



Opinion Feature Extraction via Domain Relevance

Ashwini P. Lingojar, Prof. L. J. Sankpal

Department of Computer Engineering,
Sinhgad Academy of Engineering,
Kondhwa, Pune, India

Abstract— Nowadays e-commerce sites are widely use by the consumer in order to purchase a product. There where large no of product reviews are presented by the product users about product. These reviews are in the form of text so it is very difficult for the customer to read all those review to make the purchase decision and it is also time consuming task. To overcome this problem it is necessary to have an automatic mechanism which extracts the product features from the large review corpus. It is important to extract valid opinion features to make a good purchase decision. Here we are going to extract frequent and infrequent features from the online reviews. For this we are using Intrinsic and Extrinsic domain relevance (IEDR) and feature clustering to extract opinion feature. Bothe of theses approach uses two domain review corpus in order to consider the distributional characteristic of the opinion feature across different domain. IEDR approach is use to find the valid frequent opinion feature and feature clustering is use to extract infrequent features from both domain-specific and domain- independent corpus. The extraction of features from the review is done on the basis of domain relevance score of the features from both domain specific and domain independent review corpus. Candidate features that are less generic and more domain-specific are then conformed as opinion features.

Keywords: Natural language processing, opinion mining, opinion feature extraction, IEDR, feature clustering

I. INTRODUCTION

In recent years with the explosive growth in social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, comments are postings on the Web, are increasingly using the content in these media for decision making. Decision making task is easily done by the considering the opinions of the users about the product. Opinions are key influencers of our behaviour, which central to almost all human activities. Whenever one need to make a purchase decision about the product, one want to know others' opinions. In the real world, businesses and organizations always want to find consumer or public opinions about their products and services.

Most of the opinion or sentiments are expressed in the textual form and analyzing those reviews is very difficult task. Analysis of these opinions is known as opinion mining and sentiment analysis. Sentiment analysis is the computational study of people's opinions, sentiments and attitude expressed in text. The opinion mining is the extension of data mining which utilize natural language processing techniques in order to extract people's opinion from the World Wide Web. For this reason opinion mining is the recent area of interest for researchers in the Natural Language Processing (NLP) domain.

Now people is planned to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. The Opinion mining system analyze each text and see which part contain opinionated word, which is being opinionated and who has written the opinion. Sentiment analysis analyzes each opinionated word or phrase and determines its sentiment polarity orientation, whether it is positive or negative or neutral. . In opinion mining, an opinion feature indicates an entity or an attribute of an entity on which users express their opinions. In this paper, we propose two different approaches for the identification of such opinion features from unstructured textual online reviews. As the structure of an opinion feature is distributed in a domain independent, and domain depended corpus. So there is an approach that identifies the opinion feature is called IDER (Intrinsic and Extrinsic domain relevance) approaches that evaluated the domain relevance of an opinion feature across two corpora and extract the valid opinion features along with feature clustering approach.

II. RELATED WORK

An opinion feature is defined as a object or entity on which users express their opinion. There are two types of opinion features such as implicit features and explicit features. In existing there are different approaches are presents for feature extraction which are mainly classified as supervised approach and unsupervised approach.

W.jin and H. H. Ho, discuss supervised machine learning framework that uses lexicalized Hidden Markov Model. This framework is naturally integrates linguistic features into automatic learning supported by model. This model can identify complex product specific features which are possible low frequency phrase in the review. This system can also self learn new vocabularies based on the pattern it has seen from the training data. Therefore the system is able to predict potential features in test dataset even without seeing them in the training set. This framework does not identify the role of pronoun in the mining result. [5]

Conditional Random Fields (CRFs) can employ rich features for review mining. F.Li and C. Han , provide the framework that can utilize the relationship among object features, positive opinions and negative opinions. It jointly extracts these three types of expressions in a unified way. The linguistic structure information can be naturally integrated into model representation, which provides more semantic dependency for output labels. With this framework, they investigate the chain structure, conjunction structure and syntactic tree structure for review mining. A new unified model, called skip tree CRFs, is proposed for review mining. But this does not cluster the related object features to provide more concise review summary. [2]

H. Guo Z. discusses mutual reinforcement approach to deal with the feature level opinion mining problem. It groups the product feature terms in reviews if they have similar meaning or refer to the same topic. Thus it can provide users a more sound and non-trivial opinion evaluation. Based on a pre-constructed association set, this approach is effective in finding the implicit product features, and well fit for online applications. Also, largely identify the related explicit product features which an opinion word is attached in reviews. This approach is easy to be combined with the existing explicit adjacency approaches to optimize the performance. Thus it provides a more accurate opinion evaluation. But this method of candidate product feature extraction and filtering can partly identify real product features; it may lose some data and remain some noises.

Hanshi Wang Lizhen Liu provide the novel method that uses the corresponding opinion words to extract features, and filters the noises according to mutual support scores and confidence scores. It also identifies the implicit features and clusters the features based on the knowledge of the context dependent information. Features considered, include not only the explicit features but also the implicit features. In this opinion words are divided into two categories, vague opinions and clear opinions, to deal with the task. Feature clustering depends on three aspects: the corresponding opinion words, the similarities of the features in text and the structures of the features in comment. Moreover, the context information is used to enhance the clustering in the procedure. In small scale corpora, it cannot perform well. [7] Certain Related Terms regarding this topic are as follows

1. *Natural Language Processing*

NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language. Today most of the information is stored on the internet in the form of blogs, opinions, attitude etc. and this data is in unstructured form. The extraction and analysis of huge unstructured internet content is beyond the human power and time. The content is mostly written in natural language. This situation necessitate an automatic natural language processing tool that extract and analyze the people sentiments from this unstructured texts.

2. *IEDR (Intrinsic and Extrinsic Domain relevance)*

The domain relevance of an opinion feature, which is computed on a domain-specific review corpus, is called intrinsic-domain relevance. Likewise, the domain relevance of the same opinion feature computed on a domain-independent corpus is called extrinsic domain relevance. IDR reflects the specificity of the feature to the domain review corpus while EDR characterizes the statistical association of the feature to the domain independent or generic corpus.

3. *Feature Clustering*

Clustering technique is use to identify complex relationships between features. Cluster the features with high similarity into groups to form a summary because people tend to use different words to express the same feature. Grouping those features and transforming them guarantee there is no loss of information, so that the infrequent features are also extracted by using feature clustering.

4. *Existing System Problems*

In the existing system IEDR considering only features that have high occurrence of the frequency which is frequent features, so it is less successful in dealing with the extraction of infrequent features. Frequent features are the features that people are most interested in for a given product. However, there are some features that only a small number of people talked about. These features can also be interesting to some potential customers. Therefore, it is necessary to overcome this limitation to provide valid opinion features.

III. IMPLEMENTATION

PROBLEM STATEMENT

IEDR approach is use to extract valid opinion feature from the online review corpus across two corpora, one domain specific corpus and one domain independent corpus. The IEDR utilizes the domain specific and domain independent corpus. The domain-specific opinion features will be mentioned more frequently in the domain corpus of reviews, compared to a domain-independent corpus. For each extracted feature ,the feature that are less generic and more domain specific are consider as a valid opinion feature. IEDR considering only noun phrases for extracting candidate feature so it is less successful in dealing with the extraction of infrequent features as well as non-noun features.

PROPOSED SYSTEM

The proposed system uses IEDR approach as well as feature clustering in order to extract both frequent and infrequent features from the online review corpus which is helpful for customer to make good parches decision of product.

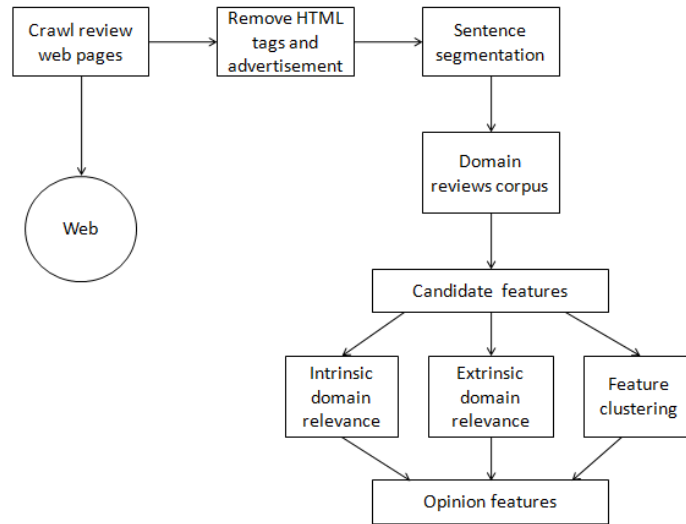


Figure gives the architectural overview of the proposed system. The system perform the extraction of features in two main steps: candidate feature extraction and IEDR and feature clustering. The input to the system is name of the product and entry page for all the reviews of the product. The output is the valid opinion features.

The system first crawls all the reviews related to the product name from the web. The obtaining reviews are in the form of web pages. Various operations are performing on the web pages in order to remove HTML tages from the web page. Then sentence segmentation is done in order to obtain the domain review corpus. Extract a list of candidate features from the domain review corpus via syntactic rule given in following table . Then for each extracted candidate feature estimate it’s IDR, which represents the statistical association of the candidate to the given domain corpus, and EDR, which reflects the statistical relevance of the candidate to the domain-independent corpus. Only candidates with IDR scores exceeding a predefined intrinsic relevance threshold and EDR scores less than another extrinsic relevance threshold. Feature clustering technique with k-means algorithm is also applied for each candidate feature in order to extract the infrequent candidate features. Then the feature extracted by IDER and feature clustering can be considered as the valid opinion feature.

Rules	Interpretation
$N + SBV \rightarrow CF$	if N has a SBV dependency relation then identify N as a CF.
$N + VOB \rightarrow CF$	if N has a VOB dependency relation then identify N as a CF.
$N + POB \rightarrow CF$	if N has a POB dependency relation then identify N as a CF.

IV. ALGORITHM REQUIRED

1. Web crawler

Input: Set of popular URLs U

Output: Repository of visited web pages R

Step 1: If (u==NULL)

then(P->U & P==NULL)

Step2: get (P*)

Step3: if(P*==R)

then return 1;

Step 4: else

add (P*->R)&&(P* !=R,U)&(P*->u)

2. Intrinsic and Extrinsic Domain Relevance

Input: A domain specific and domain independent corpus C

Output: Domain relevance scores (IDR or EDR)

For (each candidate feature c_i) do

For (document D_b in C) do

Find w_{ab} then

Cal s_a

Cal $dispa$

For (document D_b in C) do

Cal (deviab)

Compute domain relevance dr_a

Return (List of domain relevance)

3. Identifying Opinion Feature via IEDR:

Input: Domain review corpus C and domain-independent corpus I
Output: A validated list of opinion features of the user domain related events.
Step 1: Extract candidate from the review corpus C.
Step 2: For (candidate feature c_a) do
Step 3: Compute IDR score idr_a on the review corpus C
Step 4: Compute EDR score edr_a on the review corpus I
Step 5: If ($idr_a > ith$) and ($edr_i < eth$) then
Confirm candidate c_i as a feature.
Step 7: return (set of opinion feature)

4. Feature Clustering

Input: Domain reviews and features of customer review for particular product.
Output: Review matrix.
Step 1: For each review R_i in the raw review database {
Step 2: For each feature f_j in the review {
Step 3: If f_j is present in R_i then $M_{ij} = 1$ {
Else $M_{ij} = 0$
}
}
}

Grouping of features based on clustering techniques

The grouping of features based on feature clustering is done by following steps:

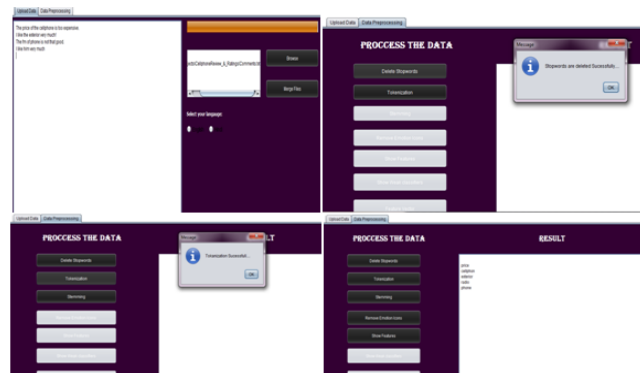
- Step1: Construct the review matrix for the review set using Algorithm 1.
Step2: Compute the frequency of occurrence for each feature.
Step3: Apply k-means clustering technique with for the dataset of frequencies of features computed in step2 and obtained groups.
Step4: Label the group with highest cluster and the groups with lower successive means.
Step5: Assign the weight to each feature.
Step6: Compute the sum of each group of cluster.
Step7: Maximum value of the sum consider as a valid features.

V. EXPERIMENTAL SETUP

Dataset: The experiment is conducted on the customer reviews of products. We were collecting the reviews from different shopping website. Products in these sites have a large number of reviews. Each of the reviews includes a text review and a title.

Evaluation Procedure: For each product, we first crawled and downloaded the reviews. These review documents were then cleaned to remove HTML tags. While cleaning the document various operations are performing on the review document in order to obtained only text review document. Then, syntactic rules are applied in the review document which is used to generate candidate features. After that, IEDR and feature clustering is applied to the extracted candidate features. To evaluate the discovered features our system performance is evaluated against precision versus recall. To evaluate the effect of corpus size on feature extraction, the systems F-measure performance versus the size of domain review corpus is used.

VI. RESULTS



For the given proposed system, the experimental results are obtained in the form of precision, recall and F-measures. These experimental results are used to demonstrate the effectiveness of our proposed approach by comparing it with the existing approaches. Here the IEDR along with feature clustering approach is used to compare with all other previous approach of feature extraction.

VII. CONCLUSION AND FUTURE WORK

We proposed a statistics approach to opinion feature extraction based on the IEDR feature-filtering criterion, which utilizes the disparities in distributional characteristics of features across two corpora, one domain-specific and one domain independent. IEDR identifies candidate features that are specific to the given review domain and yet not overly generic. Along with this, another feature clustering approach extracts the infrequent features. Both frequent and infrequent features are extracted from the online review corpus. This extracted feature is used to make good purchase decision.

This paper can be extended with the following features. The system can be enhanced by investigate new opinion mining algorithms that make good use of the IEDR extracted opinion features to summarize online reviews of products . A new two-stage feature-based opinion mining approach will be developed, where opinion features are first identified, followed by the discovery of the associated opinion polarities.

ACKNOWLEDGEMENT

I'm thankful to many persons who contributed to the completion of this research. Particularly I like to thank my guide Prof. L. J. Sankpal to help in research and Head of department Prof. B.B. Gite to allow me to continue this topic. Lastly I would like to Department of computer engineering, Sinhgad Academy of Engineering, Kondhwa to share their knowledge with me during my research work.

REFERENCES

- [1] M.Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 168-177, 2004
- [2] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.
- [3] W. Jin and H.H. Ho, "A Novel Lexicalized HMM- Based Learning Framework for Web Opinion Mining," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 465-472, 2009.
- [4] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.
- [5] S.-M. Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text, 2006.
- [6] G. Qiu, C. Wang, J. Bu, K. Liu, and C. Chen, "Incorporate the Syntactic Knowledge in Opinion Mining in User-Generated Content," Proc. WWW 2008 Workshop NLP Challenges in the Information Explosion Era, 2008.
- [7] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, pp. 9-27, 2011.
- [8] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003.
- [9] I. Titov and R. McDonald, "Modelling Online Reviews with Multi- Grain Topic Models," Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.
- [10] Y. Jo and A.H. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," Proc. Fourth ACM Int'l Conf. Web Search and Data Mining, pp. 815-824, 2011.
- [11] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu, "Structure-Aware Review Mining and Summarization," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 653-661, 2010.
- [12] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews," Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies, pp. 1496-1505, 2011.
- [13] W.X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a Maxent-Lda Hybrid," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 56-65, 2010. L. Tesnière, *Elements de la syntaxe structurale*. Librairie C. Klincksieck, 1959.
- [14] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th Int'l Conf. World Wide Web, pp. 959-968, 2008.
- [15] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 342-351, 2004.
- [16] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance," IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 3, MARCH 2014

AUTHOR PROFILE

Ashwini P. Lingojar received the B.E. degree in Computer Science & Engineering from Sipna College of Engineering & Technology, INDIA in 2013 and is pursuing M.E. in Computer Engineering at Sinhgad Academy of Engineering, Pune.

Prof. L. J. Sankpal received the B.E. degree in Computer Engineering from Marathwada University in 1991, M.E. degree in Computer Engineering from Shivaji University in 2001 and pursuing Ph D from Bharti University. She is currently working as a Assistant Professor in Sinhgad Academy of Engineering, Pune.