早稲田大学審査学位論文
博士（人間科学）

Unified Modeling and Analyzing of Personal Data and Behaviors for Individualized Information Utilization

個人化情報活用のためのパーソナルデータと挙動解析による統合モデリング手法

２０１４年７月

早稲田大学大学院　人間科学研究科

周　暁康

ZHOU, Xiaokang

研究指導教員：　金　群　教授

Abstract

With the rapid development of emerging computing paradigms, such as Ubiquitous Computing, Cloud Computing, Social Computing, and Mobile Computing, we have been continuously experiencing a fast change from all walks of our work, life, learning and entertainment. The high accessibility of SNS (Social Networking Service), coupled with the increasingly widespread adoption of wireless mobile computing devices, enables more and more populations to continuously generate larger amount of data from different environments, which represents more information in terms of an individual's behavioral habits and daily routines. It is said that a new era of big data has arrived. Among those big data, the so-called personal data, not only referring to the collection of data that is generated from an individual, but also any data that is related to an individual, has become a crucial source of innovation and value. It is indubitable that the valuable information hidden in personal big data can benefit individuals in various aspects.

Generally, information utilization is regarding various applications involving people and information together, including information seeking and recommendation for an individual, and information sharing and knowledge creation within a group. To take advantage of the considerable size of personal data for individualized information utilization, it is of crucial importance to build a well-structured user

model, further make sense of individuals' intentions or needs, analyze various information behaviors and social activities, and better understand the user contexts. However, in many cases, the large scales of personal big data, dynamically generated from a variety of systems and different devices, is always with different data structures or no structure, and coupled with lots of useless noise data. It makes it difficult to find piece of relevant information that fits users' time-varying needs. Research works have tried to find a flexible and efficient solution, but it's still far from satisfaction due to the more complexity and heterogeneity of the personal big data.

In this study, to facilitate individualized information utilization and sharing not only for the individuals, but also for the groups or communities, we concentrate on the computational approaches to unified modeling and analyzing of the personal data and behaviors. The chaotic data will be systematically organized and managed to form the associative information, including time-varying individual intentions and additional information for the descriptions of data relations. The individuals' information behaviors and social activities will be analyzed to extract the behavioral features and calculate the similarities among them. Furthermore, the related individuals will be connected in a dynamically socialized networking according to the calculation of their dynamical and potential correlations. Their multi-dimensional profiling will be built,

and the social communities will be discovered based on the outcomes from the analysis of personal data. Finally, both the behavior patterns and user correlations will be considered together to develop an integrated recommendation mechanism to provide the users with individualized support.

Firstly, a unified framework of data integration and organization called *Organic Streams*, is proposed to analyze and organize the personal big data. The new concept of organic stream, which is designed as a flexibly extensible data carrier, is introduced and defined to provide a simple but efficient means to formulate, organize, and represent the personal big data with inherent and potential relations. It can also be regarded as a logic metaphor to meaningfully analyze and process the raw stream data into an associatively organized form based on the individual needs. A heuristic mechanism is developed and applied to capture users' time-varying interests or needs, and aggregate and integrate the relevant data together to obtain the associative information.

A behavioral analysis approach is proposed to detect and calculate the social influence hidden in the individual behaviors, and model and analyze the sequential behaviors in the task-oriented processes with formal descriptions. The action patterns are extracted from an individual user's sequential behaviors toward a certain purpose, and the behavioral similarities among a group of users are then calculated and

described based on the action patterns. The perceived social influence can be utilized to analyze and describe users' social relationships, and the extracted action patterns as well as their similarities can help improve the quality of user contexts and assist the recommendations in the task-oriented processes.

Based on these basic models and methods, the *DSUN* (Dynamically Socialized User Networking) model, as a viable alternative way to obtain larger information sources and connect more and more people together, is constructed to describe and represent users' implicit and explicit social relationships, namely the characteristics -based relationships and influence-based relationships, using the valuable outcomes of analyzing personal data and individual behaviors. A set of measures are introduced and defined to measure and describe the detail of user correlations, which can dynamically calculate and build a connection between two related people in a specific time period, according to their static and dynamical feature based similarity and interactional behavior based social influence. A series of attributes are defined and analyzed to build the multi-dimensional user profiling, which can facilitate the search of information sources by finding favorable users in both global (e.g., *hub user* and *promotion user*) and personalized (e.g., *contribution user* and *reference user*) way. Three algorithms are developed to discover the multi-types of social communities considering both the dynamical user correlations (e.g., *strong correlation-based tie*

and *weak correlation-based tie*) and profiling (e.g., *user profiling-based tie*), which can recommend users to join different communities in accordance with their different intentions, so as to promote the information sharing and collaborative work.

Moreover, as an application of the proposed approaches, an integrated recommendation method is proposed to provide users with the individualized learning guidance and support. A hierarchical model is presented to describe the relations among learning actions, activities, sub-tasks and tasks within a user community for the task-oriented learning process. The *LA-Pattern* (Learning Action Pattern) is defined to discover and represent an individual user's learning behavior patterns extracted from sequences of learning actions, and the *Goal-driven Learning Group* is proposed to analyze and describe the similarities of learning behaviors among a group of users. Based on these, an integrated mechanism is developed for the goal-driven learning recommendation in accordance with the analysis of behavior patterns and user correlations, which can provide the target user with the most suitable learning action as the appropriate next learning step to complete a specific learning goal.

To demonstrate the feasibility and effectiveness of our methods, two experimental studies are conducted respectively. The experimental results with the analyses conducted using the Twitter data demonstrate the high usability and practicability of our proposed *DSUN* model which can assist personalized information

utilization and sharing in both favorable user finding and social community discovering. The empirical analysis results conducted in a community-based learning system illustrated that the calculated LA-Patterns and Goal-driven Learning Groups can correctly describe the users' learning behaviors and their similarities as well, which can be applied to frequency-based learning pattern recognition and categorization according to different learning goals. And the evaluation results showed the usefulness of our proposed recommendation method that can effectively guide users to pursue their learning purposes and facilitate the task-oriented learning process.

This study is expected to benefit both individuals and communities, not only in the systematical processing of personal big data which can capture the time-varying individual needs and generate associative information from the chaotic data to facilitate the personalized information seeking and social knowledge creation, but also in the dynamical constructing of social relationships which can help build a well-structured user model and involve increasing people into a well-connected social networking to promote the information sharing and recommending. The unified modeling and analyzing approach presented in this study can facilitate the individualized information utilization from chaotic data to associative information, and further to connected people.

TABLE OF CONTENT

Chapter 1 Introduction 1

1.1 Background 1

1.2 Purpose of this Study 4

1.3 Contributions of this Study 6

1.4 Organization of the Thesis 7

Chapter 2 Related Work 10

2.1 Social Media Application and Life Log Analysis 10

2.2 User Relationship Analysis 13

2.2.1 User Correlation Analysis 13

2.2.2 Social Influence Analysis 14

2.2.3 Social Community Discovery 15

2.3 Information Behavior Modeling and Pattern Analysis 16

2.4 Summary 18

Chapter 3 Analysis of Personal Data and Behaviors: Definition and Model 20

3.1 Organizing of Personal Stream Data 20

3.1.1 Metaphors for Organizing Process 20

3.1.2 Organic Streams 22

3.1.3 A Scenario for Enrichment of User Search Experience 31

3.2 Analysis of Individual Behaviors 34

3.2.1 Detecting Influence from Individual Behaviors 34

3.2.2 Analyzing Sequential Action Behaviors 36

3.3 Summary 39

Chapter 4 Dynamically Socialized User Networking 42

4.1 Constructing of DSUN Model 42

4.1.1 The Basic Model 42

4.1.2 User Relationship Description 44

4.1.3 User Characteristics Similarity Analysis 46

4.1.4 User Influence Behavior Analysis 50

4.2 User Correlation and Profiling Analysis 52

4.2.1 Measures for User Correlations 53

4.2.2 Attributes for User Profiling 56

4.3 Mechanisms for Social Community Discovery 62

4.4 Experiments on DSUN Model 66

4.4.1 System Architecture of User and Community Recommendation 66

4.4.2 Data Set for DSUN Model Experiments 68

4.4.3 Experimental Results of User Profiling 70

4.4.4 Experimental Results of Social Community Discovery 75

4.4.5 Discussions 81

4.5 Summary 84

Chapter 5 Task-Oriented Recommendation for Learning Support 86

5.1 Definitions and Hierarchical Model 86

5.2 Similarity of Learning Action Behaviors 89

5.2.1 Generating of Learning Action Patterns 89

5.2.2 Generating of Goal-Driven Learning Group 91

5.3 User Correlation Analysis in Learning Processes 94

5.4 Recommendation in Task-Oriented Processes 97

5.4.1 Detection of Goal-Driven Learning Action Patterns 97

5.4.2 Goal-driven Learning Recommendation Mechanism 99

5.5 Experiments on Learning Recommendation 102

5.5.1 System Architecture of Task-Oriented Recommendation 103

5.5.2 Data Set for Learning Action Experiments 104

5.5.3 LA-Pattern Analysis Results 106

5.5.4 Evaluation 113

5.6 Summary 119

Chapter 6 Conclusions 121

6.1 Summary of this Study 121

6.2 The Limitations 127

6.3 Future Works 128

Acknowledgements 129

Bibliography 130

LIST OF FIGURES

Figure 1-1 Facilitation from Associative Data to Connected People.....5

Figure 3-1 Graph Model for Stream Data.....21

Figure 3-2 Image of Organization Process of Stream Data 25

Figure 3-3 Illustration of Dynamical Division of Time Slices 27

Figure 3-4 Generation of Associative Ripples.....29

Figure 3-5 Algorithm for Generating Associative Ripples 31

Figure 3-6 The Scenario of Enhanced Information Seeking.....32

Figure 4-1 Conceptual Image of Building Dynamical Similarity-Based Relationships.....49

Figure 4-2 Illustration of Influence-Based Relationship 50

Figure 4-3 Algorithm for Generation of Strong Correlation-Based Tie 63

Figure 4-4 Algorithm for Generation of Weak Correlation-Based Tie 64

Figure 4-5 Algorithm for Generation of User Profiling-Based Tie.....65

Figure 4-6 Architecture of User Correlation/Profiling-Based Recommendation System.....68

Figure 4-7 An Image of A Specific User's Contribution and Reference Users ...73

Figure 4-8 Images of Different Types of User Communities 76

Figure 4-9 Changing in Size of Communities According to Different Thresholds

.....78

Figure 4-10 Changing in Size of User Profiling-Based Ties for Different Hub Users 80

Figure 4-11 Image of Information Dissemination in User Profiling-Based Tie ..81

Figure 5-1 Model of Task-Oriented Learning Process.....88

Figure 5-2 Algorithm for LA-Pattern Generation.....91

Figure 5-3 Algorithm for Goal-driven Learning Group Generation.....92

Figure 5-4 Algorithm for Goal-Driven Learning Action Pattern Detection 98

Figure 5-5 Conceptual Process of Learning Action Recommendation.....99

Figure 5-6 Algorithm for Learning Action Recommendation 102

Figure 5-7 Architecture of Task-Oriented Recommendation System.....103

Figure 5-8 Statistics for Goal-driven Learning Groups 108

Figure 5-9 Statistics and Analysis for LA-Patterns in Each Group 110

Figure 5-10 Evaluation Results in Comparison.....115

Figure 5-11 Usefulness Evaluation for Recommendation Results 117

Figure 5-12 Assessment Results Based on Table 5-2 118

LIST OF TABLES

Table 3-1 Descriptions of Information Behaviors 35

Table 4-1 Description for Static Features and Rules for Scoring 46

Table 4-2 Results of Top 10 Users for Basic Attributes 70

Table 4-3 Results of Top 10 Hub and Promotion Users 71

Table 4-4 Results of Top 5 Contribution Users and Reference Users for a Specific User 74

Table 4-5 Statistics for Numbers of Communities According to Different Thresholds 77

Table 4-6 Statistics for Hub Users in User Profiling-Based Ties 79

Table 5-1 Learning Actions and Their Notations 105

Table 5-2 Values for Assessment of Recommended Actions 118

Chapter 1 Introduction

In this chapter, we introduce the background of this study. The motivation and purpose of this study are discussed. We then present the major contributions of this thesis. The organization of this thesis is addressed in the end.

1.1 Background

With the continuous development of social media (such as Twitter, Facebook), more and more populations have been involved into this social networking revolution, which leads to a tremendous increase of data scale, ranging from the daily text data to multimedia data that describes different aspects of people's life. For instance, Twitter has registered more than six hundred million active users by January, 2014, who generate average 58 million tweets per day. Every month, Facebook deals with 570 billion page views, stores three billion new photos, and manages 25 billion pieces of contents [1]. It is said that the age of big data, which is permeating into all aspects of our life, work, and learning, has arrived. Big data, which "includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time" [2], are "high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable

enhanced decision making, insight discovery and process optimization" [3]. Especially, the so-called personal big data (e.g., life logs), which includes a variety of data, such as text, location, sound, images and videos, and is dynamically produced from multiple sources with different data structures or even no data structure, has become an important source of value and innovation. However, it is difficult for an individual (end-user) to utilize the useful and valuable information, such as user experience and social knowledge hidden in these records, by simply processing the large amount of raw data. Therefore, it has become a big challenge to process such massive amounts of personal data, which has attracted a lot of attentions ranging from academia, industry to government as well.

Information utilization, including information seeking and recommendation for an individual, and information sharing and knowledge creation within a group, has always been identified as an important concept during decades of development of information technology, ranging from research institutions to government organizations. Although there are a variety of definitions and applications for it, the key point is all about the individuals and information coming together [4]. Recently, the rapid development of emerging computing paradigms, such as Ubiquitous Computing, Social Computing, and Mobile Computing, enable more and more people to continuously sharing their personal contents including feelings, experience, and

knowledge from their local environments. These cooperative and pervasive data collected at the personal, urban, and global scale, which represents more information in terms of an individual's behavioral habits and daily routines [5], contains big potential value for an individual, business, domestic and national economy development as well. That is, the insights from the personal data coupled with individuals' information behaviors and social activities, can be viewed as a kind of valuable outcome to benefit the personalized information utilization.

Therefore, to facilitate individualized information utilization and sharing for both individuals and communities, we concentrate on the computational approaches to unified modeling and analyzing of the personal data and behaviors. On one hand, it is essential to find an effective way to model and organize the personal data, to capture users' time-varying intentions and needs timely, and mine the complex association among data efficiently. On the other hand, it is necessary to analyze the individuals' information behaviors and social activities from both of the networked digital space and physical human society, to be aware of the insight from the individual behaviors, and better understand the user context from a series of actions. Meanwhile, since the complex relations among big data always associate with the corresponding relationships among people, as well as their groups, the refinement of our social relationships is also requisite, not only to enlarge our professional sphere to promote

the information sharing and social knowledge creation, but also to extend our social circle to seek for more timely and accurate information related to our needs in different situations.

1.2 Purpose of this Study

As discussed above, the information dynamically generated from a variety of systems and different devices has spread more widely and quickly, and individuals are connected much closer than ever before, which leads to an explosive increase of data scale. Thus, it has become an increasing important issue to effectively analyze and organize the so-called personal big data, in order to find a way to better utilize the valuable information to provide individuals with the personalized service and support. In details, to take advantage of the considerable size of personal data for individualized information utilization, a well- structured user model for an individual, and a user networking model for a group of individuals are requisite. The data analytics is also necessary to ma , analyze the various information behaviors and social activities, and better understand the user contexts. However, the large scales of personal big data with different data structures or no structure is always along with lots of useless noise data, which makes more difficult to utilize piece of relevant information to satisfy individual -varying requirements.

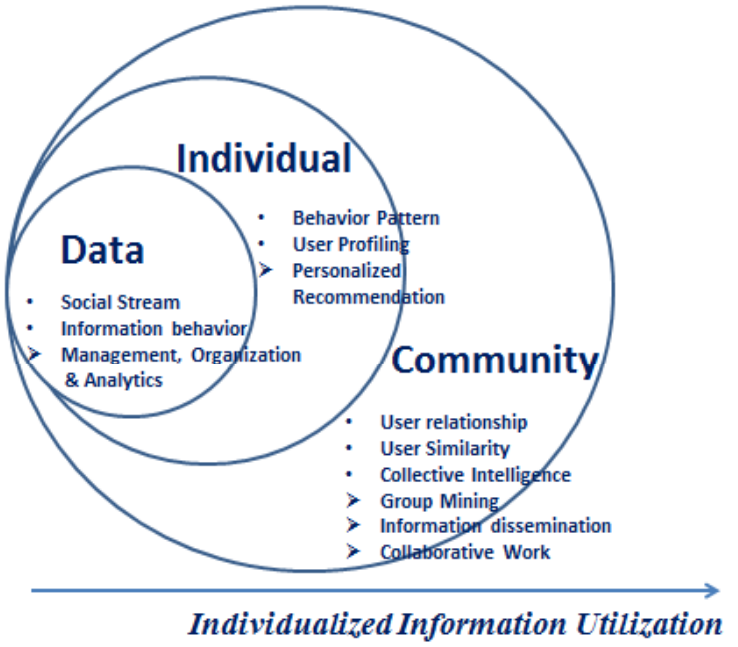


Figure 1-1 Facilitation from Associative Data to Connected People

Therefore, in this study, we delve into modeling and analyzing the personal data coupled with the individual behaviors. As shown in Fig. 1-1, firstly, the personal big data will be systematically managed and organized, which aims to meaningfully process the raw data into an associatively and methodically organized form. And users' information behaviors and social activities will be modeled to extract the behavioral features and analyze the similarities among them. Secondly, users' dynamical profiling will be built and the potential correlations will be discovered in accordance with the outcomes from the analysis of personal data and individual behaviors, which aims to better exert the capabilities of individual users and utilize the collective intelligence from social communities. Thirdly, both the behavior similarities and the social relationships will be integrated into a personalized recommendation mechanism,

which can be applied to assist the collaborative works in the task-oriented processes. Accordingly, all these aim to continuously provide an effective way for the individualized information utilization from chaotic data to associative information, and further to connected people.

1.3 Contributions of this Study

In this study, we focus on four major contributions on:

1) A unified modeling for associative organization of personal big data

A new concept of organic stream, which is designed as a flexibly extensible data carrier, is introduced and defined to provide a simple but efficient means to formulate, organize, and represent the personal big data. A heuristic mechanism is developed to capture individuals' time-varying interests or needs, and aggregate and integrate the relevant data together to obtain the associative information.

2) An behavioral analysis method for detecting of social influence and action patterns

A behavioral analysis method is proposed to detect and calculate the social influence within the individual behaviors. A mechanism is developed to automatically analyze and extract individuals' action patterns from a series of information behaviors. And the behavioral similarities among a group of individuals are calculated and described based on the action patterns.

3) A user networking model for multi-dimensional user profiling and dynamical community discovery

A *DSUN* (Dynamically Socialized User Networking) model is proposed to describe users' implicit and explicit social relationships based on the outcomes of the analysis of personal data and individual behaviors. A series of measures are defined to describe and measure the dynamical user correlations, and a set of attributes are analyzed to build the multi-dimensional user profiling, which are utilized to discover the multi-types of social communities for the facilitation of information utilization and sharing.

4) An integrated mechanism for progressive recommendation in the task-oriented learning processes

An integrated mechanism is developed to utilize the combination of user behavior patterns and user correlations in a progressive recommendation, which can provide users with the most suitable learning actions as their learning guide and support for the better utilization of personal learning experience in the task-oriented processes.

1.4 Organization of the Thesis

This thesis is organized in six chapters. In Chapter 1, we introduce the background of this study, discuss the motivation and purpose, and point out the major contributions

of this study.

In Chapter 2, we address and discuss the main issues related to this study, including issues of studies on social media application and life log analysis, analyses for user relationships and communities, and user behaviors modeling and pattern recognition.

In Chapter 3, the basic concepts and models to analyze the personal data and individual behaviors are introduced. A heuristic mechanism is developed to organize the associative information from the raw data [6], and a behavioral analysis method is proposed to analyze and model the individual behaviors [7].

In Chapter 4, the details of the constructing of the *DSUN* model [8], as well as the definitions and calculations of user correlation and profiling, are introduced. A series of algorithms and mechanisms are proposed and developed to build the user profiling [9] and discover the social community respectively. The experimental results are analyzed and discussed finally.

In Chapter 5, an integrated mechanism is developed to support the collaborative learning in the task-oriented processes, in which both user behavior patterns and user correlations are utilized to provide the individualized recommendation of learning actions [7]. The experiment and evaluation results are discussed finally.

Finally, in Chapter 6, we give a brief summary of the thesis, and conclude the

features of this study. The expected challenging issues and the future working directions are also presented.

Chapter 2 Related Work

Three main issues related to this study are walked through in this chapter. That is, issues of studies on social media application and life log analysis, analyses for user communities and relationships, and user behavior modeling and pattern recognition are addressed respectively.

2.1 Social Media Application and Life Log Analysis

There are many applications utilizing the analysis results of the data across social media [10-19], ranging from the public/individual interest detecting, to the topic-based activity/event predicting. By examining the twitter data to track rapidly-evolving public sentiment and activity, Signorini et al. showed that Twitter can be used as a measure of public interest or concern about health-related events [10]. Analysis of twitter communications in [11] revealed the experimental evidence that Twitter can be used as an educational tool to help engage the students, and mobilize the faculty into a more active and participatory role. A study addressed in [12] examined the impact of posting social, scholarly, or a combination of social and scholarly information to Twitter on the perceived credibility of the instructor, which may have implications for both teaching and learning. Black et al. [13] presented a

methodological approach and a technology architecture, which examined Twitter as a transport protocol in different socio-technical contexts, in order to capture, transfer, and analyze the twitter interactions. Byun et al. [14] developed a java-based data gathering tool with the design specifications to continuously and automatically collect social data from Twitter and filter noisy data, which can benefit the analysis of twitter messages, and further assist the detections of hot issues and topics and the discoveries of groups or communities. Wang et al. [15] proposed a hashtag-based sentiment classification method for the Topic Sentiment Analysis in Twitter, in which a graph model was introduced to deal with the hashtag-level information with three inference algorithms (loopy belief propagation, relaxation labeling and iterative classification algorithms) for classification. Kendall et al. [16] conducted a content-based analysis on twitter posts which focused on health-related fitness activities, in order to support the tool and application design with social media platform. Cogan et al. [17] proposed a method to reconstruct the complete conversations around the initial tweets, in order to analyze how users communicate on the initial post over time. Vosecky et al. [18] developed an interactive system, called Limosa, to model the comprehensive geographic characteristics of the topics discussed in Twitter, and visualize users' geographic interests. Pervin et al. [19] proposed a method as well as the implementation to detect the trending topics with a contextual meaning from the

real-time text stream in twitter posts.

Especially, as one kind of personal big data, life logs have attracted increasing attentions in recent years [20-25]. The analysis results of life logs can be utilized to provide people with more adaptive and personalized services. Yamagiwa et al. [20] proposed a system to achieve an ecological lifestyle at home, in which sensors measure the social life log by the temperature, humidity, intensity of illumination related to human action and living environment closely. Hori et al. [21] developed a context-based video retrieval system which utilized various sensor data to provide video browsing and retrieval functions for life log applications. Hwang et al. [22] developed a machine learning method for life log management, in which a probabilistic network model was employed to summarize and manage the human experiences by analyzing various kinds of log data. Kang et al. [23] defined metadata to save and search life log media, in order to deal with those problems such as high-capacity memory space and long search time cost. Shimojo et al. [24] have re-engineered the life log common data model (LLCDM) and life log mashup API (LLAPI) with the relational database MySQL and the Web services, which can help access the standardized data, in order to support the integration of heterogeneous life log services. Nakamura et al. [25] proposed a method to infer users' temporal preference according to their interests by analyzing the web browsing logs.

2.2 User Relationship Analysis

Three important aspects, including user correlation analysis, social influence analysis, and social community discovery, are introduced to address the issues of analysis of user relationships.

2.2.1 User Correlation Analysis

Researchers recently focused more on the aspect of user correlation analysis [26-33], which can benefit not only the social graph constructing, but also the user grouping process. Based on the analysis of the correlation between social and topical features in online social networks such as Flickr, Last.fm, and aNobii, Aiello et al. [26] built a user similarity network to perform the prediction of friendships. Yu et al. [27] proposed a method to provide suggestions of suitable social groups based on a user's personal photo collection, which considers both similarities of the groups and relationships among images in a user's collection. In order to find the strong relationships automatically in social networks, Aiello et al. [28] analyzed the social network in a social bookmarking system named Nobii, in order to investigate the interplay of profile similarity and link creation according to interest-based factors. Xiang et al. [29] developed a latent variable model to infer the relationship strength, in which both profile similarity and interaction activity are taken into account to improve the strength estimation. Leroy et al. [30] proposed a so-called cold start link

prediction method which could detect potential social graph by using group membership information obtained from Flickr. Wilson et al. [31] proposed the interaction graphs to quantify user interactions in the Facebook social network, in order to find out whether social links are valid indicators of real user interaction. Tang et al. [32] explored different group-profiling strategies using information extracted from the real-world social media sites for group construction, in order to better assist network analysis and navigation. Zheng et al. [33] proposed a framework for the personalized friend and location recommendation, which measure the similarity relationship between users in terms of their location histories and recommend a group of friends in a geographical information system community.

2.2.2 Social Influence Analysis

Issues regarding to the measuring of social influence have been hotly discussed recently [34-40]. Romero et al. [34] proposed and developed an algorithm that determines the users' influence and passivity according to their information activities through social media. Tang et al. [35] proposed the TAP (Topical Affinity Propagation) to model and identify the topic-level social influence across a large social network. Sang et al. [36] proposed a multimodal probabilistic model, which considered both the users' textual annotation and uploaded visual images, in order to extract the users' topic distributions and topic-sensitive influence strengths.

Achananuparp et al. [37] examined user behaviors, especially the retweeting activities among the Twitter users to model the behaviors relevant to information propagation, which was further used in the event detecting process in the Twitter environment. Tang et al. [38] proposed a framework to analyze the user influence within the online healthcare community, in which the users' reply relationship, conversation content and response immediacy were considered together to identify the influential users. Ronald et al. [39] proposed an agent-based model to describe the influence from social activities between a pair of users, which has been experimented with four input networks, in order to demonstrate the relevance of the social network structure. Gomez-Rodriguez et al. [40] proposed and demonstrated a scalable method to infer diffusion and influence through networks by tracing information cascades in the asset of blogs and news articles.

2.2.3 Social Community Discovery

User relationship modeling and analysis for the discovery of communities, have also drawn a large body of researches [41-46]. Lin et al. [41] proposed the MetaFac framework which utilized various social contexts and interactions for community structure extraction, to support the community discovery process. Leskovec and Horvitz [42] constructed a communication graph to examine characteristics and patterns based on the collective dynamics of 240 million users rather than the

individual actions or characteristics. Yin et al. [43] built a model for the latent community topic analysis, in which community discoveries were incorporated into the topic analysis in the text-associated graphs, in order to guarantee the topical coherence in the communities. Goolsby [44] introduced the so-called ad-hoc crisis community that used the social media as a crisis platform to generate community crisis maps. Zhang et al. [45] presented a unified framework which combined the author-topic model with the user friendship network analysis for the user community discovery in online social networks. Paliouras [46] focused on discovering user communities and their roles from the logs of users' activities across the social networking, which can be used to model the users' interests and personalize Web applications.

2.3 Information Behavior Modeling and Pattern Analysis

The user behavior modeling, as well as the behavior pattern analysis, has been developed by more and more researchers for its widespread availability during these years [47-55]. Razmerita [47] proposed a generic ontology-based user modeling framework (OntobUMf) to model the user behaviors, and used it for the user classification, in order to enhance the personal knowledge management. Stolfo et al. [48] used the EMT (Email Mining Toolkit) empowered with behavior-modeling techniques to compute behavior profiles of user email accounts, which can help detect

the viral propagation problem. Chen et al. [49] proposed a data mining method to extract the user movement behavior patterns, in order to predict and recommend suitable services for users in a mobile service environment. Considering both user behaviors and collaborative filtering, Liu et al. [50] proposed a semantic relatedness measure between words to retrieve related words and detect new word tasks, which can help enrich user experience and discover hidden information. Lee et al. [51] developed a non-supervised learning framework to discover the behavior patterns, in which a new cluster validity index was proposed for agglomerative iterative Bayesian fuzzy clustering, and the fuzzy-state Q-learning was proposed to learn the sequential actions. Yun et al. [52] proposed a model to mine mobile sequential patterns, in which both user moving patterns and purchase patterns were taken into account, and they further devised three algorithms (TJLS, TJPT, and TJPF) to determine the frequent sequential patterns in the mobile commerce environment. Muñoz-Organero et al. [53] utilized the behavior patterns to predict student motivation, which can further be used to predict the successful completion of an e-learning course, based on the analysis of relationships between the motivation and performance of 180 students who took an e-learning course deployed on a Moodle e-learning platform. Plantevit et al. [54] presented a method to mine sequential patterns from multidimensional and multilevel databases, which took account of different dimensions and levels of granularity, in

order to discover the regular specific patterns. Zhao et al. [55] proposed an access-pattern-driven distributed caching middleware named APRICOD, which gave more consideration on user interactions for media streaming applications.

2.4 Summary

Research works pointed out the trends of utilization of the personal data not only in the design of systems and models, but also in practical recommendation and prediction services, which can provide us with more information and knowledge in various aspects using the data mining and analyzing methods. The analysis of user relationships and discovery of communities can promote the collaborative works. And the modeling of user behaviors, as well as the calculation of influence from the social activities, can also help enhance the information seeking and sharing process. In this study, we concentrate on the unified modeling of personal data, and analyzing of individual behaviors. Comparing with the traditional methods, we focus more on capturing individuals' time-varying intentions, and further integrate and organize the associative information in accordance with the extracted intentions. We also highlight the modeling of sequential information behaviors to improve the behavioral contexts of individuals in the task-oriented processes. Differing from other user graph model, a newly developed user networking model based on the calculations of the similarity of user characteristics and the influence from interactional behavior, is built, which can

result in the multi-dimensional user profiling and multi-types of social communities for individualized information utilization. Furthermore, an integrated mechanism is developed to utilize the both behavior patterns and user correlations for the progressive recommendations, which can provide users with personalized learning guidance and support.

Chapter 3 Analysis of Personal Data and Behaviors: Definition and Model

The personal big data, including the large scales of user-generated data and a variety of individual behaviors, contains enormous amount of potential worth and value not only for the enrichment of people's social activities, but also for the enhancement of collaborative works from all walks of life. In this chapter, the basic concepts, and models are introduced to discover the valuable outcomes from the so-called personal stream data, and individuals' information behaviors as well.

3.1 Organizing of Personal Stream Data

The tremendous amount of diverse data, which is dynamically generated by individuals, and continuously spread across the social networking, can be viewed as the personal stream data. It should be organized systematically to obtain the meaningful information.

3.1.1 Metaphors for Organizing Process

Toward this purpose, we introduce the metaphors for the organization of this kind of data streams as follows.

Drop: Drop is a minimum unit of data streams, such as a message posted to

Twitter, or a status change in Facebook by a user.

Stream: Stream is a collection of drops following the timeline, which contains the messages, activities and actions of a user.

River: River is a confluence of the streams from a series of different users which are constituted through following or subscribing their followers/friends. Note that the confluence scale could be extended to followers' followers.

Ocean: Ocean is a combination of all the streams.

As mentioned above, the content posted from each user can be seen as a drop, and the drops converging together from one specific user can form a stream. Then the streams generated from this user and his/her friends can aggregate together to form the river. Finally, all the streams from all the users come together to form the ocean. The conceptual graph model is shown in in Fig. 3-1.

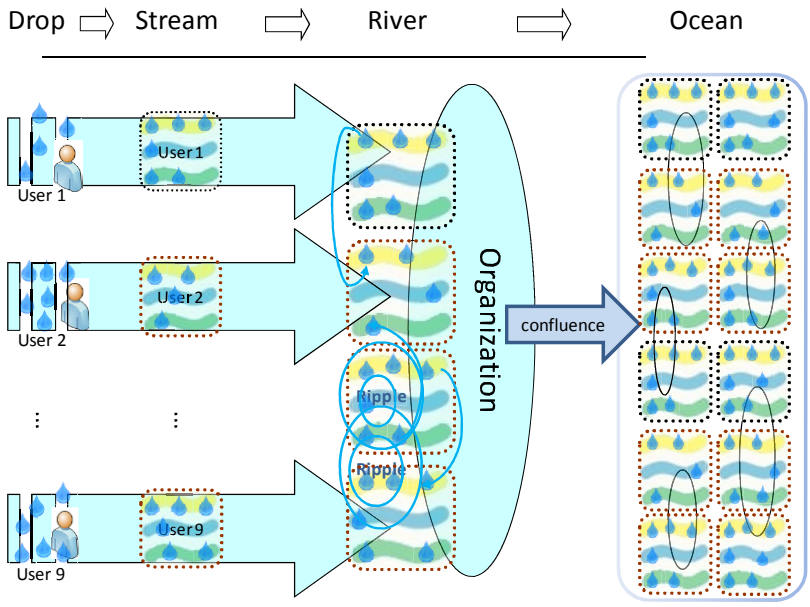


Figure 3-1 Graph Model for Stream Data [6]

Based on these, the following definitions are proposed and defined for the personal data organization process that satisfies users' current interests or needs.

Heuristic Stone: Heuristic stone represents one of a specific user's current interests or needs which may be changed dynamically in different time periods.

Associative Ripple: Associative ripple is a meaningfully associated collection of the stream data that are related to some topics with regards to a specific user's interests in the river level.

Associative Drop: Associative drop is the stream data selected into an associative ripple, which is related to one specific heuristic stone and can further composes the organic stream.

3.1.2 Organic Streams

3.1.2.1 Concept and Definition

The organic stream, which is designed as a flexibly extensible data carrier, is introduced and defined to provide a simple but efficient means to formulate, organize, and represent the personal big data. As an abstract data type, organic streams can be regarded as a logic metaphor, which aims to meaningfully process the raw stream data into an associatively and methodically organized form, but no concrete implementation for physical data structure and storage is defined. The details are addressed as follows.

Organic Stream: Organic stream is a dynamically extensible carrier of organized personal data that may contain potential and valuable information and knowledge. The formal description of organic stream can be expressed in Eq. (3.1).

$$OS = \Phi\ (Hs, Ad, R) \tag{3.1}$$

where,

$Hs = \{Hs[u_1, t_1], Hs[u_2, t_2], ..., Hs[u_m, t_n]\}$: A non-empty set of heuristic stones in accordance with different users' intentions (e.g., users' current interests or needs), in which each $Hs[u_i, t_j]$ indicates a extracted heuristic stone of a specific user u_i during a selected time period t_j.

$Ad = \{Ad_1, Ad_2, ..., Ad_n\}$: A collection of associative drops which can refer or link to each other based on the inherent or potential logicality in a methodical and associative way.

R: The multi-types of relations among heuristic stones and associative drops in the organic stream.

Furthermore, the relation *R* in the organic stream can be categorized into three major types: relation between each heuristic stone; relation between each associative drop; and relation between heuristic stone and associative drop.

Relation between Heuristic Stone and Associative Drop: This type of relation identifies the relationships between one heuristic stone and a series of associative drop,

which can be represented as *Heuristic Stone* × *Associative Drop*. It is the basic relation in the organization of organic stream, which means different granularities of heuristic stones may lead to different scales of related drops connecting together in the organic stream.

Relation between Heuristic Stone and Heuristic Stone: This type of relation identifies the relationships among the heuristic stones in the organic stream, which can be represented as *Heuristic Stone* × *Heuristic Stone*. Due to the different users and different time periods, this kind of relation can further be categorized into two sub-types:

$Hs[u_i, t_x] \leftrightarrow Hs[u_i, t_y]$: This relation identifies the relationships of the heuristic stones extracted from one user. That is, this relation is used to describe those internal relationships or changes for a specific user's intentions. Given a series of heuristic stones from a specific user u_i, represented as $\{Hs[u_i, t_1], Hs[u_i, t_2], ..., Hs[u_i, t_n]\}$, the differences from $Hs[u_i, t_1]$ to $Hs[u_i, t_n]$ changed in a sequence can demonstrate the transitions of this user's interests or needs in a specific period, which can be employed to infer his/her further intention.

$Hs[u_i, t_x] \leftrightarrow Hs[u_j, t_y]$: This relation identifies the relationships of the heuristic stones extracted among different users. That is, this relation is used to describe those external relationships among different users' intentions. Given two heuristic stones,

represented as $Hs[u_i, t_a]$ and $Hs[u_j, t_b]$ for two different users, the relationship can demonstrate the potential connections among these two users in accordance with their dynamical interests or needs.

Relation between Associative Drop and Associative Drop: This type of relation identifies the relationships among those drops that are clustered into the associative ripples and further compose the organic stream, which can be represented as *Associative Drop* × *Associative Drop*. The drops connected together based on this relation in different associative ripples can represent the whole trend as well as its changes following the timeline.

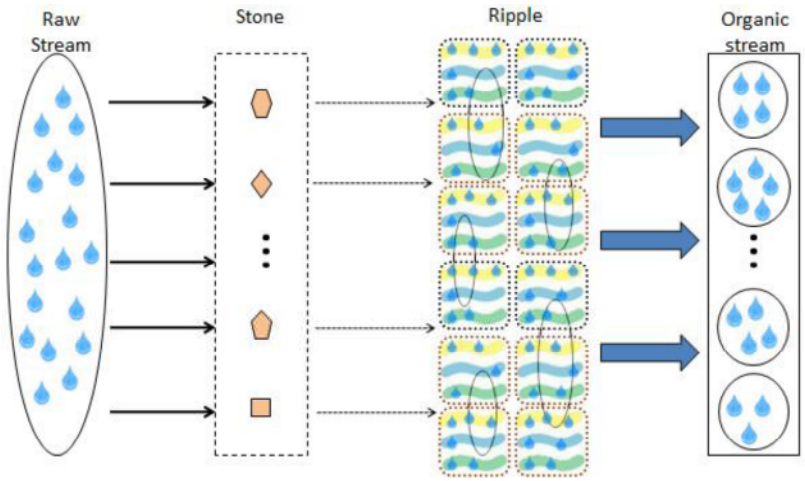


Figure 3-2 Image of Organization Process of Stream Data [56]

Fig. 3-2 shows an image of the organization of personal stream data. As discussed above, the heuristic stone is defined to represent a specific user's current interest or need, which can be discovered and extracted from his/her own streams. The associative ripple is then generated in accordance with the heuristic stones. For a

specific user, the whole timeline will be divided into several time slices, the heuristic stones are composed by the keywords that are calculated and extracted from his/her own stream data according to the TFIDF-based method. Then each of the time slices will produce an associative ripple in accordance with the heuristic stone. Specifically, each extracted heuristic stone, which can be viewed as the cluster center in each divided time slice, will be "threw" back into the river, to generate a series of associative ripples which consist of a set of related drops. Note that different granularities of the calculation of the heuristic stones will lead to different numbers of the associative ripples. The details of extraction and generation processes of heuristic stone and associative ripple are discussed as follows.

3.1.2.2 Collecting Heuristic Stones

As defined above, the heuristic stone is utilized to represent a user's current interest or need. We discover a specific user's interests or needs from his/her own stream data following the timeline.

Specifically, two types of interests, the time-evolving interest and consistent interest, are taken into account to describe users' dynamical intentions, which can be expressed in Eq. (3.2).

$$H = \{(h_i)^n | 1 \leq i \leq n, h \in \{H_T | H \tag{3.2}$$

where $(h_i)^n$ indicates the n-dimensional interests extracted from the users' personal

stream data. The parameter i, ranging from 1 to n, indicates the ranking number of the interest. H_T indicates the time-evolving interest, called *transilient interest*, which describes one kind of interests that will change during some special time periods, or be intrigued due to some hot topics or interesting events. On the other hand, H_D indicates the consistent interest, called *durative interest*, which describes one kind of interests that can be viewed as the inherent interest and will be continuously held during a long time period.

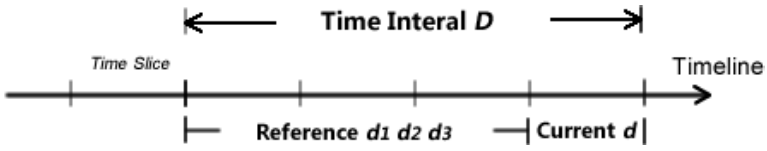


Figure 3-3 Illustration of Dynamical Division of Time Slices

To quantify and distinguish the transilient interest and durative interest for a specific user u_i, as shown in Fig. 3-3, in a selected time interval D with several dynamically divided time slices d_j, the transilient interest will be extracted in the current time slice d, while the other past time slices (e.g., d_1, d_2, and d_3) will be considered as the references. The durative interest will be extracted referring to the whole time interval D, in which each time slice will hold the same durative interest. The TFIDF-based method is developed to calculate the frequency-based weight $w(H_T)$ for the transilient interest, which is expressed in Eq. (3.3).

$$w(H_T) = \frac{N(k_i)|_{d_x}}{\sum_k N(k_i)|_{d_x}} * \frac{|D|}{C(k_i)} \tag{3.3}$$

where $C(k_i) = \left|\{j \mid if\ \exists k_i \in K_{d_j}\}\right|$

For a keyword k_i, D indicates the whole time interval, while d_j indicates each time slice, $D = \{d_j\}$. For instance, if D is set as one month (say 28 days), then d_j can be set as one week (seven days), thus $D = \{d_1, d_2, d_3, d_4\}$. $N(k_i)|_{d_x}$ indicates the frequency of the specific keyword k_i in the time slice d_x. $\sum_k N(k_i)|_{d_x}$ indicates the sum of frequency of all the keywords in the same time slice d_x. $C(k_i)$ indicates the number of time slices in which the keyword k_i has occurred in the keyword set K_{d_j}.

On the other hand, Eq. (3.4) is employed to calculate the frequency-based weight $w(H_D)$ for the durative interest.

$$w(H \qquad \frac{(k_i)|_D}{\sum_k N(k_i)|_D} \tag{3.4}$$

where, $N(k_i)|_D$ indicates the frequency of a specific keyword k_i in the whole time interval D, while $\sum_k N(k_i)|_D$ indicates the sum of frequency of all the keywords.

3.1.2.3 Generating Associative Ripples

The extracted heuristic stone is utilized to generate the associative ripples. Note that each heuristic stone may generate a series of ripples, which depends on the number of divided time slices and the granularity of the user's interest. A cluster center will be formed by the heuristic stone in the time slice divided from the whole time interval. The related drops in the river will converge to the cluster center. The distance between the drop and the center describes the relevance between them, and the drops which

have the same relevance to the center will distribute in the same circle. Fig. 3-4 illustrates the associative ripples generated in each time slice.

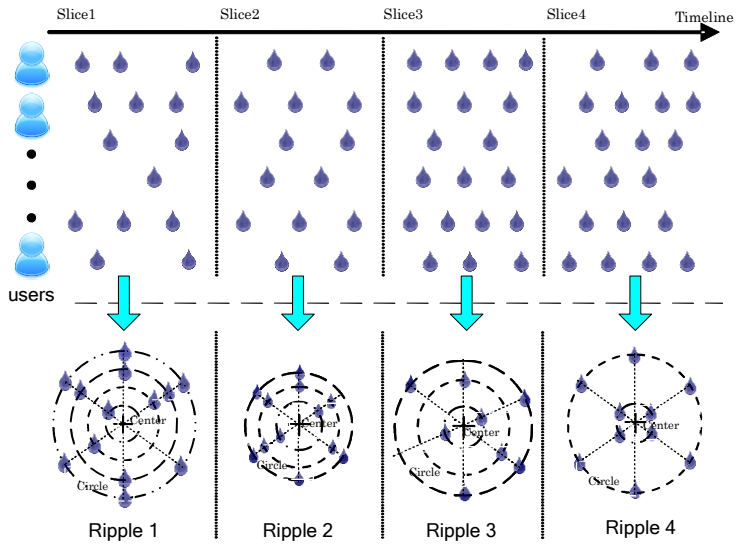


Figure 3-4 Generation of Associative Ripples [6]

As shown in Fig. 3-4, four associative ripples are generated in different time slices. The "+" indicates the cluster center in each time slice, and the "+" with several circles around it compose the ripple. All these four ripples are generated by one heuristic stone, which compose a ripple sequence. The drops distributed in each circle mean that they are relevant to the ripple in some degrees, while others are not. At the beginning, the drops from different users are distributing in the river following the timeline. After the clustering, the time sequence is broken, so that the drips do not follow the timeline any more in each ripple, but the relevancy degree to the cluster center from inside to the outside. Note that these ripples falling in the sequence still

follow the timeline.

A six-tuple (Z, Hs, $Ar_{h_i x}$, Ar_{h_i}, Ars, Q) is employed to describe the generation and composition of the associative ripple.

$Z = \{Z_1, Z_2, Z_3, \ldots, Z_m\}$: A non-empty set of input data, which can be a collection of the contents posted by all the users in a certain group.

$Hs = \{h_1, h_2, h_3, \ldots, h_n\}$: A non-empty set of the heuristic stones to represent users' n-dimensional interests, each of which can generate a series of associative ripples.

$Ar_{h_i x} = < \{Z_1, Z_2, Z_3, \ldots, Z_t\}_{cir_n} >$: A non-empty sequence of stream data sets which have clustered into one associative ripple. Each $\{Z_1, Z_2, Z_3, \ldots, Z_t\}_{cir_n}$ indicates a set of related data that distributed in a specific circle cir_n of the ripple. Note that the sequence of these data sets indicates the descending order in terms of the relevance degree regarding to the heuristic stone in the center.

$Ar_{h_i} = <Ar_{h_i 1}, Ar_{h_i 2}, Ar_{h_i 3}, \ldots, Ar_{h_i x}>$: A non-empty sequence of the associative ripples produced by one heuristic stone, which follow the timeline in sequence.

$Ars = \{Ar_{h_1}, Ar_{h_2}, Ar_{h_3}, \ldots, Ar_{h_n}\}$: A non-empty set of Ar_{h_i}, which is the final results of associative ripples for one specific user.

$\mathrm{Q}(\mathrm{Z_i}, \mathrm{h_j}) \rightarrow \mathrm{Z_i} \in \mathrm{Ar_{h_i x}}$: A matching function which is used to decide whether Z_i

belongs to $Ar_{h_i x}$ and calculate the corresponding relevancy.

The algorithm to generate the associative ripples is shown in Fig. 3-5.

Input: The stream data set *Z*,
A specific user's interest set *Hs*
Output: The associative ripple set *Ars*

Step 1: Divide the whole time period into several time slices. For each h_j, create a cluster center in each time slice
Step 2: For each Z_i to each h_j,
if satisfy $Q(Z_i, h_j) \rightarrow Z_i \in Ar_{h_i x}$,
insert Z_i with the relevance degree into the corresponding circle cir_n in $Ar_{h_i x}$
Step 3: For each h_i,
sort the corresponding $Ar_{h_i x}$ by timeline and record them in each Ar_{h_i}
Step 4: Return the associative ripple set
$Ars = \{Ar_{h_1}, Ar_{h_2}, Ar_{h_3}, \ldots, Ar_{h_n}\}$

Figure 3-5 Algorithm for Generating Associative Ripples

3.1.3 A Scenario for Enrichment of User Search Experience

The method discussed above can be employed to enrich search experience and improve the information seeking process. The stream data analyzer and organizer are developed to create the user profiling in accordance with the organization of personal data from the so-called social streams that includes stream data in the cyber world or life log data from the physical world, which can contribute to both query initialization and document matching. Fig. 3-6 reveals the overall scenario of the enhanced information seeking process.

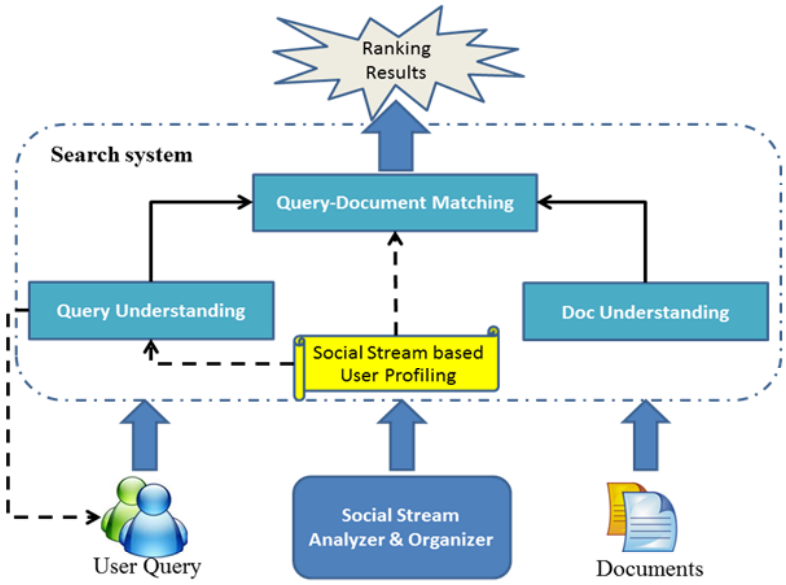


Figure 3-6 The Scenario of Enhanced Information Seeking [6]

As shown in Fig. 3-6, comparing with the traditional search system, two proposed operators, stream analyzer and organizer, are added to provide supplementary information. The personal data from social streams, such as the tweets in Twitter, are integrated into the search engine to assist search experience enrichment and search process facilitation. Users' current needs or interests are extracted with those related personal data posted in the social network platform. We provide information to the implemented search engine and go further to create the social stream based user profile. The information provided in the user profile might affect the ranking results which satisfies user's queries more precisely. In addition, users may get feedback with some other adaptive keywords based on the analysis of stream data. They may start over a new query to pursue their requirements. All these can

contribute to the enrichment of user search experience.

As mentioned above, the proposed mechanisms and methods focus on organizing those raw data into meaningful contents in accordance with users' current interests or needs. The associative information which is extracted and organized from the social streams can be utilized to influence the search ranking results and further provide users with a new perspective for information seeking and discovery. As a summary, user search experience can be enriched in the search engine as follows:

- The heuristic stone(s) shown in the search engine can be viewed as sort of hints to the search keyword and provide users with a new perspective. These keywords changing dynamically and timely can be employed as sub-queries in the next search round to improve the search keyword quality and refine the search scope, in order to seek information more related to users' current needs.
- The associative ripple(s) generated according to the heuristic stone contains some potential trends or needs among a certain group of people. These meaningfully organized information can provide users with the hot topics that people are discussing recently or the potential trends that may be pursued in the future, which can guide users to seek the more suitable and satisfied information. Especially, this function may become more helpful in some cases that people do not know what they are really concerned about when they face to a search engine,

or just have a general concept of the seeking issue and type some words to have a try.

- The search results can also be re-ranked in accordance with the heuristic stones and associative ripples. Those search results that are more related to users' current interests or needs, or the hot topic and trends will come to the front of the whole search results, which could maximally satisfy users' requirements.

3.2 Analysis of Individual Behaviors

People always conduct a variety of information behaviors to interact and communicate with each other within some certain social groups. It is necessary to analyze these individual behaviors to discover the behavioral insights hidden in their activities, which can help refine their social relationships, and benefit the individualized information seeking and recommendation.

3.2.1 Detecting Influence from Individual Behaviors

Specifically, in this study, two types of information behaviors, the influencing behavior and influenced behavior, are taken into account, which can be defined as follows.

Influencing Behavior ($IgB(\overrightarrow{u_i u_j})$): A set of information behaviors of user u_i, which indicates that user u_i tends to influence user u_j. It can also be considered as one kind of behaviors that indicates user u_i tend to delivery his/her personal information to

user u_j.

Influenced Behavior (*IdB*($\overrightarrow{u_i u_j}$)): A set of information behaviors of user u_j, which indicates that user u_j has been influenced by user u_i. It can also be considered as one kind of behaviors that indicates user u_j has received user u_i' personal information or has been in favor of user u_i' thought.

Table 3-1 Descriptions of Information Behaviors

Symbols	Value
T	Selected time period for users
$\|T\|$	Number of total days in the selected time period T
$Act_T(u_i)$	Number of days that u_i has conducted information behaviors
$NIB(u_i)$	Number of information behaviors of u_i
$NIgB(u_i)$	Number of influencing behaviors of u_i
$NIgB(u_i, u_j)$	Number of influencing behaviors of u_i to u_j
$NIdB(u_i)$	Number of influenced behaviors of u_i
$NIdB(u_i, u_j)$	Number of influenced behaviors of u_i from u_j
$NIB_{ing}(u_i)$	Number of information behaviors of u_i which have influenced others
$NIB_{ing}(u_i, u_j)$	Number of information behaviors of u_i which have influenced u_j
$NIB_{IgB}(u_i)$	Number of information behaviors of u_i which contain influencing behaviors
$NIB_{IdB}(u_i)$	Number of information behaviors of u_i which contain influenced behaviors
$NIB_{\overline{IgB}}(u_i)$	Number of information behaviors of u_i which do not contain influencing behaviors
$NIB_{\overline{IdB}}(u_i)$	Number of information behaviors of u_i which do not contain influenced behaviors

For instance, in Twitter, the information behavior "@name" can be considered as the influencing behavior. It means user u_i tends to build a connection, or delivery some information that may be related to user u_j, when u_i mentions "@ u_j" in his/her posts. The information behavior "RT @name" can be considered as the influenced behavior. It means user u_j has selected and received a sort of u_i' personal opinions,

when u_j mentions "RT @ u_i" in his/her posts. In addition, the influenced behaviors can also be viewed as the positive behaviors for information propagation.

To describe and analyze individuals' information behaviors, especially the influence-based behaviors, the frequency of information behaviors generated from each user is taken into account for the quantification, which is summarized in Table 3-1.

3.2.2 Analyzing Sequential Action Behaviors

We go further to analyze a series of individual behaviors which can be described as a sequence of action behaviors, in order to discover the behavioral similarities based on the action patterns not only to benefit an individual user, but also for a group of users in a social community.

3.2.2.1 Formal Description of Action Behaviors

To discover and model the sequence-based action patterns, the sequential action behaviors in a task-oriented process can be formalized as follows:

$act = \{U, O, Ir\}$: A non-empty set to describe the information action, which is the minimum unit for the description of information behaviors. U indicates the user who has conducted this specific action, O indicates the concrete operation of this action behavior (e.g., clicking a web link), and Ir indicates the information resources that the user U has used associated with this action behavior.

$Act = < act_1, act_2, \ldots, act_n, G>$: A non-empty set to describe the information activity, which is represented as a sequence of information actions. Especially, G, in the end of the sequence, is a special action that indicates a specific purpose of this information action sequence, while each act_i indicates the information action that belongs to this activity to complete the certain purpose.

$S\text{-}Task = < Act_1, Act_2, \ldots, Act_n, T >$: A non-empty set to describe the information sub-task, which is represented as a sequence of information activities. Each Act_i indicates the information activity that belongs to this sub-task. T indicates a specific time period selected within the whole information task, which can also be viewed as an end of time interval.

$Task = < S\text{-}Task_1, S\text{-}Task_2, \ldots, S\text{-}Task_n, \mathbb{T} >$: A non-empty set to describe the information task, which is represented as a sequence of information sub-tasks. Each $S\text{-}Task_i$ indicates the information sub-task that is divided from this task, while $\mathbb{T}$ indicates the whole time period to complete the specific information task.

3.2.2.2 Similarity Analysis of Action Patterns

The *trie* [57], an ordered tree-based structure which can be used to store a dynamic string-like data set, has been well developed and applied in information storing and retrieving. For instance, Iglesias et al. [58] have applied the *trie* data structure in behavior profile creation and recognition for a computer user. In this study, we

employ this tree-based data structure to find all the related sub-sequences with their frequency in a given information action sequence, in order to calculate the weight *w* of each action pattern. In particular, a certain action sequence with its subsequence suffixes which extend to the end of this sequence will be all inserted into a *trie*, in order to calculate the frequency of each sub-sequence during the tree building process. For example, if the whole sequence is <*A, B, C, D*>, three sub-sequences < *B, C, D*>, < *C, D*> and <*D*> shall also be inserted.

Based on these discussed above, two criteria are given to generate the action patterns.

Criteria 1 - Basic Criteria: Given a pre-defined action purpose set $G = \{G_1, G_2, \ldots, G_m\}$, and a sub-sequence q described as $< Act_1, Act_2, \ldots, Act_j, Act_n >_{u_i}^{w}$, where *w* indicates the weight of each sub-sequence. If it satisfies that $n >= 2$, $w >= 2$, and $Act_n \in G$, then q is an action pattern for user u_i, which can be described as $< act_1, act_2, \ldots, act_j >_{u_i}^{w} \rightarrow G_k$.

Criteria 2 - Incorporation Criteria: Given two sequences $q_1: < act_1, act_2, \ldots, act_n >_{u_i}^{w_x}$ and $q_2: < act_1, act_2, \ldots, act_m >_{u_i}^{w_y}$ for user u_i, if they satisfy that $w_x = w_y$, and $< act_1, act_2, \ldots, act_n > \subset < act_1, act_2, \ldots, act_m >$, then q_1 can be incorporated into q_2.

Based on these two criteria, for a specific user, a variety of action patterns can be

extract from the whole action sequence to describe the behavioral features during a selected time period. Note that the different granularities of input action sequences (e.g., one sub-task or one task) will lead to different results of action patterns, which represent the different characteristics of the action behaviors during different time periods.

The similarity among a group of users based on their action patterns can further be analyzed. That is, the whole of action patterns extracted from each user can be grouped into different categories according to the behavioral similarities, specifically, including the action sequences and the corresponding purpose. The former one represents the similarity of information behaviors among the users based on the action patterns, while the latter one indicates the same purpose that these users try to achieve within a specific time period.

Based on these discussed above, it can be viewed as that the users in the same group may have similar action sequences with different weights, which can be formalized as $\left[< Act_k >_{u_1,u_2,\ldots,u_i}^{w_1,w_2,\ldots,w_i}\right]$, to pursue the same purpose within a task-oriented process. Thus, the similar action behaviors can be shared among them in order to facilitate their collaboration works and reach the better efficiency.

3.3 Summary

In this chapter, we have introduced the basic methods to systematically model and

describe the personal data and individual behaviors. The basic concepts, models and mechanisms have been addressed to organize the so-called personal stream data, and model the individual action behaviors, in order to better understand the insight hidden in the big personal data and further benefit the individualized information utilization.

Following a graph model to describe the personal stream data in a hierarchical structure using a set of metaphors, the organic stream has been introduced and defined as an extensible data carrier to formally organize, and represent the personal big data. Three basic relations were defined and proposed to flexibly describe the inherent and potential relationships among the raw stream data. Based on these, two kinds of user interests, the time-evolving interest and consistent interest, were defined and represented as the heuristic stones, to capture users' diversified intentions. And a heuristic mechanism was developed to generate the associative ripples, which can aggregate and integrate the relevant stream data together based on the heuristic stones in an associative way. A scenario has been discussed to demonstrate how to utilize the associative information to benefit the enrichment of information seeking process.

As for the individual behavior analysis aspect, two types of information behaviors, the influencing behavior and influenced behavior, have been introduced and defined in order to detect the influence among users' interactional behaviors, which can be utilized to refine the users' social relationships, and benefit the

information and knowledge delivery during their interaction processes. On the other hand, to model users' sequential behaviors, a formal description was given to represent users' action behaviors in the task-oriented processes. Furthermore, an analysis method has been proposed to model and discover the action patterns based on the calculation of an individual user's sequential behaviors toward a certain purpose, which can be further employed to analyze the behavioral similarities among a group of users.

Chapter 4 Dynamically Socialized User Networking

As a utilization of the outcomes from the analysis of the personal data along with the individual behaviors, the *DSUN* (Dynamically Socialized User Networking) model is constructed to describe the dynamical user correlations and build the multi-dimensional user profiling. In this chapter, the details of constructing of the *DSUN* model, as well as the definitions and calculations of the user correlation and profiling, are introduced. A series of algorithms and mechanisms are proposed and developed to build the user profiling and discover the social community respectively. The experimental results are analyzed and discussed in the last.

4.1 Constructing of DSUN Model

In this section, we introduce the basic structure of the *DSUN* (Dynamically Socialized User Networking) Model, in which two basic relationships among users are defined and described, to discover and represent the potential user correlations and dynamical user profiling.

4.1.1 The Basic Model

Basically, the definition and the structure of the *DSUN* model can be defined and expressed as follows [8].

$$G_{DSUN}(U, E, W) \tag{4.1}$$

where

$U = \{u_1, u_2, u_3, ..., u_n\}$: A non-empty set of vertexes in the network model, and each u_i indicates a unique user.

$E = \{e_{ij}: <u_i, u_j> \mid$ if a relationship exists between u_i and $u_j\}$: A collection of edges that connect the vertex in U, which represent a variety of relationships among users in the network.

$W = \{w_{ij} \mid if\ \exists\ e_{ij} \in E\}$: A series of weights w_{ij}, and specifically, each weight appending on the corresponding edge is developed to identify the strength of a specific relationship between a pair of users. This value is employed to dynamically construct the model and further analyze the correlations between users.

To represent and analyze the user correlations and profiling, the compositions of the vertexes and edges in *DSUN* model can be described as follows.

For each vertex, $u_i = \{I_i, H_i, A_i\}$: I_i indicates the specific user u_i, which can be viewed as the id to identify a unique user; H_i indicates the n-dimensional interests/needs for this specific user; and A_i indicates a set of statistics-based attributes that can be used to describe the multi-dimensional profiling of this specific user. In a word, each vertex is employed to describe the corresponding user's information in an individual level.

For each edge, e_{ij} = <u_i, UC_T, u_j>: Vertex u_i is the user on the head of edge e_{ij}, which indicates the potential benefactor who may provide useful information related to user u_j's current requirements, while vertex u_j is the user on the tail of edge e_{ij}, which indicates the beneficiary who may receive the valuable information from user u_i in regards to his/her interests/needs. That is, in this connected user networking, the edge $\overrightarrow{u_i u_j}$, existing from u_i to u_j, not only indicates the potential benefit between a pair of users, but also illustrate the direction of information dissemination between these two users. UC_T is a multi-tuple including a set of measures to describe various types of relationships with different weights w_{ij} between user u_i and u_j during a selected time period T.

4.1.2 User Relationship Description

According to the theory of homophily from sociology, which is one of the fundamentals in the social networks [59], individuals are accustomed to constructing connections with those people who have the similar characteristics (e.g., interests, needs) with each other. Especially, the connection may be considered strong when one person can provide the help. Furthermore, the stronger the connection between two individuals, the more similar characteristics they may have [60]. Following these discussions, in this section, two important factors, the user characteristics similarity factor and user influence behavior factor, are taken into account to describe the

dynamical and potential user relationships.

Characteristics-based relationship: This kind of relationship is used to describe the relationships that are based on the users' individual characteristics, such as interests or needs, which can be viewed as the implicit relationships. For a pair of users, more similar characteristics they have may lead to stronger connection between them.

Influence-based relationship: This kind of relationship is used to describe the relationships that are based on the users' direct interactional behaviors, specifically, the influence behaviors between two users, which can be viewed as the explicit relationships. For a pair of users, more frequent influence behaviors they conduct with each other may result in stronger connection between them.

The user characteristics similarity based relationship should be considered as the basic relation in the *DSUN* model, which is used to determine whether two users can be connected. It can also be viewed as the prerequisite to the user influence behavior based relationship. Moreover, stronger characteristics similarity based relationships may result in stronger influence between two users. On the other hand, the user influence behavior based relationship can be considered as an essential relation in the *DSUN* model, which can be employed to detect the information delivery and sharing among related users, and illustrate the social perspective of information and

knowledge dissemination process. Moreover, the frequent influence behaviors will lead to more similar characteristics between two users.

4.1.3 User Characteristics Similarity Analysis

4.1.3.1 Analyzing of Static Features

Users' personal information, such as the basic profile published by themselves in social media (e.g., Facebook, Twitter), as well as the local information updated from their mobile devices, can be utilized as the important and fundamental resources to extract and analyze users' static features, as well as their similarities.

Table 4-1 Description for Static Features and Rules for Scoring

User Feature	Rule
Location	If user u_i and u_j are in the same area (e.g., the same city), the score is 1 If users u_i and u_j are in the close area (e.g., the nearby city), the score is 0.5 Else, the score is 0
Occupation	If user u_i and u_j's occupations are the same, the score is 1 If user u_i and u_j's occupations are in the same field, the score is 0.5 Else, the score is 0
Education	If user u_i and u_j have the same education background, the score is 1 If user u_i and u_j have the similar education background, the score is 0.5 Else, the score is 0
Age	If user u_i and u_j are the same age, the score is 1 If user u_i and u_j are contemporary, and less than 10 years apart in age, the score is 0.5 Else, the score is 0
Employer	If user u_i and u_j have the same employer, the score is 1 Else, the score is 0

Therefore, based on the user features presented in [61], we select part of "general" and "domain-specific" attributes as users' static features shown in Table 4-1, which are extendible according to different requirements in different circumstances. In addition, a set of rules are defined to assign the similarity score to the features, which will be used for the calculation of similarities.

Based on these, a feature-vector $V_f = [V_{Loc}, V_{Occ}, V_{Edu}, V_{Age}, V_{Emp}]$, which is extendible in accordance with the features that are accessible from the users' self-generated profile across social media, will be given to describe the static features for each user. For two specific users, by comparing the pair of elements in the feature-vector, each of them will be assigned with a value using the rules shown in Table 4-1. Finally, users with a higher sum of the score values will result in more similarities in terms of their static features, which will contribute to the further calculation of characteristics-based relationships.

4.1.3.2 Analyzing of Dynamical Features

Since users are continuously publishing mass of personal data across social media, it has become more important to extract their dynamical features (e.g., interests) by analyzing the user-generated data. Thus, to quantify the weight of these relationships in accordance with the users' dynamical features, the valuable outcomes from the analysis of personal data will be utilized for the relationship analysis. That is, the

users' time-varying interests or needs, which can be represented as the heuristic stones, and the related data (e.g., the data represented as the associative drop), will be employed to describe users' dynamical features and further analyze the similarity-based relationships .

In details, as the discussion of the calculation of the heuristic stones in Section 3.1.2.2, finally, based on different values of the dimension coefficient, the top-*n* scoring keywords will be selected as the heuristic stones to represent a specific user's interests for his/her dynamical features, which can further contribute to the calculation of user similarities.

4.1.3.3 Analyzing of Similarity Based on User Characteristics

The composition of associative ripple discussed in Section 3.1.2.3, is employed to analyze the dynamical similarities among users. As shown in Fig. 4-1, in one time slice divided from the whole time interval, the associative ripple consists of a series of circles, which indicates the strength of relations regarding to the specific interest, ranging from the inside to outside. That is, the selected data, distributing in different circles, is clustered to the center that indicates the specific interest in accordance with the different strengths of relevance. The closer to the center, the more relative information it may have. Thus, the users, who generate those related data, will be involved into the analysis of potential user relationships. For a specific user u_i with

one of his/her associative ripples which contain a series of ranked data, other users, who have provided more relevant data in it, will be considered as the more related or favorable users.

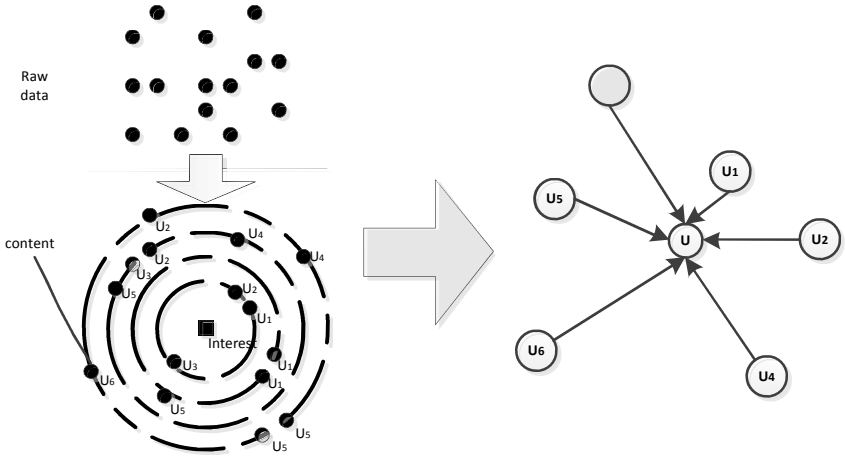


Figure 4-1 Conceptual Image of Building Dynamical Similarity-Based Relationships

Thus, combining with the static feature-based similarity we discussed above, the weight for similarity-based relationships can be quantified as following.

$$w(CSR_{ij}) = \alpha * sim_{sf}\left(u_i, u_j\right) + \beta * \frac{\sum(w_{cir}(n) * AR_n(Cir,\ u_j))}{\sum(w_{cir}(n) * AR_n(u_i))} \tag{4.2}$$

where, $sim_{sf}\left(u_i, u_j\right)$ indicates the similarity based on users u_i and u_j's static features. $AR_n(u_i)$ denotes the total number of data that are clustered on the n-level circle of user u_i's ripple, while $AR_n(Cir,\ u_j)$ denotes the number of data that is provided by user u_j on the n-level circle. $w_{cir}(n)$, which is the frequency-based weight assigned to each circle, indicates the relevance to the center of the ripple. The default value of α and β is 0.5.

4.1.4 User Influence Behavior Analysis

In light of social influence studied in [62], which illuminates how the process of social influence can work out interpersonal coordination and agreements among actors in a network of interpersonal influence, we consider that the users can influence each other who they communicate, work, make friends together with the similar characteristics through interactions. Especially, the social network influence process can be viewed as the information transmission and opinion formation process [63], which involves the interactional individuals into the repeated information transfers across the social networks [64].

Thus, among a variety of relationships arising out of information behaviors conducted across social media, we focus more on the influence-based relationships which are employed to describe the interactional influence between two connected users through their information behaviors.

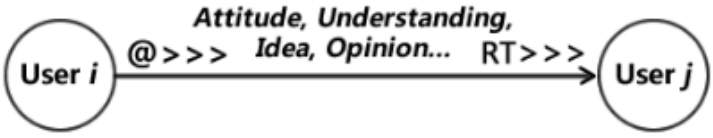


Figure 4-2 Illustration of Influence-Based Relationship

As illustrated in Fig. 4-2, the directed edge $\overrightarrow{u_i u_j}$, existing from u_i to u_j, which is built based on the similarity calculation from user u_i to user u_j in accordance with their static and dynamical features, can be viewed as a bridge of information

transmission. The personal attitude, understanding, idea, and opinion of u_i can be delivered to u_j through it. Based on these, the influencing behavior (e.g., "@name"), can be viewed as the pushing behavior to promote the transmission of information, while the influenced behavior (e.g., "RT @name"), can be viewed as the receiving behavior, which indicates the information has already been transmitted. During this influencing and influenced process, users who are interested in each other will fall in various types of agreements, and the shared information and knowledge will be diffused to the whole community.

Note that in the *DSUN* model, the influence-based relationships will be available based on the constructing of the characteristics similarity-based relationships. For a pair of users u_i and u_j, if u_i can influence u_j, or u_j has been influenced by u_i, the basic condition is that they should have a sort of similarities in term of their characteristics. In this situation, when u_j is influenced by u_i, the information transferred from u_i can be useful to support him/her.

According to these, the influence-based relationships can be quantified as follows.

Influencing degree: Given a pair of users $\overrightarrow{u_i u_j}$, the influencing degree is the value that reflects the weight of influencing relationships from user u_i to user u_j, which indicates the influencing tendency of user u_i who tends to deliver information

to user u_j. It can be recorded as *w(IgB($\overrightarrow{u_i u_j}$))*, and calculated in Eq. (4.3).

$$w\left(IgB\left(\overrightarrow{u_i u_j}\right)\right) = \frac{\sum_T NIgB(u_i,u_j)}{\sum_T NIgB(u_i)} \tag{4.3}$$

where, $\sum_T NIgB(u_i)$ indicates the sum of influencing behaviors conducted by user u_i in the selected time period T, while $\sum_T NIgB(u_i,u_j)$ indicates the sum of influencing behaviors from user u_i to user u_j in the selected time period T.

Influenced degree: Given a pair of users $\overrightarrow{u_i u_j}$, the influenced degree is the value that reflects the weight of influenced relationships from user u_j to user u_i, which indicates the scale of accepted information from user u_i according to the influenced behaviors. It can be recorded as *w(IdB($\overrightarrow{u_i u_j}$))*, and calculated in Eq. (4.4).

$$w\left(IdB\left(\overrightarrow{u_i u_j}\right)\right) = \frac{\sum_T NIB_{ing}(u_j,u_i)}{\sum_T NIB_{ing}(u_j)} + \frac{\sum_T NIdB(u_i,u_j)}{\sum_T NIdB(u_i)} \tag{4.4}$$

where, $\sum_T NIB_{ing}(u_j)$ indicates the sum of information behaviors conducted by user u_j which have influenced other users in the selected time period T, and $\sum_T NIB_{ing}(u_j,u_i)$ indicates the sum of information behaviors conducted by user u_j which have influenced user u_i during T. $\sum_T NIdB(u_i)$ indicates the sum of influenced behaviors conducted by user u_i in the selected time period T, while $\sum_T NIdB(u_i,u_j)$ indicates the sum of influenced behaviors conducted by user u_i from user u_j during T.

4.2 User Correlation and Profiling Analysis

In this section, we firstly introduce and define a set of measures to describe and

calculate the details of user correlations in the *DSUN* model. A set of attributes are then proposed to analyze and build the user profiling.

4.2.1 Measures for User Correlations

Basically, two measures are introduced to give the basic descriptions of each user in the model.

Interest degree: Interest degree is the in-degree of each vertex. Specifically, given the vertex u_i, the interest degree can be recorded as $InD(u_i)$, while the vertexes connecting to u_i in this way can be recorded as the interest degree-based set: $InS(u_i)$, where $|InS(u_i)| = InD(u_i)$.

Popularity degree: Popularity degree is the out-degree of each vertex. Specifically, given the vertex u_i, the popularity degree can be recorded as $PoD(u_i)$, while the vertexes connecting to u_i in this way can be recorded as the popularity degree-based set: $PoS(u_i)$, where $|PoS(u_i)| = PoD(u_i)$.

The interest degree of a specific user u_i indicates the number of other users from whom he/she may get more helpful information related to his/her current interests or needs. The higher the interest degree is, the more relevant information user u_i may obtain. The popularity degree of a specific user u_i indicates the number of other users who may get useful information related to their current interests or needs from user u_i. The higher the popularity degree is, the more contribution user u_i may provide.

Furthermore, a three-tuple is utilized to define the details of users' dynamical correlations. That is, three important factors, the user similarity, user influence and user interaction, are taken into account for the refinement of user relationships. Thus, the user correlation in a selected time period *T* can be expressed as:

$$UC_T = \varphi(CS, InF, InT) \tag{4.5}$$

where

CS: The user similarity correlation, which is calculated based on the characteristics-based similarity among the users, is employed to describe the basic relationship in the *DSUN* model. That is, this kind of correlation is considered as the prerequisite relationship between two users.

InF: The user influence correlation, which is calculated based on the influenced behaviors among the users, is employed to describe the beneficial influence relationship between a pair of users in the *DSUN* model. It can also be used to analyze and describe the information or knowledge delivery among a group of users.

InT: The user interaction correlation, which is calculated based on the influencing behaviors among the users, is employed to describe the interactional relationship between a pair of users in the *DSUN* model.

Based on these, two measures are proposed to calculate and refine users' social relationships as follows.

B-Influence-based Correlation degree: Given a pair of users u_i and u_j represented in *DSUN* model, B-influence-based correlation degree indicates the beneficial influence from u_i to u_j, which considers both the characteristics-based similarities and behavior-based influence between these two users.

As discussed above, we consider if the users may generate the so-called benefactive influence on other users, or get beneficial influence from other users, two conditions should be satisfied as: given two users u_i and u_j, if $u_i \in InS(u_j)$, and $\exists\, IdB(\overrightarrow{u_i u_j})$, then there should be a beneficial influence from u_i to u_j. Thus, the B-influence-based correlation degree can be quantified as follows.

$$w(InFC_{ij}) = w(CSR_{ji}) * w\left(IdB(\overrightarrow{u_i u_j})\right) \tag{4.6}$$

Interaction-based Correlation degree: Given a pair of users u_i and u_j, interaction-based correlation degree indicates the interactional influence between u_i and u_j, which considers the influencing behaviors between these two users.

As discussed above, we consider the users may influence each other during their interaction process. Thus, given a pair of users u_i and u_j, the influencing behavior $IgB(\overrightarrow{u_i u_j})$ is taken into account to calculate the interactions between them. The more influencing behaviors they conduct to the counterpart, the higher degree value they may get in term of their interactional correlation. Thus, the interaction-based correlation degree can be quantified as follows.

$$w(InTC_{ij}) = \frac{\min\ (\sum_T NIgB(u_i,u_j),\sum_T NIgB(u_j,u_i))}{|\sum_T NIgB(u_i,u_j) - \sum_T NIgB(u_j,u_i)|+1} \quad (4.7)$$

4.2.2 Attributes for User Profiling

Users with different attributes which describe and represent one aspect of the user's profiling respectively may contribute to their communities in different ways. Thus, in this section, we define a set of user attributes, and go further to analyze and describe the users' multi-dimensional profiling referring to a group of users or a pair of users.

We first define four basic user attributes as follows.

Activeness: Activeness is one of the user attributes which identifies how the users keep conducting information behaviors within a specific time period *T*. For instance, in Twitter, the users who frequently post messages, retweet other users' contents, and communicate with his/her followers, would be considered as increasing the activeness.

Positiveness: Positiveness is one of the user attributes which identifies how the users tend to influence other users through their information behaviors. For instance, in Twitter, the users who often post contents that contain "@name" would be considered as augmenting the positiveness.

Independence: Independent is one of the user attributes which identifies how the users are accustomed to delivering their original ideas, attitudes or opinions through their information behaviors frequently. For instance, in Twitter, the users who always

post plain text messages without “RT @user” would be considered as independence.

Valuableness: Valuableness is one of the user attributes which identifies how the users’ personal contents can provide useful information related to other users. It is also an important measure to identify whether a user has the potential capability to influence a mass of other users through his/her information behaviors.

Generally, the users are sharing and exchanging their personal information and knowledge through the complicated interactions. During this process, some users tend to continuously deliver their private knowledge, which will become the origination of information dissemination in a community, while others who tend to follow these users to derive useful information will gradually be influenced by them. Thus, it is important to define and identify this kind of users who look like the center in term of a group of users, which will not only benefit the analysis of information dissemination, but also the facilitation of information seeking and recommendation.

Hub user: Hub user is the user who continuously share and deliver information and knowledge through his/her information behaviors to an influential extent in which individuals can benefit from him/her directly or indirectly, so as to result in a high reputation in regard of a group of individuals within the specific limits.

The diffusion of information along within the *DSUN* model from one user to a group of users is employed as an important feature to calculate and figure out the

so-called hub user, which can be viewed as a collective and global measure of worthiness based on the influence scope of a certain group of individuals within the information dissemination process. Inspired by the studies presented in [37, 65], the diffusion degree can be defined as follows.

Diffusion degree: Given a specific user u_i, diffusion degree indicates the density of the influence scope based on his/her information behaviors, which can be quantified in Eq. (4.8). The higher the reputation degree is, the more individuals may derive valuable information in a more extensive range.

$$w_{dif_i} = \sum_{j \in IB_{u_i}} Depth(IB_j) * IdU(IB_j) \tag{4.8}$$

where, IB_{u_i} indicates the set of information behaviors of user u_i. $Depth(IB_j)$ indicates the average influence depth of an information behavior, while $IdU(IB_j)$ indicates the number of individuals that have be influenced by this information behavior.

Consequently, based on the discussion above, a hub user should satisfy the *Activeness*, *Independence*, and *Valuableness* simultaneously, and hold a high *Diffusion degree*, which can be quantified as follows.

$$w_{hub}(u_i) = \alpha * \left(\frac{Act_T(u_i)}{|T|} * \log \sum_T NIB(u_i)\right) + \beta * \left(\frac{\sum_T NIB_{\overline{IdB}}(u_i)}{\sum_T NIB(u_i)}\right) + \gamma * (NK(u_i, T)) + \theta * w_{dif_i} \tag{4.9}$$

where, $\left(\frac{Act_T(u_i)}{|T|} * \log \sum_T NIB(u_i)\right)$ is used to quantify the Activeness. $\left(\frac{\sum_T NIB_{\overline{IdB}}(u_i)}{\sum_T NIB(u_i)}\right)$ is employed to quantify the Independence. And $(NK(u_i, T))$ is used

to quantify the Valuableness, which indicates the number of matched keywords in u_i's posted contents comparing with the whole body of users during the time period *T*.

On the other hand, the issue of promotion also plays a crucial role in information dissemination process. The modeling and identifying of promotion users will help promote the referrals and ratings of information through the users' interactional behaviors in a certain community, which will finally benefit the information seeking and knowledge sharing process.

Promotion user: Promotion user is the user who can tremendously increase and promote the sharing and delivering of information that disseminates via him/her, which means a large fraction of information will get the high referrals if this kind of users are willing to deliver them through his/her information behaviors.

Promotion degree: Given a specific user u_i, promotion degree indicates the change of referrals of information after he/she has delivered them, which also shows the power of influence regarding to a group of users. The promotion degree can be quantified as follows.

$$w_{pro_i} = \sum_{j \in IdB_{u_i}} \sum_{n=1}^{Depth(IB_j)} \frac{IdU_n(IB_j)}{n} \tag{4.10}$$

where, IdB_{u_i} indicates the set of influenced behaviors of user u_i. $IdU_n(IB_j)$ indicates the number of users who have been influenced by the information behavior IB_j of user u_i in the *n*th-depth.

Consequently, based on the discussion above, a promotion user should satisfy the *Activeness* and *Valuableness* simultaneously, and hold a high *Promotion degree*, which can be quantified as follows.

$$w_{pro}(u_i) = \alpha * \left(\frac{Act_T(u_i)}{|T|} * \log \sum_T NIB(u_i)\right) + \beta * \left(\frac{NK(u_i)}{NK}\right) + \gamma * w_{pro_i} \quad (4.11)$$

Personally, for each user in the *DSUN* model, they may not concern who influence the most of users, or who benefit the most for the information dissemination. What they are really concerned may be who can provide more valuable information related to their own requirements. In order to deal with this situation, contrasting with the hub user and promotion user, who are globally visible to the whole body of users in the networking, we further define two kinds of users to assist the users' personalized information seeking process within the pairs of users.

Contribution user: Contribution user is the better benefactor u_i among the users linked to a target user u_j, who can better support user u_j with more related and valuable information, or transfer the beneficial influence to him/her through information behaviors.

Specifically, for a pair of connected users, $<u_i, u_j>$, with their linked neighborhood users, the weight of B-influence-based correlation appending on their edges, will be taken into account to calculate the contributions from u_i to u_j.

Contribution degree: Given a pair of users, $<u_i, u_j>$, contribution degree is the

value that reflects the contribution or importance of user u_i to user u_j, which can be quantified as follows.

$$CoD(u_i, u_j) = w(InFC_{ij}) * \left(\frac{1}{\sum_{u_l \in PoS(u_i)} w(InFC_{il})} + \frac{1}{\sum_{u_k \in InS(u_j)} w(InFC_{kj})}\right) \quad (4.12)$$

On the other hand, in the most of time, the users can not benefit from other users with the valuable information directly. However, they may have several similarities either in characteristics or in behaviors with some user, which will also benefit the information and knowledge sharing between these two users in a complementary way.

Reference user: Reference user is the most similar user u_j among users linked to a target user u_i, who can understand and complement with each other better, and further share the similar information and experience in a reciprocal way.

Reference degree: Given a pair of connected users, <u_i, u_j>, with a set of users U who have influenced on them, the reference degree indicates the similarity between these two users in accordance with their characteristics and influenced behaviors with users in set U, which can be quantified as follows.

$$Ref(u_i, u_j) = sim_{Ref}(V(IdB_{u_i}), V(IdB_{u_j})) \quad (4.13)$$

where, $sim(IdB_{u_i}, IdB_{u_j})$ indicates the similarity based on the influenced behaviors using the measure of Cosine similarity. $V(IdB_{u_i})$ indicates the vector of B-influence-based correlation degree calculated with other users.

4.3 Mechanisms for Social Community Discovery

We define and develop the concept of *tie*, to discover and describe the multi-types of social communities based on the analysis of the dynamical user correlation and profiling.

Tie: Tie describes a group of users who confluence together following different types of rules and connect under a certain relationship among them, in which the users can communicate and profit from each other through their interactional behaviors in a collaborative way. Specifically, in this study, two kinds of ties are proposed, which consider the user correlations and user profiling respectively.

- User correlation-based tie

Based on our discussions above, the user correlation-based tie will be further categorized into two sub-types as follows.

Strong correlation-based tie: The strong correlation-based tie (abbreviated as *strong tie* in the following discussion) is constructed based on the users' direct interactional information behaviors, which is a partition of the whole set of users.

The algorithm to discover and construct the strong tie is shown in Fig. 4-3. The strong tie is constructed in according with the direct interactions between each user, and the discovery process is similar to the breadth-first search process. That is, in a specific period, the users who interact directly more with each other will be selected

and grouped into the same community.

Input: The user set $U = \{ u_1, u_2, \ldots, u_n \}$ in the *DSUN* model

Output: The strong tie set $C_{st} = \{ C_{st_1}, C_{st_2}, \ldots, C_{st_n} \}$

Step 1: If $U \neq \emptyset$,

create a new tie set C_{st_i}, receive user u_i from the input user set U, calculate the Interaction-Correlation degree $w(InTR_{ij}) = \frac{\min\ (\sum_T NIgB(u_i,u_j), \sum_T NIgB(u_j,u_i))}{|\sum_T NIgB(u_i,u_j) - \sum_T NIgB(u_j,u_i)|+1}$ for him/her with all the neighborhood users u_j in U,

proceed to **Step 2**

else

go to **Step 5**

Step 2: If $w(InTR_{ij}) > threshold\ \delta_1$, add u_j into a queue Q,

record tie set as $C_{st_i} = C_{st_i} \cup \{ u_i, u_j \}$, and record user set as $U = U - \{ u_i, u_j \}$

proceed to **Step 3**

Step 3: If $Q \neq \emptyset$,

pop the top element u_j in queue Q, calculate the Interaction-Correlation degree $w(InTR_{jk}) = \frac{\min\ (\sum_T NIgB(u_j,u_k), \sum_T NIgB(u_k,u_j))}{|\sum_T NIgB(u_j,u_k) - \sum_T NIgB(u_k,u_j)|+1}$ with all the neighborhood,

proceed to **Step 4**

else

go to **Step 1**

Step 4: If $w(InTR_{jk}) > threshold\ \delta_1$, add u_k into a queue Q, record tie set as $C_{st_i} = C_{st_i} \cup \{ u_k \}$, and record user set as $U = U - \{ u_k \}$,

go to **Step 3**

Step 5: Return the strong tie set $C_{st} = \{C_{st_1}, C_{st_2}, \ldots, C_{st_n}\}$

Figure 4-3 Algorithm for Generation of Strong Correlation-Based Tie

Weak correlation-based tie: The weak correlation-based tie (abbreviated as *weak tie* in the following discussion) is constructed based on the users' indirect influenced information behaviors from others, which is a partition of the whole set of users.

The algorithm to discover and construct the weak tie is shown in Fig. 4-4. The weak tie is constructed in according with the similarity of influenced information behaviors between each user, and the discovery process is similar to a clustering process. That is, in a specific period, the users who have the most similar influenced behaviors will be selected and grouped into the same community.

Input: The user set $U = \{ u_1, u_2, \ldots, u_n \}$ in the *DSUN* model

Output: The weak tie set $C_{wt} = \{C_{wt_1}, C_{wt_2}, \ldots, C_{wt_n}\}$

Step 1: For each u_i in the user set U, calculate the Influence-Correlation degree $w(InFR_{ki}) = w(CSR_{ik}) * w\left(IdB(\overrightarrow{u_k u_i})\right)$ with other users, record as $V_i = [v_{i1}, v_{i2}, \ldots, v_{in}]$, let the set $V = \{ V_1, V_2, \ldots, V_n \}$

Step 2: If $U \neq \emptyset$,

create a new tie set C_{wt_i}, receive user u_i from set U, let $C_{wt_i} = C_{wt_i} \cup \{u_i\}$,

$C_{wt} = C_{wt} \cup \{C_{wt_i}\}$, $V = V - \{V_i\}$

proceed to **Step 3**

else

go to **Step 5**

Step 3: If $V \neq \emptyset$,

receive element V_j from set V, calculate the similarity of V_j with each C_{wt_i} in C_{wt},

let $V = V - \{V_j\}$, proceed to **Step 4**

else

go to **Step 5**

Step 4: If value $> \delta_2$,

add u_j into the C_{wt_i} with the highest value, record as $C_{wt_i} = C_{wt_i} \cup \{u_j\}$,

go to **Step 3**

else

create a new tie set C_{wt_j}, let $C_{wt_j} = C_{wt_j} \cup \{u_j\}$,

go to **Step 3**

Step 5: Return the weak tie set $C_{wt} = \{C_{wt_1}, C_{wt_2}, \ldots, C_{wt_n}\}$

Figure 4-4 Algorithm for Generation of Weak Correlation-Based Tie

- User profiling-based tie

On the other hand, considering the users' multi-dimensional profiling, we further propose the user profiling-based tie according to the users' different attributes and functions. In details, the concept of hub user is utilized for the discovering process.

User profiling-based tie: The user profiling-based tie is basically constructed based on the hub users with the users who have been directly or indirectly influenced by him/her, which is a covering of the whole set of users.

Input: The user set $U = \{ u_1, u_2, \ldots, u_n \}$ in the *DSUN* model

Output: The user profiling-based tie set $C_{ua} = \{C_{ua_1}, C_{ua_2}, \ldots, C_{wa_n}\}$

Step 1: For each user u, calculate the hub user degree

$$w_{hub}(u_i) = \alpha * \left(\frac{Act_T(u_i)}{|T|} * \log \sum_T NIB(u_i)\right) + \beta * \left(\frac{\sum_T NIB_{\overline{TdB}}(u_i)}{\sum_T NIB(u_i)}\right) + \gamma * \left(\frac{NK(u_i)}{NK}\right) + \theta * w_{dif_i}$$

, sort the users by the value from high to low, and record as a sequence U'

Step 2: If $U' \neq \emptyset$,

create a new tie set C_{ua_i}, receive user u_i from the input user set U',

according to the influenced behaviors, select all the users who have influenced by u_i from user set U, record as U_{in_i},

proceed to **Step 3**

else

go to **Step 4**

Step 3: Let tie set $C_{ua_i} = C_{ua_i} \cup \{u_i\} \cup U_{in_i}$, record hub user sequence as $U' = U' - C_{ua_i}$,

go to **Step 2**

Step 4: Return the user profiling-based tie set $C_{ua} = \{C_{ua_1}, C_{ua_2}, \ldots, C_{wa_n}\}$

Figure 4-5 Algorithm for Generation of User Profiling-Based Tie

The algorithm to discover and construct the user profiling-based tie is shown in

Fig. 4-5. The user profiling-based tie is constructed in according with the calculation of hub users, as well as the influenced behaviors among the related users, which can also been viewed as the information diffusion process.

4.4 Experiments on DSUN Model

In this section, we introduce the design and implementation of an application prototype system. We show and discuss the experimental results in regards to the analyses of user profiling and community respectively, in order to demonstrate the practicability and usefulness of our proposed methods.

4.4.1 System Architecture of User and Community Recommendation

We propose a prototype system which utilizes Twitter data to calculate the user correlations and profiling among a group of users, and further discover the favorable users and social communities for recommendations based on our proposed methods.

The architecture for the individualized user and community recommendation based on user correlation and profiling analysis is shown in Fig. 4-6.

This system consists of nine major components: User Characteristics Analyzer & Extractor, User Behavior Extractor, Characteristics-based Relationship Analyzer, Influence-based Relationship Analyzer, User Networking Builder, User Correlation Analyzer, User Profiling Analyzer, User Community Analyzer, and Individualized Recommender.

The User Characteristics Analyzer & Extractor is used to analyze the collected personal data and further extract the static and dynamical features, including the users' diverse interests and time-varying topics as well. On the other hand, the User Behavior Extractor is used to analyze and extract the users' information behaviors, especially the influence behaviors among their communications. Following these, the Characteristics-based Relationship Analyzer is employed to analyze and calculate the user relationships based on the similarities of characteristics, while the Influence-based Relationship Analyzer is employed to analyze and calculate the user relationships based on the behavior-based influence. Furthermore, the User Networking Builder is responsible for constructing the *DSUN* model to describe and represent the users' implicit and explicit relationships. The User Correlation Analyzer will then analyze these relationships from the *DSUN* model, and further figure out the dynamical and potential user correlations. Based on these, the User Profiling Analyzer will concentrate on building the users' profiling in accordance with a series of attributes we proposed, while the User Community Analyzer will contribute to discovering the social communities using our proposed algorithms. Finally, by the Individualized Recommender, a set of favorable users and communities which can fit a specific user's current requirements will be recommended.

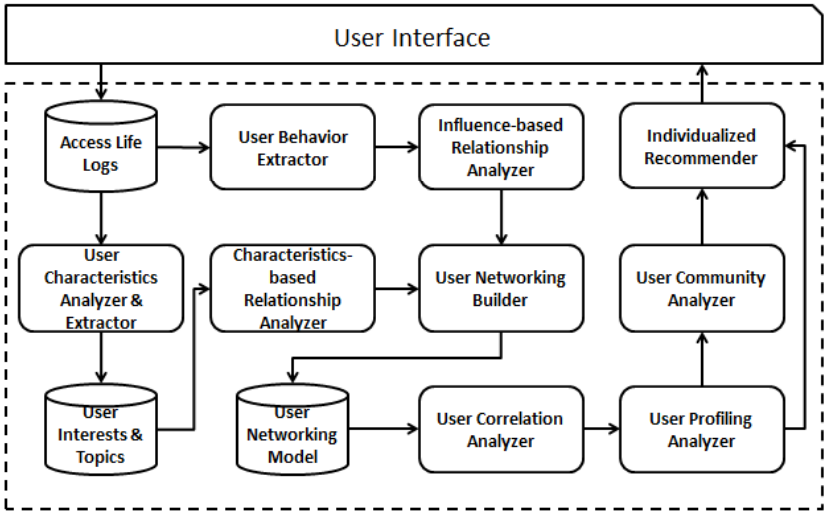


Figure 4-6 Architecture of User Correlation/Profiling-Based Recommendation System

4.4.2 Data Set for DSUN Model Experiments

As mentioned above, Twitter, one of the famous social media networking system, has been employed to collect the data set for our experimental analysis. In details, we collected data by crawling the contents generated from the users in a Twitter list na which was ranked as one of the top 140 Twitter lists, with their followees and followers. It was selected since it was one of the most popular Twitter lists, and the Tweets posted in this list were written in English with fewer other languages, which is important for us to analyze the data computationally. The collecting period is from April 1 2013 to June 5 2013. To build our *DSUN* model efficiently in a controllably computational extent, the data generated from April 28 to June 5 has been used to conduct our experiments, and the users who post the English contents can become the candidates for the modeling.

Furthermore, the selected whole time period is dynamically divided into several time slices according to the time-varying topic-based trends. That is, we extract the keywords from the posted contents and detect changing of their rank to determine the change of trends. Finally, eight time slices, T_1: April 28 - May 1, T_2: May 2 - May 6, T_3: May 7 - May 10, T_4: May 11 - May 14, T_5: May 15 - May 19, T_6: May 20 - May 29, T_7: May 30 – June 2, T_8: June 3 - June 5, are generated, in each of which one *DSUN* model can be built to describe the users' current correlations, as well as the dynamical profiling and community.

According to the basic user attributes we discussed in Section 4.2, 2461 users, whose values of Activeness are higher than 0, are employed to construct the *DSUN* model, which means these users continuously conduct information behaviors in each time slice during this period. Based on these, a total of 116,243 user-generated contents are extracted correspondingly for the user profiling and community analysis.

In addition, we give two assumptions. Firstly, after we filtering the invalid users (e.g., advertiser), as well as the Internet slang words in the contents, we assume that there are no deviant users in our data set, which means all users have conducted their information behaviors normally and generated no spam or irrelevant actions. Secondly, due to the upper limit of each user's available collecting time for the officially provided Twitter API which we used to collect the data, we assume that we collect

adequate data for each user in our experiment even if his/her posts exceed over the limitation.

4.4.3 Experimental Results of User Profiling

4.4.3.1 Analysis of Hub User and Promotion User

Table 4-2 Results of Top 10 Users for Basic Attributes

Activeness		**Positiveness**		**Independence**		**Valuableness**	
justinbieber	0.44	StoryAboutNSN	0.64	dorieclark	0.18	HannahMixdChick	0.74
HannahMixdChick	0.44	WishtoMeetJB	0.40	judahsmith	0.18	justinbieber	0.37
Forbes	0.36	armyofsb	0.39	adidasNEOLabel	0.18	X3Kim	0.37
scooterbraun	0.36	GecaBieber	0.30	FreedMyself	0.18	iamwill	0.35
hhassan140	0.36	dopebiebss	0.22	ClevverTV	0.18	scooterbraun	0.30
iamwill	0.32	PaO_LaRrAzA	0.21	newsycombinator	0.18	ddlovato	0.28
cstross	0.32	denisse__chavez	0.17	dandypantsfilms	0.18	dijahyellow	0.28
MileyCyrus	0.24	erikamkidrauhl	0.16	CandyEsparza1	0.18	Forbes	0.22
X3Kim	0.24	Boybeliebertaha	0.15	dijahyellow	0.18	theycallmejerry	0.22
BelieveTUpdates	0.19	LeslyKelsBieber	0.14	ruizabieberswag	0.18	cstross	0.22

As the basic statistics analysis, we calculate the four attributes discussed in Section 4.2.2 to describe the basic profiling of each user in the latest time slice T_8 from the whole time period. Each attribute can be employed to independently represent users' one aspect of his/her profiling in accordance with the information behaviors they conducted in the selected time period. The top ten users for each attribute are shown in Table 4-2. Note that each column of value is normalized by $\sqrt{\sum_{i=1}^{N} W^2}$ respectively, where N indicates the number of users, while W indicates the value of attribute in each column. The calculations in the following also use this normalization

method.

Based on these, we further calculate the hub and promotion users to describe and find some specific users in terms of their global and collective contributions according to Eq. (4.9) and Eq. (4.11) respectively. We set $\alpha = \beta = \gamma = \frac{1}{6}$, $\theta = \frac{1}{2}$ in Eq. (4.9), and $\alpha = \beta = 0.25$, $\gamma = 0.5$ in Eq. (4.11), as we assume that the diffusion degree and promotion degree would be more important to identify the hub and promotion user in this study. The results of top ten hub and promotion users are shown in Table 4-3.

Table 4-3 Results of Top 10 Hub and Promotion Users

Diffusion Degree		**Hub User**		**Promotion Degree**		**Promotion User**	
justinbieber	0.97	justinbieber	0.65	RAMARTIBE	0.51	RAMARTIBE	0.30
scooterbraun	0.22	scooterbraun	0.25	RT2PROMO	0.30	hesniall	0.17
AlfredoFlores	0.07	HannahMixdChick	0.23	justinbieber	0.30	justinbieber	0.15
Forbes	0.04	iamwill	0.15	hesniall	0.30	RT2PROMO	0.15
theycallmejerry	0.03	Forbes	0.15	monicahillb	0.20	warriorGaGa	0.12
iamwill	0.03	X3Kim	0.13	mkrigsman	0.20	monicahillb	0.10
MileyCyrus	0.02	cstross	0.12	EBruschini	0.20	LadyGagaINDO	0.10
BelieveTUpdates	0.02	ddlovato	0.12	LadyGagaINDO	0.20	mkrigsman	0.10
ddlovato	0.02	hhassan140	0.11	warriorGaGa	0.20	EBruschini	0.10
billboard	0.02	BelieveTUpdates	0.10	missioncontinue	0.10	ccitizen21	0.08

We give our observations and discussions for the hub users and promotion users based on their attributes as follows.

1) As for each attribute shown in Tables 4-2 and 4-3, the users with high rankings as the hub users almost have high values of each basic attribute, especially

for the Activeness and Valuableness. On the contrary, it seems that the users with high rankings as the promotion users may not keep high values of each basic attribute. It indicates that to a certain group of users, the hub users always keep active and provide valuable information, so that they can influence on a large scale of users, while the promotion users may not always keep as active as the hub users, but when they tend to post their personal contents or deliver other users' information, lots of users will also be influenced by them, which will greatly promote the information dissemination process.

2) On the other hand, since we consider the diffusion degree and promotion degree are more important than other basic attributes in the identification processes of hub and promotion user, as shown in Table 4-3, the users who obtain the higher value of diffusion degree will also keep higher rankings as the hub users, which are the same in the promotion users. Furthermore, comparing the values in both the hub user and promotion user, it seems that the distribution of hub users are more centralized than the promotion users in our data set, especially for the top three users, which also means that according to our data set, it is more obvious and easier to distinguish the hub users than the promotion users, since the values of promotion users tend to be close.

3) Specifically, some users who hold the higher values in each attribute become

both the hub user and promotion user based on our data set, which means in this case, this kind of users are extremely important in constructing the relationships among the users. Moreover, note that the user who keeps the third ranking as the hub user does not obtain a high value in the diffusion degree, but is the top one in the attribute of Valuableness. In this situation, it means although this user has not influenced on a large number of users as other hub users actually, he should be viewed as a potential hub user since he has posted lots of valuable information related to other users. Thus, it is extremely important to identify and recommend this kind of users to others, to benefit the information seeking and sharing process.

4.4.3.2 Analysis of Contribution User and Reference User

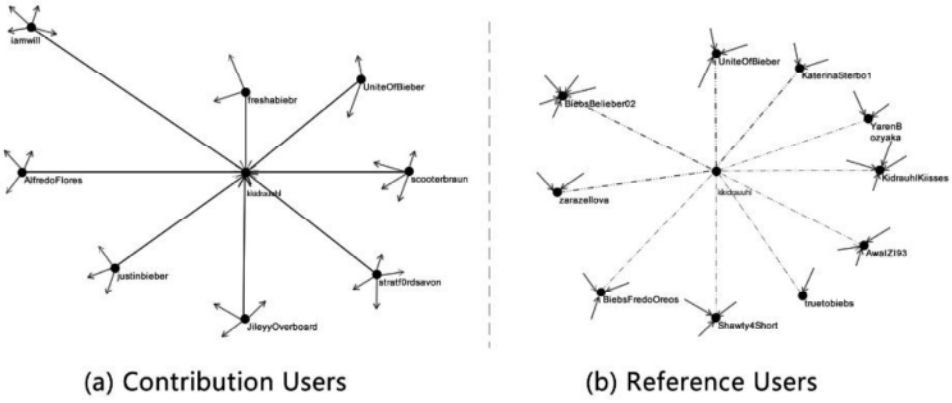


Figure 4-7 An Image of A Specific User's Contribution and Reference Users

To better support the personalized information seeking and recommendation process using our user networking model, we calculate the dynamical correlations between two users based on the proposed measures. Table 4-4 shows the best five contribution

and reference users for a specific user, and Fig. 4-7 shows the image of the dynamical connections of him/her with other users.

Table 4-4 Results of Top 5 Contribution Users and Reference Users for a Specific User

Contribution User	$InD(u_i)$	$PoD(u_i)$	$CoD(u_i, u_j)$
freshabieber	35	111	0.47
UniteOfBieber	4	234	0.36
scooterbraun	6	449	0.20
stratf0rdsavon	14	20	0.19
JileyyOverboard	3	56	0.16
Reference User	$InD(u_i)$	$PoD(u_i)$	$Ref(u_i, u_j)$
UniteOfBieber	75	10	0.39
KaterinaSterbo1	42	2	0.37
YarenBozyaka	36	1	0.36
KidrauhlKiisses	105	3	0.35
AwaIZI93	51	2	0.35

We further give our observations and discussions for the contribution and reference user to a specific user as follows.

1) Generally, according to Table 4-4, the users who can provide others with more related information would mostly keep a high popularity degree, which could be viewed as one feature of the contribution users at most of the time. Moreover, as shown in Table 4-4, due to the diversity of users' requirements, the user who holds the highest popularity degree may not be the most suitable for the specific user, which means that the calculation of contribution user can help find more suitable users to provide more related and personalized information. On the other hand, comparing

with the popularity degree, the reference users may hold a higher interest degree, which can be employed to indicate several similar influencing users related to the specific user.

2) Another important difference between the contribution user and reference user is that, comparing the scores of contribution degree and reference degree from the top to the bottom respectively, the contribution degree declines gradually from the first to the fifth, while the reference degree looks almost the same among the top five users.

3) As shown in Fig. 4-7 (a), to the specific user indicated in the center, totally, there are eight contribution users calculated in this case, while we only show the top ten reference users to him in Fig. 4-7 (b). The length of each edge indicates the weight of correlations between two users. We connect the specific user with the contribution users using the solid lines, in order to indicate their direct correlations in the *DSUN* model, while the reference users are connected using the dotted lines, which indicates their indirect correlations.

4.4.4 Experimental Results of Social Community Discovery

We go further to discuss the community discovery result using the *DSUN* model. Fig. 4-8 shows the images of three basic and important types of communities we proposed in this study.

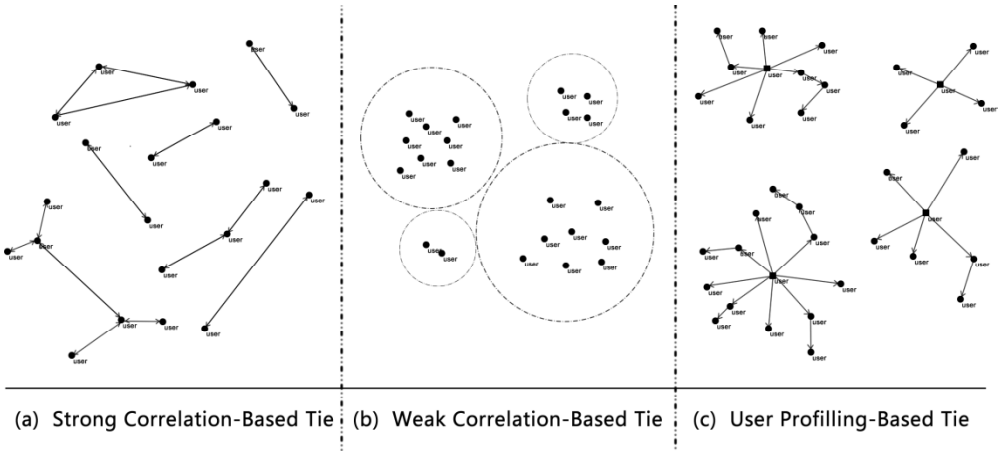


Figure 4-8 Images of Different Types of User Communities

Fig. 4-8 (a) shows a basic image of the strong correlation-based ties according to our experiment results, which looks like a series of vertexes connected with several polylines. In our data set, the sizes of each strong correlation-based tie are small. The biggest community only contains six users, while most of the communities only contain two users, which means that at most time the interactions are generated among a small number of users, especially a pair of users in most of the cases.

Fig. 4-8 (b) shows a basic image of the weak correlation-based ties according to our experiment results, which looks like a series of clusters. In our data set, the sizes and numbers of weak correlation-based tie often change dynamically.

Fig. 4-8 (c) shows a basic image of the user profiling-based ties according to our experiment results, which spread from the center to all around. In our data set, the sizes of user profiling-based tie change largely according to different hub users.

4.4.4.1 Analysis of Correlation-Based Tie

We give some detailed analysis for the correlation-based tie, specifically, the weak correlation-based tie.

Table 4-5 Statistics for Numbers of Communities According to Different Thresholds

		0.50	**0.71**	**0.86**	**0.95**
4.28-	Regular	12	13	14	15
5.01	Isolated	0	6	10	15
5.02-	Regular	13	15	17	16
5.06	Isolated	0	7	10	17
5.07-	Regular	10	12	14	21
5.10	Isolated	0	7	8	10
5.11-	Regular	9	10	12	15
5.14	Isolated	0	3	3	6
5.15-	Regular	5	8	10	16
5.19	Isolated	0	0	1	4
5.20-	Regular	7	10	11	19
5.29	Isolated	0	0	4	5
5.30-	Regular	5	8	10	15
6.02	Isolated	0	0	1	2
6.03-	Regular	6	8	9	12
6.05	Isolated	0	0	2	2

We employ four different thresholds, 0.5, 0.71, 0.86, and 0.95, to generate different sizes of communities in which community members will be influenced by different groups of users. Specifically, as shown in Table 4-5, when setting the threshold to 0.5, we only count all the communities which members are more than one user, and record them as regular communities. When the threshold is increased to 0.71, 0.86, and 0.95, these communities are separated, and new communities generate. We

count and record the communities with one member, which are also influenced by a certain group of users and separated from the original communities due to the different thresholds, as isolated communities.

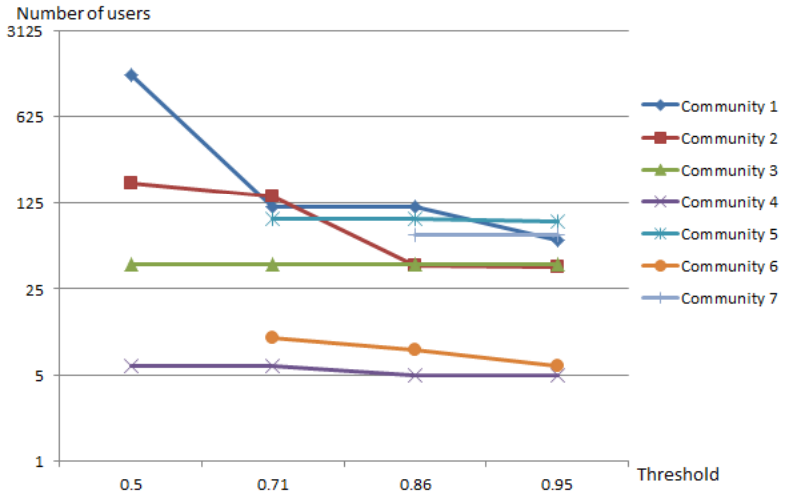


Figure 4-9 Changing in Size of Communities According to Different Thresholds

We illustrate how the size of community changes and new community generates along with the changing of different thresholds. As shown in Fig. 4-9, four communities are selected when the threshold is set to 0.5, in which Community 3 keep its size when the threshold is increased to 0.71, 0.86, and 0.95, while the sizes of other three communities all reduce along with the increasing of thresholds. In details, when the threshold is increased to 0.71, another two new communities, Community 5 and Community 6, are separated from other communities. Likewise, Community 7 occurs when the threshold is set to 0.86. Among these new generated communities, Community 6 reduces its size when the threshold is bigger, which is the same to

Community 1, Community 2, and Community 4, while Community 5 and Community 7 almost keep their sizes along with the further changing of thresholds.

4.4.4.2 Analysis of Profiling-based Tie

The constructing of profiling-based tie mainly depends on the identifying of hub users. Different set of hub users will lead to different sizes and numbers of profiling-based ties. Table 4-6 shows some statistics of a set of hub users who are employed to construct profiling-based tie in the time slice T_6, which results in most users in communities comparing with other time slices.

Table 4-6 Statistics for Hub Users in User Profiling-Based Ties

Hub user	Average Depth	Biggest Depth	Covered Users
justinbieber	1.0026	5	1546
scooterbraun	1.0036	4	752
iamwill	1.0012	2	330
AlfredoFlores	1.0010	3	271
pattiemallette	1.0030	2	240
MileyCyrus	1.0074	6	189
Forbes	1.0179	5	48

As shown in Table 4-6, totally seven users contribute to the community constructing. The biggest depth ranges from two to five in accordance with different hub users. However, due to the large numbers of covered users in each community, the average of depth is approximate to one, which indicates the users tend to share and deliver information directly from the close user whom they connected to.

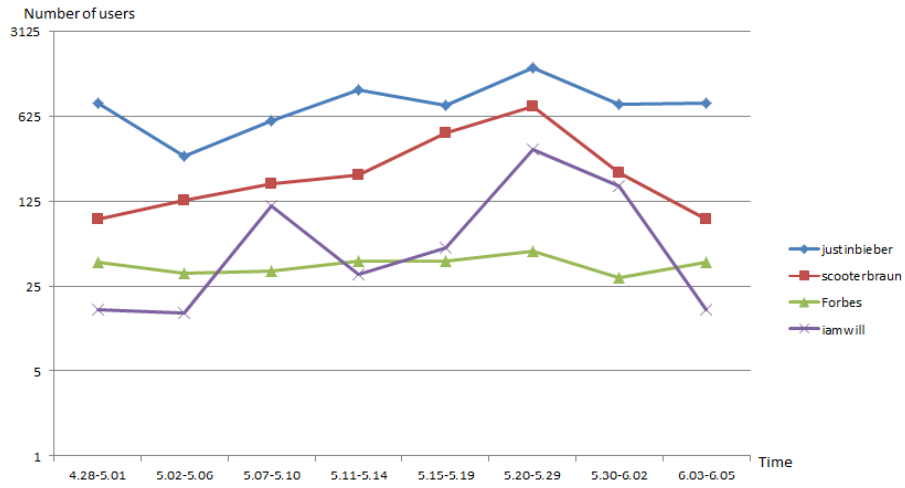


Figure 4-10 Changing in Size of User Profiling-Based Ties for Different Hub Users

We further select four users who continuously become the hub users in constructing profiling-based ties in each time slice, and demonstrate the changing of the size of communities depending on them. As shown in Fig. 4-10, during the whole time period, the top two users always keep in the top rankings, which mean they are continuously attract and influence the most numbers of users and hold the biggest communities. On the other hand, other two users keep on alternating their rankings in different time slices along with the changing of different topics.

As we discussed above, the user profiling-based tie also contributes to the facilitation of information dissemination. The bigger the community is, the better the information dissemination will be. Fig. 4-11 shows part of users in the community that is constructed by the hub user indicated in the center of the graph. In other words, in this community, the related information will originate from the hub user in the

center, and then deliver one by one through the users along with the directed edges. Especially, the promotion users indicated in this graph will improve this information dissemination process and help deliver the information and knowledge to more related users in an efficient way.

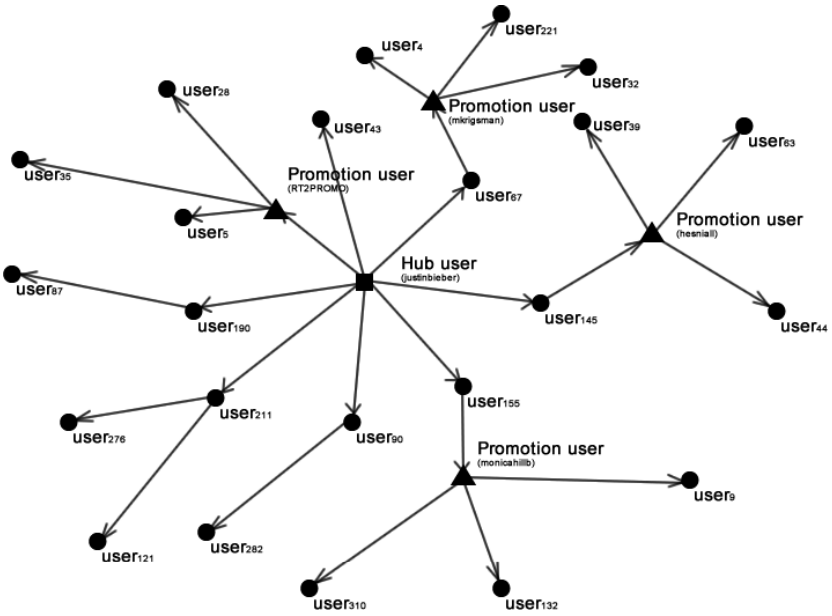


Figure 4-11 Image of Information Dissemination in User Profiling-Based Tie

4.4.5 Discussions

Based on the analysis results, to benefit a specific user (e.g., the target user mentioned in Section 4.4.3.2), firstly, a list of rankings with the corresponding values of four attributes will be calculated for him as a basic information referring to his recent state analysis. Furthermore, As shown in Table 4-4, according to the analysis of his correlations with other users, a set of contribution and reference users, sorting by their rankings, can be provided to him as well, in which the contribution user is enabled to

directly provide him with the more detailed resources, while the reference user is capable to guide him a shortcut as an indirect way to approach his purpose.

As for the community recommendation, the discovering and analyzing of three basic kinds of communities could show the users a whole of perspective (like Fig. 4-8) of users' interactional communications in different views. For instance, the community based on strong tie can facilitate the collaborative works along with the better interactions and communications with others. That is, the users involved in strong ties can have better discussions to assist their collaborative works, or motivate the social knowledge generation and facilitate the information sharing. Besides, based on the analysis of each community, the topics they discussed can also attract other users who are interested in them to join in, which can incent the motivations for the information seeking and knowledge sharing.

As for the community discovered by weak tie, in this situation, the users in the same community are influenced by a set of the same users, thus, although they may not have direct interactions with each other, the similar interests or needs can become the incentive to make them to exchange and share their personal experience, which may result in direct interactions in the community. That is, communities based on weak tie can help the users to be acquainted with more friends. Continuously, along with the communication progress of information exchanging among users, the weak

tie may evolve to the strong tie with increasing interactions. In addition, these kinds of communities, as well as their personal knowledge and experience, can be recommended to similar users who are also interested in a set of same users, to involve more users to become active users, and further benefit the information recommendation process.

Differing from former two types of ties, the discovery process of the profiling-based tie especially relies on the selection of hub users. Different orders of choosing the hub users will lead to different generation results of the user profiling-based tie. Thus, to better support the influence-based information dissemination process, we design our mechanism to involve more users into a community by selecting the hub user from high to low, and try to engage as more promotion users as possible into each community. In addition, following this way, as the process of information dissemination shown in Fig. 4-11, the user-generated information, delivered from the hub user, refined by other users in the diffusing process, promoted and shared by the promotion user, can integrate with a great degree of collective intelligence of the whole community, which can be shared with a range of abilities and knowledge that would not reside in a single individual. The derived ideas and opinions along with this process would attract more users to engage into with more personal and diverse information and knowledge, which could extremely

boost the information sharing in a socialized way.

4.5 Summary

In this chapter, we have proposed a unified method to analyze the dynamical user correlations and build the multi-dimensional profiling in accordance with the valuable outcomes from the analysis of personal data and individual behaviors, which can be utilized for the recommendations of favorable users and social communities.

Firstly, we have introduced and refined the *DSUN* model to construct the dynamical user networking in accordance with two basic relationships among a group of users. The characteristics-based relationship is employed to describe the implicit relationships based on the similarities of users' characteristics, while the influence -based relationship is employed to describe the explicit relationships based on the users' direct interactional behaviors. A set of measures have been introduced and defined to describe and measure the details of user correlations, and a set of attributes have been proposed to analyze and build the multi-dimensional user profiling in both individual and collective way. For the social community analysis, three algorithms have been developed to discover and represent three basic types of ties according to the user correlations and profiling respectively.

An application prototype system has been designed and implemented, in which the Twitter data was employed to demonstrate the high usability and practicability of

the proposed *DSUN* model. The experimental results of user profiling illuminate that the four basic attributes we proposed can correctly describe the users' basic profiling. And our newly defined hub user, promotion user, contribution user, and reference user can efficiently identify the favorable users to support a specific user in both global and individual way. The experimental results of the discovery of social community illustrate that our proposed and developed mechanisms can discover the user correlations and profiling based communities dynamically in the different time periods, which can be employed for the facilitation of personalized information utilization.

Chapter 5 Task-Oriented Recommendation for Learning Support

With the high development of social networks, collaborations in a socialized web-based learning environment has become increasing important, which means people can learn through interactions and collaborations in communities across social networks. In this chapter, two important factors, user behavior patterns and user correlations, are taken into account to facilitate the information and knowledge sharing in a task-oriented learning process. Following a hierarchical graph model for the enhanced collaborative learning within a task-oriented learning process, the LA-Pattern (Learning Action Pattern) and Goal-driven Learning Group, are introduced to extract and analyze users' learning behaviors in both personal and cooperative way. An integrated mechanism is developed to utilize both user behavior patterns and user correlations for the recommendation of individualized learning actions. The experiment and evaluation results are presented and discussed finally.

5.1 Definitions and Hierarchical Model

According to the formal description of action behaviors in Section 3.2.2, a series of definitions are introduced and defined to describe the user information behaviors in a

task-oriented learning process.

Learning Action: A learning action is composed of the minimum unit of learning operations. A learning action may consist of a series of learning operations. For example, learning new foreign language words is regarded as a learning action, in which two operations, reciting and handwriting new language words, can be employed for two minimum units of it.

Learning Activity: A learning activity is a set of learning actions, which constitute a purposeful learning process with a certain learning action sequence and time span. It is an educational process or procedure intending to motivate learning through actual experience. For example, in a foreign language lesson, a learning activity consists of learning actions - learning new words, new grammars and texts, doing exercises and quizzes. In this situation, the goal of this action sequence is to finish the final quiz. Besides, the sequence of learning actions for a learning activity can be optimized so as to satisfy different target students.

Learning Sub-Task: A learning sub-task is a set of learning activities toward a certain learning purpose in a specific learning stage. For example, each lesson could be viewed as a learning sub-task in the whole foreign language learning course.

Learning Task: A learning task is a set of learning sub-tasks. It contains all learning action sequences that represent a complete learning process.

Following the definitions above, a hierarchical model is addressed to interpret the structure and relations in task-oriented learning processes.

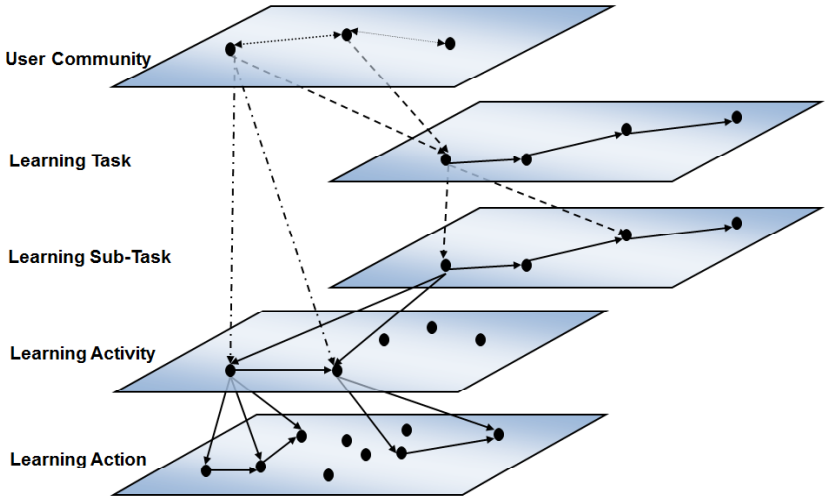


Figure 5-1 Model of Task-Oriented Learning Process [7]

Generally, as shown in Fig. 5-1, a learning course may be divided into several learning tasks with a certain sequence to achieve the final learning purpose. As for each learning task, it shall contain several sub-tasks in different learning stage with a specific sub-purpose. Continuously, in each learning sub-task, a series of learning objectives shall be established in different learning periods for different users, followed by a series of learning activities which are assigned to realize them. Likewise, the learning activity can further be divided into a sequence of learning actions which are composed of the minimum units of the learning operations. In details, they can be recorded as what has been done at what time with the necessary materials for a specific learner. Take a specific English learner for instance, a learning

action sequence can be recorded as memorizing new words in early morning, learning English grammar in the morning, and doing exercise in the afternoon.

As for the users in a learning course, one learning task may be assigned to a group of users. To pursue the similar learning purpose, users may have some discussions through a web-based communication system, which may generate a number of direct interactions (such as reply, mention) and a variety of indirect interactions (e.g., posts containing similar topics). That is, all these direct and indirect interactions, which may contain users' hidden intentions, can be utilized to analyze user correlations within a learning task during a specific period. Moreover, based on these dynamical relationships calculated among the users, the similar learning actions, especially those successful sequences of learning actions, can be shared along with successful learning experience and user-generated social knowledge.

5.2 Similarity of Learning Action Behaviors

In this section, the concepts of Learning Action Pattern (LA-Pattern) and Goal-driven Learning Group, as well as their formal definitions and related algorithms, are introduced to discover and analyze users' learning behaviors expressed by learning actions.

5.2.1 Generating of Learning Action Patterns

The Learning Action Pattern (LA-Pattern) is defined to discover and describe

individual users' information behavior patterns within a task-oriented learning process.

Following the similarity analysis of the sequential behaviors in Section 3.2.2, for each activity, < act_1, act_2, …, act_n, G>, in a learning process, if the special action G is "start taking a quiz", the < act_1, act_2, …, act_n > indicates a series of learning actions that are performed to prepare for a better outcome of the quiz. And the time period T in each sub-task indicates a learning stage within the whole learning task. For example, if the whole learning task is an English language learning course, each lesson can be viewed as a *S-Task*. Thus T indicates the learning period during this lesson, while < Act_1, Act_2, …, Act_n > indicates the learning activities that are taken for this lesson.

Based on these discussed above, the LA-Pattern, which is a sub-sequence of learning actions, can be defined as:

$$< act_i >_u^w \rightarrow G \tag{5.1}$$

where <act_i> denotes the learning action sequence which tends to constitute a purposeful learning process. w denotes the weight of this LA-Pattern, and specifically, the value indicates the frequency that this segment occurs in a whole learning action sequence (e.g., a learning action sequence generated from a learning sub-task). u denotes the user whom this learning action sequence belongs to. G denotes the

learning goal of this sequence. In this way, the LA-Patterns can be viewed as a set of sequential learning behaviors frequently occurring in a specific user's learning action sequence, which intends to complete a certain learning purpose.

In this study, for a specific user, the input set of the learning action sequences shall consist of his/her learning sub-tasks. The algorithm for LA-Pattern generation based on the similarity of the sequential behaviors is described in Fig. 5-2.

Input: The Learning action sequence set {< act_1, act_2, …, act_n>}

Output: The LA-Pattern set $\{< act_i >_u^w \rightarrow G\}$

Step 1: For each user u_i, divide the whole learning action sequence < act_1, act_2, …, act_n> into several sub-sequences in each corresponding pre-defined learning period T, which can be recorded as q_i:

$$\{< act_1, act_2, \dots, act_j >_{T_1}, \dots, < act_1, act_2, \dots, act_k >_{T_n}\}_{u_i}$$

Step 2: For each q_i, insert each sequence element with their corresponding subsequence suffixes into a *trie*, in order to build a tree-based data structure for each user u_i

Step 3: For each user u_i, traverse the *trie* from the root node to each leaf node in order to acquire all the sub-sequences with their corresponding frequency, which can be recorded as *sq*: $\{\{< act_k >^w\}_{u_i}\}$

Step 4: Filter each element in *sq* according to *Criteria 1*, if $k >= 2$, $w>=2$ and $act_k \in G$, remain the satisfied sequences, which can be recorded as *rsq*: $\{\{< act_j >^w \rightarrow G_n\}_{u_i}\}$

Step 5: For $\forall rsq_i, rsq_j \in rsq$, according to *Criteria 2*, if $w_x = w_y$, and $< act_i > \subset < act_j >$, let $rsq_j = rsq_i \cup rsq_j$

Step 6: Return the LA-Pattern: $\{< act_i >_u^w \rightarrow G\}$ for each user u_i

Figure 5-2 Algorithm for LA-Pattern Generation [7]

5.2.2 Generating of Goal-Driven Learning Group

Based on the formalization and generation of LA-Patterns, in order to analyze the similarities in a user community in accordance with each individual user's learning

behaviors, we introduce the concept of Goal-driven Learning Group.

Goal-driven Learning Group: A Goal-driven Learning Group is given to model learning behaviors of a certain group of users, who have the same learning goal and similar learning actions according to the LA-Patterns in a specific learning period, which can be formalized as:

$$\{[< act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}][G_j, T]\} \tag{5.2}$$

where, $<act_k>$ denotes the learning action sequence in a specific LA-Pattern, u_i denotes the owner of this LA-Pattern, while w_i denotes the corresponding weight. G_j denotes goal actions representing the learning purpose for this learning group, and T denotes the specific learning period.

Input: The LA-Pattern set $\{< Act_i >^{w}_{u} \rightarrow G\}$
The learning goal set $\{G_n\}$
Output: The Goal-driven Learning Group set
$\{\{[< Act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}][G_j, T]\}\}$

Step 1: For each user u_i, according to the learning goal set $\{G_n\}$, divide the LA-Pattern set $\{< act_i >^{w}_{u} \rightarrow G\}$ into several sub-sets, which can be recorded as $\{[< act_j >^{w}_{u}, \ldots, < act_k >^{w}_{u}] \rightarrow G_n\}_{u_i}$

Step 2: Group the learning action pattern set $\{[< act_j >^{w}_{u}, \ldots, < act_k >^{w}_{u}] \rightarrow G_n\}_{u_i}$ by selecting the learning action sequence $[< act_j >^{w}_{u}, \ldots, < act_k >^{w}_{u}]$ with the same learning goal G in a specific learning period T

Step 3: In each group, incorporate each $< Act_j >^{w}_{u}$ which may have the same learning action sequence, and record as $[< Act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}]$

Step 4: Return the $\{\{[< Act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}][G_j, T]\}\}$ as the final the Goal-driven Learning Group set

Figure 5-3 Algorithm for Goal-driven Learning Group Generation [7]

The algorithm to compute the Goal-driven Learning Group for the behavioral modeling of a group of users is described in Fig. 5-3.

The learning period can be defined according to different factors and granularities. For instance, according to the curriculum, the learning period can be defined as one course or one lesson, while according to the time, the learning period can be defined as one month, one week, or one day. In each period, a group of users' learning patterns will be firstly grouped by their different learning purposes which indicate as goal actions, and then be further grouped by calculating the similarity of the corresponding learning action sequence $<act_k>$. That is, users will be assigned into groups according to their same learning purpose and similar learning behaviors. Note that, this process is not a partition, which means one user can be assigned into several Goal-driven Learning Groups.

We consider three users with their learning action sequences to form an input user group as a simple example to demonstrate the generation process of LA-Patterns, as well as Goal-driven Learning Groups based on our algorithms.

Given users u_1, u_2, and u_3, with their recorded learning action sequences, s_1 = <*egiuvgiuvgi*>, s_2 = <*uvuvgiuvgi*>, and s_3 = <*uvgiuvgiuvgigi*> in the same period T = *one week*.

As for user u_1, using the *trie*-based algorithm, the sub-sequences of learning

actions with their corresponding frequencies can be calculated as $sq_{u_1} = \{< gi >^3, < egi >^1, < vgi >^2, \dots\}$.

Assume the learning purpose is denoted by learning action *i*, based on *Criteria 1* described in Section 3.2.2, the preliminary learning action pattern set can be calculated as $rsq_{u_1} = \{< g >^3 \rightarrow i, < vg >^2 \rightarrow i, < uvg >^2 \rightarrow i\}$. According to *Criteria 2* described in Section 3.2.2, the sequences $< vg >^2$ and $< uvg >^2$ satisfy $< vg > \sqsubset < uvg >$, and $w_{vg} = w_{uvg} = 2$. Thus we combine these two sequences together, and the LA-Pattern set for user u_1 can be calculated as $\{< g >^3_{u_1} \rightarrow i, < uvg >^2_{u_1} \rightarrow i\}$. Similarly, the patterns for other two users can be calculated as $\{< g >^2_{u_2} \rightarrow i, < uvg >^2_{u_2} \rightarrow i\}$, and $\{< g >^4_{u_3} \rightarrow i, < uvg >^3_{u_3} \rightarrow i\}$. Furthermore, the Goal-driven Learning Group for these three users during *one week* can be calculated as $\{[< g >^{3,\ 2,\ 4}_{u_1,u_2,u_3}, [< uvg >^{2,\ 2,\ 3}_{u_1,u_2,u_3}][i, one\ week]\}$.

5.3 User Correlation Analysis in Learning Processes

The constructing of *DSUN* model is applied to build a user networking in a learning process, in order to support the calculation of similarity among users. In details, two important factors, users' communication actions and similarities of their posted contents, which indicate direct interactions and indirect interactions among users, are taken into account to construct the user networking within a learning process. Thus, Eq. (5.3) is developed to quantify the value of w_{ij} between two vertexes u_i and u_j,

which denote two related users.

$$w_{ij} = \alpha * w_{ar} + \beta * w_{cr} \tag{5.3}$$

where w_{ar} denotes the relationships based on users' direct communication actions, and w_{cr} denotes the relationships based on similarities of users' posted contents. α and β are the importance coefficients which satisfy $\alpha + \beta = 1$.

The user relationship based on direct interactions, which can be viewed as direct relationship among users, is calculated in accordance with users' direct communications. That is, the value of w_{ar}, ranging from 0 to 1, indicates the interested degree from user u_i to user u_j based on the direct interaction times during a specific learning period (e.g., reply, mention), which can be quantified as:

$$w_{ar} = \frac{IR(u_i,\ u_j)}{IR(u_i)} \tag{5.4}$$

where $IR(u_i,\ u_j)$ denotes the interaction number from user u_i to user u_j, $IR(u_i)$ denotes the total interaction number of user u_i.

The user relationship based on indirect interactions, which can be viewed as indirect relationship among users, is calculated in accordance with the similarities of users' posted contents. That is, the intentions hidden in users' posted contents will be taken into consideration, in order to analyze the potential user relationships. Since users may not always discuss topics about their learning courses in the system, to restrict user relationships within the learning domain in user networking, we provide a

set of keywords for each lesson, in order to extract more keywords related to users' learning intentions. Eq. (5.5) is used to extract keywords that can represent users' intentions more related to a specific lesson in a selected duration *d*, which can be expressed as:

$$K = \frac{TF(t_i,d_j)}{\sum_{k=1}^{n} TF(t_k,d_j)} * \frac{|D|}{|\{j: t_i \in d_j\}|} + \varphi C_{kos} \quad (5.5)$$

where $C_{kos} = \begin{cases} 1, if\ this\ keyword\ exsists\ in\ the\ learning\ set \\ 0, otherwise \end{cases}$

In Eq. (5.5), $TF(t_i, d_j)$ denotes the frequency of a word t_i in a selected duration d_j. $|D|$ is the number of all the durations divided from the whole period. φ is the equilibrium coefficient which ranges from 0 to 1.

For each keyword which can represent a specific user's potential intention, we connect those users who post contents related to this keyword with him/her in the user networking according to the weight w_{cr}, which can be quantified as:

$$w_{cr} = \frac{n_K}{M_j} \quad (5.6)$$

where n_K denotes the frequency of the keyword that indicates the target user u_i's intention occurring in user u_j's posted contents. M_j denotes the total number of user u_j's posted words.

Based on these calculations, each extracted keyword calculated by Eq. (5.5) will represent one of a user's current intentions with its importance indicated by its weight within the current learning period. In this study, we select one keyword with high

importance (weight) as a user's intention and go further to calculate the correlations of contents posted by other users according to this intention using Eq. (5.6), in order to find those users who can provide information more related to it.

Based on these, specifically, to calculate the better benefactor u_i among users linked to a target user u_j, the contribution degree between a pair of users, exactly, from u_i to u_j, can be quantified as:

$$CoD(u_i, u_j) = w_{ij} * \left(\frac{1}{\sum_{l=1}^{n} w_{il}} + \frac{1}{\sum_{k=1}^{n} w_{kj}}\right) \quad (5.7)$$

Specifically, for a pair of connected vertexes, $<u_i, u_j>$, the value indicates in what degree the user u_i can support the user u_j in accordance with one of the current intentions. Thus, the correlation calculated based on $CoD(u_i, u_j)$ can be useful for the user u_j to better cope with his/her requirement in this situation, which means it can be applied to identify the most possible user u_i who can best support the target user u_j.

5.4 Recommendation in Task-Oriented Processes

In this section, an integrated method is proposed and developed to support a target user with the individualized learning action recommendation based on the analysis of both learning action patterns and user correlations within a task-oriented learning process.

5.4.1 Detection of Goal-Driven Learning Action Patterns

We have demonstrated how to extract the LA-Patterns from an individual, which can

represent his/her personalized learning behaviors, and how to generate the Goal-driven Learning Groups in a user community, which can describe the similarities among a group of users. Based on these, in this section, we discuss how to utilize other users' action patterns to support the detection of a specific user's learning behavior patterns, in order to facilitate the learning action recommendation process.

Input: A learning action sequence $s = < act_1, act_2, \ldots, act_n >$
A given Goal-driven Learning Group $LG = \{[< Act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}][G_j, T]\}$
Output: The goal-driven learning action pattern set $DP = \{<act_i>^w\}$

Step 1: For all the learning action sequence $< Act_k >$ in the given Goal-driven Learning Group $LG = \{[< Act_k >^{w_1,w_2,\ldots,w_i}_{u_1,u_2,\ldots,u_i}][G_j, T]\}$, build a *trie*-based structure *LG'*

Step 2: For the inputting learning action sequence from a specific user *u*, find all the sub-sequences $<act_i>$ in accordance with the *trie*-based structure *LG'*

Step 3: For each $<act_i>$, calculate the frequency as the weight *w*, which can be recorded as $<act_i>^w$

Step 4: Return the set $\{<act_i>^w\}$ as the goal-driven learning action pattern set *DP* for user *u*

Figure 5-4 Algorithm for Goal-Driven Learning Action Pattern Detection [7]

Specifically, to detect a specific user's learning behavior patterns toward a specific learning goal, the Goal-driven Learning Group is employed as a given learning action pattern set, which can help model an inputting target user's learning action sequence. That is, for a specific learning goal, we try to find all the matched sub-sequences in a target user's given learning action sequence in accordance with the learning action patterns selected in a Goal-driven Learning Group. For a higher efficiency in this process, the Aho-Corasick algorithm [66], which is one of the famous multi-string

search algorithms in the pattern matching field, is employed. The algorithm to detect the goal-driven learning action pattern is expressed in Fig. 5-4.

5.4.2 Goal-driven Learning Recommendation Mechanism

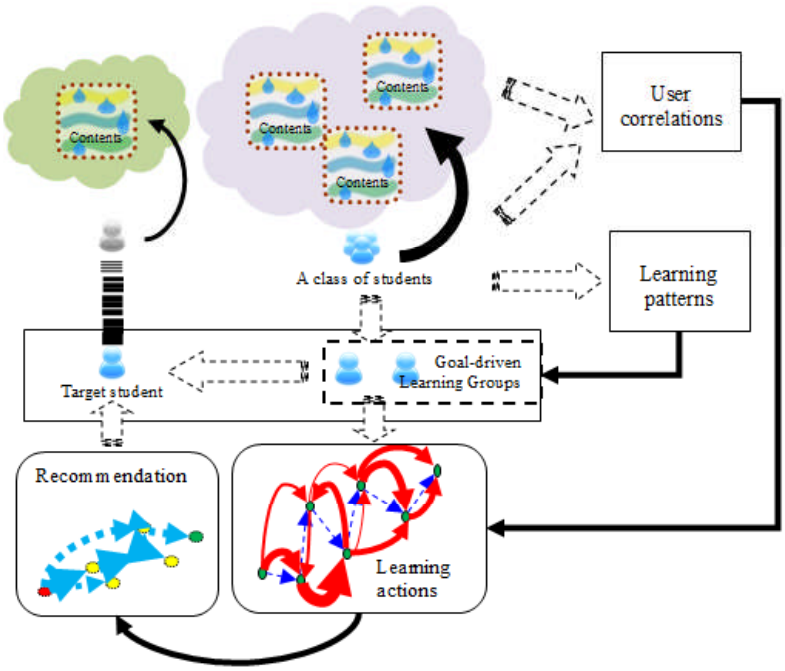


Figure 5-5 Conceptual Process of Learning Action Recommendation [7]

As shown in Fig. 5-5, both learning behavior pattern and user correlation are considered for the recommendation of learning action in a specific learning period. That is, the learning behavior pattern portion is taken into account of the information behavior factor which has been recorded as learning actions in the log data, while the user correlation portion indicates the interaction factor in both the direct and indirect way among a group of users within a learning course during a specific learning period. In details, the learning action patterns, which are used to describe the users' learning

behaviors in both the individual and collective way, can be extracted in accordance with different learning goals. Meanwhile, the user networking model can be constructed to represent users' potential and dynamical relationships based on the communication actions and posted contents, in which the specific correlations between the target user and other users can be further analyzed and extracted. Moreover, considering the detected goal-driven learning action patterns of a target user, three important weights: the weight which indicates the frequency of a LA-Pattern in a Goal-driven Learning Group, the weight which describes the users' relationships in the user networking model, and the weight which indicates the frequency of a detected goal-driven learning action pattern from the target user, can be used together to figure out the most suitable learning action from those similar users, which can be provided as the next learning step to serve the target user's specific learning purpose. Note that each recommended learning action refers to a target user for a specific learning goal within a selected learning period (e.g., one week for a lesson), which means the learning action we recommended is a specific action (e.g., view learning content regarding to the current lesson) following the current instructions, but not the generic action which will be suitable for the whole semester.

The formula to calculate the similar users is expressed as follows.

$$W_U = \gamma * \frac{\sum w_{DP_i} * w_{GP}(DP_i, u_j)}{\sum w_{DP}} + (1-\gamma) * \frac{CoD(u_j, u_i)}{\omega} \quad (5.8)$$

where, $w_{GP}(DP_i, u_j) = \frac{w_p}{\sum w_p}$

In Eq. (5.8), w_{DP} denotes the frequency-based weight for a detected goal-driven learning action pattern of the target user, $w_{GP}(DP_i, u_j)$ denotes the frequency-based weight for a LA-Pattern of user u_j in a Goal-driven Learning Group, $CoD(u_j, u_i)$ denotes the contribution degree calculated from the user networking model, and ω in the denominator is used for the normalization with a default value of 2.

For a target user u_i with the final learning action act_{u_i} in a given learning action sequence, Eq. (5.9) is used to calculate the weights of a set of learning actions that are selected from those similar users, which may further be inferred as the possible next learning actions.

$$W_{act} = \frac{\sum W_{U_j} * w_{na_i}}{\sum W_U} \tag{5.9}$$

where, w_{na_i} denotes the frequency-based weight of a learning action generated by those similar users, following the learning action act_{u_i}. That is, the learning actions with a higher weight will be recommended to the target user as the next learning action.

Based on these discussions above, the recommendation algorithm to provide the target user with the next possible learning action for the individualized learning support is described in Fig. 5-6.

Input: The target user u_i's current learning action act_{u_i}

A specific learning goal G_i

Output: The recommended next learning action $actnext_{u_i}$

Step 1: For the whole user group, calculate the LA-Patterns $\{< act_i >_u^w \rightarrow G\}$ for each user u

Step 2: For the specific learning goal G_i, generate the Goal-driven Learning Group LG: $\{[< Act_k >_{u_1,u_2,\ldots,u_n}^{w_1,w_2,\ldots,w_n}][G_i, T]\}$

Step 3: For the whole user group in a learning course, build the user networking model based on the weight $w_{ij} = \alpha * w_{ar} + \beta * w_{cr}$ in a selected timescale T.

Step 4: For the target user u_i, calculate the contribution degree $CoD(u_j, u_i) = w_{ji} * (\frac{1}{\sum_{l=1}^{n} w_{jl}} + \frac{1}{\sum_{k=1}^{n} w_{ki}})$

Step 5: For the target user u_i with his/her learning action sequence $s = < act_1, act_2, \ldots, act_{u_i} >$, generate the goal-driven learning action pattern set $DP = \{<act_i>^w\}$

Step 6: For user u_i's current learning action act_{u_i}, find the following learning actions from users in the Goal-driven Learning Group LG and record as $actnext = \{actnext_{u_a}, actnext_{u_b}, \ldots, actnext_{u_n}\}$

Step 7: Calculate the weight $W_U = \gamma * \frac{\sum w_{DP_i} * w_{GP}(DP_i, u_j)}{\sum w_{DP}} + (1 - \gamma) * \frac{CoD(u_j, u_i)}{\omega}$ for each user in the Goal-driven Learning Group LG, in order to find the similar users

Step 8: For each element $actnext_{u_n}$ in the list $actnext$, calculate the weight $W_{act} = \frac{\sum W_{U_j} * w_{na_i}}{\sum W_U}$

Step 9: Return $actnext_{u_j}$ with $Max(W_{act})$ to be the recommended next learning action

Figure 5-6 Algorithm for Learning Action Recommendation [7]

5.5 Experiments on Learning Recommendation

In this section, after introducing the architecture of the implemented system, we show and discuss the experiment results for the analysis of learning behaviors, in order to explain the practicability and usefulness of our behavior modeling. Based on these,

the evaluation results are given to demonstrate the applicability and effectiveness of our proposed recommendation method.

5.5.1 System Architecture of Task-Oriented Recommendation

The architecture for the individualized learning action recommendation is shown in Fig. 5-7. This system consists of seven major components: Learning Action Pattern Analyzer, Goal-driven Learning Group Generator, User Interaction Analyzer, User Intention Analyzer & Extractor, User Networking Builder, User Correlation Analyzer, and Individualized Action Recommender.

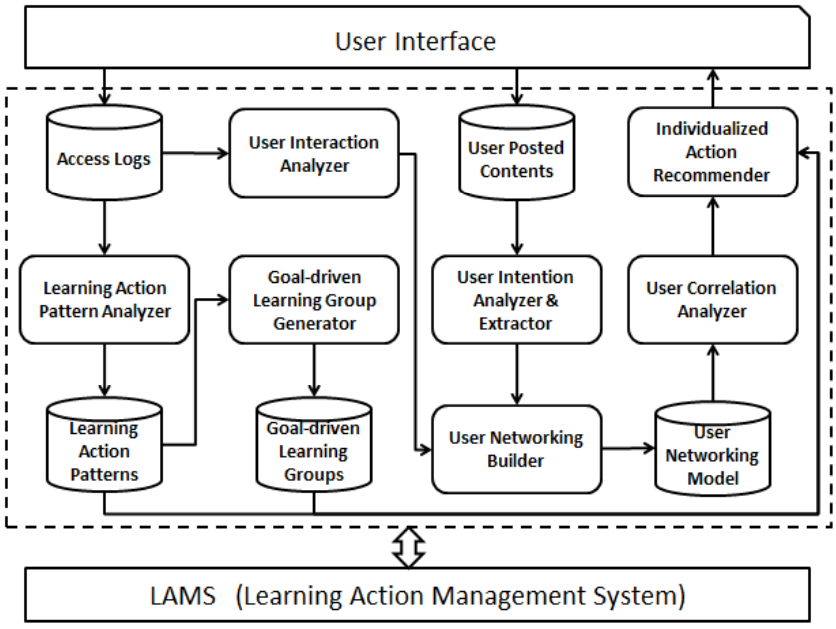


Figure 5-7 Architecture of Task-Oriented Recommendation System [7]

As for the user information behavior analysis, the Learning Action Pattern Analyzer is used to calculate the LA-Patterns according to the analysis of learning behaviors described by learning actions which are generated in a specific learning

course and stored in the Access Logs. The Goal-driven Learning Group Generator is employed to divide the users into several Goal-driven Learning Groups based on the similarity analysis of LA-Patterns. On the other hand, to calculate user correlations, the User Interaction Analyzer is integrated to compute users' direct communication behaviors, while the User Intention Analyzer & Extractor is responsible for the analysis and extraction of users' potential intentions. Based on these two modules, the User Networking Builder can construct the user networking model, which can be further applied to figure out the correlations among the target user and other users. By the Individualized Action Recommender, after comparing a list of potential learning actions, the most possible next learning actions can be calculated as a feedback sent to the target user, which can benefit the action recommendation process among the similar users. Finally, the Learning Action Management System (LAMS) manages all the learning actions and controls the whole recommendation process in this system.

We assume that there are no deviant users in our system, which means all users have used our system normally and generated no spam or irrelevant learning actions. Given that the assumption holds, we can ensure that all patterns contributed by learning action sequences are practicable and usable.

5.5.2 Data Set for Learning Action Experiments

In this study, as shown in Table 5-1, we define totally 18 kinds of learning actions in

our community-based (Moodle) learning system. This system has been used by 57 students to have the course named “Introduction to Information Processing” in a Japanese private university from September 2011 to January 2012. In order to facilitate the description in the following sections, we use 18 letters to represent them, which are also shown in Table 5-1.

Table 5-1 Learning Actions and Their Notations [7]

Learning Actions	Notation
view course introduction	a
view own learning history	c
lesson overview	d
view learning content	e
view posts in forum	g
discuss in forum	h
upload report in forum	i
update report in forum	j
delete report in forum	k
search in forum	l
View a quiz	o
start taking a quiz	p
refresh a quiz result	q
submit quiz result	r
review a quiz	s
view an assignment	u
upload a finished assignment	v
view user's profile	y

Totally, 12066 learning actions have been generated, in which, 13 kinds of learning actions have been generated among 18 kinds in total. In details, the learning action type *u*, view an assignment, containing 4465 learning action records, is the

most frequently used learning action, while the learning action type *k*, delete report in forum, containing only six learning action records, seems rarely to be used by users.

In our system, one course consists of 15 lessons. As for requirement, each student should conduct at least four actions for each lesson. Considering the absence, usually, a student should generate more than 60 actions to finish one course if not absent from a lesson. On the contrary, a student who generates too many actions (e.g., 600 actions) can be considered as a special case. In average, students generate 211 actions, and nearly four users generated less than 50 actions, which are 7, 35, 41 and 48 actions, while three users generated over 400 actions, which are 465, 668, and 732 actions.

5.5.3 LA-Pattern Analysis Results

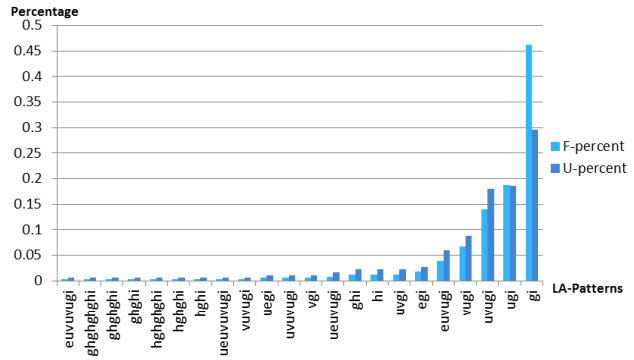
As discussed in Section 5.2.1, LA-Patterns are a series of sub-sequences of the whole learning action sequence, which end with some special learning actions as the learning goals. That is, the learning action with the characteristics that can be viewed as a specific learning purpose will be selected as the goal action manually. Thus, according to the classification of learning actions (see Table 5-1), in this study, we pre-define the following learning actions: *i* (upload report in forum), *j* (update report in forum), *p* (start taking a quiz), *q* (refresh a quiz result), *r* (submit quiz result), *v* (upload a finished assignment), to compose the learning goal set $G = \{i, j, p, q, r, v\}$,

which leads to six Goal-driven Learning Groups based on the similarity of LA-Patterns. That is, we consider the learning action sequences ending with these learning actions in set *G* to generate LA-Patterns and Goal-driven Learning Groups as well.

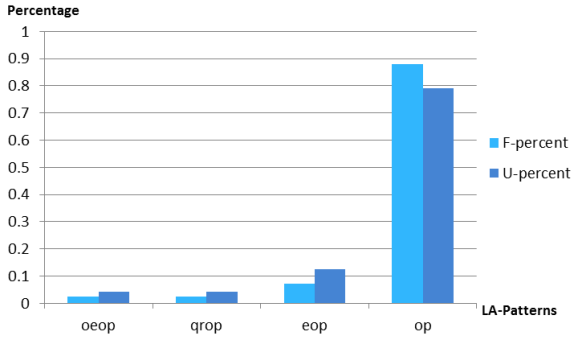
For each user in one lesson, the learning actions sequence generated following the timeline composes an *S-Task*. All these *S-Tasks* will be used to extract the learning action sub-sequences which may become the LA-Patterns. Accordingly, according to *Criteria 1 - basic criteria*, 838 learning action sub-sequences have been extracted, which can be categorized into six major types in details: *i* contains 215, *j* contains 1, *p* contains 29, *q* contains 78, *r* contains 49, and *v* contains 466, respectively. Consequentially, according to *Criteria 2 - incorporation criteria*, after incorporating these learning action sub-sequences containing each user, the six types of LA-Patterns, have been refined to 183, 1, 24, 51, 37, and 359, respectively. Finally, we have obtained totally 655 LA-Patterns from more than ten thousand learning actions, which can be further assigned into six Goal-driven Learning Groups.

We demonstrate some statistics for each pattern shown in Fig. 5-8, according to two important factors: the frequency a certain pattern occurs in all users' learning action sequences, and the number of users who have conducted a certain pattern. Note that the Goal-driven Learning Group- *j* has only one element: *hj* with the frequency 3,

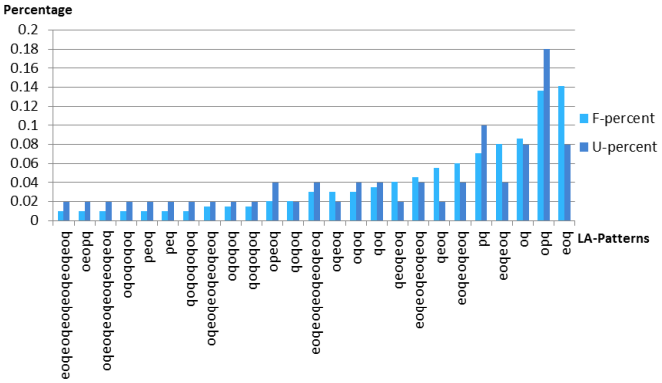
thus, we do not build a figure for it. Besides, in order to facilitate the further analysis, the values of both the frequency of each pattern and the number of users in each pattern have been converted into the percentage in each figure (a) to (e).



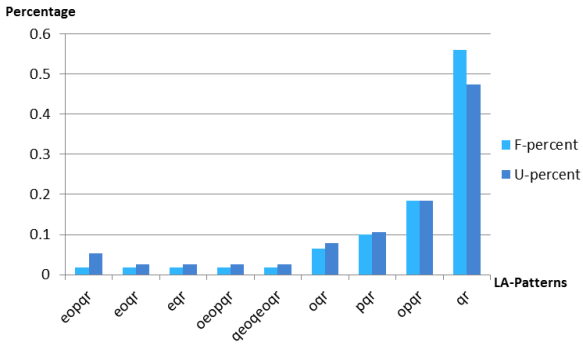
(a) Goal-driven Learning Group-*i*



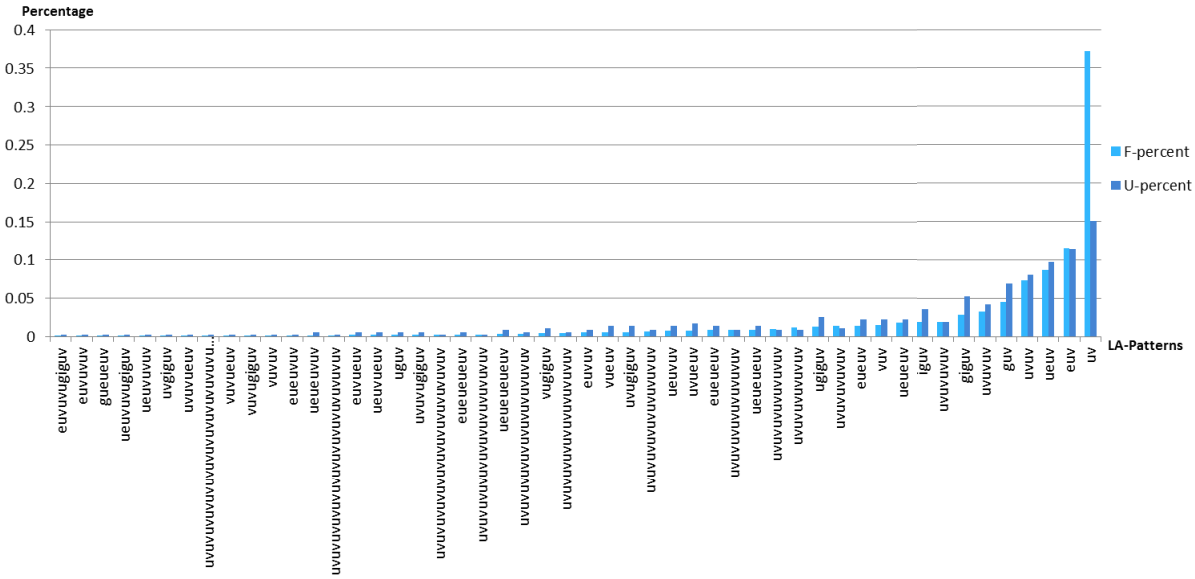
(b) Goal-driven Learning Group-*p*



(c) Goal-driven Learning Group-*q*



(d) Goal-driven Learning Group-*r*



(e) Goal-driven Learning Group-*v*

Figure 5-8 Statistics for Goal-driven Learning Groups [67]

As shown in Fig. 5-8 (a), 22 kinds of patterns have been categorized into this learning group-*i*. Considering both the frequency factor and user factor, the top three patterns in this group are *gi*, *ugi* and *uvugi*, which are 46% and 29%, 18% and 18%, 13% and 18% respectively. As shown in Fig. 5-8 (b), four kinds of patterns have been categorized into this learning group-*p*. The top one pattern *op*, occupies nearly 90% and 80% for the frequency factor and user factor respectively. As shown in Fig. 5-8 (c), 25 kinds of patterns have been categorized into this group-*q*. Differing with the former two learning groups, the top three rankings of patterns based on two factors are different. That is, according to the frequency factor, the top three ranking are *eoq*, *opq*, *oq*, and the percentage are 14%, 13% and 8%, while *opq*, *pq*, *eoq* occupy the top three positions in accordance with the user factor by 18%, 10% and 8%. As shown in Fig. 5-8 (d), nine kinds of patterns have been categorized into this group-*r*. Considering both the frequency factor and user factor, the top three patterns in this group are *qr*, *opqr* and *pqr*, which are 55% and 47%, 18% and 18%, 10% and 10% respectively. As shown in Fig. 5-8 (e), 55 kinds of patterns have been categorized into this group-*v*, which is the most among all six groups. Considering both the frequency factor and user factor, the top three patterns in this group are *uv*, *euv* and *ueuv*, which are 37% and 15%, 11% and 11%, 8% and 9% respectively.

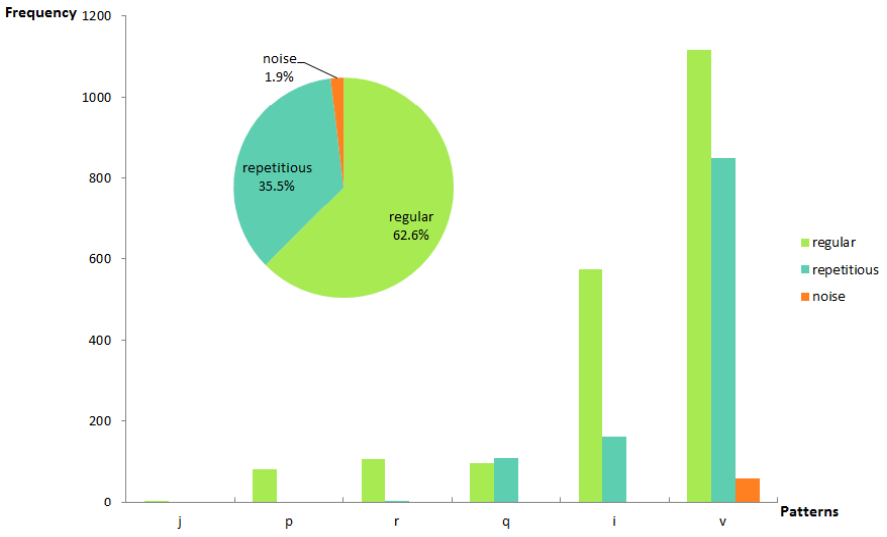


Figure 5-9 Statistics and Analysis for LA-Patterns in Each Group [7]

Among all the patterns in six groups, patterns in group-*v* occur more than two thousand times, while patterns in group-*j* occur three times. To further analyze the features of LA-Patterns, based on these statistics results, in each learning group, patterns can be divided into three categories: *regular*, *repetitious* and *noise*, which are shown in Fig. 5-9. More detailed descriptions are given as follows.

1) Regular patterns indicate these patterns that have no repeated action element in the sequence, which means each learning action in this pattern is unique.

2) Repetitious patterns indicate these patterns that have repeated action elements in the sequence, which means there is at least one learning action that has been done at least twice in this pattern.

3) Noise patterns indicate these patterns that are abnormal, which may be not correct or should not be recommended.

Note that only patterns in group-*v* and patterns in group-*p* contain some noise patterns, which is less than 2% among all patterns.

Based on these analyses, we can induce some useful insights as follows.

Generally, the shorter patterns may contain more users with higher frequency. For example, the *gi* in group-*i*, *op* in group-*p* and *uv* in group-*v*, which occupy nearly half in each group. These patterns can be viewed as a kind of common-use patterns or shortest patterns to complete a certain learning purpose. However, on the other hand, it does not mean only those patterns used by more users with higher frequency are useful. Moreover, some potential information (e.g., similarities among a small group of users) can be discovered and utilized from those patterns in spite of lower frequency. For instance, according to the patterns:

<*uvuvuvvuvuvuv*>,

<*uvuvuvvuvuvuvuv*>,

< *uvuvuvvuvuvuvuvuv*>,

there are always three users: User u_{15}, User u_{25} and User u_{27}, in these patterns. It indicates that these three users may have a sort of behavior similarities to achieve the same learning goal: upload a finished assignment. Thus, more related information should be shared within them to pursue higher learning efficiency.

Furthermore, as for the categories, *regular*, *repetitious* and *noise*, in each

Goal-driven Learning Group, holding the assumption in Section 5.5.1, the regular patterns can be viewed as basic patterns, which provide users with some basic steps as references to complete a certain learning goal, such as *ghi* in group-*i*, *opqr* in group-*r* and *euv* in group-*v*. The noise patterns refer to those patterns with none-recommended or incorrect sequence, such as *vueuv* in group-*v* and *qrop* in group-*p*. In these two groups respectively, we assume that in a well-defined LA-Pattern, learning action *v* cannot occur before *u*, and learning action *q* cannot occur before *p*, which means users should not upload a finished assignment before viewing it and users cannot redo a quiz before starting it. The repetitious patterns can be viewed as a positive means to better complete a certain learning purpose. Moreover, the so-called repeated factor can be extracted to facilitate the further learning action recommendation process. For instance, the sub-sequence *eoq*, which occurs multi-times in group-*q*, can become the repeated factor for this group. Then in the following recommendation process, when the learning action *e* is recommended to a specific user to refresh a quiz result, learning actions *o* and *q* can also be recommended to him/her.

Basically, our proposed methods mainly concentrate on the calculation of frequency of learning action sequences, rather than the meaning of each sequence. That is why some noise patterns have been extracted. However, the results discussed above certify that most of the LA-Patterns we extracted are meaningful and useful.

We mainly employed the frequency factor in the following recommendation process, and considered the meaning of each pattern as the secondary factor to calculate the weight of each learning action.

5.5.4 Evaluation

As for the learning action recommendation, for instance, in the situation discussed in Section 5.4.1, given User u_{16} with his/her recorded learning action sequence,

s = <*egigeuvuvugigueuguvgigeugueuvugigeuvugiguvuvuvuvuvueuvgiguvugig*>,

assume the learning goal is i (upload report in forum), according to the algorithm shown in Fig. 5-4, the calculated goal-driven learning action pattern set is

$P = \{< ugi >^4, < uvugi >^4, < vugi >^4, < uvgi >^2, < uvuvugi >^1, < gi >^7, < vgi >^2, < euvuvugi >^1, < ueuvugi >^1, < vuvugi >^1, < euvugi >^2, < egi >^1 \}$.

For other users, based on the analysis in the Goal-driven Learning Group-i, their frequency-based weights of each LA-Pattern, for example, $< uvgi >^{2,\ 2,\ 2,\ 3}_{u_{32},u_{41},u_{48},u_{67}}$, together with the frequency-based pattern weight of User u_{16}, can be used to figure out the similarity between two users according to their learning behaviors.

Considering that the user behavior patterns and user correlations have the same influence in the learning action recommendation process, we set γ to be 0.5 in Eq. (5.8). Finally, following Eq. (5.8) and Eq. (5.9), the learning actions with the top three weights will become the recommendation results for User u_{16} as the possible next

learning step after learning action *g* (view posts in forum) in this learning period, which are learning action *u* (view an assignment) with 0.36, *h* (discuss in forum) with 0.31 and *e* (view learning content) with 0.12.

In order to evaluate the usefulness of recommendation results, other 19 users have been selected to conduct the recommendation process mentioned above. To each user, we recommend them with three learning actions sequentially in accordance with their weight-based rankings. Finally, totally 60 learning actions are selected and provided to these 20 users respectively as their recommended next learning steps. Besides, in order to evaluate the recommendation results, the likert scale-based questionnaire, which is the most widely used approach to scaling responses in the survey research [68], is given to classify and demonstrate the usefulness of the recommendations, where the subject will be asked to rate the recommendation results in the likert scales varying from "*Very good*", "*Good*", "*Fair*", and "*Not good*". For the fairness, the instructor of this course was asked to conduct the evaluations and grade the recommended learning actions according to the four scales. The evaluation results are shown in Fig. 5-10. In addition to the recommendation results, these 20 users' original decisions are graded to make the comparisons.

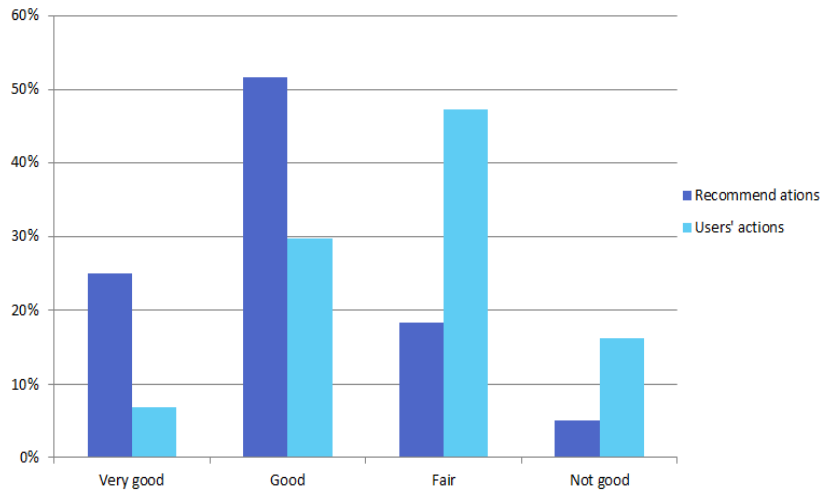


Figure 5-10 Evaluation Results in Comparison [7]

Basically, 95% of the recommended actions are considered as effective recommendations (among "*Very good*", "*Good*" and "*Fair*"). Moreover, comparing with users' original actions, over half of the recommended actions are "*Good*" which would lead to a high learning efficiency, while nearly half of users' original actions are "*Fair*" which would result in a low learning efficiency. In details, 25 percent recommended actions are "Very good" and 51.7 percent are "Good", which can be considered to be efficient and useful, while only 6.7 percent users' actions are "Very good" and 29.7 percent are "Good". On the contrary, less than 5 percent recommended actions are considered as unfavorable results, while users have made nearly 17.6 percent unfavorable decisions without the recommendations.

In addition to comparing the number of recommended actions with the number of users' original actions, the rankings of the recommendation results and the original

actions of 20 users, are taken into account for the further comparisons. That is, the user's own decisions, which are sorted by their frequencies and ranked in a sequence, are employed to compare with the recommendation results sequence sorted by their weights. The comparison results are categorized into "*Higher*", "*Same*", "*Lower*", and "*Extra*", which indicates whether the ranking of recommended actions is higher, same, lower, comparing with the rankings of users' original decisions, or the recommendation results are even new actions which have not occurred in users' decision sequences.

Based on these, in each category, we employ the four scales mentioned above to give the assessment for each learning action. That is, when a recommended learning action is assigned to "*Very good*", and has a "*Higher*" ranking than user's original decision, it means this user's choice is not good enough. And if this user follows this suggestion, he/she may achieve a better score with a higher efficiency. Thus, following this comparison, we can further demonstrate the effectiveness of our recommendation results which can better assist users' decision-making to pursue a better learning efficiency.

Finally, as shown in Fig. 5-11, 42% recommended actions achieve a higher ranking, while 25% hold their rankings as same as users' original ones. Besides, 18% actions' ranking decline, and 15% are new actions.

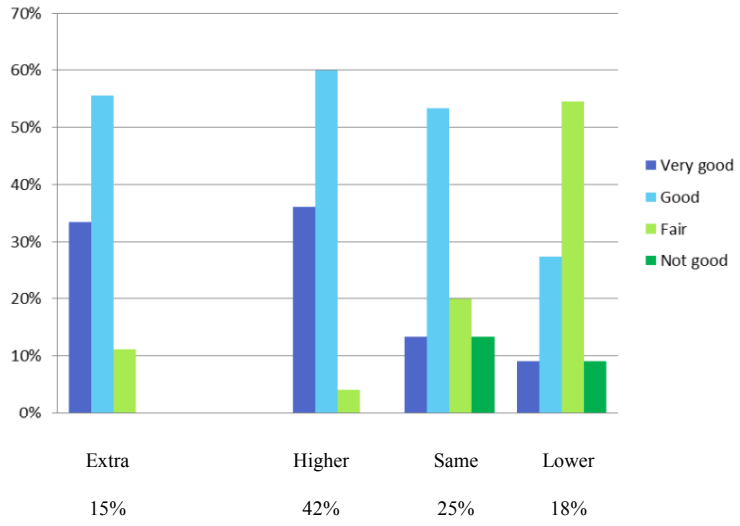


Figure 5-11 Usefulness Evaluation for Recommendation Results [7]

We give the following discussions as further analysis,

1) Among the "*Higher*" recommended actions, the efficient actions reach to nearly 96% (both "*Very good*" and "*Good*"). Moreover, 36% of them are "*Very good*" which means by following these steps, the users can achieve their learning purpose in a more efficient way.

2) Among the "Lower" recommended actions, the most actions are the "Fair" and "Not good" actions (totally nearly 64%), which means according to these suggestions, the users can rearrange or drop those unfavorable actions which may not be suitable for their current learning goal, in order to avoid an inefficient learning style.

3) As for the "Extra" recommended actions, which are new actions not in users' original decision sequence, over 88% of them are efficient actions (both "Very

good" and "Good"). These actions can be viewed as new perspectives to provide the users with more adaptive learning actions, in order to better pursue the learning purpose in a positive way.

Table 5-2 Values for Assessment of Recommended Actions [7]

	Very good	Good	Fair	Not good
Higher	2	1	-1	-2
Same	0	0	0	0
Lower	-2	-1	1	2
Extra	2	1	-1	-2

According to the comparisons discussed above, in order to illuminate the value of our recommendation more clearly, a weighted matrix is designed for the score-based assessment, which is shown in Table 5-2. That is, each recommended action will be assigned an additional weight (score) in accordance with the usefulness ("*Very good*", "*Good*", "*Fair*" and "*Not good*") and effectiveness ("*Higher*", "*Same*", "*Lower*" and "*Extra*") a .

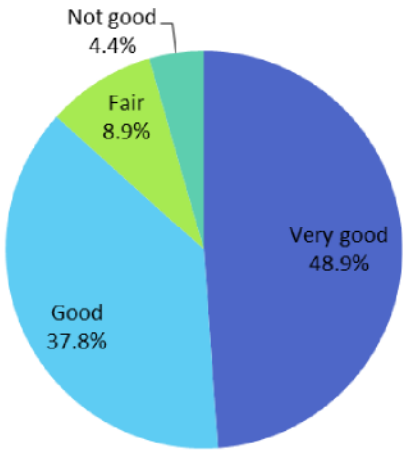


Figure 5-12 Assessment Results Based on Table 5-2 [7]

Finally, for each category of recommended actions, "*Very good*" scored totally 22 points, while "*Good*" scored 17, "*Fair*" scored 4, and "*Not good*" scored 2. For a clear illustration, the assessment results are shown in percentagewise in Fig. 5-12.

As a summary, all these results discussed above show that our recommendation method are practicable, and can effectively guide users to pursue their learning purposes in the task-oriented collaborative learning process.

5.6 Summary

In this chapter, we have presented an integrated method to model, discover and analyze users' learning behavior patterns and interaction-based correlations for the goal-driven learning action recommendations in a task-oriented learning process, in order to support the social knowledge utilization and task-oriented learning collaboration.

We have introduced a hierarchical model for the task-oriented collaborative learning process, in which the relationships among learning actions, learning activities, learning sub-tasks and learning tasks are described within a user community in an abstract level. Based on these, the LA-Pattern was proposed to discover and represent an individual user's learning behavior patterns which are described by the sequences of learning actions with their frequencies. Furthermore, the Goal-driven Learning Group was proposed to model a group of users' learning behaviors categorized by

different learning goals, in order to analyze the similarities among users in terms of learning behavior patterns. Two algorithms were developed to generate the LA-Pattern and Goal-driven Learning Group respectively. A goal-driven learning recommendation mechanism has been developed to utilize the LA-Patterns and user correlations to recommend the suitable learning actions to users as their next adaptive learning steps, which are expected to help users to complete a specific learning goal in a more efficient way.

The experimental results conducted in a community-based learning system have demonstrated that the LA-Patterns extracted by our algorithm can correctly describe the users' learning behaviors, and the further analysis based on the Goal-driven Learning Groups also showed that our method could be applied to frequency-based learning pattern recognition and categorization according to different learning goals, which can benefit the analysis of task-oriented learning behaviors with high usability and practicability. The evaluations based on the empirical results demonstrated the usefulness and effectiveness of our proposed recommendation method, which can support and facilitate the task-oriented collaborative learning processes.

Chapter 6 Conclusions

In this chapter, after a brief review of this thesis, we summarize the feature of this study. The challenging issues are then discussed. The possible future works are addressed in the end.

6.1 Summary of this Study

In this thesis, a computational approach to modeling and analyzing the personal data and individual behaviors is proposed. The unified models and integrated mechanisms are developed and applied to assist the individualized information utilization for both individuals and communities. Based on the analysis results of the personal data and individual behaviors, a user networking model is constructed to build the multi-dimensional user profiling and discover the dynamical social communities. And a recommendation mechanism is developed to provide the personalized learning guidance and support in the task-oriented processes based on a comprehensive consideration of behavior patterns and user correlations.

To systematically model and describe the personal data and individual behaviors, we concentrate on the methodical organization of personal data, and the automated identification of action patterns from the sequential individual behaviors. In details:

- The *Organic Streams* are proposed as a unified framework to formally organize and represent the personal big data, in which three basic relations are defined to flexibly describe the inherent and potential relationships among the data, and methodically organize the raw stream data into an associatively organized form.
- A heuristic mechanism is developed to capture the users' time-varying interests or needs which are represented as the *heuristic stones*, and further aggregate and integrate the relevant data together in the *associative ripples* to obtain the associative information based on the individual needs.
- A behavioral analysis method is proposed to model the individual behaviors in the task-oriented processes with formal descriptions. The action patterns are modeled and extracted based on the calculation of an individual user's sequential behaviors toward a certain purpose, and the behavioral similarities among a group of users are then analyzed and described based on the extracted patterns.

To facilitate the individualized information utilization, we delve into the analyzing and discovering of potential user correlations and dynamical user profiling in accordance with the outcomes from the analysis of the personal data with the individual behaviors, to provide users with more favorable users and communities, which can be viewed as a viable alternative way to obtain the larger information

resources. In details:

- A *DSUN* (Dynamically Socialized User Networking) model, which considers a combination of both the characteristics-based relationship and influence-based relationship, is constructed to connect more related people together by measuring their dynamical and potential correlations.
- A method is proposed to build and analyze the multi-dimensional user profiling in accordance with a set of attributes, which can help find the favorable users to facilitate a target user's information seeking in both global (e.g., *hub user* and *promotion user*) and personal (e.g., *contribution user* and *reference user*) way.
- A mechanism is developed to discover and represent three basic types of ties based on users' dynamical correlations (e.g., *strong correlation-based tie* and *weak correlation-based tie*) and profiling (e.g., *user profiling-based tie*) respectively, which can recommend users to join different social communities to satisfy their different requirements.

Furthermore, as an application of the proposed methods, an integrated recommendation method is developed, in which the behavior patterns and user correlations are taken into account to better facilitate the learning experience sharing and learning collaboration in the web-based learning environment. In details:

- A hierarchical model is addressed to describe the relations among learning actions, activities, sub-tasks and tasks in a user community for the task-oriented learning process.
- A learning behavior modeling is proposed, which includes the *LA-Pattern* (Learning Action Pattern) to discover and represent an individual user's learning behavior patterns extracted from sequences of learning actions, and the *Goal-driven Learning Group* to analyze and describe the similarities of learning behaviors among a group of users.
- An integrated mechanism is developed for the goal-driven learning recommendation based on the analysis of learning behaviors and user correlations, which can provide a target user with the next possible learning actions for the individualized learning support.

Two experimental studies are conducted respectively to demonstrate the feasibility of our proposed methods. An application prototype system has been designed and implemented to demonstrate the high usability and practicability of the *DSUN* model using the Twitter data. The experimental results based on the calculations of the attributes for user profiling illuminate that the favorable users can be efficiently identified to support a specific user in both global and individual way. While the experimental results based on the extractions of ties for user community

discovery demonstrate that our mechanisms can discover the user correlation and profiling based communities dynamically in the different time periods.

The evaluations have been conducted in a community-based (Moodle) learning system to illustrate the usefulness and effectiveness of our proposed recommendation method. The experimental results demonstrate that the *LA-Patterns* and the *Goal-driven Learning Groups* can correctly recognize and categorize the frequency-based learning patterns in the task-oriented learning processes. And the evaluation results illustrate that the mechanism for the progressive recommendations can assist users to complete a specific learning goal in a more efficient way, by providing the suitable learning actions as their next adaptive learning steps.

As a summary, the features of this study can be concluded as:

(1) Associative organization of personal data for personalized information utilization

The raw personal data is methodically and associatively aggregated and integrated into an organized form based on their inherent logicality and potential relationships. Users can obtain the associative information that fits their time-varying interests or needs in a heuristic way.

(2) Automated detecting of task-oriented action patterns to facilitate personal experience utilization

The task-oriented action patterns are modeled and extracted from users' sequential behaviors toward a certain purpose. Based on these, the user experience hidden in a series of individual behaviors represented by a sequence of actions can be shared according to the similarity of action patterns.

(3) Multi-dimensional attributes and measures for dynamical user profiling

A series of attributes based on the analysis of individuals' information behaviors, and a set of measures based on the analysis of users' correlations are defined and calculate to build the multi-dimensional user profiling for both information seeking and sharing support.

(4) Discovery of multi-types of social communities for promoting of information utilization and sharing

The dynamical user profiling and potential correlations are considered to discover the social communities from multiple perspectives. Users can be recommended into different types of communities to obtain larger information sources.

(5) Combination of behavior patterns and user correlations for progressive recommendation

The behavior patterns and the user correlations are taken into account together to calculate the similarities among a group of users. Based on these, the

progressive recommendation will then provide the target user with the next suitable action toward a certain purpose.

We highly expect the contributions of this study can facilitate the individualized information utilization from chaotic data to associative information, and further to connected people, which can benefit both individuals and communities not only for the personalized information seeking and recommendation, but also for the information sharing and social knowledge creation.

6.2 The Limitations

However, there are some limitations and unsolved issues.

- Qualities of the data set

Part of the experiment results particularly depend on the quality of the data set. Due to this reason, the experiment of strong tie-based community discovery in the *DSUN* model only resulted in few of user groups.

- Pre-definitions and pre-processing

As for the recommendation for the collaborative learning support, the learning goal actions are required to be pre-defined by the instructor in a learning system, which limits the scale of the discovery of learning patterns. Besides, the whole learning action sequence needs to be preprocessed into several sub-sequences as the input before generating the patterns, and the different granularities will lead to

different results.

- Weight setting

The weights of the coefficients are set in an ideal way (we treat the elements equally in the formulas). More experiments should be conducted to optimize the coefficients to adapt the more complex situations.

6.3 Future Works

As for future work, in addition to overcome the limitations mentioned above, we will develop and improve our system with the refined mechanisms to provide more flexible and adaptive services for the utilization of the associative information and social knowledge from more extensive collections of the cooperative and pervasive data in both cyber and physical world. Performance evaluation experiment will be conducted to improve our proposed methods and system for better individualized utilization. We will also consider developing the algorithms and mechanisms to realize the sustainable information utilization, and extract the structured knowledge to increase and maximize the value of data.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Qun Jin who has always been kindly supporting and encouraging me through my academic life during the last four years. I would also like to express my gratitude to Professor Nishimura, Professor Kikuchi and Professor Ozawa for their kind advice and support upon the completion of this thesis.

Besides, I would like to express my deepest appreciation to my parents and friends who have always provided me with spiritual support throughout my life.

I am also thankful to all the colleagues and students in the Networked Information Systems Laboratory who have participated in discussions and have collaborated with me.

Bibliography

[1] C.Q. Ji, Y. Li, W.M. Qiu, U. Awada, and K.Q. Li, "Big Data Processing in Cloud Computing Environments," in *Proc. 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN)*, Dec. 13-15, 2012, pp.17-23.

[2] "Big data: science in the petabyte era," *Nature* 455 (7209):1, 2008.

[3] M. A. Beyer and D. Laney, "The Importance of 'Big Data': A Definition," 2012.

[4] R. J. Todd, "Back to Our Beginnings: Information Utilization, Bertram Brookes and The Fundamental Equation of Information Science," *Information Processing & Management*, vol. 35, no. 6, pp. 851–870, Nov. 1999.

[5] K. Shilton, "Four Billion Little Brothers?: Privacy, Mobile Phones, and Ubiquitous Data Collection," *Communications of the ACM*, vol. 52, no. 11, pp.48-53, Nov. 2009.

[6] X. Zhou, N.Y. Yen, Q. Jin and T.K. Shih, "Enriching User Search Experience by Mining Social Streams with Heuristic Stones and Associative Ripples," *Multimedia Tools and Applications (Springer)*, vol.63, no.1, pp.129-144, Mar. 2013.

[7] X. Zhou, J. Chen, B. Wu and Q. Jin, "Discovery of Action Patterns and User Correlations in Task-Oriented Processes for Goal-Driven Learning Recommendation," *IEEE Transactions on Learning Technologies*, no.99, 2014.

[8] X. Zhou and Q. Jin, "A Heuristic Approach to Discovering User Correlations from Organized Social Stream Data," *Multimedia Tools and Applications (Springer)*,

Accepted.

[9] X. Zhou, W. Wang and Q. Jin, “Multi-Dimensional Attributes and Measures for Dynamical User Profiling in Social Networking Environments,” *Multimedia Tools and Applications (Springer)*, Accept with Minor Revision.

[10] A. Signorini, A.M. Segre and P.M Polgreen, “The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US during the Influenza A H1N1 Pandemic,” *PLoS ONE*, vol. 6, no. 5, May. 2011.

[11] R. Junco, G. Heiberger and E. Loken, “The effect of Twitter on College Student Engagement and Grades,” *Journal of Computer Assisted Learning*, vol. 27, no. 2, pp. 119-132, Apr. 2011.

[12] A.J. Kirsten, “The Effect of Twitter Posts on Students' Perceptions of Instructor Credibility,” *Learning Media and Technology*, vol. 36, no.1, pp. 21-38, Mar. 2011.

[13] A. Black, C. Mascaro, M. Gallagher and S.P. Goggins, “Twitter Zombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere,” in *Proc. the 17th ACM international conference on Supporting group work (GROUP '12)*, Sanibel Island, USA, Oct. 27-31, 2012, pp.229-238.

[14] C. Byun, Y. Kim, H. Lee and K.K. Kim, “Automated Twitter Data Collecting Tool and Case Study with Rule-Based Analysis,” in *Proc. the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS '12)*, Bali,

Indonesia, Dec. 3-5, 2012, pp.196-204.

[15] X. Wang, F. Wei, X. Liu, M. Zhou and M. Zhang, “Topic Sentiment Analysis in Twitter: A Graph-Based Hashtag Sentiment Classification Approach,” in *Proc. the 20th ACM international conference on Information and knowledge management (CIKM '11)*, Glasgow, United Kingdom, Oct. 24-28, 2011, pp.1031-1040.

[16] L. Kendall, A. Hartzler, P. Klasnja and W. Pratt, “Descriptive Analysis of Physical Activity Conversations on Twitter,” in *Proc. CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*, Vancouver, BC, Canada, May 7-12, 2011, pp.1555-1560.

[17] P. Cogan, M. Andrews, M. Bradonjic, W.S. Kennedy, A. Sala and G. Tucci, “Reconstruction and Analysis of Twitter Conversation Graphs,” in *Proc. the 1st ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial '12)*, Beijing, China, Aug. 12-16, 2012, pp.25-31.

[18] J. Vosecky, D. Jiang and W.N. Limosa, “A System for Geographic User Interest Analysis in Twitter,” *in: Proc. the 16th International Conference on Extending Database Technology (EDBT '13)*, Genoa, Italy, Mar. 18-22, 2013, pp.709-712.

[19] N. Pervin, F. Fang, A. Datta, K. Dutta and D. Vandermeer, “Fast, Scalable, and Context-Sensitive Detection of Trending Topics in Microblog Post Streams,” *ACM Trans. Management Information Systems (TMIS)*, vol. 3, no. 4, article 19, Jan. 2013.

[20] M. Yamagiwa, M. Uehara, and M. Murakami, “Applied System of The Social Life Log for Ecological Lifestyle in The Home,” in *Proc. International Conference on Network-Based Information Systems (NBIS '09)*, Indianapolis, USA, Aug. 19-21, 2009, pp. 457-462.

[21] T. Hori, and K. Aizawa, “Capturing Life-Log and Retrieval Based on Contexts,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME '04)*, Jun. 27-30, 2004, pp. 301-304.

[22] K.S. Hwang, and S.B. Cho, “Life Log Management Based on Machine Learning Technique,” in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Seoul, Korea, Aug. 20-22, 2008, pp.691-696.

[23] H.H. Kang, C. H. Song, Y.C. Kim, S.J. Yoo, D. Han, and H.G. Kim, “Metadata for Efficient Storage and Retrieval of Life Log Dedia,” in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Seoul, Korea, Aug. 20-22, 2008, pp. 687-690.

[24] A. Shimojo, S. Matsumoto, and M. Nakamura, “Implementing and Evaluating Life-Log Mashup Platform Using RDB and Web Services,” in *Proc. the 13th International Conference on Information Integration and Web-based Applications and Services (iiWAS '11)*, Ho Chi Minh City, Vietnam, Dec. 5-7, 2011, pp. 503-506.

[25] A. Nakamura and N. Nishio, “User Profile Generation Reflecting User's Temporal

Preference through Web Life-Log," in *Proc. the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*, Pittsburgh, Pennsylvania, USA, Sep. 5-8, 2012, pp. 615-616.

[26] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines and F. Menczer, "Friendship Prediction and Homophily in Social Media," *ACM Trans. the Web (TWEB)*, vol. 6, no. 2, article 9, Jun. 2012.

[27] J. Yu, X. Jin, J. Han and J. Luo, "Collection-Based Sparse Label Propagation and Its Application on Social Group Suggestion from Photos," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, article 12, Feb. 2011.

[28] L. M. Aiello, A. Barrat, C. Cattuto, G. Ruffo and R. Schifanella, "Link Creation and Profile Alignment in the aNobii Social Network," in *Proc. IEEE Second International Conference on Social Computing (SocialCom)*, Minneapolis, MN, USA, Aug. 20-22, 2010, pp. 249-256.

[29] R. Xiang, J. Neville and M. Rogati., "Modeling Relationship Strength in Online Social Networks," in *Proc. the 19th International Conference on World Wide Web (WWW '10)*, Raleigh, NC, USA, Apr. 26-30, 2010, pp. 981-990.

[30] V. Leroy, B. B. Cambazoglu and F. Bonchi, "Cold Start Link Prediction," in *Proc. the 16th ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD '10)*, Washington DC, DC, USA, Jul. 25-28, 2010, pp. 393-402.

[31] C. Wilson, A. Sala, K. P. N. Puttaswamy and B. Y. Zhao, “Beyond Social Graphs: User Interactions in Online Social Networks and their Implications,” *ACM Trans. the Web(TWEB)*, vol. 6, no. 4, article 17, Nov. 2012.

[32] L. Tang, X. Wang and H. Liu, “Group Profiling for Understanding Social Structures,” *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 3, no. 1, article 15, Oct. 2011.

[33] Y. Zheng, L. Zhang, Z. Ma, X. Xie and W. Ma, “Recommending Friends and Locations Based on Individual Location History,” *ACM Trans. the Web(TWEB)* vol. 5, no. 1, article 5, Feb. 2011.

[34] D. M. Romero, W. Galuba, S. Asur and B. A. Huberman, “Influence and Passivity in Social Media,” *Machine Learning and Knowledge Discovery in Databases*, *LNCS*, vol. 6913, pp. 18-33, 2011.

[35] J. Tang, J. Sun, C.i Wang and Z. Yang, “Social Influence Analysis in Large-Scale Networks,” in *Proc. the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, Paris, France, Jun. 28 - Jul. 1, 2009, pp. 807-816.

[36] J. Sang and C. Xu, “Social Influence Analysis and Application on Multimedia Sharing Websites,” *ACM Trans. Multimedia Computing. Communications, and Applications (TOMM),* vol. 9, no. 1s, article 53, Oct. 2013.

[37] P. Achananuparp, E. P. Lim, J. Jiang and T.A. Hoang, “Who is Retweeting the Tweeters? Modeling, Originating, and Promoting Behaviors in the Twitter Network,” *ACM Trans. Management Information Systems (TMIS)*, vol. 3, no. 3, article 13, Oct. 2012.

[38] X. Tang and C. C. Yang, “Ranking User Influence in Healthcare Social Media,” *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, article 73, Sep. 2012.

[39] N. Ronald, V. Dignum and C. M. Jonker, “When Will I See You Again: Modeling The Influence of Social Networks on Social Activities,” in *Proc. the Multi-Agent Logics, Languages, and Organisations Federated Workshops (MALLOW)*, Lyon, France, Aug. 30 - Sep. 2, 2010.

[40] M. Gomez-Rodriguez, J. Leskovec and A. Krause, “Inferring Networks of Diffusion and Influence,” *ACM Trans. Knowledge Discovery from Data (TKDD)*, vol. 5, no. 4, article 21, Feb. 2012.

[41] Y. R. Lin, J. Sun, H. Sundaram, A. Kelliher, P.l Castro and R. Konuru, “Community Discovery via Metagraph Factorization,” *ACM Trans. Knowledge Discovery from Data (TKDD)*,vol. 5, no. 3, article 17, Aug. 2011.

[42] J. Leskovec and E. Horvitz, “Planetary-Scale Views on A Large Instant-Messaging Network,” in *Proc. the International Conference on World Wide Web (W)*, Beijing, China, Apr. 21-25, 2008, pp. 915-924.

[43] Z. Yin, L. Cao, Q. Gu and J. Han, "Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, article 63, Sep. 2012.

[44] R. Goolsby, "Social Media as Crisis Platform: The Future of Community Maps/Crisis Maps," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 1, no. 1, article 7, Oct. 2010.

[45] Z. Zhang, Q. Li, D. Zeng and H. Gao, "User Community Discovery from Multi-Relational Networks," *Decision Support Systems*, vol. 54, no. 2, pp. 870-879, Jan. 2013.

[46] G. Paliouras, "Discovery of Web User Communities and Their Role in Personalization," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 151-175, Apr. 2012.

[47] L. Razmerita, "An Ontology-Based Framework for Modeling User Behavior—A Case Study in Knowledge Management," *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans,* vol. 41, no. 4, pp. 772-783, Jul. 2011.

[48] S. J. Stolfo, S. Hershkop, C. W. Hu, W. J. Li, O. Nimeskern and K. Wang, "Behavior-Based Modeling and Its Application to Email Analysis," *ACM Trans. Internet Technology*, vol. 6, no. 2, pp. 187-221, May. 2006.

[49] T. S. Chen, Y. S. Chou, T. C. Chen, "Mining User Movement Behavior Patterns in A Mobile Service Environment," *IEEE Trans. Systems, Man and Cybernetics, Part A:*

Systems and Humans, vol. 42, no. 1, pp. 87-101, Jan. 2012.

[50] Z. Y. Liu, Y. B. Zheng, L. X. Xie, M. S. Sun, L. Y. Ru and Y. Zhang, "User Behaviors in Related Word Retrieval and New Word Detection: A Collaborative Perspective," *ACM Trans. Asian Language Information Processing (TALIP)*, vol. 10, no. 4, article 20, Dec. 2011.

[51] S. W. Lee, Y. S. Kim, Z, Bien, "A Nonsupervised Learning Framework of Human Behavior Patterns Based on Sequential Actions," *IEEE Trans. Knowledge and Data Engineering*, vol.22, no.4, pp.479,492, April 2010.

[52] Ching-Huang Yun, Ming-Syan Chen, "Mining Mobile Sequential Patterns in A Mobile Commerce Environment," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 37, no. 2, pp. 278-295, Mar. 2007.

[53] M. Munoz-Organero, P. J. Munoz-Merino, C. D. Kloos, "Student Behavior and Interaction Patterns With An LMS as Motivation Predictors in E-Learning Settings," *IEEE Trans. Education*, vol. 53, no. 3, pp. 463-470, Aug. 2010.

[54] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire and Y. W. Choong, "Mining Multidimensional and Multilevel Sequential Patterns," *ACM Trans. Knowledge Discovery from Data*, vol. 4, no. 1, article 4, Jan. 2010.

[55] Z. W. Zhao and W. T. Ooi, "APRICOD: An Access-Pattern-Driven Distributed Caching Middleware for Fast Content Discovery of Noncontinuous Media Access," *ACM Trans.*

Multimedia Computing Communications and Applications (TOMCCAP), vol. 9, no. 2, article 15, May 2013.

[56] X. Zhou, J. Chen, Q. Jin and T.K. Shih, "Organic Stream: Meaningfully Organized Social Stream for Individualized Information Seeking and Knowledge Mining," in *Proc. the 5th IET International Conference on Ubi- Media Computing (U-Media2012)*, Xining, China, Aug. 16-18, 2012. (Best Paper Award)

[57] E. Fredkin, "Trie Memory," *Communications of the ACM*, vol. 3, no. 9, pp. 490-499, Sep. 1960.

[58] J. A. Iglesias, P. Angelov, A. Ledezma and A. Sanchis, "Creating Evolving User Behavior Profiles Automatically," *IEEE Trans. Knowledge and Data Engineering*, vol. 24, no. 5, pp. 854-867, May 2012.

[59] M. Mcpherson, L. Smith-Lovin and J. M. Cook, "Birds of A Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415-444, Aug. 2001.

[60] M. Granovetter, "The Strength of Weak Ties: A Network Theory Revisited," *Sociological Theory* vol. 1 pp. 201–233, 1983.

[61] G. P. Barbier, "Finding Provenance Data in Social Media," Ph.D. Dissertation. Arizona State University, Tempe, AZ, USA. Advisor(s) Huan Liu. AAI3482460.

[62] N. E. Friedkin, "A Structural Theory of Social Influence," Cambridge University Press, 1998.

[63] P. M. DeMarzo, D. Vayanos and J. Zwiebel, “Persuasion Bias, Social Influence, and Unidimensional Opinions,” *Quarterly Journal of Economics* vol. 118, no. 3, pp. 909–968, Aug. 2003.

[64] M. O. Jackson and B. Golub, “Naive Learning in Social Networks: Convergence, Influence and Wisdom of Crowds,” *Fondazione Eni Enrico Mattei Working Paper*, Jul. 2007.

[65] J. Yang and J. Leskovec, “Modeling Information Diffusion in Implicit Networks,” in *Proc. the 2010 IEEE International Conference on Data Mining (ICDM '10)*, Sydney, Australia, Dec. 14-17, pp. 599-608.

[66] A. V. Aho and M. J. Corasick, “Efficient String Matching: An Aid to Bibliographic Search,” *Communications of the ACM*, vol. 18, no. 6, pp. 333-340, Jun. 1975.

[67] X. Zhou and Q. Jin, “Analysis of Sharable Learning Processes and Action Patterns for Adaptive Learning Support,” in *Proc. the 13th International Conference on Web-based Learning (ICWL2014)*, *LNCS*, Tallinn, Estonia, August 14-17, 2014.

[68] R. Likert, “A technique for the measurement of attitudes,” *Archives of Psychology*, vol. 22, no. 140, pp. 5-55, Jun. 1932.