSSION

A. Dataset

80 Reviews were extracted from different sources for Flipkart, Amazon, Snapdeal and Myntra.

B. Results

Table1 Cosine similarity between four documents.

	Doc 0	Doc 1	Doc 2	Doc 3
Doc 0	1	0.001204	0.014298	0.012322
Doc 1	0.001204	1	0.046292	0.00659
Doc 2	0.014298	0.046292	1	0.001017
Doc 3	0.012322	0.00659	0.001017	1

After extracting the reviews the tf-idf were calculated. Cosine Similarity is a measure of similarity between two documents that ranges from 0 to 1. Cosine Similarity was used to calculate similarity between documents. Here each document is a single review. Table 1 shows results for only 4 documents, in the same way similarity was calculated for 80 reviews.

The reviews were filtered using J48 classifier. Elimination of fake reviews further improved the quality of data. Table 2 depicts that out of 80 reviews collected 29 were fake while and 51 were truthful.

Table 2 Distribution of fake and truthful reviews.

Truthful	Fake
51	29

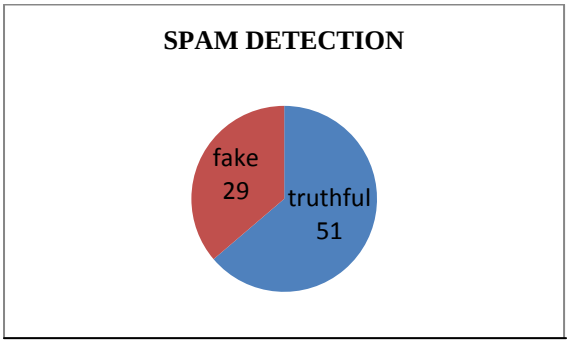

Figure2 Pie chart showing distribution of reviews

Figure 2 shows the distribution of reviews. The reviews collected were filtered on the basis of presence of bold text, comparison between products, presence of question in the review and rating deviation. These parameters helped to remove the fake reviews and hence improve the quality of crowdsourced data.

Table 3 Number of on time delivery reviews for different e-commerce websites.

Year	Snapdeal	Amazon	Myntra	Flipkart
2010	34	40	14	55
2011	20	15	35	41
2012	32	40	23	44
2013	16	21	43	22
2014	22	28	44	71
Average	24.80	28.80	31.80	46.16

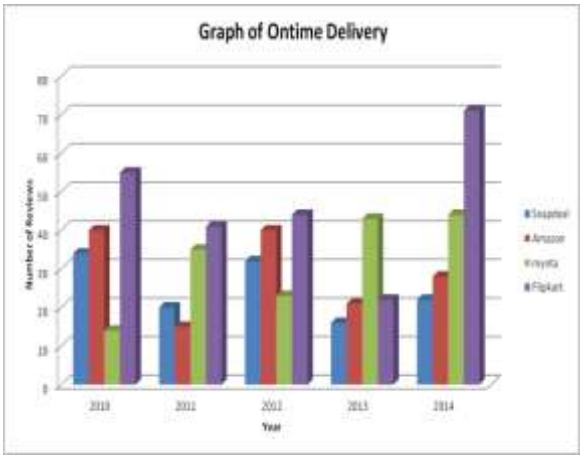

Figure 3: Comparative results of On time delivery attribute.

Figure 3 shown above gives a comparative view of 4 shopping websites based on the delivery attribute. For every 80 reviews collected for each site average reviews for delivery are shown. This graph can be plotted for a single company across different time period to judge its overall performance over the year.

V. CONCLUSION

There is a lot of scope to improve the quality of crowdsourced data. This approach to analyze reviews with help of data mining technique is an attempt to improve the usability of crowdsourced data. From business and marketing point of view this analytics is very crucial. Such kind of analysis is becoming very important with the growth in social media. Preprocessing and clustering these reviews remove the unwanted and redundant data respectively. The average on time delivery for 5 years shown in table 3 depicts that Flipkart had better delivery service than other three. This technique is useful for generating reports and will help in decision making for the analyst.

REFERENCES

[1] Adam Westerski et al "Idea Relationship Analysis in Open Innovation Crowdsourcing Systems" 8th International Conference Conference on Collaborative Computing United States , October 14-17,2012.

[2] Ahmed Rafea, Nada A. Mostafa "Topic Extraction in Social Media", IEEE Computer Society, 2013

[3] Andrea Back, Thomas P. Walter, "A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests", IEEE Computer Society , 2013

[4] Cindy Puah, Ahmad Zaki Abu Bakar et al "Strategies for Community Based Crowdsourcing".

[5] Daren C. Brabham "Crowdsourcing as a Model for Problem Solving" Convergence: The International Journal of Research into New Media Technologies Vol 14(1): 75–90, 2008.

[6] Demetrios Zeinalipour-Yazti et al,"Crowdsourced Trace Similarity with Smartphones", IEEE Ttractions on Knowledge and Data Engineering, Vol.25, No.6, June 2013.

[7] Duc Thang Nguyen, Lihui Chen et al "Clustering with Multiviewpoint-Based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, Vol.24, NO.6, June 2012.

[8] Geoffry Barbier, Reza Zafarani et al "Maximizing benefits from crowdsourced data" Springer 2012.

[9] Lipika Dey, Ishan Verma et al "A Framework to Integrate Unstructured and Structured Data for Enterprise Analytics"

Nicholas Kong, Marti A. Hearst et al, "Extracting References Between Text and Charts via Crowdsourcing" CHI 2014, One of a CHInd, Toronto, ON, Canada.

[10] Paul Andr´e, Aniket Kittur et al "Crowd Synthesis: Extracting Categories and Clusters from Complex Data" CSCW 2014 Promoting Participation and Engagement February 15-19, 2014, Baltimore, MD, USA

[11] Wesley Willett, Jeffrey Heer et al "Strategies for Crowdsourcing Social Data Analysis" CHI, USA 2012, May 5-10, 2012.

[12] Wesley Willett, Shiry Ginosar et al "Identifying Redundancy and Exposing Provenance in Crowdsourced Data Analysis", IEEE Transactions on Visualization and Computer Graphics, vol 19, no 12, December 2013.

[13] Yuxiang Zhao & Qinghua Zhu et al "Evaluation on crowdsourcing research: Current status and future direction" Inf Syst Front (2014) 16:417–434 Springer Science+Business Media, LLC 2012

[14] Zhang Li, ZHANG HONGJUAN "RESearch of Crowdsourcing Model Based on Case Study" Supported by the State Natural Science Foundation IEEE, 2011.

[15] Zheng-Jun Zha et al "Product Aspect Ranking and Its Applications" IEEE Transactions on Knowledge And Data Engineering, Vol. 26, No. 5, May 2014.