Send Orders for Reprints to reprints@benthamscience.ae

 The Open Cybernetics & Systemics Journal, 2015, *9,* 1478-1484

Open Access

Study on Genre and its Role in Discourse Information Processing

Yaping Wan[1], Xiaohua Yang[1], Wentao Mo[2], Zhiming Liu[1], Juan Zhang[1,*] and Zhi Li[3]

[1]*School of Computer Science and Technology, University of South China, Hengyang 421001, China;* [2]*Software Center, State Nuclear Power Research Institute, Beijing 100029, China;* [3]*Beijing Aerospace Technology Institute, Beijing 100074, China*

Abstract: The owners of social networks and researchers who are interested in text information processing recognize and use genre to identify network resource and participating in mutually understood communicative acts. Many studies have shown that, the genre can improve the ability to access resource in the information pool, such as music and book community. Some linguists describe genre as a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form. However, in the current studies, how to use genre to mark information is always a difficult thing. A discourse genre is a subtle and complex concept. To some extent, it represents perplexing relationship between a large number of communicative events among social network. In this paper the genre is seen as a multidimensional phenomenon and is denoted using vector. We use vector distance to measure the relationship strength of social networks communicative events. It can be proved automatic discourse classification according to genre in social information sharing, transfer and knowledge communication provides a higher level quality of service(QoS). Finally, we select movie community to investigate the key role of genre in classifying discourse and seizing users interests. The results show that user various behaviors stickiness in network community and discourse genre relationship intensity have potential common features. The more similar discourse genre is, the more similar of social network users all kinds of behavior. Through these we can identify new genre and illustrate that genres are used for coordinating information addressing aspects of coordination mechanisms such as divisibility, concurrency, accessibility and timing that help people improve the coordination of information process.

Keywords: Genre vector, information process, measurement, user behavior.

1. INTRODUCTION

The rapid development of global computer and communication technology and popularization of Internet due to the exponential growth in the amount of information. The registered number of social network [1, 2] is growing year by year, and its content is constantly updated, which provides an unprecedented real experimental platform for the study of large-scale social network. The traditional classification techniques use text content build indexing model, but the social network contains a variety of structured and unstructured data (such as the various video, audio, graphics, image information), which makes it difficult to use the traditional classification methods to recommended or obtain information. This situation demands for tools able to ease searching, retrieving, and handling such a huge amount of data. Among those tools, automatic genre classifiers have a particularly important role, since they could be able to automatically index and retrieve all kinds of data in a human-independent way. This is very useful because a large portion of the words used to describe Web content is inconsistent or incomplete.

In recent years, Computational Linguistics, a lot of text classification researchers recognize the importance of the genre classification, and the emergence of an important theoretical turn, this is to say, scholar's interest is from the content classification to pay equal attention to the genre. Genre has become one of hotspots in the field of computational linguistics, information science [3, 4]. The genre is explicit formal specification of a shared conceptual model, which can be used as a clear description of the definition of social networking information resources and can support knowledge sharing and reusing. It's summed up the genre at least has the following advantages:

- Information interaction. Through genre, human can express themselves more clearly, at the same time to obtain the resources they need to, especially to those non-textual resources which are usually difficult to get through the way of text content searching, such as music, video, image. Genre will make communications more easily recognizable and understandable by recipients.
- Resource annotation. To the owners of all kinds of web resources, how to identify information resource for the information processing that is convenient for the user to obtain cyber source is the first issue to seriously considered. Automatic genre identification is one of the key factors in improving the often inadequate results of search engines, as the user would be able to specify the desired Web genre along with a set of keywords.

1874-110X/15

- Filtering and Ranking. The organization of documents, bookmarks, or digital document identifiers can occur topic-centered, genre-centered, or in a combined fashion. Having identified the underlying paradigm one can provide user guidance for filing, give hints or special views for browsing and searching, and identify classes that are not properly organized. Genre information provides meta knowledge for automatic Web page abstraction, which is concerned with the preparation of Web pages in a consistent and clearly arranged form.

However, the description of the genre is a complex and challenging work from the social discourse. Genre is the results of the human mind abstract generalization in a certain historical period. That the limitations of human recognizing and its own dynamic evolution make comprehensive, accurate, and effectively summarized large and complex genre category is very difficult. Second, there is no strict logical reasoning and proof about genre classification, so its classification criteria is rather vague and the boundaries is overlapping, which make it is difficult to protect the independence and uniqueness of the genre category [5]. Some researchers had begun to study the genre very early, but still in the exploratory stage. Even there are some good applications, the domain-specific theory and methods are difficult to be widespread use [6, 7].

The aim of the paper is two-fold. First the paper serves as a theoretical exploration of the genre model in general. It attempts to establish whether the measurement is suitable for capturing the essence of web-mediated genres or whether the digital context of genres may call for a reconsideration of- or at least provide new insights into- the constituents of the genre model. Second, even though a systematic characterization of web-mediated genres is outside the scope of this paper, we use the books tag as exemplary material in our theoretical discussion and in that way provide a tentative characterisation of the books tag as a genre. The reasons for choosing the books tag are it is a web-generated genre in the sense that it came into existence with the advent of the WWW and has no direct parallel outside the Web. Our goal is to build user interest model based on genre and user behavior.

2. RELATE WORK

Since the end of the twentieth century, researchers recognize the importance of text classification in accordance with the genre. Genre research in linguistics had a long history, however, a more comprehensive and systematic genre study is after the foundation of modern linguistics. Especially in recent years, theoretical and applied research on the genre has gradually penetrated into the field of information science.

Swales was one of the earliest scholars studied genre. He defined genre as "purpose, form". He believed genre was made up of a group of communicative event, which have some common goals. These goals are identified by the parent discourse groups, and thus form the basic principles of the genre [8]. Swales discussed the relationship between genre identification and communication goals, pointing out that communication target is used as discourse classification standard. He thought genre analysis is text-driven and context-driven process [9]. Kevin Crowston [5] agreed Orlikowskiand Yates's definition [10] because he took into account all three aspects of genre that they recognized as fundamental: content, form, and purpose. They thought genre as a multidimensional phenomenon, which took into account not only the document attributes, but its role in human endeavor. Rosso explored the use of genre as a document descriptor in order to improve the effectiveness of Web searching [11]. They conducted three user studies to develop a genre palette and show that it is recognizable to users. In an online experiment in which 257 participants categorized a new set of 55 pages using the 18 genres, on average, over 70% agreed on the genre of each page.

Inger Askehave and Anne Ellerup Nielsen [12] considered the genre characteristics were on-linear, multi-modal, web-mediated documents. They found that most genre research has focused on the characteristics of "printed" texts, whereas less has been one to apply the genre theory to digital genres. Santini [13] Strictly distinguished between the differences of "text type" and genre. She presented an automatic genre classification model that implements a flexible classification scheme. Then in 2011, Santini presented a genre classification model which was cross-tested with a number of genre collections [14]. From the view of this article, genres are based on more or less tight conventions that allows people to reconstruct or infer the context in which texts have been produced, together with their purposes and functions. In the application of genre, Freund [6], Vidulin [15] and Rosso, [16], etc. used genre to represent the document targets. The mainstream of current search engine optimization is that using genre to denote information retrieval query goal. TGSE Seminar (Towards Genre-Enabled Search Engines) devoted to the use of genres to improve search engine in 2007. Some researchers have used the genre to filter the search engines return results [6, 15]. Freund's findings also showed that the relationships of the document genre and user current task can effectively improve the quality of information retrieval.

3. CHARACTERISTICS OF USER INTEREST MODEL BASED ON GENRE AND USER BEHAVIOR

As an important feature of information process object, genre constructs another dimension which is perpendicular to topic. Genre reflects relationship of information objects and tasks, and is developing with the expansion of user knowledge. The complexity of the information object genre determines the user interest model based on genre has many different characteristics.

i) Roughness. Genre model is derived from the user's own information task in which it can be constantly refinement and evolution by the user feedback. Different with knowledge engineering of general meaning, genre model does not come from domain experts and knowledge engineering. Regularity, accuracy and authority that information process task is eager for are certainly not as good as experts doing.

ii) Finiteness. The user interest model which is restricted by user information tasks depicts object that is user information involved. The type of information process object is limited, so the type of genre is limited, especially compared with all genres.

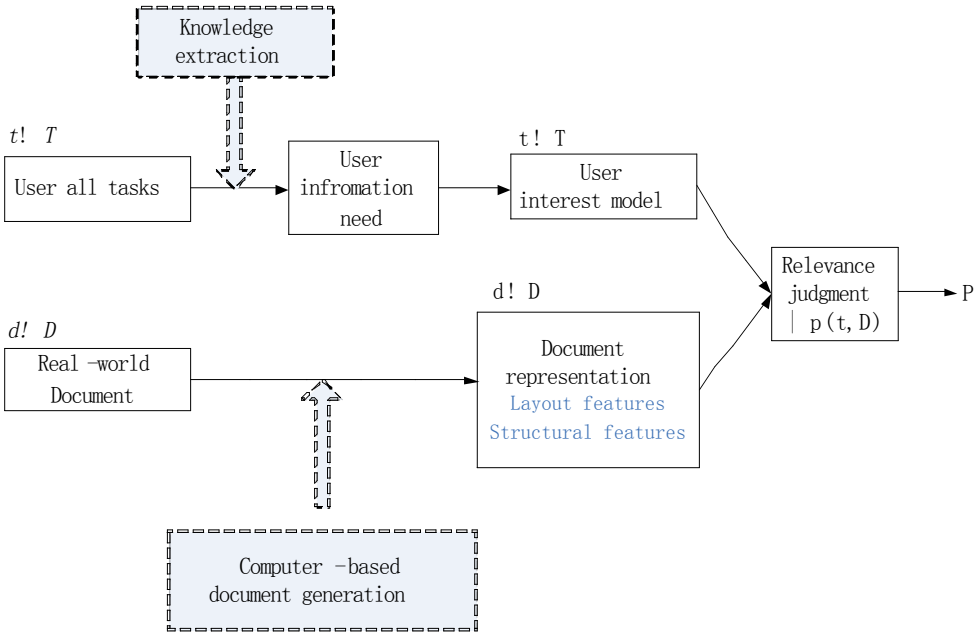

Fig. (1). Traditional user interest model in the recommendation system.

iii) Intersection. The user information tasks may involve a number of areas. The divergent thinking and the complexity of task tend to make task consumer information object not limited in one area, but crossing of multiple information objects. So genres of information space have the characteristics of cross and merge.

iv) Dynamics. Genre expresses user information tasks and a cluster of the same type object. With the increasing of user information tasks, information object samples are increasing. The genre model that is dynamic update according to the changes of user information task and the development of information objects has a evolution process which are from coarse to fine and from shallower to deeper.

v) Timeliness. The shortest period may be just a single query about a user to some areas of concern, but as long as the whole life. There will be great differences in time and the user's interest has certain period of validity. The user interest model itself has a certain life cycle, so the dynamic changes of user information objects and tasks also reflect the validity of model.

3.1. Traditional User Interest Model in Information Recommendation System

User interest models represent certain kinds of discourse models which captures seek-specific aspects of a real-world discourse such that an information need or a query task at hand can be efficiently addressed. Recommendation system can accurately and efficiently recommend various web resource in accordance with the user interest model.

So, how to represent the use interest model is one of the key technologies in information recommendation domain. The traditional use interest model will be the first to introduced in this section as Fig. (**1**).

Definition 1 (User Interest Model in the Recommendation System). Let D be a set of documents, and let T be a set of user information tasks. A user interest model P for D and T is a tuple <D, T, ρP>, whose elements are defined as follows:

1. D is the set of representations of the documents D. d∈D may capture layout aspects, the logical structure, or semantic aspects of a document d∈D.

2. T is the set of task contexts or formalized information needs.

3. ρP is the retrieval function and quantifies, as a real number, the relevance of a document representation d∈D with respect to a query representation t∈T:

$$\rho P: T \times D \rightarrow P$$

3.2. User Interest Model Based Genre in Information Recommendation System

Though various user interest models have been proposed, but it is difficult to accurately represent the user task context that analyzes conceptual relations of the kind shown in Fig. (**1**). Actually, real world not only has structured documents, but all kinds of unstructured discourses which mainly refer to voice, video, image, etc. And the relation between real-world discourses D and associated discourse representations D should allow any kind of transformation. it is illustrated here as shown in Fig. (**2**).

Definition 2 (User Interest Model Based Genre). Let D be a set of discourse which Specifically refers to those unstructured information or resources, and let T be a set of user information tasks. A user interest model P for D and T is a tuple <D, T, ρP>, whose elements are defined as follows:

1. D is the set of representations of the discourses D. d∈D.

2. T is the set of task contexts or formalized information needs.

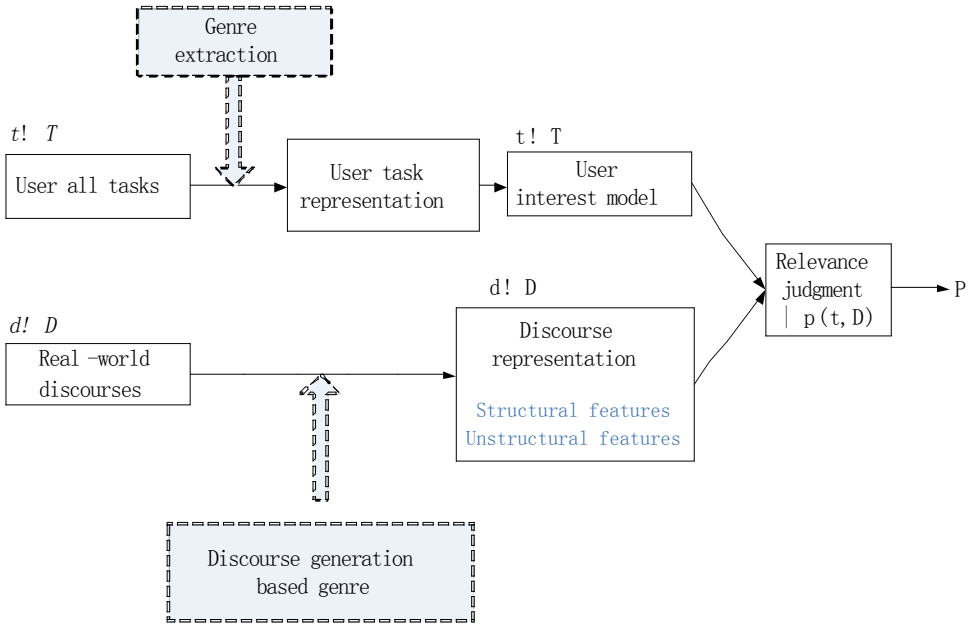

Fig. (2). User interest model based genre.

3. ρP is the retrieval function and quantifies, as a real number, the relevance of a discourse representation d∈D with respect to a task representation t∈T:

$$\rho P: T \times D \rightarrow P$$

The development of interest models is an active research field with various open questions. In spite of its simplicity the vector space model is quite successful; recent work focuses on probabilistic models as well as on models that rely on hidden variables. With respect to special purpose tasks, such as genre classification, even less is known concerning the user's information need and the adequacy of genre models. Though genre models are special document or discourse models they are constructed quite differently: different kinds of simple and complex features, among others from the field of natural language processing, are combined and statistically optimized to capture the "gist" of a genre class.

4. MEASUREMENT OF GENRE

4.1. Initialization and Evolution

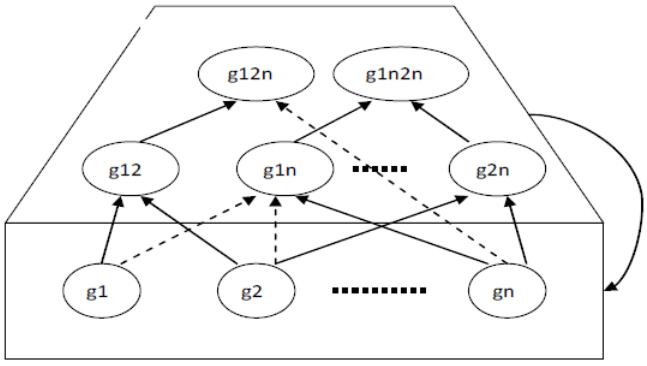

Fig. (3). Evolution of genre.

As can be seen from Fig. (**3**), the construction of genre model is divided into two stages. The first phase is initialization of model. The relationships of initial information task and object are obtained through statistical language method and the initial Genre module diagram is established. The next stage is evolution of genres. The new genre classes are generated and interactive evolution by user information task. The user information tasks establish contacts in different information objects, and add constantly the samples of genre class. That machine learning and pattern mining of supervision can make genre diagram with the task interaction and sample changes constantly adjust. New genre class that consistent with the roughness and rough membership will sooner or later translate into genre base vector.

4.2. Representation and Calculation of Genre Vector

Halliday, the founder of systemic functional linguistics school, firstly gives the definition of discourse. He states that "the word text is used in linguistics to refer to any passage, spoken or written, of whatever length, that does form a unified whole"[17]. This paper does not intend to explore the concept of discourse from a linguistic view, and our goal is only to introduce it to the information space in which discourse will be taken as task context that is related to user's interest and behavior in social network. This view is of a piece with the above expression. Discourse genre is a cross-mixing of the base genre. Therefore, we give the following definition,

Definition 3. Given discourse space U, G={g1,g2,…,gn} is the set of base genres of the U, the genre Dg of the discourse D is an n-dimensional vector (dg1,…,dgn), in which dgi ∈ [0,1],

dgi=0，if D has no characteristic of base genre gi，

dgi>0，if D has characteristic of base genre gi。

Definition 4. If discourse D's n n-dimensional vectors are linearly independent, n Independent vectors become n base vector to measure the n-dimensional discourse space.

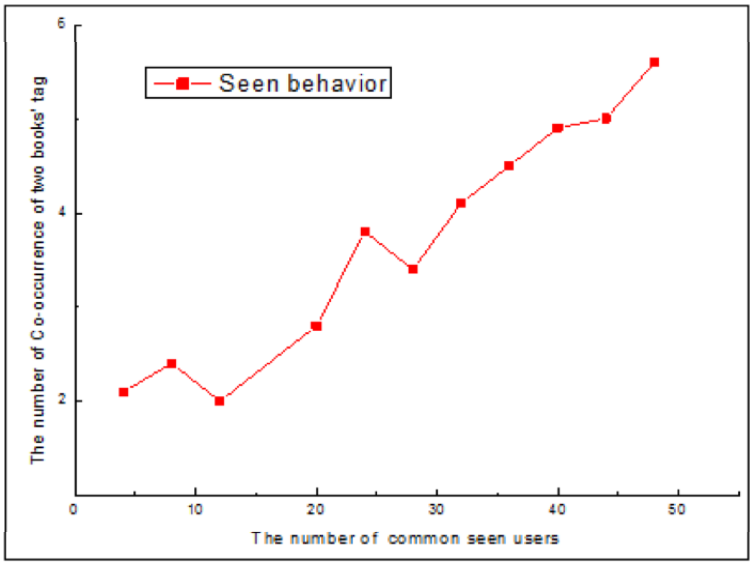

Fig. (4). Relationship of user seen behavior and books tag.

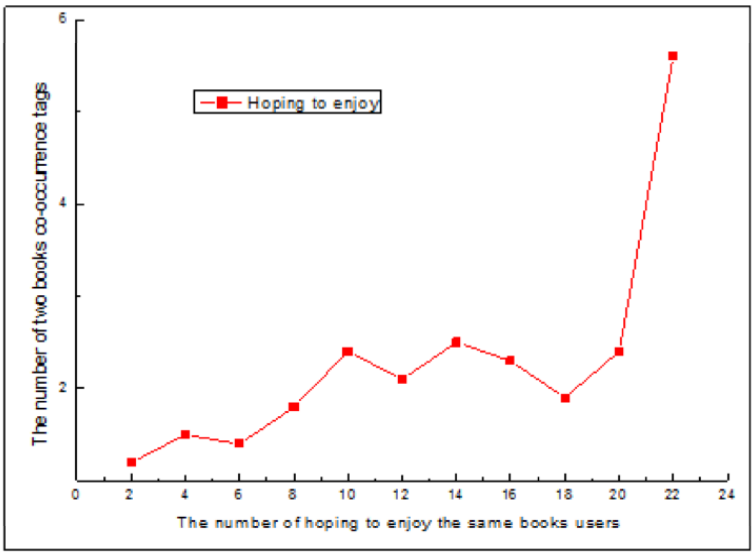

Fig. (5). Relationship of hoping to enjoy behavior and books tag.

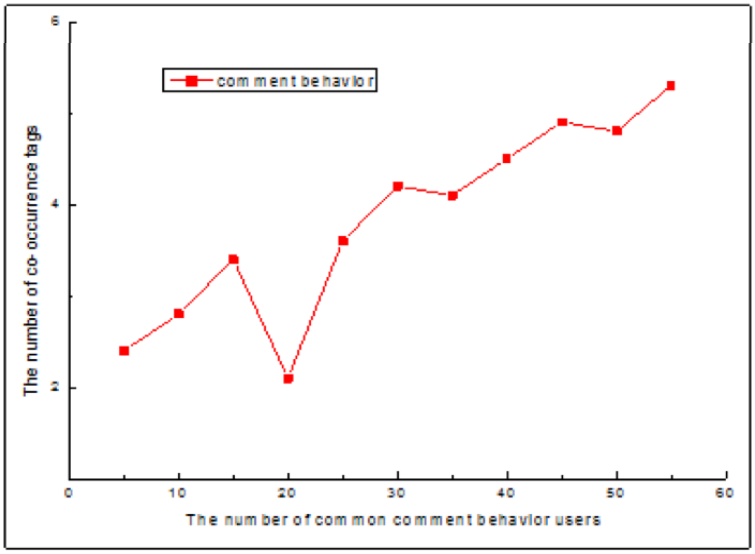

Fig. (6). Relationship of comment behavior and books tag.

Table 1. Douban book genres.

Novel (5614836)	**Essay (1251173)**	**Prose (3776563)**	**Fairy tale (2383568)**
Poem (3263941)	Masterpiece (5203323)	Cartoon (2354168)	Picture book (1758621)
Suspense (1245872)	Reasoning (1452145)	Romance (985647)	Science fiction (758649)
Martial arts (1287321)	Fantasy (653281)	History (751893)	Philosophy (561892)
Biography (358968)	Design (238957)	Memoirs (365879)	Music (587624)
Travel (128597)	Inspirational (389576)	Workplace (128765)	Education (752416)
Food (1869421)	Food (1869421)	Health (2486273)	Home (385694)
Management (853695)	Financial (178546)	Business (235847)	Marketing (156234)
Financial management (215546)	Stock (456231)	Popular Science (52648)	Internet (1025844)

Property 1. Between any two discourse D1 and D2 in discourse space, define using genre vector distance measure genres similarity betweenD1 and D2. The smaller the distance is, the higher the genre similarity of D1 and D2. Assume that the genre vectors of two discourses D1, D2 were,

Dg1= (dg11,dg12,dg13,…,dg1n)

Dg2= (dg21,dg22,dg23,…,dg2n)

Where, n is the number of base genre.

The similarity of D1 and D2 is $\mathrm{sim}(D_{g1}, D_{g2})$,

$$\mathrm{sim}(Dg1, Dg2)=\cos(Dg1, Dg2)=\frac{\sum_{k=1}^{n}(dg_{1,k}\times dg_{2,k})}{\sqrt{\sum_{k=1}^{n}dg_{1,k}^2}\times\sqrt{\sum_{k=1}^{n}dg_{2,k}^2}} \quad (1)$$

In equation (1), we use the cosine vector to measure the discourse genre relationship. Through it we can identify new discourse genre and implement automatic discourse genre tag. Further, we can render genre information a powerful filter technology for information processors in network community. To discourse which genre is known, using vector space model to identify the genre. For those unknown genres of discourses by analyzing the potential common characteristics between genre and user behavior, we can use user behavior to measure the genre, and therefore leads to the following Property 2,

Property 2. The number of co-occurrence among community users behavior can quantify the strength relationship between discourse genres. This proved discourses that most users frequently operate (such as read, click) have some common base genres.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Experiment of Genre and User Task Context

This paper selects Douban (http://book.douban.com) data as the basis of experiment, in which the different labels, the number of users and user behavior data are selected to statistics and analysis. And we use MySql database to stored the download data in the location.

This experiment in which we collected a total of 2783 books using the 36 popular tags which are selected from hundreds of tags. The same as above, these 36 tags are as shown in Table **1**.

The digital in Table **1** following each genre expresses the total number of the books of this genre in the Douban book community.

Experiments designing are as follows:

Purpose of the experiment: a survey on the number of people that any two films are common seen and the common tags that two films had verified the potential commonalities relationship of genre and user behavior.

Experimental Procedure:

(1) Firstly, we download original data, users, behavior data of user on the book (mainly the four user behavior data: see, read, want to see, the critic).

(2) Filter users: to filter out the users who had seen a certain amount of books and had kept a certain preference.

(3) Filter book:

Step 1: We filter out all book which constituted the film set F1 and were seen by the set of users U1;

Step 2: The books which are selected from F1 which had been seen over a certain value by U1 constituted the set F2;

Step 3: That the number of users who had seen (or the user behavior of seeing, book critic) any two books which were selected from the collection F2 and the number of common tags are gotten. The entire experimental results are shown in Figs. (**4-6**).

At the same time, the statistical of user behavior data which user are seeing book also had been collected, but the amount of online data is too little to discover the laws, so this paper temporarily does not analyze and discuss them.

As can be seen from Figs. (**4-6**) the behavior which users had seen these books is of most consistent with our basic design ideas. This is due to the kind of data is the most in all four behaviors we collected. Statistical thinking reveals that the greater the capacity of the sample is, the smaller the sampling error. On the other hand, users want to see and book critic behavioral data also are basically in line with our hypothesis. This shows that the negotiations between the social network and Linguistics discourse have common characteristics and user behavior in social networks and genre features have great similarities. In other words, the more similar discourse genre is, the more similar behavior of users of social network is.

CONCLUSION

With the enhancement of network penetration and the popularity of network applications, the scale of social networking users will be further expanding, and more and more users will extend real life relationships to network. This study shows that users in the social network information behavior have a high degree of user viscosity. Users always focus on the discourse genre of their own interest, and the discourse set most users are interested is greater similarity in the genre. Also due to the use of different purposes, the user behaviors exhibited. Overall, the community user behavior on social networking sites are more dispersed, but the discourse genre relationships embodied communication in a fixed time when a particular discourse genre is concerned is still the center of user behavior.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Project supported by National Natural Science Foundation of China (61402220), The Hunan province science and technology Program (2013FJ3030), The Hunan Province Philosophy Social Science Fund (14YBA335), The Hunan Colleges and Universities Ideological and Political Education Research Fund (2014[14C10]), Doctor Start Fund of the University of South China (2011XQD39), construct program of the key discipline in University of South China, construct program of the key laboratory in University of South China, construct program for Science and technology innovative research team in University of South China.

REFERENCES

[1] K. Weize, L. Yiqun, Z. Min, and M. A. Shaoping, "Answer quality analysis on community question answering", *Journal of Chinese Information Processing*, vol. 25, no. 1, pp. 3-8, 2011.

[2] W. Yu-Xiang, Q. Xiu-Quan, L. Xiao-Feng, and M. Luo-Ming, "Research on context-awareness mobile sns service selection mechanism", *Chinese Journal of Computers*, vol. 33, no. 11, pp. 2126-2135, 2010.

[3] F. Zhi-fei, L. Hong-fei, Y. Zhi-hao, and Z. Jing, "Automatic classification of chinese text genre", *Journal of Chinese Information Processing*, vol. 20, no. 2, pp. 24-32, 2006.

[4] B. Stein, and S. M. zu Eissen, "Retrieval models for genre classification", *Journal of Information Systems*, vol. 20, no. 1, pp. 93-119, 2008.

[5] K. Crowston, and B. H. Kwasnik, "A framework for creating a facetted classification for genres: addressing issues of multidimensionality", In: *Proceedings of the 37th Hawaii International Conference on System Sciences*, Big Island, HI, USA, 2004,pp. 1-9.

[6] L. S. Freund, "*Exploiting Task-Document Relations in Support of Information Retrieval in the Workplace*", PhD thesis, University of Toronto, 2008.

[7] V. Vidulin, M. Luštrek, and M. Gams, "Using genres to improve search engines", In: *Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP*", Borovets, Bulgaria, September, 2007.

[8] M. Swales, *Genre analysis: English in academic and research settings*, Cambridge University Press, Cambridge, 1990.

[9] I. Askehave, and M. Swales, "Genre identification and comminicative purpose: a problem and a possible solution", *Applied Linguistics*, vol. 22, no. 2, pp. 195-212, 2001.

[10] W. J. Orlikowski, and J. Yates, "Genre repertoire: the structuring of communicative practices in organizations", *Administrative Sciences*, vol. 33, no. 1, pp. 541-574, 1994.

[11] M. A. Rosso, "User-based identification of web genres", *Journal of the American Society for Information Science and Technology*, vol. 59, no. 7, pp. 1053-1072, 2008.

[12] I. Askehave, and A. E. Nielsen, "Digital genres: a challenge to traditional genre theory", *Information Technology & People*, vol. 18, no. 2, pp. 120-141, 2005.

[13] M. Santini, "Automatic genre identification: towards a flexible classification scheme", In: *BCS IRSG Symposium: Proceedings of Future Directions in Information Access*, Glasgow, Scotland, 2007.

[14] M. Santini, "Cross-testing a genre classification model for the web", *Text, Speech and Language Technology*, vol. 42, no. 3, pp. 87-128, 2011.

[15] V. Vidulin, M. Luštrek, and M. Gams, "Using genres to improve search engines", In: *Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP*", Borovets, Bulgaria, September, 2007.

[16] M. A. Rosso, "User-based identification of web genre", *Journal of the American Society for Information Science and Technology*, vol. 59, no. 7, pp. 1053-1072, 2008.

[17] M. A. K. Halliday, *Cohesion in English*, Longman, London, 1976.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015