Sentiment analysis in E-Commerce using Recommendation System

Mr. R. Kumaran M.E [AP/IT]
dept.of Information Technology
V.S.B.Engineering College
Karur,Tamil Nadu

Ms. T. Yamuna, B.TECH
dept.of Information Technology
V.S.B.Engineering College
Erode,Tamil Nadu

Ms. L. Monisha, B.TECH
dept.of Information Technology
V.S.B.Engineering College
Erode,Tamil Nadu

Ms. P. Maheswari, B.TECH
dept.of Information Technology
V.S.B.Engineering College
Karur,Tamil Nadu

Abstract:- **Sentiment analysis gives significant information for decision making in various domains. Online shopping has become more and more popular because of its variety of types, lowest price and models. Now a days many people use online shopping to purchase the product through the Internet. Sentiment analysis also known as reviewing people opinion about the product. It uses natural language processing and computational linguistics to extract subjective information from the given data and classify opinions. The sentiments include ratings, reviews and emoticons. The proposed system use stochastic learning algorithm which analyze various feedbacks and user reviews which are classified as negative, positive and neutral. Our method was evaluated against real user data collected from an online website. Hybrid Recommendations is one of the important module of the system which helps overcome the drawbacks of the traditional Collaborative and Content Based Recommendations. We implemented supervised learning algorithm (Support Vector Machine) to improve the accuracy of results. It gives better performance than existing method. The proposed system helps the people to find out correct review of the product.**

Keywords: Sentiment analysis, reviews, stochastic learning algorithm, hybrid recommendation, support vector machine.

I. INTRODUCTION

Many devices such as Laptops and Smart phones using internet have made online shopping very easy. Online shopping has become more and more popular because of its variety of types, lowest price and fast logistic systems. Now a days many people use online shopping to purchase the product through the Internet. Sentiment analysis also said to be opinion classification which refers to the use of natural language processing and computational linguistics to extract subjective information from the given data and classify opinions. E-commerce means electronic commerce. It means dealing with goods and services through the electronic media and internet. E-commerce involves with in a business to help the people and by using the information technology like Electronic Data Interchange (EDI).

E-Commerce website related to the vendor on the Internet, who trades products or services directly to the customer from the portal. The portal uses a digital shopping cart or digital shopping basket system and allows payment through credit card, debit card or EFT (Electronic fund transfer) payments.

A massive internet growth will be added to E-commerce. Internet and smart phones are becoming an integral part of every life supply chain is also becoming leaner and smarter as digital platforms are helping to better connect with the customers which significantly reduces the waste and supporting to online businesses. A payment gateway is an e-commerce application service provider that authorizes credit card payments for e-businesses, online retailers, bricks and clicks, or traditional brick and mortar. The life of online business is the payment routes which comprises credit card, debit card, online banking payments, electronic funds transfer. The world is transforming from cash to digital money and thus there is a need of payment gateways for sustainable future e-commerce.

II. PROPOSED SYSTEM

A recommendation system has been implemented based on hybrid approach of stochastic learning and context based engine. The proposed framework have tried to combine the existing algorithms for recommendation to come up with a hybrid one. It improves the performance by overcoming the drawbacks of traditional recommendation systems. The proposed framework presented a implementation of a product recommendation system based on hybrid recommendation algorithm. The main advantages of this framework is to provide a visual organization of the data based on the underlying structure and a significant reduction in the size of the search space per result output. This framework also provide a simple method to search the products anywhere and anytime. Ratings, reviews and emoticons are analyzed and categorized as positive and negative sentiments. It uses very less time to predict all the ratings and reviews and classify it as positive negative and neutral. It will be used to calculate all the reviews which are collected from the E-commerce websites. It gives better performance than the existing method based on hybrid recommendation. It takes less time to calculate the accurate results as positive, negative and neutral.

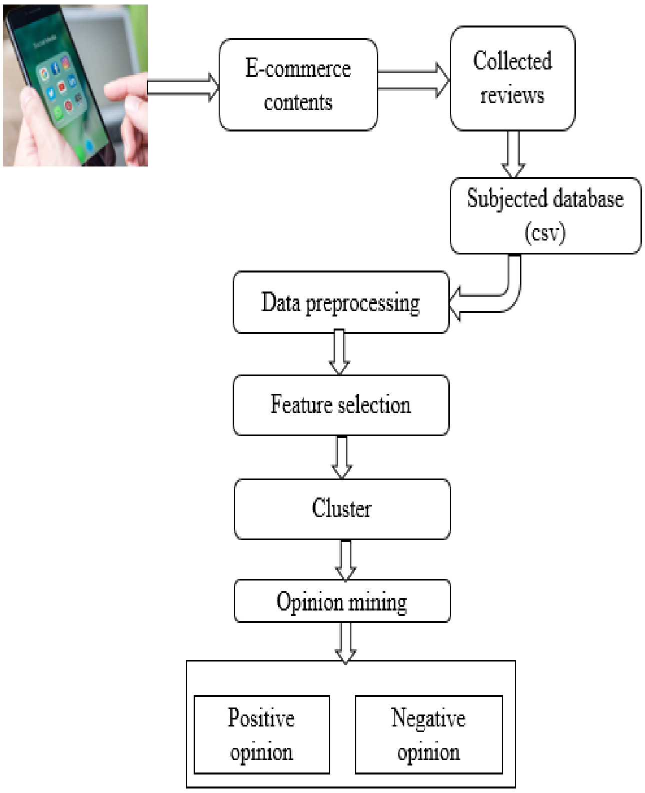

Figure 1: Architecture Of Proposed System

To begin with, we shed the light on our target social network Twitter. The microblogged Twitter was established in July 2006, and twitter's users has exceeded 140 million. In general, it has the facility of allowing users to post brief comments to a maximum of 140 characters, also known as tweets.

Recently, Twitter's users anticipate by publishing more than Million tweets daily about a variety of interests. For instance, one user can describe briefly his personal experience in using a particular mobile phone or purchasing any product. We choose Twitter because of the following features users can manage relationships in form of lists called followings/followers friends. These lists show posts from other which reflect their opinions and ideas. People can publish their opinions in divers topic such as political issues, books, movies, shopping and celebrities etc. in brief posts so called tweets.

Twitter has a public API supported by many different languages. The availability of this API gives developers the ability to develop software's to access the Twitter server and utilize their information. The Wide spread of advanced technological systems like phones and other smart devices allows instant access to Twitter.

Many products are available online. Most of the e-shopping sites inspire shopper to address their reviews about products to express their ideas on many aspects of the products. This gives rise to immense collection of feedbacks on web. These reviews contain rich and beneficial knowledge and have become main resource for both buyers and firms. Buyers usually look for quality report from online reviews before purchasing a product and firms can use these review as feedback for better product development, buyer relationship management and for the development of new marketing approach. A product may have number of aspects. Some of the product aspects are more significant than the others and have strong influence on the eventual buyer's decision making as well as firm's product development strategies.

Identification of important product aspects become necessary as both buyers and firms are benefited by this. Buyer can quickly make purchasing decision by paying attention to the significant aspects as well as firms can focus on improving the quality of these aspects and thus enhance product position accurately.

III. METHODOLOGIES

A. Twitter Services

Twitter has a public API supported by many different languages. The availability of this API gives developers the ability to develop software's to access the Twitter server and utilize their information. This enriches the service as it has a wide range of current information available from its users. Use of web and e-shopping web sites is developing very fast. This gives rise to immense collection of feedbacks on web. These reviews contain rich and beneficial knowledge and have become main resource for both buyers and firms. Buyers usually look for quality report from online reviews before purchasing a product and firms can use these review as feedback for better product development, buyer relationship management and for the development of new marketing approach. A product may have number of aspects.

i)Twitter Network Creation:

At first the Twitter network is created. In this the user can post their tweets and also able to view their timeline. The admin is able to collect all the tweets and information of the users.

ii)Classification:

The classification process is used to identify the category of the data's the classification is used to identify impossible data combinations, missing data's, out of range value, etc. The classification is used to remove the damaged data's, and the empty data's in the overall dataset.

iii)Clustering:

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It first select the data in the input. Then it validate the input data for the analyzing process.

iv) Feature Extraction:

Feature extraction is used to reduce the amount of resources required to describe a large set of data feature extraction starts from an initial set of measured data and builds derived values. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features Determining a subset of the initial features is said to be feature selection / Extraction.

B.Sentiment Similarity Analysis

The methods take the similarity analysis as an important and basic content, which consider the sentiment and emotion as the evaluation factors for trust. Additionally, sentiment and affective similarity analysis have been studied extensively in natural language understanding, data mining and statistical analysis .The existing methods of sentiment analysis based similarity exploration can be divided into three levels, which are document level, sentence level, and entity and feature level. All three levels are based on opinion lexicon, which is a collection of specific keywords or sentiment lexicons (extracted from gathered reviews) with parts-of-speech tags and treaded as the basis for analyzing the reliability of reviews Additionally, sentiment and affective similarity analysis have been studied extensively in natural language understanding, data mining and statistical analysis The existing methods of sentiment analysis based similarity exploration can be divided into three levels, which are document level, sentence level, and entity and feature level. Neutral usually means that no opinion is given. The analysis, both at the document level and at the sentence level, cannot exactly discover those specific objects whether people like or dislike. At entity and feature level, the approach concerns directly about the opinion itself instead of looking at language constructs (documents, paragraphs, sentences, clauses, or phrases). Hsu [2] adopted a sentiment word database to extract sentiment-related data from microblog posts and used these data to investigate the effect of different types of sentiment-related words on product recommendations. They proposed an approach ISTS that can exploit two factors from online social network: the sentiment orientation in friends posts about certain items and the trust relations between friends [18]. Li and Dai [19] proposed a promising methodology to handle the trust mechanism for P2P network. They let parties rate each other after the completion of transaction, and use the aggregated ratings of a given party to derive a trust score.While indirect trust is used widely in long path connected users through intermediate users.

C.Sentiment Similarity Based User Trust Relationship Calculation Framework:

The users are usually the consumers who have involved in E-commerce activities. They may have purchased some items or services and posted reviews on these objects, as shown in figure 2. Typically, a user can post multiple reviews on multiple items. Therefore, these reviews for specific items can be expressed in several texts. These reviews can usually be obtained by collecting network information. To find the trust, including direct and propagation, based on sentiment similarity of reviews by users in E-commerce systems, we propose a generally four-step computing framework. Firstly, the entity-sentiment word pairs are extracted from reviews. The step is a key process to further deal with the sentiment similarity analysis and direct trust computing. The extraction of entity-sentiment word pairs in part 1 is mainly to analyse the vocabulary of the text, extract the entity words and sentiment words which describing the object. We use NLProcessor linguistic parser for entity word and high frequency word combined with public lexicon for sentiment word. A mutual information formula is used to calculate the relationship between each entity word and the sentiment word. And then we can find those words pairs with close relationship. Secondly, we perform sentiment similarity calculations for two user-related reviews based on entity-sentiment word pairs. Calculation of sentiment similarity in part 2 is to compute the similarity degree of different reviews texts. This step is to use the obtained entity-sentiment word pairs for comparative analysis. Thirdly, we use a new formula to calculate direct trust between two users with a common review objects. Calculation of direct trust in part 3 is to compute direct trust have review son same item or object. The calculation method mainly includes two aspects, one is sentiment similarity, and the other is the user's rating of the object.

Finally, calculation of propagation trust in part 4, assume users as nodes, and the links includes basic information about reviewers, such as user name, user ID, and object information of the review which are based on their direct trust, as edges to create a trust network. Then we use an improved shortest path algorithm to calculate the propagation trust links between each pair of user nodes. The main goal of the existing sentiment analysis methods is to cluster the sentiments of users, commonly dividing people's sentiments to things into several types. Even at the entity and feature levels, its main purpose is to divide the user's sentiments into likes or dislikes. However, the above methods concern directly in overall trend which is insufficient when we calculate the trust based on sentiment similarity. It is necessary to analyze the specific attitudes on specific objects in reviews.

The similarity analysis based on sentiment has become an important research approach to establish trust relationship. Many studies have shown that there is highly correlation between trust and similarity.They demonstrated that individuals with similarities also have a high degree of trust in certain areas. These similarities include interest, content, behavior, etc. At meanwhile, through the analysis of the data from the FilmTrust Web site, the results show that when the similarity of users changes within a certain range, the trust between users changes accordingly. This change indicates that there is very strong relationship between trust and similarity. They used interest similarity between nodes to weight domain local trust recommendation. These innovative studies proved that there is a correlation between trust and similarity, and they had presented the corresponding calculation method. The entity and feature levels, its main purpose is to divide the user's sentiments into likes or dislikes. However, the above methods concern directly in overall trend which is insufficient when we calculate the trust based on sentiment similarity. It is necessary to analyze the specific attitudes on specific objects in reviews.

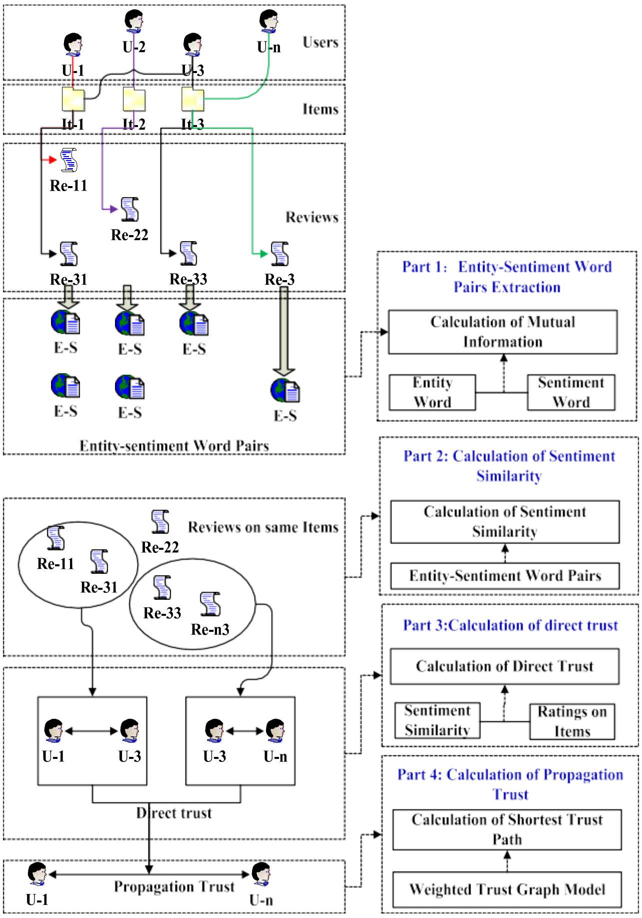
Users
U-1
U-2
U-3
U-n
Items
It-1
It-2
It-3
Reviews
Re-11
Re-22
Re-31
Re-33
Re-3
E-S
Entity-sentiment Word Pairs
Part 1: Entity-Sentiment Word Pairs Extraction
Calculation of Mutual Information
Entity Word
Sentiment Word
Reviews on same Items
Re-11
Re-22
Re-31
Re-33
Re-n3
Part 2: Calculation of Sentiment Similarity
Calculation of Sentiment Similarity
Entity-Sentiment Word Pairs
Part 3:Calculation of direct trust
Calculation of Direct Trust
Sentiment Similarity
Ratings on Items
U-1
U-3
U-3
U-n
Direct trust
Part 4: Calculation of Propagation Trust
Calculation of Shortest Trust Path
Weighted Trust Graph Model
U-1
Propagation Trust
U-n

Figure 2: Trust Calculation Framework Based On Sentiment Similarity

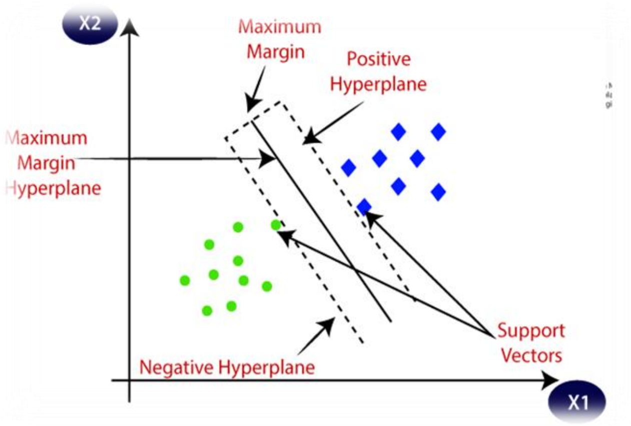

Figure 3: Classification of decision boundary based on support vector machine algorithm

A.DATASET COLLECTION

The experimental dataset is collected from Amazon.com. It contains 143.7 million reviews on 24 categories of product.The dataset includes basic information about reviewers, such as user name, user ID, and object information of the review, e.g. the product name or the product ID, the specific text of the review and the user's rating status.

Table 1:Review data structure

Field	Description
reviewerID	ID of the reviewer
asin	ID of the product
reviewerName	name of the reviewer
helpful	helpfulness
reviewText	text of the review
overall	rating of the review
summary	summary of the review
unixReviewTime	time of the review
reviewTime	time of the review

Table 2: Selected dataset of reviews

Field	Description
asin	ID of the product
title	name of the product
price	price in US dollars
imUrl	url of the product image also-viewed/
related products	bought sales rank information (related to also bought, also viewed, bought together, buy after viewing)
salesRank	brand name
brand categories	list of categories the product

TABLE 3. Selected dataset of reviews from amazon.

Category	Reviews	Items
Books	22.5M	2.37M
Electronics	7.82M	498K
Sports and Outdoors	1.32M	532K
Video Games	3.26M	51K
Baby	915K	71.3K

According to the data description we use three fields: related, overall and helpfulness, which is in the original experiment dataset, to judge whether the trust exists between users. If two users have the same related products (also bought, also viewed, bought together, buy after viewing),product metadata file that contains item description, product category information, price, brand, and image fea- tures and also-viewed / also-bought .The whole dataset is more than 350GB, and we just collect reviews on five categories (books, electronic, sports and out- doors, video games and baby) of products as test data for our experiments data will be correlated. According to the relationship between *precision* and *recall*, *F value* is used to represent the common impact of the two indicators on the results.

B.MEASURES

If the trust of any two nodes obtained by the sentiment mining algorithm is greater than a preset non-zero threshold value and trust relationship exists actually at the same time, then we deem the result of calculation is correct, and otherwise it is incorrect. If the results acquired by the pro- posed algorithm do not show trust links (including direct and propagation) relationship between two users, but the trust between them is true actually, this situation is called trust relationship missing. The evaluation index can be used to evaluate the trust links relationship found by the sentiment analysis algorithm;

that is, the validity of the algorithm is expressed by the ratio of the correct trusted links.

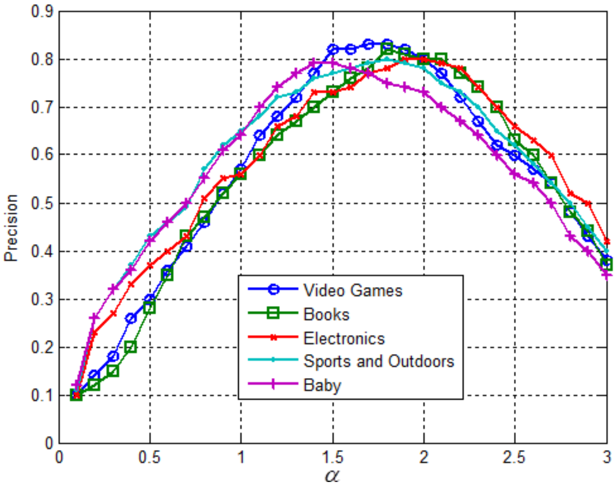

Figure 4: precision of direct trust analysis with different α.

The evaluation index can be used to evaluate the trust links relationship found by the sentiment analysis algorithm; that is, the validity of the algorithm is expressed by the ratio of the correct trusted links.

C.EXPERIMENTAL RESULTS AND ANALYSIS

We divide the review dataset into two parts, which is randomly divided into two non-overlapping parts, namely, training dataset and test dataset. We use 80% of the dataset for training and extracting entity-sentiment word pairs, computing sentiment similarity and trust. We use the left 20% of the dataset for experiment of accuracy and effectiveness of direct and propagation trust. Both direct trust relationship and the propagation trust are analysed experimentally to get trust between users. In the direct trust experiment, we use the control factor to test different accuracy of direct trust with different value of α. The range of α in our experiment is set to [0, 3].

The experiment of *precision*, *recall* and *F value* on different α are shown in figure 4, figure 5 and figure 6. These figures indicate that the factor of α plays a significant role in all results. Specifically, the values of *precision*, *recall, and F Value* grow apparently with the increase of α in the range of [0.1, 1.75]. Meanwhile, with the value of α increase in the range of [1.75, 3], the values of the three indictors decrease obviously.

The weights of ρ along with the increase of α, trust is different with distinct categories, but the overall trends are alike no matter α in range of [0.1, 1.75] or [1.75, 3]. These imply that the weights of ρ are important for the trust accuracy. Because the value of ρ increases monotonically with the value of α, when the value of α increase within a certain range, the accuracy of trust also increases, but when a certain maximum value is reached, the accuracy decreases conversely.Therefore, it is wise to consider the proportionality of the factor α comprehensively. At the same time, it can be seen that the influences of α on distinct categories of commodity vary apparently. This reveals the distinguish between the tendencies towards trust and people's views of different items. Training sample dataset plays a very important role in entity-sentiment word pair mining that we utilize in the experiment. We conduct an experiment to find the relationship between the number of training samples and the accuracy of trust between users. Choosing the appropriate threshold value ε is an important part in improving the propagation trust links analysis. But in real E-commerce system, the lengths of the trust path are different when people make trust judgment on diverse types of commodities. For example, some people prefer to consider a great many pages of reviews before making the final purchase decision, while others just like to reference to a few reviews. Therefore, it is really a challenge to find a perfect ε value.

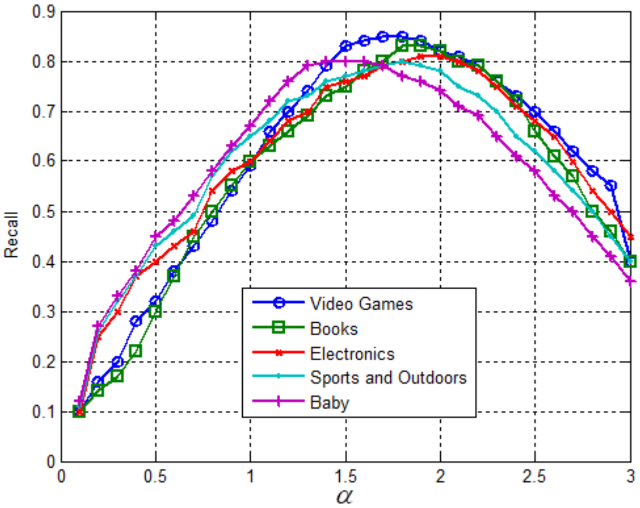

Figure 5: Recall of direct trust analysis with different α.

Training sample dataset plays a very important role in entity-sentiment word pair mining that we utilize in the experiment. We conduct an experiment to find the relationship between the number of training samples and the accuracy of trust between users. According to figure 4 to figure 6, when the value of α is close to 1.5, the accuracy is higher. Therefore, we set α=1.5 to perform the accuracy under different number of reviews samples. Total number of the selected training dataset contains more than 10 10^5 reviews, and this ensures the reliability of our experiments.

In propagation trust experiment, the trust path length between every two nodes is controlled by a threshold ε.T is a value obtained through experiences. The larger the ε, the shorter trust path length is allowed, and the less nodes can be included in the trust path. Whereas smaller ε value means relatively lower sentiment similarity values can be deemed as the existence of trust between users and long trust paths.

IV. CONCLUSION AND FUTURE WORK

The proposed framework presented a novel implementation of a product recommendation system based on hybrid recommendation algorithm. The main advantages of this framework is to provide a visual organization of the data based on the underlying structure and a significant reduction in the size of the search space per result output. This framework also provide a simple method to search the products anywhere and anytime. Ratings, reviews and emoticons are analyzed and categorized as positive and negative sentiments. Search the products based on price based filtering and reviews based filtering. Our method was evaluated against real user data collected through an online website, by using a subset of the movies liked by each user as input to the system. Hybrid Recommendations is one of the main modules of the system which helps overcome the drawbacks of the traditional Collaborative and Content Based Recommendations.

Not each user gives their reviews on each item, so the user's reviews data are usually sparse for a particular item. How to explore similarity of users with extremely sparse reviews data, e.g. by designing more efficient algorithm to overcome the challenge; (2) the degree to which people trust others is different for different things. Under more stringent requirements, it is also necessary to distinguish the categories of trust targets in details. how to include other information, for example, purchase item category, brand and other activities, into user sentiment calculation framework and (3) how to incorporate temporal factors to capture users' similarity change will be the focus of future research.

V.REFERENCES

[1] H. Liu, F. Xia, Z. Chen, N. Y. Asabere, J. Ma, and R. Huang, ``TruCom: Exploiting domain-specic trust networks for multicategory item recommendation," IEEE Syst. J., vol. 11, no. 1, pp. 295304, Mar. 2017.

[2] P.-Y. Hsu, H.-T. Lei, S.-H. Huang, T. H. Liao, Y.-C. Lo, and C.-C. Lo, ``Effects of sentiment on recommendations in social network," in Electron Markets. Berlin, Germany: Springer, 2018, pp. 110, doi: 10.1007/s12525- 018-0314-5.

[3] C. Qin, W. Siyi, and A. Lin, ``The joint beta distribution with refund rate in online C2C trust building: A theoretical study on Taobao," in Proc. Int. Conf. E-Learn. E-Technol. Educ. (ICEEE), Lodz, Poland, Sep. 2012, pp. 191196.

[4] S. Kraounakis, I. N. Demetropoulos, A. Michalas, M. S. Obaidat, P. G. Sarigiannidis, and M. D. Louta, ``A robust reputation-based computational model for trust establishment in pervasive systems," IEEE Syst. J., vol. 9, no. 3, pp. 878891, Sep. 2015.

[5] P. De Meo, E. Ferrara, D. Rosaci, and G. M. L. Sarné, ``Trust and compactness in social network groups," IEEE Trans. Cybern., vol. 45, no. 2, pp. 205216, Feb. 2015.

[6] M. G. Ozsoy and F. Polat, ``Trust based recommendation systems," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Niagara Falls, ON, Canada, Aug. 2013, pp. 12671274.

[7] L. Sheugh and S. H. Alizadeh, ``A fuzzy aproach for determination trust threshold in recommender systems based on social network," in Proc. 9th Int. Conf. E-Commerce Developing Countries, Focus E-Bus. (ECDC), Isfahan, Iran, Apr. 2015, pp. 15.

[8] Y. Ruan, L. Alfantoukh, and A. Durresi, ``Exploring stock market using twitter trust network," in Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl., Gwangiu, South Korea, Mar. 2015, pp. 428433.

[9] C.-N. Ziegler, Social Web Artifacts for Boosting Recommenders, vol. 487. Berlin, Germany: Springer, 2013.

[10] C.-N. Ziegler and J. Golbeck, ``Investigating interactions of trust and interest similarity," Decis. Support Syst., vol. 43, no. 2, pp. 460475, 2007.