Competitor identification by the use of sentiment classification based on user reviews

Dr.N.Ananthi[1], Aishwarya.H[2], Aruna S[3], Ashwini Rithanya T[4]

[1]Associate Professor, Department of Information Technology, Easwari Engineering College
[2]UG scholar, Department of *Information Technology*, Easwari Engineering College
[3]*UG scholar, Department of* Information Technology, *Easwari Engineering College*
[4]*UG scholar, Department of* Information Technology, *Easwari Engineering College*

Abstract— **It is essential to identify the competitiveness between the items where each item has numerous features. In our project we use product reviews as well as the hotel reviews for the implementation in which the Hotel Domain Sentiment Classification can be extended to offer Service recommendation to the users based upon the requirements. The users would be adding the reviews of the items based on their intention. Then the collection of the data is done on the unstructured datasets over multiple domains sand then the Natural Language Processing(NLP) is applied to identify similar kind of reviews for the product. Then the Collaborative Filtering Technique is applied to identify the best items in various domains based upon the user reviews and then the items are sorted. Once the sorted item list is obtained, the competitiveness of the items are to be identified. This is performed based upon the intersection between the user reviews of the similar kind of products that are provided by the n number of users for every product. An user based CF Algorithm is adopted in order to generate appropriate recommendations. This aims at calculating a personalized rating of each candidate service for the user, and then presents a personalized service recommendation list thus recommending the most appropriate services to the user. Finally, the percentage of the competitiveness between the products is calculated based upon either the total number of users or the total number of user reviews.**

.

Keywords- ***Sentiment Analysis, Collaborative Filtering,Aspect Rating, Structural Correspondence Learning, Natural Language Processing***

I. INTRODUCTION

Sentiment Analysis can be described as a type of Natural Language Processing which includes obtaining the feeling of a user or a group of users expressed in various comments, requests or questions posted by them on the internet. It involves building a system that can collect the user opinions and then examine and classify them according to the polarity of the post. In other words, sentiment analysis aims to determine the view of a speaker or writer on particular subject. Liu [1] defined a sentiment as a quintuple “, where oj is a target object, fjk is a feature of the object oj, soijkl is the sentiment value of the opinion of the opinion holder hi on feature fjk of object oj at time tl, soijkl is +ve, -ve, or neutral, or a more granular rating, hi is an opinion holder, tl is the time when the opinion is expressed.” Sentiment Analysis has applications in various fields. For example, in marketing it helps in determining the success or failure of a new product launch or any new commercial campaign or determining that which version of a product is liked more in which part of the world. Various companies can use this data to determine their future strategies regarding a particular product or service. Sentimental Analysis can be based on a document, sentence or a phrase. In document based sentimental analysis, sentiment of the whole document is calculated as a whole and summarized according to the polarity. In sentence based sentiment analysis, individual sentences are classified as positive, negative or neutral whereas phrase based sentimental analysis assigns a polarity to the individual phrases contained in a sentence. The first requirement of sentimental Analysis is to find the subject towards which the opinion is expressed. After that the sentiment is classified as positive (which denotes satisfaction or happiness on behalf of user), negative (which shows rejection or disappointment) or neutral (which denotes no strong sentiment involved). Then the sentiment can be given a score which denotes the degree of positive or negative response from the user. There are various challenges involved with sentimental analysis. Subabrata [2] categorized these challenges as following:

A. Implicit Sentiment Sometimes a sentence may carry a strong sentiment without containing any sentiment bearing word in it. For e.g. One has to be on a lot of medications to make such a documentary

B. Domain Dependency Some words have different polarity when used in different domains. For e.g. The movie was inspired from a Hollywood movie. I got inspired by this book.

C. Thwarted Expectations Sometimes the writer builds up a positive context and refute it in the end. For e.g. Excellent performances, very good music, stunning cinematography, all in vain because of lack of imagination of the writer/director.

D. Pragmatics The pragmatics of the user needs to be identified. For e.g. It was good to see India destroy Australia in final. The match destroyed my interest in sports.

E. World Knowledge Sometimes the knowledge of an entity which is used in the sentence is required to identify the sentiment. For e.g. He is just as good a person as Dracula One has to know about Dracula to understand the correct sentiment behind this sentence.

F. Entity Identification There may be multiple entities in a sentence it is important to identify that the sentiment is directed towards which entity. Chelsea is better than Man. Utd. This statement is + ve for Chelsea and –ve for Man. Utd.

G. Negation Handling negation is very difficult. One method is to reverse the polarity of every word that comes after a negative word (e.g. not). For e.g. I do not like this movie. However this method will fail for Not only was the food delicious, the service was excellent.

II. LITERATURE SURVEY

Comparative analysis in Sentiment Analysis Technique:

The research work in the comparative analysis has been increasingly addressed in the recent years.. Comparative analysis is used to identify the competitive positions in businesses and results over a period of time. The literature surveys the sentiment analysis and opinion mining. In 2014, Chetan Kaushik proposed various techniques of sentiment analysis and the challenges associated with it. With the increase in the use of sentiment analysis in r social media sites, an increasing interest can be seen regarding Sentimental Analysis and Our main requirement is to develop a technique that can differentiate between the positive, negative or neutral sentiments underlying in a text. By performing analysis to identify the sentiments behind any text, one can predict the mood of the people about a particular product or any kind of services.

Competitor mining:

Competitor mining refers to identifying competition between the products and many kind of services. In our project competitor mining helps to the users to find the best product with the least effort. George Valkans and Theodoros Lappas proposed this strategy to find how do we quantify the competitiveness between two items? Who are the true competitors for the given item? What are the features of the item that affects its competitiveness? Despite the impact of this problem to many domains, only a limited amount of work has been done in order to get the best solution. In this paper we find the competition between the best hotel services based on aspect rating of the features and also the analysis of products at various different domains.

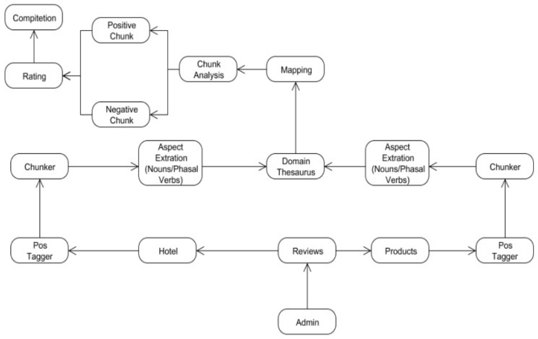
Other methods of analysis:

In this paper we use collaborative filtering to generate appropriate recommendations. It aims at calculating a personalized rating of each candidate service for a user, and then presenting a personalized service recommendation list and recommending the most appropriate services to him/her. Then we need to find the percentage of competitiveness between the products that can be calculated based on no of users reviews products/total no of users. Badrul Sarwar proposed this technique in his journal Item-based Collaborative Filtering Recommendation algorithms which projects different techniques for computing item-item similarities and different techniques for obtaining recommendations from them.

III. PROPOSED FRAMEWORK

In this paper we propose a competitor analysis framework by using Sentiment Classification based on the user reviews. That is market monitoring Web agent system, namely Market Watcher Agent, for based on the design of an automated gathering business information relevant to a company in an automated approach. The technology is designed to assist competitor analysis that has the following important roles in strategic planning. The users would be adding the reviews of the items based on their intention. Then the collection of the data is done on the unstructured datasets over multiple domains sand then the Natural Language Processing is applied to identify similar kind of reviews for the product. Then the Collaborative Filtering Technique is applied to identify the best items in various domains based upon the user reviews and then the items are sorted. Once the sorted item list is obtained, the competitiveness of the items are to be identified. This is performed based upon the intersection between the user reviews of the similar kind of products that are provided by the n number of users for every product. An user based CF Algorithm is adopted in order to generate appropriate recommendations. This aims at calculating a personalized rating of each candidate service for the user, and then presents a personalized service recommendation list thus recommending the most appropriate services to the user. Finally, the percentage of the competitiveness between the products is calculated based upon either the total number of users or the total number of user reviews.

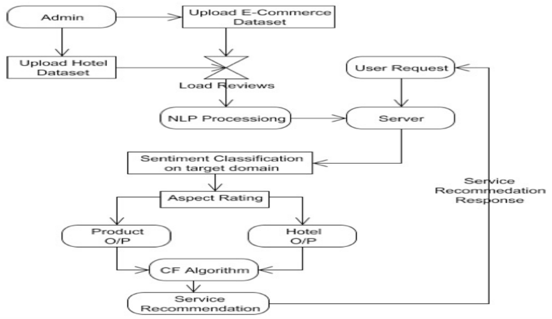
Figure 1. The workflow of the CA framework using Sentiment classification.



The workflow of the CA framework using Sentiment classification is presented in Figure 1. All the steps are supposed to fulfill the roles of CA as an ultimate goal as mentioned above. The workflow can be briefly divided into three phases, namely Pre processing of data sets, loading reviews, sentiment analysis, aspect rating and service recommendation. Data mining and decision support techniques are applied that convert collected information into meaningful business intelligence. The business intelligences are in turn, presented and used by the management through reporting and dashboard technologies respectively. If needed, users can preset certain rules so that instant alerts can be sent to the managers' mobile phones or PDA for immediate attention.

The Competitor Analysis Architecture is presented in the Figure 2. Structural correspondence learning (SCL) first selects a set of pivots, common features to both source and the target domains, using some criteria. One approach for selecting pivots is to select all features that occur more than a pre-defined number of times in both domains. Alternatively, a word association measure such as the mutual information (MI) could be used to measure the degree of association of a feature to a domain name, and select common features that have a high degree of association between both the source and the target domains By first predicting the pivots, and then learning a classifier using those predicted pivots as additional features, SCL attempts to reduce the Mismatch between features in the source and the target domains.

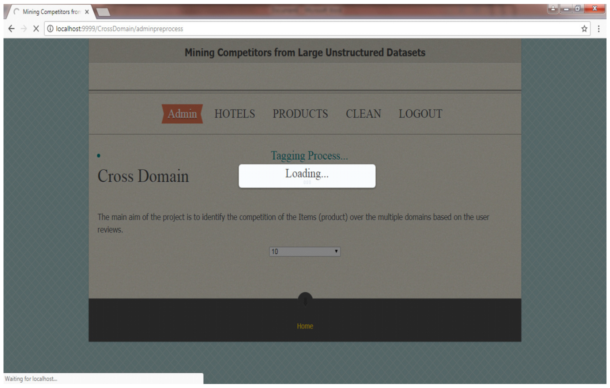
Figure 2. Competitor Analysis System Architecture



IV. IMPLEMENTATION

POS TAGGING FOR USER REVIEWS

Huge Collection of data is retrieved from open source datasets that are publicly available from web applications like Trip Advisor and Amazon .The Data's are in CSV or TSV Format. The CSV(Comma separated values) files were read and manipulated using Java API that itself developed by us which is developer friendly ,light weighted and easily modifiable. The User review for two different domains were loaded as a CSV or TSV file ,parsed using api and then each review by each customer is processed sequentially. The reviews were given one by one to POS Tagger which splits each word in the review and tags it based on the Parts of Speech the word belongs.


Mining Competitors fron ×
localhost:9999/CrossDomain/adminpreprocess
Admin
HOTELS
PRODUCTS
CLEAN
LOGOUT
Home
Waiting for localhost...

Mining Competitors from Large Unstructured Datasets

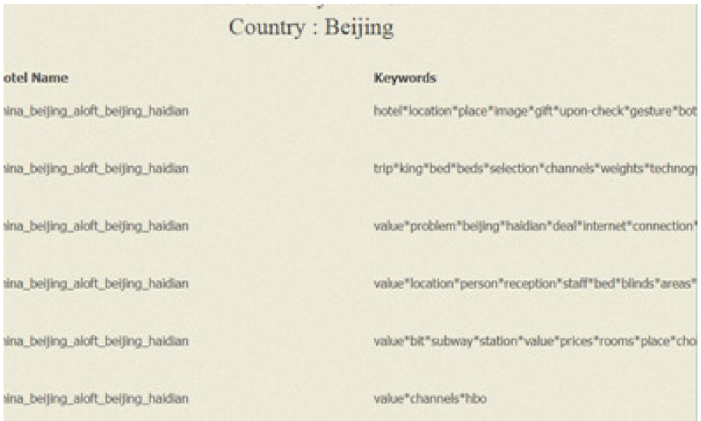
Cross Domain

Tagging Process...

Loading...

The main aim of the project is to identify the competition of the Items (product) over the multiple domains based on the user reviews.

10

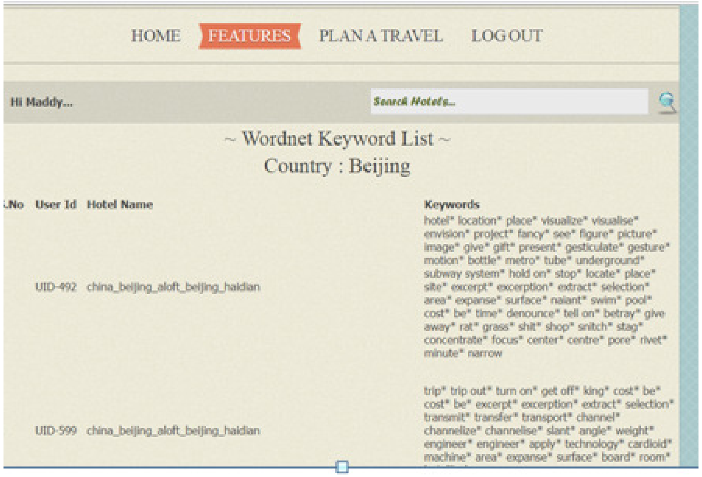


Country : Beijing

otel Name	Keywords
iina_beijing_aloft_beijing_haidian	hotel*location*place*image*gift*upon-check*gesture*bot
iina_beijing_aloft_beijing_haidian	trip*king*bed*beds*selection*channels*weights*technog
iina_beijing_aloft_beijing_haidian	value*problem*beijing*haidian*deal*internet*connection*
iina_beijing_aloft_beijing_haidian	value*location*person*reception*staff*bed*blinds*areas*
iina_beijing_aloft_beijing_haidian	value*bit*subway*station*value*prices*rooms*place*cho
iina_beijing_aloft_beijing_haidian	value*channels*hbo

CHUNKING THE REVIEWS AND ASPECT EXTRACTION

Chunking Process is done on each and every review of all and the products. The Chunking Process will take POS tagged output as input for Grouping the Words based on meaning of the Review. Chunking Process is done so that we can easily extract the sentiment embeddings associated with the Aspects of the particular review. The meaningful words that should be read continuously for proper understanding of the review are marked with square bracket. Now the Aspects in each review are extracted from the POS Tagger result. The Noun and Phrasal Verbs are the key Attributes in any sentence. So those things were extracted from the tagged reviews and marked as Aspects of the particular review by a user. Now mappings are done to properly annotate the user review and associated Aspects with the Chunks in it.



~ Wordnet Keyword List ~
Country : Beijing

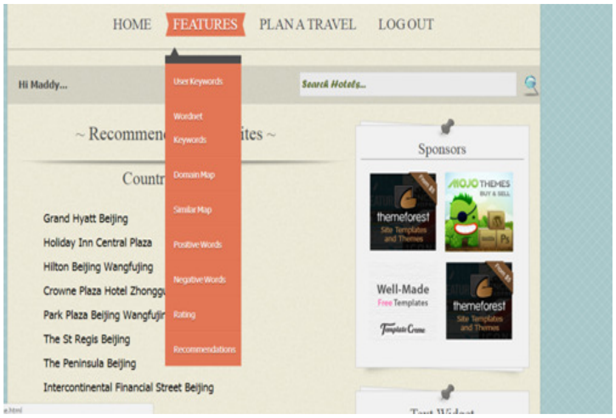
.No	User Id	Hotel Name	Keywords
	UID-492	china_beijing_aloft_beijing_haidian	hotel* location* place* visualize* visualise* envision* project* fancy* see* figure* picture* image* give* gift* present* gesticulate* gesture* motion* bottle* metro* tube* underground* subway system* hold on* stop* locate* place* site* excerpt* excerption* extract* selection* area* expanse* surface* natant* swim* pool* cost* be* time* denounce* tell on* betray* give away* rat* grass* shit* shop* snitch* stag* concentrate* focus* center* centre* pore* rivet* minute* narrow
	UID-599	china_beijing_aloft_beijing_haidian	trip* trip out* turn on* get off* king* cost* be* cost* be* excerpt* excerption* extract* selection* transmit* transfer* transport* channel* channelize* channelise* slant* angle* weight* engineer* engineer* apply* technology* cardioid* machine* area* expanse* surface* board* room*

BUILDING DOMAIN THESARAUS ON TARGET DOMAINS

A Domain Thesaurus is built depending on the Candidate Services and Keyword List. Candidate Services and Keyword List is dependent on the Target domains and it can be prepared before porting the classifier to Target domain. Expert Knowledge is needed for preparing the domain Thesaurus. The Domain Thesaurus can be Updated Regularly in order to get accurate Results of the Recommendation System. Now the Aspects extracted are subjected to domain grouping based on the main target domain. In this we demonstrate the sentiment classification for different domains using a single classifier without any training on any other target domain. The domain Thesaurus for any domain is built and classified. The results obtained are error free. In NLP techniques used for extraction of aspects and relevant classifications given on main domain using domain Thesaurus

SERVICE RECOMMENDATION BY IDENTIFYING DOMAINS

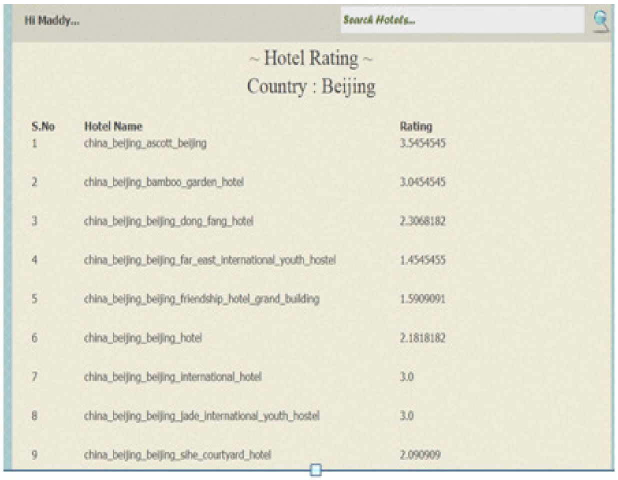
The Chunked Reviews of the User is retrieved and the Keywords (Aspects) corresponding to the User is Analyzed for its Valence and Arousal. Valence Means Weather the Keywords Means a positive or Negative thing and Arousal answers, how much it is?. Ratings are given for each domain in Target based on the Valence and Arousal for each User of each review. For product reviews the Overall Rating is now manipulated by taking average values of each rating of several users of a particular product. In Hotel Domain we extend ranking to give Personalized Service recommendation to user based on requirements to user. Ranking is done for all hotels based on Ratings by similar users using CF (Collaborative Filtering) and will be sorted based on Bubble Sort Algorithm to have the most appropriate personalized Recommendation for the User. Then we need to find the percentage of competitiveness between the products that can be calculated based on no of users reviews products/total no of users.



~ Recommen ites ~

Countr

Grand Hyatt Beijing
Holiday Inn Central Plaza
Hilton Beijing Wangfujing
Crowne Plaza Hotel Zhongg
Park Plaza Beijing Wangfuji
The St Regis Beijing
The Peninsula Beijing
Intercontinental Financial Street Beijing



~ Hotel Rating ~
Country : Beijing

S.No	Hotel Name	Rating
1	china_beijing_ascott_beijing	3.5454545
2	china_beijing_bamboo_garden_hotel	3.0454545
3	china_beijing_beijing_dong_fang_hotel	2.3068182
4	china_beijing_beijing_far_east_international_youth_hostel	1.4545455
5	china_beijing_beijing_friendship_hotel_grand_building	1.5909091
6	china_beijing_beijing_hotel	2.1818182
7	china_beijing_beijing_international_hotel	3.0
8	china_beijing_beijing_jade_international_youth_hostel	3.0
9	china_beijing_beijing_sihe_courtyard_hotel	2.090909

V. CONCLUSION

We have presented the formal definition of the competitiveness between two items, which we validated quantitatively as well as qualitatively. Our formalization is applicable across multiple domains, thus overcoming the short comings of the previous approaches. We have considered a large number of factors that have been widely largely overlooked in the past such as the position of the items in the multi-dimensional feature space as well as the preferences and the opinions of the users. Our work introduces the end to end methodology for the data mining such as the collection from a large datasets of the customer reviews. Based on our competitiveness definition we address the computationally challenging problem of finding out the top -k competitors of the given item .This proposed system is efficient and is applicable to a wide variety of multiple domains with a very large population of the items. The efficiency of our methodology was also verified by the experimental evaluation on real datasets from the hotel domain as well as the product domain. Our experiment also revealed that only a small number of reviews is sufficient to confidently estimate the different types of the users in the given market as well as the number of users that belong to each type.

References

[1] M. E. Porter, Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, 1980.

[2] R. Deshpand and H. Gatingon, "Competitive analysis," Marketing Competitors," Letters, 1994.

[3] B. H. Clark and D. B. Montgomery, "Managerial Identification of Competitors," Journal of Marketing, 1999.

[4] W. T. Few, "Managerial competitor identification: Integrating the categorization, economic and organizational identity perspectives," Doctoral Dissertaion, 2007.

[5] M. Bergen and M. A. Peteraf, "Competitor identification and competitor analysis: a broad-based managerial approach," Managerial and Decision Economics, 2002.

[6] J. F. Porac and H. Thomas, "Taxonomic mental models in competitor definition," The Academy of Management Review, 2008.

[7] M.-J. Chen, "Competitor analysis and interfirm rivalry: Toward a theoretical integration," Academy of Management Review, 1996.

[8] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the web," in ICDM, 2006.