QUALITY ASSESSMENT OF CUSTOMER REVIEWS EXTRACTED FROM WEB PAGES : A REVIEW CLUSTERING APPROACH

P.S Hiremath[1], Siddu P. Algur[2], S. Shivashankar[3]
[1] Gulbarga University, Dept. of P.G. Studies and Research in Computer Science, Gulbarga, Karnataka, India
[2] BVB College of Engg. and Tech., Department of Information Science and Engg., Hubli, Karnataka, India
[3] Karnatak Science College, Department of Computer Science, Dharwad, Karnataka, India

Abstract

The number of customer reviews that a product receives is growing at very fast rate. Customer reviews posted on the websites vary greatly in quality. In this paper, we make an attempt to assess a review based on its quality, to help the customer make a proper buying decision. The quality of customer review is assessed as most significant, more significant, significant and insignificant. A novel and effective web mining technique based on review clustering is proposed for assessing a customer review of a particular product. This is performed in two steps : (1) Cluster the reviews into four groups by applying k-means clustering technique and compute the cluster weights. (2) Assess quality of the given reviews and classify them by considering the cluster weights. Experimentation has been done using publicly available review databases for four different products. The results are analyzed and the efficacy of the proposed method has been demonstrated.

Index Terms: Customer reviews, Feature extraction, Feature weights, Cluster weights, Web mining, Clustering technique.

1. INTRODUCTION

Of late, the web has become an excellent source for posting customer reviews. The customers can now post reviews of products at merchant sites and express their views on almost everything. In the past few years, there has been an increasing interest in mining and assessing the customer reviews [1 - 3]. However, the customer reviews posted at online shopping sites vary greatly in quality. Thus, it is very essential to have a mechanism which is capable of assessing the quality of reviews for purchase decision or marketing intelligence. Identifying the quality of customer reviews is useful for both potential buyers and product manufacturers. For a potential buyer, it is more convenient and less time consuming to see at a glance feature by feature comparison of customer reviews. For a product manufacturer, it helps to find the strengths and weaknesses of his/her own products and also that of the competitors.

There are three main review formats commonly found on the web. Different review formats may need different techniques to identify and assess the quality of the reviews.

Format 1 : Pros and Cons
- The reviewer is asked to describe Pros and Cons separately. e.g., C|net.com uses this format.

Format 2 : Pros, Cons and detailed review
- The reviewer is asked to describe Pros and Cons separately and also write a detailed review. e.g., Epinions.com uses this format.

Format 3 : Free format
- The reviewer can write freely, i.e., no separation of Pros and Cons. e.g., Amazon.com uses this format.

The opinion orientations (positive or negative) of features are known from Format 1 and 2 because pros and cons are separated and thus there is no need to identify them.

In this paper, we propose a novel and an effective web mining technique for assessing the customer review of a particular product based on the review clustering. Given a product name and a set of URL's of web pages that contain customer reviews on the product, it works in two stages:

Stage1: Cluster the reviews into four groups by applying k-means clustering technique and compute the cluster weights.

Stage2: Assess quality of the given reviews and classify them by considering the cluster weights.

Experimental results show that the proposed technique can measure the quality of review and assess it accordingly. The efficiency of the task of customer review summarization can be enhanced by identifying and eliminating the insignificant reviews and thus retaining only significant ones.

The rest of the paper is organized as follows. The section 2 presents the related work. In the section 3, we present the proposed technique of quality assessment of customer review based on a clustering technique for assessing the customer reviews. The section 4 shows the experimental results. The section 5 gives the conclusion.

2. RELATED WORK

Lot of research has been done in text summarization and terminology identification. The authors Dejong [4], Tait [5] and Radev and McKeown [6] propose text summarization using template instantiation. This technique needs to design a template by identifying and extracting primary elements and facts in a document. Paice [7], Kupiec, Pederson and Chen [8], Hovy and Lin [9] have focused on text summarization using text extraction, which is based on representive sentences. Kan and McKeown [10] have proposed a combined approach by merging template instantiation and text extraction. Jacquemin and Bourigault [11], Justeson and Katz [12], Daille [13] and Church and Hanks [14] have focused on terminology identification using symbolic approach.

Many researchers are working on information extraction from texts. Their main focus is on machine learning and NLP methods for extraction or classification of entities and relations. Extending the same, the other area of research is opinion/review extraction from web pages and opinion summarization based on product features. Dave, Lawrence and Pennock [15] have proposed semantic classifier for product reviews, but it does not mine features of the product.

Liu et. al. [16] have proposed a technique to analyze customer reviews of Format 3. Their focus is on identifying the product reviews and summarizing by determining the orientation of each review. This technique is based on unsupervised item set mining. Further, this approach cannot be applied to reviews of Format 2 for obtaining accurate results, because a review contains short and incomplete sentences. Morinaga et al. [17] proposed a system to know the reputation of the product, but it does not focus on analysis of the reviews. Liu et al. [18] proposed a system to (i) compare customer reviews of many competing products, and (ii) identify product features from reviews. The technique is based on NLP and supervised pattern discovery. It identifies product features of reviews of Format 2 consisting of only pros and cons. They also provide a visualization system which can be applied to all review formats.

The major problem with existing studies on assessment of reviews is that they consider all reviews irrespective of the significance of each review. Hence, classification of reviews based on significance is an important task. Turney [19] proposed a system that classifies reviews as “thumbs up” for useful review or “thumbs down” for unuseful review by using an unsupervised learning algorithm. Pang et al. [2] proposed a supervised learning algorithm for the same problem.

In Kim et al. [20], a system for assessment of quality of reviews, is proposed using regression models. They derive ground-truth from user votes for helpfulness and then train the model and test it. Liu et al. [21] proved the biases present in the voting system and proposed a system for classification of review region by defining standard specification of quality of reviews. The extraction of customer reviews from web pages using visual clue based extraction procedure VSAP has been investigated by Hiremath et al. [22]. Further, Hiremath et al.[24] proposed a system to automatically extract and assess the quality of the review using quartile measure and to identify a customer review as Most Significant review, More Significant review, Significant review and Insignificant review.

3. PROPOSED TECHNIQUE

To assess and classify the customer reviews for a specific product, product features must be identified and extracted accurately from the Web pages. Since the product features are often domain dependent, it is desirable that the feature extraction system is as flexible as possible. We propose a novel and effective technique to extract the customer reviews from the web pages and classify them into different groups based on their quality using clustering technique. Any method of rating the reviews based on the helpful votes from the customers fails to provide a clear guideline for what a good review consists of [21]. We define four types of review qualities, which are determined by applying the k-means clustering technique [23]. The clustering algorithm partitions the features into four (k=4) clusters, i.e., groups such that the similarity within a group is larger than that among other groups. The four types of reviews are:

(i) Most Significant Review (MSR): It is the one which corresponds to the maximum cluster weight. A feature weight is defined as the sum of corresponding feature values in the reviews of a cluster divided by the total number of reviews in that cluster. A cluster weight is defined as the sum of all the feature weights of that cluster.

(ii) More Significant Review (MoSR): It is the one which corresponds to the next maximum cluster weight and whose value is less than the first one.

(iii) Significant Review (SR): It is the one which corresponds to the next maximum cluster weight and whose value is less than the second one.

(iv) Insignificant Review (ISR): It is the one which corresponds to all the remaining reviews and whose cluster weight is less than the third one.

The system model of the proposed technique, namely**,** Quality Assessment of Customer Reviews Extracted from Web Pages : A Review Clustering Approach ,is depicted in the Fig.1. It consists of the following components :

- Feature extraction
- Review matrix construction
- Review clustering and cluster weight computation.
- Review quality assessment and classification.

The output of each component is the input for the next component.

3.1. Feature Extraction

It extracts the features from the given set of reviews (pros and cons as separate reviews) extracted from the web pages with the format shown in the Table 1(a) and (b), respectively [16].

Review No.	Content of the review
R_1 (pros)	Great picture quality, price, great zoom ratio, nice control layout, nice LCD size
R_2 (pros)	Price, ease of use, nice quality photos, LCD screen, small size 2xZoom.

Table 1(a). The format of pros review database with sample records

Review No.	**Content of the review**
R_1 (cons)	Battery usage, included software could be improved, included 16MB memory is stingy, need extended warranty
R_2 (cons)	Unreliable, long delays between pictures. Bad interface which forces you to press OK between pictures.

Table 1(b). The format of cons review database with sample records

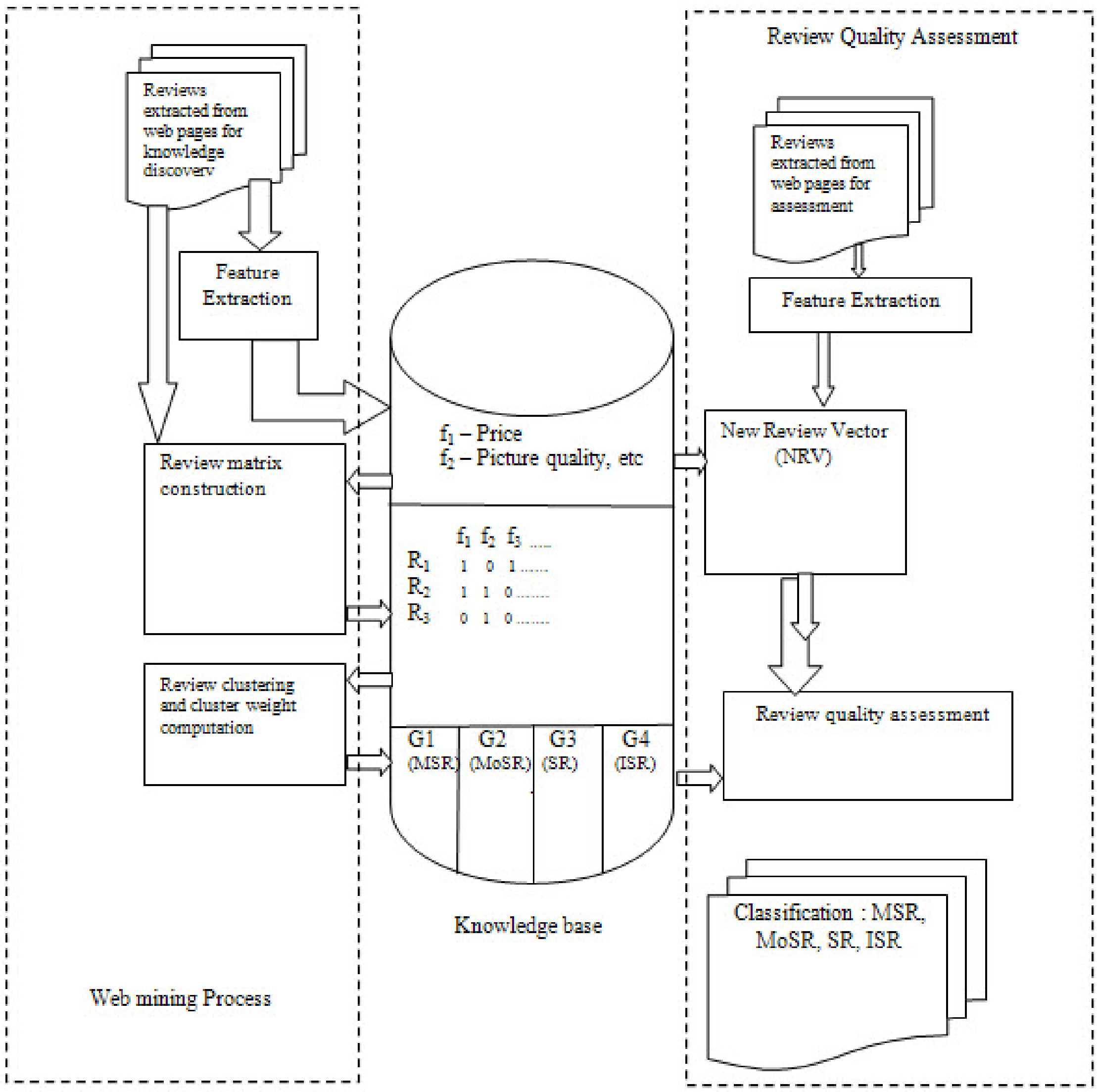

Fig 1. System model of proposed technique

In our study, we have used customer reviews on the product, namely, digital camera, which are crawled from the website, namely, epinion.com, as our dataset. The data set consists of one thousand two hundred (1200) reviews on five (5) types of digital cameras. Fifty percentage (50%) of these reviews are used for knowledge discovery and the remaining fifty percentage (50%) of the reviews are used for assessing the quality of review.

As discussed in the section 1, there are three common review formats. In our work, we focus on the reviews with Format 2. Due to the separation of pros and cons, there is no need to decide the orientation for reviews as discussed in [8]. The existing methods of [18] are used to extract the product features from the customer reviews with Format 2. In [18], the authors extract the product features from reviews with Format 2, using POS tagger and also check for grouping synonyms. The method in [21] makes use of edit distance to compare the similarities between the surface strings of two mentions, and uses contextual similarity to reflect the semantic similarity between the two mentions. Thus, the features extracted from the customer reviews are stored in the feature database after checking whether that feature is already existing or not. A sample of extracted features stored in the database is shown in the Table 2.

Feature number :	$\mathbf{f_1}$	$\mathbf{f_2}$	$\mathbf{f_3}$	$\mathbf{f_4}$	$\mathbf{f_5}$	$\mathbf{f_6}$	**-**	$\mathbf{f_{63}}$
Feature name :	Price	Picture Quality	Zoom	Speed	Battery life	Memory Card	--	warranty

Table 2. A sample of extracted features

3.2 Review matrix construction:

The inputs for this component are the set of raw reviews and the feature set extracted in the earlier step. Consider that there are a total of m customer reviews for a particular product and n features are extracted from each of the reviews. We construct a review matrix M of order of m x n using the Algorithm 1.

Algorithm 1 : Algorithm for review matrix construction.

For each review R_i in the raw review database
{
For each feature f_j in the review
{
If f_j is present in R_i then $M_{ij} = 1$
else $M_{ij} = 0$
}
}

A sample review matrix constructed for few reviews (illustrated in the Tables 1 and 2) using the Algorithm 1 is given in the Table 3.

	Price	**Picture Quality**	**Zoom**	**Speed**	**Battery life**	**Memory Card**	**-**	**warranty**
Review no	$\mathbf{f_1}$	$\mathbf{f_2}$	$\mathbf{f_3}$	$\mathbf{f_4}$	$\mathbf{f_5}$	$\mathbf{f_6}$		**fn**
R_1	1	1	1	0	0	0		
R_2	1	1	0	0	0	0		
R_3	0	0	0	0	1	1		
R_4	0	0	0	0	0	0		
R_5	1	0	0	0	0	0		
:								
R_m	-	-	-	-	-	-	-	-

Table 3. A sample review matrix

3.3 Review clustering and cluster weight computation

Now, we propose to group the reviews into four groups by applying a k-means clustering technique with k = 4 and absolute difference of two data values as the distance measure, for the data set of reviews present in review matrix. The input to this component is the review matrix constructed in the Algorithm 1. The algorithm for grouping reviews based on clustering technique is given in the Algorithm 2.

Algorithm 2 : Proposed algorithm for grouping of reviews by clustering

Step1 : Construct the review matrix M for the review set using Algorithm 1.

Step2 : Apply k-means clustering technique with k = 4 for the review set and obtain four clusters of reviews.

Step3 : For each cluster, compute cluster weight W_g, g=1 to 4, as shown below :

a) Compute feature wise sum of the reviews in g^{th} cluster, given by

$$Y_{gj} = \sum_{i=1}^{p} X_{ij} \text{, for j=1 to n}$$

where

n = number of features of the product

p = number of reviews of product in the g^{th} cluster

Y_{gj} = sum of j^{th} feature of all the reviews belonging to g^{th} cluster.

X = sub matrix of review matrix M for g^{th} cluster.

b) Compute the feature weight vector WV_g for g^{th} cluster given by

$$WV_g = (WV_{g1}, WV_{g2}, \ldots WV_{gn})$$

where $WV_{gj} = Y_{gj} / p$, for j = 1 to n, are the j^{th} feature weights for g^{th} cluster of reviews

c) Compute cluster weight W_g for g^{th} cluster given by

$$W_g = \sum_{j=1}^{n} WV_{gj}$$

Step 4: Label the clusters as G_1, G_2, G_3 and G_4 groups with decreasing order of their cluster weights W_g, g = 1, 2, 3, 4. The corresponding feature weight vectors WV_g, g = 1, 2, 3, 4, are the representative vectors of the clusters G_1, G_2, G_3 and G_4, respectively. The G_1, G_2, G_3 and G_4 contain MSRs, MoSRs, SRs and ISRs, respectively.

The feature weight vectors WV_g, g = 1, 2, 3, 4, are used for the quality assessment of reviews, which is described in the next section.

3.4. Review Quality Assessment

The third step of the proposed technique is to find out the group to which a given review belongs based on its quality. A review from the raw review database, and the feature weight vector of each group are the inputs for rating the review quality assessment. The algorithm for review quality assessment is given in the Algorithm 3.

Algorithm 3 : Algorithm for Review quality assessment.

1. Identify and extract the features appearing in the given review and store it in the New Review Vector (NRV).
2. Compute dot product $NRVS_g$ of NRV and WV_g, g = 1, 2, 3, 4, given by

$$NRVS_g = \sum_{j=1}^{n} (WV_{gj})(NRV_j)$$

3. Let $NRVS_{max}$ = max ($NRVS_1$, $NRVS_2$, $NRVS_3$, $NRVS_4$)
4. Determine the review quality assessment using the following criteria :
   If $NRVS_{max} = NRVS_1$, then the Review is Most Significant (MSR).
   If $NRVS_{max} = NRVS_2$, then the Review is More Significant (MoSR).
   If $NRVS_{max} = NRVS_3$, then the Review is Significant (SR).
   If $NRVS_{max} = NRVS_4$, then the Review is Insignificant (ISR).

4. EXPERIMENTAL RESULTS

For the purpose of experimentation, we apply the proposed technique to see how effective it is in assessing the quality of review from pros and cons in reviews of Format 2. We also show its effectiveness on the task of customer review summarization. The proposed cluster analysis of reviews based on the k-means technique is implemented and tested by taking reviews from the web pages and assessing them as most significant, more significant, significant, and insignificant review. The performance of this technique is compared with the quartile measure technique [24].

We consider 1200 customer reviews of digital camera, out of which 50% of the reviews are used as training set to evaluate the representative vectors by clustering the reviews into four types by employing the clustering technique, and the remaining 50% of the reviews are used to assess the review quality.

From the training set of reviews, we build the knowledgebase as given in the Table 4. The 63 features such as picture quality, price, battery, etc., could be identified. The second, third, fourth and fifth columns of the table show the weight vectors of groups G_1, G_2, G_3 and G_4, respectively. Each rows of these columns indicate the weights of the features belonging to that group, e.g the weights 0.3976, 1, 0.0723, 0.012, etc, of features f_1, f_2, f_3, f_4, etc, belong to group G_1, similarly the weights 0.5088, 0.0, 0.0614, 0.0175, etc, of features f_1, f_2, f_3, f_4, etc, the weights 1.0, 0.0, 0.0841, 0.0187, etc, of features f_1, f_2, f_3, f_4, etc and the weights 0.0, 0.0, 0.0743, 0.0338, etc of features f_1, f_2, f_3, f_4, etc, belong to groups G_2, G_3 and G_4, respectively. The last row of the Table 4 shows cluster weights W_g for each of the groups G_1, G_2, G_3 and G_4. These cluster weights, i.e., 2.5411, 2.4653, 2.2796 and 1.2876, are used to assess the reviews as MSR, MoSR, SR and ISR.

Features	G_1	G_2	G_3	G_4
f_1	0.3976	0.5088	1.0000	0.0000
f_2	1.0000	0.0000	0.0000	0.0000
f_3	0.0723	0.0614	0.0841	0.0743
f_4	0.012	0.0175	0.0187	0.0338
f_5	0.0723	0.0263	0.1402	0.2061
f_6	0.0602	0.1316	0.0467	0.1182
f_7	0.0000	0.0088	0.0093	0.0473
f_8	0.012	0.0263	0.028	0.0338
f_9	0.0361	0.0088	0.0748	0.0236
f_{10}	0.0000	0.0000	0.0000	0.0034
f_{11}	0.0361	0.0088	0.0093	0.027
f_{12}	0.0241	0.0351	0.0561	0.0642
f_{13}	0.0000	0.0439	0.0093	0.0236
f_{14}	0.3253	1.0000	0.0000	0.0000
f_{15}	0.0000	0.0000	0.0093	0.027
f_{16}	0.0000	0.0000	0.0187	0.0372
f_{17}	0.0000	0.0000	0.0000	0.0034
f_{18}	0.012	0.0000	0.0000	0.0068
f_{19}	0.0843	0.0965	0.0841	0.0709
f_{20}	0.0361	0.0439	0.0467	0.0507
f_{21}	0.0000	0.0439	0.0654	0.0608
f_{22}	0.012	0.0088	0.028	0.0068
f_{23}	0.0000	0.0000	0.0093	0.0034
f_{24}	0.0000	0.0088	0.0000	0.0000
f_{25}	0.012	0.0439	0.028	0.0068
f_{26}	0.012	0.0000	0.0000	0.0000
f_{27}	0.012	0.0088	0.028	0.0101
f_{28}	0.0000	0.0000	0.0187	0.0068
f_{29}	0.0361	0.0088	0.0093	0.0068
f_{30}	0.012	0.0088	0.028	0.0000
f_{31}	0.0000	0.0088	0.0187	0.0068
f_{32}	0.0000	0.0000	0.0000	0.0034

Features	G_1	G_2	G_3	G_4
f_{33}	0.0000	0.0088	0.0093	0.0034
f_{34}	0.0000	0.0000	0.0374	0.0068
f_{35}	0.0000	0.0000	0.0000	0.0034
f_{36}	0.0000	0.0175	0.0093	0.0169
f_{37}	0.0000	0.0000	0.0000	0.0034
f_{38}	0.0000	0.0000	0.0000	0.0034
f_{39}	0.0000	0.0088	0.0374	0.0101
f_{40}	0.0000	0.0000	0.0000	0.0034
f_{41}	0.0000	0.0175	0.0093	0.0000
f_{42}	0.012	0.0000	0.0000	0.0000
f_{43}	0.0000	0.0000	0.0093	0.0000
f_{44}	0.012	0.0000	0.0000	0.0000
f_{45}	0.0000	0.0000	0.0000	0.0034
f_{46}	0.0000	0.0000	0.028	0.0068
f_{47}	0.0361	0.0175	0.0654	0.0541
f_{48}	0.0000	0.0000	0.0000	0.0034
f_{49}	0.012	0.0175	0.0187	0.0203
f_{50}	0.0000	0.0088	0.0000	0.0000
f_{51}	0.0000	0.0088	0.0000	0.0000
f_{52}	0.0000	0.0000	0.0000	0.0034
f_{53}	0.0000	0.0000	0.0000	0.0034
f_{54}	0.0361	0.0088	0.0000	0.0405
f_{55}	0.012	0.0000	0.0000	0.0000
f_{56}	0.0723	0.1316	0.1402	0.0338
f_{57}	0.012	0.0000	0.0187	0.0169
f_{58}	0.012	0.0088	0.0093	0.0169
f_{59}	0.0241	0.0088	0.0000	0.0135
f_{60}	0.0000	0.0175	0.0093	0.0135
f_{61}	0.0000	0.0000	0.0000	0.0135
f_{62}	0.012	0.0263	0.0093	0.0203
f_{63}	0.012	0.0000	0.0000	0.0101
WV_g	2.5411	2.4653	2.2796	1.2876

Table 4. The knowledgebase of the groupings and weights of the features for digital camera, containing representative vectors for each group, obtained by the proposed method.

For a given review of digital camera, e.g., “Great picture quality, price, great zoom ratio, nice control layout, nice lcd size”, the extracted features are picture quality (f_2), price (f_1), zoom (f_3), control (f_{30}) and lcd (f_{12}). Construct a vector of 1 x 63 for the give review with the values of all the features appearing in the review as 1, and others are zero. Using the knowledge base (Table 4) and Algorithm 3, we obtain the vector NRVS = (1.506, 0.0141, 1.1682, 0.1385). Further, we assess the quality of the review based on the maximum value and categorize it accordingly. Since, the maximum value $NRVS_{max}$ = 1.506 corresponds to G_1, the given review is assessed as most significant review (MSR).

The Table 5 shows the summary of the assessment results of the remaining 50% of reviews of digital camera. The second column of the table shows the number of reviews classified to each group using k-means clustering technique, e.g. there are 252, 112, 98 and 158 reviews as MSR, MoSR, SR and ISR, respectively. Similarly, the third, fourth and fifth columns give the similar information corresponding to other products, namely, car, mobile and mp3, respectively. The sixth, seventh, eight and ninth columns give the similar information corresponding to the same products using feature clustering approach [25].

The experimental results show that, there are considerable number of reviews belonging to insignificant group, which do not influence a buying decision significantly. Hence, such reviews may be ignored while summarization of customer reviews. We observe that the proposed method based on review clustering identifies more number of reviews as insignificant as compared to classification of reviews based on feature clustering technique. Since the reviews are grouped using k-means clustering, the efficacy of the proposed method is better in terms of identifying the number of insignificant reviews.

Group name	Proposed Technique (Review clustering)				CAR Technique [25] (Feature Clustering)			
	Camera	**Car**	**Mobile**	**MP3**	**Camera**	**Car**	**Mobile**	**MP3**
MSR (G_1)	83	96	42	110	252	210	228	106
MOSR (G_2)	114	100	51	78	112	195	95	28
SR (G_3)	107	56	68	53	98	73	44	140
ISR (G_4)	296	348	299	261	158	122	93	228

Table 5. Comparison of the number of reviews classified to different groups obtained by proposed method and CAR technique [25] for Camera, Car, Mobile and mp3 player.

5. CONLUSION

In this paper, we have proposed a novel and effective web mining technique for assessing the customer review extracted from a web page for a particular product based on the clustering of reviews. The quality assessment of a customer review is categorized as most significant review, more significant review, significant review or insignificant review. This is performed in two steps : (1) Extract the features of reviews and group the reviews by a clustering technique and then compute cluster weight for each of the groups, (2) Assess the review by projecting its review vector on to the feature weight vectors of all the four groups. The experimental results show that usually there are large number of reviews belonging to insignificant group. The reviews belonging to insignificant group may be ignored while review summarization, and thus optimizing the process of review quality assessment. The clustering of reviews is found to be more effective than clustering of features of a review based on feature weights.

6. REFERENCES

[1] Hu Minqing and Bing Liu, Mining opinion features in customer reviews. AAAI. Vol.3 No.2 (2004), pp. 76-85.
[2] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP-2002. Vol No. (2002), pp.79-86.
[3] Ana-Maria Popescu and O. Etzioni, Extracting product features and opinions from reviews. HLT-EMNLP. (2005), pp. 32-33.
[4] G. Dejong, An Overview of the FRUMP System. In Stratregies for Natural language Parsing publisher (1982), pp. 149-176.
[5] J. Tait, Automatic summarising of english text. PhD Dissertation, Cambridge: University of Cambridge, 1983.
[6] D. Radev and K McKeown, Generating natural Language summaries from multiple on-line sources. Cisticsomputational Lingu 24, no. 3 (September 1998), pp.469-500.
[7] C. D. Paice, Constructing Literature Abstracts by Computer: techniques and prospects. Information Processing and Management 26 (1990), pp.171-186.
[8] J. Kupiec, J. Pedersen and F. Chen, A Trainable Document Summerizer. SIGIR-1995. Vol No. (1995), pp. 68-73.
[9] E. Hovy and C. Y. Lin, Automated Text Summarization in SUMMARIST. ACL Workshop on Intelligent, Scalable Text Summarization. (1997), pp.89-96.
[10] M. kan and K. McKeown, Information Extraction and Summerization:Domain Independence through Focus Types. Columbia University Technical Report CUCS-030-99, 1999, pp. 55-62.
[11] C. Jacquemin and D. Bourigault, Team extracting and automatic indexing. In Handbook of Computational Linguistics, edited by R Mitkov. Oxford University Press, 2001.
[12] J. S. Justeson and S. M. Katz, Technical Terminology: some linguistic properties and an alogorithm for identification in text. Natural Language Engineering 1, no. 1 (1995), pp. 9-27.
[13] Daille, Study and Implementation of Combined Techniques for Automatic Extraction of terminology. In The Balancing Act: Combing Symbolic and Statistical Approaches to Language Processing. Cambridge: MIT Press.
[14] K. Church and P. Hanks, Word association norms, mutual information and lexicography. Computational Linguistics 16, no. 1 (1990), pp. 22-29.
[15] K. Dave, S. Lawrence and D Pennock, Minig the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW-2003. (2003), pp.519-528.
[16] Hu Minqing and Bing Liu, Mining and summarizing customer reviews. KDD.Vol.18 No.2 (2004), pp. 119-135.
[17] S. Morinaga, K. Yamanishi, K. Tateishi and T. fukushima, Mining Product Reputation on the Web. KDD-2002. Vol No. (2002), pp. .
[18] Liu Bing, Hu Minqing and Cheng Junsheng, Opinion observer: Analyzing and Comparing Opinions on the Web. WWW-2005. Chiba, (2005), pp. 342-351.
[19] P. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL-2002. (2002), pp.76-83.
[20] Kim Soo-Min, Patrick Pantel, Tim Chklovski and Macro Pennacchiotti, Automatically Assessing Review Helpfulness. EMNLP-2006. (2006), pp.423-430.
[21] Liu Jingjing, Yunbo Cao, Chin-Yew Lin, Yalou Huang and Ming Zhou, Low-Quality Product Review Detection in Opinion Summarization. Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, publisher (2007), pp. 334-342.
[22] P.S. Hiremath, S.S. Benchalli, Siddu P. Algur and Renuka Udapudi, Mining Data Regions from Web Pages COMAD-2005. (2005), pp. 130-138.
[23] X. Wu, V. Kumar, Ross, J. Ghosh, O. Yang, H. Motoda, G. Mclachlan, A. Ng, B. Liu, P. Yu, Zhi-Hua Zhou, M. Steinbach, D. Hand and D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no. 1, (January 2008), pp. 1-37.
[24] P.S. Hiremath, Siddu P. Algur, S. Shivashankar, Web based Quality Assessment of Customer Reviews using Quartile Measure, International Journal of Recent Trends in Engineering, Vol. 1, No. 1, (May 2009), pp. 194-199.
[25] P.S. Hiremath, Siddu P. Algur, S. Shivashankar, Cluster Analysis of Customer Reviews Extracted from Web Pages, to appear in International Journal of Applied Computer Science and Mathematics.