

Chapter 2

Related Work

In group participated web applications, comments and ratings made by users can be collected by system owner or web spider program. This information reflects people's opinions on specific objects. Mining this information can produce new knowledge that individual cannot provide in the past. In this chapter, we first review two research areas related to user opinion analysis and prediction, including recommendation algorithm and sentiment analysis.

The interaction of different people in social network forms several types of relationship, by which the network constructed shows some interesting statistical properties and evolution laws. The structure of the network has a significant impact on the propagation of information. These properties and laws provide theoretical basis and methodology support to make prediction on collective view through trust network. Thus, we will also review some related research on dynamic network mechanism in this chapter.

In the field of computer science, trust is a widely used term. Mainly related research sub-areas include: federated authentication, credible third party payment platform, reputation system. Jøsang et al. [23] summarized these research sub-areas as "trust management" research area. In the end of this chapter, we will give a brief review on these research sub-areas and describe the relationship between them and the research work in this book.

2.1 Recommendation Algorithm

Recommendation algorithm provides sorting prediction and rating prediction through the analysis of users' preference.

By analyzing users' preference, sorting prediction aims at providing a recommendation list with limited length for specific user from mass information to improve user's experience. Given user x and its request q , content set I , length of content list l , sorting prediction filters out the content set I_s which satisfies specific condition:

$$I_s = \arg \max_{I_s \in I \wedge |I_s|=l} \sum_{i \in I_s} P(i \text{ is relevant} | u_x, q) \quad (2.1)$$

and makes it most relevant to the target user. Sorting prediction can be divided into free prediction and conditional prediction according to the scope. The former is to filter among the whole content set without clear request. For example, Google provides news recommendation service. The latter is to filter among the content subset which is relevant to user's specific query. For example, in the movie rating website, the system recommends movie list according to the type of movie selected by user.

Sorting prediction can help users find their interested information, but it is difficult to evaluate the quality of the information. So in many web applications, users are allowed to rate the content. Rating is expressed by internal value and every value represents a clear semantic. For example, 1 means very bad, 5 means very good. Rating prediction is used to complete the following task: given user x and an unrated object y to determine what the possible rating is. In probability, this can be expressed as calculating the probabilities of each rating value and taking the rating with the maximum probability as the output, as Eq. (2.2) shows.

$$\rho'(u_x, i_y) = \arg \max_v P(\rho(u_x, i_y) = v | u_x) \quad (2.2)$$

As rating is always expressed by consecutive integers, it is not a good choice to treat each different value as an independent class label. Thus, expectation form like Eq. (2.3) is always adopted in literatures to do the weighted calculation.

$$\rho'(u_x, i_y) = \sum_v v \cdot P(\rho(u_x, i_y) = v | u_x) \quad (2.3)$$

In practical Web systems, rating prediction and sorting prediction are always combined with each other. The system usually provides a recommendation list and makes rating prediction on each item in the list. For example, in product rating website, the system usually recommends product for users and provides a comprehensive rating on the product. When combining rating prediction with sorting prediction, we should consider the priority problem of them. When sorting prediction gets higher priority, it may appear that items in the recommendation list have high relevance but the predicted rating is low. In the application of product recommendation, it doesn't make any sense to recommend bad product to users. When rating prediction gets higher priority, it may appear that items in the recommendation list have low relevance. This situation should also be avoided. A possible solution is to do sorting prediction first and set a threshold value for the rating of the items in the list. Thus, only item with rating which is higher than the threshold will be displayed.

Both sorting prediction and rating prediction come down to the calculation of the utility value $\rho'(u_x, i_y)$ of a recommendation candidate to a target user [7]. In sorting prediction, utility value corresponds to the relevance degree. While in rating prediction, utility value corresponds to a specific rating value.

Recommender systems are usually classified into three categories, based on how recommendations are made: content-based methods, collaborative filtering (CF) methods, and hybrid approaches. Content-based recommendation calculates the utility value based on the target user's historical visited data and the text features of the candidate items for recommendation. Collaborative filtering (CF) tries to find desired items based on the preference of the set of similar users. In order to find out like-minded users, it compares other user's ratings with the target user's ratings. And then the target user will be recommended items that people with similar tastes and preferences liked in the past. It is not necessary to analyze the contents of items, therefore, it can be applied to many kind of domains there a textual description is not available. Hybrid approach combines collaborative and content-based methods, which helps to avoid certain limitations of content-based and collaborative systems. In this section, we will introduce content-based method and hybrid approach in brief. In the next section, we will introduce CF method in detail, due to its successful using in business and highly correlation to our work.

2.1.1 Content-Based Recommendation

Content-based recommendation calculates the utility value based on the target user's historical visit data and the text features of recommended items. These kinds of approaches can be divided into two categories, instance-based learning algorithm and model-based learning algorithm. Both of them need to express user's preference and recommended candidates as the vector of attributes.

We can get the attribute vector of recommended candidates by text processing technology or by manual labeling (For example, Web 2.0 websites widely adopt the Tagging function). TF-IDF [24] (term frequency/inverse document frequency) is a representative algorithm for text factorization. Let $\vec{\omega}_y = (\omega_{y1}, \omega_{y2}, \dots, \omega_{yl})$ represents the attribute vector of the recommended text item i_y , in which ω_{yl} represents the l th attribute's weight. Suppose that the number of texts can be recommended is N , and keyword k_l is occurred in n_l texts. Let f_{yl} represent the times of keyword k_l occurred in the text i_y , then the term frequency TF_{yl} is defined as follows:

$$TF_{yl} = \frac{f_{yl}}{\max_l f_{yl}} \quad (2.4)$$

The max function returns the maximum number of occurrences of all the key words in text i_y . Besides, as keywords which occurs frequently in all documents bring little information, IDF_l is used to weight TF_{yl} :

$$IDF_l = \log \frac{N}{n_l} \quad (2.5)$$

$$\omega_{yl} = TF_{yl} \times IDF_l \quad (2.6)$$

In practical applications, the text is usually represented by a certain number of keywords. For example, in Fab [25] system, 100 key words with highest weight are used to represent a web page. Given target user u_x , we can calculate all the attribute vectors of texts created or rated by the user with Eqs. (2.4–2.6). Combining with information of user rating, we can get the vector representation of user preference: $\vec{w}_x = (\omega_{x1}, \omega_{x2}, \dots, \omega_{xl})$. There are many kinds of approaches to calculate the \vec{w}_u , such as simple weighted calculation, Rocchio [26] and Winnow [27] algorithm. After getting the attribute vector representation of user preference, we can get the utility value of text by comparing it with the content vector of text which can be recommended. Cosine measure is a commonly used similarity measurement. The specific formula of cosine measure is as follows:

$$\rho'(u_x, i_y) = \cos(\vec{w}_x, \vec{w}_y) = \frac{\vec{w}_x \cdot \vec{w}_y}{|\vec{w}_x| \cdot |\vec{w}_y|} \quad (2.7)$$

Model-based approaches archive each user's prediction model through training sample set. Multiple attribute values of the information are taken as the input, and the utility value is taken as the output. Typical model-based algorithms include Bayesian classification [28, 29], clustering, neural network, decision tree [29] and support vector machine [30]. A novel prediction method is proposed in [31], it generalizes a related model from user click behavior and web page data to recommend web page. Although the model is content-based, it implements content decoupling by ingenious model generalization and can recommend new web pages different from the user concerned before.

Content based recommendation system mainly has the following disadvantages [3]: First, attribute extraction problem. A piece of information must be expressed as a certain number of attributes or keywords. We can automatically complete this job for text type information (news, web pages) by keywords extraction techniques. But for other types of information, such as movie, music, product, it is impossible to automatically get an ideal attribute representation and manual attribute definition. This requires a lot of domain knowledge for system designers and system maintenance worker, which bring extra knowledge acquisition burden. The second problem comes with the recommendation scope. As only items similar to the past preference of the user will be recommended, items not similar to user's preference will not be recommended even if the user may be interested in these items. To solve this problem, some recommendation systems introduce randomness. The last problem comes with new users. The recommendation system can calculate out new user's preference only after he has provided enough ratings. Thus new user will not get ideal recommendation results in the initial period, thereby affecting their willingness to use the system.

2.1.2 Collaborative Filtering

By measuring the user's access history (include download, purchase, click, rating), collaborative filtering recommend the target user with new items preferred by other users similar to the target user. This research began in the beginning of last century. Resnick et al. [32] built a large-scale collaborative filtering system in Grouplens to recommend high quality news to users. Compared to content-based recommendation, collaborative filtering doesn't require semantic analysis on the recommended content, and overcoming the limitation of recommendation scope in content-based recommendation [3]. Thus, collaborative filtering is widely used in many large-scale commercial systems, including Google News [4], Netflix.com and Amazon.com [5]. Some researchers [18, 19] use the term "trust" clearly in the description of their algorithm, whose main idea is the use of trust transitivity to solve the sparse data problem. As these algorithms only consider the correlation of users' historical ratings, the term 'similarity' may be a more appropriate description.

Algorithms for collaborative recommendation can be grouped into four general classes: neighborhood-base method, latent factor model, graph-based model and socialization recommendation. Because this part of content is highly correlated to this book, we will introduce it in detail in the next chapter.

2.1.3 Hybrid Methods

Several recommendation systems use a hybrid approach by combining collaborative and content-based methods, which helps to avoid certain limitations of content-based and collaborative systems [25, 33–38]. Different ways to combine collaborative and content-based methods into a hybrid recommender system can be classified into four types.

1. Implementing collaborative and content-based methods separately and combining their predictions.

In this approach, first we can combine the outputs or ratings obtained from individual recommendation systems into one final recommendation using either a linear combination of ratings [34] or a voting mechanism [35]. Alternatively, we can choose one of the individual recommenders which is perform "better" than others based on some recommendation "quality" [39, 40].

2. Adding content-based characteristics to collaborative models.

Several hybrid recommendation systems are based on collaborative methods but also maintain the content-based profiles for each user. These content-based profiles are used to calculate the similarity between two users. This kind of technique can overcome some sparsity problems of a purely collaborative method

since not many pairs of users had a significant number of commonly rated items [35]. Reference [41] employs an approach in using the variety of different content-analysis agents that acts as additional participant in a collaborative filtering community. As a result, the users whose ratings agree with some of the agents' ratings would be able to receive better recommendations.

3. Incorporating some collaborative characteristics into a content-based approach.

The most common and popular approach in this category is to use some dimensionality reduction techniques on content-based profiles. Reference [42] uses latent semantic indexing to create a collaborative view of a collection of user profiles, this results in a performance improvement compared to the pure content-based method.

4. Constructing a general unifying model that incorporates both content-based and collaborative characteristics.

This kind of methods is very popular in recent years. Reference [33] proposed using content-based collaborative characteristics in a single rule-based classifier. References [43] and [44] proposed a unified probabilistic method for combining collaborative and content-based recommendations, which is based on probabilistic latent semantic analysis. Reference [45] uses Bayesian mixed-effects regression models to employ Markov chain Monte Carlo methods for parameter estimation and prediction.

2.2 Sentiment Analysis

Textual data contains two main kinds of information: facts and the emotions. Research area such as search engine and text mining focuses on exploring facts from the text, while sentiment analysis studies on the computation methods of opinions, moods and emotions from the text [46].

The rapid development of web applications makes it easy to obtain amounts of text information which contains the users' opinions. All these information are distributed in the application systems such as BBS, social networks, blogs and review sites and so on. Collecting this information by crawling and doing sentiment analysis, we can get the objects' (e.g., products) web word-of-mouth. The manufacturers can know about the users' basic impression on their products, the products' features which are satisfactory or dissatisfactory, then they can improve the products or deal with crisis management based on these knowledge.

Due to the difference of granularity, sentiment analysis is classified as document level, sentence level and feature level [47]. Document level classified one comment as $ve+$ and $ve-$. The user's comments can be translated to a corresponding value by document level analysis for further analysis. Sentence level classifies any single sentence in the document as objective fact and subjective attitude and maps every subjective attitude to $ve+$ or $ve-$. Although document

level or sentence level analysis has its own emphasis, we don't have enough commentary to find out what the user like or dislike from one comment. Feature level can help us obtain the main opinion which the user has for the main characteristic of reviewed object.

The main technique used in sentiment analysis is natural language understanding [1]. How to map human's abundant emotional expression to numerical value accurately is still a challenging task. We have to point out that sentiment analysis will also be affected by spam [48]. As described in [Chap. 1](#), identifying spam simply depending on the text features is hard. At present, there isn't any perfect solution for this problem [46]. Finally, sentiment analysis focuses on obtaining structured expression of user's attitude from un-structured comments. Certain tasks, such as recommending some contents which the user may be most interested in based on his historical comments or ratings, or generating personalized group opinion analysis results based on the subjective preferences of the user, are beyond the main scope of sentiment analysis. They all belong to the topics of recommendation engines mentioned above. Because this part of content is highly correlated to this book, we will introduce it in detail in the [Chap. 4](#).

2.3 Dynamic Network Mechanism

Users' interaction in the web systems forms a network which contains different types of relationships. These networks show significantly different statistical properties and evolution rules comparing with a randomly generated network. Grasp and application of these rules will directly affects the design and evaluation of collective view prediction algorithm which takes trust network as input.

Strictly speaking, the network N is a connected, acyclic, directed graph which meets the following requirements: (1) there is a subset of vertices X and the in-degree of each node is 0; (2) there is a subset of vertices Y which is disjoint with X , and the out-degree of each node in Y is 0; (3) each edge has a non-negative weights, called edge capacity. The graph G is a two-tuple (V, E) , where V is the set of nodes, and E is a subset of $V \times V$ called edge set. If the element in E is ordered, then G is a directed graph. If the element in E is unordered, namely $(i, j) \in E \Leftrightarrow (j, i) \in E$, then G is an undirected graph. As people often use the term 'network' as the synonymous of 'graph' in research area, in this book we take 'network' and 'graph' as the same term.

2.3.1 Statistic Characteristics

There have been extensive studies on the statistical properties of various types of network, including social network, cited network, food chain network, Internet and so on. The results show that, although there are many differences in scale and filed,

most networks follow some common rules. The most important three properties are “long tail distribution”, “small diameter” and “clustering effect”.

2.3.1.1 Long Tail Distribution

Long tail distribution means the number of nodes whose degree (out-degree and in-degree in directed graph) is d , denoted as N_d , follows the power law distribution $N_d \propto d^{-\gamma}$, and γ is the power rate index [49]. This kind of distribution has been observed in phone call network [50], web pages link network [51–55], the Internet [56], cited network [57], web social networks [58], and other research filed. The study also found that the parameter γ is generally between 2 and 3. For example, the out-degree of web network follows the distribution where $\gamma_{in} \approx 2.1$, the in-degree follows the distribution where $\gamma_{out} \approx 2.4$ [59], while the autonomous system follows the distribution which $\gamma \approx 2.4$ [56]. However, some research results [60] also show that not all networks follow power law distribution, the reason can be explained by “DGX” distribution [61].

Scale-free network is the concept mentioned frequently with long tail distribution in many literatures. The definition of scale-free is: given random variable x and probability distribution $p(x)$, there is a function $g(b)$ that makes $p(bx) = g(b)p(x)$ for all x and b . In a scale-free network, the shape of distribution remains the same when scaling the range of the observed random variables. The research in [62] shows that only the network which follows power law is scale-free, while some other common distributions, for example, normal distribution doesn’t have this nature.

2.3.1.2 Small Diameter

According to the definition of graph theory, if there is a path whose length is up to d between each pair of node (u, v) in graph G , the diameter of G is d . Due to the presence of outliers in the actual network, effective diameter [63] is used to denote the diameter of the network approximately. If the length between 90 % node pair (u, v) in graph G is up to d' , the effective diameter of G is d' .

The study found that the Internet, Web page link network, as well as many real or Web social network all meet the characteristics of small diameter [52, 59, 64–68]. For example, the effective diameter of the MSN network is 6.6 [49], which follow the previous sociological “Six Degrees of Separation” theory.

2.3.1.3 Clustering Effect

In many real networks, the relationship between the nodes have transmissibility characteristic, i.e., if node u links to node v and node v links to node z , then u will link to z with a higher probability comparing with the randomly generated network

which has the same distribution (node degree). This effect of the network is measured by clustering coefficient.

The clustering coefficient of a node is defined as follows: given node v in graph G whose degree is d , the clustering coefficient C_v of v is the ratio of the actual number of links between v 's neighbors, denoted as d , to the number of links that may exist $d(d-1)/2$ [49]. Similarly, we can define C_d as the average clustering coefficient of all the nodes in graph G whose degree is d ; the clustering coefficient of graph G is the mean of all nodes' clustering coefficient.

Some studies found that the clustering coefficient of actual network is higher than that in random networks with identical distribution. In addition, the clustering coefficient C_d decreases with the increases of d , and follows the power law $C_d \propto d^{-1}$ [69, 70]. Clustering effect can be used to discover the organizational structure [70–73]. Usually, the nodes with low degree in the network belong to different dense sub-graphs; the sub-graphs are connected by hubs.

2.3.1.4 Other Characteristics

Some other characteristics of real network have also been found in recent years. Some studies show that complex network is composed by the frequent appearance of basic network motifs [74, 75], while the random network with identical distribution does not have this characteristic. Recent studies also show that there is significant impact of network motifs to the transmission pattern of information [49]. The social network is also “self-healing” [59, 76], i.e., if we remove nodes randomly from the network, the connectivity of the network will not be significantly affected. But merely to remove several nodes with high degree, the connectivity of the network would be severely reduced. Self-healing is consistent with clustering effect, for sub-graphs are often connected with each other by a small number of nodes whose degree is high. Therefore, network will be divided into a number of independent sub-graphs after removing these nodes.

2.3.2 Evolution Law

Social network not only has the above static statistical properties, its evolution also follows some common patterns, including “Densification power law” and “Shrinking diameter”.

With the increase of the number of nodes, the social network becomes more and more dense, i.e., the average degree of the node will increase with the increasing of nodes number. Generally, the densification of network follow the power law: $e(t) \propto n(t)^a$, where $e(t)$ and $n(t)$ represent the number of edges and nodes at time t respectively, and the range of parameter a is between 1 and 2. Research work in [77] shows that the Internet and paper citation network follow the densification

power law with $a = 1.2$ and $a = 1.6$ respectively. A study of social network [78] also showed that the number of nodes (the number of users) increased rapidly in growing period, and the increasing speed was gradually slowed down in mature period.

In traditional theory, the diameter of the network is growing slowly with the increase of nodes number. However, the results in a study on the Internet and paper citation network [78] show that the effective diameter of the network decreases slowly with the increase of nodes number. Intuitive explanation for this phenomenon is that the degree of nodes increases after joint into the network, which makes the probability of finding shorter path between node pair increase accordingly.

2.3.2.1 Evolution Model

A number of explanations have been proposed on network evolution law. The Ordos model is an earlier proposed network evolution model: given n nodes, each node pair is connected by a new edge with the same probability. However, the graph generated by this model does not satisfy the long-tail distribution which is the main characteristic of real network. People also proposed the Prior Subsidiary model [53, 79] and Evolving copying model [51, 80]. In both models, each time adding a new node u to graph G , m new edges will be created. In Prior Subsidiary model, each newly created edge is connected to node v whose degree is $d(v)$ with the probability $p_u(v) \propto d(v)$. In evolving copying model, the new node is connected to m randomly selected nodes with probability β and is connected to the m neighbors of m randomly selected nodes with probability $1 - \beta$. Graph generated by Prior Subsidiary model, an evolving copying model all meet the long-tail distribution, but diameter contraction characteristic can't be established. In small-world model, a regular grid is generated first, and then the tail node of each edge is changed to another randomly selected node with probability p . A variety of networks with different characteristics can be obtained by changing the parameter p from the regular network ($p = 0$) to the random network ($p = 1$). The network generated by lower p has more partial structures constituted by short links, at the same time the diameter is longer. Higher value of p will destroy the local structure of the network, and generate shorter diameter.

Forest fire model [49] combines the advantages of the Prior Subsidiary model and Evolving copying model. It uses two parameters, the forward combustion probability p and the rearward combustion probability r , to simulate a new node "burning" the edges already existed to join the network. The experimental analysis result shows that the network generated by forest fire model, satisfies the long tail distribution and diameter contraction at the same time. The nodes with high degree act as a bridge to shorten the distance between nodes.

The above models focus on generating network which satisfies specific statistical properties, in order to explain the reasons of existing actual network. Given an actual network, how to select the model parameters in order to fit the actual

statistical properties of the network? Exponential random graph [81] provides a method to select optimal model from the candidates set. But this method focuses on the measurement of the local structure in graph (e.g., which statistical properties of the node determines the generation of an edge?), and this approach is not scalable in large-scale network. The parameter estimation methods for Evolving copying model have also been proposed in [82]. But experiments show that Evolving copying model is not sufficient to simulate many actual networks. Kronecker graph model [49] use matrix Kronecker product to obtain model parameters. The complexity of the algorithm is the linear function of edge number in the network. It can effectively simulate large-scale real network, such as Epinions.com trust network.

2.3.2.2 Link Prediction

Link prediction is another important topic of network evolution law. Link prediction tries to use the intrinsic characteristics of network to model the evolution law. More specifically, it refers to “given the network snapshot at time t , predict the edges that will be introduces into the network since time t to a given future time t' ” [83]. A common application of link prediction is to recommend friends in social network website. Another useful application is to predict the possibility of two users become coauthor. For example, when a new entry is created in the wiki site, system can recommend this record to the user who is suitable to edit it based on users’ past edit history.

Formally, the input of link prediction is a relational network G , give a weight to each node pair $\langle S, V \rangle$ which does not exit public edge between them, and then sort by weight in ascending order, the node pairs with higher weight is the link may occur in the future. “Nearest Neighbor” and “Public Nearest Neighbor” are two typical weight calculation methods. A simple strategy of Nearest Neighbor treats the shortest path between $\langle S, V \rangle$ as the weight. It is the manifestation of network clustering effect. If node u links to node v and v links to node z , u may be connected to z with higher probability than the randomly generated network (with the same distribution). This approach does not take the number of accessible paths into account. In common sense, the more accessible paths between two nodes, the closer relationship between them. Therefore, a more reasonable method Katz [83], is to use the weighted summation of the accessible paths between all the node pair as weight. Public Nearest Neighbor method calculates the common neighbors of all nodes. The more common neighbors of node u and node v , the higher probability they are linked in the future.

A recent study [49] showed that, in MSN social network, we can do link prediction accurately by combining Public Nearest Neighbor rule and user latest activity time.

2.3.3 Web Information Cascades

An information cascade refers to an activity or a view point is widely adopted due to the influence of others [84]. The research in information cascade began in the field of sociology, economics and epidemiology, such as the study of Diffusion of Innovations [85] in sociology, “trend leader” in viral marketing [86], and vaccination people in epidemiological [87]. In recent years, with the development of social network, this topic also began to be concerned in the field of computer science, for example, the study of the rule of cited article in the blog [88].

The main reason that the study of web information cascade receives consideration is that, if the impact pattern between nodes in network can be decided, many decisions can be expressed as optimization problem, thereby generating direct economic benefits. Take viral marketing for example, given the marketing budget and the influence function of users in network and using a simple hill-climbing search algorithm, we can get the optimization approximate solution with a degree not less than 63 %. We can use the similar method to optimize the problem of how to place the monitors of city water pollution monitoring point.

It is different of the study in information cascade between computer science and sociology or economics. A significant difference is that the latter focuses on small-scale groups, the degree of familiarity between groups are often higher than users in the website. Some research results about blog [89–91] show that the mode of transmission that found in sociology and economics is not significant in blog. A recent study [49] found that the local structure of network has a direct impact on the dissemination of information. In addition, there are also significant differences between these local structures in different application contexts.

2.4 Trust Management

The concept of trust has been used in many research fields in computer science, including distributed authentication and authorization, trusted payment platform, and reputation system. Jøsang et al. [23] call them as trust management, in other words “The activity of creating systems and methods that allow relying parties to make assessments and decisions regarding the dependability of potential transactions involving risk, and that also allow players and system owners to increase and correctly represent the reliability of them and their systems.” According to this definition, trust-based collective view prediction can be attributed to the trust management research areas, despite there are significant differences between our concerns and the three sub-areas discussed above.

Distributed authentication and authorization focus on finding a unified method to do cross-system authentication (single sign-on) and authorization in order to break the technical barriers between heterogeneous systems. The early research in this area dates back to the protocol-based system PGP public key certificate [92]

and X.509 [93]. Blaze et al. [22] and others extended authentication method based on public key cryptography system to support expression and verification of security policy. In addition, the rise of automated trust negotiation [94] as well as joint authentication and access control [87, 95–97] in recent years can also be attributed to this sub-field. In these research areas, confirmation the authenticity of the identity is a major concern. Grandison [98] called it “identification trust”. The identification trust is a certain relationship, the result of it is either positive or negative, and there is no intermediate state. In our study trust is an uncertain relationship, it shows the strength of trust, and it is used to generate the opinion prediction of a group.

Trusted payment platform is designed to provide users with a reliable online trading environment, to reduce the negative impact of anonymity and cross-region of Internet. Alipay in China and PayPal are the typical applications of trusted payment platform. The core idea is to increase the accountability for Internet transactions. Trusted payment platform ensure the reliability of the transactions through a series of measures. Take Alipay for example, users need to authenticate through real-name system before transaction. The real-name authentication methods include registering Alipay one-card and identifying through bank accounts. User’s real identity information can be obtained through these two ways. In transaction process, users first took goods, and then pay money into his Alipay account. Sellers shipped after receipt of the notification message. Until the buyer received the goods and sent the satisfactory information to Alipay, the sellers will get the payment in their own Alipay account. When trade disputes happen, the buyers and sellers submit their evidence to Alipay and the specialized arbitration agency will do mediation. In addition, the application website of Alipay, such as Taobao, can also provide transaction commitment such as “Quality problems of seven days” to enhance the credibility of the shop.

Reputation system calculates the collecting user feedback, so that the participants in the system are able to determine whether the behavior of others follow the business rules, thereby avoiding potential risks. In addition, the reputation system motive users to participate in the collaboration by setting reasonable rules. Reputation system has been used successfully in e-commerce sites such as Taobao and eBay. Some P2P software such as eMule and Tribler [99] also use reputation systems to prevent the occurrence of the “free-rider” [100]. Reputation system has some similarities with our study, but there are significantly differences between them in research methods and purposes. Generally, the former usually takes user’s behavior patterns as a starting point, trying to find a reasonable reputation function. Assume the user as a rational subject, the output of user behavior reputation function will have a positive impact, thereby reducing the possibility of bad behavior. Golbeck [17] pointed out that the standard of whether a user’s behavior is good or bad in reputation system is relatively uniform. But in the group opinion prediction problem, there is no absolute standard of good and bad for a user’s behavior. For the same item (such as movie), it is common for different users give opposing comments. Reference [101] proposed a transverse evaluation result on

the effectiveness of the reputation system. Reference [102] classified and summarized the reputation calculation method.

2.5 Summary

In this chapter, we reviewed three research areas related to our study. The conclusion and approach in those research areas formed the foundation of our work. And we will improve the proposed models and algorithms to for usage.

Emotional analysis can convert the text comments into ratings which can reflect user opinions, so that we can focus on collective view prediction in the latter case. In the subsequent discussion of this article, we will treat comments and ratings as synonyms. In other words, we assume that any unstructured document-level review can be mapped to a real number by the sentiment analysis, or mapped to multiple real numbers through the analysis of characteristics of a class.

The output of Collaborative filtering and trust-based collective view prediction is the same. The commonly used evaluation metrics in collaborative filtering, such as prediction accuracy and coverage, can be used directly in our research. In addition, we also explore in what condition can similarity is defined as trust, and we also design a second-order Markov chain-based collaborative filtering algorithm. The algorithm is further used to combine with trust-based prediction algorithms to improve the accuracy of the prediction.

The power law distribution of the social network, small diameter and other characteristics also played a key role in our study. Small diameter feature ensures that we can find enough near neighbors to do prediction within a few steps through breadth-first search. The power law distribution characteristics determines that a little items in the Web system usually have a lot of ratings, and these popular items is the weak point to break robustness of the collaborative filtering algorithm [6]. In the trust network, a few nodes also have high-degree which impact a lot on our prediction algorithm. So when evaluating the robustness of algorithm, we must consider the impact of attacks against these high-degree nodes especially.

In the following section, we use Beta probability distribution which usually used in reputation systems research to represent the trust [103]. We also introduce two operators in belief theory [104] to do trust measurement. In addition, the results of research in the field of distributed authentication and authorization for cross-system provide a standardized solution for trust-data sharing in collective view prediction.

Trust-based Collective View Prediction

Luo, T.; Chen, S.; Xu, G.; Zhou, J.

2013, XI, 146 p. 41 illus., Hardcover

ISBN: 978-1-4614-7201-8