Implementation of Co-Occurring Phrase based Text Mining Technique for Analyzing Review Result of Product

Mr. Vijay M. Shekhat
Computer Engineering Department
Noble Group of Institutions (GTU),
Junagadh, India

Prof. Chetan R. Chauhan
Computer Engineering Department
Noble Group of Institutions (GTU),
Junagadh, India

***Abstract*-- Customer satisfaction is important in this competitive environment manufacturer need to regularly upgrade their product and for they need to fulfil the customers need. Review is very good method from where we can identify customer's view about product and what kind of problem they facing in using particular product. If we will not take care of this then customer will shifted towards other alternatives available in the market. But along with internet usage number of reviews is increasing every day and for that manually analysis is not effective in both the context time and cost. So it is need of some method which automatically does this work and provides us with required output. So for solving this problem we select this thesis title which will helpful to society.**

This thesis focus on co-occurring phrase based technique to analysis customer review of product in which we use method of keyword based filtering in first phase so that noisy reviews are filter out and save further computation.

Then in second phase we propose phrase based architecture which will applied on output of first phase and further filter reviews based on negative sets of keywords.

Finally we use supervised learning approach to update the sets of keywords so that we will improve system continuously.

Keywords— Text Mining, Phrase Based Technique

1. INTRODUCTION

As customer satisfaction is important in this competitive environment manufacturer need to regularly upgrade their product and for they need to fulfil the customers need. Review is very good method from where we can identify customer's view about product and what kind of problem they facing in using particular product. If we will not take care of this then customer will shifted towards other alternatives available in the market. But along with internet usage number of reviews is increasing every day and for that manually analysis is not effective in both the context time and cost. So it is need of some method which automatically does this work and provides us with required output. So for solving this problem we select this thesis title which will helpful to society.

Also it will help full for consumer which is searching for the right product. He/she can go through review analysis and decide whether that product is his interest or not.

Eighty percent of data in the world is at this time stored in unstructured textual format. Although some techniques such as Natural Language Processing (NLP), can achieve partial text analysis, there are at present no computer programs on hand to analyze and understand text for varied information extraction requirements fully.

Therefore text mining is a dynamic and emerging area. The world is rapidly becoming information demanding, in which particular information is being collected into very huge data sets. For example, Internet contains a huge amount of online text data. This data is quickly modified and rise.

It is not feasible to manually arrange such vast and quickly growing data. The requirement to extract meaningful and related information from such big data sets has led to an significant requirement to build up computationally efficient text mining algorithms.

An example problem is to automatically allocate text documents to predefined sets of categories depending on their content. Other examples of problems involving large data sets include searching for required information from scientific citation databases like MEDLINE, search, filter, and categorize web pages by subject and routing related email to the proper addresses.

PROBLEM OF ANALYSIS OF LARGE AMOUNT OF DATA

In early days human resource was utilized for finding useful information from data. But as we all know that every day huge amount of data are generated and shared in various organization and companies. Also by use of internet amount of text and other data are increasing regularly for example social networking, marketing, online shopping, emails etc.

This data are useful in one or other way for some work like decision making process, surveying, marketing etc. So we need to analyze that data. But doing that work manually is not feasible because it increase cost very drastically.

This requirement of analysis will encourage researchers to work on finding some way of automatically analysing this data.

PROBLEMS OF VARIOUS FORMATS OF DATA

As we know that each organization put their data in specific formats for example relational data base, files, records web page etc. When we want to combine data from various sources we can't use simple method of extraction of data. For that we need to find appropriate methods to handle various formats of data. So for solving this problem various knowledge discovery and data mining approaches are developed by different scientist and researchers.

KNOWLEDGE DISCOVERY FROM DATA (KDD)

Knowledge discovery is refers to process of extracting useful information from large set of database. With the increase importance of collecting data, there is need for different techniques to help researchers, analysts and decision makers in finding useful information from the quickly rising volumes of data.

In the past 50 years the concept of finding or discovering useful interesting patterns in data has been addressed by different research groups and individual researchers. We hope here to give a better idea of how KDD relates to these other approaches. Such approaches have been given different names, such as exploratory data analysis, information discovery, information harvesting, data archaeology, and data pattern recognition.

DATA MINING

Data mining is the process of discovering interesting knowledge, such as association, patterns and significant structures, from large amount of data stored in databases, data warehouses, or other information repositories. Due to the large amount of available data and urgency of finding information from that large data for many business and decision support this field is continuously growing day by day.

TEXT MINING

Text mining is a process to finding useful information from large set of unstructured information. We know that now a day's regularly information amount is increases, most of them are unstructured document which is maintain by each one as per their requirement. For example website data, journals, papers, email, product reviews etc.

Old method for retrieving information from data base is not applicable here because the unstructured data. So in past days this work is manually done by the human. But it increase cost of the work and as it is time consuming to find related information from large set of information. Person doing this job must go through whole data set and need to find required information one by one.

As time passes new technique of KDD and data mining is evolved this made good improvement in this area. This improvement will reduce time and cost of the work done manually. It encourages researchers to find more useful and better technique.

Text mining is also known as text data mining which picks better quality data from the text. It includes clustering, categorization, document summarization, analysing sentiment etc. Text mining is process which use set of algorithms for getting structured information from unstructured text and methods used to analyse this information.

There are large numbers of application of text mining for example automatic filtering spam mail using some keyword, such as message may discarded or send it to spam based on some repeating keywords. Another example is made survey using information available on the various resources.

2. DEALING WITH TEXT DATA

Dealing with text data is not straight forward task as like dealing of relational database. There for we need to develop such method which can handle such data.

There are various methods available to deal with text data developed by various researchers. These methods are commonly called text mining technique.

These techniques is basically divided into three categories

- Keyword based text mining technique.
- Phrase based text mining technique
- Pattern based text mining technique

KEYWORD BASED TEXT MINING TECHNIQUE

This technique used to find useful information from large set of text data by using some important keywords. These keywords are chosen very carefully based on the required information.

This technique will decide whether this text or part of text is of interest or not on the basis of frequency or the occurrence of key words.

PHRASE BASED TEXT MINING TECHNIQUE

It can be easily understand that group of keywords is more meaning full then single keyword. So in phrase based technique instead of analysing single keyword we use group of keywords together.

PATTERN BASED TEXT MINING TECHNIQUE

In pattern based technique some kinds of patterns are indentified which is formed by several terms. And based on that pattern it is decided that whether it is required document or not.

3. PROPOSED WORK BASED ON CO-OCCURRING PHRASE BASED TECHNIQUE

There are various methods are available for text mining which is performing well in one or more aspects but not properly handle all the aspects. Also many systems available are complex and it requires more CPU time so that many different researchers try to develop method for solving this problem.

Here our work is to develop system for analysing product review. Reviews are generally small paragraph containing few sentences. So many methods which works on frequency of occurring terms or pattern is might not work well in this area so we are trying to take this thing into account for finding efficient and easy solution for this problem.

Due to the small size we are considering phrase based technique is more efficient for this as we generally not get big common patterns in the reviews. So we will try to do this task by co-occurring phrase based technique.

Our whole work is divided into three phase in first phase we just using keyword based algorithm to filter reviews which is not of our interest. So that that review will separated and no need to further west computation power. And in second phase we will use co-occurring terms as a phrase and remove reviews which is not of our interest. And finally in the third phase we use supervised learning technique which updates sets of keywords so that next time we get better output.

PHASE - 1

As shown in figure 1 phase one is simple and straight forward in which we are going to filter those reviews which is not of our interest. Particularly we will filter reviews which not telling problem. This is done based on keyword based method in which if keyword is present then that review is of interest and if keyword is not present then review is not of our interest.

For keywords we are going to maintain two sets of term namely positive term set (PT), and negative term set (NT), we will store keyword based on which particular review is selected as interested review along with review as an attribute.

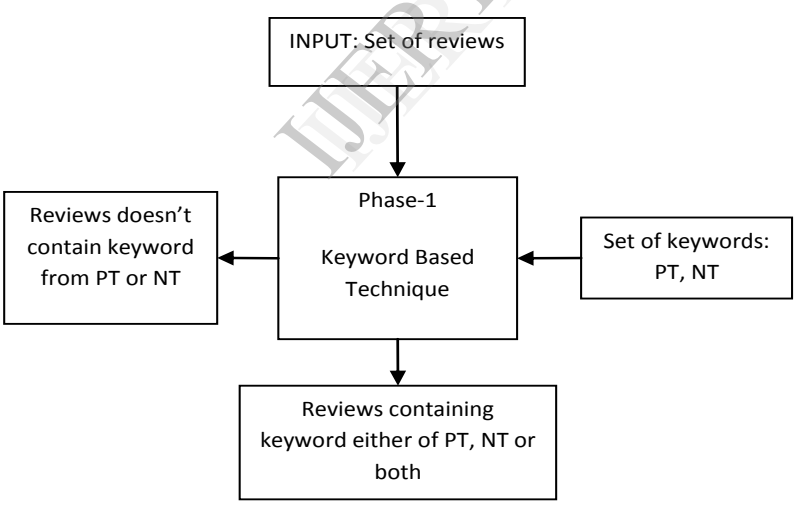

Fig. 1. Proposed Architecture of Phase – 1.

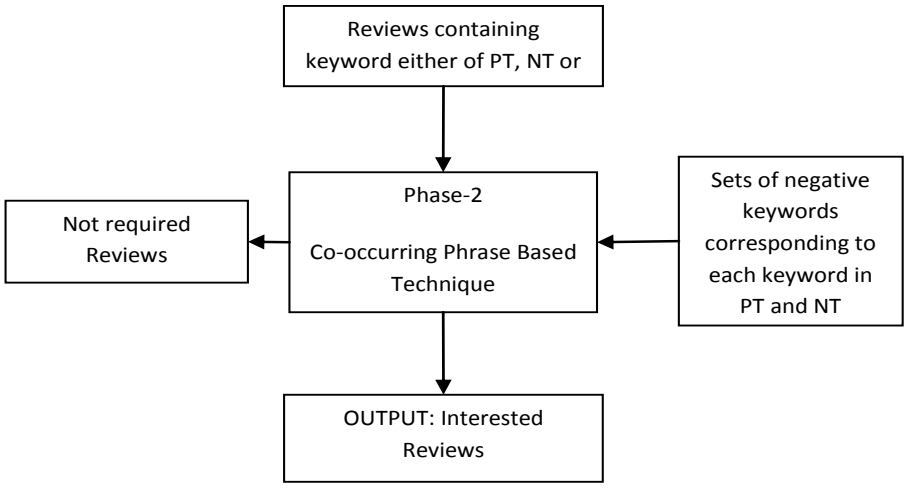

Fig. 2. Proposed Architecture of Phase – 2.

PHASE – 2

As shown in figure 2 in phase – 2 we are maintain sets of key words. One set for each keyword in PT and NT is maintain, that sets contains keywords which is negate the meaning of the keyword. Now based on saved keyword as attribute in phase – 1 we select corresponding set and find co-occurring phrase. If co–occurring term of stored keyword is present in corresponding set then we filter out that review from interesting reviews. And if we not found any co-occurring term which belongs to keyword's corresponding set than that review is consider as interested review. After properly finding interested and not interested review we calculate number of reviews which are problem telling out of total number of review and based on that we can give percentage ratings to that product.

PHASE – 3

As shown in figure 3 in phase – 3 we will apply supervised learning algorithm to update keywords. In this technique we use to take feedback from supervisor who is using system. If any review is found not interested in interesting set which is reported by user then we will provide them on which keywords it is selected as a interested review and ask them for keywords which negates the our selected keywords and update our sets of key words as per supervisor's guidance.

4. RESULT ANALYSIS

For result analysis we use F-measure which is the combination of Precision and Recall and Entropy which is decides how homogeneous a cluster is.

Precision is calculated as follows:

$$P = \text{Precision}(i.j) = \frac{N_{ij}}{N_j}$$

Recall is calculated as follows:

$$R = \text{Recall}(i.j) = \frac{N_{ij}}{N_i}$$

Where

N_{ij}: number of members of class i in cluster j

N_j: number of members of cluster j

N_i: number of members of class i

Now from precision and recall we calculate F-measure as:

$$F(i) = \frac{2PR}{P+R}$$

For overall F-measure we take weighted average of F-measure of each class.

For calculating Entropy we use standard formula:

$$E_j = -\sum_i P_{ij} \log(P_{ij})$$

For total entropy we take sum of entropy of each cluster.

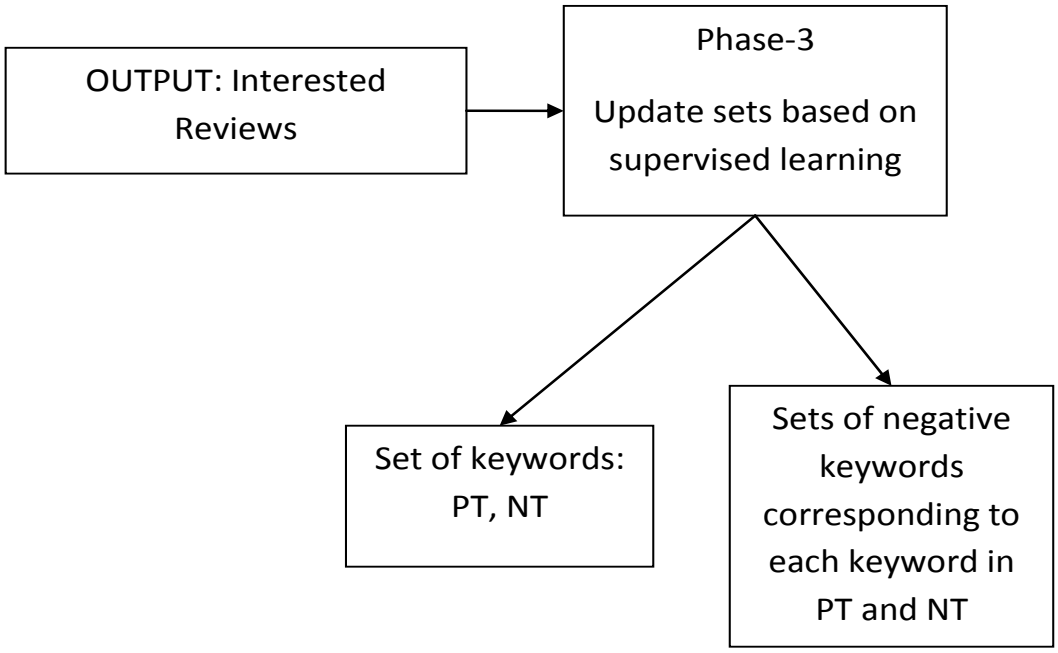

Fig. 3. Proposed Architecture of Phase – 3.

COMPARISON RESULT OF EXISTING SYSTEM WITH OUR SYSTEM

Here we compare our result with two technique which use DIG (Directed Graph) for finding phrases in the document. Here we increase F-measure and decrease.

TABLE I. COMPARISON OF PROPOSED SYSTEM WITH EXISTING SYSTEM.

	F-Measure	Entropy
Single Term Similarity	**0.709**	**0.351**
Combine Similarity	**0.904**	**0.103**
Proposed System	**0.924**	**0.101**

GRAPHICAL REPRESENTATION

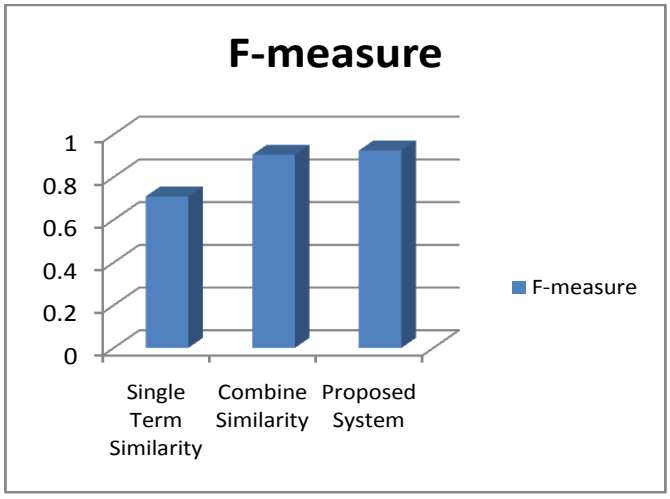

Fig. 4. Comparison chart for F-measure.

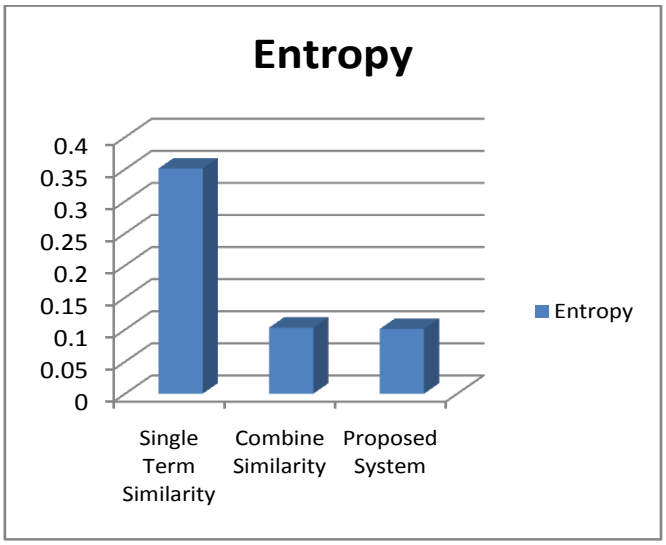

Fig. 5. Comparison chart for Entropy.

COMPARISON BASED ON SPACE REQUIREMENT

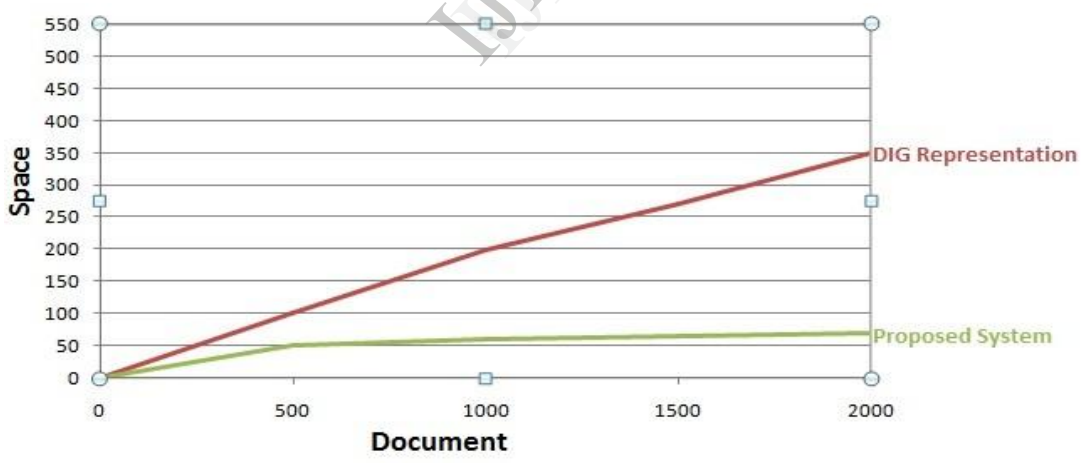

Fig. 6. Comparison chart based on space requirement.

REFERENCES

[1] B. T. Bartel, G. W. Cotterl, R. K. Belew, "Automatic Combination of Multiple Ranked Retrieval Systems",SIGIR, 1995
[2] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task", SIGIR, 1992
[3] David D. Lewis, "Evaluating and Optimizing Autonomous Text Classification Systems", SIGIR, 1995
[4] Dipti Shyam Charjan, "Review of Text Mining Method: Investigation and Analysis", IJACAE, 2013
[5] Firat Tekiner, Yoshimasa Tsuruoka, Jun'ichi Tsujii, Sophia Ananiadou, John Keane,"Parallel Text Mining for Large Text Processing", CSNDSP, 2010
[6] H. Ahonen-Myka, O. Heinonen, M. Klemettinen, and A. I. Verkamo," Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery", IJCAI, 1999
[7] H. Ahonen-Myka, O. Heinonen, M. Klemettinen, and A. I. Verkamo,"Mining in the phrasal frontier",PKDD, 1997
[8] Jorge Villalon, Rafael A. Calvo," A Decoupled Architecture for Scalability in Text Mining Applications", IUCS 2013
[9] Khaled M. Hammouda, and Mohamed S. Kamel, "Efficient Phrase-Based Document Indexing for Web Document Clustering", IEEE 2004
[10] M.Suganthy 1, K.Rupika 2 ,J. Sharmiliya Fransuva3, "text mining for pattern identification", IJFSET, 2013
[11] N. Menaga B. Hemapriya, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IJCTT, 2013
[12] N. Zhong, Y. Li, and S. Wu, "Effective Pattern Discovery for Text Mining", IEEE, 2012
[13] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, 2013
[14] Ronen feldman, ido dagan, haym hirsh, "Mining Text Using Keyword Distributions", JIIS, 1995
[15] Sheng-Tang Wu, Yuefeng Li, Yue Xu, "Deploying Approaches for Pattern Refinement in Text Mining",IEEE, 2006
[16] S. Ghosh, S. Roy, and S. K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms",IJARCCE, 2012
[17] S.Revathi, and Dr.T.Nalini, "Clustering and Classification Augmented with Semantic Similarity for Text mining", IJCSI, 2013
[18] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen. "Automatic pattern-taxonomy extraction for web mining" WI'04, 2004.
[19] V. Ramanathan, T. Meyyappan, "Survey of Text Mining", ICTBM 2013
[20] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, "Knowledge discovery in databases: an overview", AI Magazine 1992
[21] Xiaodong Ji," Multi-resolution Keyword Mining Algorithm based on Frequent Pattern Technique", Atlantis Press, 2010