Inf Syst Front (2010) 12:607–629
DOI 10.1007/s10796-009-9199-3

Personalisation in web computing and informatics: Theories, techniques, applications, and future research

Min Gao · Kecheng Liu · Zhongfu Wu

Published online: 21 July 2009

Abstract Recently, personalised search engines and recommendation systems have been widely adopted by users who require assistance in searching, classifying, and filtering information. This paper presents an overview of the field of personalisation systems and describes current state-of-the-art methods and techniques. It reviews approaches for (1) user profiling, including behaviour, preference, and intention modelling; (2) content modelling, comprising content representation, analysis, and classification; and (3) filtering methods for recommendation, classified into four main categories: rule-based, content-based, collaborative, and hybrid filtering. The paper also discusses personalisation systems in different domains, and various techniques and their limitations. Finally, it identifies several issues and possible directions for further research that can improve recommendation capabilities and enhance personalised systems.

Keywords Personalisation · Recommendation · User profile · Information filtering

1 Introduction

The volume of information over the Internet is increasing at a tremendous rate. Users are usually confronted with situations in which they have been exposed with too many options to choose from. They need help to explore and to filter out irrelevant information based on their preferences (Montaner et al. 2003). The corresponding requirement is to describe and capture relevant information for directing users based on specific interests and needs (Chedrawy and Abidi 2006). This support is identified as personalisation.

Personalisation is defined as automatic adjustment, re-structuring, and the presentation of tailored information content for individuals (Perugini and Ramakrishnan 2003). It builds customer loyalty by creating meaningful one-to-one relationships and understanding user needs in different contexts (Riecken 2000, Frias-Martinez et al. 2006). It is a process of gathering and analysing user information for delivering the right information at the right time (Chiu 2001).

In summary, personalisation is the ability to provide tailored content and services to individuals based on knowledge about their preferences and behaviour (Kuo and Chen 2001, Liang et al. 2007, Adomavicius and Tuzhilin 2005). Lots of research has been done in the personalisation domain, and there are several surveys and reviews of this research. In this paper, we have reviewed them in a different way.

This work was supported by the National Social Sciences Foundation of China under Grant No. ACA07004-08, Postdoctoral Science Foundation of China under Grant No. 20080440699, Projects of Chongqing under Grant No. 2008BB2183, KJ071601 and 2008-ZJ-064, conducted while M. Gao was visiting scholar at the University of Reading from Nov. 2007 to Nov. 2008, sponsored by China Scholarship Council

M. Gao (✉)
College of Computer Science, Chongqing University,
Chongqing 400044, China
e-mail: gaomin@cqu.edu.cn

K. Liu
Informatics Research Centre, University of Reading,
Reading RG66WB, UK
e-mail: k.liu@reading.ac.uk

Z. Wu
College of Computer Science, Chongqing University,
Chongqing 400044, China
e-mail: wzf@cqu.edu.cn

This paper firstly analyses the trends and topology of personalisation, and then presents surveys on the applications, methodologies, and techniques of personalisation in web computing and informatics. The contribution of the paper includes three aspects. Firstly, a novel classification schema is summarised for the survey of the previous personalisation literature. The schema includes three essential parts of personalisation: user profiling, content modelling, and information filtering. Secondly, it presents a comprehensive discussion on the methodologies and the techniques to achieve the three parts in detail. Thirdly, it identifies several issues and possible directions for further research on personalisation and recommendation systems. In brief, they are the analyses of contextual information, user groups, and domain knowledge for personalisation systems (also called customisation, personalised recommendation system). The paper provides a strong foundation for researchers to construct personalised systems.

In the following sections, we analyse the trends and topology of personalisation (Section 2), review personalisation domains and applications (Section 3), analyse the state of the art in personalisation methods (Section 4) and techniques (Section 5), identify several directions and issues for further research (Section 6), and draw a conclusion by identifying research directions and issues in Section 7.

2 Trends of personalisation and review methodology

The target of this section is to explain (1) the trends of personalisation through an analysis of the publications in a variety of authoritative databases on personalisation up to 2008 and (2) The review methodology of the paper.

2.1 Trends of personalisation

To explain the trends of the development of personalisation in recent years, we analysed the popularity of the field of research on personalisation up to 2008 if they addressed the issue of personalisation on theories, techniques, and applications. As can be seen from Fig. 1, the four curves show changes in the amount of publications with respect to personalisation from 2000 to 2007. The numbers in the *x* axis are publication years. The numbers in the *y* axis are the number of publications. The curves represent the publication information on personalisation in the most authoritative databases of science and technology, including (1) IEEE (Institute of Electrical and Electronics Engineers) (2) ACM (Association for Computing Machinery), (3) Springer, and (4) Science Direct. We used “personalisation”, “customise”, or “personalised” as the keywords (index items in IEEE), and used “computer science” and “engineering” as the selected subjects for the searches in these databases.

The numbers of articles during the past eight years are depicted in Fig. 1. The curves are mostly similar in showing a gradually increasing trend in the number of papers on personalisation. The increase is more noticeable during the last three year period. It appears that personalisation has become more important, and increasingly more researchers have worked on the field in recent years. Except reviews and cure algorithmic researches, several domains, which include digital library, e-commerce, e-learning, news recommendation system, and search engine, are involved in these publications. Because the demand for personalisation and recommendation systems is keeping increasing recent year, more and more researchers have been focusing on this topic, and the personalisation has been identified as the biggest thing in the next ten years in these domains (Weller 2007, Zhang 2007).

2.2 Review methodology

There are already many surveys of personalisation. For instance, Eirinaki and Vazirgiannis (2003) reviewed user profiling and Web usage mining processes and methods in e-commerce. Murthi and Sarkar (2003) reviewed personalisation from two aspects: the role of personalisation in firm value systems and user preference learning

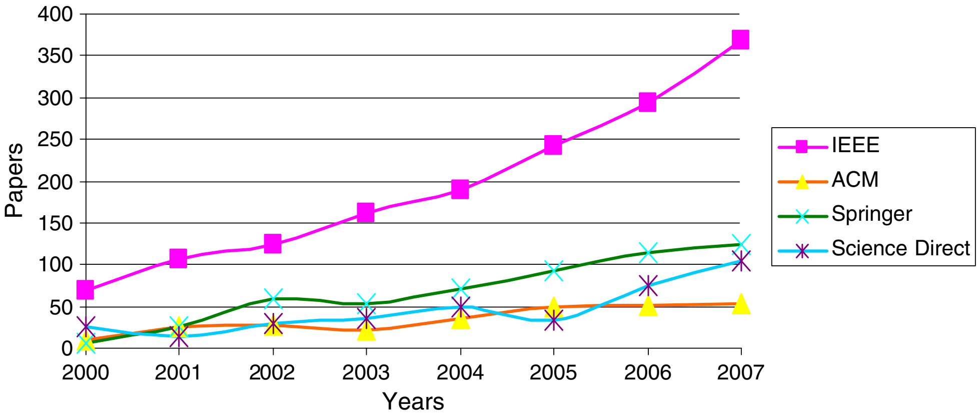

Fig. 1 Statistics of the publications on personalisation in IEEE, ACM, Springer, and Science Direct

approaches. Montgomery and Smith (2008) reviewed personalisation in four areas: definitions in marketing, conceptual framework for its models, methodologies for user preference, and effectiveness analysis. They then identified some examples representing the state of the art for the practice of personalisation. Adomavicius and Tuzhilin (2005) gave a review on three recommendation filtering approaches: content-based, collaborative, and hybrid. Jeevan and Padhi (2006) gave a survey of tools and techniques of personalisation in digital libraries.

In this paper, we review the previous research from three aspects: user profiling, content modelling, and information filtering. Since personalisation is about selecting or filtering information or product items for individuals, personalisation systems act as mediators between items and users (See Fig. 2). Consequently, most personalisation systems have three main procedures (Kim et al. 2002, Pazzani and Billsus 2006): (1) To make user profiles for users (2) To model the content (the description of information or products) for items (3) Data processing, using several filtering approaches to find out which items match with which users. We regard user profiling, content modelling, and information filtering as the essential parts of personalised systems. Therefore, we focus on these three categories for our review on the previous studies of personalisation.

After analysing previous research on the above three categories, we developed a detailed classification schema to review the previous researches on personalisation. Figure 3 represents the topology. The modelling of users and contents is the basis for constructing a personalised system. Behaviour, interest, and intention modelling are three important aspects of user profile modelling (Nasraoui 2005, Horvitz et al. 1998, Kramer et al. 2000, Jung et al. 2005). Content modelling includes content classification and analysis approaches besides the similar profiling methods used in user profiling (Joachims et al. 1998, Wei et al. 2008). It is also called content profiling, document modelling, or item profiling. Once the user profiles and content models have been obtained, systems can filter information and customise interfaces and services for users. Four main filtering approaches for personalised systems are: rule-based filtering, content-based filtering, collaborative filtering, and hybrid filtering (Liang et al. 2007, Adomavicius and Tuzhilin 2005). Details on the methodologies of user profiling, content modelling and information filtering are discussed in Section 4. The leaf nodes in Fig. 3 are the techniques that they require for application.

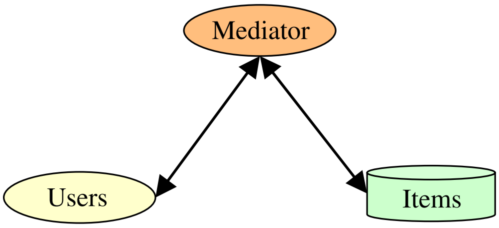

Fig. 2 A basic supply and demand model

Since personalisation is closely paired with the techniques and applications for which it is employed (Montgomery and Smith 2008), we will introduce applications with different techniques in the next section.

3 Personalised systems in different domains

There are several domains in personalisation research, including digital libraries, e-commerce, e-learning, news recommendation system, and search engines. A personalisation system is a computer-based application that build user profiles from past usage behaviour to provide relevant recommendations. The objective of personalisation systems is to retrieve information which is of interest to users from a large repository (Liang et al. 2007). Since most literature focuses on the first four domains, we analyse the role of personalisation and typical applications in these domains with a brief introduction to approaches and techniques.

3.1 Digital library

Digital Libraries (DL) became popular about 15 years ago. In general, DL is defined as collections of information that have associated services delivered to user communities using a variety of technologies (Callan and Smeaton 2003). Due to the great variety of information stored by DLs, users expect more intelligent services when they access and search for information (Frias-Martinez et al. 2006). One of the key elements on which these services are based is personalisation.

Personalisation allows a library appear as a set of separate smaller libraries to individual users (Jeevan and Padhi 2006). Three basic services provided by DLs are (1) personalisation of services, (2) interface personalisation, and (3) personalisation of content (Frias-Martinez et al. 2006).

Three steps (Morgan 2002) are adopted to achieve the above three services. Firstly, subject specialists create short lists of services for (1), interface parts and links for (2), and information resources (3) relevant to users. Some items in lists for (3) are recommended based on their importance to subject areas. Secondly, users who visit library websites are presented with a set of information resources organized by title, subject area, popularity, recommendation, and ease-of-use. Thirdly, every visitor has the option to create an account by submitting a username and selecting services for (1), interface parts, links, and styles for (2), and subjects for (3) from their respective lists.

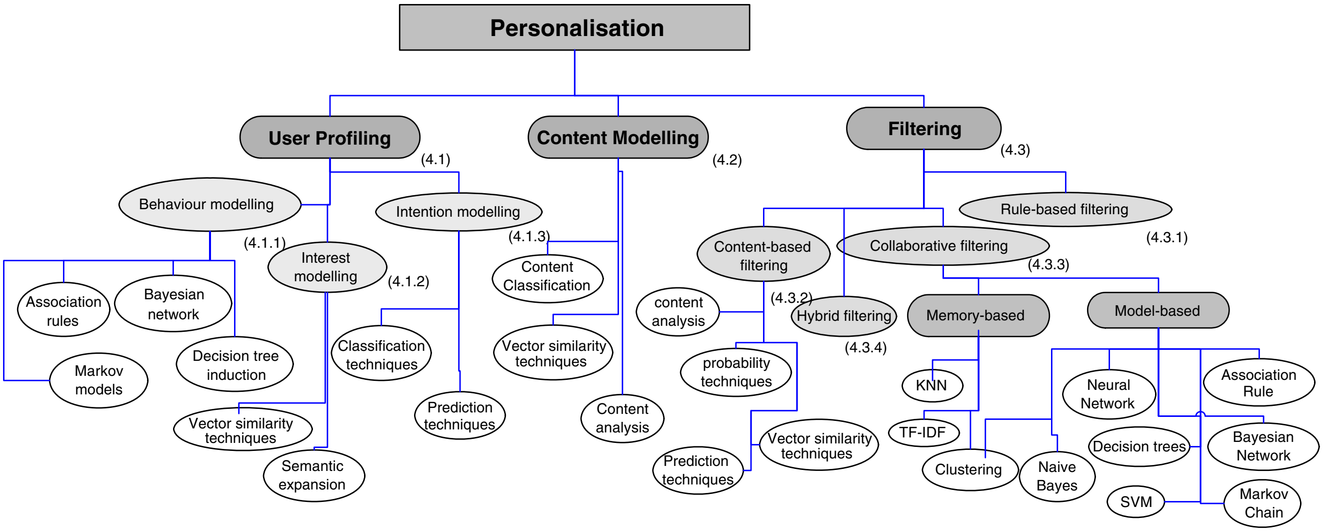

Fig. 3 Topology of personalisation approaches

These three steps are used broadly in DLs. However, they are based on manual selection of settings. How to automatically achieve them is complex, especially for (3) personalisation of content. Content means course-related materials, personalised information, and interactive guides for help, training, and support. Keyword-based user profile and content modelling is usually applied to content-based personalisation.

3.2 E-commerce

In the e-commerce domain, personalisation has emerged as a tool implementing strategies to lock in existing customers and to win new customers (Kuo and Chen 2001).

Customer profiling is proposed as the most important step in achieving one-to-one marketing in e-commerce. Narayanan et al. (2002) proposed a task-based personalised search form and a cognitive model to collect user profile. Three kinds of profiles have been proposed to represent online customers (Niu et al. 2002), which include basic profile, preference profile, and rule profile. Schubert and Koch (2002) discussed four steps of the customer profile life cycle: modelling, data input, data processing, and information output. Youm and Cho (2001) regarded the purchasing pattern and preference are of equal importance in the customer profile to recommend products by item dependency map.

To give more support to the design and implementation of user modelling, from a novel perspective, Fink and Kobsa (2000) discussed the core features of selected commercial systems from three aspects: company profile, product profile, and similar products. The product profile is analysed in detail including functionality, data acquisition, representation, extensibility, integration of external users and usage information, privacy, architecture, supported software and hardware, and the client base.

To recommend appropriate products, some approaches are proposed for product filtering. Schubert and Koch (2002) proposed a collaborative filtering method to achieve personalised interface and store. Wu et al. (2001) proposed a prototype system with a content-based collaborative filtering approach. It helps to form user communities for targeted advertising and recommendations. A hybrid approach bases on collaborative filtering and Web usage mining was proposed to achieve more appropriate product recommendations (Ha 2002).

3.3 E-learning

Personalisation is often portrayed as the “next big thing” in e-learning (Weller 2007). E-learning has potential to offer personalisation on a large scale. The ultimate aim of e-learning personalisation is to offer learning objects to meet the needs of learners in a timely fashion. Data mining aids the process by identifying patterns of behaviour.

For offering personalised textbooks for learners, a “Slicing Book” approach was proposed to slice textbooks and to compose the slices according to the logical relationships and user requirements (Dahn and Schwabe 2002). To construct personalised course content, a personalised Web tutor tree was used to adapt teaching in accordance with individual student ability in a distance learning environment (Tang et al. 2000). Personalised learning paths were proposed to achieve a course generation engine (Carchiolo et al. 2003). The engine can create a tree that includes all possible paths starting from the learner’s possessed knowledge towards desired knowledge. For organising and composing course notes, Cheong et al. (2002) designed a web-based collaborative multimedia content authoring and management system. Quarati (2003) combined reusable techniques with recommendation techniques to realise content best fitting for the needs of different students. From an instructor perspective, Hsu et al. (2003) proposed a result feedback model that analysed relationships between student learning time and corresponding test results to give instructors crucial information relating to course content refinement.

From an applied perspective, some researchers have proposed several architectures and systems to construct personalised learning content. A probabilistic framework for finding principled solutions were developed by Yu et al. Probabilistic active learning methods were used to actively query users, thereby solve the "new user problem" and reduce computational cost by working on a selected subset of user profiles while keeping high accuracy (Yu et al. 2004). An architecture for defining reusable adaptive educational content was presented, which allowed content to be easily interchanged and reused across applications (Karagiannidis et al. 2001). An instructional framework was proposed by Papanikolaou and Grigoriadou (2003) to unify the lesson generation process.

3.4 News recommendation and search engine

Online personalisation information systems allow for personalised access to online information such as personalised portals to world-wide news, filtering systems, recommendation services, and web search.

Much research has been focused on news filtering since the end of the last century. Researchers at MIT have developed an experimental online newspaper system called "Fishwrap". "Fishwrap" can help freshmen balance individual requirements for personalisation with the need to gain knowledge about the world (Chesnais et al. 1995). "PointCast" is a personalised news delivering service, which was developed as a screen saver that displays headlines in customised categories (Ubois 1996). "SmartPush", news content is classified using semiautomatic tools, pre-defined vocabularies, and metadata to match against user profiles (Jokela et al. 2001). Liang and Lai (2002) presented an approach that analysed browsing content and time to estimate user interests. Their experiment results show that the proposed system outperforms traditional headline news compiled by a news editor in terms of both objective performance indices and customer satisfaction. A semantic client-side personalised pre-fetching system called "NewsAgent", was designed for Internet news services by Xu and Ibrahim (2004).

Information retrieval is used to indentify relevant items for recommendation, which begins with a user querying and searching relevant documents from a large database (Liang et al. 2007). The main function of information retrieval is to match user queries with features of documents. Information retrieval techniques have mostly been applied to Internet search engines in recent years. Considering user interests, Scime (1997) presented a method for query construction and results analysis. They provided a user with a list of choices based on his/her determination of importance. From a technical perspective, Schmitt and Oberlander (2002) combined personalisation and query processing techniques to satisfy user demand for quick search and comprehensive results. Wang et al. (2002) proposed a knowledge-driven search system for multidimensional data analysis. The system adopted association rules and sequential pattern-based data mining. Wallace et al. (2003) pointed out that semantic content analysis, user profiling, and adaptation will provide users with effective content-searching capabilities. Liu et al. (2004) proposed a context-based personalisation searching approach which takes contextual information to disambiguate user queries.

3.5 Personalisation applications

Typical personalised applications in several domains which were mentioned in Sections 3.1–3.4 and their main features are represented in Table 1.

The "MyLibrary" system allows users to set up personalised and shared libraries through reorganising resources contained in the Research Library Web site (Di Giacomo et al. 2001). It provides basic personalisation mechanisms regarding user-driven services to create a portable Web page and list of available information resources (Jayawardana et al. 2001, Jeevan and Padhi 2006). The Personalised collaborative digital library in "CYCLADES project" formalises an environment in which users search for information, organise information space, collaborate with other users sharing similar interests, and get recommendations (Renda and Straccia 2005).

To produce accurate recommendations, Mooney and Roy (2000) utilised information extraction for text categorisation in a content-based book recommending system. Unlike traditional collaborative filtering, item-to-item (item-based) collaborative filtering is applied in Amazon.com. The online computation scale of this algorithm is independent of the number of customers and number of items in the product catalogue (Linden et al. 2003).

In the Adaptive Hypermedia Services System, a multiple models approach for dynamic composition and delivery of reusable learning objects was proposed by (Conlan et al. 2002). It describes an adaptive metadata-driven engine that composes sliced educational materials dynamically. An intelligent tutoring system based on fuzzy item response

Table 1 Domains and Features of Personalisation Systems

Domains	Systems	Features	Ref.
Digital library	MyLibrary	Customised and shared digital library, rule-based recommendation	(Di Giacomo et al. 2001)
	CYCLADES	Collaborative digital library environment	(Renda and Straccia 2005)
E-commerce	Content-based Book Recommending System	Content-based recommendation	(Mooney and Roy 2000)
	Amazon	Item-based collaborative filtering	(Linden et al. 2003, Das et al. 2007)
	Multiple-interests and Multiple-content Recommendation System	Collaborative filtering based on both items and users	(Li et al. 2005)
E-learning	Adaptive Hypermedia Services	Multi-model approach to the dynamic composition and delivery of reusable learning objects	(Conlan et al. 2002)
	Intelligent Tutoring System	Based on fuzzy Item Response theory	(Chen and Duh 2008)
News recommendation and search engine	Intelligent News Filtering Organisation System	Based on readers' revealed preferences	(Mock and Vemuri 1997)
	ANATAGONOMY	Learns preferences from user browsing behaviour	(Sakagami and Kamba 1997)
	Google News	Combines recommendations from different algorithms using a linear model	(Das et al. 2007)

	Internet Search Advisor	Knowledge-driven search system;
		Multidimensional data analysis;
		Data mining based on association rules and sequential patterns.
			(Lai et al., 2003)

theory was implemented by Chen and Duh (2008). It recommends appropriate course content to learners whilst taking into account whether a learner's ability matches the difficulty levels of recommended course content well. The SMCM (Semantics Modelling Method for Content Management) prototype was proposed to design and organise learning content to assist learners constructing knowledge in a specific social context.

INFOS (Intelligent News Filtering Organisation System) reorganises the order of news based on reader interests (Mock and Vemuri 1997), ANATAGONOMY system learns preferences from browsing behaviour of a user, and uses a learning engine and a scoring engine to produce personalised news (Sakagami and Kamba 1997). To achieve scalable online collaborative filtering, Google news system combines different algorithms in a linear model for recommendations (Das et al. 2007). Li et al. (2005) implemented a system using collaborative filtering based on both item and user for multiple-interest and multiple-content recommendation.

The discussion in this section concentrates on the roles and functions of personalisation in several domains. It briefly introduces a variety of methodologies for maximum user convenience and personalised resource usage. In the next two sections, methodologies and techniques are introduced and analysed from a technical perspective.

4 Methodologies for personalisation

In a personalisation system, a "user profile" is important because it records relevant information such as preferences and behaviour patterns of individual users. Details of user profiling methodologies are represented in Section 4.1. Similar to user profile modelling, content modelling, analysed in Section 4.2, is also a key element of personalisation systems. Finally, Section 4.3 reviews the most used recommendation approaches.

4.1 User profiling

When analysing how a personalised system makes recommendations for a user, one key issue is the user profile. The user profile stores appropriate approximations of an individual's information, such as basic information (e.g. age, gender), information usage behaviour, interests, and intention, which is represented by terms of keywords, concepts, and features (Pretschner and Gauch 1999, Park and Chang 2008). There are four design decisions involved in user profile generation and maintenance. These include representing profiles, generating initial profiles, capturing feedback and profile learning (Montaner et al. 2003).

1). Profile representation is the first step and all other techniques depend on it. Once this step is decided, the other techniques can be defined. It should be in the form of a general universal representation method, for example, XML documents. A system needs to know as much as possible from a user to provide satisfactory results from the very beginning. 2). It needs to use a suitable technique in order to generate an accurate initial profile (Montaner et al. 2003), e.g., by questionnaires or setting default values. 3). To capture user profile information that varies and requires active engagement of the user at different degrees, we will distinguish it by asking users (fill-in-profile, explicit feedback, or ratings), watching users (interaction with the web sites, e.g., click stream or transaction), and analysing the data (Schubert and Koch 2002). 4). Finally, the system gathers relevant feedback to learn the user's tastes, interests, or preferences (pattern discovery). The data collected from watching the user is usually not suitable to be directly used in information filtering algorithms. Therefore, learning techniques are needed to extract relevant information. In these processes, users are usually classified into different types or groups. The derived information is stored in the user's profile for further processing. Machine learning techniques are ideal for this process because they are designed to capture patterns and to represent what has been learnt from input data with a structural representation.

To take an example, profiles for user *u* and items *i* are based on a set of terms, named *ContentBasedProfile (u)* and *Content (i)* respectively. More formally, let *Content (i)* be content profile of an item, i.e., a set of attributes. It is usually computed by extracting a set of features from content, such as TF-IDF (term frequency-inverse document frequency) method. Let *ContentBasedProfile (u)* be the profile of user *u* containing tastes and preferences of the user. The profile is obtained by analysing the content of items previously bought and rated by the user. It can also be computed from individually rated content vectors using TF-IDF. The advanced profiling methods based on data mining have been used mainly in the context of system usage analysis, i.e., to discover user behaviour, preference, and intention analysis (Adomavicius and Tuzhilin 2005).

Personalisation has already been applied to some applications, which includes content personalisation, interface personalisation, interaction personalisation, and other related services. Personalisation of content aims at developing systems that are able to automatically provide personalised systems according to user preferences, behaviour, and intention. This process is closely related to information filtering. Interface personalisation systems tailor the user interface according to a user's characteristics. These characteristics are: 1). physical devices used for accessing systems and 2). the stereotype which the user fits. The interaction personalisation systems aim to create personalised navigation and service flow.

Within the context of personalisation description, there are many potential dimensions (Frias-Martinez et al. 2006) that a user profile will be described by (See Table 2).

In Table 2, the "personal data" includes gender, age, language, culture, and sex. In general, some of these factors affect the perception of interface layout. "Cognitive style" indicates the way in which a given user processes information. It is used to adapt services to the way the user processes information, especially in web-based services (Ford and Chen 2000) and e-learning environment (Magoulas et al. 2003). "Device information" captures the user's hardware and network environment. The information affects personalisation in two ways: the size of the visual display unit and download speed. "Context" captures the physical environment from where a user is accessing a system (e.g., location, time). "History" records a user's past actions in the system, which will be applied to personalisation based on the assumption that users' future actions can be inferred by examining their past pattern of behaviour. "Interests" represents topic categories, usually in the form of terms. "Interaction experience" indicates the user's knowledge about interacting with the system.

Table 2 Main Dimensions of User Profile

Name	Description	Effect
Personal data	Basic information includes age, language, culture and sex.	Interface
Cognitive style	The way in which a user processes information	Interaction
Device information	Hardware and network environment	Interface and content
Context	Physical environment when a user is accessing the system	Infer the user's intention
History	The user's past interaction with the system	Infer the user's behaviour and interests
Behaviour	The user's behaviour pattern	Content and interaction
Interests	Topics the user is interested in	Content
Intention/Goal	The intentions, goals or purposes of users	Content and interaction
Interaction experience	The user's knowledge on interacting with the system	Interface
Domain knowledge	The user's level of knowledge in a particular topic	Infer the user's intention

"Domain knowledge" indicates the knowledge levels of a user in particular topics. In all of these, modelling user behaviour, interest and intention are three complicated profiling methods (Liang et al. 2007).

4.1.1 Behaviour modelling

Historical data record user behaviour of browsing or transactions on a web site. User behaviour modelling involves the discovery of patterns from one or more web servers (Frias-Martinez et al. 2006). Analysing this data is helpful to determine navigation paths, cross marketing strategies, and the effectiveness of promotional campaigns.

There are already some methods for behaviour modelling. Nanopoulos, et al. (2001) used (1) association rules to model web user history and predict next requests. (2) Markov models provide another possible approach to capturing user historic behaviour in a website and allow for the implementation of a link prediction service (Sarukkai 2000). Markov chains can be used to model URL access patterns dynamically. The patterns are observed in the nearest navigation logs. (3) Decision tree induction techniques are the most widely used prediction methods, and may be used to predict the next interaction (Kim et al. 2002, Cho et al. 2002). However, in many cases prediction results are only locally optimal because of most systems only predicting one step forward. Sun et al. (2002) propose an approach which predicts multiple steps forward. Detailed techniques of these methods are described in Section 5.

4.1.2 Interest modelling

Preference is defined by a function *pref(i)* which represents how much a user likes or dislikes a given item *i* by analysing his/her behaviour history (Jung et al. 2005). Many studies have been performed on users' interests or preferences. There exist three popular methods (Schubert and Koch 2002) for extracting user preferences: direct, semi-direct, and indirect extraction. (1) The direct approach asks users to tell explicitly what they like. For example, listing all categories and asking users to check those of interest to them. (2) The semi-direct approach asks users to rate all documents they had read and gains their preference through these ratings. (3) The indirect approach captures user preferences from past browsing data, such as hyperlink clicks or time spent on reading a document.

The indirect approach for interest modelling is classified into three types depending on how they represent preferences, which are vector similarity, probability, and association rules.

(1) Vector similarity is used in both collaborative filtering and content-based filtering. In collaborative filtering, a user's preference is represented by high ratings to given items rated by himself/herself and similar users (Breese et al. 1998, Resnick et al. 1994). Content based approaches, such as TF-IDF (Joachims 1997, Billsus and Pazzani 1999, Blei et al. 2003) and SVM (Support Vector Machines) (Billsus and Pazzani 1999, Ruvini 2003) use vector similarity to measure how similar a specific piece of content is to content selected by a given user.

(2) Preference is represented by the probability of a user selecting a given item (Breese et al. 1998). The probability is adopted to predict user future selection, for example, in a Bayesian network (Friedman et al. 1997, Joachims 1997).

(3) Preference is represented by the strength of association between an item and the user's history in association rule mining methods (Nanopoulos et al. 2001, Brin et al. 1997).

Joachims (1997) and Jung et al. (2005) concluded that the results from the Bayesian classifiers are relatively not good as that of TF-IDF and SVM, whereas they show smaller differences in common documents. Reasons for the differences are sparse description, omitted synopsis, and limited training set. Bayesian classifiers use products of probability terms and it is dominated by terms with low probability values. This makes it vulnerable to sparse data. In addition, experimental results show that TF-IDF is more robust than SVM in sparse data sets (Jung et al. 2005).

4.1.3 Intention modelling

The intention (expressed as the purpose, aim, and goal), means what one intends to accomplish or attain, or the reason for which that user is searching information (Frias-Martinez et al. 2006). Intention modelling is constructed of a model to identify the goal of a user when interacting with a system. For example, the consumers can be divided roughly into two groups: one with the intention of buying and the other without the intention of buying. Chen et al. (2002) classifies intention into action intention and semantic intention. The former is a low level intention, e.g. by the analysis of mouse click and keyboard typing, whether the user has the intention of buying can be deduced. "I want to buy a mug from eBay for my friend" is a semantic intention. The reason for user searching mug is to buy a friend a mug. Then his/her goal is to find an artistic or elegant mug.

Intention modelling is largely based on a classification system that has a set of predefined categories (Frias-Martinez et al. 2006). To model user intention, Ruvini (2003) presented an approach that infers the goal of a search using SVMs. Another solution which provides good results is Bayesian networks (Horvitz et al. 1998, Chen et al. 2002). “Office Assistant” (Horvitz et al. 1998) combines SVM and Bayesian Networks for predicting user goals. Other valid approaches are decision trees, neural networks, and semantic expansion (Frias-Martinez et al. 2006, Blanco-Fernandez et al. 2008). To achieve better personalised recommendations, Chen, et al. (2002) proposed a hybrid approach, which uses Naïve Bayes and association rule to model user action intention (behaviour pattern) and find the proper concepts of corresponding keywords respectively. The model is trained incrementally and is used to predict future actions. Besides keyword features, Chen et al. (2002) proposed an algorithm utilising “WordNet” to generalise learned rules for similar words.

In some publications, user intention modelling is treated similarly as behaviour and interest modelling. However, intention modelling is at a higher level than behaviour and interest modelling. The intention analysis is further analysis based on the behaviours and interests and other three elements: (1) users’ query content, (2) organisational structure of the system, and (3) context (the information is used to characterise the situation of a participant in an interaction). In another word, behaviour and interest modelling provide a basis for intention modelling.

4.1.4 Analysis for user profiling

The user profile is usually defined as a vector of *n* features. For example, a user profile is described simply by the user’s ratings on items. Some context-aware personalisation systems treat contextual information as extended features of rating-based profiles (Adomavicius et al. 2005). Recommendation systems based on this kind of profile have grown in popularity in recent years. However, this kind of profile causes the system to suffer from the “Cold start” problem. Sometimes the system cannot draw any conclusion for users because it has not yet gathered sufficient information (Schein et al. 2002). The advanced profiling methods we mentioned in prior sections can solve the problem if they are adopted in the rating-based recommendation systems.

4.2 Content modelling

In personalisation systems, document modelling is a key element. In this context, some kinds of indications are needed on the topic of a particular document, which are usually expressed in the form of keywords. There are two possibilities to obtain these keywords, (a) metadata already has a field that contains them for content-free items and (b) keywords are obtained using document modelling techniques like TF-IDF.

The main up to date document analysis techniques are Latent Semantic Analysis/Indexing (LSA/LSI) and Probabilistic Latent Semantic Analysis/Indexing (PLSA/PLSI). LSA is used to obtain term-concept and concept-document relationships from a term-document matrix (Hofmann 2004). PLSA is a statistical view on LSA to estimate probability from a collection of documents (Hofmann 1999, Jin et al. 2004).

Several methods for content classification have been proposed, each with a different approach for comparing new documents to the reference set. These include (1) comparisons between a variety of frequently-used vector representations of documents, e.g., SVM, KNN (k nearest neighbours), linear least-squares fit, TF-IDF; and (2) joint probabilities of the words being in the same document, for example, naïve Bayesian, decision trees, and neural networks (Gauch 2003).

The detailed discuss of the techniques in this section are discusses in Section 5.1.

4.3 Information filtering recommendation methods

A problem that a user is often confronted with is that there are enormous quantities of information on the internet. Recommendation methods are essential in this context. For example, a busy customer who knows what he wants, too many choices may cause a barrier to purchase. Imagine that a customer has already configured and bought a car, which also pleases another customer with similar taste. It is easier for the second customer to just order the same configuration than to run through the whole selection process again (Schubert and Koch 2002, Montaner et al. 2003).

Adomavicius and Tuzhilin (2005) pointed out the recommendation problem can be formally formulated as follows: let U be the set of all users and let I be the set of all possible items that will be recommended. Let f be a utility function that measures the rating (or usefulness) $r_{u,i}$ of user u on item i. Let R be possible ratings for I, the function f can be seen as a mapping from U and I to R, f: $U \times I \rightarrow R$.

There are four filtering approaches for making recommendations: (a) rule-based filtering, based on “if this, then that” rules processing (Choi and Han 2008), (b) content-based filtering, based on a comparison between items and users profiles (Park and Chang 2008), (c) collaborative filtering, which serves relevant material to customers by comparing their own personal preferences with others (Instone 2000), and (d) hybrid method, which combines content-based and collaborative filtering. Details of these methods are in the following sections.

4.3.1 Rule-based approaches

Rule-based approaches allow information systems to specify rules *f(u,i)* based on demographics or static profiles of users, which are collected through a registering process by asking users a series of questions. Then, pre-specified if-then rules (See Fig. 4) are applied to select relevant information for recommendation (Liang et al., 2007). It relies on predefined groups or classes of users to determine what content should be displayed or what services should be provided. For example, online brokerages often classify their accounts by sex and age, and then provide different user services, products, or preferential treatment.

In rule-based approaches, filtering rules from domain experts (e.g., marketing and e-learning) is a core component in providing personalised recommendations. The effectiveness of rule-based approaches mainly depends on the quality of knowledge in a rule base. It is, however, difficult to obtain valuable marketing rules *f(u,i)* from marketing experts and to validate the effectiveness of extracted rules. In addition, the rule-based approach poses significant maintenance issues. Machine-learning techniques can be used to reduce the problems of knowledge-acquisition and knowledge-maintenance problems in rule-based approaches to personalised recommendations (Kim 2001). In addition to the knowledge engineering bottleneck problem, the methods used for generating user profiles are another weakness. The input is usually a subjective description of users or their interests by the users themselves, and is thus prone to bias (Mobasher 2007).

4.3.2 Content-based approaches

Content-based approaches facilitate the discovery and filtering of information by comparing user profile with the description of items (Chedrawy and Abidi 2006). Content-based filtering systems will compare items' profiles with a user's profile (See Fig. 5) in order to estimate which items the user may be interested in (Min and Han 2005). For instance, a content-based news recommendation system will extract essential items of news using some text classification methods, and then match them against the user profile to select a set of promising news recommendations.

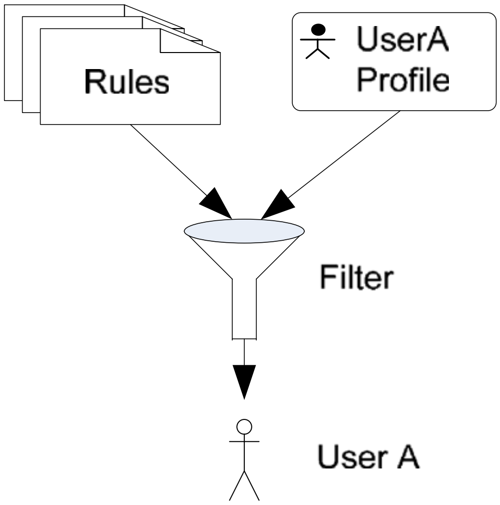

Fig. 4 Rule based filtering

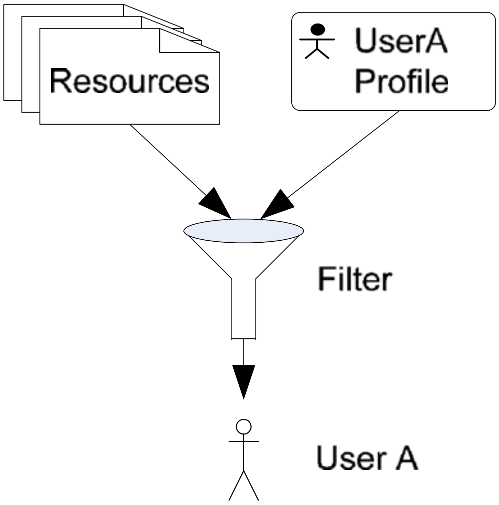

Fig. 5 Content-base filtering

Formally, in content-based recommendation methods, the utility function *f(u,i)* of user *u* on item *i* is estimated based on the utilities *f(u,i_k)* that were rated in the past and i_k are similar to item *i* (Adomavicius and Tuzhilin 2005). For example, in the case of a book recommendation application, the system learns which books user *u* has rated highly in past. Then only the books with a high similarity rating to the user's preferences will be recommended. The similarity can be determined in several ways. One of the best known measures is Cosine similarity (See Section 5.1.1).

Content-based filtering techniques are most effective at text-intensive domains. Such systems include "News Weeder", "Infofinder", and "News Dude" (Park and Chang 2008). However, they are not suitable for multimedia content, such as images, pictures, and sounds. Moreover, it is difficult to identify other related and potential interests. (Mobasher 2007).

4.3.3 Collaborative approaches

Collaborative filtering is a complementary technology to content-based filtering (Das et al. 2007), which uses preferences of similar users in the same preference group as a basis of recommendation (See Fig. 6). It is an approach to making recommendations by finding correlations among shared likes and dislikes of system users. It has the capability of finding items of potential interest from previous ratings of other users (Liang et al. 2007). In this way, the mechanism for filtering information is changed and improved on demand.

Formally, utility function *f(u,i)* is estimated based on the ratings of item *i* by those users who are similar to user *u*.

For example, in a book recommendation system, the system tries to find similar users to user *u*; only books that are most liked by similar users will be recommended.

User-based collaborative filtering was the most successfully used technique for building recommendation systems in the past (Konstan et al. 1997, Li et al. 2005). However, its computation cost grows linearly with the number of users and items i.e. the approach suffers serious scalability problems. Another significant limitation is the sparsity of the dataset (Mobasher 2007, Das et al. 2007). To remedy these weaknesses, varieties of optimisation strategies have been proposed (Das et al. 2007), which include similarity indexing, dimensionality reduction, and offline clustering of users.

In addition, the item-based collaborative filtering is an extension of user-based collaborative filtering (Kitts et al. 2000). It builds an item-item similarity matrix rather than a user-user similarity matrix for recommendations (See Fig. 7). It can quickly recommend items because of the pre-computed model. They have been shown to produce recommendation results that in some cases are comparable to traditional, user-based collaborative filtering recommendation systems (Li et al. 2005). Mobasher (2007) points out that the similarity between two items can be calculated by comparing their ratings rated by the same users.

Several papers categorise collaborative personalisation systems into memory-based and model-based systems, based on the types of recommendation techniques used for rating estimation (Das et al. 2007, Hofmann 2004), (Adomavicius and Tuzhilin 2005). The main difference between them is that the model-based techniques calculates utility predictions not on some heuristic rules, but on a model learned from underlying data using statistical and machine learning techniques (Adomavicius and Tuzhilin 2005). Memory-based methods are sufficient for many real-world problems; however, they still have several shortcomings. Hofmann (2004) pointed out: (a) Accuracy of recommendation obtained by memory-based methods will be suboptimal; (b) Nothing is actually learned from available user profiles and no general insight is gained because no explicit statistical model is constructed; (c) such solutions not scale well because they require a lot of memory and processing time; (d) it is difficult to systematically tailor memory-based algorithms to a specific task. Model-based methods will address these shortcomings. A model-based method will (a) achieve higher prediction accuracies, (b) compress data into a compact statistical model that automatically identifies user communities, (c) be able to compute preference predictions in real time, and (d) give a system designer more flexibility in specifying the objectives of the application.

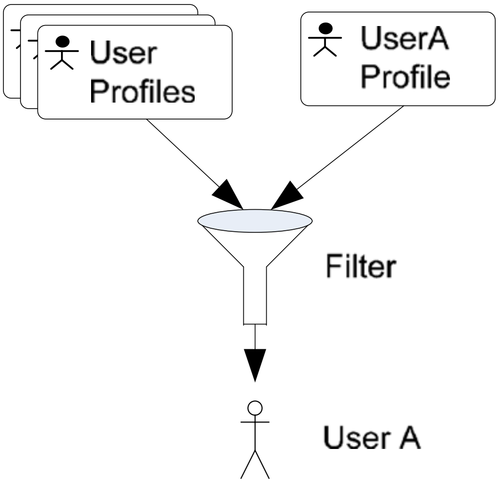

Fig. 6 Collaborative filtering (User-based)

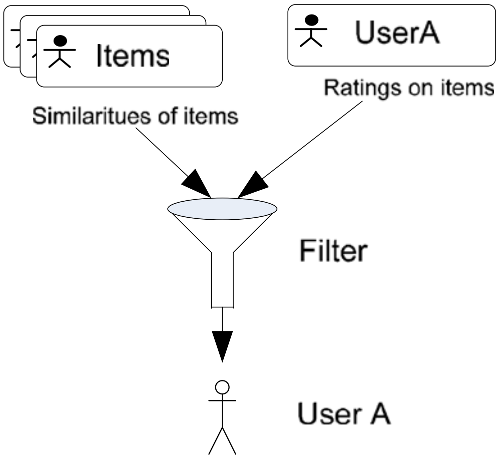

Fig. 7 Collaborative filtering (Item-based)

4.3.4 Hybrid approaches

Generally speaking, rule-based systems will capture common reasons for making recommendations, but they do not offer the same detailed personalised recommendations that are available with other recommendation approaches (Pazzani and Billsus 2006). Content-based filtering has proven to be better than collaborative filtering when recommending textual documents; whereas collaborative filtering does not require a content description of an information item, which is popular with e-tailors that sell physical products such as 'Amazon.com' (Linden et al. 2003). This method usually performs best when explicit non-binary user ratings for similar objects are available (Mobasher 2007). Furthermore, collaborative filtering will find similarities among different users. The integration of these two types of filtering is reported to exhibit good performance in some domains (Liang et al. 2007, Chedrawy and Abidi 2006, Liu and Shih 2005) such as e-commerce and digital libraries.

To avoid certain limitations of content-based and collaborative approaches, several systems use hybrid approaches by combing collaborative and content-based filtering (Balabanovic and Shoham 1997, Claypool et al. 1999, Melville et al. 2002). On the one hand, collaborative approaches will solve shortcoming of content-based systems such as the lack of subjective data, user ratings, and novelty. On the other hand, content-based approaches will solve the limitations of collaborative approaches including the "new items" problem, data sparsity problem, and complex computation problem (Montaner et al. 2003). Different ways to combine collaborative and content-based methods are classified as follows (Adomavicius and Tuzhilin 2005): (1) implementing collaborative and content-based methods separately and combining their predictions, (2) incorporating some content-based characteristics into a collaborative approach, (3) incorporating some collaborative characteristics into a content-based approach, and (4) constructing a general unifying model that incorporates both content-based and collaborative characteristics. Table 3 is a summary of these approaches.

The essential issue of personalised recommendation methodologies is pattern discovery and requirement prediction (Frias-Martinez et al. 2006). Machine learning techniques are ideal because they are designed to capture patterns and represent what has been learned from the input data. They are usually used when a system needs knowledge from its previous experience (Witten and Frank 2002). Regular patterns and rules are important for user modelling, content modelling, and information filtering. Those techniques make it possible to recognise patterns and rules automatically. The techniques will be discussed in the next section.

5 Techniques for personalised recommendation

There is already a lot of research on how machine learning techniques help to solve problems in personalisation. Based on different phases of the personalisation process, the techniques are categorised into user profiling, content modelling, and filtering techniques. Based on the kind of recommendation system, the techniques are classified into content-based, collaborative, and hybrid filtering. Collaborative personalisation systems are categorised as memory-based and model-based, based on the types of recommendation techniques used for rating estimation (Adomavicius and Tuzhilin 2005). These categories overlap each other; one technique can be applied in different personalisation processes, can also be used in both content-based and collaborative filtering. It is difficult to strictly classify them.

In this section, we will analyse the features and limitations of (1) techniques for content-based recommendation systems in Section 5.1, including techniques for user profiling and content modelling, and (2) techniques applied to memory-based and model-based collaborative recommendation systems in Section 5.2, including KNN, K-means, Naïve Bayes, Bayesian networks, Markov chains, association rules, decision trees, and neural networks.

5.1 Techniques for content-based recommendation

Content-based recommendation is based on the past items which were preferred by a user. Consequently, it applies the techniques mentioned in the user profiling and content modelling sections, which include terms weight-

Table 3 Summary of Filtering Approaches for Recommendation

Approaches	Applicable Scope	Advantages	Weakness	Systems
Rule-based approach	The system aims to achieve basic services for personalisation.	It is easy for systems to use if-then rules to select personalised information. It captures common reasons for making recommendations.	It is not easy for administrators to pre-specify rules; it is not flexible; it is for groups (women, children, etc), not real personalisation.	MyLibrary (Di Giacomo et al. 2001)
Content-based approach	Text-intensive domains	It is most effective at text-intensive domains.	Not suitable for multimedia information; cannot find potential interests of users.	News Weeder, Infofinder, and News Dude (Park and Chang 2008)
Collaborative approach	non text-intensive domains with lots of historical data	Popular in the e-commerce domain; can find potential interests for users; not limited to textual information.	"Cold start" problem, new user problem, new item problem.	Amazon (Linden et al. 2003, Das et al., 2007) CYCLADES (Renda and Straccia,2005)
Hybrid approach	The systems want to achieve advanced services for personalisation.	Combines the advantages of content-based and collaborative methods.	It may be too complicated to achieve.	ExplaNet (Masters et al. 2008)

ing, content analysis, classification, and user behaviour analysis.

5.1.1 Computing weights of terms and similarity

In content-based personalised recommendation systems, users and items are usually represented as term vectors. There are several methods of computing the weights of the terms. TF-IDF is recognised as the best known method according to prior studies of (Joachims 1997, Billsus and Pazzani 1999, Blei et al. 2003).

- *TF(*t_i*;*d_j*)* means the number of occurrences of term t_i in document d_j (so called term frequency).

  $$TF\left(t_i; d_j\right) = freq(t_i)$$

- *DF(*t_i*)* (Document frequency) counts how many documents include term *i*.
- *IDF(*t_i*)* means the importance of the term t_i in all documents. The fewer documents include t_i, the more important the term t_i becomes. |D| is the number of documents.

  $$\mathrm{IDF}(t_i) = \log\frac{|D|}{DF(t_i)}$$

The process of computing TF-IDF is listed as follows.

1) A basic vocabulary of terms is chosen;
2) Calculate *TF(*t_i*;*d_j*)* for every term and every document in the corpus (document collection);
3) Count *DF(*t_i*)* for every term;
4) Compute all *IDF(*t_i*)* for every term;
5) All the columns contain *tf*idf* values for each of the documents which constitute a matrix *X*.

Moreover, utility function *f(u,d)* is usually represented in content-based recommendation by a scoring heuristic method, such as cosine similarity measure Eq. 1 (Adomavicius and Tuzhilin 2005). Where *K* is the total number of keywords; term vectors $\overrightarrow{w_u}$ and $\overrightarrow{w_d}$ represent user profile and document; *i* is a term in both $\overrightarrow{w_u}$ and $\overrightarrow{w_d}$.

$$sim\left(\overrightarrow{w_u}, \overrightarrow{w_d}\right) = \cos\left(\overrightarrow{w_u}, \overrightarrow{w_d}\right) = \frac{\overrightarrow{w_u} \cdot \overrightarrow{w_d}}{||\overrightarrow{w_u}||_2 \times ||\overrightarrow{w_d}||_2} = \frac{\sum_{i=1}^{K} w_{i,u} w_{i,d}}{\sqrt{\sum_{i=1}^{K} w_{i,u}^2}\sqrt{\sum_{i=1}^{K} w_{i,d}^2}} \quad (1)$$

The difficulty in TF-IDF is identifying the keywords or items in a document since there are plenty of words. While the TF-IDF reduction is helpful to address this shortcoming, it only provides a small amount of reduction in description length and reveals little in the way of statistical document structure. Several other dimensionality reduction techniques are proposed. For example, latent semantic Analysis (LSA) uses a singular value decomposition of *X* matrix to identify a linear subspace in the space of TF-IDF features, that will not only capture most of the variance in the collection, but also achieve significant compression in large collections (Blei et al. 2003). In addition, from semantic respective, linear combination of the TF-IDF features and derived features of LSI will capture some aspects of basic linguistic notions such as synonymy and polysemy.

5.1.2 Content analysis

Latent Semantic Analysis is a well-known technique for content analysis, which partially addresses polysemy and synonymy problems. It represents contents and queries not by terms but by underlying (latent hidden) concepts. This hidden structure is not a fixed many-to-many mapping between terms and concepts, but depends on the corpus and term correlations it embodies (Hofmann 2004). The key idea of LSA is to map high-dimensional vectors to low-dimensional ones in a latent semantic space. The goal of LSA is to find a data mapping which provides information well beyond a syntactic level and reveals semantic relations. LSA is a valuable analysis tool with a wide range of applications (Hofmann 1999).

For content analysis, suppose there is a given collection of documents *D=* $\{d_1,...,d_n\}$ with terms from a vocabulary *T=* $\{t_1,...,t_m\}$. It will be summarised in a $n*m$ co=occurrence table of counts $A=(freq(d_i,t_j))_{ij}$, where $freq(d,t) \in IN$ denotes how often term *t* occurred in document *d*. *A* is called term document matrix. The rows and columns of *A* represent document and term vectors respectively. Vector-based document representation is mapped in a low-dimensional space obtained by Singular Value Decomposition (SVD) of the term-document matrix *A* (Papadimitriou et al. 2000). The simplified vector space representation will preserve most relevant information.

To discover latent variables in a document, PLSA provides a more flexible approach than LSA (Hofmann 2003, Hofmann 1999, Jin et al. 2004). PLSA has a solid statistical foundation. The starting point for PLSA is a statistical model called aspect model. For instance, document *d* is a mixture of underlying (latent) *K* aspects, and each aspect is represented by a distribution of words *p(w|t)* (Hofmann 1999). *t* is a topic. In brief, if *probability(t|d)* and *probability(w|t)* is estimated, and then *probability(t|query)* can be estimated, then we can compute the similarity of document to the query. PLSA can obtain these probabilities from a document collection. It is supposed that there are hidden factors underlying co-occurrences among two sets of objects (Hofmann 2003). In PLSA, an expectation-maximisation (EM) algorithm is used to calculate the relationships between hidden factors and two sets of objects. It is experimentally proved that PLSA consistently outperforms LSA (Hofmann 1999).

5.1.3 Content classification

Content classification techniques are applied to divide items (e.g., news, papers, and books) into a number of classes or groups; the items in the same class or group are similar in some way. There are a number of techniques for content classification, such as SVM, association rules, and decision trees.

SVM SVM is a kind of generalised linear classifier. A special property of SVM is that it will separate a n-dimensional space into two data sets by a hyperplane, which minimises classification errors and maximises the margin between the two sets (Joachims 1997).

In a personalisation system, assume a user has provided ratings for n products. Suppose that the rating will either be 1 (user likes the product) or —1 (user does not like the product). Then it is a binary classification problem. The task is to separate products that a user likes from those that he or she dislikes. The task can be fulfilled by SVM. For making recommendations, a training set including tuples of the form $\{(I_j, r_j)\}_{j=1}^{n}$ is given to learn the SVM. Here, I_j (as input) are items, j is the feature vector, and r_j (as output) is the corresponding rating. If the SVM model is obtained for a user, it will determine whether to recommend another item to this customer by some equation (Cheung et al. 2003).

In contrast to neural networks, SVM can always find a global solution. And it always constructs classifications with a high quality. However, SVM is currently considered slower at run time than other techniques with similar methods' performance (Cheung et al. 2003). Also, the binary classification is not enough for a personalised recommendation.

Association rule Association rule learning is a popular method for discovering relations. Based on strong rules (Piatetsky-Shapiro 1991), Agrawal et al. (1993) introduced association rules for recognising relationships between items from a large scale of transaction data. For example, a rule found in sales data indicates that if a customer buys both tomatoes and beef, he is likely to also buy onions. Such information will be used as the basis for decisions about marketing activities.

Given a set of transactions where each transaction is a set of items (item set), an association rule implies the form $X \Rightarrow Y$, where X and Y are item sets; X and Y are called body and head respectively. *Support(X,Y)* Eq. 2 for rule $X \Rightarrow Y$ is the percentage of transactions that contain both item sets X and Y among all transactions. *Confidence(X,Y)* Eq. 3 for the rule is the percentage of transactions that contain item set X (Han and Kamber 2006, Kim et al. 2002).

$$Support(X, Y) = Count(X, Y)/|T| = P(XY) \tag{2}$$

$$Confidence(X, Y) = Count(X)/|T| = P(X) \tag{3}$$

The *Support(X,Y)* represents the usefulness of the rule. The *Confidence(X,Y)* denotes the certainty of the rule. Association rule mining is the discovery of all rules that are above a user-specified minimum *support* and *confidence*. For example, *{beef, tomato}*⇒*{onion}* is an association rule because both *support* and *confidence* of the rule are above the minimum support and confidence.

In user behaviour and intention modelling, through an analysis of historical interactions between users and a system, association rule sets will be created to record personalised relationships between links and content. Then the rule set can be applied to analyse user intention (which links are well associated with the previous link) and to recommend products (which items are associated with the items in shopping-basket). In addition, through analysing user interest and similar users' association rule sets, potential new relevant items will be discovered and displayed to the user.

Decision tree Decision tree is a prediction model that includes both items and target values. In a decision tree, each non-leaf node represents a test on an attribute of cases, each branch is linked to an outcome of a test, and each leaf node denotes a class prediction.

In a personalisation system, a decision tree can be applied to find rules for a predetermined target through analysing historical interactions between users and items in the website. Once the decision tree has been obtained, the system will recommend items for users with a high accuracy.

Compared with association rule learning, decision tree mining has only one predetermined target, which aims to discover a small set of rules that can be used to obtain an accurate classifier for users (Liu et al. 1998). In contrast to neural networks, the decision tree is well suited to multi-dimensional applications and has strong explanatory ability (Kim 2001, Kim et al. 2002).

Decision tree learning is very helpful to content classification in terms of performance, simplicity, and understand-ability when there is a small number of structured attributes (Pazzani and Billsus 2006). However, it is usually not ideal for unstructured text classification tasks because such tasks frequently involve a large number of relevant features. Therefore, a decision tree's tendency to base classifications on as few tests as possible will lead to poor performance on text classification.

5.1.4 User behaviour analysis

A number of techniques can be applied for user behaviour prediction, such as association rules (Section 5.1.3.2), decision tree (Section 5.1.3.3), neural networks (Section 5.2.4), and Markov chains.

A Markov chain is a stochastic process with the Markov property. Having the Markov property means, given the present state, future states are independent of the last previous states. That means the present state captures all information that influences the future evolution of the process. At each second, a system will change state to another state, or remain at the same state. This change is called a transition.

A discrete Markov chain model can be represented by tuple$<S; A; \lambda>$. S is a state space. A represents a transition matrix that includes possible transition from one state to another. Similarly an element of $A*A$ (A^2) represents the probability of transitioning from one state to another in two steps. And λ represents the initial probability distribution of the states in S. The fundamental property of a Markov model is dependent on the last previous state. If vector $s(t)$ denotes probability vector for all states at time t, then $s(t)=s(t\text{-}1)*A$ (Sarukkai 2000).

In the context of user profile modelling, Markov chains can be used to capture user behaviour in a website and implement a link prediction service. Based on the last previous state, URL access patterns can be modelled by Markov chains.

Although Markov chains have been traditionally used to characterise asymptotic properties of random variables, the transition matrix can also be used to estimate short-term predictions (Sarukkai 2000).

In user behaviour prediction, a probability distribution can be created to predict which link is likely to be clicked next. Variants of the Markov process can be calculated to accommodate weighting of more than one history state, e.g., A^2, A^3. Each of the previous links are used to predict a user's future behaviour (link choices) and combined in a variety of ways.

The Markov chain approach has two main limitations (Sarukkai 2000). (1) Dimensionality: The matrix A is very large ($N*N$ for N URIs). It is not scalable for very large sites. (2) The need for a large amount of training data. Since the approach is statistical, the applicability of the model is dependent on the amount of data available. Therefore it is not suitable for new personalisation systems.

5.2 Techniques for collaborative systems

For memory-based collaborative systems, after obtaining weights and utilities of words in user profiles, a number of techniques can be used to classify users, such as the nearest neighbours and clustering. The personalised systems can then recommend products most liked by the peers of the user. Differing from memory-based methods, model-based methods use collections of ratings to learn a model that is then used to make rating predictions. The techinques will be discussed in detail in this section.

5.2.1 KNN

KNN is a kind of instance-based learning for classifying objects based on closest training examples in a feature space (Billsus and Pazzani 1999). It is amongst the simplest of machine learning algorithms. It finds k nearest neighbours of an object, and then assigns the object to the most common class among its neighbours (Frias-Martinez et al. 2006).

The similarity function used by KNN algorithms depends on the type of data. For example, the Euclidean distance metric is often applied for structured data; Cosine similarity measure is often used when vector space model is used (Pazzani and Billsus 2006).

For information filtering, KNN classification approach compares a target user's profile with other users' in order to find the top k users who have similar tastes or interests (Mobasher 2007).

In the context of document classification, KNN can be applied to rank the k nearest neighbours for a given document. Then the category which the k neighbours are included in is the class of the input document (Yang 1999).

KNN algorithm is robust to noisy training data and effective if training data is large. However, Mobasher (2007) pointed out that (1) kNN takes neighbourhood calculations an online process. Following the increased numbers of users and items, it will lead to latency for providing personalised information. (2) The sparse nature of the datasets makes it difficult to get effective clusters. The increase of the number of items in a database decreases the likelihood of significant overlap of rated items among users, which will result in less reliable correlations.

5.2.2 K-means

K-means is a well known geometric clustering algorithm. Given a set of points, it uses a local search approach to partitioning the points into k clusters by minimising the sum of square diviations between data and the corresponding cluster centroids (Hartigan and Wong 1979). A set of k initial cluster centres is chosen arbitrarily.

Each point is then assigned to the centre closest to it, and the centres are recomputed to be the centres of the new point groups. This is repeated until the process has stabilised.

K-means can be applied to construct user groups with similar preferences or cluster different content or services. Consequently, the systems can make personalised recommendation according to the similar users' suggestions or selections.

K-means is effective in producing good clustering results for many practical applications (Alsabti et al. 1998, Arthur and Vassilvitskii 2006). However, k-means costs computation time proportional to the products of the number of patterns and clusters. It is a problem especially for large historical data.

5.2.3 Bayesian probability

Bayesian probabilistic models can be used to make rating predictions, which include naïve Bayes model and Bayesian network model.

Naïve Bayes uses joint probabilities of words and categories to estimate probabilities of categories for a document (Breese et al. 1998). It assumes these words are independent. This makes it far more efficient than exponential complexity of other Bayes approaches (Yang 1999). Two frequently used formulations of Naïve Bayes are: the multivariate Bernoulli and the multinomial model (Pazzani and Billsus 2006).

For text classification, multivariate Bernoulli assumes every document is represented as a binary vector over the space of all words from a vocabulary V. The multinomial model records word frequency information rather than the binary representation. Maximum likelihood estimates will take word frequencies into account.

A Bayesian network represents a set of variables and their probabilistic dependencies (Adomavicius and Tuzhilin 2005).

According to Horvitz et al. (Horvitz et al. 1998), Bayesian networks can be used to capture relationships between the goals and needs of a user and the observations about a sequence of actions and words in a user query. Through computation and use of probability distributions over a user's goals, a Bayesian network can provide appropriate assistance for dynamically tailoring information or services.

In contrast to the Naïve Bayes, Bayesian Networks describe a joint probability distribution for a set of features. Generally, a Bayesian Network provides better performance in classification than Naïve Bayes classifier. However, the computational complexity for building Bayesian Network becomes impractical when training data becomes large (Chen et al. 2002).

5.2.4 Neural networks

Neural networks are popular due to their ability to classify patterns (Shepherd et al. 2002). The learning or self-adapting ability allows them to store and recognise input states and sequentially generate appropriate outputs. A well-trained neural network can recognise incomplete input pattern in a much simpler way (Im and Park 2007).

User profiling can be seen as a pattern recognition process. It needs the techniques for of pattern classification, fault tolerance, graceful degradation, and so on. Neural networks are good at all these features (Chen and Norcio 1997). In the process of user profiling, each user has a unique profile characterised by a set of attribute values. In the training process, characterised by link weights, the network state represents the characteristics or attributes of a user or a user class (Shepherd et al. 2002, Im and Park 2007).

For user classification, each neural unit represents the centroid of a cluster in a feature space. So a network with k units can be applied to partition the feature space into k clusters (Albanese et al. 2004). After completing the training phase, neural networks can be used for classifying users.

However, neural networks are not so commonly used in personalisation systems as other learning strategies, e.g. decision tree (Shin et al. 2000), because of its shortcoming of being a "black box", which means systems do not know how it arrives at a given result.

5.3 Techniques and their features in a glance

Table 4 summarises the characteristics of recommendation systems techniques, presented along four dimensions. The first dimension captures some features of these techniques. The next dimensions describe the limitations of the techniques. The last column displays potential solutions.

6 Directions and Issues for Further Research

Personalised recommendation systems can be extended in several ways. From a methodological perspective, this includes (a) context analysis for personalisation, (b) facilitating user grouping with user profiling data in social networks, and (c) enhancing collaborative filtering with domain knowledge. From a technical perspective, this may include (a) providing an advanced recommendation techniques, (b) maintaining a balance between implicit and explicit feedback, (c) supporting multi-criteria ratings and (d) providing more flexible and effective recommendations. We will not review and discuss research opportunities from the technical perspective because they have been discussed

Table 4 Summary of Techniques for Content-based and Collaborative Personalisation Systems

Techniques	Features	Limitations	Potential Solutions
TF-IDF	The best known measures for keyword weights (Adomavicius and Tuzhilin 2005).	Difficult to identify key words or items in a document since there are many of words (Blei et al. 2003).	TF-IDF reduction; Latent semantic indexing (Isozaki and Kazawa 2002).
KNN	The most popular technique for classification. Robust to noisy training data and effective if training data is large (Billsus and Pazzani 1999).	Online neighbourhood formation process; difficult to get effective clusters because of the sparse nature of dataset (Mobasher 2007).	To combine with content-based approaches (Adomavicius and Tuzhilin 2005); to removes useless features before other procedures.
K-means	Partition points into k clusters by minimising the sum of square diviation between data and the corresponding cluster centroids (Hartigan and Wong 1979).	Non-determinacy; sensitive to initial condition; local optimum; each attribute has same weight; arithmetic mean is not robust to outliers (Teknomo 2008).	Large numbers of data (Teknomo 2008); median instead of mean overcomes the outlier's problem (Teknomo 2008); to combine with other techniques such as weight calculation (Adomavicius and Tuzhilin 2005).
Naïve Bayes	It assumes these words are independent; the computation far more efficient than exponential complexity of non-Naïve Bayes approaches (Yang 1999).	The independence assumption is often violated in the real world (Yang and Webb 2002).	To combine with other techniques, such as Tree Augmented Naive Bayes(Friedman et al. 1997).
Bayesian network	Represents a set of variables and their probabilistic dependencies (Friedman et al. 1997).	Exponential time for the general case (Haddawy, 1994, Yang and Webb 2002). Obtaining experimental data is often time consuming and costly (Tong and Koller 2001).	Combine with active learning algorithms (Tong and Koller 2001).
Markov chain	Future states are only dependent on the current states; allow systems to dynamically model URL access patterns (Sarukkai 2000).	Dimensionality (Markov chain matrix is typically very large) and needs large amount of training data (Sarukkai 2000); sensitive to parameter estimation (Nuel 2006).	Modelling site-specific transition models and within-site transition models; clustering sites before applying the Markov chain analysis; using sparse matrix storage representations. (Gamerman and Lopes 2006)
Association rule	Popular method for discovering relations (Agrawal et al. 1993).	High computational complexity; difficult to decide the threshold for support and confidence values (Hipp et al. 2002).	To combine with other techniques (Cho et al. 2002); develop additional rule quality measures (Hipp et al., 2002).
Decision tree	A predictive model which maps observations about an item to conclusions about its target value (Pazzani and Billsus, 2006); quality depends on both the classification accuracy and the size of the tree (Cho et al. 2002).	Not ideal for unstructured text classification tasks (Pazzani and Billsus 2006)	To combine with other techniques (Cho et al. 2002).
Neural networks	Learning or self-adapting ability (Frias-Martinez et al. 2006); recognise incomplete input pattern (Im and Park, 2007).	Local optimisation; black-box problem (Shin et al. 2000).	To combine with other techniques, e.g., case-based reasoning (Im and Park 2007).
SVM	Always finds a global solution; constructs a hyperplane, which minimises classification errors and maximises the margin between two input data sets (Joachims 1997, Bomhardt, 2004) ; performed very well in a high-dimensional dataset (Cheung et al. 2003).	It is currently considered slower at run time than other techniques with similar generalisation performance levels; black-box problem (Frias-Martinez et al. 2006, Isozaki and Kazawa 2002).	To remove useless features before other procedures (Isozaki and Kazawa 2002).

in some literature e.g. (Balabanovic and Shoham 1997, Adomavicius and Tuzhilin 2005, Adomavicius and Kwon 2007). In the remainder of this section, we will describe directions and issues from methodological perspective.

6.1 Context analysis for personalisation

Current personalisation systems make recommendations based only on user and item information and do not take contextual information into consideration. However, in some situation, contextual information is important (Gao et al. 2008). For instance, a customer may have different preferences for types of movies he wants to watch depending on whether he is with his partner or not (Adomavicius and Tuzhilin 2001).

In order to take contextual information into consideration, Chedrawy and Abidi (2006) proposed a method which uses multiple perspectives to represent context. Adomavicius and Tuzhilin (2001) proposed a general method to express multidimensional space (including contextual information) and calculate the utility function.

However there are still many issues in context-aware personalisation. Firstly, the context attributes in their approaches are the same for every user. Secondly, the contextual parameters are static, and therefore do not vary with user actions. Personalised and dynamic contexts are important for recommendation (Gao and Zhongfu 2009). For example, in the case of personalised content delivery on a web site, time and place are usually important contextual parameters to determine what content needs to be recommended to users, but they are not that important for some special users working at SOHO (small office/home office).

Pragmatics is a study considered with how context influences the meanings of signs (Liu 2000). From a pragmatic perspective, the meanings of signs will be different in different contexts; and the contexts are dynamic and depend on user purposes.

Consequently, the method of obtaining personalised and dynamic contexts is significant in constructing more efficient personalised recommendation systems.

6.2 Facilitating User Grouping with User Profiling Data in Social Networks

As we mentioned before, collaborative filtering approaches benefit from similar users preferring similar items. Thus finding similar users to a user is most important for collaborative filtering. It can be seen as finding a virtual community where users have similar interests.

Correspondingly, existing communities are useful for personalised recommendation. Recently, more and more people take part in social networks, playing games, sharing photos, obtaining knowledge, chatting, studying, and shopping. There are already a number of groups in social networks, e.g., Fackbook (http://www.facebook.com), QQ Group (http://group.qq.com), and Yahoo Group (http://groups.yahoo.com).

One of the most famous search engines, "Google" (http://www.google.com) has acquired a stake in Chinese popular social portal Tianya.cn (http://www.tanya.cn). Tianya offers twenty million registered users a variety of services including user blogs, classifieds, photos hosting, news, sports, and university information. Depending on the social network, Google has taken an important step to support personalised services, the next generation of Internet search (Zhang 2007).

In this context, two important lessons should be learned which are not stressed sufficiently in current personalisation literature. Firstly, personalisation and social communities are closely related. On the one hand, social networks are a source for collecting comments and user profile information that is needed for personalisation. On the other hand, personalisation is needed to make the information collected in communities useable. Secondly, personalisation is not only about grabbing information from a customer and using it to provide a personalised offer, but also concerns the building of long-term relationships between users, communities, and personalised services.

Common characteristics can be extracted from groups and communities in order to extract useful personal information. For example, if a person is a member of the "Disney Club Penguin" site frequently, when he/she goes to a book shop, books for children will have a high likelihood of being recommended (Zhang 2007). Park and Chang (2008) proposed a method to obtain customer profile models based on group behaviours such as clicks, basket insertions, purchases, and interest fields.

Analysing personal information across different groups in communities is a good idea. However, there are several issues: (1) the personal privacy problem: while privacy is already a problem in personalised recommendation systems, it is even more serious across communities. Current privacy management interfaces are woefully inadequate and their importance is only now being recognized (Ramakrishnan et al., 2001). How can systems or virtual communities open and use users' personal information and historical usage data in an appropriate way? (2) the relationship problem. What are the relationships between communities? Is semantic web (Berners-Lee et al. 2001) or pragmatic web (De Moor 2005, De Moor et al. 2002, Singh 2002) a potential method? (3) Computational complexity problem (Das et al. 2007). The data that comes from data of communities will be huge, how can these distributed data sets be processed in a scaleable method? How can we simplify the computational complexity problem? We believe that research into these interesting problems will help in developing the next generation personalisation systems.

6.3 Enhancing collaborative filtering with domain knowledge

Although collaborative filtering methods have achieved good results in personalised recommendation systems, they can be augmented by knowledge-based techniques, such as case-based reasoning and semantic analysis to improve recommendation accuracy and to address some of the limitations (e.g. new user, "new item" problems, "cold start"). The problem-solving principles of case-based reasoning make them a candidate for integration with similarity based information filtering methods. Chedrawy and Abidi (2006) proposed a case-based method for information personalisation using a hybrid method of item-based collaborative filtering and case-based reasoning. Liang et al. (2007) proposed a spreading mechanism for semantic expansion. The authors adopted an ontology-based semantic-expansion approach to building user profiles. By extracting domain knowledge from Web pages, this algorithm produces intermediate results that will be finally integrated in a machine-understandable format such as ontology.

Domain knowledge is significant for applying personalisation in different domains. For example, in the e-learning domain, when a user wants to learn something, personalised systems must know what he wants, the distance between what he wants and his prior knowledge, and what pedagogical relationships exist between different knowledge components e.g., the "Computing Curricula 2001" (CC2001) project report includes recommendations for both required core and elective components to be included at all levels of the curriculum. After obtaining results, systems recommend learners appropriate learning objects. In service discovery and composition, if a system wants to achieve personalised web service recommendations, not only does the system require information regarding the user and the item profile, but also IOPE (input, output, precondition, and effect) elements and OWL-WS (web ontology language for web services) .

Although a need for knowledge acquisition is a well-known bottleneck, knowledge in several domains is available in some structured machine-readable forms, e.g., ontology. However, domain knowledge is not designed for personalisation and information filtering. Therefore, the problem of applying domain knowledge to support personalised recommendation remains a problem.

7 Conclusions

Research in personalisation has made significant progress and has contributed to many successful personalisation systems. Large numbers of user profiling, content modelling, and recommendation filtering approaches are discussed. The current personalised recommendation systems still require further improvements to make personalisation approaches and techniques more effective and applicable to a broader range of real-life applications.

This paper presented topology and trend of personalisation through an analysis of a number of publications on the subject of personalisation in several important academic databases. It analysed a host of typical application domains and personalised service systems, including digital libraries, e-commerce, e-learning, news filtering, and search engine domains and applications. Various approaches were reviewed for user profiling, content modelling, and personalised recommendation. Several advanced modelling methods were summarised, including user behaviour, interest, and intention modelling. Content analysis and classification methods were also introduced. Four recommendation filtering approaches were discussed including rule-based, content-based, collaborative, and hybrid filtering. Then several machine learning techniques such as clustering and association rule mining which performed functions for user profiling, content modelling, and information filtering were discussed. Finally, we introduced possible extended methods that can provide better recommendation capabilities from methodological perspective.

We hope that the analysis and discussion in this paper will provide a starting point for researchers to construct more effective and useful personalisation and recommendation systems.

Acknowledgments We gratefully acknowledge valuable input from Dr. Lily Sun, Director of Postgraduate Studies, and Wenge Rong, PhD student, both from the School of Systems Engineering, the University of Reading. Thanks are also due to Hubert Grzybek from the Informatics Research Centre for his help in polishing the use of English.

References

Adomavicius, G. & Kwon, Y. O. (2007) New recommendation techniques for multicriteria rating systems. *IEEE Intelligent Systems*, 48–55.

Adomavicius, G., & Tuzhilin, A. (2001). *Multidimensional recommender systems: a data warehousing approach*. Lecture Notes in Computer Science: Proceedings of the Second International Workshop on Electronic Commerce. 2232.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on, 17*, 734–749.

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS), 23*, 103–145.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*, 207–216.

Albanese, M., Picariello, A., Sansone, C. & Sansone, L. (2004) Web personalization based on static information and dynamic user behavior. *Proceedings of the 6th annual ACM international workshop on Web information and data management*, 80–87.

Alsabti, K., Ranka, S. & Singh, V. (1998) An efficient k-means clustering algorithm. *First Workshop on High-Performance Data Mining,* 10.

Arthur, D. & Vassilvitskii, S. (2006) How slow is the k-means method? *Proceedings of the twenty-second annual symposium on Computational geometry*, 144–153.

Balabanovic, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM, 40*, 66–72.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*, 28–37.

Billsus, D. & Pazzani, M. J. (1999) A hybrid user model for news story classification. *Proceedings of the seventh international conference on User modeling table of contents*, 99–108.

Blanco-Fernandez, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., Lopez-Nores, M., Garcia-Duque, J., et al. (2008). A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. *Knowledge-Based Systems, 21*, 305–320.

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bomhardt, C. (2004) NewsRec, a SVM-driven personal recommendation system for news websites. *Web Intelligence, 2004. Proceedings. IEEE/WIC/ACM International Conference on.*

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Learning, 9*, 309–347.

Brin, S., Motwani, R. & Silverstein, C. (1997) Beyond market baskets: generalizing association rules to correlations. *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, 265–276.

Callan, J. & Smeaton, A. (2003) Personalization and recommender systems in digital libraries. *Joint NSF-EU DELOS Working Group Report, available at: www.dli2. nsf. gov/internationalprojects/working_group_reports/personalisation.html.*

Carchiolo, V., Longheu, A., Malgeri, M. & Mangioni, G. (2003) Courses personalization in an e-learning environment. *Advanced Learning Technologies, 2003. Proceedings. The 3 rd IEEE International Conference on.*

Chedrawy, Z. & Abidi, S. S. R. (2006) Case based reasoning for information personalization: using a context-sensitive compositional case adaptation approach. *Engineering of Intelligent Systems, 2006 IEEE International Conference on.*

Chen, C. M., & Duh, L. J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert Systems With Applications, 34*, 2298–2315.

Chen, Q., & Norcio, A. F. (1997). Modeling a user's domain knowledge with neural networks. *International Journal of Human-Computer Interaction, 9*, 25–40.

Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W. Y., & Wenyin, L. (2002). User intention modeling in web applications using data mining. *World Wide Web, 5*, 181–191.

Cheong, S. N., Kam, H. S., Azhar, K. M. & Hanmandlu, M. (2002) A web-based collaborative enabled multimedia content authoring and management system for interactive and personalized online learning. *Computers in Education, 2002. Proceedings. International Conference on*, 872–873.

Chesnais, P. R., Mucklo, M. J. & Sheena, J. A. (1995) The Fishwrap personalized news system. *Community Networking, 1995.'Integrated Multimedia Services to the Home'., Proceedings of the Second International Workshop on*, 275–282.

Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems, 35*, 231–243.

Chiu, W. (2001) Web site personalization.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems With Applications, 23*, 329–342.

Choi, O., & Han, S. Y. (2008). Personalization of rule-based web services. *Sensors, 8*, 2424–2435.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. & Sartin, M. (1999) Combining content-based and collaborative filters in an online newspaper. *ACM SIGIR Workshop on Recommender Systems*.

Conlan, O., Wade, V., Bruen, C. & Gargan, M. (2002) Multi-model, metadata driven approach to adaptive hypermedia services for personalized elearning. *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 100–111.

Dahn, I. & Schwabe, G. (2002) Personalizing textbooks with slicing technologies-concept, tools, architecture, collaboration use. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume*.

Das, A. S., Datar, M., Garg, A. & Rajaram, S. (2007) Google news personalization: scalable online collaborative filtering. *Proceedings of the 16th international conference on World Wide Web*, 271–280.

De Moor, A. (2005) Patterns for the pragmatic web. *Conceptual Structures: Common Semantics for Sharing Knowledge: 13th International Conference on Conceptual Structures.* Kassel, Germany.

De Moor, A., KEELER, M. & RICHMOND, G. (2002) Towards a pragmatic web. *Conceptual Structures: Integration and Interfaces: 10th International Conference on Conceptual Structures, ICCS 2002, Borovets, Bulgaria, July 15–19, 2002: Proceedings*.

Di Giacomo, M., Mahoney, D., Bollen, J., Monroy-Hernandez, A. & Ruiz Meraz, C. M. (2001) MyLibrary, a personalization service for digital library environments. *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries (= ERCIM Workshop Proceedings 01/W03). Dublin*.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology, 3*, 1–27.

Fink, J., & Kobsa, A. (2000). A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web. *User Modeling and User-Adapted Interaction, 10*, 209–249.

Ford, N. & Chen, S. Y. (2000) Individual differences, hypermedia navigation, and learning: an empirical study. *Journal of Educational Multimedia and Hypermedia,* 9.

Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2006). Automated user modeling for personalized digital libraries. *International Journal of Information Management, 26*, 234–248.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.

Gamerman, D. & Lopes, H. F. (2006) *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, Chapman & Hall/CRC.

Gao, M. & Zhongfu, W. (2009) Incorporating pragmatic information in personalized recommendation systems. *The 11th International Conference on Informatics and Semiotics in Organisations.* 156–164.

Gao, M., Zhongfu, W. & Liu, K. (2008) Pragmatic Grid for personalized resource provision. *Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on.*

Gauch, S. (2003). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems, 1*, 219–234.

Ha, S. H. (2002) Helping online customers decide through web personalizatio. *IEEE INTELLIGENT SYSTEMS*, 34–43.

Haddawy, P. (1994) Generating Bayesian networks from probability logic knowledge bases. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 262–269.

Han, J. & Kamber, M. (2006) *Data mining: concepts and techniques*, Morgan Kaufmann.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *Applied Statistics, 28*, 100–108.

Hipp, J., Guntzer, U. & Nakhaeizadeh, G. (2002) Data mining of association rules and the process of knowledge discovery in databases. *Industrial Conference on Data Mining*, 15–26.

Hofmann, T. (1999) Probabilistic latent semantic analysis. *matrix,* 50, 2.

Hofmann, T. (2003) Collaborative filtering via gaussian probabilistic latent semantic analysis. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 259–266.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS), 22*, 89–115.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D. & Rommelse, K. (1998) The Lumiere project: bayesian user modeling for inferring the goals and needs of software users. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 256–265.

Hsu, H. H., Chen, C. J. & Tai, W. P. (2003) Towards error-free and personalized Web-based courses. *Advanced Information Networking and Applications, 2003. AINA 2003. 17th International Conference on*, 99–104.

Im, K. H., & Park, S. C. (2007). Case-based reasoning and neural network based expert system for personalization. *Expert Systems with Applications, 32*, 77–85.

Instone, K. (2000) Information architecture and personalization. *White Paper, Argus Associates, INC., Dec.*

Isozaki, H. & Kazawa, H. (2002) Efficient support vector classifiers for named entity recognition. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7.

Jayawardana, C., Hewagamage, K. P., & Hirakawa, M. (2001). Personalization tools for active learning in digital libraries. *MC Journal: The Journal of Academic Media Librarianship, 8*, 1.

Jeevan, V. K. J. & Padhi, P. (2006) A selective review of research in content personalization *Library Review, Volume 55, Number 9, 2006 , pp. 556–586(31),* 55, 556–586.

Jin, X., Zhou, Y. & Mobasher, B. (2004) Web usage mining based on probabilistic latent semantic analysis. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 197–205.

Joachims, T. (1997) A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning,* 143–151.

Joachims, T., Nedellec, C. & Rouveirol, C. (1998) Text categorization with support vector machines: learning with many relevant. Springer.

Jokela, S., Turpeinen, M., Kurki, T., Savia, E. & Sulonen, R. (2001) The role of structured content in a personalized news service. *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on.*

Jung, S. Y., Hong, J. H. & Kim, T. S. (2005) A statistical model for user preference. *IEEE Transactions on Knowledge and Data Engineering*, 834–843.

Karagiannidis, C., Sampson, D. & Cardinali, F. (2001) An architecture for defining re-usable adaptive educational content. *Proceedings of the 2nd International Conference on Advanced Learning Technologies*, 21–24.

Kim, J. K., Cho, Y. H., Kim, W. J., Kim, J. R., & Suh, J. H. (2002). A personalized recommendation procedure for Internet shopping support. *Electronic Commerce Research and Applications, 1*, 301–313.

Kim, J. W. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *International Journal of Electronic Commerce, 5*, 45–62.

Kitts, B., Freed, D. & Vrieze, M. (2000) Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 437–446.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM, 40*, 77–87.

Kramer, J., Noronha, S., & Vergo, J. (2000). A user-centered design approach to personalization. *Communications of the ACM, 43*, 44–48.

Kuo, Y. F., & Chen, L. S. (2001). Personalization technology application to Internet content provider. *Expert Systems with Applications, 21*, 203–215.

Lai, H. J., Liang, T. P., & Ku, Y. C. (2003). Customized Internet news services based on customer profiles. *Proceedings of the 5th international conference on Electronic commerce*, 225–229.

Li, Y., Lu, L., & Xuefeng, L. (2005). A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce. *Expert Systems With Applications, 28*, 67–77.

Liang, T.-P. & Lai, H.-J. (2002) Discovering user interests from Web browsing behavior: an application to Internet news services. *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on.*

Liang, T.-P., Yang, Y.-F., Chen, D.-N. & Ku, Y.-C. (2007) A semantic-expansion approach to personalized knowledge recommendation. *Decision Support Systems,* In Press, Corrected Proof.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: item-to-item collaborative filtering. *IEEE Internet computing, 7*, 76–80.

Liu, B., Hsu, W. & Ma, Y. (1998) Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 80–86.

Liu, D. R., & Shih, Y. Y. (2005). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *The Journal of Systems & Software, 77*, 181–191.

Liu, F., Yu, C., & Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering, 16*, 28–40.

Liu, K. (2000) *Semiotics in information systems engineering*, Cambridge University Press.

Magoulas, G. D., Papanikolaou, & Grigoriadou, M. (2003). Adaptive web-based learning: accommodating individual differences through system's adaptation. *British Journal of Educational Technology, 34*, 511–527.

Masters, J., Madhyastha, T., & Shakouri, A. (2008). ExplaNet: A Collaborative Learning Tool and Hybrid Recommender System for Student-Authored Explanations. *Journal of Interactive Learning Research, 19*, 51–74.

Melville, P., Mooney, R. J. & Nagarajan, R. (2002) Content-Boosted collaborative filtering for improved recommendations. *Proceed-*

ings of the Eighteenth National Conference on Artificial Intelligence, 187–192.

Min, S. H., & Han, I. (2005). Detection of the customer time-variant pattern for improving recommender systems. *Expert Systems with Applications, 28*, 189–199.

Mobasher, B. (2007) Data mining for web personalization. *The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Berlin Heidelberg New York: Springer Verlag*.

Mock, K. J., & Vemuri, V. R. (1997). Information filtering via hill climbing, WordNet, and index patterns. *Information Processing and Management, 33*, 633–644.

Montaner, M., Lopez, B., & DELA Rosa, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review, 19*, 285–330.

Montgomery, A. & Smith, M. D. (2008) *Prospects for personalization on the Internet*, SSRN.

Mooney, R. J. & Roy, L. (2000) Content-based book recommending using learning for text categorization. *Proceedings of the fifth ACM conference on Digital libraries*, 195–204.

Morgan, E. L. (2002) *Making information easier to find with MyLibrary*, Infomotions, Inc.

Murthi, B. P. S., & Sarkar, S. (2003). The role of the management sciences in research on personalization. *Management Science, 49*, 1344–1362.

Nanopoulos, A., Katsaros, D., & Manolopoulos, Y. (2001). Effective prediction of web-user accesses: a data mining approach. *WEBKDD, 1*, 2001.

Narayanan, S., Bailey, W., Tendulkar, J., & Daley, R. (2002). Design of model-based interfaces for a real world informationsystem. *Systems, Man and Cybernetics, Part A, IEEE Transactions on, 32*, 11–24.

Nasraoui, O. (2005) World Wide Web personalization. *Encyclopedia of Data Mining and Data Warehousing, Idea Group*.

Niu, L., Yan, X. W., Zhang, C. Q. & Zhang, S. C. (2002) Product hierarchy-based customer profiles for electronic commerce recommendation. *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on,* 2.

Nuel, G. (2006). Pattern statistics on Markov chains and sensitivity to parameter estimation. *Algorithms for Molecular Biology, 1*, 17.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences, 61*, 217–235.

Papanikolaou, K. A. & Grigoriadou, M. (2003) An instructional framework supporting personalized learning on the Web. *Advanced Learning Technologies, 2003. Proceedings. The 3 rd IEEE International Conference on*, 120–124.

Park, Y.-J. & Chang, K.-N. (2008) Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications,* In Press, Corrected Proof.

Pazzani, M. J. & Billsus, D. (2006) Content-based recommendation systems. *The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science,* 4321.

Perugini, S., & Ramakrishnan, N. (2003). Personalizing web sites with mixed-initiative interaction. *IT Professional, 5*, 9–15.

Piatetsky-Shapiro, G. (1991) *Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley editors*, MIT Press, Cambridge, MA.

Pretschner, A. & Gauch, S. (1999) Ontology based personalized search. *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, 391–398.

Quarati, A. (2003) Designing shareable and personalisable e-learning paths. *Information Technology: Coding and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference on*, 454–460.

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y., & Karypis, G. (2001). Privacy risks in recommender systems. *IEEE Internet Computing, 5*, 54–62.

Renda, M. E., & Straccia, U. (2005). A personalized collaborative Digital Library environment: a model and an application. *Information Processing and Management, 41*, 5–21.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. (1994) GroupLens: an open architecture for collaborative filtering of netnews. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175–186.

Riecken, D. (2000). Introduction: personalized views of personalization. *Communications of the ACM, 43*, 26–28.

Ruvini, J. D. (2003) Adapting to the user's internet search strategy. *Proceedings of the 9th International Conference on User Modeling (UM2003), Pittsburgh*, 55–64.

Sakagami, H., & Kamba, T. (1997). Learning personal preferences on online newspaper articles from user behaviors. *Computer Networks and ISDN Systems, 29*, 1447–1455.

Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. *Computer Networks, 33*, 377–386.

Schein, A. I., Popescul, A., Ungar, L. H. & Pennock, D. M. (2002) Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval.* ACM New York, NY, USA.

Schmitt, B. & Oberlander, S. (2002) Evaluating and enhancing meta-search performance in digital libraries. *Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on*, 93–102.

Schubert, P. & Koch, M. (2002) The power of personalization: customer collaboration and virtual communities. *Proceedings of the Eighth Americas Conference on Information Systems (AMCIS)*, 1953–1965.

Scime, A. (1997) Taxonomic information retrieval (TAXIR) from the World Wide Web: knowledge-based query and results refinement with user profiles and decision models. George Mason University.

Shepherd, M., Watters, C. & Marath, A. T. (2002) Adaptive user modeling for filtering electronic news. *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, 1180–1188.

Shin, C.-K., Yun, U. T., Kim, H. K., & Park, S. C. (2000). A hybrid approach of neural network and memory-based learning to data mining. *Neural Networks, IEEE Transactions on, 11*, 637–646.

Singh, M. P. (2002). *The pragmatic web: preliminary thoughts*. Amicalola Falls State Park, GA: NSF-OntoWeb Workshop on Database and Information Systems Research for Semantic Web and Enterprises.

Sun, X., Chen, Z., Liu, W. & Ma, W. Y. (2002) Intention modeling for web navigation. *Proceedings of the 11th World Wide Web Conference (WWW)*.

Tang, C., Lau, R. W. H., Li, Q., Yin, H., Li, T. & Kilis, D. (2000) Personalized courseware construction based on web data mining. *Proceedings of the International Conference on Web Information Systems Engineering*, 204–211.

Teknomo, K. (2008) K-Means clustering tutorials. http://people.revoledu.com/kardi/tutorial/kMean/.

Tong, S., & Koller, D. (2001). Active learning for structure in Bayesian networks. *Proceedings of the International Joint Conference on Artificial Intelligence, 17*, 863–869.

Ubois, J. (1996). Cast system. *Internet World, 7*(94–98), 100.

Wallace, M., Karpouzis, K., Stamou, G., Moschovitis, G., Kollias, S. & Schizas, C. (2003) The electronic road: personalized content browsing. *IEEE MULTIMEDIA*, 49–59.

Wang, G. T., Xie, F., Tsunoda, F., Maezawa, H. & Onoma, A. K. (2002) Web search with personalization and knowledge. *Multimedia Software Engineering, 2002. Proceedings. Fourth International Symposium on.*

Wei, C. P., Yang, C. S., & Hsiao, H. W. (2008). A collaborative filtering-based approach to personalized document clustering. *Decision Support Systems, 45*, 413–428.

Weller, M. (2007). *Virtual learning environments: using, choosing and developing your VLE*. Oxon: Routledge.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record, 31*, 76–77.

Wu, K. L., Aggarwal, C. C. & Yu, P. S. (2001) Personalization with dynamic profiler. *Proceedings of the third International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems.*

Xu, C. Z. & Ibrahim, T. I. (2004) A keyword-based semantic prefetching approach in internet news services. *IEEE Transactions on Knowledge and Data Engineering*, 601–611.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*, 69–90.

Yang, Y. & Webb, G. I. (2002) A comparative study of discretization methods for naive-bayes classifiers. *Proceedings of PKAW*, 159–173.

Youm, S. H. & Cho, D. S. (2001) Personalized recommendation based on item dependency map. *Industrial Electronics, 2001. Proceedings. ISIE 2001. IEEE International Symposium on,* 1.

Yu, K., Schwaighofer, A., Tresp, V., Xu, X. & Kriegel, H. P. (2004) Probabilistic memory-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 56–69.

Zhang, Z. (2007) Community— the dawn of personalisation technology in the next generation Internet. http://googlechinablog.com/2007/08/blog-post_20.html 2008.

Min Gao is currently a Ph.D. candidate at Chongqing University, China. She received her M.Sc. degree in Computer Theory and Application from Chongqing University in 2005. Her recent research is primarily in personalization and recommendation in web computing and informatics.

Kecheng Liu Fellow of the British Computer Society and full Professor of Applied Informatics, is Director of the Informatics Research Centre at the University of Reading, United Kingdom. He has published over 150 papers in the journals and conferences, and serves in editorial boards of several journals. He is one of the key figures in Organisational Semiotics. His research includes information systems, pervasive informatics and intelligent systems

Zhongfu Wu is a senior professor, and Director of Network and Grid Research Institute, Chongqing University, China. He has been active in the research of e-learning, information systems and security, and high-performance computing for more than forty years and been broad known at research domain of Computer Science.