Yang and Yecies *Journal of Big Data* (2016) 3:3
DOI 10.1186/s40537-015-0037-9

Journal of Big Data
a SpringerOpen Journal

RESEARCH Open Access

# Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews

Jie Yang[1*] and Brian Yecies[2]

*Correspondence:
jiey@uow.edu.au
[1] SMART Infrastructure Facility, Faculty of Engineering and Information Sciences, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia
Full list of author information is available at the end of the article

**Abstract**

Analysis of online user-generated content is receiving attention for its wide applications from both academic researchers and industry stakeholders. In this pilot study, we address common Big Data problems of time constraints and memory costs involved with using standard single-machine hardware and software. A novel Big Data processing framework is proposed to investigate a niche subset of user-generated popular culture content on Douban, a well-known Chinese-language online social network. Huge data samples are harvested via an asynchronous scraping crawler. We also discuss how to manipulate heterogeneous features from raw samples to facilitate analysis of various film details, review comments, and user profiles on Douban with specific regard to a wave of South Korean films (2003–2014), which have increased in popularity among Chinese film fans. In addition, an improved Apriori algorithm based on MapReduce is proposed for content-mining functions. An exploratory simulation of results demonstrates the flexibility and applicability of the proposed framework for extracting relevant information from complex social media data, knowledge which can in turn be extended beyond this niche dataset and used to inform producers and distributors of films, television shows, and other digital media content.

**Keywords:** Social media, User-generated content, Big data analytics, Content mining, Parallel association-rule mining

## Background

The last decade has witnessed the dramatic expansion of online social networks (hereafter OSNs) at the global level. More and more people are employing OSNs in their day-to-day lives to access information, express opinions and share experiences with their peers. As a result, massive volumes of content are generated every day from numerous social media channels. A typical example is Facebook, which was recently reported to receive 10 million new photographs every hour [1].

A significant proportion of online content is associated with the film domain, as many OSNs (such as Rotten Tomatoes, FilmCrave and Twitter) provide cinema fans with convenient mechanisms for posting and sharing their opinions or comments about movies online. Prospective audiences are increasingly inclined to rely on online reviews to make their own viewing choices, as well as a list of films upon which they might comment (but not necessarily see). As a result, investigating user-generated content (hereafter

UGC) to discover significant patterns generated by these online audiences is becoming increasingly common [2–5]. The benefit of this type of research is derived from several aspects. For instance, UGC analysis reveals how electronic word-of-mouth (hereafter eWOM) can be utilized as a powerful communication tool and social networking channel for spreading awareness of a given film in both the offline and online worlds. In addition, such analysis closes the gap between the film producer and audiences by offering a better understanding of consumers' opinions. In turn, the provision of media contents can be customized in terms of production, distribution, exhibition, and associated promotional advertising. By predicting audience preferences and future behaviors, not only can industry stakeholders promote their contents more effectively, but an improved user experience can be offered, and audiences can be better assisted in finding films related to their particular interests.

A survey of the relevant literature reveals that investigators have devised a number of approaches to UGC analysis related to film media [3, 4, 6–8]. For example, the sentiment analysis applied to movie reviews promises a better understanding of audience opinion. UGC-based prediction systems have also been proposed in relation to film ratings and box office performance. In addition, film recommendation systems can be regarded as another important outcome of UGC analysis. Some of these approaches are reviewed in "Related work".

However, regardless of the growing interest in UGC analysis in general, there has been little research into the potential social benefits and commercial applications of Chinese language eWOM as a tool for developing and utilizing UGC. In addition, most existing work focuses on a particular aspect of the available data (primarily movie reviews) or specific metrics (such as box office statistics or movie ratings) that are analyzed with an average single machine (i.e. a PC with 4 cores and 4 gigs of memory). These results are insufficient to provide a complete framework to cover all components of audience experience, or measure the effects of both internal and external factors all while performing the analytical processes in a timely and efficient way. More importantly, with the exponential accumulation of UG.C, the challenges associated with the big four "V" problems data volume (numbers of films, users, and generated reviews), variety (different data formats), velocity (streaming comment data), and veracity (language uncertainty) continue to multiply. This presents a typical scenario for Big Data processing, which is difficult to address using traditional analysis methods. Thus there is an increasing need to develop alternative frameworks for conducting UGC analysis. Our aim is to develop an efficient and practical but novel technique for investigating large Chinese-language data sets associated with audience responses to international media contents one that is superior to those available to researchers using traditional qualitative and quantitative survey or analysis instruments.

With this end in view, this paper describes an innovative collection and analytical tool, termed Douban-Learning, that is, a brand-name rather than a claim about the system's intelligence, which has been designed to facilitate UGC-based data mining in the film domain. The major point of difference that distinguishes our inventive work from conventional methods is that the proposed framework is designed to address Big Data problems associated with a reliance on a single-machine equipped with average hardware and software, while at the same time satisfying the type of costly data storage and

computational requirements covered in previous studies [9, 10]. The major contributions offered here can be summarized as follows:

- An efficient framework, Douban-Learning, is implemented for large volumes of social media data processing based on the Hadoop platform. User-generated contents are collected, distributed, stored and processed on the Hadoop distributed file system (HDFS);
- An asynchronous scraping crawler is implemented via a multiple-task queue, which facilitates data collection in an efficient and simultaneous manner;
- Multiple heterogeneous features are generated to represent raw data records related to a variety of film details (i.e. actor, director, writer, and story elements typically found in English-language comments on most OSNs), movie reviews, and user profiles. A novel extraction, transformation and load (ETL) process is introduced to facilitate quantification;
- An improved Apriori algorithm based on MapReduce is proposed to increase the flexibility and efficiency of Big Data mining.

The framework proposed here can in turn inform strategies not only for producing and promoting films featuring particular actors, plot elements, and locations, etc., but also for modifying or localizing stories for specific markets, regions, and target audiences.

The remainder of this paper is organized as follows. In "Related work", we briefly review some existing work for UGC analysis relevant to our work. We then introduce the Douban-Learning framework in "Experimental results", where three major modules are discussed in terms of data harvesting, feature generation and content mining. Our proposed framework is evaluated in "Experimental results", and in "Conclusion" we offer our conclusion and suggest further prospects for the system.

## Related work

This section offers a brief review of state-of-the-art research in terms of UGC-based analysis. First, we investigate sentiment analysis, gleaned from users' comments. We then discuss the UGC-based prediction system for movie ratings or box office performance. We further investigate one existing popular application of UGC analysis: recommendation systems. Finally, some existing processing architectures and ability for mining UGC are also provided.

### Sentiment analysis

One of the most important roles of UGC mining is to understand users' attitude or preference via sentiment analysis. Although sentiment analysis has been a major topic of natural language processing for many years, only recently has it attracted the interest of the Web-mining research community. One reason for this change is the increasing popularity and availability of large collections of topic-oriented data across online review sites, microblogs, and social networking sites. For our purposes, sentiment analysis is mainly applied to movie reviews in order to extract subjective information from the comments made by particular audiences and categorize them as positive, neutral or negative. To express the matter in mathematical terms: let $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ be a tuple representing

$n$-textual features extracted from one review, and $y \in \{positive, neutral, negative\}$ be the class label. Sentiment analysis aims to train a classifier that extracts the decision rule subject to the following constraint:

$$y = f(\mathbf{x}) + e, \tag{1}$$

where $f(\cdot)$ is an unknown decision function to be estimated by the classifier, and $e$ is the corresponding error. Essentially, textual features ($\mathbf{x}$) and classification function ($f(\cdot)$) play the largest and most critical role in accurately determining the sentiment of a given review. A number of different textual features or classifiers have been proposed to supplement the previous studies on which our study builds.

In [3], a heuristic sentiment analysis of movie reviews is proposed. Given the word class, the method explores textual features using the combination of "Adverb", "Adjective", and "Verb" for the review classification. Samad et al. provide a hybrid algorithm to determine individuals' views by combining *support vector machine* (SVM) and *particle swarm optimization* (PSO) algorithms as the classifier [6]. Two typical features, word frequency and *term frequency-inverse document frequency* (TF-IDF), are extracted for training purposes. Experimental results show that the hybridization of SVM with PSO improves classification accuracy, compared to conventional SVM classifiers. A review summarization algorithm is proposed in [11], in which the *latent semantic analysis* (LSA) method is used to identify textual features, which are accessible to experts and non-experts alike because of their appearance in general sites such as the popular Internet Movie Data Base (aka IMDB) online site and mobile application. As a result, movie reviews are represented using feature-based summarization. A comprehensive study of movie tweets is found in [12]. Around 10,000 tweets are collected and labelled manually, and are then each converted into a binary textual feature to train an SVM classifier. Other sentiment analysis work can also be found in [13–15].

#### Prediction system

Movie revenue and box office statistics are critical measurements for quantifying a film's commercial success. Yet, audience reviews from raw UGC cab also offer a practicable model of predicting film performance. As with sentiment analysis, Eq. (1) can also be used as a general model for the prediction system. However, the class label $y$ now applies to the film revenue rather than its sentiment.

In [4], 14 keywords (such as "love", "wonderful", "best") are extracted as textual features. Then a Na?ve Bayes classifier is trained to predict trends at the box office. Yu et al. propose an approach based on *probabilistic latent semantic analysis* (PLSA) to extract sentiment factors, and two autoregressive models are then implemented for prediction [7]. Experimental analysis shows that positive sentiments are strong indicators of film performance. Similarly, in [16] the PLSA approach is again employed to identify textual features. Furthermore, a Fuzzy logic method is employed to quantify extracted features, followed by the application of a regression model for prediction. Another fuzzy-based hierarchy method is employed in [17] to generate features characterized in terms of information quality and source credibility. Real-world data samples are then collected from four OSNs, including Aditya's, Rotten Tomatoes, FilmCrave and Teaser Trailer, to conduct the prediction. The effect of another important OSN (Twitter) on movie sales

is also investigated [18]. Experimental results show that positive tweets are indeed associated with higher movie sales, whereas negative film comments usually reflect lower movie sales. (Consider that prediction models in [7, 16, 18] are based on the (logistic) regression model.) Some other classifiers, such as neural networks, are proposed in [19, 20]. The superior simulation results obtained demonstrate the efficacy of neural network-based models over the regression prediction.

#### Recommendation system

Movie recommendation systems, one of the most popular applications of UGC analysis, aim to suggest new movies to audiences based on their established preferences (generated from historical users' contents). Most existing recommendation systems fall into two categories: collaborative filtering (CF) and content-based (CB) methods.

CF methods make recommendations based on a group of users outside the sample group with similar film preferences. Given a user list $\mathcal{U} = \{u_i | \forall u_i \in \mathcal{U}\}$, for any one target user $u_i$, CF methods generate a sorted user list $\widehat{\mathcal{U}}$, which satisfies the following conditions:

- $\widehat{\mathcal{U}} \subset \mathcal{U}$, and $u_i \notin \widehat{\mathcal{U}}$;
- $sim\{u_j, u_i\} \geq sim\{u_k, u_i\}$, subject that $\forall u_j, u_k \in \widehat{\mathcal{U}}$ and $j < k$.

The variable $sim\{u_j, u_i\}$ represents the similarity between user $u_j$ and $u_i$. By finding the most similar user(s) to $u_i$, the film recommendation is made by aggregating the historical watching information from $\widehat{\mathcal{U}}$. Examples of the CF-based movie recommendation system include [5, 21, 22].

By contrast, the CB recommendation system takes into account movie metadata such as film genres, actors, directors, and basic descriptions [23–25]. That is to say, the correlation between movies is utilized as the key criteria for movie recommendations. Let the list $\mathcal{M} = \{m_i | \forall m_i \in \mathcal{M}\}$ be the previously reviewed film list for the $i$-th user. Let the $content(\cdot)$ function represent movie metadata, i.e., a set of pre-defined attributes or features characterizing movies. Accordingly, CB methods estimate a user's ($u_i$) preference for any movie $m_j$ ($m_j \notin \mathcal{M}$) based on its similarity with $\mathcal{M}$:

$$sim\{\mathcal{M}, m_j\} = \sum sim\big(content(m_i), content(m_j)\big). \tag{2}$$

More recently, some hybrid algorithms have been proposed to improve the accuracy of recommendations. For instance, in [8], both CF and CB-based recommenders are employed in parallel. A $K$-nearest-neighbourhood algorithm is implemented to estimate similarity. Meanwhile, clustering algorithms are also combined with the CF method to group together similar movies before recommendation [26]. Li et al. further suggest using a fuzzy K-means algorithm to cluster films with similar profiles [27]. These hybrid algorithms demonstrate their superiority over traditional recommendation systems by addressing problems such as data sparsity and cold start.

### Other studies

The UGC mining-based analysis can also be applied to discover relevant knowledge and to improve decision-making processes for individuals and organizations. In this context, some examples of existing architecture and processing ability include business intelligence, marketing, and disaster management. For instance, in [28] the authors present a marketing campaign approach to using Facebook UGC. A streaming model is established for predicting the number of visits, profit or even return on investment (ROI) with respect to advertising elements. Another mechanism for analyzing the impact of advertising is carried out in [29]. By analyzing raw contents such as timestamps, brand-term frequency and individual responses, the results reveal customer feelings about a brand, as well as other economic and social variables impacting on a company. Furthermore, research in [30, 31] shows that contemporary companies that take advantage of UGC analysis seem to outperform their competitors, and report commercial benefits such as cost effectiveness and improved efficiencies. In addition, the study in [32] proposes a two-model approach to examining post-disaster recovery using social media data. User generated content is firstly separated into active and passive perspectives, then a communication mediation and cultivation model is applied. Analytical results demonstrate that social media creates positive effects in post-disaster recovery.

### Summary

Despite the amount of research devoted to UGC analysis, most of the work conducted to date has focused on particular aspects of film contents. At the same time, little analysis has been conducted on Chinese culture-based user contents, either from the point of view of users' demographic profiles or their language habits. More importantly, existing work is mainly conducted by using an average single machine or even manual calculation. Many challenges associated with data volumes, variety, velocity and veracity still remain unresolved in conventional UGC analysis.

Taking all these aspects into account, the Douban-Learning framework is proposed as an efficient Big Data processing tool for Chinese social media platforms, with the ability to cover data collection, feature extraction and content mining. The proposed framework has a number of advantages which increase its utility in real applications:

- Unlike most conventional methods, Douban-Learning facilitates massive data analysis;
- Douban-Learning allows experts to apply multiple metrics to describe or explain features which can potentially yield results from multiple perspectives;
- Douban-Learning produces analytical results using association results, findings that are easily interpretable and clearly expressed for use in decision-making.

## Douban-Learning framework

In this section, we first provide some background information on the study area, including one famous Chinese OSN that is, Douban and its users' attitudes to South Korean (hereafter Korean) films. Douban film forums, which have been engaging with an increasing wave of Korean films and stars since the Korean and Chinese governments and their respective film industries began planning a co-production treaty in 2011, are

by far the largest aspect of this OSN. Next, the Big Data processing framework is introduced and the implementation of its three main modules is discussed. The implemented quantification algorithm and Hadoop platform are also introduced here.

#### Douban's Boost to Korean Cinema

Since the late 1990s Korean cinema has become one of the most dynamic national cinemas in the world. Korean films have made transnational connections across Asia and beyond, through their stylistic trends and experimentation with narrative and genres [33]. Given the fundamental role and importance that Korean cinema occupies on the global stage, it is unsurprising that a parallel phenomenon exists in the online world. In this paper we focus narrowly on Korean films and their influence on *Douban*, one of the biggest interest-oriented Chinese OSNs. This social website attracts more than 100 million active visitors per month, and has amassed over 65 million registered users. In 2015, it is accessed by over 30 % of Chinese Internet users, making this platform a major magnet for film fans across China. Douban users are able to disseminate their opinions on a wide range of these international films, and to make recommendations to their followers and friends. The result is the generation of vast quantities of self-interested user records.

However, documenting these public-available online records and mining them for useful information about Chinese movie audiences and their behavior toward international films is a challenging undertaking especially with a single machine. It involves the collection of massive amounts of data created by movie reviewers, ranging from individual users to geographically based catchments that grow over time. Furthermore, to gain a comprehensive understanding of Chinese-based UGC in this area, a variety of factors must be considered, such as specific features of films and movie reviews, as well as user profiles. In summary, the Douban OSN offers great potential of the Big Data application for analyzing Chinese UGC in ways that other studies such as [34] have yet to master.

To explore this rapidly changing arena, we selected a total of 114 Korean films released between 2003 and 2014. This subset includes the top 10 performing films in each year according to Korean box office statistics, which are publically available on the Korean Film Council online database [35]. Whilst these films were conspicuously popular among Korean fans, the case is not necessarily the same for Chinese fans, and thus this particular dataset offers a relatively unbiased opportunity to investigate the nature of their reception in user comments on Douban. (The open nature of the Korean and Chinese-Douban dataset and its potential for re-use makes it possible for independent observers and readers to replicate and build upon the results discussed below.)

#### Douban-Learning framework

In this section, the Douban-Learning framework is proposed as a means of discovering critical patterns of audience behavior in Chinese UGC. The proposed framework consists of three main stages or modules, which are summarized as follows:

1. *Data crawler module:* raw data records are firstly collected and distributed via an asynchronous scraping crawler; all collected records are then uploaded to the Hadoop platform and stored on the HDFS.

2. *Feature generation module:* this module is used to identify significant attributes from collected raw samples and producing high-level features;
3. *Content mining module:* an improved association rule-mining algorithm is implemented in this module, allowing us to categorize and interpret UGC with multiple heterogeneous features.

A more detailed description about individual modules and their functionalities is presented below. The proposed Douban-Learning framework is also shown in Fig. 1.

#### *Data crawler*

To begin with, a scraping crawler module is implemented in the first stage in our framework process to collect raw samples. Douban allows access to its data via public Application Programming Interfaces (APIs), while the samples that it provides are compressed in the JSON format. Despite the general availability of this data, a systematic procedure to eliminate redundant or unnecessary content via an asynchronous scraping crawler has been developed to overcome these constraints.

The scraping crawler module is then implemented to consist of one global controller and multiple workers, which are configured in different computer nodes. The controller is used to manage account details, distribute multiple IP addresses, keep track of tasks, and inspect status reports. The workers, on the other hand, execute scraping tasks concurrently. More precisely, when the list of target movies has been determined, the controller generates a task queue and then randomly assigns the priority for different movies. Note that films with a higher priority will be assigned to workers before those with a lower priority.

Next, an idle worker is initialized with a valid account and IP address for an assigned film. The worker creates further scraping threads for various contents from Douban. When the API constraint is applied, the worker is halted by creating a breaking point and recording the current status. The global controller will later reactivate halted workers

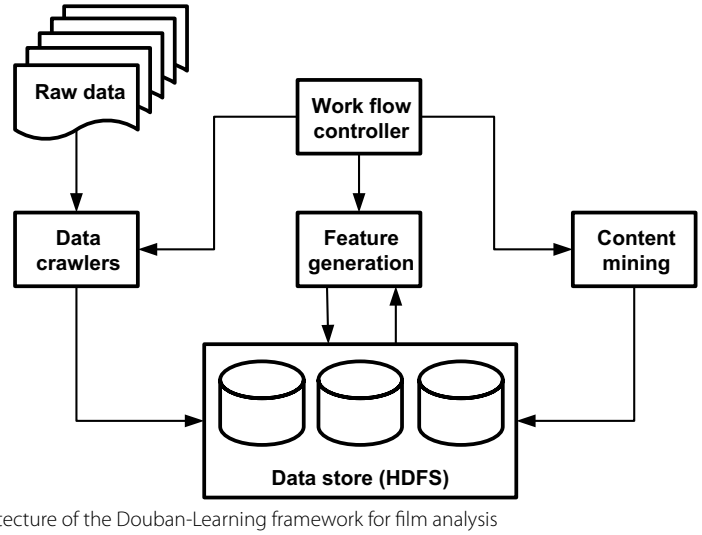


**Fig. 1** Architecture of the Douban-Learning framework for film analysis

until a pre-defined condition (here a particular time period) is satisfied. This distributing–working–waiting–reactivating process is recursively repeated for each worker until the assigned task is completed. The worker is then released and is ready for the next collection.

In this data collection stage, three categories of Douban contents are considered: film details, review comments and user profiles. (The raw attributes and relationships between contents are represented in Fig. 2). Collected samples are then distributed and stored separately according to the data category. That is, for a particular movie, all comments are recorded in one log file, while the related user profile is kept in another log document. The special symbol "#" is used to separate raw attributes. Note that a global file is created individually to store the basic film details in the controller node, whereas review comments and user profile data are stored in worker nodes. Therefore, the results from the first stage of our framework are collected records, which are then uploaded to the Hadoop platform and stored on the HDFS.

#### *Feature generation*

The second stage in our framework involves the identification or extraction subjective information from raw samples, which we then convert into usable features. Accordingly, feature generation module is implemented in this stage using a Big Data extraction, transformation and load (ETL) tool (Hive [36]). A high-level feature can be regarded as a user-defined hierarchical representation of these initial raw attributes. In accordance with the record category, related feature lists are generated to cover the same aspects: film details, review comments, and user profiles (as summarized in Table 1). Again, these features are generated as they are typically found in English-language comments on most OSNs. Among these, some features can be directly extracted from raw attributes, whereas others require high-level aggregation.

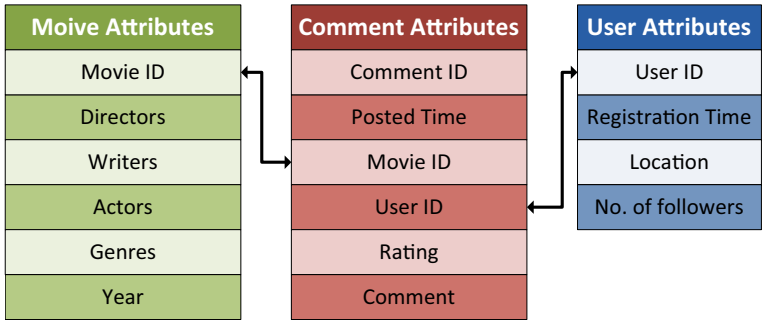


**Fig. 2** Data structure and relationships between Douban contents–films, reviews and users

**Table 1 Summary for generated features representing raw attributes**

| Category | Features |
|---|---|
| Film | Genres, movie rating, popularity |
| Review | Actor, director, writer, story, emotion, review rating |
| User | Location, activity, membership duration, leadership |

To simplify the explanation, a normalization procedure is introduced first. Let $\mathcal{S}$ be a finite set: $\mathcal{S} = \{s_i | \forall s_i \in \mathcal{S}\}$. The normalized function $\|\cdot\|$ is defined as follows:

$$\|s_i\| = \frac{s_i}{\max(\mathcal{S})}, \tag{3}$$

where $\max(\mathcal{S})$ represent the maximal value from $\mathcal{S}$. Herein the normalization procedure is implemented using the $\max(\cdot)$ function from Hive.

For the raw film information, as shown in the first row of Table 1, note that the "genres" feature can be extracted from the original genre attribute. In terms of movie rating ($R_i$) and popularity ($P_i$), they can be measured as follows:

$$R_i = \left\| \frac{\sum_{j=1}^{c_i} r_{i,j}}{c_i} \right\|, \tag{4}$$

$$P_i = \|c_i\|, \tag{5}$$

where $r_{i,j}$ represents the rate received from the $j$-th comment to the $i$-th film, and $c_i$ is the number of total comments for the $i$-th film.

The calculations used in Eqs. (4) and (5) reflect two basic assumptions: first, the average film rating is strongly associated with the ratings drawn from individual comments; second, the more comments a film receives, the more popular it is rated. The problem with this binary approach to feature measurement is that both outcomes from Eqs. (4) and (5) are continuous and not operable in real-world applications. Therefore, to quantify extracted features, the definition of a describable set is set out as follows:

**Definition 1** Let $\mathcal{Q} = \{q_1, q_2, \cdots, q_Q\}$ be a $Q$-item set for a particular domain $\psi$. Its describable set $\mathcal{D} = \{d_1, d_2, \cdots, d_D\}$ is a $D$-item set that satisfies the following conditions:

- $d_i (\forall d_i \in \mathcal{D})$ is a linguistic label that used to describe a certain status of $\psi$;
- $\mathcal{D}$ has less items than $\mathcal{Q}$, i.e., $Q < D$;
- $\psi(q_j) = d_i$, subject to $F(q_j, d_i) < F(q_j, d_k)$ $(\forall d_i, d_k \in \mathcal{D},\ i \neq k)$, where $\psi(\cdot)$ is a description function that mapping $q_j$ to an unique item $d_i$, and $F(\cdot)$ is a distance measure function (For simplicity, the Euclidean distance is employed herein).

As the definition shows, there exists a many-to-one relationship between the original and describable sets; that is, one item from $\mathcal{D}$ is used to describe or map more than one item in $\mathcal{Q}$. At the same time, each item from $\mathcal{D}$ is a linguistic label to describe a unique status of $\psi$. We further assume that these items (labels) are sorted in a particular order of status. A possible describable set for movie ratings, for instance, could be: $\mathcal{D} = \{high, medium, low\}$. And a typical set for describing film popularity is $\mathcal{D} = \{popular, medium, unpopular\}$.

Note that domain knowledge for $\psi$ is required to determine statuses ranging from high to low, and then to represent each status with one item in $\mathcal{D}$. Different business or operational requirements may result in a variety of statuses, thereby producing various

describable sets. The advantage of the describable set is to convert continuous data using discrete labels, thereby enhancing business operation and understanding of the significant trends present in UGC.

In general, a mapping algorithm is introduced to quantify continuous features by representing them with the describable set; see Algorithm 1. Later, this mapping algorithm is applied to quantify film features, such as movie ratings ($R_i$) and popularity ($P_i$), respectively.

**input** : The original set $\mathcal{Q}$, its describable set $\mathcal{D}$, and maximal iteration $T$;
**output**: A quantified feature ***y***.

*Initialize $D$ clusters by assigning randomly $q_j$ to one cluster ($\forall q_j \in \mathcal{Q}$), and $D$ is the number of items in $\mathcal{D}$;*
**for** $t = 1$ **to** $T$ **do**
    *Compute the probability of $q_j$ with respect to any one cluster;*
    *Assign $q_j$ to the $i$-th cluster that is with the biggest probability and establish the mapping function: $\psi(q_j) = i$;*
    *Recompute all cluster centers and record them as $\Delta_i$, $i = [1, D]$;*
    *Rank clusters using the center value: $I = sort(\Delta)$, and then map clusters with items: $d_i = I_i$;*
**end**
*Feature quantification: $y_j = I_{\psi(q_j)}$;*

**Algorithm 1:** The mapping algorithm for feature quantification

The next step in the feature generation stage is to extract informative features from film reviews (as shown from the second row in Table 1). To this end, the word segmentation and sentiment analysis technique is implemented. The word segmentation technique involves the division of complete sentences into their syntactical and semantic components. Whereas in English a space is usually used to separate words, Chinese has a very different structure whereby word division is either non-existent or operates in different ways. Thus, any procedure adopted for Chinese word segmentation must consider the specific language habits of Chinese users.

In this paper, we employ the LTP-Cloud service [37] for Chinese word segmentation. The LTP-Cloud service is an open-sourced Chinese language processing service, including part-of-speech tagging, named entity recognition, and so on. In addition, for each specific film studied, multiple corpora were also manually created to identify actor, director, writer, and story, with each category designed to elicit information about a particular aspect of the movie. Further details can be found in our preliminary research findings [38, 39].

Some general keywords used in these corpora are shown in Table 2. Note that, in practice, the actual names of actors, directors, or writers for specific movies are included as well. Word segmentation is then carried out on the basis of these corpora. In adopting this approach, our analysis of the dataset constitutes a major advance on previous studies, which rely either on data available in English-language sources (which present fewer

**Table 2 Corpus used in association with selected film attributes**

| Corpus | Keywords |
|---|---|
| Actor | 英雄(hero),女主角(actress),男主角(leading male character), 配角(supporting role),演员(actor) |
| Director | 导演(director) |
| Writer | 作家(writer),编剧(scriptwriter) |
| Story | 特效(special effects),场景(scene),台词(actor's lines),剧情(drama),逻辑(logic),故事(story), 节奏(rhythm),制作(production),情节(episode) |

problems than a Chinese-language dataset) or are restricted to a single, and thus limited, dictionary of Mandarin terms.

Based on the word segmentation exercise, four high-level features (i.e., "actor", "director", "writer", and "story") are generated, reflecting the incidence of keywords from four corpora in a single review. As an example, the word segmentation result from the user comment "这故事听着着实老套，但是导演和演员们的真诚都能带给观众发自心底的温暖" ("This story sounds old fashioned, but the director and actors' sincerity can be felt by the audience") is shown in Fig. 3. The words or phrases isolated within each box represent the word segmentation stage, while the matched keywords for the related corpus are found in the relevant boxes with dotted lines. Note that, for this example, a sequence of one, one, and one keyword is matched to the "actor", "director", and "story" corpora, respectively. Thus, the related four-feature vector is represented as: (1, 1, 0, 1), as no writer keywords were found.

Moreover, the sentiment analysis method in [40] is employed in this paper to classify the sentiment from comments. Then outcomes from sentiment analysis are taken as the value of the review "emotion" feature. At the same time, for the "review rating" feature, Eq. (6) is employed to convert the raw "Rating" attribute into a binary feature:

$$CR_{i,j} = \begin{cases} 1, & \text{if} \quad r_{i,j} > R_i \\ 0, & \text{if} \quad r_{i,j} \leq R_i, \end{cases} \tag{6}$$

where $r_{i,j}$ is the rating received from the $j$-th comment to the $i$-th film, and $R_i$ represents the average rating for the $i$-th film (see Eq. (4)). That is, individual ratings are divided into two groups according to their relationship to the average film rating. This binary approach to converting the rating attributes aids the processes in this preliminary study by simplifying the ratings as one of the key feature categories (listed in Table 3). Later, to quantify the six review features discussed above, Algorithm 1 is again employed.

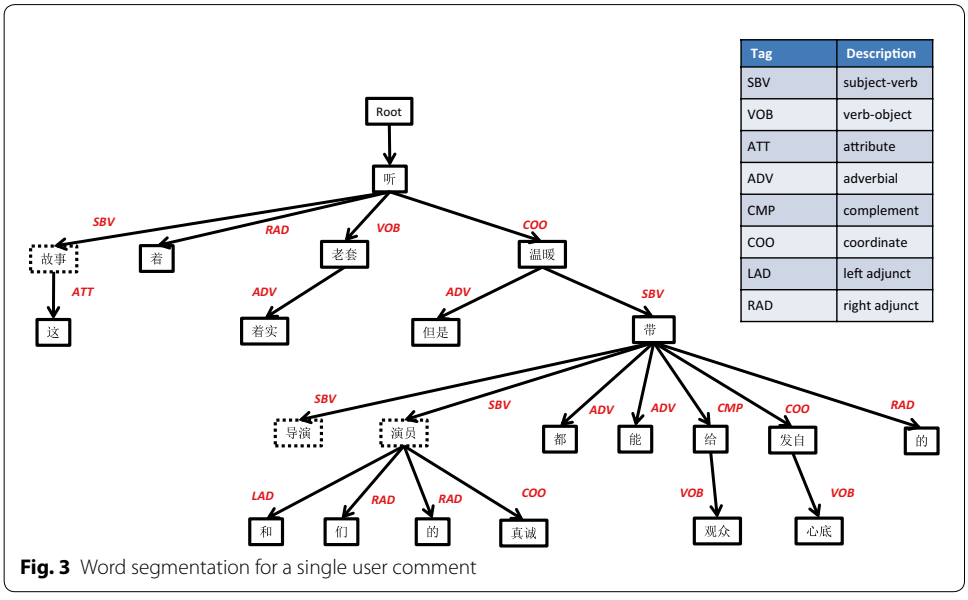


**Fig. 3** Word segmentation for a single user comment

For the user profile, four features (again shown in Table 1) are generated to simulate a given user's interests and behavior location, activity, membership duration, and leadership. Among these, "location" is a spatial feature derived from users' demographic information. However, the original location attribute collected is at the city level. We then develop an aggregation function in Hive to map users either from city to province (for Chinese users), or from city to country (for international users).

Furthermore, to measure other user's features, such as activity, membership duration, and leadership, the following definitions are proposed:

- The "*activity*" feature is computed from users' previous viewing lists. That is, the more Korean films watched by a user, the higher their activity level;
- The "*membership duration*" feature is used to measure the period of time elapsed since a user's registration. Thus, the earlier the date of registration with Douban, the longer the time-span;
- The "*leadership*" feature is to estimate the amount of influence exerted by a given user, which is related to the number of followers gained by a user since registration.

- Let $u_i$ be the total number of watched Korean films, $d_i$ the registration date, and $f_i$ the current number of followers for the $i$-th user, respectively. Based on three assumptions, the features of activity ($A_i$), membership duration ($S_i$), leadership ($L_i$) for user $u_i$ can be estimated as follows.

$$A_i = \|u_i\|, \tag{7}$$

$$S_i = \left\|date(d_i, T_{now})\right\|, \tag{8}$$

$$L_i = \left\| \frac{f_i}{date(d_i, T_{now})} \right\|, \tag{9}$$

**Table 3 Extracted high-level features and relevant quantification outcomes for Douban contents**

| Features | Quantification outcomes |
|---|---|
| Genres | 剧情(drama),喜剧(comedy),动作(action),惊悚(Thriller),爱情(love),犯罪(crime),历史(history),悬疑(Mystery) |
| Movie rating | High, medium, low |
| Popularity | Popular, medium, unpopular |
| Actor, director, writer, story | Mentioned more than twice, mentioned once, none |
| Emotion | Positive, neutral, negative |
| Review rating | High, low |
| Location | Name of Chinese provinces or foreign country |
| Activity | High, medium, low |
| Membership duration | Long, medium, short |
| Leadership | High, medium high, medium low, low |

where $T_{now}$ is a constant for current date, and the function $date(\cdot)$ is used to compute the number of days from $d_i$ to $T_{now}$. As before, Algorithm 1 is applied to quantify the continuous features.

In summary, by combining the aggregation function and proposed quantification algorithm, the feature generation module produces high-level features in the second stage of our proposed framework. A total of 13 features are produced from the raw samples representing three categories of contents on Douban: film, review and user. Table 3 shows the final outcome of the generated features.

#### *Content mining*

In the third and final stage of the Douban-Learning framework, we will find out significant correlations and extract new knowledge from high-level features in Table 3. Thus, the content mining module is implemented herein using association rule mining.

Rule mining is one of the most popular data mining tools used for such purposes due to its simplicity and efficiency. In general, association rules are used to describe dependence or correlation among features (or items). Thus, an association rule-based algorithm was considered for this module. A typical rule takes the form $\mathcal{A} \longrightarrow \mathcal{C}$, where $\mathcal{A}$ and $\mathcal{C}$ represent the antecedent and consequent set of the rule, respectively, and $\mathcal{A} \cap \mathcal{C} = \emptyset$. The rule implies that all items from $\mathcal{A}$ have a high probability of being associated with items from $\mathcal{C}$. *Support*, *confidence*, and *lift* are critical measurements for evaluating association rules, which are defined as follows [41]:

**Definition 2** Given $N$ data records, the support of $\mathcal{A}$ ($supp(\mathcal{A})$) is the proportion of records which contain all items from $\mathcal{A}$, which can be computed as follows:

$$supp(\mathcal{A}) = \frac{|\mathcal{A}|}{N}, \tag{10}$$

where $|\mathcal{A}|$ is the number of records contains $\mathcal{A}$. The confidence of the rule $\mathcal{A} \longrightarrow \mathcal{C}$ is computed as:

$$conf(\mathcal{A} \longrightarrow \mathcal{C}) = \frac{supp(\mathcal{A} \cup \mathcal{C})}{supp(\mathcal{A})}. \tag{11}$$

The lift of a rule ($\mathcal{A} \longrightarrow \mathcal{C}$) is the ratio between support and confidence, which can be computed as follows:

$$lift(\mathcal{A} \longrightarrow \mathcal{C}) = \frac{conf(\mathcal{A} \longrightarrow \mathcal{C})}{supp(\mathcal{C})}. \tag{12}$$

The measurements for support and confidence of rules reflects the item frequency, while lift is used to check the flexibility of rules (or the interestingness of rules [42]).

Nevertheless, traditional association rule mining algorithms, such as Apriori, encounter many practical problems when processing large data sets. For instance, Apriori employs a "bottom-up" strategy to produce different levels of frequent-item

sets. In addition, it requires repeated scanning of the entire dataset until all possible combinations of frequent sets are found. This typically leads to large-scale or high-dimensional results that exceed the processing capacity of a single computer. As a result, some improvements have been proposed to facilitate Apriori parallelization using *MapReduce*.

MapReduce is an effective process in the Hadoop platform for parallel computing and Big Data processing. A MapReduce process consists of two main phases, *map* and *reduce*. In the map phase, the input *key-value* pairs are processed individually by a map function and produce a second set of intermediate key-value pairs. The new pairs are then clustered according to their keys and provided as the input for the reduce function in the reduce phase. A third set of key-value pairs is then derived as the final output.

MapReduce-based processes dramatically improve the reliability and efficiency of conventional Apriori algorithms [43, 44]. However, a major limitation is that a large number of candidate association rules are generated, with many antecedent and/or consequent conditions. These conditions are neither easy to process (due to the large number of rules), nor straightforward to understand or interpret due to their complexity. To solve this problem, in this paper we limit the category of features contributing to the consequent set. That is, only some features (to be pre-defined by experts) are permitted to be included in the rule consequence process. To this end, a three-step parallel Apriori algorithm is proposed, of which is shown from Algorithm 2, Algorithm 3 and Algorithm 4, respectively. The proposed Apriori algorithm first applies the MapReduce process to generated Frequent-1 itemsets (step 1) and candidate Frequent-$k$ itemsets (step 2), respectively. The association rule sets are then obtained by combining frequent itemsets (step 3) as the support-confidence framework [45]. That is, rules satisfied minimal support and minimal confidence are selected. Furthermore, we also add a constraint by limiting the feature category in the consequent set. Therefore, the valid rules from the proposed Apriori algorithm will be a subset of the full rules. Again, this constraint is used to reduce the huge number of rules for better interpretation.

**input** : Number of total records $n$, minimal support $\mu$ and minimal confidence $c$;
**output**: Frequent-1 itemsets ($\mathcal{F}_1$).

**Procedure:** *Frequent-1 itemsets*
  **Function** *Map(Lineid, LineContent):*
    **for** $item$ *in LineContent* **do**
      *emit(item, 1)*;
    **end**
  **Function** *Reduce(item, valueList):*
    sum =0;
    **for** $value$ *in valueList* **do**
      sum += value;
    **end**
    **if** $sum > (\mu \times n)$ **then**
      *emit(item, sum)*;
    **end**
$\mathcal{F}_1$.append(item);

**Algorithm 2:** Producing Frequent-1 itemsets

```
input : Number of total records n, minimal support μ and minimal confidence c;
output: Frequent-k itemsets (F_k).
Procedure: Candidate Frequent-k itemsets
    k = 1;
    while F_k ≠ ∅ do
        Function Map(Lineid, LineContent):
            for item in Combine(F_k) do
                if item in LineContent then
                    emit(item, 1);
                end
            end
        Function Reduce(item, valueList):
            sum =0;
            for value in valueList do
                sum += value;
            end
            if sum > (μ × n) then
                emit(item, sum);
            end
        k = k + 1;
        F_k.append(item);
    end
```

**Algorithm 3:** Producing Frequent-$k$ itemsets

```
input : Number of total records n, minimal confidence c and consequent feature set I;
output: Association rule sets
Procedure: Association rule generation
    Function Map(Lineid, LineContent):
        for items in F_k do
            A, C=Separate(F_k);
            emit(A, C);
        end
    Function Reduce(A, C):
        if C ⊂ I then
            if conf(A ⟶ C) > c × n then
                emit(rule:A ⟶ C);
            end
        end
```

**Algorithm 4:** Generating association rule sets

## Experimental results

This section presents experimental results following application of the improved Apriori algorithm to features extracted for content mining. The cloud infrastructure employed is presented in "Cloud infrastructure", and the experimental setup and data sets are presented in "Experimental setup". The performance of the proposed framework is then evaluated in "Performance analysis".

### Cloud infrastructure

A virtual cluster is a simple but fast environment in which to build up the Hadoop framework. In the cloud infrastructure we implemented, a Dell server with Intel Xeon E5-2630 1.8 GHz cores and 32G memory is employed. A virtual cluster consisting of four nodes is then deployed. For each node, two virtual CPU and 4GB of memory is allocated. In addition, one node is set up as the master machine for Hadoop, while the remainder are used as slaver nodes. In addition, for the Hadoop platform, the 2.5 version

is installed. At the same time, the global controller for the data crawler is deployed in the same master machine as Hadoop, while workers are distributed to the slaver nodes. The Hive [36] tool is also implemented in the platform to pre-process raw data samples.

### Experimental setup

Table 4 shows the summary statistics for the data harvested from Douban for 114 Korean films until April 2015. During this period, a total of 714,946 comments are collected from 228,806 distinct users. Each film received, on average, around 6271 comments. In addition, of the 714,946 comments, 54,939 were made without allocating a rating. These statistics resulted in an average film rating of 3.7 (on a range of 1–5) for all films.

Figure 4 illustrates the geographic distribution of users who commented on the 114 Korean films in the study. About 80 % of users left location information, which for Chinese nationals is identified at the provincial level, including the municipalities of Beijing, Tianjin, Shanghai, and Chongqing (treated as provinces), while overseas users are labelled by their country of origin. Unsurprisingly, most Chinese users were located in developed cities and provinces such as Beijing, Shanghai, Guangdong, Zejiang, and Jiangsu, while users from the U.S. formed the largest single overseas group. Generally speaking, in these centres the user experience is enhanced by greater access to faster Internet speeds and a sophisticated public infrastructure for film exhibition. Therefore,

**Table 4 Descriptive statistics relating to film data collected**

| Variable | Mean | Standard deviation | Total |
|---|---|---|---|
| Number of comments | 6271 | 1489 | 714,946 |
| Rating | 3.7 | 1.2 | $2.468 \times 10^6$ |
| Number of users | 2007 | 1278 | 228,806 |

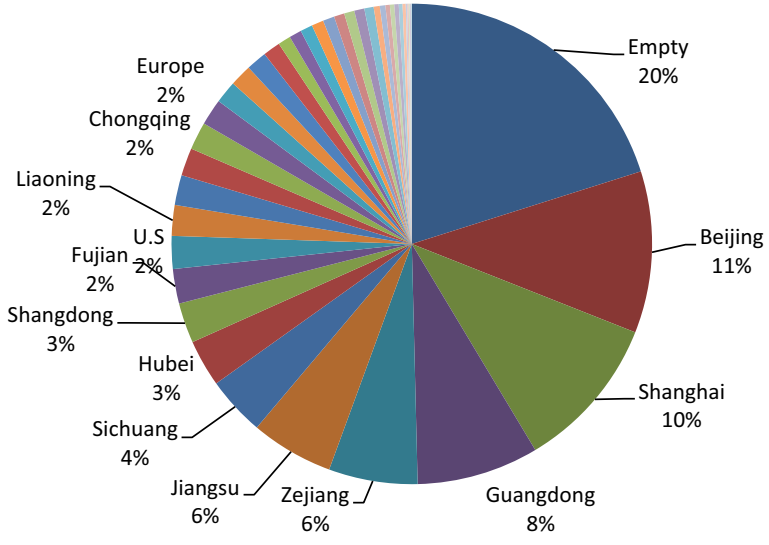


**Fig. 4** Geographic distribution of users commenting on the 114 Korean films in the study

the data illustrates significant participation from a greater number of users from these areas.

### Performance analysis

This section deals mainly with the effect of the minimum support threshold ($\mu$) on the performance of the improved Apriori algorithm. A smaller value for $\mu$ is more prone to generating more rules than a larger value for $\mu$. Nevertheless, a vast number of rules is difficult to understand and process, not to mention the computational costs involved. Our aim was then to test the robustness of the proposed algorithm (in particular the execution time and number of rules generated) according to various parameters. To this end, the impact of the minimum support threshold was considered by using different values for $\mu$ i.e., $\mu$ was set at 10, 20, 30, 40, and 50 %, respectively, while minimum confidence $c$ was set at 50 %. In other words, any rules satisfied $\mu$ and $c$ were selected.

Furthermore, in the proposed rule-mining algorithm, all features associated with films, reviews and users are included in the antecedent set; while film and review features are selected for the consequent feature set, with a view to understanding what factors might effect user comments or their preference. The proposed rule-mining algorithms are then applied to the full data set. The full data samples contain 660,007 comments (comments without allocating ratings are removed). For each comment, 13 features are extracted as mentioned in Table 3. Therefore, a total ($660{,}007 \times 13$) sample matrix is formed to represent the full data set. Apart from the proposed Apriori algorithm, we also benchmarked it with other algorithms for mining association rules, including traditional Apriori [46], Eclat [47], MapReduce-based Apriori (MRA) [48], and Spark-based Apriori (SBA) [49]—whose implementations are summarized at [50]. Both traditional Apriori and Eclat algorithms are single machine-based, which employ the generated-and-test mining strategy. On the other hand, MRA and SBA utilize the state-of-art cloud computing technologies. In particular, SBA is a cutting-edge algorithm based on Spark [51], in which data can be processed and cached in the machine memory.

Table 5 shows comparisons between five mining algorithms in terms of the number of rules extracted. As observed, with a decrease in the minimum support threshold, more rules are generated. For instance, the proposed Apriori algorithm produced 581 and 12,466 rules for $\mu = 50$ % and 10 %, respectively. Meanwhile, the proposed Apriori algorithm generated the minimal number of rules in five cases. That is because we only allowed certain features to form the consequent set, thereby eliminating unnecessary rules for easier interpretation. On the other hand, both traditional Apriori and Eclat

**Table 5 Comparison of number of generated rules with three association rules mining algorithms**

| Algorithm | $\mu = 50$ % | $\mu = 40$ % | $\mu = 30$ % | $\mu = 30$ % | $\mu = 10$ % |
|---|---|---|---|---|---|
| Apriori | 712 | × | × | × | × |
| Eclat | × | × | × | × | × |
| MRA | 712 | 2531 | 3984 | 5127 | 16,215 |
| SBA | 712 | 2531 | 3984 | 5127 | 16,215 |
| Proposed | 581 | 1438 | 225 | 3432 | 12,466 |

algorithms failed to generate association rules for any case of $\mu \leq 40$ %. That is, with a sample matrix of the (660,007 × 13)-dimension from the given problem, the amount of data exceeds the processing capacity of the traditional Apriori and Eclat techniques using a single machine, which leads to significant computational cost and extensive memory requirements.

In addition, Table 6 shows comparisons between execution times for different algorithms. The reported execution time reflects the entire mining process including loading data, rules mining, and the generation for output reporting. Unsurprisingly, the traditional Apriori and Eclat algorithms required much more time than other cloud computing-based mining algorithms when a huge data set is presented. When $\mu = 50$ %, for instance, the computational time for traditional Apriori and the proposed algorithm was approximately 5 days and 85.2 s, respectively. The much longer processing time recorded for traditional Apriori is related to the procedures for generating frequent itemsets. Again, both traditional Apriori and Eclat approaches fail to solve the problems of $\mu \leq 40$ %. In contrast, all parallel mining algorithms perform stably in terms of execution time. All association rules are found approximately 100 s. Furthermore, the proposed algorithm performs better than MRA by taking the less computational time, indicating its flexibility and suitability for Big Data mining investigated herein. In addition, the average execution time for the proposed (84.02 s) is slightly slower than the SBA algorithm (82.84 s). One reason for this is the SBA approach is processing the data within the machine memory, while the proposed algorithm is applying MapReduce on disk. The implementation of the proposed algorithm using Spark will improve its performance, and thus we leave this aspect of our work for future research.

#### Discussion

For generated association rules, we are more interested in rules with high lift (as defined in Eq. (12)), a measurement which reflects the interest value of rules [42]. Herein we summary some rules with high lift in Table 7. The minimum support threshold $\mu$ and confidence $c$ were set at 30 and 50 % respectively.

A few observations can be made: First, in terms of film features, there is a notable absence of significant rules that relate to a film's rating, popularity, or to a specific director or actor. This may be because the specific cohort under investigation at this time pays less attention to a commercial film's production context (for reasons that could be explored in a further study). Second, however, movie genre appears to be a significant feature (see Rule 1). The romance–drama (剧情) had the widest appeal among this

**Table 6 Summary of execution time as a function of $\mu$**

| Algorithm | $\mu = 50$ % | $\mu = 40$ % | $\mu = 30$ % | $\mu = 30$ % | $\mu = 10$ % |
|---|---|---|---|---|---|
| Apriori | 5 days | × | × | × | × |
| Eclat | × | × | × | × | × |
| MRA | 125.8 | 123.7 | 124.6 | 128.1 | 126.5 |
| SBA | 82.1 | 83.2 | 82.9 | 82.7 | 83.3 |
| Proposed | 85.2 | 84.8 | 82.2 | 81.6 | 86.3 |

**Table 7 Descriptive statistic of collected film data**

| Index | Rule | Confidence | Support | Lift |
|---|---|---|---|---|
| 1 | (genre: 剧情) ⟶ (emotion: positive, actor: never, director: never, story: never) | 0.787 | 0.513 | 1.022 |
| 2 | (membership duration: medium) ⟶ (genre: 剧情, emotion: positive) | 0.561 | 0.512 | 1.101 |
| 3 | (activity: high) ⟶ (emotion: positive, director: never) | 0.864 | 0.355 | 2.150 |

particular film community. This mirrors the fact that presently the largest percentage of films made in China are romance dramas.

These two observations have significant implications for the future of China's feature film industry given that since China joined the WTO in 2001, cooperation between the Chinese and Korean film industries has drawn them and their fans closer together. This relationship has blossomed via a handful of policy-driven co-produced romance dramas. It has also resulted in a much larger number of informal collaborations, including other genre films made by Chinese companies using Korean visual effects firms, foreign cast and crews, and shooting scenes on location in one or both countries. As a result of these "willing collaborations", Korean films now loom large among Douban users. In short, Douban users disseminate their opinions on a wide range of these international collaborative films, and to make recommendations to their followers and friends, thus potentially influencing film culture on a wide-scale.

Third and finally, in terms of user profiles, we discovered that there is an absence of observable and significant rules relating to "location" and "leadership" (see Rules 2 and 3). Medium users (i.e., membership duration is medium) are more interested in the genre and therefore respond with positive comments. A user's established preference for Korean films (i.e., activity is high) also leads to positive comments on new Korean films, although this group generally cares less about a film's director and more about its cast and story.

With these three observations in mind, the framework for our study could potentially be extended beyond Big Data processing not only of text-based film-related content generated by a variety of international users commenting on other country's films across different social network systems, but to encompass reams of data generated by next-generation media sources. For example, in the not-too-distant future, additional sources of user/customer sentiment analysis are likely to include real-time interactions experienced via new Hybrid Broadcast Broadband TV (HbbTV) platforms which will effectively remove the spotlight from online text-based social networking sites such as Douban. The ability to directly engage interactive feedback from audiences via multiple consoles, smart TVs and smart devices, for instance, will provide a just-in-time capacity to both content producers and also the traditional distribution channels that are currently the subject of social media sentiment analysis. Having said this, we trust that the preliminary methods and tools explored in our present study will underpin some of the new research initiatives on the horizon. The ability to process the types of structured and unstructured data at volume and scale derived from Douban and other SNS platforms is presently giving birth to a new field, whatever its future might be. Harvested at scale across multiple languages, platforms and geographies, eWOM is already being

aggregated and utilized by Western media pioneers like NetFlix and HBO in their content-generation and media strategies. Giant Chinese online and mobile video services such as Sohu, Youku, Sina Video, Tencent Video, and LeTV where Big Data harvesting and analysis is prevalent are surely heading in a similar direction. Perhaps some of the exploratory lessons offered here will provide them with additional food for thought.

## Conclusion

In this paper, we propose a complex framework for Big Data processing that can not be achieved with a single-machine utilizing average hardware and software. Three modules are introduced that are capable of crawling raw online records, generating key features to represent original samples in useful ways, and then running an association rule-mining algorithm on clouds for further content mining.

The proposed framework is implemented using the cutting-edge Hadoop platform, which is used as the fundamental tool for storing and processing harvested data sets. Thirteen high-level features are generated from three categories (film details, reviews, and user profiles) using aggregation functions, and the data is further quantified using the description set. More importantly, an improved parallel Apriori algorithm is proposed to discover significant correlations among these thirteen key features, with a view to expanding the analytical methods to a larger data set, that is, all film (or other popular culture) comments on Douban, and/or other future-generation OSNs.

In the wake of this preliminary and somewhat novel study, the proposed framework offers a flexible capability and efficient applicability for the processing of large amounts of social media data that in turn can be fed back to producers and distributors of both commercial and user-generated digital media contents.

Research on mining user-generated content, however, is still in its infancy, and therefore progress in this exciting arena must continue. The research work presented in this paper has only investigated a small area using big data techniques. There are many possibilities for future research directions and improvements, including the implementation of content mining using other rule mining algorithms (such as the Frequent Pattern Growth (FP-Growth) strategy), as well as using Spark or Apache Tez platform to achieve better performance. One of very next tasks is to investigate users' network structure to identify leadership links and trends among Douban's expansive user and follower networks.

**Authors' contributions**

JY and BY are the principal researchers for the work proposed in this paper. JY's contributions include collecting data, initial drafting of the article, and coding implementation. BY plays an important role in providing the background investigation and editing the article. Both authors read and approved the final manuscript.

**Author details**

[1] SMART Infrastructure Facility, Faculty of Engineering and Information Sciences, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia. [2] School of Humanities and Social Inquiry, Faculty of Law, Humanities and the Arts, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia.

**Acknowledgements**

This article draws on research currently being conducted in association with an Australian Research Council Discovery project, Willing Collaborators: Negotiating Change in East Asian media production DP 140101643.

**Competing interests**

The authors declare that they have no competing interests.

Received: 16 April 2015 Accepted: 21 December 2015
Published online: 13 January 2016

**References**

1. Mayer-SchÖnberger V, Cukier K. Big data: a revolution that will transform how we live, work and think. New York: Houghton Mifflin Harcourt Publishing Company; 2013.
2. Koh NS, Hu N, Clemons EK. Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures. Electron Commer Res Appl 2010; 9(5):374–85 (Special Section on Strategy, Economics and Electronic Commerce).
3. Singh VK, Piryani R, Uddin A, Waila P. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In: International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013; p. 712–7.
4. Yao R, Chen J. Predicting movie sales revenue using online reviews. In: IEEE International Conference on Granular Computing (GrC), 2013; p. 396–401.
5. Chang A, Liao J-F, Chang P-C, Teng C-H, Chen M-H. Application of artificial immune systems combines collaborative filtering in movie recommendation system. In: International Conference on Computer Supported Cooperative Work in Design (CSCWD); 2014. p. 277–82.
6. Samad A, Basari H, Burairah H, Ananta GP, Junta Z. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Eng. 2013;53:453–62.
7. Yu X, Liu Y, Huang X, An A. Mining online reviews for predicting sales performance: a case study in the movie domain. IEEE Trans Knowl Data Eng. 2012;24(4):720–34.
8. Amolochitis E, Christou IT, Tan Z-H. Implementing a commercial-strength parallel hybrid movie recommendation engine. IEEE Intell Syst. 2014;29(2):92–6.
9. Zang W, Zhang P, Zhou C, Guo L. Comparative study between incremental and ensemble learning on data streams: case study. J Big Data. 2014;1:1–5.
10. Liu X, Wang X, Matwin S, Nathalie J. Meta-mapreduce for scalable data mining. J Big Data. 2015;2(1):14.
11. Liu C-L, Hsaio W-H, Lee C-H, Lu G-C, Jou E. Movie rating and review summarization in mobile environment. IEEE Trans Syst Man Cybern Part C Appl Rev. 2012;42(3):397–407.
12. Wong FMF, Sen S, Chiang M. Why watching movie tweets won't tell the whole story? In: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, New York, NY, USA; 2012. p. 61–6.
13. Singh VK, Piryani R, Uddin A, Waila P. Sentiment analysis of movie reviews and blog posts. In: IEEE 3rd International Advance Computing Conference (IACC); 2013. p. 893–98.
14. Hodeghatta UR. Sentiment analysis of hollywood movies on twitter. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2013. p. 1401–4.
15. Mouthami K, Devi KN, Bhaskaran VM. Sentiment analysis and classification based on textual reviews. In: International Conference on Information Communication and Embedded Systems (ICICES); 2013. p. 271–6.
16. Gupta N, Abhinav KR. Fuzzy sentiment analysis on microblogs for movie revenue prediction. In: International Conference on Emerging Trends in Communication, Control, Signal Processing Computing Applications (C2SPCA); 2013. p. 1–4.
17. Yeap JAL, Ignatius J, Ramayah T. Determining consumers' most preferred ewom platform for movie reviews: a fuzzy analytic hierarchy process approach. Comput Human Behav. 2014;31:250–8.
18. Rui H, Liu Y, Whinston A. Whose and what chatter matters? the effect of tweets on movie sales. Decision Support Syst. 2013;55(4):863–70.
19. Delen D, Sharda R. Predicting the financial success of hollywood movies using an information fusion approach. Indus Eng J. 2010;21(1):30–7.
20. Ghiassi M, Lio D, Moon B. Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Syst Appl. 2015;42(6):3176–93.
21. Barrio JB, Rubio XA. Geolocated movie recommendations based on expert collaborative filtering. In: Proceedings of the Fourth ACM Conference on Recommender Systems, New York, NY, USA; 2010. p. 347–8.
22. Singh V, Mukherjee M, Mehta G. Combining collaborative filtering and sentiment classification for improved movie recommendations. In: Multi-disciplinary Trends in Artificial Intelligence. Lecture Notes in Computer Science, vol. 7080; 2011. p. 38–50.
23. Kawase R, Nunes BP, Siehndel P. Content-based movie recommendation within learning contexts. In: International Conference on Advanced Learning Technologies (ICALT); 2013. p. 171–3.
24. Nessel J, Cimpa B. The movieoracle-content based movie recommendations. In: International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 3; 2011. p. 361–64.
25. Dumitras A, Haskell BG. Content-based movie coding—an overview. In: IEEE Workshop on Multimedia Signal Processing; 2002. p. 89–92.
26. Wang Z, Yu X, Feng N, Wang Z. An improved collaborative movie recommendation system using computational intelligence. J Visual Lang Comput. 2014;25(6):667–75.
27. Li Q, Kim BM. Clustering approach for hybrid recommender system. In: Proceedings of IEEE/WIC International Conference on Web Intelligence; 2003. p. 33–8.
28. Trattner C, Kappe F. Social stream marketing on facebook: a case study. Int J Soc Humanist Comput. 2013;2(1):86–103.
29. Jansen BJ, Zhang M, Sobel K, Chowdury A. Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol. 2009;60(11):2169–88.
30. Harri AL, Rea A. Web 2.0 and virtual world technologies. J IS Educ. 2009;20(2):137–44.
31. Eisenfeld B, Fluss D. Contact centres in the web 2.0 world. CRM Mag. 2009;13(2):48–9.

32. Cheng JW, Mitomoa H, Otsukab T, Jeonc S. The effects of ict and mass media in post-disaster recovery—a two model case study of the great east japan earthquake. Telecommun Policy. 2013;39(6):515–32.
33. Yecies B. Inroads for cultural traffic: breeding korea's cinematiger. In: Black D, Epstein S, Tokita A, editors. Complicated currents: media production, the Korean wave, and soft power in East Asia. Melbourne: Monash University EPress; 2010.
34. Nagwani NK. Summarizing large text collection using topic modeling and clustering based on mapreduce framework. J Big Data. 2015;2(1):6.
35. https://www.kobis.or.kr/kobis/business/main/main.do.
36. https://hive.apache.org/.
37. http://www.ltp-cloud.com/intro/en/.
38. Yecies B, Yang J, Berryman M, Soh K. Marketing bait (2015) Using smart data to identify e-guanxi among china's 'internet aborigines'. In: Film Marketing into the twenty-first century. British Film Institute.
39. Yecies B, Yang J, Berryman M, Soh K. Korean female writer–directors and smart analysis of their reception on china's social media scene. In: Women Screenwriters: An International Guide. Palgrave Macmillan (forthcoming)
40. Yuan C, Zhuang Y, Li H. Semantic based chinese sentence sentiment analysis. Int Conf Fuzzy Syst Knowl Discov (FSKD). 2011;4:2099–103.
41. Zhang C, Zhang S. Association rule mining: models and algorithms. In: Lecture Notes in Computer Science; 2307. Lecture Notes in Artificial Intelligence, New York: Springer, C2002. 2002.
42. Brin S, Motwani R, Jeffrey D. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data; 1997. p. 265–76.
43. Lin M-Y, Lee P-Y, Hsueh S-C. Apriori-based frequent itemset mining algorithms on mapreduce. In: Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication. ICUIMC '12; 2012. p. 1–8.
44. Li N, Zeng L, He Q, Shi Z. Parallel implementation of apriori algorithm based on mapreduce. In: 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD); 2012. p. 236–41.
45. Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective. IEEE Trans Knowl Data Eng. 1993;5(6):914–25.
46. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases; 1994. p. 487–99.
47. Liu L, Li E, Zhang Y, Tang Z. Optimization of frequent itemset mining on multiple-core processor. In: Proceedings of the 33rd International Conference on Very Large Databases; 2007. p. 1275–85.
48. Othman Y, Osman H, Ehab E. An efficient implementation of apriori algorithm based on Hadoop–Mapreduce model. Int J Rev Comput. 2012;12:59–67.
49. Qiu H, Gu R, Yuan C, Huang Y. Yafim: a parallel frequent itemset mining algorithm with spark. In: Proceedings of the 2014 IEEE International Parallel & Distributed Processing Symposium Workshops; 2014. p. 1664–71.
50. https://jackyanguow.wordpress.com/2015/11/04/source-codes-for-parallel-association-rule-mining/.
51. https://spark.apache.org/.