

KEY USERS AND BOX OFFICE ANALYSIS IN AN INTEREST BASED VIRTUAL COMMUNITY

I-HsienTING, Shyue-Liang WANG, Po-Jen CHEN

*Department of Information Management
National University of Kaohsiung
No. 700, Kaohsiung University Road
811 Kaohsiung City, Taiwan
e-mail: {iting, slwang}@nuk.edu.tw, seraph0331@hotmail.com*

Abstract. In recent years, with the growth of the Internet technology, the users of virtual community not only play the role of the information receiver but also a very important one to provide information. However, there is large amount of information aggregated daily and therefore information overloading has become a very serious problem. Under this situation, how to find information efficiently is also a very important issue. In this paper, we believe users in a virtual community may affect each other, especially those with high influence who have been called as *Key Users*. Therefore, we observe the biggest virtual community of movies on the Internet which is named IMDb (The Internet Movie Database). An architecture also has been proposed that combines Social Networks Analysis and the features of IMDb to discover those users who have high influence in the virtual community. We collected 17 months (January 2010 to May 2011) from IMDb including 17 366 users and 243 074 reviews. By applying the method we proposed, there are about 22 key users and 111 reviews were discovered. We also use the box office of the movies to justify our results.

Keywords: Social networks analysis, IMDb, key users, box office, virtual community

1 INTRODUCTION

Virtual Community now has become a very popular application in the Internet, due to the explosive development of the Internet and communication technology [2]. Fur-

thermore, the development of social networking website and Web 2.0 also increase the influence over the Internet users. Nowadays, users in many popular web applications, such as Wikipedia, Yahoo! Knowledge, web forums and blogs play the role not only as information receivers but also information providers. Some popular social networking websites, such as Facebook or Twitter, cannot even survive without the input of users. Therefore, many companies have paid attention to the opinions and ideas given by the users. From the companies' point of view, it is essential to study the users' behaviour and how to apply the research findings.

Many tools and techniques have been applied to understand the users on the Internet, such as web agent, data mining and web mining, machine learning and artificial intelligence, etc. By applying these tools and techniques, it is very easy to collect information and behaviour of the Internet users. However, it is not sufficient if we cannot understand the meaning behind the collected information and users' behaviours. Unfortunately, the data currently aggregated on the Internet is too big to be analysed and therefore important users behaviours are hard to be discovered. Thus, it is a problem if we cannot explore the big data and extract the useful information.

On the Internet, many virtual communities are formed based on different interests, such as cuisine, game, movie, music, etc. These platforms also allow users to search product information or comments. These comments have been proved as very important reference and key point for users to make decisions [6, 10]. Thus, many companies are now trying to discover the so-called key users who are influential over these virtual communities. They believe that the confirmations of these key users could be beneficial to their marketing strategy. It would therefore be very necessary to study the relationship between the key users and product sales.

Three major goals are aimed to achieve in this paper according to this background. First, we will propose an approach to discover key users. In order to understand the behaviour and relationship of users, Social Networks Analysis (SNA) has been claimed as the most suitable methodology. It has been applied very widely in different areas, such as on-line recommendation system and information diffusion. In these researches, SNA was also applied to discover the roles and positions of users [3, 21, 11]. Second, we will propose an approach to measure the influence of key users according to the box office of movies. In this paper, IMDb (The Internet Movie Database, <http://www.imdb.com>) will be selected as the interest based virtual community, which is the biggest movie database in the world. A very important reason for us to select IMDb as the target is that it provides a rating value to measure the usefulness of a review. The rating value is generated according to users' voting, which is a very useful measurement for SNA. Third, in the paper we will also discuss how to take the discovered key users into practice.

In this paper, we set two hypotheses according to the goals discussed above. Hypothesis one is that the searching time of users through the system can be reduced based on only key users reviews. In this research, we would like to prove the discovered key users are useful for IMDb users to reduce the searching time of

reviews. Hypothesis two is that the useful rating is the key for measuring the users trust of reviews. These two hypotheses will be tested in the research.

The structure of the paper is organized as follows. In the first section, we will introduce the motivation and background of the paper. Related literature and works will be reviewed in Section 2, including literature on virtual community, interest based virtual community, social networks analysis and related researches and applications of social networks analysis. In Section 3, the research model and method of the paper will be proposed and the experiment design of the paper described. The experiment design and analysis of experiment result will be included in Section 4. The evaluation and discussion of key users will be included in Section 5. The paper is concluded in Section 6.

2 LITERATURE REVIEW

2.1 Virtual Community

There are different definitions of the term “virtual community”. According to Rheingold, virtual community is defined as a relationship group that contains enough people and enough discussions [19]. Baym proposed that virtual community is composed by four structures:

1. A virtual community creates different communication forms to express opinions.
2. The identification of members in a virtual community is unique.
3. Members in a virtual community can establish relationship through various approaches.
4. There are regulations and rules in a virtual community.

[5, 12] Adler and Christopher defined virtual community as a group of Internet users with the same interest in a public cyberspace, such as WWW, to exchange opinions, to communicate and to exchange information [1].

The definition of virtual community by Armstrong and Hagel can also be applied for categorizing virtual community. They divide virtual communities into four different categories according to different behaviour forming factors. The four categories are *Transaction-Oriented*, *Interest-Oriented*, *Relationship-Oriented* and *Fantasy-Oriented* [4]. Most of popular virtual communities nowadays such as Yahoo! and Facebook are actually combining these four categories rather than focusing on a particular one.

2.2 Interest Based Virtual Community

With the definition of virtual community discussed previously, interest based virtual community will be selected as the main focus in this paper. Currently, there are many interest based virtual communities on the Internet. In this paper, our focus

is the virtual community based on the interest of “*Movie*”. In general, the users behaviour are quite different from other interested based virtual communities. In movie based virtual community, users always review or comment the items which are interesting to them. The interaction between users is not a common behaviour in the movie based virtual community.

Recently, there are many researches focused on movie based virtual community. Ahmed et al. have tried to visualize the social relationship in IMDb, one of the famous online movie database. In their research, they also try to discover the relationship between events and movies, such as the event of September 11 attacks [3]. Some researchers also applied the content-based data in IMDb to strengthen the weakness of current recommendation systems [9]. Debnath in 2008 analysed the data to discuss the important factors such as director, cast, cost, category and user in IMDb that may affect users’ viewpoint. The result of the analysis shows that user is the most influential factor in IMDb [6]. In this paper, we therefore will also use IMDb as the analysis target and take the important factors proposed by Debnath into account.

2.3 Social Networks Analysis

The Social Networks Analysis (SNA) has been studied long. In 1925, Lewin started focusing on related researches of SNA. They intended to describe the interaction and relationship between people by using simple geometry graph (line and point). In 2000, Scott proposed the so-called “*Sociogram*”. Figures 1 and 2 are two sample sociograms. Figure 1 is a directed sociogram and Figure 2 is an undirected one [20].

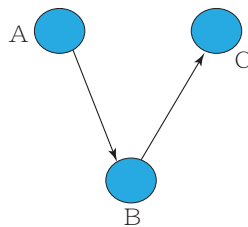


Figure 1. A directed sociogram

In a sociogram, *Node* is the basic component to form a social network which is usually used to present a user or event, etc. For example, in Figures 1 and 2, A, B, and C denote three nodes in the sociogram. *Edge* is a component used to describe the relationship between nodes. For example, in Figure 1, node A thinks B is his/her friend and B thinks C is his/her friend. However, B does not think A is his/her friend and C does not think B is his/her friend. In Figure 2, no relationship direction is designated in the sociogram, it only denotes relationships exist between A and B and B and C.

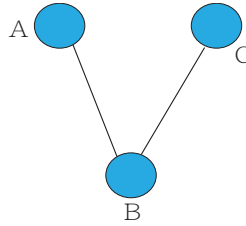


Figure 2. An undirected sociogram

In recent years, the research on SNA has changed the direction with the development of Web 2.0 and virtual community. In web 2.0, users play the role not only as information-receivers but also as information-providers. Furthermore, the amount of data has also increased rapidly due to many interactions between users aggregated on the Internet every day. Therefore, it is more complicated to analyse a social network. Many social networks analysis measurements have been performed to describe complex social network. In this section we will also introduce some important measurements for social networks analysis.

2.3.1 Centrality

In a social network, it is always an important issue to analyse the position of a node to explain its importance in a social network. Therefore, the measurement of *Centrality* has been proposed to measure which nodes are in the central position of a network [7]. Centrality can also be used to locate the most powerful node in a network that is able to affect other nodes. There are different approaches to measure the centrality if the viewpoints are different, such as *Degree Centrality* and *Closeness Centrality*, etc.

$$D - Centrality(i) = \frac{Degree(i)}{\sum_{i=1}^n Degree(i)} \quad (1)$$

The equation for centrality measurement is shown in Equation (1). In Equation (1), i denotes a node in a social network and *degree* is the total number of edges between node i and other nodes. Therefore, the higher centrality value of a node means more nodes to be connected with (including *in-degree* and *out-degree*) the node [20].

2.3.2 Small World

In 1967, the psychologist Milgram proposed famous *Degrees of Separation* and *The Small World* [17]. The small world theory means the degree for each two nodes is not more than six (in the US), whether they know each other or not. Therefore, in a small world, we only need a few connectors to reach the whole network. In 1988, Watts also pointed out that many social networks in the real world basically correspond to the small world theory [21], such as markets, economic units, etc.

Watts also suggested the characteristics and rules of small world theory. First, in a small world, the clustering coefficient is usually high. Clustering coefficient is a degree to measure the closeness of the nodes in a social network. In Equation (2), $E(x)$ denotes how many nodes are linked in a network x .

$$\text{Clustering Coefficient} = \frac{E(x)}{E_{\max}} \tag{2}$$

Second, the average degree of separation in a small world is usually lower, or has a lower *Diameter* in a network. The degree of separation means the shortest path of each pair of nodes in a network and *diameter* means the shortest path in a network. Watts also pointed out that the relationship social network is a graph in between of random graph and regular graph. Table 1 shows the different categories of graphs. Small world graph is a graph that contains the characteristics of high clustering coefficient and low degree of separation.

	Low Clustering Coefficient	High Clustering Coefficient
Low degree of separation	Random graphs	Small world graphs
High degree of separation	N/A	Regular graphs

Table 1. The categories of graphs (clustering coefficient-degree separation)

2.3.3 Density

Density is also a very common measurement used in most of social networks researches. The definition of density is to measure the distance between the graph and a complete graph [20]. Figure 3 shows examples of measuring the density of a graph.

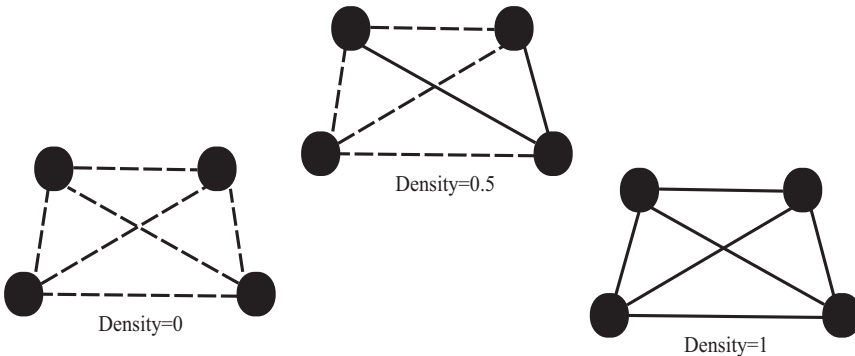


Figure 3. Examples of density measurement

The measurement of density is usually used for studying the clustering situation of a network. Lewis et al. combined density and regression model to discuss the *tie* relationship in Facebook and concludes that the density of friends in Facebook is 22.4% [15]. Kumar et al. also observed the social networks structure (nodes and density) with changing of time in the famous social networking website Yahoo! and Flickr. In their researches, they found that users would do their best to invite friends to join the same social network right after their joining a group or social network. Therefore, the density and degree of the social network will continually increase. However, the speed will slow down after a period of time [13].

2.4 Social Networks Analysis Applications

With the growth of virtual community and social networking websites, more and more interactions among users are generated. Many researchers start to apply social networks analysis based on the data from virtual community [14]. Qin et al. in 2012 proposed a research by using the reviews and comments in YouTube to analyse the behaviour of website users. The analytic results will then be exercised to design a recommendation system [18].

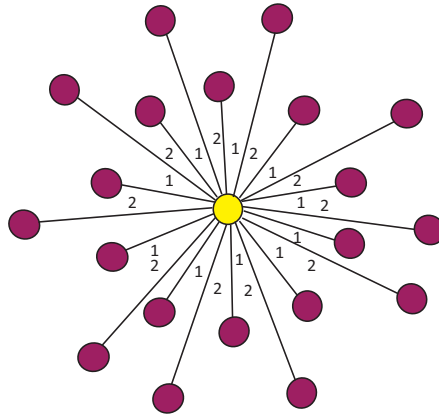


Figure 4. A sociogram of users-movie (figure source: [18])

In Qin's research, the central node of Figure 4 is a movie from YouTube and other nodes around it are users who have co-reviews. In this research, the finding suggests that users did not interact with each other directly but interacted (review, rating) through an indirect intermediate (movie). In this paper, we will use the same concept to analyse the users interaction in IMDb.

There are three basic factors in social networks: actors, relationships and link-ages [8, 16]. However, with the growth of social network, it has more and more complex relationships in the network. Therefore, another issue is to filter and down-size the data. In this research, we intend to discover the key users in a virtual

community which is a big task due to its bulky database. Thus, an efficient approach to discover the key users that can represent the social network is essential for this research.

Zhang et al. proposed a research about discovering key users in an interest based virtual community. In their research, an on-line movie review website *Epinion.com* has been selected as the focus. They picked a set of key users to test their influence on their friends. Linkage based model has been evaluated as the most successful model to discover key users under different environment [22].

3 RESEARCH MODEL AND METHOD

According to the motivation, introduction and literature review discussed previously in this paper, an approach is going to be applied to find the key users in a movie based virtual community and to analyse the box office of movies as well. In Figure 5 the research and experiment process of the proposed approach is shown.

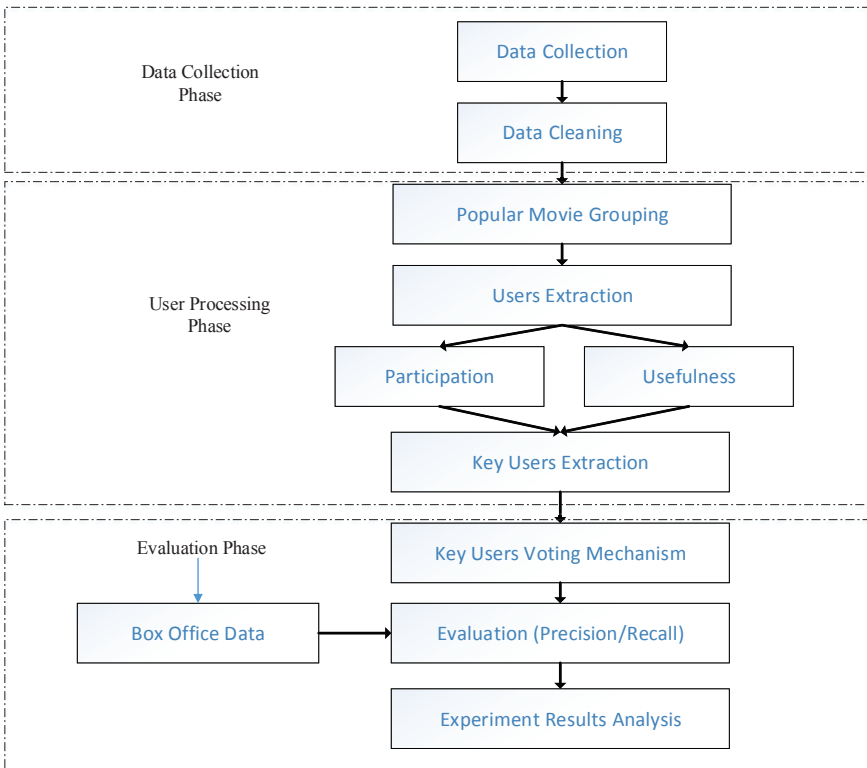


Figure 5. The research and experiment process of the paper

In Figure 5, three main phases are presented, including data collection phase, users processing phase and evaluation phase. In phase one, data collection and cleaning are two steps which will be introduced in detail later. In phase two, the main idea is to discover and extract key users from the movie database. We firstly will try to discover the most popular movies in order to downsize the data. From these popular movies, users who have commented on or reviewed to these movies will be extracted. The next step is to extract key users according to two measurements, participation and usefulness of reviews. In phase three, the main idea is to evaluate the accuracy of the key users we discovered. A voting mechanism will be used to evaluate the key users and also to evaluate the relationship between key users review and box office. In this paper we suppose that if a movie has high score from key users review and comment then the movie's box office is good. Precision and recall will be used to evaluate the quality of the discovered key users.

The company could save time when predicting the potential and box office of a movie with the help of digging out the key users of IMDb using the research model. That is also the reason why we need to evaluate the relationship between key users and box office. Consequently, company or cinema can just try to extract and analyse the key users comment for decision making, such as marketing strategy, advertisement, cinema arrangement, etc. Entire database's investigation would no longer be necessary.

3.1 Data Collection

The data source of this paper is from IMDb which is the biggest movie information and review database. We collected 17 months data from January 2010 to May 2011. In IMDb, movies are categorised into different genres, such as drama, action, comedy, love story, sports, etc. In IMDb, users are also allowed to post comment or review to a movie, as well as scoring and voting. Figure 6 shows a sample movie and its information in IMDb database. In this figure, basic information of the movie category, director, writers, cast, movie release date etc. are displayed. Ratings and reviews from users and the reviews' usefulness to others are also exhibited.

Figure 7 is the movie information raw data from IMDb for the movie in Figure 6. A program could be easily designed to extract the useful information such as its category, director, cast and writers from the data.

Figure 8 is the review information raw data from IMDb for the movie in Figure 6. Useful information such as the usefulness of a review, reviewing time, reviewer himself and content of the review can be drawn out from the plain text files likewise.

In this paper, we design a Perl based program to process and clean the collected data. The pre-processed data are shown in Tables 2 and 3. Table 2 shows the pre-processed movie information data, including the title, the release date and the rating. Table 3 shows the pre-processed user information data, including user name, movie, date, rating and useful. These data will then be used in next section for users process to identify and discover key users.



Figure 6. A sample movie and its information in IMDb (source: IMDb website)

Title	The Girl with Dragon Tattoo
Date	2011/12/20
Rating	8.3 (17 250)

Table 2. The pre-processed movie information data

3.2 Users Processing

In the users' processing step, we will first try to define the most popular movies in IMDb. We assume that key users' comments and reviews will drive other users to watch the movie and therefore the movie with good reviews will usually become popular movie with successful box office. Thus, we would like to dig out popular

User Name	Movie (year)	Date	Rating	Useful
User 2509775	Alice in Wonderland (2010)	2010/8/18/	7/10	1/2
User 2509775	Million Dollar Baby (2004)	2010/10/2	10/10	1/1

Table 3. The pre-processed user information data

```

714 </a>
715 <span class="ghost"></span>
716 <a href="externalreviews?ref_wtt_ov_st" title="457 IMDb critic reviews" > <span itemprop="reviewCount">457 critic</span>
717 </a>
718 <span class="ghost"></span>
719 <a href="criticreviews?ref_wtt_ov_st" title="41 review excerpts provided by Metacritic.com" > 41
720 </a>
721 <a href="http://www.metacritic.com" target="_blank"> Metacritic.com
722 </a>
723
724 </div>
725 <div class="clear"></div>
726 </div>
727 </div>
728 <div class="description">
729 Journalist Mikael Blomkvist is aided in his search for a woman who has been missing for forty years by Lisbeth Salander, a young o
730
731 <div class="text-block" itemprop="director" itemscope itemType="http://schema.org/Person">
732 <h4 class="inline">Director:</h4>
733 <a href="/name/nm0000999/?ref_wtt_ov_dr" itemprop="url"><span class="itemprop" itemprop="name">David Fincher</span></a>
734 </div>
735 <div class="text-block" itemprop="creator" itemscope itemType="http://schema.org/Person">
736 <h4 class="inline">Writers:</h4>
737 <a href="/name/nm0001873/?ref_wtt_ov_wr" itemprop="url"><span class="itemprop" itemprop="name">Steven Zaillian</span></a>
738 <a href="/name/nm2297183/?ref_wtt_ov_wr" itemprop="url"><span class="itemprop" itemprop="name">Stieg Larsson</span></a>
739 </div>
740
741 <div class="text-block" itemprop="actors" itemscope itemType="http://schema.org/Person">
742 <h4 class="inline">Stars:</h4>
743 <a href="/name/nm0088819/?ref_wtt_ov_st" itemprop="url"><span class="itemprop" itemprop="name">Daniel Craig</span></a>
744 <a href="/name/nm1913734/?ref_wtt_ov_st" itemprop="url"><span class="itemprop" itemprop="name">Rooney Mara</span></a>
745 <a href="/name/nm0001626/?ref_wtt_ov_st" itemprop="url"><span class="itemprop" itemprop="name">Christopher Plummer</span></a>
746 <span class="see-more inline nobr">
747 <a href="fullcredits?ref_wtt_ov_st" itemprop="url"> See full cast and crew</a> &raq;
748 </span>
749 </div>
750 </div>

```

Figure 7. A sample movie information raw data (source: IMDb website)

```

3778 <span itemprop="review" itemscope itemType="http://schema.org/Review">
3779 <strong itemprop="name">Despite claims to the contrary, a necessary re-interpretation of the story</strong>
3780 <span itemprop="reviewRating" itemscope itemType="http://schema.org/Rating">
3781 <meta itemprop="worstRating" content="1" />
3782 <meta itemprop="ratingValue" content="8" />
3783 <meta itemprop="bestRating" content="10" />
3784 </span>
3785 <div class="comment-meta">
3786 27 December 2011 | by <a href="/user/ur23911039/?ref_wtt_urw" ><span itemprop="author">Jlars77</span></a>
3787 <meta itemprop="datePublished" content="2011-12-27" />
3788 (United States)
3789 &ndash; <a href="/user/ur23911039/comments?ref_wtt_urw" >See all my reviews</a>
3790 </div>
3791 <div>
3792 <p>
3793 <span itemprop="reviewBody">After the announcement that The Girl with the Dragon Tattoo was getting an English-language film treatment, I decided that the hype had
3794 built up to a point where I just had to read the source material for myself. Though it is not without flaws, Dragon Tattoo is an excellent story with the important mission of raising
3795 awareness concerning violence against women. Here days after finishing the book I watched the Swedish film. The hype train had me excited for an outstanding thriller. <br /><br />The
3796 hype train let me down. <br /><br />It was left cold and somewhat irritated by the Swedish adaptation. A ton of important plot elements were left out, some were inexplicably added
3797 (Blomkvist's memories of the island became far too important and contrived), and Rapace felt all wrong as Lisbeth. She was brilliant and violent, but lacked the quiet pensiveness
3798 of the original character. She did not come off as autistic and emotionally disturbed, just bratty and rude. Worst of all, I was constantly confused by the extremely rushed, strange new
3799 take on the story. <br /><br />As a lover of foreign films, I normally grind my teeth when I hear that America is developing a remake. However, I found myself desperate for this one. I
3800 needed a movie that actually gave me the experience of reading the book for the first time, that made me care for Lisbeth and that truly disturbed me. Thankfully, the American
3801 adaptation (not a remake) delivered exactly what I was looking for. Those who say this version is unnecessary or a rehash must have seen a very different Swedish Girl with the Dragon
3802 Tattoo than I saw. The American take is jam-packed with scenes that were either skimmed or completely left out of the Swedish version. Yet, despite being more robust, the English-
3803 language Dragon Tattoo is incredibly paced, feeling less rushed yet hitting all the important plot points. The characters have time to develop and grow on you, the clue-finding makes
3804 more sense, and the killer is more horrifying. Screenwriter: Zaillian knows exactly what to leave out and what to change (though the ending, which mirrors the book's ending, could
3805 have been arranged better). Craig, Mara, and even Plummer are spot-on in their roles and feel more fleshed-out as characters. Mara, in particular, inhabits Larsson's Lisbeth in a
3806 way Rapace did not. She captures Lisbeth's silent, borderline-autistic nature perfectly. Her fragile body and alien appearance even match the book's description. She allows
3807 herself to be vulnerable, but clearly respects it over time. It's her a captivating performance. <br /><br />If someone were to ask me, personally, which version to see, I would have to
3808 say without hesitation that this is the rare occasion where the American adaptation is superior. I did not think it was possible to stay so true to the story under three hours.</p>
3809 </div>

```

Figure 8. A sample review information raw data (source: IMDb website)

movies from the database to downsize the data. In this research, we apply the idea of co-review by YouTube to discover the most co-reviewed movies as the popular movies.

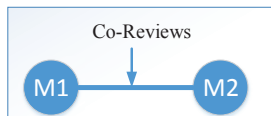


Figure 9. The co-review concept

Figure 9 is the concept of the co-review approach. In the figure, there are two movies: M1 and M2, and co-reviewer of the two movies exists. According to the co-review concept: a movie has higher degree (linkage) and would hold more reviews. The co-review concept can also help to filter out the users who only posted 1 review

to downsize the data. In this paper, top movies with more than 300 degrees are selected according to the criteria based on the analytic result in Figure 12 which will be discussed in the next section. With regards to the key users discovery, two selection criteria are named: “User’s review” and “Usefulness index”. In the first step, users that posted more reviews than the average number of the total reviews will be picked out. Then, the key users can be discovered according to the usefulness index.

4 EXPERIMENT RESULTS

In the experiment, we collected data from IMDb during the period from January, 2010 to May, 2011 (17 months). During this period, hundreds and thousands of movies were released, including independent movies, micro-movies, etc. Movies released only in the US are selected in order to simplify the data. Under this background, the total number of movies is 402 and there are 17366 users with 243074 reviews.

4.1 Popular Movies Identification

In this paper, the first step is to identify popular movies from the database. According to the co-review concept of Figure 9, we analyse the entire movie network by using indirected sociogram in order to understand the co-review structure of the movies. The sociogram is generated by using a very famous SNA software, UCINET (<http://www.analytictech.com/ucinet/>). As shown in Figure 10, the total movies in the network have 339 nodes and 38000 edges. In this figure, we obviously can find most co-review behaviours happen in some particular movies. In the sociogram, the more central-located movies are considered as more popular movies.

In order to discover the most popular movies, we try to zoom in the sociogram and distinguish the sociogram into different levels according to the degree. In this research, the higher degree means more co-reviews. Figure 11 shows the sociogram with different degree levels. In the figure, each different color represents 25% (100 degrees) interval scale. The movies with red colors are the movies with a degree value more than 300. The total number of movies in this level is 52, that is 15.3% of all movies which received reviews (52/339).

Figure 12 is the users’ reviews distribution diagram, in which the x-axis indicated the number of movies that users posted reviews and the y-axis represents the number of users. It shows that most of the reviewers post only few reviews, 903 users only posted one review meanwhile only one user posted 51 reviews. It means the number of key users in the IMDb is small. This analytic result matches the hypothesis that had be proposed in the paper earlier: there only 2606 out of 17366 users have posted reviews for those popular movies. Most of the users only posted reviews for one to three movies. In the following section, we will discuss how to select the key users from the IMDb database.

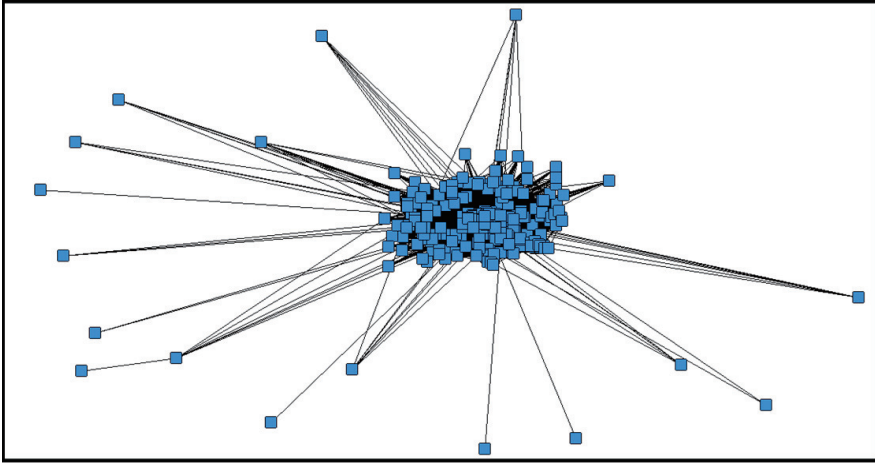


Figure 10. The sociogram of popular movies

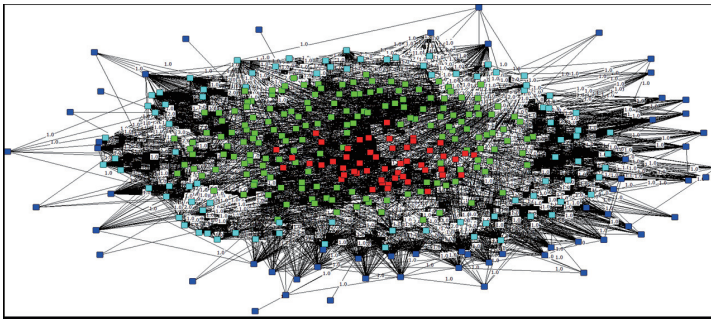


Figure 11. The sociogram distinguished by different degree levels

4.2 Key Users Discovery

In the previous section, we discovered that 2606 users posted reviews for popular movies. In order to select the key users, we set up a two-step approach which includes “participation” and “usefulness value”.

The data analysis for key users’ selection can be found in Table 4. In Table 4, we select data from 100 % of users, top 50 % participation of users and 25 % participation of users. Under the three different data sources, we summarise (itemize) the total users, maximum participation, minimum participation, average participated movies and average usefulness in the table.

Based on the data, we select top 50 % participation users from the 2606 users in the first step. The reason why 50 % is the threshold for selecting users will be explained later. In Table 4, if we select 25 % participation of users, then the

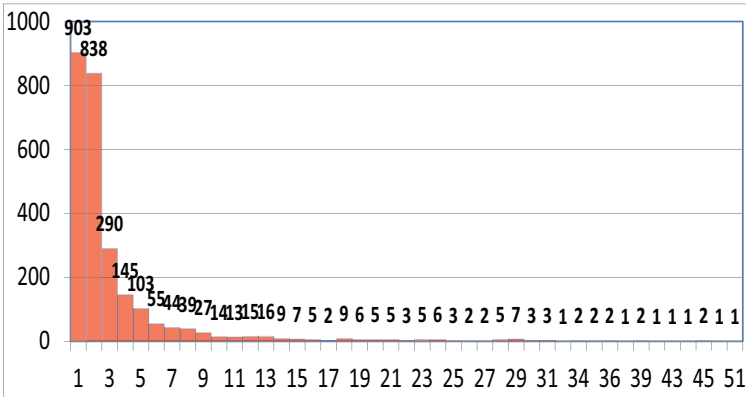


Figure 12. Users' reviews distribution diagram

Data source	100 % of users	50 % of users	25 % of users
Total users	2606	576	161
Maximum participation	51	51	51
Minimum participation	0	4	10
Average participated movies	3.4331	9.57	19.826
Average usefulness	9.14	9.426	6.41

Table 4. Data analysis for key user's selection

average participated movies is 19.826 and average usefulness is 6.41. If we select 50% participation of users, then the average participated movies is 9.57 and the average usefulness is 9.46. Although the average participated movies for 25% of users is more than 50% of users, the average usefulness of 25% of users is lower than 9.14 of 100% users. Therefore, 50% of users is chosen because of acceptable average participated movies and over-average usefulness. After the first step of key users selection, the remaining number of users is 576.

In the second step of key user's exploration discovery, 80/20 rule in small world theory, which is also known as "the Power Law" is exercised. It states that roughly 80% of the outputs aggregate in top 20% of the inputs. Therefore, we apply the theory and believe the top 20% users who have higher usefulness in IMDb are the users who have biggest influence.

According to this rule, we selected the top 20% of users based on the usefulness. The selected key users are listed in Table 5. In the table, UID (the unique user's ID in IMDb) as well as Usefulness, Reviews and AVG-Usefulness is included. The total final key users in Table 5 are 22 users.

UID	Usefulness	Reviews	AVG-Usefulness
21833424	1725	5	345
22236153	1699	5	339.8
3312099	1381	5	276.2
6199900	1262	6	210.3333
2898980	1465	9	162.7778
19404364	1065	7	152.1429
4828685	502	5	100.4
3112639	629	7	89.85714
12449122	529	6	88.16667
6624649	336	4	84
3387663	593	9	65.8889
3580907	215	4	53.75
5555913	364	7	52
2115026	311	6	51.8333
1622466	959	20	47.95
22836725	414	9	46
3798706	1387	31	44.74194
16839649	218	5	43.6
12838671	215	5	43
14676697	171	4	42.75
18439852	170	4	42.5
9932690	189	5	37.8

Table 5. The final discovered key users (22 users)

5 KEY USERS EVALUATION

402 movies in a whole during the data collective period not only make viewers difficult to choose the worth-watching movies by going through all the reviews but also it is a huge job for the companies to frame strategies. Therefore, we can only rely on the key users' reviews to finalize a worth-watching-movie list if the representativeness of those discovered key users is strong enough. In such case, the impact of the discovered key users must be evaluated to ensure the quality of the proposed approach.

A key user evaluation process is shown in Figure 13. Firstly, a so-called voting mechanism is designed. If a movie among the 402 movies has been reviewed by a key user, then it will be counted as voted. If any one of the movie has not been reviewed by any of the 22 key users, then it would be counted as non-voted. Secondly, the actual box office of the movie would be collected to see its popularity. Finally, the voted and not voted movies will be evaluated by using recall and precision and compared to the real box office data.

Table 6 shows the collected box office data including information about maximum, minimum and average box office, number of good selling movies and the

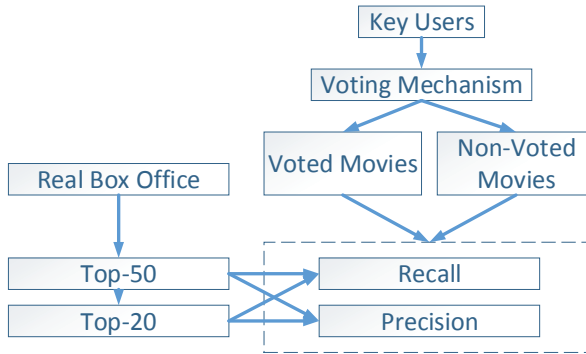


Figure 13. Key users evaluation process

Number of movies	397 (5 without box office data)
Maximum box office	760 507 625 (USD)
Minimum box office	1 087 (USD)
Average box office	36 952 363.91 (USD)
Good selling movies (more than 35 M USD)	125 (31.4 %)
Maximum box office of well selling movies	760 507 625 (USD)
Minimum box office of well selling movies	35 024 475 (USD)

Table 6. Movie box office data

maximum and minimum box office of the good selling movies. However, it is not easy to give a precise number of how much box office revenue is the threshold to judge whether the movie is a good selling one. As a consequence, we are using top 20 or 50 box offices to evaluate the key users in this paper instead of figuring out one threshold to define good selling movie. After being voted, 122 movies have been selected as voted movies and other 280 movies identified as not voted movies. Among the 122 voted movies, 33 movies appeared in the top 50 box office and 13 movies reappeared the top 20 box office. The precision and recall analyses of top 50 and top 20 box office are shown in Tables 7 and 8.

Top 50 movies	Voted movies	Not voted movies	Total
True	33	17	50
False	87	265	352
Total	122	280	402
Precision	27.049 % (33/122)		
Recall	66 % (33/50)		

Table 7. Precision and recall evaluation (top-50)

Tables 7 and 8 show the precision and recall value for top 50 and top 20 movies. In Table 7, we can find the precision is 27.049 % and the recall is 66 %. In Table 8, the precision is 11.417 % and the recall is 70 %. In this paper, we consider the recall value should be more important than precision value. From the result, the key users can help us to catch the 33 out of top-50 movies and 14 out of top-20 movies. It is very helpful for the audiences to reduce the time for selecting movies according to the approach that proposed in this paper.

Top 20 movies	Voted movies	Not voted movies	Total
True	13	6	20
False	108	274	382
Total	122	280	402
Precision	11.475 % (11/122)		
Recall	70 % (14/20)		

Table 8. Precision and recall evaluation (top-20)

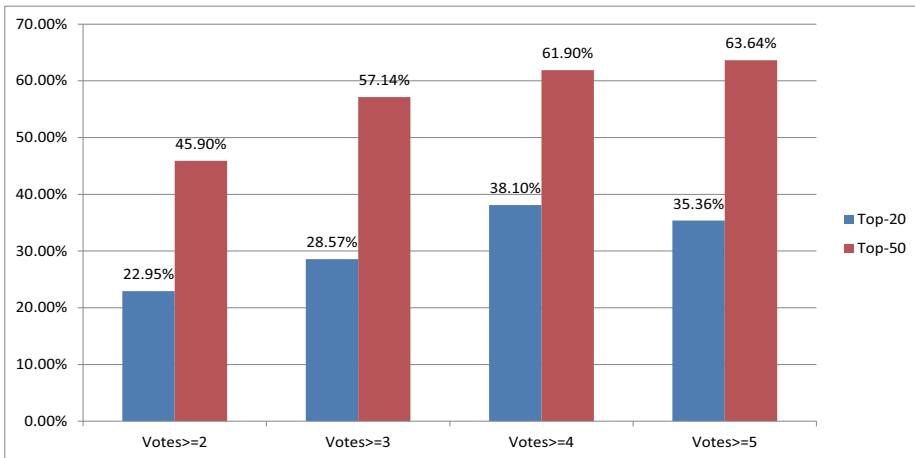


Figure 14. Votes and accuracy

Figure 14 shows another analysis of the accuracy when the votes for movies are different. In the figure, it shows that if more than three key users vote for a movie, then the probability of the movie as a top one is very high, especially for top-50 box office movies.

Another analysis is that if we change the threshold for selecting key users from top 20 % to 40 % users, the recall value of top-50 box office movies increased from 66 % to 76 %. Evidently, the more users from 17366 we selected, the less insecurity for the company to make important decision. Under this situation, it is suggested that companies could select more key users. The analysis result is shown in Table 9 listing the recall value of top-50 movies under different key users selection threshold.

Threshold	20 % users	30 % users	40 % users
Key users	22	34	59
Voted movies	122	157	216
Top-50 movies	33	36	38
Top-20 movies	14	14	14
Top-50 recall	66 %	72 %	76 %

Table 9. The recall analysis for different key users selection threshold

6 CONCLUSION

In this paper, we have proposed an approach to discover key users from IMDb, which is a very popular online movie virtual community. Based on the co-review concept and usefulness value, key users are successfully discovered and the quality of key users has also been evaluated after a series of experiments. The proposed evaluation process is also interesting, especially the voting mechanism. According to the voting mechanism and comparing to the real box office, accuracy of key users' voting can therefore be evaluated. From the experiment results, the recall value of the voted movies (based on top 20 % key users) in the top 20 best selling movies is 70 %. The result is considered as a good one, since it can reduce the loading of viewers to discover best selling movies.

There are two main contributions in the paper. First, the searching time of users can be reduced through applying the key users discovery approach. Second, the verified usefulness value can be a good index to discover key users. According to the experiment results, key users can be discovered very efficiently from the top 20 % users who have higher usefulness value.

Future researches are suggested to focus on the following points.

1. Other possible factors are suggested to be included, such as the cast factor, semantic analysis, etc.
2. In this paper, we only analyse an interest based virtual community, future researches can be considered to study other cases, such as virtual community for food, restaurant or 3C.
3. In this research, we did not take the movie categories into account, such as action, crime, family, comedy, etc.

It would be very interesting to discover key users for each different movie category.

REFERENCES

- [1] ADLER, R. P.—CHRISTOPHER, A. J.: Internet Community Primer Overview and Business Opportunities. Harper Collins Publishing Inc. 1999.
- [2] ADMIC, L. A.—ADAR, E.: Friends and Neighbors on the Web. Social Networks, Vol. 25, 2003, No. 3, pp. 211–230.

- [3] AHMED, A.—BATAGELJ, V.—FU, X.—HONG, S.-H.—MERRICK, D.—MRVAR, A.: Visualisation and Analysis of the Internet Movie Database. Proceedings of the 6th International Asia-Pacific Symposium on Visualization, Sydney, Australia 2007, pp. 17–24.
- [4] ARMSTRONG, A. G.—HAGEL III, J.: Net Gain: Expanding Markets Through Virtual Communities. Harvard Business School Press, MA, USA 1997, pp. 82–97.
- [5] BAYM, N.: Interpreting Soap Operas and Creating Community: Inside An Electronic Fan Culture. In Kielser, S. (Ed.): Culture of the Internet, Mahwah, NJ, USA 1997, pp. 103–119.
- [6] DEBNATH, S.—GANGULY, N.—MITRA, P.: Feature Weighting in Content Based Recommendation System Using Social Network Analysis. Proceedings of the 17th International World Wide Web Conference, Beijing, China 2008, pp. 1041–1042.
- [7] FREEMAN, L. C.: Centrality in Social Networks Conceptual Clarification. Social Networks, Vol. 1, 1978–1979, pp. 215–239.
- [8] HANNEMAN, R. A.—RIDDLE, M.: Introduction to Social Network Methods. Available on: <http://faculty.ucr.edu/hanneman/nettext>.
- [9] JAKOB, N.—WEBER, S. H.—MÜLLER, M. C.—GUREVYCH, I.: Beyond the Stars: Exploiting Free-Text Users Reviews to Improve the Accuracy of Movie Recommendation. Proceedings of The 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong, China 2009, pp. 57–64.
- [10] JUNG, J. J.: An Empirical Study on Optimizing Query Transformation on Semantic Peer-to-Peer Networks. Journal of Intelligent & Fuzzy Systems, Vol. 21, 2010, No. 3, pp. 187–195.
- [11] JUNG, J. J.: Ubiquitous Conference Management System for Mobile Recommendation Services Based on Mobilizing Social Networks: A Case Study of u-Conference. Expert Systems with Applications, Vol. 38, 2011, No. 10, pp. 12786–12790.
- [12] JUNG, J. J.: Evolutionary Approach for Semantic-Based Query Sampling in Large-Scale Information Sources. Information Sciences, Vol. 182, 2012, No. 1, pp. 30–39.
- [13] KUMAR, R.—NOVAK, J.—TOMKINS, A.: Structure and Evolution of Online Social Networks. In Yu et al. (Eds.): Link Mining: Models, Algorithms and Applications, Springer-Verlag 2010, pp. 337–357
- [14] KUPERMAN, M.—ABRAMSON, G.: Small World Effect in An Epidemiological Mode. Physics Review Letter, Vol. 86, 2001, pp. 2909–2912.
- [15] LEWIS, K.—KAUFMAN, J.—GONZALEZ, M.—WIMMER, A.—CHRISTAKIS, N.: Tastes, Ties, and Times: A New Social Network Dataset Using Facebook.com. Social Networks, Vol. 30, Issue 4, 2008, pp. 330–342.
- [16] LIYONG, W.: An Adaptive Evolution Mechanism for Growing Social Networks. In Proceedings of Information Management, Innovation Management and Industrial Engineering, Sanya, China 2008, pp. 320–324.
- [17] MILGRAM, S.: The Small World Problem. Psychology Today, Vol. 1, 1967, No. 1, pp. 61–67.
- [18] QIN, S.—MENEZES, R.—SILAGHI, M.: A Recommendation System for YouTube Based on Its Network of Reviews. In Proceedings of the 2010 IEEE Second In-

- ternational Conference on Social Computing, Minneapolis, Minnesota, USA 2010, pp. 323–328.
- [19] RHEINGOLD, H.: *The Virtual Community: Homesteading on the Electronic Frontier*. The MIT Press, Reading, MA, USA 2000.
 - [20] SCOTT, J. P.: *Social Network Analysis: A Handbook*. Sage Publication, Thousand Oaks, CA, USA 2000.
 - [21] WATTS, D. J.: Collective Dynamics of “Small World” Networks. *Nature*, Vol. 393, 1998, pp. 440–442.
 - [22] ZHANG, Y.—WANG, Z.—XIA, C.: Identifying Key Users for Targeted Marketing by Mining Online Social Network. In *Proceedings of the 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshop 2010*, pp. 644–649.



I-Hsien TING received his Ph.D. from Department of Computer Science, The University of York, United Kingdom in 2007. From 2007, he joined Department of Information Management of National University of Kaohsiung in Taiwan as an Assistant Professor. His research interests include data mining, web mining, social networks analysis and mining, AI for games, etc. He has published more than 50 papers related to the research areas. He is also very active and organized many international conferences and plays the role of program committee member in more than 50 international conferences. Currently, he is the associate

editor in many international journals, such as *Social Networks Analysis and Mining*; he is referee of many top international journals.



Shyue-Liang WANG received his Ph.D. from State University of New York at Stony Brook in 1984. From 1984 to 1994, he joined the University of New Haven and New York Institute of Technology as Assistant/Associate Professor. From 1994 to 2002, he joined I-Shou University in Taiwan and served as Director of Computing Center, Chairman of Information Management Department, and Director of Library. From 2003 to 2007, he rejoined NYIT. From 2009 to 2011, he was Professor and Chairman of Information Management Department at National University of Kaohsiung, Taiwan. He is now Vice President and

Dean of the College of Management. He has published over 180 papers in the areas of data mining, privacy preservation, soft computing, and served as PC member and session chair of more than 80 international conferences. He is a member of the Board of Chinese American Academic and Professional Society, USA.



Po-Jen CHEN received his Master degree in 2012 and B. Sc. degree from Department of Information Management of National University of Kaohsiung. His research interests are social networks analysis, web mining and web techniques. He is now doing his military service.