Method of social network analysis and visualisation based on reviews

Xiaowei Xu¹, Wei Liu¹, Ye Tao², Xiaodong Wang¹, Jilong Wu¹
¹Department of Computer and Technology, Ocean University of China, Qingdao, People's Republic of China
²School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, People's Republic of China
E-mail: xuxw525@ouc.edu.cn

Abstract: With the development of e-commerce platforms, the huge amount of transaction volume produces a lot of commentary information. To study the relationship between users and user comments and corresponding products, a social network analysis method based on user comments is proposed. The authors regard the network connection between goods and user comments as a problem on the network and use the spectral clustering algorithm to maintain the structural characteristics of the network. This method can visualise the network structure features such as user clustering and distribution, which can help to understand the network from a macro perspective and be directly applied to the visualisation of other comment-based networks. It is significant for analysing the relationship between users and user comments and corresponding products. The comparative visualisation experiments are also discussed, to validate the effectiveness of the proposed method.

1 Introduction
With the development of the Internet, e-commerce has gradually entered a mature stage. Under this trend, various e-commerce platforms have emerged, and the number of online transactions of users has increased. During the whole transaction process, before the users finally purchase the product, they usually leave a large amount of information on the e-commerce platform, which can reflect the users’ interaction behaviour pattern with the product. The interactions between users and products can be divided into ‘browse’, ‘favourites’, ‘add to Cart’ and ‘purchase’. Currently, mainstream e-commerce platforms typically analyse user behaviour based on this information, and build classifiers through machine learning and deep learning techniques to predict purchase intent [1]. Alternatively, using data mining methods to analyse a large amount of data created by a user's online search information, which facilitates a better understanding of the user's shopping habits and preferences, thereby predicting the user's shopping behaviour [2]. The focus of these studies is to study the user's historical behaviour to predict the user's future behaviour, and to lack a discussion of the relationship between users and the relationship between product information. How to reasonably express the relationship between users–users and users–commodities becomes a research hotspot.

The emergence and rapid development of social networking applications have largely changed the ways to use the Internet. The social network is a data structure composed of relationships between multiple nodes and nodes [3]. The node represents a person or a participant in a social network [4]. Using the relationship among nodes, social networks can link a variety of social relationships, not only strangers who meet by chance, but also family members who are closely related [5]. Today's e-commerce platform has more social elements, and the communication between users is also social. Considering users and commodities in the e-commerce platform as nodes in the social network, the analysis method of the social network is used to analyse the relationship between the e-commerce platform products and users. This method has become a viable method for studying e-commerce users–users and users–commodities relationships.

There is a large amount of interactive information between the goods and users in the e-commerce platform. How to select the appropriate information to study the users–users and the users–commodities relationship becomes a difficult point of research. Online reviews are specific feedback, coming from consumers about their experience with the products [6]. Mining user comments from review information will improve the user's preference model and improve recommendation accuracy [7]. Therefore, the description of the users–users and the users–commodities relationship through the information in the online comment is required to study the relationship between the e-commerce users and the goods.

In this article, a method of social network analysis and visualisation is proposed to solve the problem of reasonably expressing the relationship between users–users and users–commodities. There are several advantages:

• The impact of the comment is presented by quantified values.
• Intuitively expresses the relationship between user reviews.

The rest of the paper is organised as follows: Section 2 briefly introduces the data description and processing, which mainly describes the source and content of the data. Section 3 describes the research and modelling methods, and introduces PageRank (PR) and spectral clustering methods. Section 4 presents the experiment and analysis our method. Firstly, this section uses the appropriate comment information to study the relationship between users. Secondly, the PR method will be used to derive a ranking of the influence of comments in a single product, which is used to determine the relationship between the user and the product. Thirdly, the clustering analysis of user comments is carried out by means of spectral clustering, and the results of user clustering and distribution are obtained, and the relationship between users and users is analysed. Finally, Section 5 concludes the paper and discusses further research directions.

2 Data description and processing
This dataset consists of movie reviews from Amazon. The data span a period of >10 years, including all ~5 million reviews up to October 2012. Reviews include product and user information, ratings, and a plaintext review [8]. Each review in this data set includes 'produId', 'userId', 'profileName', 'helpfulness', 'score', 'time', 'summary' and 'text'. In this paper, we use four of them, including 'produId', 'userId', 'helpfulness' and 'score'. For the acquired data, we need to perform data pre-processing. The specific steps are as follows:

• Converting the txt file of the original dataset to a csv file.

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
• Classifying user reviews by movie.
• Selecting the four fields you need for your experiment, including ‘produId’, ‘userId’, ‘helpfulness’, ‘score’.
• Dividing user reviews by movie name.
• Dividing movies by user review.
• Utilising the movie's user comments

3 Methods research and modelling

3.1 PageRank

The core idea of the algorithm is if a page sort is linked by many other web pages, the page is more important. If this highly important web page links to another web page, the importance of the linked web page will increase [9]. The PR algorithm gives each web page an initial value PR value to represent the importance of the web page in the initial state. Many web pages in the network can be viewed as a directed graph, and each web page in the network can be viewed as a node. The PR algorithm iterates with the given PR value until it converges. Since the movie review network and the network structure of web reference are similar, and the PR algorithm is a link relationship-sorting algorithm, PR can be used for the ordering of the movie commentary influence. It is generally believed that the same score is given to the film in the same movie review and the higher quality of the comment (more help) indicates an identity relationship. If a film has many influential comments, it means that the film has a high influence.

3.2 Clustering

Spectral clustering is an algorithm that evolved from graph theory and it has been widely used in clustering [10]. The main idea of the algorithm is to treat all the data as points in space, which can be connected by edges. The edge weight value between the two distant points is lower, and the edge weight value between the two closer distances is higher. By cutting the graph composed of all the data points, the cut graph is different. The weights between the sub-graphs are as low as possible, and the weights in the sub-graphs are as high as possible, thus achieving the purpose of clustering.

The spectral clustering algorithm kernel is an optimisation subproblem $t_w$ that transforms the clustering problem into a graph. As an example, each data object is regarded as a vertex $V$ in the graph, and the edge $E$ between the objects is given a certain weight value $W$ according to the similarity between the objects so that a weighted graph based on the sample data is obtained $G = (V, E)$. The principle of dividing the graph $G$ is to maximise the similarity between the sub-graphs divided, and the similarity between the different sub-graphs is the smallest.

The process of the spectral clustering algorithm can be summarised in the following three steps:

• Constructing a sample similarity matrix based on the input similarity matrix transformation rules.
• Calculating the eigenvalues of the constructed sample similarity matrix, the eigenvectors and establish the vector space
• Using the corresponding clustering algorithm to cluster the feature vectors in vector space.

3.3 Cluster modelling

Definition 1: The square of the Euclidean distance between two data samples is expressed as $D_{ij}$, which is defined as

$$D_{ij} = \sum_{m=1}^{M} (x_{im} - x_{jm})^2 \quad (1)$$

Definition 2: The Gaussian kernel function is defined as (2), where $\sigma$ is the ruler parameter

$$k(x,y) = \exp\left(-\frac{\|x - y\|_2}{2\sigma^2}\right) \quad (2)$$

The classical spectral clustering algorithm is shown in Fig. 1.

**Input:**
- Sample set: $S = \{s_1, s_2, s_3, s_4, \ldots, s_n\}$
- Number of clusters: $k$
- Scale parameter: $\sigma$

**Output:**
- Clustering result

**Step:**

![Spectral clustering modelling process](image_url)

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
Constructing a similar matrix $W \in \mathbb{R}^{n \times m}$, the elements in the matrix are defined as follows. The similarity function $W_{ij}$ is used to measure the similarity between the two sets of data. If the two sets of data are similar, the similarity function has a higher value. The similarity function generally uses a Gaussian similar function

$$W_{ij} = \exp\left(-\frac{\| s_i - s_j \|^2}{2\sigma^2}\right)$$  \hspace{1cm} (3)

$W_{ij}$ represents the similarity between the $i$th data $S_i$ and the $j$th data $S_j$, $\sigma$ represents a natural constant, and $\sigma$ represents a bandwidth of a Gaussian similar function.

When $i = j$, $W_{ii} = 0$.

- Constructing degree matrix $D$, the elements $D(i, j)$ on the main diagonal of the matrix $W_{ij}$ are the sum of the $i$th row elements of the similarity matrix, then construct a constructed Laplacian matrix $L = D^{-1/2}AD^{-1/2}$.
- Performing eigenvalue decomposition on Laplacian matrix $L$, calculate eigenvalues, find the eigenvectors $x_1$, $x_2$, $x_k$ corresponding to the first $k$ eigenvalues, and then construct the matrix $X$

$$X = [x_1, x_2, x_3, \ldots, x_k]$$  \hspace{1cm} (4)

- Normalise the row vector of matrix $X$, the normalised matrix is $Y$

$$Y_{ij} = \frac{X_{ij}}{\sum X_{ij}^2}$$  \hspace{1cm} (5)

Mapping data to a range of 0 to 1 makes data processing faster and easier.

- Each row in the matrix $Y$ is treated as a sample in the space $R_k$, where the number of samples is $n$ and the sample dimension is $k$, and the feature vectors are clustered using $k$-means or other classical clustering algorithms.
- The sample point $s_i$ is divided into the $j$th class if and only if the $i$th row of the matrix $Y$ is divided into the $j$th class.

## 4 Experiment and analysis

### 4.1 Construct adjacency matrix

We use adjacency matrices, in other words, store data in a two-dimensional array and use a matrix to build the model, which gives each vertex and other vertices a chance of representation of edges. If there is an edge, they have a point of 1, otherwise 0. (Of course, if it is a graph of the edge of the right, the intersection stores the weight of their side). The relationship between users is represented by a directed adjacency matrix (Fig. 2).

The step of constructing the adjacency matrix is as follows:

- Selecting a movie (symbol is ‘movie A’)
- Getting User list of user comments in the movie ‘user_list’
- Pressing user to get a list of movies (m1–mn) that have been commented by user (remove ‘movie A’)
- Getting the same reviews of ‘user’ in the movie ‘movie_i’ as the user rating ‘user_i’
- Determining if ‘ui’ has help, if there is any influence on user, add ‘user_i’ to user’s influence list
- Whether the ‘user’ in the ‘user_list’ in ‘movie_A’ appears in the influence list of other ‘users’, and if so, the other ‘users’ have influence on the ‘user’.

### 4.2 Identify commentary influence with PR

Select a movie and follow the steps in Section 4.1 to derive the adjacency matrix. From the adjacency matrix generated, we can construct a transition matrix such that a column with an outer $m$ chain will have a value of $1/m$ in each $m$ unit. In addition, get the PR score through the transfer matrix to identify influential comments.

In general, a PR value is given to each comment in advance. Since the PR value is a commented probability of influence, it is generally $1/m$, where $m$ is the total number of comments. The sum of the PR values of all comments is 1. After a PR value is given in advance, it is iterated through the following algorithm until it reaches a smooth distribution.

As shown in Table 1, we list the top 10 of the result. The higher the value of PR in the table, the higher the impact of this comment.

### 4.3 Spectral clustering experiment

Spectral clustering can visually display the distribution of user comments using visualisation methods. The same colour in the figure represents the same class. As is Fig. 3 shown, it is clear from the clustering diagram that user comments are mainly distributed in two areas. The left area indicates the aggregation area of user comment. Fig. 4 shows the clustering results obtained using the $k$-means algorithm [11].

Through the visualisation of the movie user comments in Figs. 3 and 4, the following conclusions can be drawn:
• The distribution of active users and inactive users is divided into two levels, and the two types of users are each concentrated. This shows that there are significant differences between influential users and inactive users.
• Inactive users are more closely distributed than influential users, which means that inactivity is inactive in most cases. In addition, active users have no influence on comments in some cases.
• There are some other categories of users in both types of users, and there is no absolute distinction between the two types of users. In some cases, the two sides can transform into each other.
• Compared with the $k$-means algorithm [11], the classification of spectral clustering algorithms is more detailed. The spectral clustering algorithm is better than the $k$-means clustering algorithm in processing high-dimensional data clustering, and it is more advantageous to deal with clustering of sparse data.

5 Conclusion
This paper proposes a review-based method of social network analysis and visualisation. The product and user comments are networked with each other, PR and other methods quantify the influence of user comments, and the spectral clustering algorithm is used to maintain the structural characteristics of the network. This method can visualise the relationship, visualise the user clustering and distribution, and the influence of the comments; thus helping the researchers understand the network structure from a macro perspective. It has important reference significance for the analysis of the relationship between users and their comments and commodities.

6 Acknowledgments
This research is supported by the National Key Research Development Program of China (No. 2017YFB1104205).

7 References
[1] Vieira, A.: ‘Predicting online user behaviour using deep learning algorithms’, Comput. Sci., 2015, 46, (8), pp. 127–135
[2] Curne, C., Preis, T., Eugene Stanley, H., et al.: ‘Quantifying the semantics of search behavior before stock market moves’, Proc. Natl. Acad. Sci. USA, 2014, 111, (32), pp. 11600–11605
[3] Fortunato, S.: ‘Community detection in graphs’, Phys. Rep., 2009, 486, (3), pp. 75–174
[4] Gest, S.D., Graham-Bermann, S.A., Hartup, W.W.: ‘Peer experience: common and unique features of number of friendships, social network centrality, and sociometric Status’, Soc. Dev., 2001, 10, (1), pp. 23–40
[5] Fang, B., Xu, J., Li, J., et al. ‘Online social network analysis’, Publ. House Electron. Ind., 2015, 30, (2), pp. 187–199
[6] Mudambi, S.M., Schuff, D.: ‘What makes a helpful review? A study of customer reviews on Amazon.com’, MIS Quarterly, 2010, 34, (1), pp. 185–200
[7] Chen, L., Chen, G., Wang, F.: ‘Recommender systems based on user reviews: the state of the art’, User Model. User-Adapt. Interact., 2015, 25, (2), pp. 99–154
[8] McAuley, J., Leskovec, J.: ‘From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews’, WWW, Rio de Janeiro, Brazil, 2013, pp. 897–908
[9] Shi, Y.H., Eberhart, R.: ‘A modified particle swarm optimizer’. 1998 Proc. of the IEEE Conf. on Evolutionary Computation, Anchorage, 4-9 May 1998, pp. 69–73
[10] Jain, A.K., Murty, M.N., Flynn, P.J.: ‘Data clustering: a review’, ACM Comput. Surv., 1999, 31, (3), pp. 264–323
[11] Malik, H., Laghari, N., Sangrahi, D.M., et al.: ‘Comparative analysis of hybrid clustering algorithm on different dataset’. 2018 8th Int. Conf. on Electronics Information and Emergency Communication (ICEIEC), Beijing, 2018, pp. 25–30

Fig. 4 Visualisation of $K$-means

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)