Identifying Competitors based on user reviews using Sentiment Classification Methods

Dr. V.P. Gladis Pushpa rathi, D.R.Pradheep, M.Sreenivasulu Reddy, C.Venkatesan, R.P.Karthik

Abstract— For any competitive business to get success, it should fetch longing customers to make a profit out of it. Determining the competition of businesses from their services perspective claims to have a high success rate in terms of accuracy. The proposed system allows the customers choose the business which is highly competitive in the market. It can be achieved through the use of various sentiment classification techniques to rate the business from the user reviews data sets. Parts of Speech tagging, a Natural Language Processing procedure applied on the posted reviews classifies the meaningful words. Tagged words are chunked together to build domain thesaurus to map the words as per aspect. User posted reviews are listed into domain mapping and word net classification. Rating of the business is determined by user reviewed positive and negative keywords of the related domain in the scale of 1 to 10. With the help of collaborative filtering algorithm, we filter user provided requirements and recommend the best-rated business service. Recommendation of the service or product is achieved by sorting the filtered services list. Identifying, competitiveness between the business products helps to determine the popularity among its customers for business leaders. Bringing the new features to the products or service based on user requirements improves its appealing to the customers and profit of the business.

Index Terms— Collaborative Filtering Algorithm, Competitiveness, POS tagging, Structural Correspondence Learning, Sentiment Classification, User Reviews.

1 Introduction

DATA mining can be termed as course of anatomizing secluded patterns of data according to distinct perspectives for breaking down into applicable information, which is composed and assembled in routine areas, such as data warehouses, for efficient analysis, data mining algorithms, easing business decision making and other subsequent requirements to eventually gash costs and improve turnover. Sentiment Analysis (SA) or Opinion Mining (OM) is the reckoning study of people's opinions, stance, and emotions apropos of an entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two articulation SA or OM are interchangeable. document-level, sentence-level, and aspect-level are three important classifications in SA. Document-level SA aims to classify the whole document as a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. Aspect-level terms for positive or negative or neutral.

Collaborative filtering is the method to channel the data considering utilization of systems including coordinated effort among consolidated operators, perspectives, support information, and so forth. Activities of collaborative filtering ordinarily include substantial informational collections. Collaborative filtering calculations bears: clients online cooperation, client intrigues portrayal and calculations fit to parallel with individuals interests in a comparative mold.

A theory of competitiveness can be put forward on the following observation: competing for the attention and business of the same groups of customers for two similar items define the competitiveness. Based on their solicitation to the various customer segments in their market, competitiveness between two items can be coined. Our approach overcomes the reliance on previous work on scarce comparative evidence mined from the text. The identification of the various types of customers in the desired business domain, as well as for the assessment of the percentage of clients that belong to each group. A passing scalable framework for opting the top-k competitors of a required item in most forms of large datasets.

We extracted datasets belong to trip advisor provide hotels data in Beijing city and Chicago city. Datasets pre-processed accordingly to the aspect related classification. Product reviews posted on Amazon pulled out for the different set of items with product id as the primary key tag. We operated with these datasets to justify the proposed system.

2 Literature Review

2.1 Sentiment Analysis

In 2015, Kranti Ghag proposed the Sentiment Analysis deals with identifying and aggregating the sentiment or of opinions expressed by the users. Sentiment Analysis could be carried out by techniques that may or may not use a lexicon for polarity identification. Training data set i.e. already tagged opinions may or may not be used. Some sentiment analysis could be language dependent or some could be language independent, also called as multilingual sentiment analyzers.

2.2 Finding all Competition Products

In 2015, Yu-Chi Chung propounded a set of customer preferences, want to help the company to design set of competitive products so that the products can satisfy as many customer requirements as possible and the cost of producing the products is within a specified threshold.

2.3 Finding Top-k Competitors

In 2012, Yannis Kotidis put forwarded the reverse top-k queries to identify the top-k most influential products to customers, where influence is defined as the cardinality of the reverse top-k result. This definition of influence is useful for market analysis, since it is directly related to the number of customers that value a particular product and consequently to its visibility and impact in the market.

2.4 Efficient Processing of Skyline Queries

In 2005, C.Y.Chan proposed that given a set of points, the skyline comprises the points that are not dominated by other points. A point dominates another point if it is as good or better in all dimensions and better in at least one dimension. We address the novel and important problem of evaluating skyline queries involving partially-ordered attribute domains.

2.5 Extraction of Comparative Opinionated Sentences

In 2017, Jing JI and Jian Jin profounded that with the help of the techniques on sentiment analysis, opinionated sentences referring to a specific feature are first identified from product online reviews. Then, for the selection of a small number of representative yet comparative opinionated sentences, information representativeness, information comparativeness and information diversity are investigated.

3 Proposed System

In our proposed system to identify the competitiveness between the items, in each item having the number of features. We use product reviews as well as hotel reviews for implementation of our system. Hotel domain sentiment classification can be extended to give service recommendation to users based on their requirements. The user should be adding the reviews of the item based on their intention. Then we collect the data in unstructured data sets over the multiple domains and apply the NLP to identify the similar kinds of reviews on the products. Then apply the collaborative filtering technique to identify the best items in the various domains based on the user reviews and sort the items. Now we have the sorted items list then we need to identify the competitiveness of the items. So we are going to calculate the competitiveness of the products, based on the intersection between the similar kinds of product features reviews that are provided by the 'n' number of uses for each product. A user-based CF algorithm is adapted 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.

4 Modules

4.1 POS Tagging of User Reviews

Huge Collection of data is retrieved from open source data sets that are publicly available from web applications like TripAdvisor and Amazon. The data are represented 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, lightweight and easily modifiable. The user review for two different domains was 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 accordingly.

4.2 Chunking the Reviews and Aspect Extraction

Chunking process is done on every review of all the products. It 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 embedding associated with the aspects of the 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 review by a user. Now mappings are done to properly annotate the user review and associated aspects with the chunks in it.

4.3 Building Domain Thesaurus on Target Domain

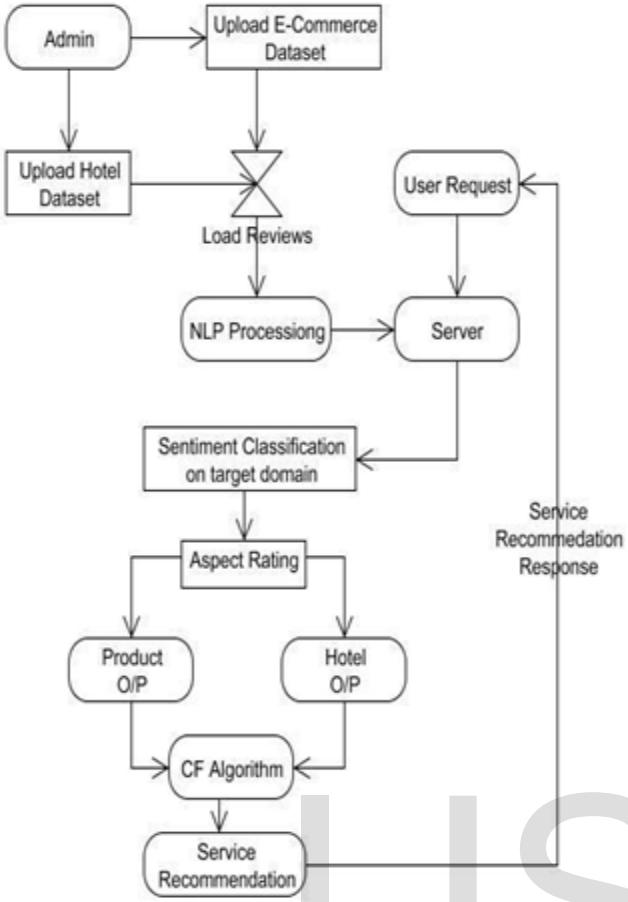
A Domain Thesaurus is built depending on the Keyword Candidate List and Candidate Services List. Keyword Candidate List and Candidate Services List are interdependent on the Target domains and it can be prepared before porting the classifier to Target domain. Expert Knowledge should be given for preparing the domain Thesaurus. The Domain Thesaurus can be updated regularly to get accurate results of the recommendation system. Now the Aspects extracted are subjected to domain groping based on the target domain.

4.4 Service Recommendation & Identifying Competitions

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. 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 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.

5 Architecture Diagram

A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. A system architecture can comprise system components, the expand systems developed, that will work together to implement the overall system.



6 RESULTS

6.1 Tagged Outpur Sample

Fig. 2. shows the tagged data output

6.2 Chunked Outout Sample

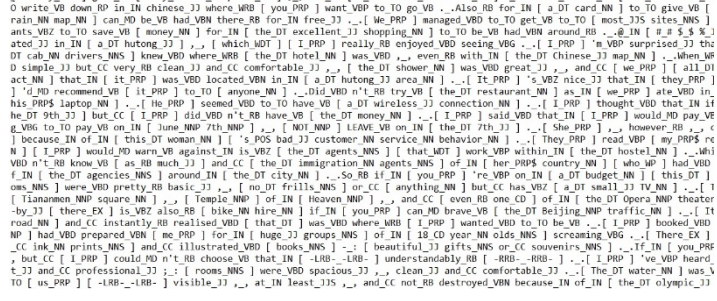


Fig. 3. shows the chunked data output

6.3 Wordnet Keywords of Listed Hotels

~ Wordnet Keyword List ~
Country : Beijing

S.No	User Id	Hotel Name	Keywords
1	UID-146	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* naiant* 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
2	UID-352	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* hotel* place

Fig. 4. shows the wordnet keywords

6.4 Positive Keywords of Listed Hotels

~ Positive Words ~
Country : Beijing

S.No	Hotel Name	Chunk	Domain	Words
1	Beijing&China_beijing_ascott_beijing&UID-102	Great_JJ room_NN	room	great-6
2	Beijing&China_beijing_ascott_beijing&UID-102	a_DT nice_JJ kids_NNS movie_NN	familyfriends	nice-6
3	Beijing&China_beijing_ascott_beijing&UID-102	excellent_JJ service_NN	service	excellent-6
4	Beijing&China_beijing_ascott_beijing&UID-102	great_JJ food_NN	food	great-8
5	Beijing&China_beijing_ascott_beijing&UID-102	good_JJ prices_NNS	value	good-6
6	Beijing&China_beijing_bamboo_garden_hotel&UID-147	Shaky_JJ start_NN great_JJ hotel_NN	food	great-6

Fig. 5. shows the postive keywords

6.5 Negative Keywords of Listed Hotels

~ Negative Words ~
Country : Beijing

S.No	Hotel Name	Chunk	Domain	Words
1	Beijing&China_beijing_bamboo_garden_hotel&UID-147	Shaky_JJ start_NN great_JJ hotel_NN	food	shaky-4
2	Beijing&China_beijing_bamboo_garden_hotel&UID-147	a_DT few_JJ dinners_NNS	food	few-3
3	Beijing&China_beijing_beijing_dong_fang_hotel&UID-649	a_DT few_JJ DongFang_NNP hotels_NNS	food	few-2
4	Beijing&China_beijing_beijing_friendship_hotel_grand_building&UID-269	a_DT few_JJ drinks_NNS	food	few-1
5	Beijing&China_beijing_beijing_friendship_hotel_grand_building&UID-269	a_DT small_JJ supermarket_NN	shopping	small-3

Fig. 6. shows the negative keywords

6.6 Competitiveness between Hotels

~ Recommecondations ~
Country : Beijing

S.No	Hotel Name	Rating	Check	Competition
1	Holiday Inn Central Plaza	5.381614	☐	
2	The Peninsula Beijing	5.0966797	☐	
3	Hilton Beijing Wangfujing	5.0085225	☐	Competition = 73 %
4	Grand Hyatt Beijing	4.9257812	☑	
5	Park Plaza Beijing Wangfujing	4.6171875	☐	
6	Crowne Plaza Hotel Zhongguancun	4.590909	☐	
7	Shangri La Kerry Centre Hotel	4.318182	☑	

Fig. 7. shows the competitiveness between hotels

7 CONCLUSION

Propounded system operated upon the hotel and product

domain datasets, which could be further extended to any type of competitive business involved in the market depends highly on user reviews. The system worked on rating the services or products based on user reviews which determines the value of the services or product rather on old-fashioned star rating. User recommendation system and competitiveness highly grasp to hold the customers to choose the right service or products on their own in the given sorted ordered list. Competition percentage appeals the customers to use the highly sophisticated services or products available among the markets.

To enhance the system in near future, the system could be built using Big Data components to handle for the collection of data from live websites for processing the same. Also, we can implement an algorithm to fetch the validation of the reviewers and rate the product or service accordingly. Rating scale could be updated with constraints which would give high weightage scale for valued reviewers and differentiate among all the reviews. The system could be extended to the market business competitors to identify the user requirements and develop the products based on their wish list.

ACKNOWLEDGMENT

The authors wish to thank Dr.T. Chandrashekar, Dr.V. Soundararajan, and Dr.R. Sugumar. This work was supported in part by a grant from Velammal Institute of Technology.

REFERENCES

[1] Kranti Ghag ; Ketan Shah, Comparative analysis of the techniques for Sentiment Analysis, Advances in Technology and Engineering (ICATE), 2013.

[2] Yu-Chi Chung, I-Fang Su, Chiang Lee and Pin-Chieh Huang, Finding All Competitive Products Using the Dominant Relationship Analysis, Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2015.

[3] Akrivi Vlachou , Christos Doulkeridis , Yannis Kotidis, Kjetil Nørvåg, Reverse Top-k Queries. Basic Research Funding Program 2008 of AUEB.

[4] C.-Y. Chan ; P.-K. Eng ; K.-L. Tan, Efficient processing of skyline queries with partially-ordered domains, Data Engineering, 2005. ICDE 2005. Proceedings.

[5] Ping Ji, Jian Jin, Extraction of comparative opinionated sentences from product online reviews,12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 2015.

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

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

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

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

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

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

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

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

[14] Z. Ma, G. Pant, and O. R. L. Sheng, "Mining competitor relationships from online news: A network-based approach," Electronic Commerce Research and Applications, 2011.

[15] S.Bao,R.Li,Y.Yu,andY.Cao,"Competitor Mining With The Web," IEEE Trans. Knowl. Data Eng., 2008. [12] G. Pant and O. R. L. Sheng, "Avoiding the blind spots: Competitor identification using web text and linkage structure," in ICIS, 2009.

[16] George Valkanas ; Theodoros Lappas ; Dimitrios Gunopulos, Mining Competitors from Large Unstructured Datasets, IEEE Transactions on Knowledge and Data Engineering, 2017.

[17] LD. Zelenko and O. Semin, "Automatic competitor identification from public information sources," International Journal of Computational Intelligence and Applications, 2002.

[18] R. Decker and M. Trusov, "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing, vol. 27, no. 4, pp. 293–307, 2010.

[19] C. W.-K. Leung, S. C.-F. Chan, F.-L. Chung, and G. Ngai, "A probabilistic rating inference framework for mining user preferences from reviews," World Wide Web, vol. 14, no. 2, pp. 187–215, 2011.

[20] Kelvin Lerman, Sasha Blair-Goldenshon, Ryan McDonald, Mining comparative opinions from customer reviews for Competitive Intelligence, IEEE Transcations on Data Engineering, 2009.