Implementation of Recommender System Using Feature-Based Sentiment Analysis

Nyein Ei Ei Kyaw, Thinn Thinn Wai
University of Information Technology
Yangon, Myanmar
nyeineieikyaw@uit.edu.mm, thinnthinnwai@uit.edu.mm

Abstract

A recommender system aims to provide users with personalized online product or service recommendations to handle the increasing online information overload problem and improve customer relationship management. Collaborative Filtering (CF)-based recommendation technique helps people to make choices based on the opinions of other people who share similar interests. This technique has been suffering from the problems of data sparsity and cold start because of insufficient user ratings or absence of data about users or items. This can affect the accuracy of the recommendation system. User-generated reviews are a plentiful source of user opinions and interests. The proposed personalized recommendation model uses feature base sentiment analysis using ontology that extracts the semantically related features to find the users' individual preferences rather than rating scores in order to build user profiles that can be understood by user-based collaborative filtering recommendation model. The proposed model intends to alleviate data sparsity problem and to improve accuracy of recommender system by finding user preferences from review text.

Keywords- Collaborative Filtering (CF), Data sparsity, Review text

1. Introduction

The growth of the internet has made it much more difficult to effectively extract useful information from the available online information. We are suffering from information overload and being at a loss for the presence of too much information. A personalized recommendation system is one of the effective ways to solve this problem and has been used in many applications [6]. The mainstream of traditional recommendation approaches is usually based on the commonality among users i.e., similar users or entities are found by measuring the similarities of the common rating scores of users. However, the insufficiency of relevant data such as sparsity significantly weakens the effectiveness of these approaches due to the fact that there are often a limited number of common ratings among users.

User-generated online reviews have evolved into a pervasive part of e-commerce nowadays, as well as an essential focus of business intelligence and big data analytics. Both the online retail websites, like Amazon.com and Taobao.com, and the forum websites, such as Dianping.com and TripAdvisor.com are collecting tremendous amounts of online reviews.

Except for the ratings by users, the user reviews can offer much finer-grained information and have become a rich source to help detect the users' preferences. Most of the reviews contain users' opinions on various aspects of the target products/ services (referred to as entities). A user's preferences for the aspects of a certain entity are of great value in developing personalized recommendations [10].

Feature-based opinion mining is an attempt to identify the features of the opinion and classify the sentiments of the opinion for each of these features. The feature-based opinion mining of product reviews is a difficult task, owing to both the high semantic variability of the opinions expressed and the diversity of the characteristics and sub-characteristics that describe the products and the multitude of opinion words used to depict them [8].

The rest of the paper is organized as follows: Section 2 describes related works. Section 3 explains the background theory. Section 4 explains in details the architecture of the proposed system. Section 5 presents the performance evaluation of the proposed system. Section 6 describes the conclusion of paper.

2. Related Works

Several papers have addressed the problems to meet the personalized requirement of a user in various ways. Pallavi R. Desai, B. A. Tidke proposed a system that presents personalized recommendation lists and recommend the most appropriate items to the user by using weights of keywords are used to indicate user' preferences and a user-based collaborative filtering algorithm is adopted with Opennlp to generate appropriate recommendations [1]. Khushboo R. Shrote, Prof. A.V.

Deorankar proposed a system in which feedback analysis is done using sentiment analysis to recommend services. Keywords are used to indicate what the users prefer [2]. Susan Thomas, Jayalekshmi S proposed a system in which sentimental analysis on the reviews is done using Naïve Bayes, a machine learning technique to distinguish between the positive and negative reviews. It also uses MongoDB database to store the review detail [3]. Shakhy.P.S1, Swapna.H2 proposed a recommendation system which considers not only user reviews but also the temporal information about the location of the services. It uses Apache Mahout learning library and MongoDB to store reviews [4]. Dr. Kogilavani Shanmugavadivel and their colleagues proposed a system deals with the implementation of personalized rating to the services for hotel reservation system and booking of cars. This system performed opinion mining on the review at the sentence level using Bayes theorem and negation rule algorithm [5].

All the papers applied sentiment analysis on the reviews by using machine learning algorithms and then similarity of previous users and active user are computed and finally recommend the top N services to the new user. The candidate service sets (features of service) that system provide and domain thesaurus (similar terms associated candidate service) are manually specified. This is time-consuming and cannot relate the semantic meaning of the terms (features) more accurately.

3. Background Theory

3.1. Collaborative Filtering

Collaborative Filtering (CF) is considered the most popular and widely implemented technique in the recommender system. The underlying assumption of CF is that people with similar preferences will rate the same objects with similar ratings. Existing CF solutions can be categorized into two main classes: (i) memory-based and (ii) model-based methods. Memory-based solutions leverage similarities in users' behaviors and preferences to make inferences about missing values in the rating matrix. Memory-based algorithms (also known as Neighborhood-based) rely on the notion of similarity among users, or items, to predict the possible interest of a user on items that he/she has not seen (or rated) before.

Memory-based CF solutions are typically divided into two main categories: user-based and item-based. The user-based approach is based on the assumption that similar users typically rate the items in a similar way. Item-based CF focuses on the similarities among items. It is based on the assumption that similar items are rated in a similar way by the same user.

Model-based methods exploit the matrix values to learn a model, similarly to a classifier that trains a model from labeled data. The learned model is then used to predict the relevance of new items for the users. At present, collaborative filtering systems have a wide range of applications and provide customers with a good experience, but they still face a number of major problems such as data sparsity. When the data is very sparse, the accuracy of the recommendation from the collaborative filtering algorithm declines, which is a very big problem [9].

3.2. Sentiment Analysis

Sentiment analysis or Opinion mining is a part of Natural Language Processing which is used to analyze the opinions expressed by the different users. Sentiment analysis can be performed on three different levels namely at the document level, sentence level, and feature level.

Sentiment classification techniques can be divided into machine-learning approaches and dictionary-based approaches. However, despite the fact that machine learning approaches have made significant advances in sentiment classification, applying them to news comments requires the use of labeled training data sets. Dictionary-based approaches, on the other hand, can provide significant advantages, such as the fact that once they have been built, no training data are necessary [7].

3.3. Ontology

An ontology is a formal description of concepts in a domain of discourse (classes), properties of every concept describing various features and attributes of the concept, and restrictions on attributes. The concepts refer to various entities that may be any product or an organization. The use of ontology in feature-level opinion mining is to distinguish the domain related features by defining the classes in the domain and giving the relationships between the classes and instances [11].

4. Architecture of the Proposed System

The proposed system consists of three parts. They are

1. Ontology-based features identification.
2. Polarity identification using SentiWordNet.
3. User base collaborative filtering approach.

In the proposed system reviews text of users are firstly collected as the dataset. And then, preprocessing step is performed on the reviews text. After that, domain ontology is constructed to extract the features which have the semantic meaning from the reviews text.

In the second part, each of the specified features is classified into positive or negative polarity by using sentiment lexicons.

In the third part of the system, a recommendation process is performed. It has three sub-processes: user similarity computation, rating prediction on each item and ranking the items. Finally, the proposed system produces top N recommendation lists to the user.

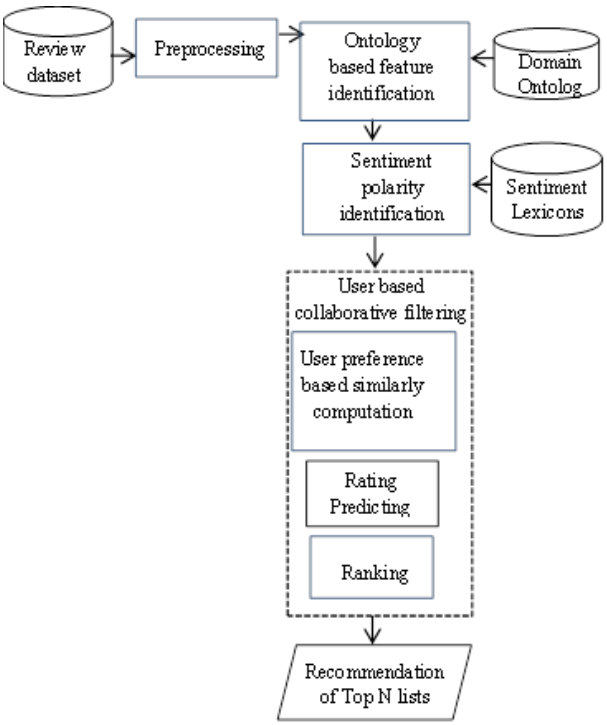

Figure1. Proposed system architecture

4.1. Reviews Collection

In this phase, reviews about hotels from online website "https://www.kaggle.com/datafiniti/hotel-reviews" are collected as the dataset for the proposed recommendation system.

4.2. Preprocessing

After collecting the hotels' reviews from the respective website, preprocessing steps are carried out on these review sentences. For preprocessing, tokenization, stemming and stop words removing are carried out by using NLTK (Natural Language Toolkit).

Firstly, NLTK tokenizer tokenizes a sentence into words and punctuation. After that NLTK's Porter stemmer removes the commoner morphological and inflexional endings from words in English.

In the stop words removing process, stop words such as "a, an, the, that…etc" are removed. Hotel features in reviews are usually nouns or noun phrases, while user opinions are usually adjectives or verbs. POS tagging helps extracting such information from reviews. POS tagging is performed by using NLTK POS tagger to tag each word such as noun, adjective, verb, adverb, conjunction, preposition, and interjection. Sample tagging results from NLTK POS tagger is shown in Figure 2.

('hotel', 'NN'), ('wa', 'NN'), ('comfort', 'NN'), ('breakfast', 'NN'), ('wa', 'NN'), ('good', 'JJ'), ('-', ':'), ('quit', 'NN'), ('a', 'DT'), ('varieti', 'NN'), ('.', '.'), ('room', 'NN'), ('aircon', 'NN'), ('did', 'VBD'), (''n't'', 'RB'), ('work', 'VB'), ('veri', 'RB'), ('well', 'RB'), ('.', '.'), ('take', 'VB'), ('mosquito', 'NN'), ('repel', 'NN'), ('!', '.'), ('realli', 'JJ'), ('love', 'JJ'), ('hotel', 'NN'), ('.', '.'), ('stay', 'VB'), ('on', 'IN'), ('the', 'DT'), ('veri', 'NN'), ('top', 'JJ'), ('floor', 'NN')

Figure 2. Sample tagging sets of hotel review

4.3. Domain Ontology Construction

Ontology is used to find the domain related features from the review sentences. Domain ontology is constructed by identifying concepts (classes), individuals, data type and object properties of the domain using the POS tagged words that resulted from the preprocessing step. Ontology construction is performed by using the Protégé 3.4 tool. Sample ontology of the hotel domain is shown in Figure 3.

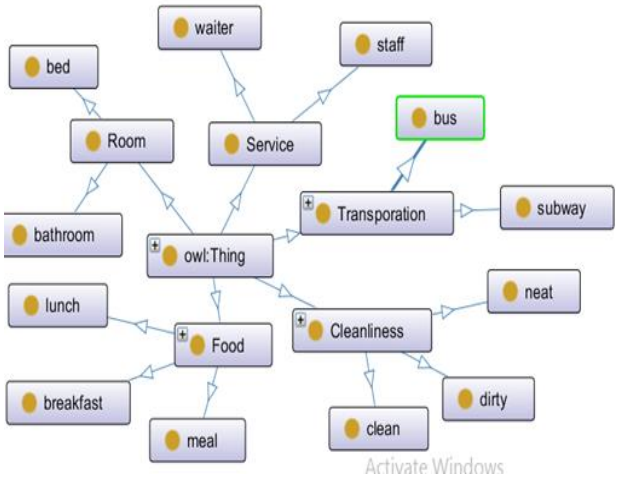

Figure 3. Sample ontology of hotel domain

4.4. Features Identification

In the features identification process, the features of user interests are extracted by using domain ontology. In

this process, noun words that are identified by POS tagger is compared with the concepts of the domain ontology, and then these words are extracted as features.

4.5. Sentiment Word Extraction and Polarity Identification

In this process, opinion words such as adjective, adverb, and verb extracted from the review sentences are used to classify the positive or negative opinion of the users. Sentiment polarity identification process is carried out by using these opinion words and SentiWordNet 3.0.

4.6. Finding User Preferences

Based on the user's opinion, overall sentiment polarity score, the entity sets that the user reviewed and user opinions on each aspect are used to calculate users' preferences. This calculation is performed by the following equation [12].

$$User\ Preference\,(u_i, f_k) = \frac{\sum_{e_i \epsilon E_i} (S_{ij}\, S_{ijk})}{\sum_{e_i \epsilon E_i} S_{ij}^2 \sqrt{\sum_{e_i \epsilon E_i} S_{ijk}^2}}$$

Where,

S_{ijk} represents the opinions that user u_i comment on aspect f_k for entity e_i.

S_{ij} represents the overall sentiment polarity score that user u_i assign to entity e_i.

E_i is the entity set that user u_i reviewed.

4.7. Similarity Calculation to Predict Rating and Ranking Items

After getting the users' preferences, user based collaborative filtering algorithm which is based on user preferences is used to predict rating and ranking items.

Firstly user similarity calculation is carried out by using the preferences of the target user and other users. This calculation is done by using Cosine similarity method.

$$Sim(u_i, u_m) = \frac{\sum_{k=1}^{|F|} (up_{ik} - \overline{up}_i)\ (up_{mk} - \overline{up}_m)}{\sqrt{\sum_{k=1}^{|F|} (up_{ik} - \overline{up}_i)^2 (up_{mk} - \overline{up}_m)^2}}$$

Where,

up_{ik} is the preference of target user u_i for aspect set F= (f1, f2,..., fk).

$\overline{up}_i$ is the preference of target user u_i for entity e_i.

up_{mk} is the preference of other user u_m for aspect set F= (f1, f2,..., fk).

$\overline{up}_m$ is the preference of other user u_m for entity e_i.

And then, items that have higher candidate score are ranked to the user. This score is calculated using the following equation based on the user preference with respect to other users with similar preference [12].

$$CS(u_i, e_j)_{=\overline{S_i}} + \frac{\sum_{u_m \epsilon u_M} sim(u_i, u_m) \cdot (S_{mk} - \overline{S}_m)}{\sum_{u_m \epsilon u_M} sim(u_i, u_m)}$$

Where,

$\overline{S_i}$ represents the overall sentiment polarity scores of user on entity e_j.

u_M represents set of users who write reviews about entity e_j.

S_{mk} represents the opinion of similar user on aspect f_k.

$\overline{S}_m$ represents the overall sentiment polarity scores of similar user on entity e_j.

5. Performance Evaluation

In the proposed system, performance evaluation is carried out by calculating precision and recall. Precision is the proportion of recommended items in the top N set that are relevant. Larger the precision better the recommendations. It is calculated by the following equations [14].

$$\text{Precision} = \frac{\text{Number of recommended items that are relevant}}{\text{Total number of recommended items}} \quad (1)$$

Recall is the proportion of relevant items found in the top N recommendations. It is calculated by the following equation.

$$\text{Recall} = \frac{\text{Number of recommended item that are relevant}}{\text{Total number of relevant items}} \quad (2)$$

The evaluation was based on the threshold value 3.5 which is specified by user and 200 rows from dataset for ten hotels. The actual rating greater than 3.5 regard as relevant items and user preference greater than 3.5 regard as recommended items. Number of recommended items that are relevant achieved from the intersection of the relevant items and recommended items.

Based on these assumptions, precision and recall are calculated by equations 1 and 2 respectively. Precision 75% of the recommended items was actually relevant to the user. Recall 70% of relevant items were recommended in the top N lists. The results are shown in Figure 4.

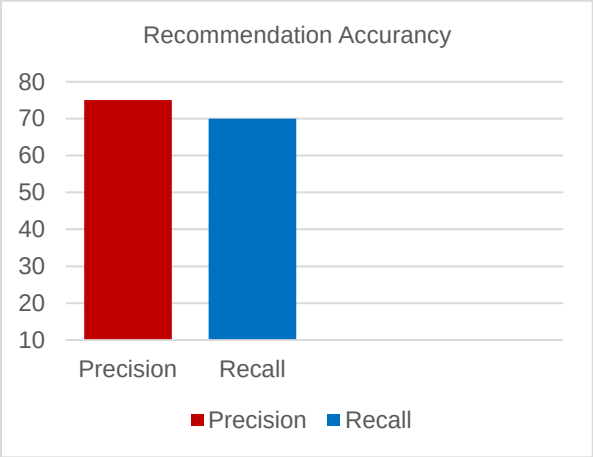

Figure 4. Performance evaluation of proposed system

6. Conclusion

In this paper, the proposed recommender system employs feature based sentiment analysis which use ontology at the feature extraction process to collect the semantic meaning of the features in review text. Then, customer preference and customer similarity are calculated based on feature words and sentiment polarity. Finally, the proposed system produces top N recommendation lists to the users.

7. References

[1] Pallavi R. Desai, B. A. Tidke, "A Survey on Smart Service Recommendation System by Applying Map Reduce Techniques", International Journal of Science and Research (IJSR) 2014.

[2] Khushboo R. Shrote, Prof. A.V. Deorankar, "Sentiment Analysis Based Feedback Analysed Service Recommendation method For Big Data Applications", International Journal of Scientific & Engineering Research 2016.

[3] Susan Thomas, Jayalekshmi S., "Recommendation System with Sentimental Analysis using Keyword Search", international journal for advance research in engineering and technology 2015.

[4] Shakhy.P.S1, Swapna.H2,"Improved Keyword Aware Service Recommendation System for Big Data Applications", International Journal of Innovative Research in Computer and Communication Engineering 2015.

[5] Dr. Kogilavani Shanmugavadivel, Dr. Thangarajan Ramasamy , Dr. Malliga Subramanian," Semantic Ranking Based Service Recommendation System using MapReduce on Big Datasets" , International Journal of Advances in Computer and Electronics Engineering 2017.

[6] Cheng Xiao, Dequan Zheng, Yuhang Yang, Automatic Domain-Ontology Structure and Example Acquisition from Semi-Structured Texts Sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009.

[7] Anisha P Rodrigues, Dr. Niranjan N Chiplunkar,"Mining Online Product Reviews and Extracting Product features using Unsupervised method", 978-1-5090-3646-2/16/$31.00 ©2016 IEEE.

[8] Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M.Moreno, V., Fraga, A., Sánchez-Cervantes, J.L., Feature-Based Opinion mining through ontologies, Expert Systems with Applications (2014), doi: http://dx.doi.org/10.1016/j.eswa.2014.03.022.

[9] Mattia G. Campana, Franca Delmastro "Recommender Systems for Online and Mobile Social Networks: A survey", *IIT-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy, 2017 Elsevier B.V. All rights reserved.*

[10] Yue Ma, Guoqing Chen, Qiang Wei, "Finding users preferences from large-scale online reviews for personalized recommendation", Springer Science+Business Media New York 2016.

[11] Drashti Naik, Jitali Patel, "Feature Extraction from Product Review Using Ontology", International Journal of Computer Sciences and Engineering Vol.5 (8), Aug 2017, E-ISSN: 2347-2693.

[12] Nan Jing, Tao Jiang, Juan Du, Vijayan Sugumaran, "Personalized recommendation based on customer preference mining and sentiment assessment from a Chinese e-commerce website", Springer Science+Business Media, LLC 2017

[13] https://www.kaggle.com/datafiniti/hotel-reviews.

[14] https://medium.com/@m_n_malaeb.