

1-1-2011

Development of personalized online systems for web search, recommendations, and e-commerce

Hao Wen

Ryerson University

Follow this and additional works at: <http://digitalcommons.ryerson.ca/dissertations>



Part of the [Mechanical Engineering Commons](#)

Recommended Citation

Wen, Hao, "Development of personalized online systems for web search, recommendations, and e-commerce" (2011). *Theses and dissertations*. Paper 732.

This Dissertation is brought to you for free and open access by Digital Commons @ Ryerson. It has been accepted for inclusion in Theses and dissertations by an authorized administrator of Digital Commons @ Ryerson. For more information, please contact bcameron@ryerson.ca.

DEVELOPMENT OF PERSONALIZED ONLINE SYSTEMS FOR WEB SEARCH, RECOMMENDATION, AND E-COMMERCE

By

Hao Wen

M.Sc., Lakehead University, Thunder Bay, Canada, 2005

B.Eng., North China Electric Power University, Baoding, China, 1996

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Mechanical Engineering

Toronto, Ontario, Canada, 2011

© Hao Wen, 2011

Author's Declaration Page

I hereby declare that I am the sole author of this dissertation.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

Signature

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this dissertation.

Please sign below, and give address and date.

DEVELOPMENT OF PERSONALIZED ONLINE SYSTEMS FOR WEB SEARCH, RECOMMENDATION, AND E-COMMERCE

By
Hao Wen
Mechanical Engineering, Ryerson University, 2011

Abstract

Personalized online systems for Web search, news recommendation, and e-commerce, are developed. The process of personalization of online systems consists of three main steps: determining a user's needs, classifying products or services, and matching the user's needs with suitable products or services. A multi-feature based method to automatically classify Web pages into categories of topics hierarchically representing the Web pages is proposed. An approach to modeling and quantifying a user's interests and preferences using the user's Web navigational data is presented. The approach is based on the premise that frequently visiting certain types of content or Web sites indicates that the user is interested in related content or retrieving information from those sites.

A personalized search system utilizing a Web user's interest, preference and search context models is developed. A Web user's interest and preference models are constructed and updated by analyzing the user's navigational data and automatically classifying Web pages. A user's search context model is used to determine how the user's interest and preference models impact his or her search behavior. An algorithm to re-rank search results generated by a conventional search engine is designed to provide a personalized Web search service. A hybrid recommender system for personalized recommendation of news on the Web is developed. Based on the similarities between Web pages and users' models of interest and preference, the Web pages are recommended to the users who are likely interested in the related topics. Moreover, the technique of collaborative filtering is employed, which aims to choose the trusted users and incorporate machine intelligence combined with human efforts. Once trusted users are determined, their behavior on the Web is considered as the manual recommendation part of the system. A method of classifying Web customers for planning customized e-marketing is proposed. The proposed e-marketing approach can be divided into four steps: determining a customer's general interest model, ascertaining a customer's local browsing model, classifying Web customers, and designing a personalized marketing and promotion plan for e-commerce based on the customer classification. Various experiments are carried out to demonstrate the effectiveness of the proposed approaches and systems.

Acknowledgements

I would like to acknowledge and extend my gratitude to many people for their support and help during my study and research at Ryerson University.

First and foremost, I owe a great amount of gratitude to my supervisors, Professor Liping Fang and Professor Ling Guan. This dissertation would not have been possible without their financial support, kind guidance and great patience. They not only gave me the knowledge and skills I needed to complete my dissertation, but also imparted to me their invaluable experience in every aspect of research. I am honored to have been their graduate student.

I am very grateful to the members of my dissertation examination committee: Professor E. Santos from Dartmouth College, Professor C. Ding, Professor M.W. Mohamed Ismail, and Professor S. Zolfaghari. Their comments and suggestions helped improve the quality of this dissertation.

I would like to thank my colleagues, especially Adrian Bulzacki and Jordan Sparks, in the Ryerson Intelligent Decision Support Systems Laboratory and the Ryerson Multimedia Research Laboratory. They inspired me in many ways during discussions.

I would like to extend my special appreciation to my friends who spent their time on my experiments. They showed sincere friendship and great patience.

I would like to give my gratitude to the Department of Mechanical and Industrial Engineering, Ryerson University, and the Ontario Graduate Scholarship (OGS) Program for their financial support.

This dissertation is dedicated to my parents, fiancée, and brother. I thank my parents for everything they have done for me, and I am indebted to my fiancée, Caixia Yang, for her love and support.

Table of Contents

Author's Declaration Page	ii
Borrower's Page	iii
Abstract.....	iv
Acknowledgements.....	v
Table of Contents.....	vi
List of Figures.....	ix
List of Tables	xi
Nomenclature.....	xii
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Objectives of the Research.....	6
1.3 Outline of the Dissertation.....	9
Chapter 2 Literature Review.....	11
2.1 Webpage Classification.....	11
2.2 User Modeling.....	15
2.3 Personalized Web Search.....	20
2.4 Recommender Systems.....	25
2.5 Electronic Marketing.....	30
Chapter 3 Web Page Classification.....	36
3.1 Introduction.....	36
3.2 Outline of the Proposed Web Page Classification Method	39
3.3 Term Extraction and Weight Calculation.....	41
3.4 Web Page Classification Based on Single Component.....	44
3.4.1 Web Page Classification Based on Meta Information	44
3.4.2 Web Page Classification Based on Effective Content	45
3.4.3 Web Page Classification Based on Directory Service.....	47
3.5 Classification Fusion.....	48
3.6 Experimental Results.....	52
3.7 Summary.....	56
Chapter 4 Web User Modeling.....	58

4.1 Introduction.....	59
4.2 Collecting a Web User's Data.....	61
4.3 Analysis of a Web User's Navigational Data.....	64
4.4 Web User Modeling.....	64
4.4.1 Naïve Bayes Theory.....	65
4.4.2 User's Interest Modeling.....	66
4.4.3 User's Preference Modeling.....	67
4.5 Experimental Results.....	69
4.6 Summary.....	71
Chapter 5 Personalized Web Search.....	73
5.1 Introduction.....	74
5.2 Conventional Web Search Engines.....	77
5.3 Personalized Web Search Engine.....	81
5.3.1 User's Search Context Model.....	81
5.3.2 Proposed Personalized Search Ranking.....	83
5.4 Experimental Results.....	87
5.5 Evaluation.....	91
5.6 Summary.....	96
Chapter 6 News Recommender System.....	97
6.1 Introduction.....	97
6.2 Architecture of the Proposed Recommender System.....	99
6.3 The Proposed Recommender System.....	101
6.3.1 User Modeling for the Recommender System.....	102
6.3.2 Web Content Recommendation.....	105
6.4 Experimental Results and Discussion.....	107
6.4.1 Evaluation of Recommender Systems.....	108
6.4.2 Sample Results.....	109
6.4.3 Evaluation of the Proposed News Recommender System.....	111
6.5 Summary.....	114
Chapter 7 Personalized E-Marketing.....	115
7.1 Personalization in E-Commerce.....	116
7.1.1 Benefits of Personalization.....	118

7.1.2 Personalized E-Commerce Models.....	120
7.2 E-Marketing.....	125
7.3 Customer Classification for E-Marketing.....	128
7.3.1 Customer Interest Modeling.....	129
7.3.2 User Local Browsing Modeling.....	130
7.3.3 Customer Classification for E-Commerce.....	131
7.4 Personalized E-Marketing Planning.....	132
7.5 Evaluation Criteria.....	133
7.6 Experimental Results.....	135
7.7 Summary.....	137
Chapter 8 Conclusions and Future Research.....	139
8.1 Main Contributions of the Dissertation.....	139
8.2 Future Research.....	142
References.....	144

List of Figures

Figure 1.1 Overall architecture of the research.....	8
Figure 3.1 The schematic of the proposed automatic classification.....	40
Figure 3.2 The screen scan of the ontology database WordNet.....	42
Figure 3.3 Examples of lexical identity.....	47
Figure 3.4 An example of retrieving manual classification for a Web page using the Google directory service.....	48
Figure 3.5 Calculating the degree of belief of Web page q belonging to category j	51
Figure 4.1 The architecture of the proposed user modeling system.....	60
Figure 4.2 The interface of the user modeling system.....	63
Figure 4.3 An example of collected user data.....	63
Figure 4.4 A tester's degree of interest on the three sport subjects.....	71
Figure 5.1 The architecture of the proposed personalized search ranking system (PSRS).....	75
Figure 5.2 An example of retrieving Web information by links in the CBC Web site.....	79
Figure 5.3 The interface of the Google directory-based search engine.....	80
Figure 5.4 The interface of the Bing query-based search engine.....	81
Figure 5.5 The rate estimation of the Web pages in a search result.....	86
Figure 6.1 The architecture of the proposed recommender system.....	100
Figure 6.2 The schematic of the Web page automatic classification method.....	104
Figure 6.3 The process of user modeling for the recommender system.....	104

Figure 6.4 The precision-recall curve of the system obtained through the participants' rating...	112
Figure 6.5 Precision rates with and without the proposed CF method.....	113
Figure 7.1 Value of online orders in Canada.....	116
Figure 7.2 An example of link personalization.....	124
Figure 7.3 An example of interface personalization.....	124
Figure 7.4 An example of content personalization using hybrid filtering.....	125
Figure 7.5 A partial screenshot of Amazon Web site.....	127
Figure 7.6 The main steps of the proposed customer classification framework.....	129
Figure 7.7 A customer's local browsing model.....	136
Figure 7.8 A customer's degree of potential purchase for electronics.....	137

List of Tables

Table 3.1 Precision rate of determining the effective content area.....	54
Table 3.2 Precision rates of three component classifications and classification fusion.....	54
Table 3.3 Precision rates of the classification fusion on a second level category.....	56
Table 4.1 A user's interest model.....	70
Table 4.2 A user's partial preference model.....	70
Table 5.1 An example of constructing a user's search context model.....	89
Table 5.2 An example of personalized re-ranking.....	90
Table 5.3 Users' average first-click on the personalized search results and the conventional search results.....	92
Table 5.4 Average ranking-efficiency of the personalized search and the conventional search	95
Table 6.1 Example of recommended news for a user at a certain time	111
Table 7.1 General e-commerce and personalized e-commerce	117
Table 7.2 Interaction models of personalized e-commerce systems	122
Table 7.3 A sample personalized e-marketing and promotion plan	133
Table 7.4 A user's partial degrees of interest	136

Nomenclature

A_u	accuracy rate of user classification
$B_{j,k}$	number of times that a customer has purchased product k in category j
C	dependent class variable
$CG(req)$	set of customers satisfying requirement req
d_j	document j
$DO(q)$	degree of popularity for Web page q
$dob(q,j)$	degree of belief of Web page q belonging to category j
$DP_j(q)$	probability of Web page q belonging to category j
$DS_s(q)$	relationship if Web page q is located in Web site s
$DU(j)$	degree of a user's interest on category j
$DW(s)$	degree of a user's preference for retrieving information from Web site s
E_{pm}	efficiency rate of marketing planning
$ER(m(k))$	estimated rate of the k^{th} Web page in a search result
F_n	feature variable n
$freq_{ij}$	number of occurrences of term t_i in document d_j
g_i	document frequency for term t_i in a collection
h	number of hours
idf_i	inverse document frequency of term t_i
I_{np}	number of items sold during the normal marketing time slice
I_p	number of items sold during the personalized marketing time slice
I_{rcmd}	number of relevant Web news items in a recommendation list
$I_{relevant}$	number of relevant Web news items in the entire list
j, j_1, j_2, j_3	category
l	number of days
M	number of documents in a collection
m_r	resulting vector of Web pages in search session r
N_i	number of Web pages that contain term t_i in all categories
n_{ij}	number of Web pages that contain term t_i in category j

N_{np}	number of customers who purchase any product during the normal marketing time slice
N_p	number of customers who purchase any product during the personalized marketing time slice
np_j	number of products in category j on an Internet shop
$p(in_j)$	degree of interest on category j
$p(in_s)$	degree of preference on Web site s
$p(not_in_j)$	probability that a user is not interested in category j
$p(not_in_s)$	probability that a user prefers not to retrieve information from Web site s
P_{high}	purchase rate of the customers with high DPP
P_{low}	purchase rate of the customers with low DPP
P_{nm}	the m^{th} product in the n^{th} category
$pr(q)$	probability of recommending Web page q to a user
q	a Web page
qk_r	vector of query keywords in search session r
q_{link}	navigational link of query page q
r	search session
req	range of degree of potential purchase
req_{max}	maximal value of req
req_{min}	minimal value of req
R_{num}	number of Web news items that are recommended to a user
$RR(m(k))$	re-rating value of the k^{th} Web page in a search result
s	a Web site
SC	set of categories defined in the system
S_{j_term}	set of terms in a document d_j
S_{q_cont}	text of q 's effective content information
S_{q_meta}	text of q 's meta information
SS	set of Web sources
$T(q)$	time factor for Web page q
tf_{ij}	term frequency of term t_i over document d_j

t_i	term i
$T_{j,k}$	total time that a customer has spent on browsing product k in category j
tn	number of participants
T_{np}	browsing time by all customers during the normal marketing time slice
T_p	browsing time by all customers during the personalized marketing time slice
U	set of customers
u_k	a customer
$view_q$	observation of opening Web page q
v_r	context vector of Web pages in search session r
W_{ij}	weight of term t_i over document d_j or category j
wk	number of weeks
$x_{1,i}$	participant i 's average rate of first-click on the personalized search results
$x_{2,i}$	participant i 's average rate of first-click on the conventional search results
η_1	factor for meta information classification
η_2	factor for Web content classification
η_3	factor for directory service
η_4	influence factor of a user's interest model on personalized search
η_5	influence factor of a user's preference model on personalized search
η_6	factor of importance on customer
η_7	factor of importance on item
η_8	factor of importance on browsing time

Chapter 1

Introduction

1.1 Background

The Internet is a global system consisting of a vast number of interconnected computer networks all over the world. It connects computers located in government organizations, schools, libraries, corporations, and homes through wired or wireless communications based on the Transmission Control Protocol and Internet Protocol (TCP/IP). Due to many conveniences offered by the Internet, affordable computer devices and internet access, there are billions of users and personal computers connected to the Internet, and are constantly growing. Canada's National Statistical Agency reported that "In 2009, 80% of Canadians aged 16 and older, or 21.7 million people, used the Internet for personal reasons, up from 73% in 2007 when the survey was last conducted" (Statistics Canada, 2010). By using the Internet, individuals can obtain news, send electronic mail, make audio and/or video calls, conduct distance learning, play games with people around the world, etc. It can be affirmed that the Internet has become an indispensable tool for the daily lives of people. Specifically, people rely on the wide variety of services that adhere

to the Internet. In this dissertation, any application service that requires data transfer between clients through the Web is defined as an online service.

Due to the Web market demands and rapid progress of software and network technologies, a significant number of online services have been developed and are widely used, including email, Web search engines, news distribution system, wiki systems, e-libraries, blog sites, e-bank Web sites, e-shopping Web sites, e-marketing systems, e-museum, and entertainment system. The nearly infinite amount of information and a large number of online services bring a digital lifestyle in which the Internet is helpful for people to satisfy their daily and working needs. However, everything has two sides. Because of the large amount of Web information and the individual diversity, a uniform interface and way of offering service will inevitably reduce some individuals' experiences of using the online services. With the explosive growth of the Internet and the rapidly increasing number of internet users, the personalization of online services of both academia and industry have been paying much more attention to the proficient skills and navigational needs of users. Eirinaki and Vazirgizannis (2003) define Web personalization as the process of customizing a Web site to the needs of specific users. They emphasize that a personalized Web site can take advantage of knowledge obtained from the analysis of a user's navigational behavior/data in correlation with the collected Web information such as the Web structure and content. This research expands the definition of Web personalization to all types of online services. A personalized online service is a targeted service that can provide a specific interface and/or result data to an individual or a group

of Web users based on the matching of available data sets and the collected profile of the users. The importance of the personalization of online services can be observed through a few cases.

Case 1: For a certain query, a conventional search engine provides the same search result to all users, regardless of different profiles and different interests of users. For example, when a user searches for "apple" using a conventional search engine, the search engine may provide the information of Apple Company on the top of the list, since it does not know whether the user is asking for the information of the apple fruit or the Apple Company. Therefore, the effectiveness of such search engines can be improved by integrating a user's model of interests/needs into the ranking process.

Case 2: As a repository of news and information, the Web has some inherent advantages over other forms of media, such as newspaper, radio, and television. The most significant advantage of distributing news on the Web itself is the infinite expandability of the Web. However, a news Web site can offer a large amount of daily news and information that exceeds a user's ability to consume it. At a traditional news Web site, a user may browse through it and filter out the uninterested items, which is time-consuming and tedious. For this reason, it is necessary and practical to develop a personalized news recommendation system that is based on a user's interests and preferences.

Case 3: Since the 1990s, e-business and e-commerce models have been implemented by many companies, for instance, AmazonTM, Apple iTunesTM store, and eBayTM. In order to make profit efficiently, an effective e-marketing strategy is required

by internet shops. Although many e-marketing strategies such as recommending best-selling items, random recommendation, and presenting similar items are available, the e-marketing strategy of recommending the items based on a user's desires is obviously the most effective way if the user's models can be accurately constructed.

From these cases, it can be seen that online service providers are challenged by how to effectively satisfy the various needs of different users. Traditional online services usually leave the process of refinement to users, which means users may spend time on searching for the relevant results themselves. In order to relieve users from tediously searching massive data sets, approaches to personalized online services attempt to intelligently find out the relevant results for each user/group-user according to user profiles. In other words, the personalization of online services is one of the solutions to improve the performance of information systems.

Today, through the work of different researchers, many methods and techniques regarding Web personalization have been developed (Yong *et al.*, 2005; Stamou and Ntoulas, 2009; Wen *et al.*, 2010). Because of the variety of online services, different types of personalization systems may have different goals, data sets, user sets, etc. For example, a personalized Web search engine aims to rank Web pages based on a user's information needs, while a personalized online shop is supposed to recommend to a user certain items that the user may purchase. These two types of online services have different goals: to satisfy user's information needs and to sell products; different data sets: Web pages and products; and different user sets: information diggers and potential

customers. Thus, the approaches to personalization may be apparently varied in different online service systems. However, from the perspective of research, personalization systems are more or less connected since they have three common elements: human, content, and a matching process (Liu *et al.*, 2004; Ardissono *et al.*, 2005; Wei *et al.*, 2005; Teevan *et al.*, 2010). In other words, most personalization systems need to face the challenges of effectively distinguishing users, classifying contents, and matching content with user profiles. Therefore, research on Web personalization covers a broad range of topics, such as user modeling, content based classification, collaborative filtering (CF), and system integration.

The topic of user modeling always occupies an important position in Web personalization because one of the prerequisites of conducting Web personalization is to determine a user's specific needs. In terms of the way to construct user models, it can be divided into two types: the explicit method and the implicit method. With an explicit user modeling method, a user's profile is structured directly from survey data or other user information. The advantages of explicit user modeling are a low computational cost and high accuracy method, while the disadvantages are dependency on user effort and limited scalability. Contrarily, a system of implicit user modeling can build a user's models without or with little participation from the user. The process of implicit user modeling is usually based on the analysis of a user's normal navigational data, instead of intentional enquiry. For example, Ng *et al.* (2007) presented a method to learn a user's preference by analyzing the user's click-through data on a search engine. Currently, the technique of

implicit user modeling has been widely used in many personalized systems, which implies the practicality of the implicit modeling method. Nevertheless, these personalized systems do not share a uniform model of constructing a user profile.

This research involves applying a method of implicit user modeling to different types of online services. A user's generally navigational data is used to construct a user's interest and preference models. By investigating the similarity between a user's models and available contents provided by the online services, the contents which are likely to satisfy the user's needs are presented to the user.

1.2 Objectives of the Research

The overall objective of this research is to provide a solution to the personalization of various online services. The proposed solution is based on the technique of user modeling and the integration of user models and online services such as Web search engines, Web news recommendation systems, and e-marketing systems. A user's interest and preference models are constructed by analyzing the user's navigational data. Three different types of personalized online services are developed by combining the user's models and the traditional online services. There are several sub-objectives of this research:

1. To develop a Web page classification method for analyzing a user's viewed Web pages. Using the proposed classification method, Web pages can be grouped into categories based on topics.

2. To design a system of user modeling that is used to identify a user's interests and preferences. A user's navigational data is tracked and analyzed by the proposed Web page classification system. A user's models of interests and preferences are constructed based on the statistics of the viewed items.

3. To develop a personalized Web search engine based on the proposed user modeling method. The proposed personalized search engine can re-rank the Web pages obtained by a conventional search engine according to a user's interest model.

4. To construct a personalized Web news recommendation system. News pages are classified using the proposed method of Web page classification. Based on the similarity between a news page and a user's models of interest and preference, the system will determine whether the news page is recommended to the user.

5. To propose an architecture of a personalized e-marketing system. In the system, a content-based filtering method is used to generate a personalized e-marketing strategy.

As mentioned above, this is related to several research fields, such as text information classification, user modeling, personalization, Web search engines, recommendation systems, and e-marketing systems. Figure 1.1 shows the overall architecture of this research. There are four components in the proposed architecture of

personalization system: Web page classification, user modeling, online services, and personalized online services.

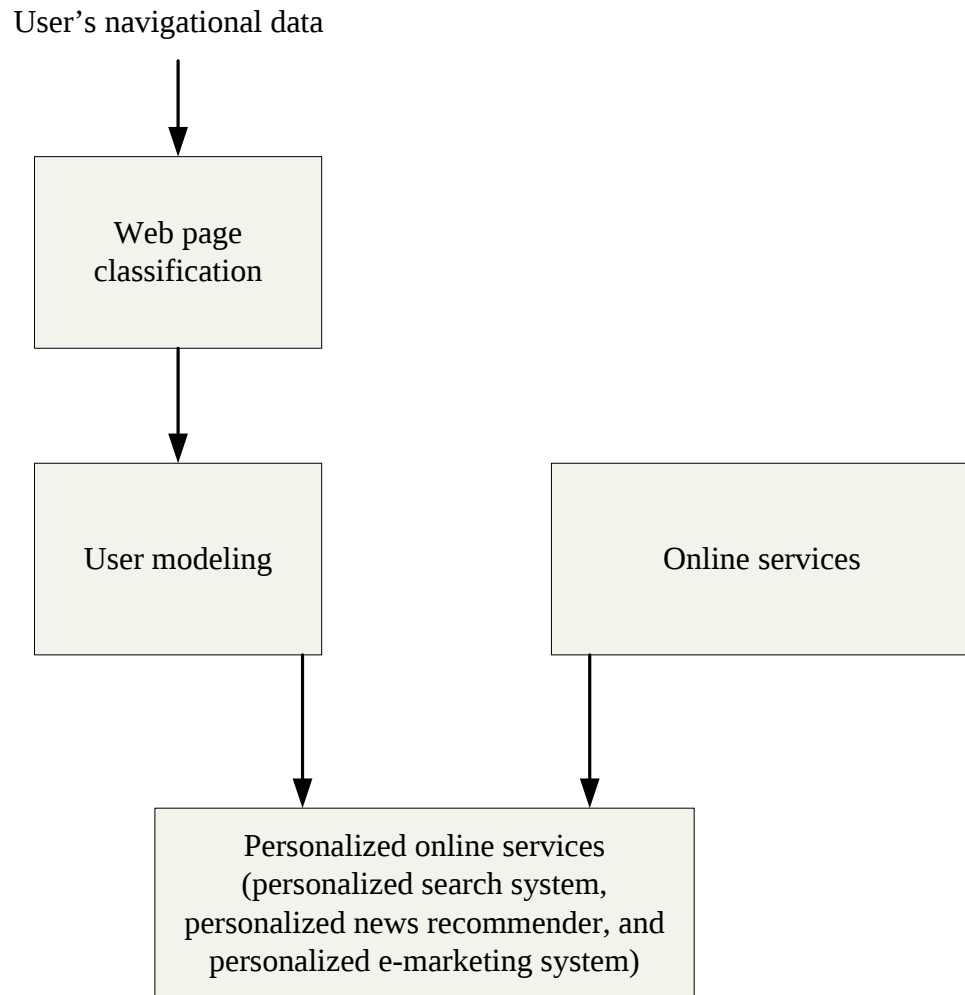


Figure 1.1 Overall architecture of the research

In the proposed architecture, a user's navigational data is tracked by the system as the raw data to model the user. Using the proposed Web page classification method, the viewed Web pages of a user are redirected to related category topics. A user's interest and preference models are constructed and updated by analyzing how often the user views the Web pages of each category. Three types of online services, search engines, news recommenders, and e-marketing systems are associated with this research. Each of them is integrated with the proposed user modeling method to provide a personalized service. The proposed personalized search engine aims to re-rank the search results based on a user's models. The personalized news recommender is developed to offer users the news of their interests. The proposed personalized e-marketing system is expected to present to a user the items that the user would potentially purchase.

In the rest of this dissertation, the proposed Web page classification method, the proposed user modeling system, and the personalized systems for search, news recommendation, and e-marketing will be discussed in details.

1.3 Outline of the Dissertation

This dissertation is organized as follows. In Chapter 2, related research on Web page classification, user modeling, personalized search systems, personalized Web news recommendation systems, and personalized e-marketing systems are reviewed. The proposed Web page classification method is presented in Chapter 3. How to build a user's

interest and preference models based on the user's navigational data is discussed in Chapter 4. A personalized Web search method based on the proposed user modeling method is presented in Chapter 5. The proposed Web news recommender system is discussed in Chapter 6, while the proposed personalized e-marketing system is presented in Chapter 7. Contributions are summarized in Chapter 8.

Chapter 2

Literature Review

As mentioned in Section 1.2, this dissertation is related to several research topics. In this chapter, a literature review is presented which is composed of five components. These five components are: Web page classification, user modeling, personalized Web search, personalized Web news recommender, and personalized e-marketing.

2.1 Web Page Classification

Web page classification can be considered as a specific case of text document classification or clustering. Ever since computers were invented, researchers have paid attention to text document classification. The computer revolution and the rise of the Internet in the 1990s have created a large amount of digital text documents, which have inspired researchers to develop a number of approaches to text document classification (Wang and Chiang, 2007; Isa *et al.*, 2009; Lan *et al.*, 2009).

The Dewey Decimal Classification (DDC) is a proprietary system of library classification developed by Melvil Dewey in 1876 (Scott, 2005). It has also been widely used in digital book classification with some revisions. The DDC attempts to organize all knowledge into ten main classes. The ten main classes are each further subdivided into ten divisions, and each division into ten sections, giving ten main classes, 100 divisions and 1000 sections. DDC's advantage in using decimals for its categories allows it to be both purely numerical and hierarchical. Yu and Xu (2008) experimentally compared the performances of blocking spam emails using four machine learning algorithms. The terms in an email are extracted and compared with certain, frequently used words that typically appear in spam emails. Depending on an appropriate threshold of similarity, an incoming email can be classified into two categories: accepted or rejected. Kim *et al.* (2005) proposed a dimension reduction method to reduce the dimension of the document vectors. They also introduced a decision function for the centroid-based classification algorithm and support vector classifiers to handle the classification problem where a document may belong to multiple classes. They concluded that with several dimension reduction methods that are designed particularly for clustered data, higher efficiency for both training and testing could be achieved without sacrificing the accuracy of text classification predictions even when the dimension of the input space is significantly reduced. Aggarwal *et al.* (2004) discussed the merits of building text categorization systems by using supervised clustering techniques. They claimed that completely unsupervised clustering has the disadvantage of being difficult to sufficiently isolate fine-grained classes of documents relating to a coherent subject matter. Therefore, the

information from a preexisting taxonomy is utilized to supervise the creation of a set of related clusters. They concluded that with partially supervised clustering it is possible to have some control over the range of subjects that one would like the categorization system to address with a precise mathematical definition of how each category is defined. They also suggested the supervised clustering as a priori knowledge to the definition of each category. Liu *et al.* (2009) proposed an imbalanced text classification method based on a term weighting approach. They found that categories with fewer examples are under-represented and their classifiers often perform far below satisfactory. They use a simple probability based term weighting scheme in order to better distinguish documents in minor categories. They claimed that the proposed approach using the term weighting method could improve the performance of minor categorization, while not jeopardizing performance for major categorization.

Although a number of methods for text document classification have been proposed, they cannot be directly employed to classify Web pages without any modification; because Web pages are semi-structured text documents with their own unique characteristics. For example, a Web page may be too short to be analyzed; a Web page may contain a lot of noise information such as advertisements. Therefore, researchers are trying to find solutions to work on Web page classification more effectively.

Among these works, one was presented by Kan and Thi (2005). They use text content features and uniform resource locator (URL) to classify Web pages for school

websites. Due to school Web sites having dedicated topics, they only need to categorize the Web pages into a few classes, such as faculty, course, student, and research pages. After using a support vector machine (SVM) based on maximum entropy to define the feature set, the Web pages can be automatically classified into the broad categories.

Some previous works have also put efforts on classifying Web pages hierarchically (Choi and Peng, 2004; Yang and Lee, 2004; Koraljka and Marianne, 2009). The work of Yang and Lee (2004) aimed to automatically create Web directories and their structure. They applied a text mining process on a corpus of Web pages to identify some important topics in this corpus. In their approach, a self-organizing map (SOM) neural network is constructed to cluster Web pages. In addition, they applied a recursive process to develop hierarchies of the identified topics and create a hierarchical structure of Web directories.

Some researchers applied the concept of ReliefF (Robnik-Sikonja and Kononenko, 2003; Wang and Makedon, 2004) in their automatic classification approaches. Jin *et al.* (2007) presented the work using ReliefF and hidden Naïve Bayes model for automatic Web pages categorization. Their work focused on finding relevant words for improving Web page classification performance based on ReliefF decision tree. After relevant words having been determined, a hidden parent in the Naïve Bayes structure is created for each word/attribute, which combines the influences from all other words/attributes.

Ru and Horowitz (2007) proposed a framework of classifying Web forms on e-commerce Web sites. In their proposed method, Web forms are identified by scanning for form tags. A neural network based classifier is applied to separate forms into different categories, such as search form and communication form. It can be observed that there is little noise information in such classification approaches.

Based on the discussion of related works, it can be found that little work has been carried out on the general Web page classification to topics with abundant noise information. The goal of this research is to attempt to integrate three collaborated classifiers for Web page classification in a noisy Web environment. In this work, the dedicated classifiers based on the directory feature, meta information feature, and Web page content feature are utilized and integrated.

2.2 User Modeling

In personalized online services based on a user's models, the content similarity between the user's profile and the products (Web pages) or services are analyzed and used to provide relevant products that satisfy the user's needs. In many personalized Web systems, a user's profile or model implies the user's attributes of interest and preference. The process of creating a user's profile or models is called user modeling. Different systems may have different structures of user models, since they have their own purposes, interfaces, user groups, etc. For example, a digital library may have a structure

of user models consisting of gender, age, and education level, while an online shop needs to model a user's age, location, and purchase interests.

In order to collect the user information and create user models/profiles, there are two approaches that are widely used: the explicit approach and the implicit approach. For the purpose of user modeling, an explicit approach is an approach to obtain a user's models/profiles through direct queries or surveys. An email subscription system usually inquires as to the user's topics of interest. This is a typical online service system using the explicit method to create a user's profile. The direct communication between an online service system and its users in the explicit approach can help the system to retrieve a user's accurate information very quickly. However, using the explicit approach to create a user profile has several limitations: the profile cannot update itself when the user's interests and preferences change over time; a user may be tired of answering inquiries during the construction of a profile; any change in the structure of a profile needs the user's participation. Therefore, implicit approaches that are flexible and capable of updating itself have been studied and applied to various personalized online services (Claypool *et al.*, 2001; Liu *et al.*, 2004; Chirita *et al.*, 2005; Frias-Martinez *et al.*, 2006; Tan *et al.*, 2006; Choi and Ahn, 2009; Jiang and Tuzhilin, 2009; Zo and Ramamurthy, 2009; Godoy *et al.*, 2010).

Because many users are unwilling to participate in the time-consuming process of constructing user interest and preference models, implicit approaches are increasingly applied in personalized online services to automatically build user models without

interrupting Web user's navigation (Liu *et al.*, 2004; Shen *et al.*, 2005; Agichtein, 2006; Radlinski and Joachims, 2007; Teevan *et al.*, 2010). The implicit approach to user modeling is a process to create and update user models by collecting and analyzing user data. A significant difference between explicit and implicit approaches is that implicit approaches need to calculate the collected user data for user modeling, while explicit approaches directly fit the collected user data into user models. Unlike explicit approaches which adopt the category (topic) structure for user profiles, implicit approaches can use either the structure of the category or the structure of bags of words to create user profiles. In the papers by Speretta and Gauch (2005), Ma *et al.* (2007), and Stamou and Ntoulas (2009), a user's interest and/or preference models are constructed by applying the structure of the category. In order to achieve higher effectiveness and accuracy, a user's interest model is usually hierarchically structured into categories. During a Web user's navigation, the user's navigational data including queries, feedback, and viewed Web pages, is tracked and analyzed by the personalized search system. Therefore, the user's interest model is identified by classifying the user's navigational data into hierarchical categories. In a user's Web search session, the Web pages obtained from traditional search engines are re-ranked based on the similarity between the user's interest model and the resulting Web pages.

The bag-of-words model is usually used in document classification, user modeling, natural language processing, and other applications of information retrieval. In the bag-of-words model, a piece of text instead of a set of categories is represented as the

description for a target such as a Web user's interests, preferences, images, or video clips. Sugiyama *et al.* (2004) assumed that a user's preference model consists of two aspects: persistent preferences and ephemeral preferences. Therefore, a user's preference model is constructed by summarizing the user's navigational history including both long-term and short term preferences into a vector of terms. Luxenburger *et al.* (2008) discussed how to match user task profiles and needs in a personalized Web search. The language model is used to represent a user's task profile, which collects the user's various navigational components such as queries, result clicks, click-stream documents, and query-independent browsed documents, while user needs are represented by the queries in the current search session. Daoud *et al.* (2008) studied how to aggregate concept-based short-term interests to represent a Web user's long-term interest model. Chirita *et al.* (2006) argue that desktop information is also substantial to build a user's interest model. In their work, both the navigational information and the desktop information of a user were utilized to construct the user's rich model of interests.

As discussed above, different Web systems may need different attributes of users to create user profiles. However, many online systems attempt to capture users' interests and preferences in order to provide a better personalized online service. Some researchers have presented their methods of building an implicit user interest model. Qiu and Cho (2006) proposed a framework to investigate the problem of personalizing Web searches based on users' past search histories without involving users' efforts. In their model, a user's interests in Web pages are formalized and correlated with the user's clicks on

search results. Then, a refined user interest model is built based on the correlation. Gunduz and Qzsu (2003) also proposed a user interest model for Web page navigation. Their approach relies on the premise that the visiting time of a page is an indicator of the user's interest in that page. They first partitioned user sessions into clusters such that only sessions which represent the similar aggregate interests of users are placed in the same cluster. Then a model-based clustering approach is employed to partition user sessions according to similar amount of time in similar pages. The resulting clusters are used to recommend pages to users that most likely contain the information which is of interest to them at that time. Although many approaches have been proposed, there is a lack of user interest models that can easily be integrated into various online services. This research is motivated by such intent that a user's interest and preference models are developed based on category topics and Web sources that can be utilized by various personalized Web systems.

One objective of this dissertation is to develop an approach to model and quantify a user's Web search interests using the user's navigational data. The approach is based on the premise that frequently visiting certain types of content indicates that the user is interested in that content. As a user surfs on a Web page, the meta-information and the content of the Web page are combined to represent the navigational content. A method of classifying Web pages using significant terms and weights is utilized in order to identify the Web page viewed by the user (Wen *et al.*, 2008a). This research employs the Naïve Bayes theory to update a user's interest and preference models.

2.3 Personalized Web Search

Since the Internet went public, a significant number of online services have been developed and widely used. Web search engines are the most popular and useful among these. In terms of retrieval method to a database and interface to users, the Web search engines can be grouped into either directory style or query style (Brin and Page, 1998; Capra and Perez-Quinones, 2005). In a typical directory-based Web search engine, the Web sites collected by the search engine are classified hierarchically through analysis of their content. The typical interface of a directory-based Web search engine usually lays out the titles of the top categories. After one or more queries by a Web user, a list of the Web sites attributed to a certain category is presented to the user. As a result of the explosive development of the Internet since the 1990s, directory-based Web search engines have gradually exposed their disadvantages: the massive number of Web sites in each category makes it nearly impossible for users to look through the vast list of Web sites; the required human effort for manual Web page classification and the complexity of automatic Web page classification decrease the efficiency of directory-based Web search engines.

In a typical query-based Web search engine, the Web pages collected by crawling software or client submissions are indexed by all terms appearing in the page's title and effective content (Baeza-Yates and Ribeiro-Neto, 1999). When a Web user's request is presented in the form of keywords, a series of Web pages, ranked by the match between the requested keywords and the terms in the Web pages, are presented to the user.

Query-based Web search engines have become dominant over the last decade because they have significant advantages over directory-based Web search engines: the algorithms for identifying and matching terms are relatively simpler than creating algorithms for Web page classification; there is only one step from a user's query to the display of some results, which shows the high efficiency of query-based Web search engines in some cases. However, query-based Web search engines have not reached optimal potential because of the diversity of individuals. For example, when a query of "latest sports news" is submitted to a query-based Web search engine, the top Web site presented by the search engine is the same for all users. This ignores the fact that an individual may prefer to retrieve sports news from a certain Web site rather than from others. Therefore, a personalized search engine system is needed to improve the search quality.

Research on personalization of Web search engines has been widely carried out, especially after traditional search engines demonstrated their usefulness and commercial value. Various solutions and approaches to the personalization have been presented by researchers, including personalized Web searches based on content analysis and based on an individual's search behavior, search re-ranking based on the hyperlink structure of the Web, and personalized searches based on user groups. In this section, related work on the personalization of Web search engines is presented and discussed.

In personalized searches based on a user's content choice, the content similarity between a user's intents and the search results (Web pages) are computed and used to re-

rank Web pages. As discussed in Section 2.2, a user's interests and intents can be estimated using either the explicit or implicit approaches of user modeling. A user's interests can be expressed by hierarchical category topics or bags of words. In addition to the user's content-based information, the user's behavior-based information is usually exploited to construct user models for personalized online services (Shen *et al.*, 2005; Teevan *et al.*, 2010). Among various data sets of the Web user's navigational behaviors, the data vector of a user's query and the data vector of the user's response to resulting the Web pages are often used to generate the user's search behavior model for personalization of Web searches. Tyler and Teevan (2010) utilized large scale query logs to investigate re-finding behavior in Web search. They showed that the re-finding behavior could be exploited as a means to improve the search experience. Shi and Yang (2007) proposed a query extension method based on the vector of the user's query log. In their work on query, a list of related queries is suggested from analysis of the user's previously submitted queries. Many works are also focused on employing a user's click-through information in order to personalize Web navigation. Shen *et al.* (2005) proposed a search-behavior based on a re-ranking method that aims to improve the effectiveness of a single search session. Once a user clicks on a Web page among the search results, the un-clicked Web pages will be re-ranked based on the similarity between the clicked Web page and each un-clicked Web page. Leung *et al.* (2008) proposed a click-through data based personalized query suggestion system. The proposed system extracts concepts from clicked Web pages, from which the user's search intent can be modeled. Therefore,

related query suggestions can be presented by merging the initial query and the user's intent model.

Some personalized Web search approaches rely on the search results of an existing search engine. The additional ranking of Web pages is carried out based on the search results given by conventional search engines. Most of those methods reviewed earlier in this section follow this process. However, many other works focus on the direct modification of page ranking algorithms of existing search engine systems. Page *et al.* (1998) discussed how to compute PageRank for large numbers of pages, and then presented the personalization of PageRank. Although there are many algorithms of ranking Web pages proposed for information retrieval systems (Chan and Chen, 2007; Meiss *et al.*, 2008; Bhatia and Kumar, 2009; Hwang *et al.*, 2010), the algorithm of PageRank is the most broadly adopted among them. Therefore, many methods of Personalized PageRank (PPR) have been proposed in the last decade (e.g. the methods of adjusting the global PageRank algorithm based on an individual's models). Haveliwala (2003) presented a modified PageRank algorithm that computes the topic-sensitive PageRank scores using the context information containing the query terms. Guo *et al.* (2007) utilized two factors to bias PageRank results for personalization: the length of time that a user spent visiting a page and the frequency that a page was visited. Su *et al.* (2008) proposed an approach for personalized page ranking by integrating the mining of the user browsing behaviors and PageRank in order to bridge the gap between global search results and local preferences. With respect to the feasibility of PPR, Pathak *et al.*

(2008) discussed how to build indices for real-time PPR and the trade-off between index size, preprocessing time, and query speed.

With the rapid development of social Web applications, some strategies that incorporate the commonality of a group of users have also been presented to help with personalized Web search. Using the approach of personalized searches based on user groups, Web users are usually clustered into different groups of varied topics by analysis of their profile data. Some methods have been proposed to achieve personalized Web search by taking into account the similarity of Web users. Smyth (2007) presented a collaborative Web search method that merges the results from a conventional search engine and the community's search knowledge base. Biancalana and Micarelli (2009) proposed a personalized web search approach based on a social tagging service. The social tags are explored by the system for the query expansion of a conventional search engine. Xue *et al.* (2009) proposed a method of collaborative personalized searches that groups Web users through a cluster-based language model. Personalized searches are carried out based on employing both the group information and the specialties of individuals for global ranking.

A personalized web search approach is developed in this research that involves implicit user modeling (user content), behavior based analysis, and rank modification. The purpose of this work is to propose a method that can utilize various aspects of user modeling to improve personalized Web searches.

2.4 Recommender Systems

Recently, significant attention has been paid by both academia and the industry to Web-based recommender algorithms because of their potential for being applied to personalized online services (Adomavicius and Tuzhilin, 2005a). For example, the method of personalized recommendation has been applied to the fields of e-commerce and e-business, which has proven the feasibility of recommendation systems through practical implementation. Among those systems incorporated with the method of personalized recommendation, the online retailer Web site Amazon (<http://www.amazon.com>) is a typical recommendation system that successfully utilizes an extensive range of different recommendation methods. The primary objective of a Web-based recommendation system is to alleviate a user's manual effort by automatically filtering the Web products in a large information repository based on the relevance between the user's needs and the products. Therefore, the technique of personalization performs an important role in recommendation systems by addressing the issue that a personalized recommendation can be achieved only if the users are distinguishable. Three techniques of recommendation have been widely studied and utilized; these are the content-based technique, collaborative filtering, and the hybrid technique (Burke, 2002; Herlocker *et al.*, 2004; Konstan, 2004; Yong *et al.*, 2005). These three techniques and their applications are discussed in the following part of this section.

A content-based recommendation system brings forth recommendations based on two aspects: a user's preferences derived from either the explicit or implicit methods, and

the features associated with the Web products stored in the system. The research on content-based recommendation systems is usually focused on the user-specific classification problem (Santos *et al.*, 2003; Frias-Martinez *et al.*, 2006) and integrating user modeling approaches into recommendation systems (Godoy *et al.*, 2010). Lancieri and Durand (2006) proposed a method to model a Web user's behavior. Their approach focused on the traces of a user's activities and the analysis of Web contents. A comparative analysis of Internet navigation traces, including hyperlinks and keywords, is used to characterize an individual's behavior as the user is accessing the Web. Choi and Ahn (2009) presented a method for identifying a customer's preferences and recommending the most appropriate product. In their proposed recommendation system, analyzing a user's real-time behavior data from his or her Web usage, such as the content being viewed, the placing of online orders, and purchasing products, constructs the user's model. Therefore, a user's preference model is not directly generated by the explicit query from the user, but as a result of the information on the user's usage behavior concerning the products. Liang *et al.* (2008) adopted a semantic-expansion approach to create a user's profile by analyzing the documents previously read by the user. They concluded that the proposed semantic-expansion approach outperforms the traditional keyword approach in collecting user interests. A number of works are also focused on the systematical integration of user classification and the recommendation method in various fields.

The technique of content-based recommendation can be widely integrated into various types of information systems, such as e-commerce, e-libraries, and news

distribution systems. Gao and Li (2007) proposed an online intelligence recommendation system for personalized internet shopping. In their proposed system, an interaction method is designed for the system to obtain a user's needs for less frequently-purchased products. Subsequently, fuzzy logic and marketing strategies are used to recommend high potential and suitable promotion products for each individual customer. Weng *et al.* (2009) designed a recommendation system for movie promotion. In their proposed system, the integrated contextual information is considered as the foundation concept of the multidimensional recommendation model. Online analytical processing (OLAP) is employed in the system to solve the contradicting problems among hierarchy ratings. Wang and Shao (2004) proposed a Hierarchical Bisecting Medoids (HBM) algorithm to cluster users based on time-framed navigation sessions. Thereafter, the time-framed cluster based user modeling is utilized in the proposed e-learning recommendation system. In the last decade, several approaches to news recommendation systems using content-based technique have been presented (Hsieh *et al.*, 2004; Ha, 2006; Lee and Park, 2007; Li *et al.*, 2010). Ha (2006) focused a news recommendation system on the basis of users' preferences, where the user's preference score prioritizes recommended articles according to the relevance between a user's preferences and the articles. Lee and Park (2007) designed a mobile Web news recommendation system (MONERS) that incorporates attributes of news articles and user preferences with regard to category topics. Li *et al.* (2010) modeled news recommendation systems as a contextual bandit problem.

In a recommendation system, if feature extraction and classification of content are available and feasible, content-based filtering can be applied. Collaborative filtering (CF), on the other hand, is usually adopted when the content representation is ambiguous or complicated. Therefore, CF is an approach used to avoid complicated content-classification for a recommendation system. There are two types of CF: user-based CF and item-based CF. In a user-based CF system (Miller *et al.*, 2004; Yu *et al.*, 2004; Cho *et al.*, 2007; Jiang and Tuzhilin, 2009; Bernstein *et al.*, 2010), a user's rating function is represented as a vector. Then a method of calculating similarity such as the cosine similarity measurement is used to compute similarities between all user pairs. Consequently, the users are able to be grouped by the similarity estimation. The last step of conducting a user-based CF is to determine the recommended products based on the users' common behaviors or preferences within the same user group. Item-based CF has gained popularity and success because of its theoretical and computational simplifications (Cheung *et al.*, 2004; Wei *et al.*, 2005; Bobadilla *et al.*, 2010). For example, Amazon (www.amazon.com) is a commercial recommendation system that utilizes item-based CF in many aspects. In an item-based CF recommendation system, the similarities of products are computed, which is different from the user-based CF recommendation systems. A simplified solution to the item-based CF system is to track a user's navigational behaviors, then products which have a high similarity value to the user's selection of browsing items are recommended.

In order to improve the effectiveness and the efficiency of a system, hybrid recommendation systems are motivated by applying both content-based filtering and collaborative filtering in a single framework (Burke, 2002). Therefore, hybrid recommendation systems are usually built based on both user grouping and product classification. Numerous approaches to integrating collaborative filtering and content-based filtering have been proposed. Ardissono *et al.* (2005) developed a multi-agent infrastructure for Web-based recommendation systems. In their proposed system, a user can dynamically enable and disable the built-in recommendation engine due to the flexibility of the multi-agent infrastructure. Moreover, the system can suggest information to possibly heterogeneous tourist groups. Choi *et al.* (2010) proposed a recommendation system that utilizes the nearest and farthest neighbors of a target customer to yield a reduced dataset of useful information in order to avoid scalability and sparseness problems when confronted by tremendous volumes of data. Recently, much attention has been paid to the safety and privacy issues of recommendation systems (Hurley *et al.*, 2007; Mobasher *et al.*, 2007). Mobasher *et al.* (2007) investigated the safety issues of user-based, item-based, and hybrid recommendation systems. They concluded that both user-based and item-based algorithms are highly vulnerable to specific attack models, while the hybrid algorithms may provide a higher degree of robustness.

A Web-based hybrid news recommender system that can combine the advantages of both content-based filtering and collaborative filtering is developed in this work. The

proposed recommender system utilizes the techniques of Web page classification, implicit user modeling, content-based filtering, and collaborative filtering. The goal of the proposed system is to provide a tool that is able to relieve Web users from repetition and tediousness of Web surfing to some degree.

2.5 Electronic Marketing

The World-Wide-Web (WWW) is a global information system which can be accessed via the Internet. Because of the remarkably rapid progress in information technology, people can use WWW significantly easier and faster than before (Aspray and Ceruzzi, 2008). The continuous decrease in prices of digital equipment makes it possible to digitize most information and present it online. This includes: text information, pictures, soundtracks, and video clips. Therefore, the Web has become the most effective and powerful information source for individuals (Ridley, 2009). The Web is also becoming an important platform for business, because convenience is making it more likely for people to shop online (Papastathopoulou and Avlonitis, 2009). The change from on-site commerce to online commerce makes it important to develop effective e-marketing and promotion models for Web service companies.

E-marketing or internet marketing is one of the key components of e-commerce, which is the marketing of products or services over the Internet (Varadarajan and Yadav, 2009). Compared to traditional marketing, e-marketing has an intrinsic advantage in that

it can reach global customers with relatively lower advertising budgets, because the media of e-marketing, the Web, is worldwide and affordable to average individuals. Currently, online purchases by both businesses and consumers are dramatically increasing, which makes e-marketing a strategic way to achieve significant benefits and increase customer satisfaction for enterprises (Chaffey, 2009). However, some limitations of e-marketing prevent it from being a high efficient approach. One of the obstacles is that an enterprise can hardly provide customized services and promotions to an individual without knowing details about the person, which likely leads to less profit for an enterprise (Long and Tellefsen, 2007).

In order to make Web services more effective, researchers have recently paid significant attention to the integration of Web service and user modeling. In these approaches, a Web page can treat a user profile intelligently based on the individual's interests, intent, or context (Zhu *et al.*, 2003). Generally, in terms of the way in which user information is acquired and analyzed, there are two types of user modeling for a personalized Web service: implicit and explicit user modeling (Allen, 1990), which has been described in Section 2.2. The method of explicit user modeling is more accurate than implicit user modeling at the time of completing a survey (Zigoris and Zhang, 2006). However, for a user's general interest model, it is not practical to ask the user to express interest for every single category. Due to the complexity of an individual's intents and behaviors, meaningfully modeling a Web customer's interests, preferences, intents, or prospective behaviors remains a competitive challenge to researchers (Laroche, 2010).

Currently, a common methodology for grouping Web customers for Web services is combining content-based and collaboration based user classification. For instance, when a Web customer is searching and browsing an internet shop, such as Amazon, the user's navigational data/products are used to cluster the customer into a user group who has viewed the same products. Other group members' navigational data are recommended to the customer for the next round of navigation. Obviously, an e-marketing system based on this type of Web user classification is unable to satisfy the increasing requirement for accurate targeting for personalized e-commerce (Kalaighnam *et al.*, 2008).

The research reported in this dissertation is driven by the intent to effectively integrate Web user classification with e-marketing to provide personalized Web services. One objective of this research is to propose a framework for a dedicated user classification system for personalized e-marketing and promotion planning, which is motivated by improving profit and expanding impact for Web service companies. In the proposed framework, a Web customer's navigational data is used to create the general interest and local browsing models for the customer. The degree of potential purchase (DPP) is calculated based on the customer's models. Therefore, customers can be classified according to their DPPs, which paves the way for a personalized e-marketing strategy.

Web user classification has attracted significant attention from researchers, due to the demand for rapid progress in the development of personalized Web services. If a Web customer can be precisely classified and understood by a service system, customized

service with accurate targeting and high efficiency becomes possible and practical. In the last two decades, several methodologies for Web user classification have been proposed. The survey method is a classic and effective way to directly determine customer intent and preference. This method is valid for both traditional services and Web service systems (Agichtein, 2006; Ridley, 2009). Haug (2006) discussed the advantages and limitations of Web surveys applied in population censuses. Haug (2006) indicates that the online population census is superior with regard to the costs of data entry, editing, interface, and accessibility. However it has some limitations such as identity validation, security of connection, and confidentiality. In 2007, the National Library of the Netherlands published a survey report on Web archiving (Ras and Bussel, 2007). Web users were immediately classified by answering the questionnaires. Five main user groups are mentioned in the report, which are: researcher (student, researcher, sociologist, historian, linguist, etc.), journalist, lawyer (lawyer and patent agent), consumer, and a group of other targets. The precise user classification made the survey analysis very reliable and convincing. Therefore, the report claims that the user test offered very valuable information.

Though explicit user modeling and user classification methods have the advantage of high accuracy, their limitations of low convenience and restricted scalability make them unsuitable for complex environments. Therefore, researchers are paying significant attention to implicit and hybrid user classification methods. Digital libraries have become a popular platform for developing personalized Web services. A number of user

classification methods have been developed and implemented for personalized digital libraries (Theng *et al.*, 1999; Cohen *et al.*, 2000; Di Giacomo *et al.*, 2000; Frias-Martinez *et al.*, 2006; Avancini *et al.*, 2007). These approaches for personalized digital libraries show various interfaces, search engines, and performance. Nevertheless, they basically use the common methodology of user modeling. In these digital library systems, the servers use the user's login information, including their username and their IP address, to distinguish the users. Based on the user's registration information, a plain classification method is used to group the users by gender, age, department, faculty, and position. Subsequently, the content-based user modeling method is used to estimate the user's interests, preferences, and behaviors in the library. The users can be classified by analyzing their user models; for example, users who have viewed the same book are grouped. Similarly, the idea of content-based user modeling is also widely used in general Web user classification. In the approach of Web user clustering presented by Rangarajan *et al.* (2004), a clustering algorithm based on the ART1 version of the adaptive resonance theory is applied. The URLs accessed by a Web user are extracted from the Web log files on the server. The vector generalized from the URLs most frequently accessed by all cluster members represents each user cluster. Liu and Keselj (2007) presented a method of classifying user navigation patterns using the combined mining of Web server logs and Web contents. In this approach, the textual content of Web pages is captured through extraction of character N-grams, which are combined with Web server log files to derive user navigation profiles. Some dissimilarity measure methods are used to calculate the profile dissimilarity.

Recently, due to rapid technological advances in the Web and information technology, e-commerce has become an important pillar for the development of a company. The prevalence of personal computers boosts the prosperity of the Internet business-to-consumer (B2C) market. In order to improve performance, personalized e-commerce models have been widely utilized by Web service companies. Researchers are also putting effort into seeking highly efficient personalized e-commerce models (Lee *et al.*, 2002; Changchien *et al.*, 2004; Wen *et al.*, 2010). Lee *et al.* (2002) by presenting an intelligent agent-based system for personalized recommendations in e-commerce. The proposed system consists of two sub-systems: user modeling and recommendation. An agent-based methodology is used for customer modeling, information gathering, information managing, and evaluation. A multi-attribute decision making method is used for recommending products to a user. The flawed performance of user preference modeling is also discussed. Changchien *et al.* (2004) developed an online personalized promotion decision support system for e-commerce. The proposed system has three modules: marketing strategies, promotion patterns model and personalized promotion products. Customer behaviors from the three categories are analyzed through data mining techniques and statistical analysis. Different marketing plans are used for customers in different categories. There are a number of personalized Web service models in the literature. Many of them suffer from obscurity of the user modeling approach. Therefore, developing an effective Web user classification method has become a critical issue for improving the quality of Web services.

Chapter 3

Web Page Classification

Document classification is a research topic in information science. The objective of document classification is to group digital documents into different categories based on analyzing one or more features of each document. Since Web pages are a type of digital document, Web page classification is considered a specific document classification problem. Unlike pure-text classification, which is mainly focused on analyzing literal contents, Web page classification deals with complex digital objects that have more features and noise.

3.1 Introduction

Along with the rapid progress of electronics, communication networks, and computer technology, the Internet has become one of the largest platforms for information production, repository, distribution, and communication. Every day, tremendous numbers of Web sites and Web pages are created, both by companies and individuals, due to easy

access of Web content management systems and the low costs of Web space (Shuen, 2008). However, while the net brings people great convenience, it is also accompanied with new technical challenges, such as how Web users can efficiently retrieve specific Web pages that they are interested in from the immensely overwhelming information depository. Commonly, there are three approaches by which a Web customer can access the locations of Web contents: memory, hint, and tools. For Web sites that a Web user often browses on, the user can always save the location of the Web sites either in their mind or within their favorites list in the browsing software. The hint approach means that a user retrieves a Web page through the address link embedded at other Web pages. The limitations of the approach of memory and hint are the requirement of much human effort and randomness, respectively, which increasingly encourages Web users to choose the approach of tools to retrieve Web information. Currently, internet search engines are the most popular tools for Web information retrieval. In terms of data structure and search method, search engines can be divided into two types: query-based and directory-based search engines (Embleton and Heinrich, 2008).

The query-based search engines are developed through four stages: collecting all terms and phrases of Web pages, filtering out redundant words, indexing Web pages by terms and phrases, and developing an interface. Therefore, internet users can retrieve the Web pages based on whether the query keywords appeared within the Web pages. Query-based search engines are relatively more efficient and effective, because their algorithms of extracting terms and indexing pages are straightforward and relatively simple; not too

much human effort is required for maintenance; and the compact and friendly interfaces are usually a great convenience to users (Ricardo and Berthier, 1999). However, there are still limitations with a query-based search engine. One of the most obvious limitations of such search engines is that the relevant Web pages may not be retrieved when the query keywords and the extracted terms of the Web page are not matched literally without further knowledge refining them (Conesa *et al.*, 2008). The other type of search engines, e.g., directory-based search engines, can somehow make up for this deficiency (Gerstel *et al.*, 2007). In a classical directory-based search engine, Web sites registered to the search engine or randomly crawled are classified hierarchically into categories. An internet user's exerted operation with directory-based search engines includes three steps: logging into a directory-based search engine, looking for the category which represents his or her desire and accesses to the category, and the selection of Web pages from the list in the category or using keywords to narrow down the list in the category (Comer, 2007). By using directory-based search engines, a user's search behavior is naturally bounded to the category he or she is interested in. Furthermore, no particular term-matching is required, compared to the query-based search engines. Therefore, the possibility and precision of retrieving relevant Web pages through directory-based search engines can be higher to some extent, presuming that all of the Web pages are included and classified correctly. However, most of the directory-based search engines cannot satisfy such a condition due to the fact that tremendous human efforts to classify all of the Web pages into appropriate categories based on their contents are not practical. Compared to query-based search engines, directory-based search engines require more human effort to classify and index

the pages. This limitation results in low quality and quantity of sorted Web sites to some directory-based search engines. For instance, to save human effort, most directory-based search engines only classify Web sites instead of Web pages, which cause all Web pages belonging to a Web site being partitioned into the same category, without considering whether or not there exist deviations between the Web pages in the Web site. Therefore, an efficient and effective classification method with little human effort for Web pages is highly demanded.

3.2 Outline of the Proposed Web Page Classification Method

A model of automatically classifying Web pages fusing three processes of classifying components of the Web pages is presented in this chapter. Because of the increase in noise information on Web pages, a method of Web page classification without any filter can barely classify an avalanche of new Web pages effectively. The proposed approach utilizes some features representing a Web page's meaning based on the Web page structure and a user's browsing habits. Diverse Web page features are extracted in the proposed approach, such as analyzing the text features of Web page titles and meta-search keywords, identifying the effective content from Web pages, and consulting a Web directory service. Through fusing the results from these three dedicated classifiers, Web pages are described and hierarchically classified into one or more categories. The schematic representation of the Web page classification method is illustrated in Figure

3.1. It includes feature extraction, parallel feature classification, and classification fusion.

The remaining portion of this chapter explains the process of the proposed method.

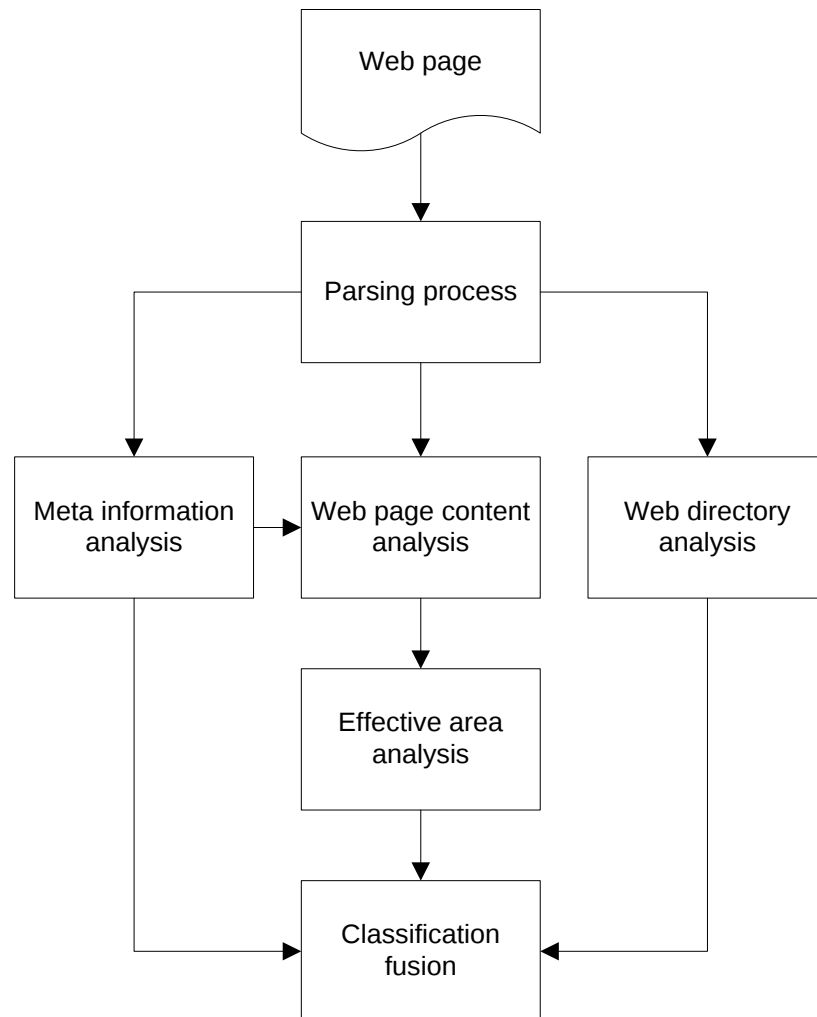


Figure 3.1 The schematic of the proposed automatic classification

3.3 Term Extraction and Weight Calculation

In text document classification, terms (important words and phrases) always play the most important and fundamental role; thus, the most features of a text document are extracted depending on certain terms. There are various methods to extract terms according to specific requirements of different classification systems. The simplest approach is to consider all words and phrases appeared in the text documents as terms. The popular idea is to determine valid terms based on their part of speech and the word frequency on training text documents, i.e., conjunction words, auxiliary words, and the words with either too high or too low frequency among the documents are not included in the term index (Ricardo and Berthier, 1999). This research adopts an alternate way to determine terms. There are three sources constituting the terms for the proposed automatic classification method of this research. The first part is based on the set of people's names in training text documents. The second part is obtained from an increasingly well-designed ontology database. The advantages of using an ontology database lie in the fact that: the ontologies themselves already eliminate meaningless words such as preposition words, auxiliary words, and pronoun words; the ontologies are extensible; and the ontologies have already been classified by the topics. In the experiments, this research makes use of the ontology from WordNet (Aggarwal *et al.*, 2004; Miller, 2009), because of its comprehensiveness and accessibility. All terms in the domain categories and term categories in the WordNet are considered as terms for this research. An example of using the WordNet to extract terms concerned with the topic of

"computer science" is shown in Figure 3.2. The third part is the noun words which appear in the training text documents with a frequency between 30% and 60%.

computer science, computing -- (the branch of engineering science that studies (with the aid of computers) computable processes and structures)

TOPIC TERM->(noun) computer#1, computing machine#1, computing device#1, data processor#1, electronic computer#1, information processing system#1

TOPIC TERM->(adj) addressable#1

TOPIC TERM->(adj) on-line#2

TOPIC TERM->(adj) off-line#2

TOPIC TERM->(adj) interoperable#1

TOPIC TERM->(adj) parallel#2

TOPIC TERM->(adj) serial#4, in series#1, nonparallel#1

TOPIC TERM->(adj) real-time#1

TOPIC TERM->(adj) stovepiped#1

TOPIC TERM->(adj) machine readable#1, computer readable#1

TOPIC TERM->(adj) open-source#1

TOPIC TERM->(noun) allocation#3, storage allocation#1

TOPIC TERM->(noun) data encryption#1

TOPIC TERM->(noun) desktop publishing#1

TOPIC TERM->(noun) access#5, memory access#1

TOPIC TERM->(noun) accumulator#3, accumulator register#1

TOPIC TERM->(noun) background#7, desktop#2, screen background#1

TOPIC TERM->(noun) backup#4, computer backup#1

TOPIC TERM->(noun) buffer#3, buffer storage#1, buffer store#1

TOPIC TERM->(noun) bulletin board system#1, bulletin board#1, electronic bulletin board#1, bbs#1

TOPIC TERM->(noun) cache#3, memory cache#1

TOPIC TERM->(noun) central processing unit#1, CPU#1, C.P.U.#1, central processor#1, processor#3, mainframe#2

TOPIC TERM->(noun) computer#1, computing machine#1, computing device#1, data processor#1, electronic computer#1, information processing system#1

TOPIC TERM->(noun) computer circuit#1

TOPIC TERM->(noun) computer network#1

TOPIC TERM->(noun) control key#1, command key#1

TOPIC TERM->(noun) counter#7

TOPIC TERM->(noun) cursor#1, pointer#3

TOPIC TERM->(noun) dedicated file server#1

TOPIC TERM->(noun) dialog box#1, panel#6

TOPIC TERM->(noun) DIP switch#1, dual inline package switch#1

TOPIC TERM->(noun) disk controller#1

Figure 3.2 The screen scan of the ontology database WordNet

In information retrieval and text mining, the *tf-idf* (term frequency–inverse document frequency) method has been proposed and applied broadly for quite a long time (Salton and Buckley, 1988). In fact, it employs a statistical measure to determine how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. Variations of the *tf-idf* weighing scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. A classic *tf-idf* method can be formulized by:

$$tf_{ij} = freq_{ij} / \max_{e \in S_{j_term}} freq_{ej} \quad (3.1)$$

$$idf_i = \log (M / g_i) \quad (3.2)$$

$$W_{ij} = tf_{ij} \times idf_i \quad (3.3)$$

where $freq_{ij}$ is the number of occurrences of term t_i in document d_j ; S_{j_term} is the set of terms in document d_j ; e indicates term t_e in document d_j ; M is the number of documents in the collection; g_i is the document frequency for term t_i in the whole document collection; tf_{ij} is the term frequency of term t_i over document d_j ; idf_i is the inverse document frequency of term t_i ; and W_{ij} is the weight of term t_i over document d_j .

According to Equation (3.2), it can be seen that the terms appearing in all categories will be weighted as 0, by which some trivial words are eliminated from the set of terms. However, since this research employs the ontology WordNet to determine the

terms, an alternative method is used to calculate the document frequency. Let N_i be the number of Web pages that contain term t_i in all categories, and n_{ij} be the number of Web pages that contain term t_i in category j , then idf_i in Equation (3.2) can be modified to n_{ij}/N_i . Therefore, the weight of term t_i over category j can be calculated by

$$W_{ij} = tf_{ij} \times (n_{ij}/N_i) \quad (3.4)$$

3.4 Web Page Classification Based on Single Component

3.4.1 Web Page Classification Based on Meta Information

In this research, the meta information of a Web page denotes the title and the meta name tagged in the Web page's source code, and the anchor text, if available. For example, the official Web site of Amazon (www.amazon.com) has the title "Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more", and the meta name "Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics,....., just about anything else". The combinational data set of these two strings forms the meta information of the Web page.

Based on the meta information of the Web page and the term weight calculation method described in Section 3.3, a classifier can be utilized to determine the potential category towards Web pages, which is given by

$$Cat_Meta(q) = \arg \max_j \sum_{i \in S_{q_meta}} W_{ij} \quad (3.5)$$

where q is the query Web page; Cat_Meta is the function to determine the potential category towards the query Web page based on meta information; S_{q_meta} indicates the text of q 's meta information; and W_{ij} is the weight of the i^{th} term in the j^{th} category.

3.4.2 Web Page Classification Based on Effective Content

Along with the rapid progress of WWW, Web pages have been increasingly implanted with a great diversity of entertainment and commercial elements. Since this work only focuses on the main text content of Web page, advertisement banners and forward links are considered as noise information which needs to be eliminated. In order to filter out such noise, some rules are followed based on the usual structure of a Web page:

- (a) Searching for lexically identical sentences in the content area with the title in meta-information. Figure 3.3 shows two situations of lexical identity.
- (b) If there is a match between the title in meta-information and a phrase in the content area, the area contains the matched phrase is considered as the effective content area.
- (c) If there is no matching between the title in meta-information and any phrase in the content area, the following sub-steps are executed.

- (c.1) The text area with a longer height is treated as the effective content area.
- (c.2) The text area with a wider width is considered as the effective content area.
- (c.3) If different columns have a similar height/width, the middle area is treated as the effective content area.
- (c.4) If the aforementioned rules are not applicable to the Web page, the text area with the maximum quantity of terms is considered as the effective content area.

After the effective content area has been determined, the same algorithm introduced in Section 3.4.1 is utilized to classify the content, which can be given by

$$Cat_Cont(q) = \arg \max_j \sum_{i \in S_{q_cont}} W_{ij} \quad (3.6)$$

where Cat_Cont is the function to determine the potential category towards the query Web page based on effective content information; S_{q_cont} indicates the text of q 's effective content information.

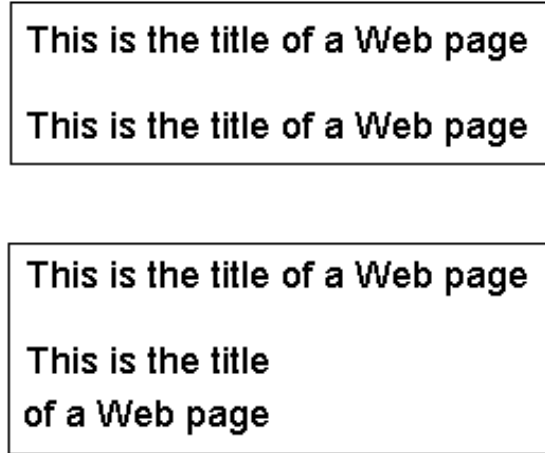


Figure 3.3 Examples of lexical identity

3.4.3 Web Page Classification Based on Directory Service

As described in Chapter 2, current Web directory services cannot provide accurate identification to every given Web page, especially Web pages in the sub-directory of a Web site. Nevertheless, manual classification for the Web site can represent the main topic. Therefore, the category information provided by Web directory services is considered as a feature for a specific Web page (Kan and Thi, 2005). Among a number of Web directory service providers, this research employs the Google directory service. The process can be formulated by:

$$FD(q) = GD(q_{link}) \quad (3.7)$$

where $FD(q)$ represents the directory feature of the query page's parent site; q_{link}

indicates the navigational link of the query page; and *GD* is the function characterizing the process of Google directory service.

Figure 3.4 shows an example of retrieving the manual classification information for a Web page using the Google directory service.



Figure 3.4 An example of retrieving manual classification for a Web page using the Google directory service

3.5 Classification Fusion

The last step of the proposed automatic classification is to fuse the results from various feature analyses that have been described in Sections 3.4.1 to 3.4.3. If the classification results of both meta-information and effective content processing are identical, the

system is determined to comply with their classification results. The fusion method is formulated to be

$$Cat(q) = Cat_Meta(q) \quad (3.8)$$

$$(if \ Cat_Meta(q) = Cat_Cont(q))$$

where Cat is the fusion function to determine the relevant category towards a query Web page.

If there exists a conflict between meta-information processing and effective content processing, a decision-making process is carried out. It assumes that Web page q is classified into category j_1, j_2 , and j_3 by the three separate classifiers respectively. Then the degrees of belief of Web page q belonging to category j_1, j_2, j_3 are calculated and compared to determine the more relevant category. The degree of belief of Web page q belonging to category j is formulated to be

$$dob(q, j) = \eta_1 \frac{\sum_{i \in S_{q_meta}} W_{ij}}{\sum_{c \in SC} \sum_{i \in S_{q_meta}} W_{ic}} + \eta_2 \frac{\sum_{i \in S_{q_cont}} W_{ij}}{\sum_{c \in SC} \sum_{i \in S_{q_cont}} W_{ic}} + \eta_3 |FD_j(q)| \quad (3.9)$$

where $dob(q, j)$ indicates the degree of belief of Web page q belonging to category j ; c indicates any category that can be recognized by the system; SC is the set of categories defined in the system; $FD(q)$ indicates the classification result by the directory service, e.g., the value of $|FD_j(q)|$ is 1 when the Web directory service considers Web page q

belonging to category j ; otherwise, the value of $|FD_j(q)|$ is 0; η_1 is the factor for meta information classification, which is equal to 1; η_2 is the factor for Web content classification, which is equal to the precision rate of determining the effective content area on training Web pages; and η_3 is the factor for the directory service, which is equal to the precision rate of the directory service on training Web pages.

Figure 3.5 illustrates the process of calculating the degree of belief that Web page q belongs to category j . The coefficient η_1 indicates the confidence degree of correctly extracting meta-information from a Web page's source code. It is set to 1 because the structure of a Web page complied with the Hypertext Transfer Protocol (HTTP) is fixed, so that the meta-information is extracted properly. Comparatively, the determination of effective content area may be inaccurate due to the flexible style of Web pages. Therefore, the coefficient η_2 applied to the Web content classification is set to the degree of precision of choosing the effective content area, which can be implemented by obtaining the degree of accuracy from the training Web page database.

When the degrees of belief for each category presented by each separate classifier are calculated, the fusion method can be formulated as:

$$Cat(q) = \arg \max_{j \in \{j_1, j_2, j_3\}} (dob(q, j)) \quad (3.10)$$

(if $Cat_Meta(q) \neq Cat_Cont(q)$)

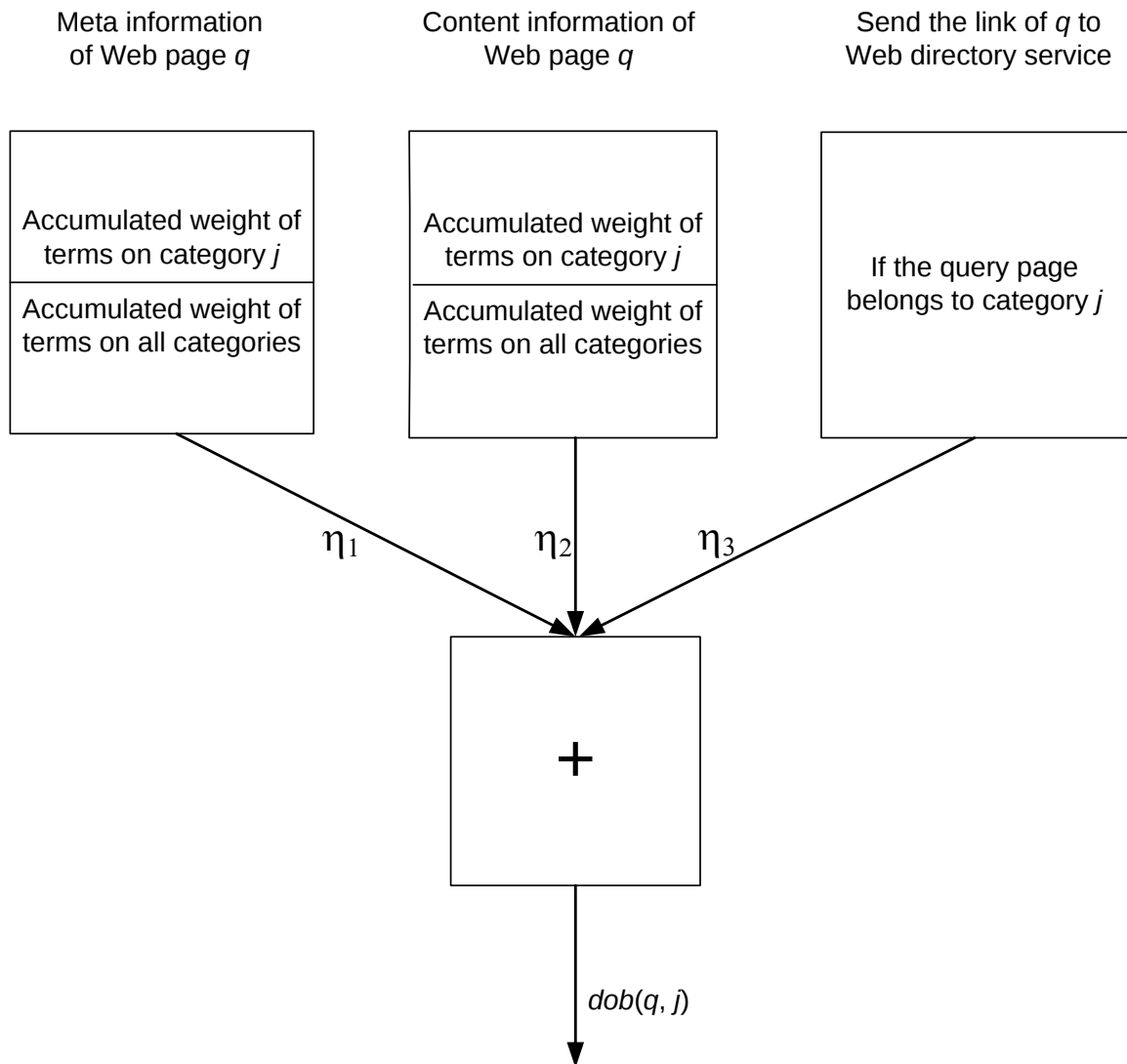


Figure 3.5 calculating the degree of belief of Web page q belonging to category j

3.6 Experimental Results

This work utilizes internet news RSS feeders as the tool for collecting text databases to train the weights of terms. The well categorization and frequent updating of news Web sites make them a ready source of training text database. Moreover, the use of RSS removes most redundant HTTP tags and noise information such as headers, footers, and commercial advertisements. There are more than 7000 news Web pages that have been collected mapping to 12 main categories, which are arts, business, computers, culture, entertainment, health, home, movie and music, science, shopping, society, and sports. In order to show the capacity of hierarchical classification of the proposed method, a classifying process of a 2nd level category is also considered. The weights of terms for four types of sports are also calculated under the main category of sports. As far as the efficiency of choosing the effective content area is concerned, manual judgment and statistics are employed in the work. Up to 1100 Web pages belonging to five types of Web sites are chosen and applied with the method of determining the effective content area. The reason of separating the five types of Web sites is for the convenience of analyzing the proposed methods. The process of obtaining the efficiency of a Web directory service is similar to acquiring the effective content area. In this case, manual work is exploited to decide whether a Web page is accordant with the categorization of a Web site by the Web directory service.

The experimental results of different steps are presented in this part. Table 3.1 shows the precision rate of determining the effective content area on five types of Web

sites. It shows that the proposed rules about choosing the effective content area of news Web pages exhibit a good performance, because a news Web page usually has the same title in both meta-information and content, which provides a useful hint for the system.

Table 3.2 gives the precision rates of directory based classification, meta information based classification, effective content area based classification, and classification fusion. It shows a good performance on identifying a shopping Web page through the Web directory service, owing to the specificity of most shopping Web sites. On the contrary, the diversity of the Web pages in a BBS Web site or a Blog Web site results in a very low precision rate of the directory based classification.

The result of Web page classification through comparing accumulated weights of terms from a Web page's meta-information implies that a well organized news or shopping Web site often has an appropriate connection between its meta-information and the Web content. It also indicates that the proposed text classification method based on the maximum accumulated weights of terms can distinguish text whether there is only a small quantity of noise information. However, with respect to blog and BBS Web pages, the lack of specific annotation for the meta-information and the diversity of their content decrease the precision rate of content classification.

Table 3.1 Precision rate of determining the effective content area

Type	Precision Rate
News Web page	96.6%
Shopping Web page	85.3%
BBS Web page	78.6%
Blog Web page	76.2%
Others	71.1%

Table 3.2 Precision rates of three component classifications and classification fusion

Type	Precision Rate			
	Directory based classification	Meta-information classification	Content information classification	Classification fusion on 1 st level
News Web page	63.2%	82.3%	85.7%	89.4%
Shopping Web page	91.4%	84.5%	76.1%	87.3%
BBS Web page	8.9%	36.1%	61.8%	67.3%
Blog Web page	6.3%	27.2%	53.2%	59.1%
Others	2.1%	6.1%	22.7%	25.2%

The result of the Web page classification based on analyzing a Web page's content shows that the performance of the text classification on news Web pages is better than that on shopping Web pages, because of the lower precision of determining the effective content area on shopping Web pages. The vagueness of the content on some BBS Web pages and Blog Web pages makes them difficult to be classified into topic categories. Moreover, the complexity of BBS web pages and Blog Web pages decreases the precision of identifying the effective content area, which makes the precision rate of content classification lower.

Table 3.2 also presents the results of classifying Web pages on the first level categories using the decision fusion integrating the three classifications. It shows that the overall precision rate on each type of Web page is higher than the result obtained from classification based on a single feature.

Table 3.3 is an example result of classifying news Web pages on the second level category. It continues proving that the proposed classification method can achieve a good performance in the genre of news Web pages.

According to the experimental results on the proposed Web page classification system, it is found that the performance of the proposed method is influenced by the genre of Web pages. It performs differently for different genres of Web site, due to the difference in the Web page's structure and nature. Obviously, the noise information in a blog Web page is possibly much more than that of a serious news Web page. Additionally, a blog may contain various topics in a Web page, which makes the Web

page categorization much more difficult. Nevertheless, the proposed classification method can achieve a good performance under the condition of Web environment with normal noise information. The proposed fusion algorithm can increase precision rate in all genres of Web site, which shows the effectiveness of the fusion algorithm.

Table 3.3 Precision rates of the classification fusion on a second level category

Sports	Precision Rate
Tennis	94.7%
Basketball	93.2%
Soccer	94.1%
hockey	92.9%

3.7 Summary

This chapter has presented a framework to automatically classify Web pages based on three different features. First, the approach uses three features to classify Web pages separately. Afterwards, classification results from different classifiers are fused and re-classified. In these experiments, the proposed method is applied to automatically classify Web pages in twelve genres. The experimental results show that the proposed method can effectively increase the precision rate, compared to the classification method using single

features of Web pages. It also indicates that the proposed classification method yields a good performance on classifying both news and shopping Web pages. However, in the case of the Web environments with a great quantity of noise information, the performance of the proposed method is somewhat decreased. Nevertheless, the proposed classification system shows great potential for application on Web services demanding unsupervised Web page categorization.

Chapter 4

Web User Modeling

User modeling is a research field belonging to human factors. It aims to construct models of human attributes and behaviors in a specific industry or commercial environment. In an online service, a user's data is collected through the Internet, by which the online service can derive the user's goals, interests, preferences, information needs, etc. A user model can be defined as the computer representation of this information. Meanwhile, the systems that construct, maintain, and update user models are called user modeling systems. The result of a user model is usually stored in a specific computer document that is called a user profile. When the content of a user's model is needed for personalized services, it can be extracted from the user profile. As a key process of personalization on the Web, user modeling has been widely used in various personalized online services, such as Web searches, e-libraries, e-museums, and e-shops.

Currently, many approaches for user modeling have been proposed for Web personalized systems. However, most of the approaches are designed for their own proprietary purposes and online systems. For the consideration of generalization, this

research aims to develop a user modeling system that can capture a user's interests and preferences for various personalized Web systems.

4.1 Introduction

An approach to model and quantify a user's interests and preference using the user's navigational data is presented in this chapter. The approach is based on the premise that frequently visiting certain types of content indicates that the user is interested in that content. Figure 4.1 shows the architecture of the proposed user modeling system. Using this system, a user's navigational data is tracked by the system for analyzing the user's interests and preferences. Web pages viewed by a user's are tracked by the system. These Web pages can be classified using the proposed classification method introduced in Chapter 3, while the source information of the Web page can be obtained directly from the http information of the page. Using Bayes theorem, a user's interest model can be obtained and updated by analyzing the content of the Web pages viewed by the user. The user's preference model can also be built and updated by analyzing the information sources of the viewed Web pages.

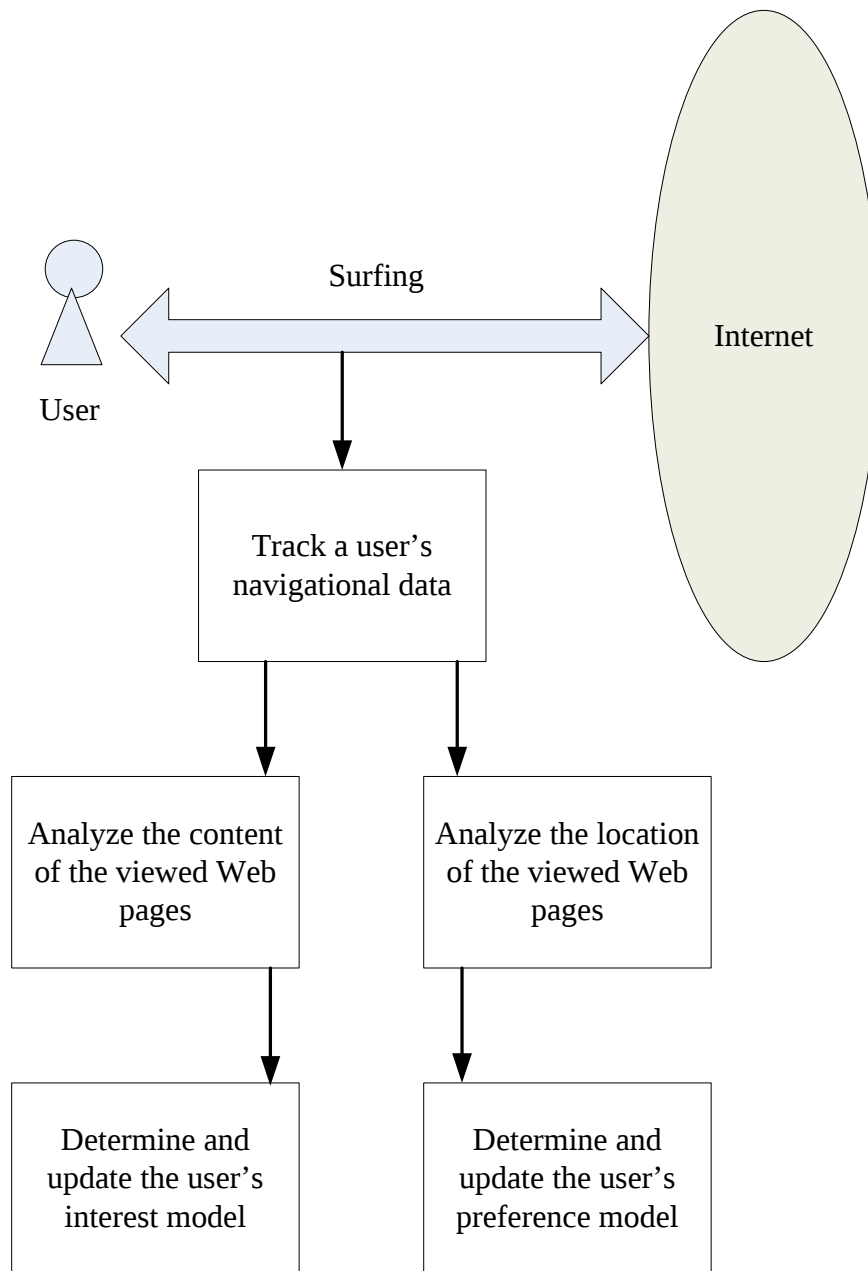


Figure 4.1 The architecture of the proposed user modeling system

4.2 Collecting a Web User's Navigational Data

In order to analyze a Web user's navigational data, a method of navigational tracking is necessary. Several types of approaches to tracking personal navigational information have been utilized in user modeling systems.

In an explicit user modeling system, a user's data is usually obtained by signup information or survey/questionnaire answers. When a user registers to use an online service, he/she may be asked to provide personal information such as gender, location, birthday, career, income, interests, and preferences. In some Web systems, a user's profile is directly constructed based on this collected personal data. The disadvantages of using signup information for user modeling are the limited scalability of signup information and infrequent updating of users' data. In order to acquire more information from users, marketing researchers usually employ another explicit method of collecting user data: online survey, e.g., design a series of questions for users to answer. In this way, specific and in-depth personal data can be collected without or with little analysis. However, many Web users are not willing to spend time on these tedious and time-consuming surveys.

Setting up a cookie in a user's computer is an easy and low-cost method. However, it can only track a user's navigational data in a certain Web site based on this method. To install a tracking tool in a user's computer is another option. For example, many Internet Explorer add-on programs can track the typing information in the address

bar. But many people are not willing to install any tracking software on their computers for the consideration of anti-virus and security reasons. Therefore, this research chooses the third option of tracking a user's navigational data: proxy server.

When a user browses the net through a proxy server, the proxy server can monitor any information between the user and the Web, which greatly satisfies the needs of the user modeling system. The disadvantages of employing a proxy server for tracking are the relatively high-cost and the limited number of simultaneous users. This research aims to find out the theoretical effectiveness of the proposed systems, so that only a dozen of test users are asked to participate. Therefore, the disadvantages of using a proxy server can be neglected in this research. Figure 4.2 shows the interface of the proxy server for monitoring user's navigational data. Users are required to type Web links into the address input-area of the user modeling system, instead of the address box of the navigation software. The proposed user modeling system can monitor a user's navigational activities and store the collected Web addresses viewed by the user into a history-data document. Figure 4.3 shows an example of collected data from a Web user.

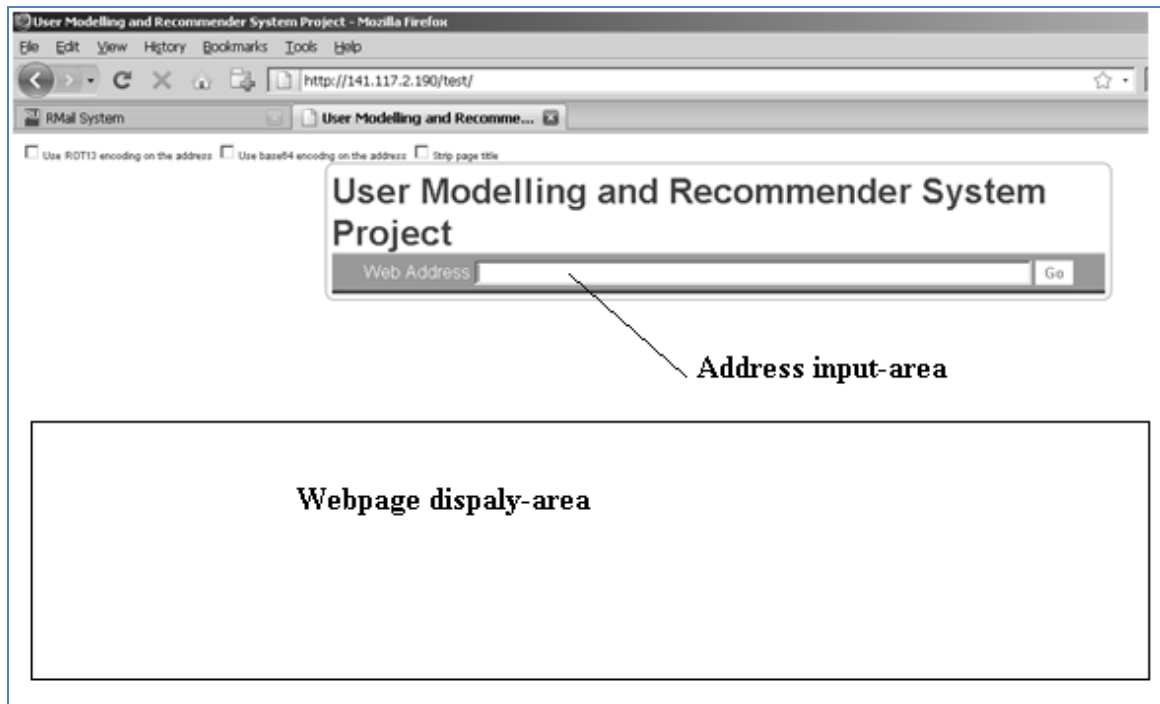


Figure 4.2 The interface of the user modeling system

```

http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/popular_bar.jpg
http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/close_btn.jpg
http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/hp_sports_sprite.gif
http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/weather_sprite.gif
http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/Red_market_status.gif
http://i.cdn.turner.com/cnn/.element/img/3.0/personalization/blue_arrow.jpg
http://www.futureshop.ca
http://www.futureshop.ca/App_Themes/Default/_samplelayout.css
http://www.futureshop.ca/App_Themes/Default/default.css
00305/splash.jpg
http://b.scorecardresearch.com/b?c1=2&c2=6034726&c3=&c4=www.futureshop.ca/splashpage.aspx&
http://www.futureshop.ca/Projects/_Content/Headlines/Assets/Weekly/20100305/splash_bg.jpg
http://www.google.ca

```

Figure 4.3 An example of collected user data

4.3 Analysis of a Web User's Navigational Data

When a user's navigational data is captured by the proposed user modeling system, the system will identify the category topics that are relative to the Web pages viewed by the user. Using the Web page classification method presented in Chapter 3, the analysis of a Web user's navigational data can be described in the following steps:

- (a) A user opens a Web page, which is monitored by the user modeling system.
- (b) The multimedia part of the Web page is filtered out, e.g., images, video clips, flash contents, etc., are removed by the system.
- (c) Determine the meta information and the effective content of the Web page.
- (d) Classify the Web page based on meta information, effective content, directory information, respectively.
- (e) Fuse the separated classification results.
- (f) Map the Web page to a category topic.

Therefore, the Web pages viewed by a user are mapped to relative category topics.

4.4 Web User Modeling

The user interest model referred to in this work represents the degree to which a Web user is interested in each category topic. By using the method of Web page classification

presented in Chapter 3, a Web user's navigational information can be grouped into different topic categories. Frequent visits of Web pages belonging to a topic category indicate that the user is interested in that topic. Mathematically, the Naïve Bayes theorem can be applied to this situation for quantitative measurement.

4.4.1 Naïve Bayes Theory

The Naive Bayes Model is a simple and well-known method for performing supervised learning of a classification problem (Kotsiantis and Pintelas, 2005; Ghosh *et al.*, 2006; Mittal *et al.*, 2007). It is based on a simple result from Bayes' theorem in statistics. The method is relatively easy to code in software. Bayes' theorem can be formulated as:

$$p(C | F_1, F_2, \dots, F_n) = \frac{p(C)p(F_1, F_2, \dots, F_n | C)}{p(F_1, F_2, \dots, F_n)} \quad (4.1)$$

where C is a dependent class variable and $F_1, F_2, \dots, F_n$ are feature variables.

Assume the system satisfies the condition of the naïve conditional independence:

$$p(F_x | C, F_y) = p(F_x | C) \quad (4.2)$$

Then Bayes' theorem can be rewritten as:

$$p(C | F_1, F_2, \dots, F_n) = \frac{p(C) \prod_{i=1}^n p(F_i | C)}{p(F_1, F_2, \dots, F_n)} \quad (4.3)$$

where $p(F_1, F_2, \dots, F_n)$ is a scaling factor, which might be a constant if the values of the feature variables are known. Equation (4.3) is called the Naive Bayes Model.

Compared to the Bayes model, the Naïve Bayes model is more manageable for computations, since it is partitioned into the class prior $p(C)$ and the independent probability distributions $p(F_i|C)$.

4.4.2 User's Interest Modeling

When a user is browsing Web page q , and q is determined to be relevant to category j through the auto-classification method described in Chapter 3, the user's likelihood of interest on category j can be updated by the following equations:

$$p_1 = p(in_j) \times p(view_q | in_j) \propto p(in_j | view_q) \quad (4.4)$$

$$\begin{aligned} p_2 &= p(not_in_j) \times p(view_q | not_in_j) \\ &\propto p(not_in_j | view_q) \end{aligned} \quad (4.5)$$

$$p(in_j) = p_1 / (p_1 + p_2) \quad (4.6)$$

where $p(in_j)$ indicates the user's degree of interest on category j ; $p(not_in_j)$ is the probability that the user is not interested in category j , and $p(not_in_j)=1-p(in_j)$; $p(in_j|view_q)$ is the probability that the user is interested in category j under the condition of opening Web page q ; $p(view_q|in_j)$ is obtained by the historical frequency of the interested categories, which is equal to $f(interested_category)/total_view$, where $f(interested_category)$ is the average click rate of the user's interested category, and $total_view$ is the overall click rate of all the navigations; and $p(view_q|not_in_j)$ is obtained by $f(not_interested_category)/total_view$, where $f(not_interested_category)$ is the average click rate of the user's non-interested category.

When a user is browsing Web page q , and q is determined not to be relative to category j , Equations (4.4-4.6) are still applicable, with only the changes that $p(view_q|in_j)$ is obtained by $f(not_interested_category)/total_view$, and $p(view_q|not_in_j)$ is obtained by $f(interested_category)/total_view$.

4.4.3 User's Preference Modeling

The Web preference model indicates the degrees of preference to Web sites (Web sources) where a user prefers to retrieve information. Similar to the assumption that has been made in the interest model, it is assumed that frequently visiting Web pages from a Web site indicates that the user is willing to retrieve information from that Web source. The same algorithm used to identify a user's interest model can also be employed to

estimate a user's preference on the Web site. A user's degree of preference on Web site s is updated by the following equations:

$$p_3 = p(in_s) \times p(view_q | in_s) \propto p(in_s | view_q) \quad (4.7)$$

$$\begin{aligned} p_4 &= p(not_in_s) \times p(view_q | not_in_s) \\ &\propto p(not_in_s | view_q) \end{aligned} \quad (4.8)$$

$$p(in_s) = p_3 / (p_3 + p_4) \quad (4.9)$$

where $p(in_s)$ indicates the user's degree of preference for retrieving information from Web site s ; $p(not_in_s)$ is the probability that the user is not interested in retrieving information from Web site s , and $p(not_in_s) = 1 - p(in_s)$; $p(in_s | view_q)$ is the probability that the user is interested in retrieving information from Web site s under the condition of opening Web page q ; $p(view_q | in_s)$ is obtained by the historical frequency of viewing information of the Web sites, which is equal to $f(preferred_website)/total_view$, where $f(preferred_website)$ is the average click rate of the user's preferred Web sites, and $total_view$ is the overall click rate of all the navigations; and $p(view_q | not_in_s)$ is obtained by $f(not_preferred_website)/total_view$, where $f(not_preferred_website)$ is the average click rate of the user's non-preferred Web site.

4.5 Experimental Results

In order to simplify the system, a one layer-based topic structure is applied to the proposed interest modeling. A user's models are dynamic and updated in real-time, because the user's navigational data is tracked and analyzed every time the user's Web activity is identified. A test user's normalized interest model at a certain time is shown in Table 4.1. For the purpose of normalization, the sum of a user's degree of interest on topic categories is set to 1. A test user's partial preference model at a certain time is shown in Table 4.2. Since an information source is defined as a Web site in this work, the preference model is a data structure with an enormous number of elements. Nevertheless, most users browse Web information within a small number of Web sites. For the sake of reducing the computational burden, a degree of preference below 0.005 is restated as 0 in the user modeling process.

The proposed method can also be used in a hierarchical structure-based modeling approach. Figure 4.4 shows the experimental results of a tester's degree of interest on the three sports subjects by using the proposed method during a period of time. This user model is constructed under the first level category topic of sports.

Table 4.1 A user's interest model

Category	Arts	Business	Computers	Culture	Entertainment	Health	Home	Movie and Music	Science	Shopping	Society	Sports
Degree	0.0212	0.0320	0.1531	0.0587	0.1092	0.0206	0.0163	0.1356	0.1239	0.1042	0.0832	0.1423

Table 4.2 A user's partial preference model

Information source	Yahoo	Google	CNN	Youtube	MSN	Ebay	Kijiji	CBC
Degree	0.1029	0.0954	0.0801	0.0616	0.0302	0.0116	0.0105	0.0093

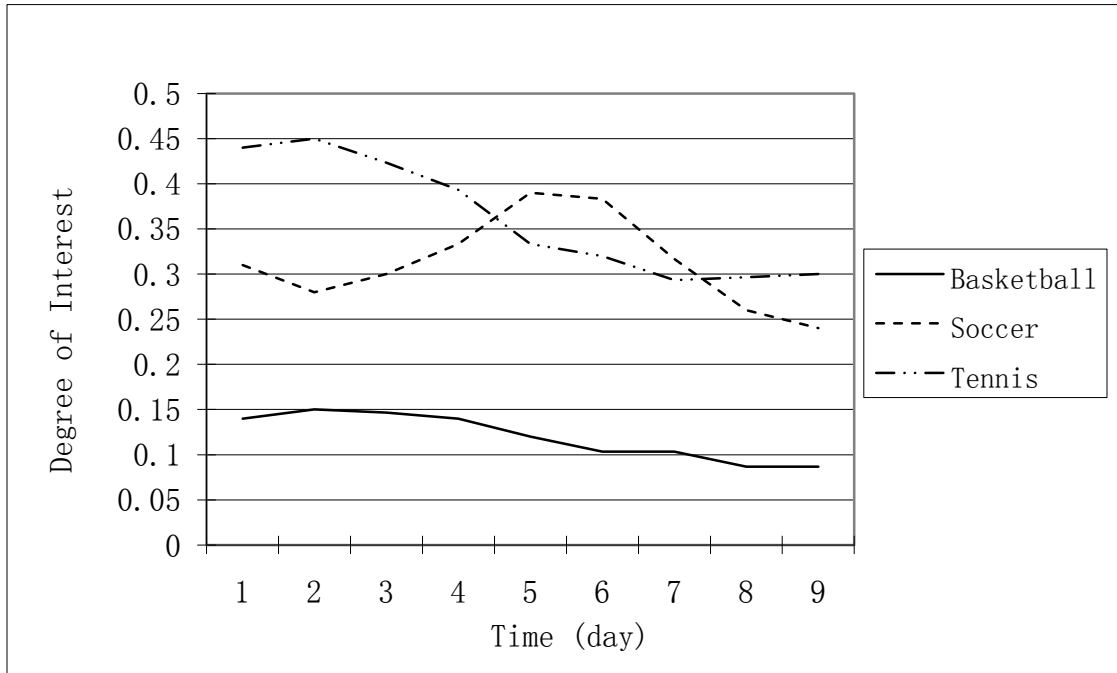


Figure 4.4 A tester's degree of interest on the three sport subjects

4.6 Summary

A framework to model a user's interests and preferences based on the user's navigational data has been presented. The approach uses the cumulative weights of terms to classify Web pages that the user is browsing. Once the content of the Web page is classified and determined, the user's interest model is updated using the Naïve Bayes Theory. A demonstration interface has been designed and implemented in order to test the proposed method. In the experiments, the proposed method is applied to quantify a user's degrees

of interest on twelve subjects and degrees of preferences on different Web sites. It also gives an experimental example of constructing a second level model of user interests. The experimental results showed that the proposed method could effectively model the user's interests. The proposed model can be integrated with personalized Web services.

Chapter 5

Personalized Web Search

Web search engines are one of the most useful and popular online services used today. There are two types of search engines: directory-based and query-based search engines. Normally, directory-based search engines have more complicated interfaces than query-based search engines. Moreover, directory-based search engines are challenged by how to automatically classify Web contents efficiently and reduce human effort. Contrarily, query-based search engines have compact interfaces and a structure that is easy to maintain, making them easy to implement and use. However, a traditional search engine faces the difficulty of distinguishing intents of users. For example, a traditional search engine can hardly determine a user's intent if a polysemy word is queried. The personalization on Web systems is a method to relieve this difficulty. A search engine can be integrated with a user modeling system, by which the likely interests and preferences of users can be utilized for re-ranking search results.

5.1 Introduction

The architecture of the proposed personalized Web search method is shown in Figure 5.1. A user's interest model, preference model, and search context model are the main components of the proposed personalized search ranking system (PSRS). The user modeling approach consisting of a user's interest and preference models is based on the premise that frequently visiting certain types of content indicates that the user is interested in that content, which is discussed in Chapter 4. The proposed approach can be divided into three steps: monitoring the user's navigational data; using the Web page classification method developed to determine a Web page's content; and employing the Naïve Bayes theory for updating the user's interest model. The same three steps can also be utilized to determine a user's Web preference model. In this work, the Web preference model assigns a value to a Web site (information source) based on the degree to which a user prefers to retrieve information from that Web site.

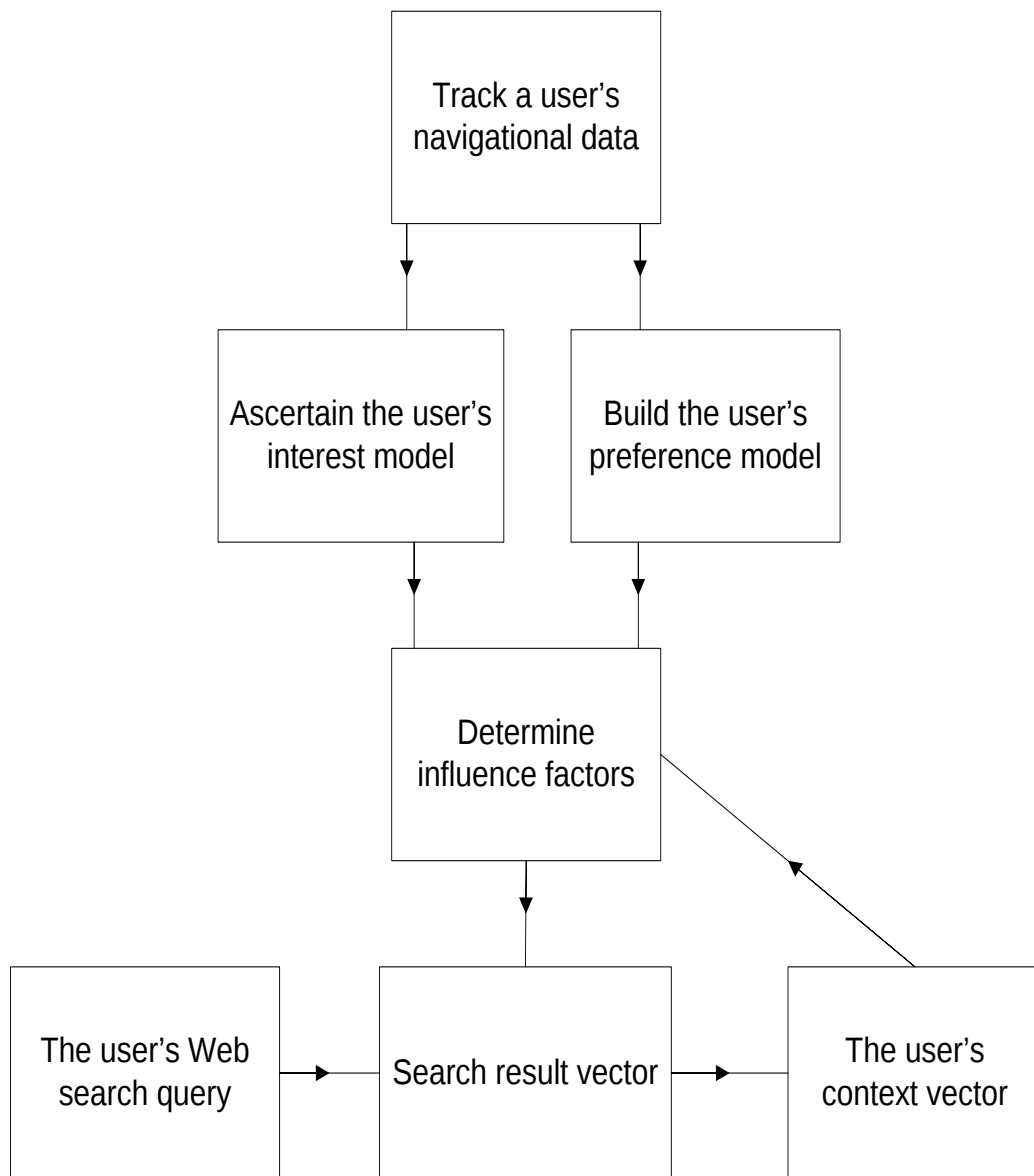


Figure 5.1 The architecture of the proposed personalized search ranking system (PSRS)

A method of auto-classifying Web pages discussed in Chapter 3 is employed as a part of the proposed user modeling process. The work employs a well-designed ontology database, WordNet (Aggarwal *et al.*, 2004; Miller, 2009), to represent valid terms for the automatic classification method. The *tf-idf* (term frequency–inverse document frequency) method is adopted to determine how important a word is to a document (category), e.g., the weight of the term to the category. First, three components of a Web page are classified separately: classification of meta information, classification of effective content area, and classification by the Web directory service. The classifications of meta information and effective content area are based on the cumulative weight of the terms, while a Web directory service can directly classify a small number of Web pages on the Internet. The classification results from the three components are fused to classify the Web pages into categories.

Tracking and recording a user's selections from the search results built up the user's search context model. In each Web search session, the Web pages viewed by a user in sequence represent the user's judgment on the search result of the query. The set of a user's selections in search sessions forms the user's search context model.

In a Web search session, the user's selection (context model) may differ from the vector of Web pages presented by the search engine. The system will check each jump-over's ranking of Web pages, in order to verify whether the user's interest or preference likely caused the jump-over. The influence factors of a user's interest model and preference model can be determined by statistically analyzing the jump-over.

In the proposed personalized Web search system, by incorporating the influence factors and the user's interest and preference models, the rank of Web pages in a search session is recalculated. In other words, if a Web page is found to be more related to the user's interest and/or preference models, and the user is usually found to be influenced by his interests and preferences in search sessions, the rank of the Web page will be raised significantly. Details on the search context model and the re-ranking of search results are discussed in this chapter.

5.2 Conventional Web Search Engines

It is a well known fact that the Web is a huge repertory, consisting of various types of Web sites and online documents. The basic data structure of the Web is the connection between a URL address and an online target, such as a Web site, a Web page, or a digital document. Therefore, the straightforward strategy of retrieving online information is direct navigation that requires a user to input a link address into the browser. This method is often used while a user logs into the frequently visited Web sites. However, this strategy is not efficient for Web sites with complex structure, and not applicable for the Web sites that a user cannot correctly input the URL addresses. Another strategy of retrieving Web information is to click a link to open a Web page. In a Web site where the contents are well-organized and the links to the contents are provided to users, this method is effective and adopted by users. Figure 5.2 shows an example of retrieving Web

information by links in the CBC Web site. However, without the organization of links, it is difficult for users to access Web sites of interest by clicking links in between them. Therefore, an information retrieval tool, which is known as a search engine, is used for users to efficiently obtain the Web information of interest.

A directory-based search engine has three processing steps: collecting information of Web sites, classifying Web sites, and presenting links to Web sites hierarchically. Figure 5.3 shows the interface of a directory-based search engine. In this type of search engine, Web links are categorized into different topics. Therefore, a user can reach the links of interest by selecting the relevant topics. Along with the rapidly increasing number of Web sites, directory-based search engines have gradually shown their limitations. Firstly, to classify Web sites either increases machine computational burden or requires a lot of manual works. Secondly, it is difficult to present the relevant Web sites to users, because the number of Web sites belonging to a category topic is very large. Thirdly, a directory-based search engine only classifies Web sites and provides links of Web sites, which cannot guide a user directly to the Web pages of interest.

Query-based search engines are more popular than directory-based search engines because they are intuitive, simple to use, and easy to maintain. A query-based search engine can be considered as a tool for retrieving information from a database. The input of a search activity is a set of query terms, while the output is a set of Web links whose targeting Web pages contain the query terms. A query-based search engine usually has a

compact interface, which is illustrated in Figure 5.4. In the search engine, the search results are presented right after the query terms are submitted.

SMART SHIFT Weekly Think Tank Making the world safer with smart technology	TOXIC WASTE The treasure in tailings Harvesting wealth from oilsands waste	PARTICLE PHYSICS Quark soup Large Hadron Collider detects 'Big Bang' matter
More Tech News from Canadian Press • Poynt inks deal with India's Time Internet; service to start in early 2011 • CounterPath net loss cut by more than half to US\$800,000 as revenues surge • Poynt inks deal with India's Time Internet; service to start in early 2011 • Ottawa approves Apple's iBookstore for Canada after Investment Canada Act review • Dell to buy Compellent for \$884 million as it looks to boost position in data storage industry • Netflix signs licensing agreements for numerous TV, film programs • It's all about the PlayBook even though RIM to report strong quarterly results • Ex-Goldman Sachs programmer convicted of stealing code for high-speed trading • Parents blue after kids rack up big iTunes bills buying expensive game add-ons • Consumers live in 'on demand' universe with telecoms to provide even more content	More Science News from Canadian Press • Oilsands don't deserve all criticism, but oversight must improve: scientists • Atomic weight of 11 elements on what was a constant periodic table are changing • International spacecraft spies possible ice volcano on giant Saturn moon Titan • Florida researchers for first time capture X-ray images coming from lightning • NASA: Long-running unmanned spacecraft nearing edge of solar system in newest milestone • NASA: Long-running unmanned spacecraft nearing edge of solar system in newest milestone • New theory for origins of Saturn's rings: Icy remnants from a moon before it plunged to death • Nature's coming attraction: Geminid meteor shower, best of year with 100-plus meteors per hour • 2nd study says Haitian cholera from South Asia, carried in by outsiders, but no group named • SpaceX spacecraft carried secret payload - a wheel of cheese - on historic orbital flight	

Figure 5.2 An example of retrieving Web information by links in the CBC Web site

Google directory [Directory Help](#)
☒ Search only in News ☐ Search the Web

News

Categories

Alternative (158)	Directories (137)	Museums and Archives (60)
Analysis and Opinion (485)	Extended Coverage (33)	Newspapers (4216)
Breaking News (108)	Headline Links (77)	Personalized News (17)
By Region (10)	Internet Broadcasts (56)	Satire (98)
By Subject (219)	Journalism (1593)	Television (112)
Chats and Forums (24)	Journals (56)	Weather (255)
Colleges and Universities (1008)	Magazines and E-zines (436)	Weblogs (169)
Current Events (275)	Media (2371)	

Related Categories:

[Computers > Internet > On the Web > Web Portals](#) (46)
[Society > Folklore > Literature > Urban Legends](#) (64)

Web Pages

Viewing in Google PageRank ord

- [CNN - Cable News Network](#) - <http://www.cnn.com>
Includes US and international stories and analysis, weather, video clips, and program schedule.
- [Aljazeera](#) - <http://english.aljazeera.net/>
English version of the Arabic-language news network. Breaking news and features plus background materia
- [The Guardian](#) - <http://www.guardian.co.uk/>
Home of the Guardian, Observer and Guardian Weekly newspapers plus special-interest web sites. Each in and free archives.
- [The New York Times](#) - <http://www.nytimes.com/>
Online edition of the newspaper's news and commentary. [Registration required]

Figure 5.3 The interface of the Google directory-based search engine

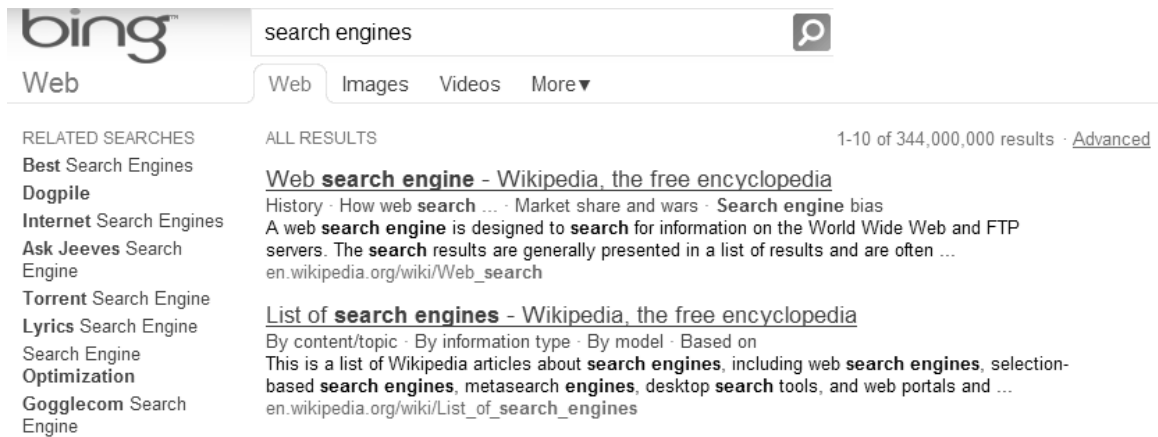


Figure 5.4 The interface of the Bing query-based search engine

5.3 Personalized Web Search Engine

This research attempts to identify the relationship between a user's Web search behaviors and his or her interest and preference models, so that a personalized Web search method can be developed based on this relationship. In other words, based on analyzing a user's search behaviors, the system will determine how the interests and preferences of the user influence the search results. The system will then provide a personalized search solution for the user based on that influence.

5.3.1 User's Search Context Model

In a query-based search engine, a document vector of Web pages is linked to a set of

query keywords. When a user sends a query to the Web search engine, a set of Web pages will be fed back with the highest ranked Web page at the top. This query operation can be formulated as:

$$Query(qk_r)=m_r \quad (5.1)$$

where *Query* represents the function of query operation; qk_r is the vector of query keywords in search session r ; and m_r is the resulting vector of Web pages in search session r , which is also represented as $(m_r(1), m_r(2), \dots, m_r(lw))$, where lw is the number of relevant Web pages, $m_r(1)$ is the Web page with the highest rank score calculated by the Web search engine, and $m_r(lw)$ has the lowest rank score among the given Web pages.

In a conventional search engine system, the ranking of the documents for a query is the same to all the users regardless of whether the users have their own judgment on the resulting Web pages. An individual's own ranking preference is represented by his or her navigational context on the resulting Web pages, which is inferred from the assumption that an individual usually first views the best matched Web page from his or her perspective, then the Web page of the second best match, etc. Therefore, the context vector of the Web pages viewed by a user during the search session r can be described as v_r , e.g., $(v_r(1), v_r(2), \dots, v_r(lo))$, with $v_r(1)$ being the initial Web page opened by the user, $v_r(2)$ the second Web page viewed, and $v_r(lo)$ the last Web page viewed by the user in search session r . In this work, v_r is considered as the r^{th} component of the user's search context model, which represents the user's personal ranking in search session r .

5.3.2 Proposed Personalized Search Ranking

A user's interest and preference models are built and updated based on his or her long-term navigational data. For each user's very first search query, the results presented to the user are the exact duplicates of one generated by the conventional search engine. Subsequently, the vector of Web pages viewed by a user is compared with the vector of Web pages resulting from the Web search engine. The user's search behavior data is utilized to determine how the user's interest and preference models impact his or her personal judgment on the given Web pages. The re-ranked results will then be presented to the user and the user's click-through is tracked. Any ranking change from the given search results to the user's context model indicates the dissimilarity between the search engine's ranking algorithm and the user's judgment. For example, if the vector of Web pages $m_r (= (m_r(1), m_r(2), m_r(3), \dots, m_r(lw)))$ is the search result of a user's r^{th} search session, and the user's search context vector in this session is $(m_r(3), m_r(1))$, e.g., the user viewed Web pages $m_r(3)$ and $m_r(1)$ successively, then there are two instances of ranking jump-over observed, which are $m_r(3)$ over $m_r(2)$, and $m_r(3)$ over $m_r(1)$. The pseudo code of determining the total number of instances of ranking jump-over from the search result to the user's context model is given in the next page:

```

read user's context vector  $v=(v(1), v(2), \dots v(lo))$ ,
read search result  $m$ ,
for  $i=1, lo, i++$ 
    find  $n$ , that  $m(n)=v(i)$ ,
    for  $j=1, n-1, j++$ 
        if  $m(j)$  is prior than  $v(i)$  in the user's context model then
            no ranking jump-over,
        else
            ranking jump-over is determined,
             $Sum_{jumpover} = Sum_{jumpover} + 1$ ,
        end if
    end for
end for

```

The next step is to utilize the ranking jump-over to determine the influence of factors in a user's interest model and preference model. Using the same example discussed above, the Web pages involved in ranking jump-over can be mapped to categories: $Cat(m_r(1))$, $Cat(m_r(2))$, and $Cat(m_r(3))$, by using the classification method introduced in Chapter 3. If the user's interest model shows that the user's degree of interest in $Cat(m_r(3))$ is higher than the degree of interest in $Cat(m_r(1))$, then the ranking jump-over of $m_r(3)$ versus $m_r(1)$ is considered as the case that the user's interest model

influences his or her search behavior. After taking into account the ranking jump-overs in all search sessions, the influence factor of a user's interest model can be determined statistically:

$$\eta_4 = (\Sigma \text{jump-overs influenced by a user's interest}) / (\Sigma \text{jump-overs}) \quad (5.3)$$

The same method can also be applied to determine the influence factor of a user's preference model:

$$\eta_5 = (\Sigma \text{jump-overs influenced by a user's preference}) / (\Sigma \text{jump-overs}) \quad (5.4)$$

Since the conventional search engines only provide ranking results without the appearance rates of Web pages, an inverse distance measure is used to estimate the rates in a search session (Teevan *et al.*, 2010), which is formulated as:

$$ER(m(k)) = 1 / (\alpha + \beta \log(k)) \quad (5.5)$$

where $ER(m(k))$ is the estimated rate of the k^{th} Web page in a search result; α is the initial coefficient which is set to 2 in this work; and β is the attenuation coefficient which is set to 1. The choice of the coefficients is taken into account for the relatively smooth decreasing and dominance of the top ten ranks, which is shown in Figure 5.5.

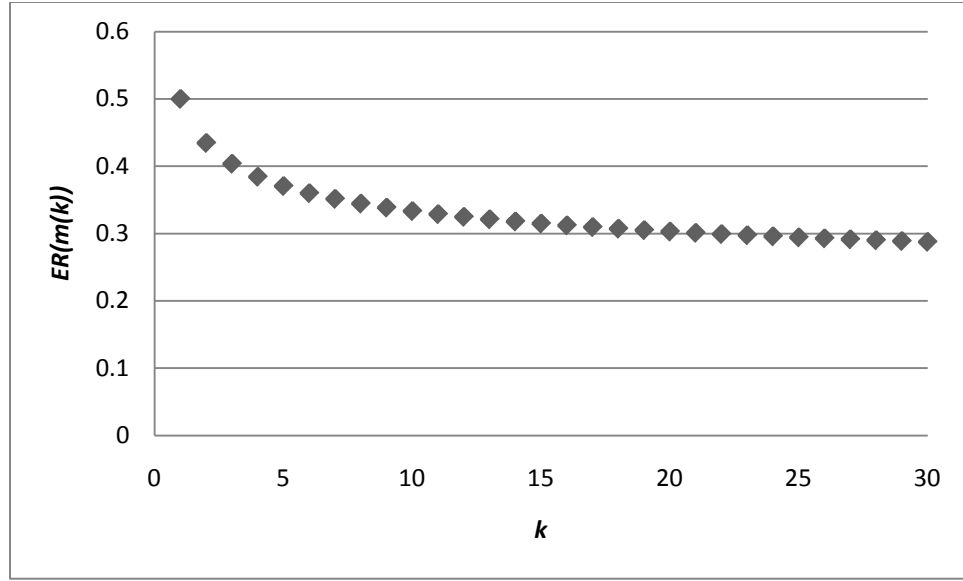


Figure 5.5 The rate estimation of the Web pages in a search result

Once a user's interest model, preference model, and influence factors have been determined, a personalized search rank algorithm can be applied, which is

$$RR(m(k)) = ER(m(k))(1 + \eta_4 UserInterest(Cat(m(k))))(1 + \eta_5 UserPreference(IS(m(k)))) \quad (5.6)$$

where RR represents the re-rating function; $UserInterest$ calls equation (4.6) to retrieve the user's degree of interest in category $Cat(m(k))$; $IS(m(k))$ is the function of identifying the information source (Web site) of Web page $m(k)$; and $UserPreference$ calls equation (4.9) to retrieve the user's degree of preference on Web site $IS(m(k))$.

Ultimately, the web pages are arranged by their re-rating value $RR(m(k))$, e.g., the Web page with the maximum re-rating value is listed at the top. Therefore, the personalization of a search engine is implemented by re-ranking the Web pages based on user models.

5.4 Experimental Results

In this work, a user's click-through in a search session is considered as the user's search context model. In order to clarify the construction of a user's search context model, an example of a user's click-through is shown in Table 5.1. The central column shows the top 10 Web page results presented by the Google search engine with the query keyword "personalization". The right column shows the user's click-through during this search session, e.g., the user viewed the 7th Web page and 1st Web page, successively. Therefore, the user's search behavior in this session can be modeled as {"7 Personalization Gallery Microsoft", "1 Personalization Wikipedia"}. When a user clicks the 7th Web page first, ranking jump-overs occur. It is necessary to identify whether the jump-overs are likely influenced by the user's interest and/or preference. Using the proposed Web page classification method to analyze the content of the 7th and other top 6 Web pages, it is found that the 7th Web page belongs to the "Computers" category, while 4 of the top 6 Web pages belong to other categories. Since the user's interest model shows that the user is more interested in the "Computers" category than any other topic, the 4 cases of jump-

over are supportive to increase the influence factor of interest model, e.g., η_4 is updated to $(\text{sum of jump-overs influenced by a user's interest} + 4) / (\text{sum of jump-overs} + 6)$. The same procedure for updating the influence factor of preference model η_5 is conducted simultaneously.

The purpose of developing a personalized search system is to change the search ranking results based on a user's models, so that the user can retrieve information of interest more effectively. Table 5.2 shows an example of the proposed personalized re-ranking of search results based on the user's models. From the results of re-ranking, it is shown that Web pages that are relevant to a user's topics of interest like "Computers", and a user's preferred Web site like "Google", are rearranged to the top of the list.

Table 5.1 An example of constructing a user's search context model

	Search engine = Google Query keyword = "personalization"	Click order
1	Personalization - Wikipedia, the free encyclopedia En.wikipedia.org/wiki/Personalization	2
2	Personalized Gifts from Personalization Mall www.personalizationmall.com	
3	What is Personalization? - Definition from Whatis.com searchcrm.techtarget.com/definition/personalization	
4	Top 5 Web Trends of 2009: Personalization www.readwriteweb.com/.../top_5_web_trends_of_2009_personalization.php	
5	personalization - Official Google Blog googleblog.blogspot.com/search/label/personalization	
6	Personalization is Over-Rated (Jakob Nielsen's Alertbox) www.useit.com/alertbox/981004.html	
7	Personalization Gallery - Windows 7 themes, wallpapers, and ... windows.microsoft.com/en-CA/windows/.../personalize	1
8	Help Center Facebook www.facebook.com/help.php?page=1068	
9	Facebook's Instant Personalization Is the Real Privacy Hairball gigaom.com/.../facebooks-instant-personalization-is-the-real-privacy-hairball/	
10	HOW TO: Disable Facebook's "Instant Personalization" [PRIVACY] mashable.com/.../disable-facebook-instant-personalization/	

Table 5.2 An example of personalized re-ranking

	Search engine = Google Query keyword = "user model"	Search results of personalized re-ranking	
1	Standard user model - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Standard_user_model	Standard user model - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Standard_user_model	1(1)
2	Figuring Out What They Expected - Joel on Software www.joelonsoftware.com/uibook/.../fog000000058.html	Figuring Out What They Expected - Joel on Software www.joelonsoftware.com/uibook/.../fog000000058.html	2(2)
3	Images for user model	Application User Model IDs (AppUserModelIDs) (Windows) msdn.microsoft.com/en-us/library/dd378459(VS.85).aspx	3(8)
4	Django User authentication in Django Django documentation docs.djangoproject.com/en/dev/topics/auth/	User Model - Radiant CMS Google Groups groups.google.com/group/radiantcms/browse.../12b2bd18b7fa0fe7	4(7)
5	Flickr User Model, v0.3 Flickr - Photo Sharing! www.flickr.com/photos/bryce/58299511/	Extending User object in Django: User model inheritance or use stackoverflow.com/.../extending-user-object-in-django-user-model-inheritance-or-use-userprofile	5(6)
6	Extending User object in Django: User model inheritance or use stackoverflow.com/.../extending-user-object-in-django-user-model-inheritance-or-use-userprofile	Django User authentication in Django Django documentation docs.djangoproject.com/en/dev/topics/auth/	6(4)
7	User Model - Radiant CMS Google Groups groups.google.com/group/radiantcms/browse.../12b2bd18b7fa0fe7	Flickr User Model, v0.3 Flickr - Photo Sharing! www.flickr.com/photos/bryce/58299511/	7(5)
8	Application User Model IDs (AppUserModelIDs) (Windows) msdn.microsoft.com/en-us/library/dd378459(VS.85).aspx	Images for user model	8(3)
9	Extending the Django User model < WebIT.ca webit.ca/2010/05/extending-django-user/	Extending the Django User model < WebIT.ca webit.ca/2010/05/extending-django-user/	9(9)
10	The User-mode Linux Kernel Home Page user-mode-linux.sourceforge.net/ - Cached - Similar	The User-mode Linux Kernel Home Page user-mode-linux.sourceforge.net/ - Cached - Similar	10(10)

5.5 Evaluation

In order to evaluate the proposed personalized Web search method, 12 participants were asked to surf the Internet and conduct Web searches for their daily requirements under the system's supervision. All participants have at least a Bachelor's degree. The participants' interest, preference, and context models were constructed during their Web navigation. In order to alleviate the computational burden, the top 30 Web pages are re-ranked by the personalized search system. Even though the explicit measurements of relevance can clarify the precision rate and recall rate judged by the users, an implicit measurement is applied to evaluate the performance of the proposed personalized Web search method. Due to the fact that a user's experience of a practical search is the greatest concern, the implicit measurement approach is chosen. Table 5.3 shows the comparison of the users' average first-click on the personalized search results and the conventional search results tracked in 100 practical search sessions. The average first-click of the participants on the personalized re-ranking Web pages is 3.09, while it is 3.80 when the clicked Web pages are mapped to the conventional search results. Therefore, the average improvement of the first-click provided by the personalized re-ranking method is 0.71.

In order to check the statistical significance of the results shown in Table 5.3, a paired *t*-test (Box *et al.*, 1978) is conducted based on the 12 participants' average rates of first-click. Denote participant *i*'s average rates of first-click on the personalized and conventional search results by $x_{1,i}$ and $x_{2,i}$, respectively. There are three steps for carrying

Table 5.3 Users' average first-click on the personalized search results and the conventional search results

User ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
Conventional search	3.41	4.13	2.98	3.31	5.05	3.54	4.91	2.89	4.21	4.07	3.32	3.79	3.80
Personalized search	2.56	3.62	2.60	2.71	3.88	2.85	4.21	2.70	2.96	3.25	2.89	2.94	3.09
Improvement	0.85	0.51	0.38	0.60	1.17	0.69	0.70	0.19	1.25	0.82	0.43	0.85	0.71

out the paired t -test:

- (a) Given two data sets $X_1=\{x_{1,1}, x_{1,2}, \dots, x_{1,tn}\}$ and $X_2=\{x_{2,1}, x_{2,2}, \dots, x_{2,tn}\}$, where tn is the number of participants, construct a new data set $X_D=X_1-X_2$. For a paired t -test, the two data sets must have the same number of elements.
- (b) Determine t -value of the two data sets by computing

$$t = \frac{\overline{X_D}}{S_D/\sqrt{tn}} \quad (5.7)$$

where t is the t -value of the two data sets; $\overline{X_D}$ is the mean value of data set X_D ; and S_D is the standard deviation of X_D .

- (b) Determine p -value from the t -value using the t -table. In a statistical analysis, if the p -value is below a selected threshold for statistical significance (often the 0.01, 0.05, or 0.10 level), the result is considered to be statistically significant.

By conducting the paired t -test, the observed difference of the mean value between the personalized and conventional search models in Table 5.3 is found to be statistically significant having $t=7.83$ and $p<0.01$.

In order to evaluate the performance of the proposed system based on a user's overall click-through, the method of Discounted Cumulative Gain (DCG) is employed in this work (Jarvelin and Kekalainen, 2000). Considering that the higher ranked Web pages have higher scores in the evaluation, the i^{th} Web page in a search result set can be scored by:

$$Score(i) = 1/\log_2(1+i) \quad (5.8)$$

During a search session, a user will click the Web pages if they think it is relevant to the query. Therefore, the ranking-efficiency can be obtained by weighing the user's click-through in a search session. Equation (5.9) presents a measurement of ranking-

efficiency for a user's search session:

$$E = (\sum (Score(i)Click(i))) / \sum Click(i) \quad (5.9)$$

where $Score(i)$ is the score of Web page i in a user's search session as defined in Equation (5.7); and $Click(i)$ indicates if the user opens the i^{th} Web page. $Click(i)$ is 1 when the user opens Web page i , otherwise $Click(i)$ is 0.

Table 5.4 shows the comparison of the average ranking-efficiency of both the personalized and conventional searches through the investigation of the participants' 100 rounds of search sessions. The average ranking-efficiency of the conventional search engine based on the participants' search behaviors is 0.44. The personalized search method improves the ranking-efficiency to 0.52. Therefore the ranking-efficiency of the personalized search is 18% higher than the conventional search. Moreover, there is no obvious regression of the ranking-efficiency observed on any participant's search data, which implies the generality of the proposed re-ranking method.

Table 5.4 Average ranking-efficiency of the personalized search and the conventional search

User ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
Conventional search	0.51	0.42	0.52	0.44	0.36	0.43	0.31	0.53	0.39	0.43	0.45	0.46	0.44
Personalized search	0.58	0.51	0.63	0.49	0.49	0.51	0.41	0.57	0.46	0.49	0.51	0.50	0.52
Improvement	0.07	0.09	0.11	0.05	0.13	0.08	0.10	0.04	0.07	0.06	0.06	0.04	0.08

5.6 Summary

A novel method for a personalized Web search system has been proposed in this chapter. A Web page auto-classification method is utilized to analyze a user's navigational data, which helps to build and update the user's interest model. The user's preference model is constructed by analyzing the information sources (Web sites) of the user's navigational data. The user's click-through during a search session is interpreted as the user's search context model which is used to determine how the user's interest and preference models influence his or her search behavior. Finally, the personalized re-ranking algorithm is implemented by recalculating the score of a Web page based on the user's interest and preference models. The experimental results show an obvious improvement of the search ranking provided by the proposed personalized search method. The proposed system utilizes both content-based and behavior-based modeling approaches for the Web search personalization.

Chapter 6

News Recommender System

6.1 Introduction

The rapid development of electronics and the World Wide Web (WWW) have significantly impacted our society and daily lives (Health, 2008). Currently, the Web is widely used for individuals and organizations in various fields, such as e-banking, education, e-commerce, research, news distribution, entertainment, and communication (Adam *et al.*, 2002; Albers and Clement, 2007; Chen *et al.*, 2007; Li *et al.*, 2008; Van Den Heuvel and Papazoglou, 2010; Von Borstel and Gordillo, 2010). Over the past few of years, the amount of online information brought and distributed by online services has been rapidly expanding. Moreover, due to the increasing popularity of the WWW, the acceleration of this information explosion on the Web keeps growing. Accordingly, Web users are requesting more and more retrieval tasks with their increasing familiarity with utilizing the Web. However, the Web is well known as a poorly organized and indexed information repository due to the extreme openness and diversity of the Web's structure and information.

Currently, a combination of using Web search engines and manual Web

navigation is one of the most commonly used methods for searching and retrieving contents on the Web (Trestian *et al.*, 2010). Although search engines have become the most popular tool to retrieve information from the Web, they do not take the initiative in providing information to a user without the process of a query. Most people frequently retrieve information from Web sites they are familiar with, whose contents are updated from time to time; news sites, product information pages, and personal blog sites for instance. It is often tedious and time consuming to repeatedly check each Web site for any new content. Therefore, a Web feed format called Really Simple Syndication (RSS) was designed and used to automatically distribute frequently updated Web files to users (Pera and Ng, 2008). RSS is one of the solutions used to notify subscribers of any new updated content on the Web. By using the RSS based Web content notification and downloading systems, a user can manage the updated files from multiple Web sites in a much simpler way. The resulting files are grouped based on their Web sources and presented in an organized email structure. However, RSS cannot take charge of most individuals' navigation on the Web due to its inherent limitations. Obviously, one limitation of RSS is that not every Web site integrates RSS into its online service. In addition, updated contents are simply organized by their information sources (Web sites) to RSS subscribers, which is non-flexible and unintelligent. Therefore, significant attention has been paid to recommender systems as the other approach to active information distribution (Adomavicius & Tuzhilin, 2005a). This approach is focused on delivering information of interest to Web users with little manual effort.

Accordingly, the main objective of this chapter is to develop a Web content recommender system. A user's long term interest and preference models are constructed based on the user's navigational history and integrated with the recommender system. The similarity between Web content and the user's models is used to determine whether the content will be provided to the user. A user collaboration method is designed to improve the effectiveness of the proposed recommender system.

6.2 Architecture of the Proposed Recommender System

The architecture of the proposed hybrid recommender system for news recommendation on the Web is shown in Figure 6.1. In the system, a Web user is distinguished by identifying his or her interest and preference models. A user's navigational data is monitored and analyzed to conduct user modeling. The automatic classification method presented in Chapter 3 is utilized to categorize the Web contents browsed by a user. In the proposed Web page classification method, the ontology base WordNet determines the terms (Aggarwal *et al.*, 2004; Miller 2009), and the weights of terms are calculated by the *tf-idf* (term frequency-inverse document frequency) method. Three components of a Web page, meta information, effective content area, and the Web address, are extracted and classified separately. The cumulative weights of the terms are utilized for the classification of meta information and effective content areas, while a small number of Web pages on the Internet can be directly classified through a Web directory service. A

fusion method is used to combine the classification results for the three components.

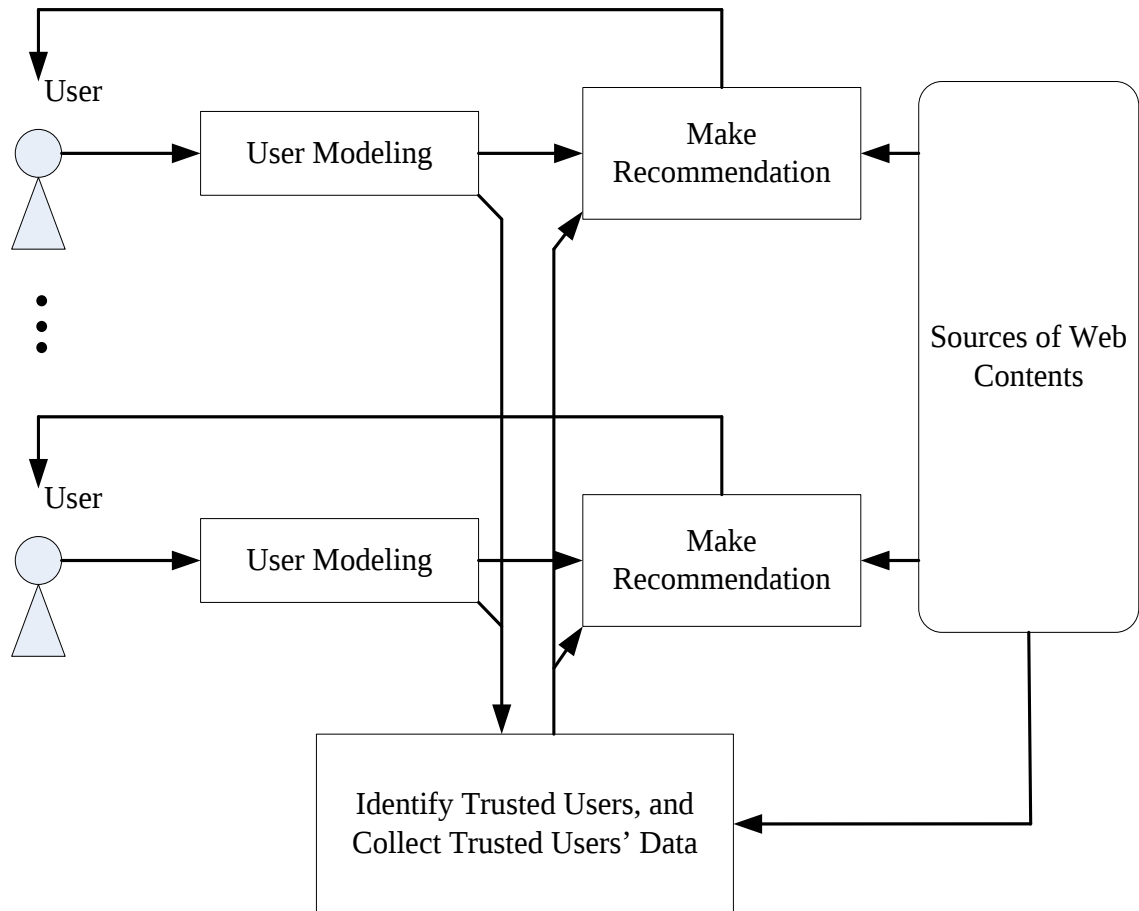


Figure 6.1 The architecture of the proposed recommender system

It is assumed that a user is interested in a certain type of content if the user frequently visits that type of content. The user modeling method presented in Chapter 4 is

utilized in the proposed recommender system. It consists of two steps: determining the content of a Web page using the Web page classification method; and utilizing the Naïve Bayes model for updating the user's interest and preference models. In this work, a user's preference model scores a Web site based on the degree to which the user prefers to retrieve information from that Web site.

The recommendation rating of the proposed system can be divided into two steps. First, a content-based algorithm is utilized to determine the probability of recommending Web content to a user, considering the factors of the user's interest and preference models, the Web content, and the time limitation. Second, the method of collaborative filtering is used to modify the probability of recommending Web content. The system will distribute some test Web content, which has been well classified and identified by users. The users who send back positive responses are considered as the trusted users. Additionally, the Web content browsed by more trusted users will obtain higher scores in the recommending process. A preliminary version of the recommender system is reported by Wen *et al.* (2009)

6.3 The Proposed Recommender System

The personalization of online services usually requires the identification of users based on their interests, preferences, navigational data, purchased products, etc. The content-based recommendation method is available if user modeling is feasible. Many techniques

of modeling a Web user's interests and preferences have been presented (Boutemedjet and Ziou, 2008; Frias-Martinez *et al.*, 2006; Teevan *et al.*, 2010). There are two types of approaches to user modeling: explicit and implicit approaches. In an explicit approach, a user is asked to present his or her preferences directly, while in an implicit approach, the system usually builds up a user's models by analyzing the user's tracked information. A news recommender system can be integrated with either of the two user modeling methods. For example, a news subscription system is a typical recommender system that uses the explicit user modeling method. This work employs the implicit approach to user modeling that is presented in Chapter 4 for news recommendation.

6.3.1 User Modeling for the Recommender System

In the proposed recommender system, the construction of a user's interest and preference models is based on analyzing the user's navigational data. In order to identify a Web page browsed by a user, a method of auto-classifying Web pages based on various features (Wen *et al.*, 2008b) is applied, which has been explained in Chapter 3.

The schematic representation of the classification solution used in the system is illustrated in Figure 6.2. The classification method includes three steps: content separation, parallel feature classification, and a combination of the results. Initially, a Web page's hyperlink information, meta information, and content information are analyzed for classification, respectively. Then, a fusion algorithm generates the final

classification result. In this approach, in order to classify text information such as a Web page's meta and content information, weights of terms in categories are used for computing the cumulative weights of terms in the target text.

In this work, the user interest model indicates the degree of a user's interest in each topic category, while the preference model represents the user's degree of preference for a Web source (Web site) to acquire information. Based on the premise that frequently visiting Web pages belonging to a certain category indicates that the user is interested in that topic, a user's interest model can be constructed by analyzing the user's navigational data. After applying the method of Web page classification, the Web pages browsed by a user can be categorized into different topic categories. Therefore, a quantitative measurement for creating and updating a user's interest model is available by utilizing the Bayes theorem. The process of modeling a user's interests and preferences for the personalized recommender is illustrated in Figure 6.3. The algorithm for updating a user's models is given and explained by Equations (4.4) to (4.9).

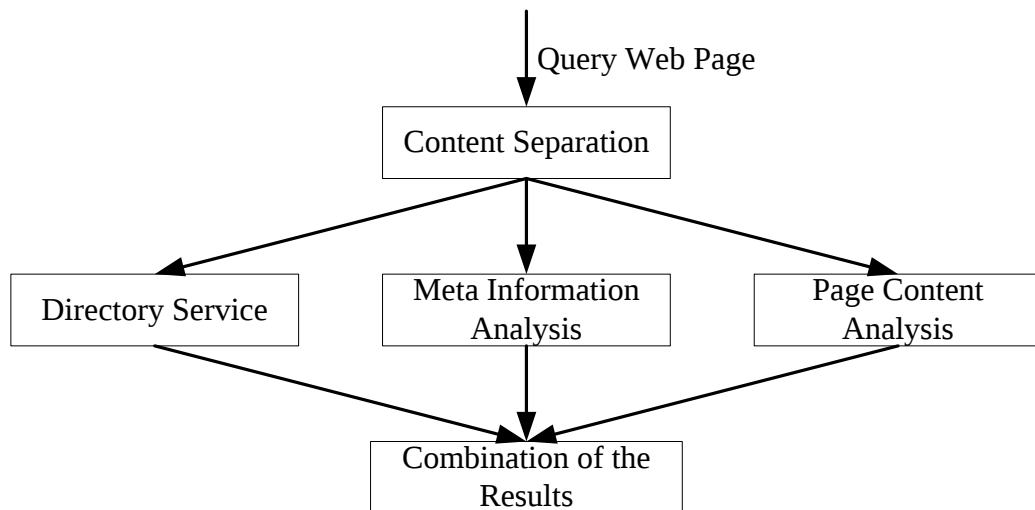


Figure 6.2 The schematic of the Web page automatic classification method

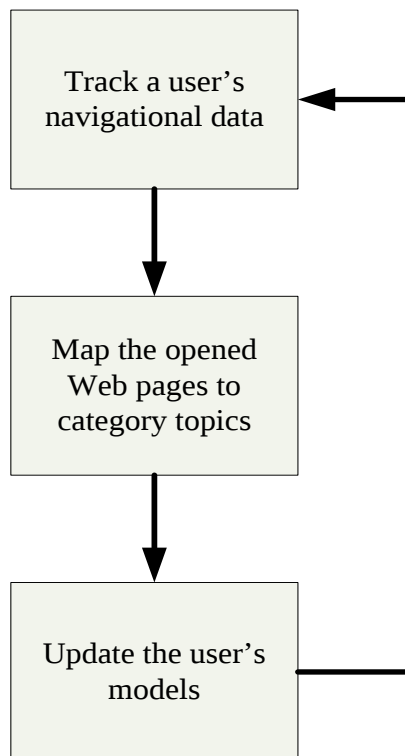


Figure 6.3 The process of user modeling for the recommender system

6.3.2 Web Content Recommendation

This work attempts to recommend Web news content to a user based on their interest and preference models, Web contents, and the behaviors of other users. After a user's interests and preferences have been identified, they are used to determine if certain Web content will be recommended to a user. For an item of Web news, three components of the content need to be quantified: the degree of the Web content belonging to the categories, the source of the Web content, and the time factor. Equation (3.8) or (3.10) can be applied to calculate the degree of the Web content belonging to the categories. The quantification of mapping the Web content to Web sites is straightforwardly determined by its hyperlink. Assuming Web page q is located in Web site s , the degree of the relationship between Web page q and Web site s is assigned as 1, e.g., $DS_s(q)=1$, while the degree of the relationship between Web page q and other Web sites is 0. The time factor is taken into account to indicate the real-time value of news on the Web. It is empirically formulated by

$$T(q) = 0.5 + 0.5 / (1 + wk + \frac{l + h / 24}{7}) \quad (6.1)$$

where $T(q)$ is the time factor for Web page q ; wk is the number of weeks that Web page q has been released; l means the l^{th} day after wk weeks since the release of Web page q ; and h indicates the h^{th} hour after wk weeks and l days since news page q has been released.

Based on a user's models and the analysis of news items, the probability of

recommending a news page to the user can be formulated as

$$pr(q) = T(q) \times \sum_{s \in SS} (DW(s) \times (0.5 + 0.5DS_s(q))) \times \sum_{j \in SC} (DU(j) \times DP_j(q)) \quad (6.2)$$

where $pr(q)$ indicates the probability of recommending Web page q to a user; $T(q)$ is the time factor for Web page q ; SS is the set of all Web sources; $DW(s)$ is the degree of a user's preference for retrieving information from Web site s ; $DS_s(q)$ is the relationship if Web page q is located in Web site s ; SC is the set of all categories considered in the work; $DU(j)$ is the degree of a user's interest on category j ; and $DP_j(q)$ is the probability of Web page q belonging to category j .

Considering that inevitable errors may arise by the current techniques of user modeling and Web page auto-classification, a method of collaborative filtering (CF) is utilized to adjust the recommender system. The process of adjustment based on CF is as follows.

- (a) The classified Web pages that are either clustered manually or acquired from well classified Web sites are considered as test pages.
- (b) Based on interest models, test Web pages are provided to the corresponding users.
In other words, the test Web pages belonging to a certain category topic are delivered to the users who are likely interested in that topic according to the users' models.
- (c) Users' feedback on the test Web pages is collected by the system. The users who

positively respond to the test pages are regarded as trusted users on the related topic until the next round of testing.

- (d) Navigational data from the trusted users is collected by the system. The likelihood of an item belonging to a certain category is increased if trusted users give many positive responses to the item. Moreover, the degree of popularity for news items is calculated based on the click rate of the trusted users.

The degree of popularity for a Web page can be determined by whether or not the users who are interested in a certain topic category recommend the Web page. Based on the process of CF, the recommendation equation (6.2) therefore is modified to:

$$pr(q) = T(q) \times \sum_{s \in SS} (DW(s) \times (0.5 + 0.5DS_s(q))) \times \sum_{j \in SC} (DU(j) \times DP_j(q)) \times (1 + 0.5DO(q)) \quad (6.3)$$

where $DO(q)$ is the degree of popularity for Web page q , which is the percentage of the response from the trusted users.

6.4 Experimental Results and Discussion

In order to evaluate the performance of the proposed Web news recommender system, evaluation experiments and results are presented in this section.

6.4.1 Evaluation of Recommender Systems

Because of intrinsic features of recommender systems, the statistical-based comparison of recommender systems is difficult to apply. Herlocker *et al.* (2004) discussed several reasons why quantitative evaluations of different recommender systems are usually incomparable. First, different recommender systems usually use different data sets. For example, it is not practical that retail recommender systems have identical products. news recommender systems often have different news sources, collection rules, and languages. Advertisement recommender systems have different clients and targets. Second, evaluation of recommender systems is difficult because of the differences in rating properties, such as density and scale. In other words, each recommender system may use a dedicated criteria of rating, which keeps the systems from a fair comparison. Third, the goals for evaluation of recommender systems may differ. Some evaluations concentrate on judging the accuracy of their systems, which emphasizes the capacity of prediction. However, some researchers argued that user satisfaction should be the eventual measuring criterion for recommender systems. Based on this point of view, commercial systems usually measure user satisfaction by sale numbers, while non-commercial systems make inquiries about users' satisfaction upon using these systems.

Some methods of measurement have been proposed to evaluate the performance of Web news recommender systems with various motivations and goals. The number of times that users access recommended news menus is used by Lee and Park (2007) to evaluate their news recommender system for the mobile Web. Their recommender system

provides three types of news menu to users: categories, recommended, and current menus. The read article ratio by menu is used to evaluate the effectiveness of the system. Instead of using the click rate of menus, Li *et al.* (2010) use a click through rate (CTR) to evaluate their personalized news article recommender system. A CTR is the ratio of the number of users who click on an item and the number of times the item is delivered to users. They believe that an effective recommender system is able to improve the overall CTR. Chen *et al.* (2009) consider that the evaluation of the top 10, 20, and 30 recommended news items is more valuable than the average evaluation of overall news items. Therefore, they only provide news listening rates of the top 10, 20, and 30 recommended items to evaluate their phonic Web news recommender system. Bomhardt (2004) proposed a Support Vector Machine (SVM) driven personal recommender system for news Web sites, in which recall and precision rates are used to evaluate the overall prediction quality of the system. The SVM-driven recommender system can reach a precision rate up to 61.69% with the consideration of the top 30 recommended news items.

6.4.2 Sample Results

For the Web news recommender system presented in this chapter, the weights of terms have been computed by utilizing approximately 7000 Web pages from well-categorized Web news sources, which are used to automatically classify Web pages into twelve topic

categories. Test users' interest and preference models are constructed and updated by tracking and analyzing their navigational data. Example interest and preference models of a test are shown in Table 4.1 and Table 4.2, respectively. The Web news captured from several news Web sites are provided to test users based on their user models and the content and location of the news.

In the experiments, news content was collected from Web sources of categorized news, such as CNN, CBC, and Yahoo News. Although news providers have classified the news items obtained from these Web sites, they are re-classified by the Web page classification method in order to consider the scalability issue of news sources including well-categorized and non-categorized content, except when the news items are used as the test information for identifying the trusted users. After the collected news is classified, the recommendation process presented in Section 6.3.2 is conducted based on a user's interest and preference models. Generally, a recommender system will present a user with only the part of the content that is determined to be relevant to the user. However, in the experiments, all collected news arranged according to the probability of recommendation scoring from high to low are presented to a user, which is for effectively evaluating the recommender system discussed in Section 6.4.1. Table 6.1 shows an example of the top 5 recommended news items for a Web user at a certain time.

Table 6.1 Example of recommended news for a user at a certain time

	Title of Recommended News
1	Buildings iPhone app has a solid foundation for architecture lovers (Yahoo)
2	Djokovic returns for Davis Cup semi-final (Yahoo)
3	France reach Davis Cup final (Yahoo)
4	Chivas USA wins on penalty kicks (Yahoo)
5	Schools use video games as teaching tools (CBC)

6.4.3 Evaluation of the Proposed News Recommender System

In order to evaluate the proposed Web news recommender system, 12 participants were asked to surf the Internet for the purpose of building their interest and preference models. Then the news collected from news Web sites was organized by the probability of recommendation based on the participants' models. Over a period of 15 to 20 days, the 12 participants were asked to use the prototype system for rating the relevance of the recommended news items at least once a day. Since it was not practical for the participants to rate the relevance of all news items by reading through them, the participants were asked to view all the titles of the news items in the list, and then decide whether they were willing to read through an item of news. Therefore, the precision and recall rates can be calculated by analyzing the participants' rating. Assuming the system recommends R_{num} pieces of news to a user, and the user rates I_{rcmd} pieces of news from the recommended list as relevant content, while the user rates $I_{relevant}$ pieces of news from

the entire list as relevant content, the precision and recall rates can be calculated by:

$$\text{Precision Rate} = I_{rcmd} / R_{num} \quad (6.4)$$

$$\text{Recall Rate} = I_{rcmd} / I_{relevant} \quad (6.5)$$

Based on the average precision and recall rates obtained from the participants' rating, the precision-recall curve of the system is drawn in Figure 6.4, where the particular pair of precision and recall rates is indicated by a small filled square for each number of recommended items. It shows that the system performs at a good precision rate if the number of recommended items is set to 30 or below. The precision rate drops when the number of recommended items is increased, which also implies that the news content of interest to a user is more likely to appear at the top of the list.

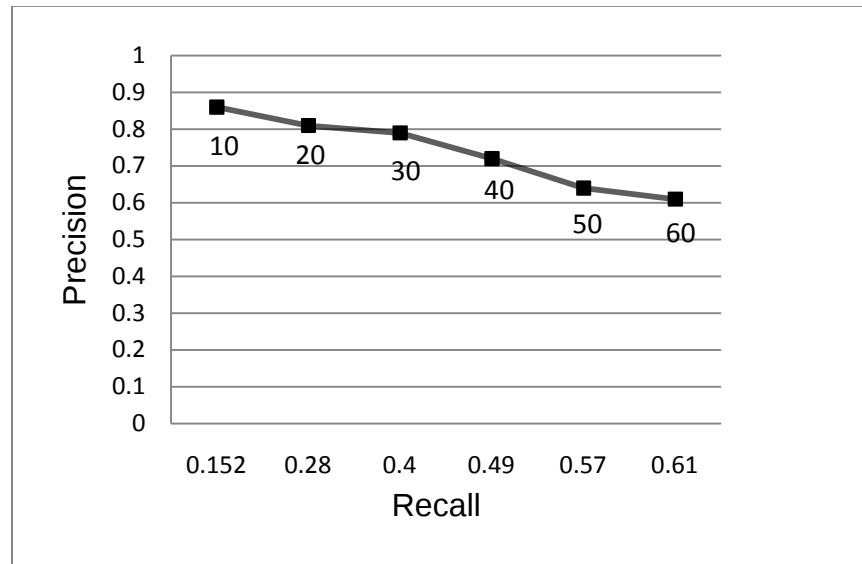


Figure 6.4 The precision-recall curve of the system obtained through the participants' rating

In order to evaluate the effectiveness of the proposed CF integrated in the system, the precision rates of the system using two types of recommendation algorithms are calculated for a comparison study. In other words, based on the same rating data from the participants, the recommendation results generated by Equation (6.2) and Equation (6.3) are used to calculate the precision rate of the system, respectively. Figure 6.5 shows the comparison results of the precision rates with and without the proposed CF method. It can be noticed that the precision rate of the hybrid system is several percentage points higher than the content-based recommender system. Since the proposed CF method is developed for a large-scale recommender system, the performance of the CF method could be fairly evaluated if a large number of users interact with the system.

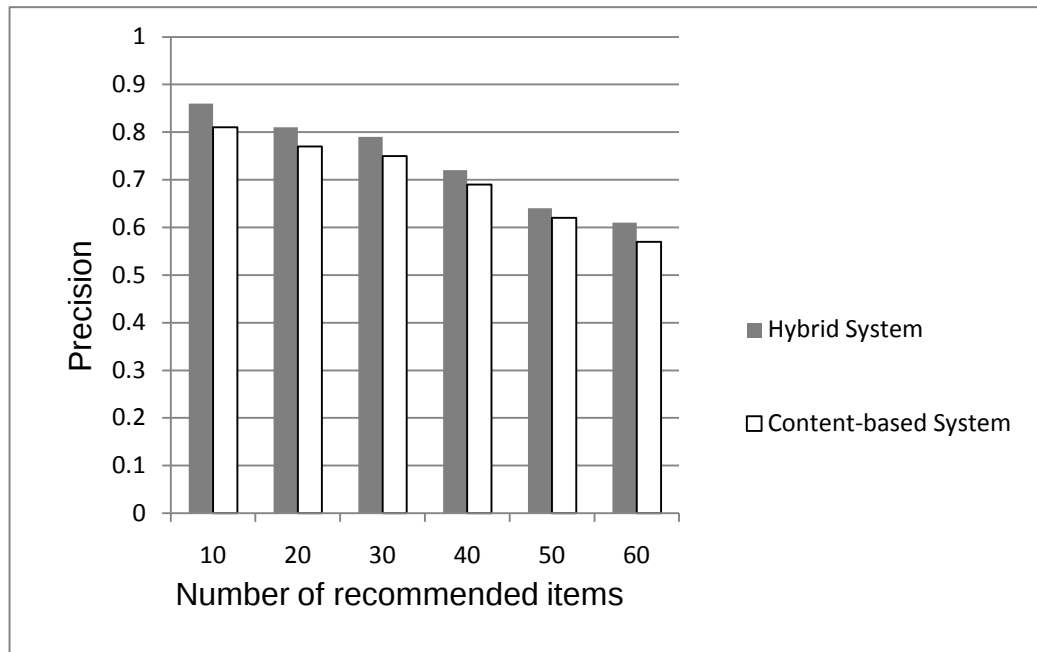


Figure 6.5 Precision rates with and without the proposed CF method

6.5 Summary

A hybrid recommender system for personalized recommendation of Web news to users has been presented. A user's navigational data is tracked and analyzed through a Web page auto-classifying process, which is used to construct and update the user's interest model. The hyperlinks extracted from the user's navigational history are used to build the user's preference model. Web news collected from various news sites is classified by the Web page classification method, and then the probability of recommendation is calculated for a user by matching the contents of news and the user's models. In addition, test information is sent to users in order to identify the trusted users, which is an attempt to improve the performance of the recommender system. The experimental results show the effectiveness of the hybrid recommender system for personalized recommendation of news from the Web for users. A solution of integrating user modeling, content-based filtering, and CF for a personalized recommender system has been proposed.

Chapter 7

Personalized E-Marketing

Electronic commerce (e-commerce) is a type of commerce that conducts the selling and purchasing of products and services over computer networks. The amount of global trade conducted electronically has grown extraordinarily with widespread internet usage. Canada's National Statistical Agency reported that "Canadians used the Internet in 2009 to place orders for goods and services valued at \$15.1 billion, up from \$12.8 billion in 2007" (Statistics Canada, 2010). Figure 7.1 shows the increasing trend in the value of online orders from 2005 to 2009 in Canada (Statistics Canada, 2010). Many companies have realized the importance of e-commerce and its many benefits. In order to effectively exploit this business model, several sub-fields of e-commerce have been studied recently, including electronic marketing (e-marketing), online transaction processing, supply chain, etc. E-marketing is the marketing model over the Internet. It has some advantages over traditional marketing, such as low cost, instant connection, and global targeting. However, since many e-commerce companies have a large number of customers, e-marketing is challenged by how to effectively target customers. Personalization in e-

commerce and e-marketing is an approach to offering personalized services to customers, which aims to improve the efficiency and targeting in e-commerce and e-marketing.

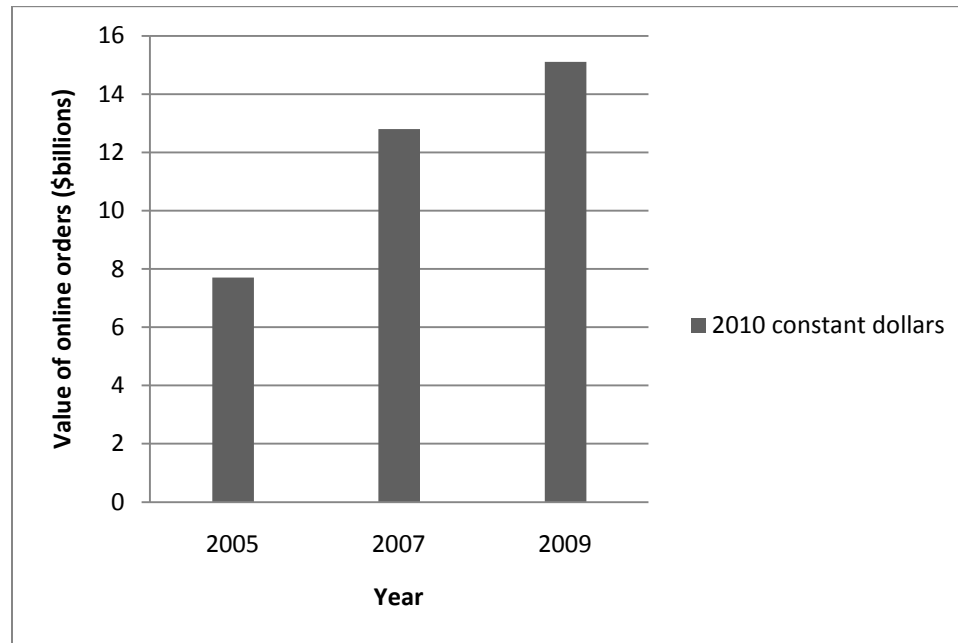


Figure 7.1 Value of online orders in Canada

7.1 Personalization in E-Commerce

In the context of an e-commerce or e-marketing strategy, personalization is the process of exclusively creating a tailored e-shopping experience for every customer or a specific group of customers (Adomavicius and Tuzhilin, 2005b). The ideal personalization scenario in e-commerce is that an e-shop knows the specific interests and needs of each customer, and can present the relevant products or services in its inventory to customers.

Table 7.1 shows the descriptions and comparison of general e-commerce and personalized e-commerce (Mulpuru *et al.*, 2007).

Table 7.1 General e-commerce and personalized e-commerce

	Type of Targeting	Practical Application
General E-Commerce	One e-commerce solution to all customers	Http://www.gm.ca uses the same interface and presents the same types of cars to all customers.
Personalized E-Commerce	A specific solution to a specific customer	Http://www.amazon.com presents different products to customers based on navigational histories or purchase records.
	A specific solution to a group of customers	Http://www.rogers.ca provides different homepages and service options to customers in different provinces.

7.1.1 Benefits of Personalization

Personalization strategies have increasingly gained interests of retailers and researchers, owing to the benefits that are associated with personalized systems for e-commerce. A successful application of personalization can improve the performance of an e-commerce system in the perspectives of increasing sales and obtaining customer loyalty. Some benefits that are brought by personalization are described as follows (Goy *et al.*, 2007).

- **Winning potential customers.** When a consumer navigates into an e-shopping Web site, it is sometimes as the result of a Web link redirection. In most cases, a Web search engine may lead a consumer whilst querying for specific products, brands, types, or on-sale information. If the e-shopping site cannot capture the attention of this type of customer at the first place, it will lose potential customers because they will turn back to the search engine for other search results. However, a personalized e-commerce system may provide more information related to a user's intents and interests than her or her original expectation, which is helpful to win the customer in the competitive e-market place.
- **Increasing customer loyalty.** The term "customer loyalty" is usually defined as a customer's behavior of repeatedly using a service or purchasing in a shop. There are many ways an e-shop can obtain customer loyalty that have a common goal: meet customer satisfaction as much as possible. When a customer's intents and interests are accurately modeled by a personalized

system, he or she may find the items of interest in the recommended list and be relieved from the time-consuming search on the Web. The experience on this customer-centric service can impress a customer with high-efficiency and convenience. Therefore, personalization is an important method to increase customer loyalty. It is also a crucial part of Customer Relationship Management (CRM), which is a research topic involving the improvement of customer loyalty by managing elements of the business relationship.

- **Decreasing marketing cost.** For a business company, it always pursues a high-efficiency of marketing investment versus benefit return. E-shops have spent more on online marketing approaches than ever before, such as email and pay-per-click advertisement. Obviously, the goal of placing advertisements is to receive more profit in return. A conventional method of marketing is to deliver information to either all customers or a random group of customers, which lacks targeting. On the contrary, the strategy of personalization aims to classify customers and present the relative information to customers. Apparently, it can avoid unnecessary marketing investment and increase the ratio of return over cost.
- **Increasing sales.** Based on the three benefits aforementioned, it can be seen that the strategy of personalization can help an e-shop by winning potential customers, keeping loyal customers, and efficiently placing advertisements. In other words, an e-shop can attract and satisfy customers by applying

personalization. Generally, an e-shop can increase sales benefiting from the increasing amount of customers and efficient recommendation method.

7.1.2 Personalized E-Commerce Models

Currently, internet shops have become an integral part of people's everyday life and an important marketing channel for retailers. Online retailers take advantage of the Internet for its relatively low costs, instant response, and easy establishment (Roland and Kannan, 2003). The strategy of personalization is usually adopted to e-commerce, because it can improve system efficiency and increase sales. However, some difficulties of effectively applying personalization to e-commerce systems are challenging online retailers and researchers. One of the obstacles is how to filter redundant information. A personalized system can track a user's navigational data by various means, such as with an internet cookie, navigational record, survey, and purchase history. Although a user's interests can be modeled by analyzing the collected personal data, the redundant information in the collection may decrease the accuracy of the user models. For example, during a customer's navigational session, the pop-up pages should be considered the redundant information. The second difficulty is how to follow a user's behavior and update his or her models promptly. A customer's data is always changed continuously for many reasons. Some of these changes are related to their interests of purchase, which is important for a personalized e-commerce system. Online retailers need to find an effective method to monitor a user's data and keep it updated. The third difficulty is how

to identify a user's intents and interests. As discussed in Chapter 3, there are two types of user modeling approaches: explicit and implicit methods. Since it is not practical to ask a customer for his or her intents every time when he or she browses an e-shop, implicit methods are usually adopted in e-commerce systems. Therefore, e-shops and researchers are challenged by how to predict a customer's likely future purchases according to previous and current navigational data.

Personalization in e-commerce applications can be discussed from several aspects, such as how to acquire personal information, how to provide a personal interface, how to cluster customers, and how to match products to user models. Table 7.2 shows several software technologies of how to interact with customers and acquire personal information (Wu *et al.*, 2003). The method of explicit profiles is usually used in personalized services that require only a small amount of personal information. Cookies are used to identify a customer or a machine in the Web. Personal tools are usually provided in portal Web sites and Web trading systems. And the method of tracking a user's navigation is widely used in implicit-based Web systems, such as recommender systems, e-shops, and streaming video services.

Table 7.2 Interaction models of personalized e-commerce systems

Explicit Profiles	Customers are asked to register and provide personal information such as shipping address, interests, etc. to e-commerce Web sites. A customer's personal information is stored in a Web server. Therefore, e-shops can use user profiles for personalized services.
Cookies	Web cookies, also known as browser cookies in Web applications, are small pieces of text document stored in customers' local machines. A cookie is sent as an HTTP header by a Web server to a customer's local machine. It can be used for identifying customers, and storing site preferences, shopping cart contents, or any information that can be utilized by a personalized e-commerce system.
Personal Tools	Some Web sites provide personal tools to users, which allows users to create personalized services or interfaces themselves. For example, in Yahoo! (www.yahoo.com), customers can use the provided tools to set up their pages containing links of personal interest.
Navigational History in Servers	When a customer browses an e-commerce Web site, his or her navigational data in the Web site can be tracked by the server. Some implicit-based personalized systems utilize the tracked navigational data to build users' models. For example, Amazon (www.amazon.com) tracks a customer's navigation in the Web site, so that it can recommend related items to the customer based on the viewed products.

An e-commerce Web site has many features and components, such as Web addresses, interfaces, and content. Therefore, personalization can be implemented to an e-commerce system through different approaches. Link personalization is an approach to insert additional relevant links for a user. In this approach, a user can generate links directing to the contents of interest in a Web site. Figure 7.2 shows an example of link personalization that is applied at the Excite! Web site. Interface personalization is an approach to customizing the page display for a user. Using the personal tool provided by a Web site, a user can adjust topics, frame size, frame location, font type, font size, color template, background, etc. Figure 7.3 shows an example of interface personalization applied at the Google Web site. Since content is one of the most important features for an e-commerce system, content personalization attracts a lot of attention in e-commerce. The purpose of content personalization is to provide optimized information or products for users based on user profiles or models. Similar to the filtering techniques of news recommender introduced in Chapter 6, three filtering methods are employed in content personalization: content-based, collaborative, and hybrid filtering. Figure 7.4 gives an example of content personalization using the hybrid method in the YouTube Web site. In YouTube, when a user opens a streaming video, the relative video clips with high numbers of clicks are promoted to the user.

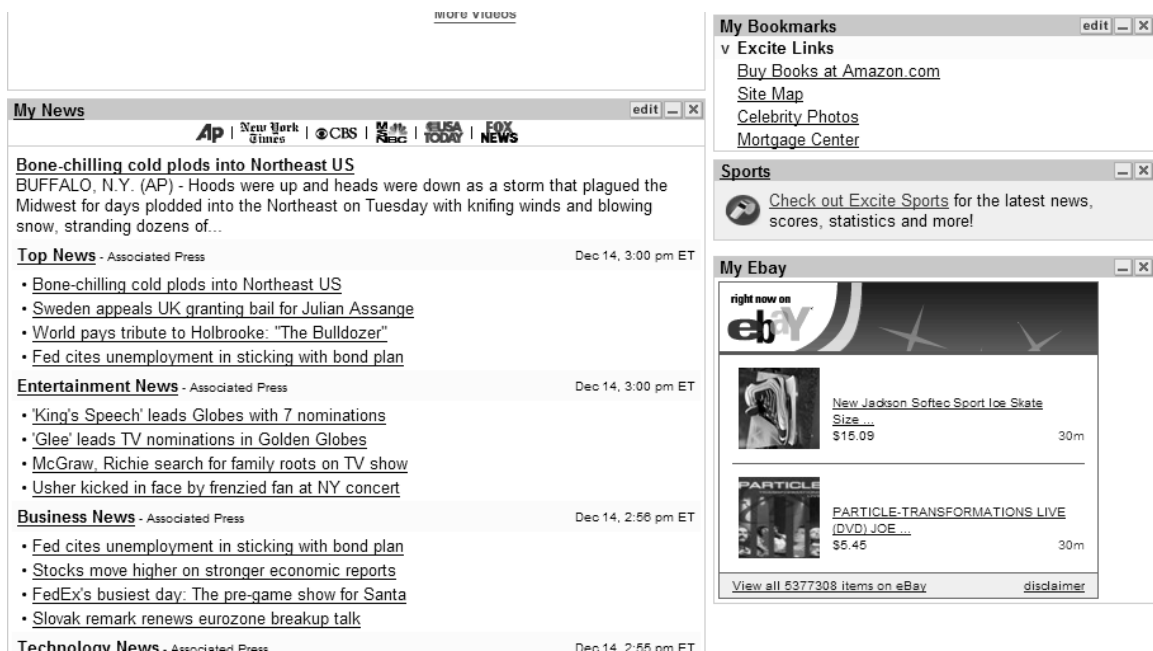


Figure 7.2 An example of link personalization

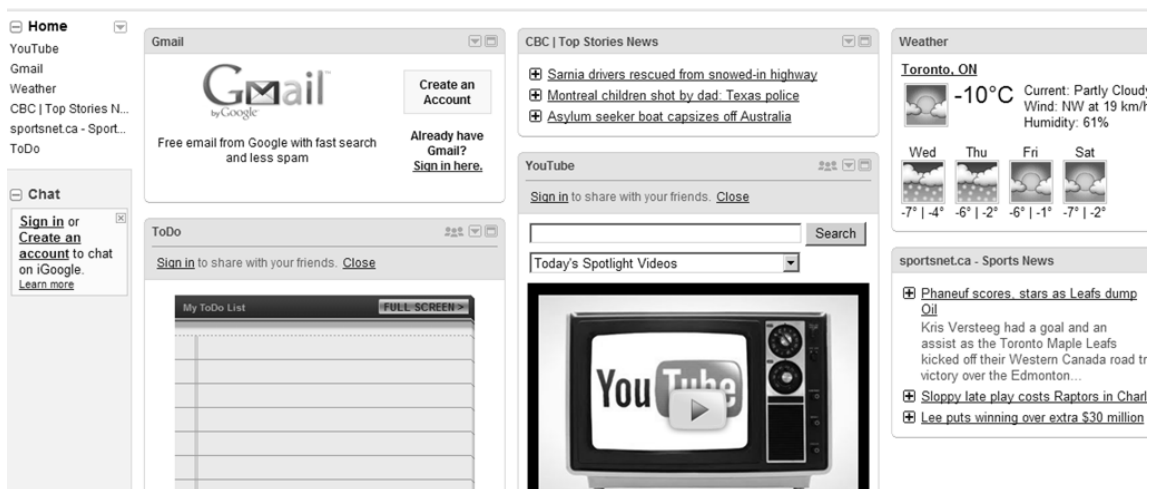


Figure 7.3 An example of interface personalization



Figure 7.4 An example of content personalization using hybrid filtering

7.2 E-Marketing

E-marketing has been recognized and utilized as an important facet of marketing due to its inherent advantages such as low costs, speedy implementation, and easy customer tracking. There are several mediums in which e-marketing can be facilitated: telephone, radio, television, and the Internet. In this dissertation, only internet based e-marketing is considered and discussed. Compared to traditional marketing, e-marketing has the following properties:

1. The delivery of information, services, and products is completely or partially through the Internet.
2. Electronic payment is usually adopted by e-marketing systems.
3. The technologies towards automatic transactions and procedures are usually utilized in e-marketing systems.
4. E-marketing systems can be easily expanded.

In other words, e-marketing can be defined as the technology of using electronic mediums such as the Internet to achieve marketing and promotion objectives. The speedy internet-based interactions between e-marketing systems and customers make it possible to achieve high efficiency for the systems and satisfy customer needs. Generally, any online service can be coupled with an e-marketing system, for instance: e-mails, Web sites, bulletin boards, instant message systems, etc. The email-based e-marketing systems have been used widely because of their effective targeting. A customer is more likely to view the contents in his or her email inbox than the contents from any other information source. Nevertheless, the email-based e-marketing systems are criticized for the pervasive email scams and spam. With the rapid progress of e-business and e-commerce, Web site based e-marketing, especially the e-shopping site-based e-marketing, has played an important role in e-business systems. Amazon is a great example of an e-shopping company that successfully incorporated the technique of e-marketing into their shopping Web sites. Figure 7.5 shows a partial screenshot of Amazon Web site.

Customers Who Viewed This Item Also Viewed



Product Image	Product Name	Price
	Lenovo Ideapad Y560 06462QU 15.6-Inch Laptop with 500GB and 32GB...	CDN\$ 1,041.18
	Lenovo IdeaPad Y560 06462CU Laptop (Bilingual) (Black)	CDN\$ 1,114.99
	HP Pavilion dv6-1230ca 15.6 inch Entertainment Notebook PC	CDN\$ 1,036.99
	Lenovo Ideapad Y560 06462KU 15.6-Inch 3D Laptop (Black)	CDN\$ 1,192.70

Customers Viewing This Page May Be Interested in These Sponsored Links [\(What is this?\)](#)

- [IdeaPad Y560 Laptop Sale](#) 
www.Lenovo.com/ShipsFree - Buy **IdeaPad Y560 Laptop** w/ Intel® Core™ i7. Save \$250, Order Today!
- [Buy Lenovo in Canada](#) 
www.FrontierPC.com - Immediate Shipping from Toronto Vancouver, Calgary & Halifax Canada
- [Laptop Computers On Sale](#) 
MemoryExpress.com - Free Shipping in Canada On All **Laptop** Computers

Figure 7.5 A partial screenshot of Amazon Web site

The screenshot of the Amazon Web site presents some progress and trends of current e-marketing systems:

1. The personalization of e-marketing is pursued for a better customer targeting.
2. Both customers and products are classified for the personalization of e-marketing.

3. Both content-based and collaborative-based filtering methods are utilized for e-marketing systems.
4. The criteria of matching a customer's needs to products can be based on diverse attributes, such as price, brand, function, etc.

It can also be observed that such e-marketing systems are item-centric, because the classification of customers is based on the items viewed by customers. Such e-marketing systems ignore the process of modeling a customer's long-term interests and preferences. Instead, the systems only derive a customer's interest model as his or her recently viewed items. This method can simplify the design of the e-marketing systems and decrease the computational burden. However, it will also decrease the effectiveness of the e-marketing systems and may not satisfy a customer's needs. Therefore, a model of personalized e-marketing system based on a customer's long-term interest model is introduced in this chapter.

7.3 Customer Classification for E-Marketing

The proposed personalized e-marketing framework has four components: determining a customer's general interest model, determining a customer's local browsing model, classifying Web customers, and creating a personalized marketing plan for e-commerce (Wen *et al.*, 2009). Figure 7.6 shows the main steps of the proposed framework.

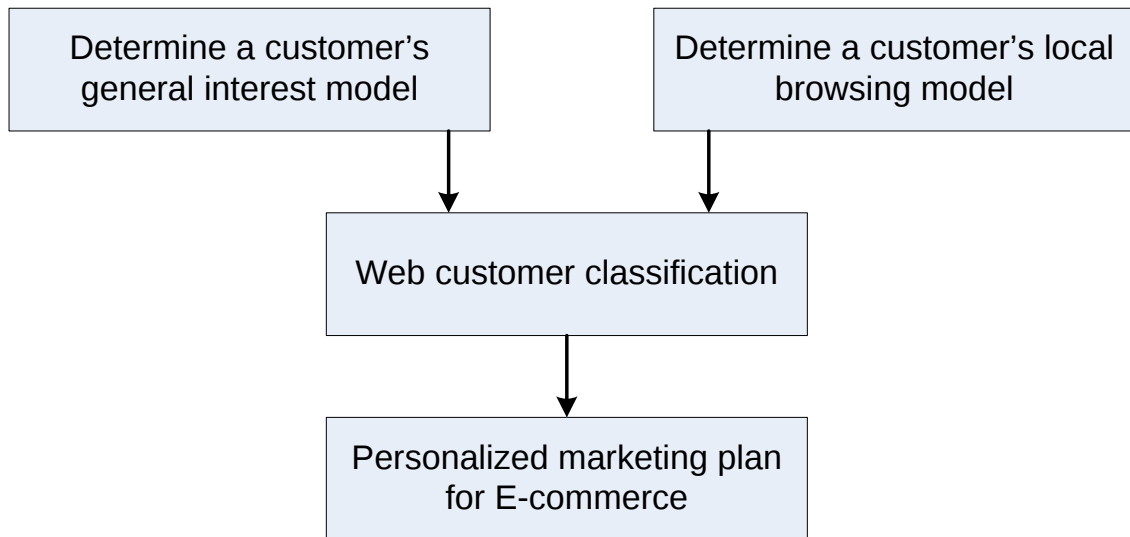


Figure 7.6 The main steps of the proposed customer classification framework

7.3.1 Customer Interest Modeling

The generation of a customer's interest model is based on the user modeling method discussed in Chapter 4. The process of building a Web user's interest model can be divided into the following three steps:

1. Tracking the user's navigational data,

2. Using the cumulative weight to determine the content of the user's browsed Web pages, and
3. Employing the Naïve Bayes model to update the user's interest model.

7.3.2 User Local Browsing Modeling

The proposed approach aims to develop a general personalized e-commerce model, which implies that this work will not be limited to a single type of Web service system. However, in order to eliminate descriptive complexity and confusion, the case of an internet shop is used as the main platform in this research. The products in an internet shop can be represented in a two layer set: $[j_1[P_{11}, P_{12}, \dots, P_{1m}, \dots], j_2[P_{21}, P_{22}, \dots, P_{2m}, \dots], \dots, j_n[P_{n1}, P_{n2}, \dots, P_{nm}, \dots], \dots]$, where j_n is the n^{th} category, and P_{nm} means the m^{th} product in the n^{th} category. Through monitoring a Web customer's behavior and navigational history on the internet shop, the customer's browsing model regarding the internet shop can be represented as $u_k[[j_1[P_{11}(T, B), P_{12}(T, B), \dots, P_{1m}(T, B), \dots], j_2[P_{21}(T, B), P_{22}(T, B), \dots, P_{2m}(T, B), \dots], \dots, j_n[P_{n1}(T, B), P_{n2}(T, B), \dots, P_{nm}(T, B), \dots], \dots]]$, where u_k represents Web customer k , and $P_{nm}(T, B)$ is the statistical result showing the overall time customer u_k has spent browsing product P_{nm} and the number of times that customer u_k has bought product P_{nm} from the internet shop within a certain period of time.

7.3.3 Customer Classification for E-Commerce

Applying personalized e-commerce requires classifying customers based on customers' unique behaviors and needs. Therefore, it is a crucial issue for an internet shop to identify a Web customer's implicit objectives when the customer is browsing the shop's Web site. Based on the customer interest model and the customer local browsing model described earlier, a method for classifying a customer for e-commerce is proposed, which can be formulated as:

$$dpp(u_k) = \sum_{j \in SC} (DU(j) / (1 - \ln \frac{\sum_{k=1}^{np_j} T_{j,k}}{T})) * (1 + \sum_{k=1}^{np_j} B_{j,k}) \quad (7.1)$$

$$CG(req) = \{u_k \in U : req_{min} \leq dpp(u_k) < req_{max}\} \quad (7.2)$$

where $dpp(u_k)$ is the degree of potential purchase (DPP) on the Web site for customer u_k ; SC is the set of categories; $DU(j)$ is the degree of customer's interest on category j ; np_j is the overall number of products in category j on the internet shop; T is a constant representing a certain period of time; $T_{j,k}$ is the total time that customer u_k has spent on browsing product k in category j during period T ; $B_{j,k}$ is the number of times that customer u_k has purchased product k in category j ; req represents a range of degrees for a potential purchase; $CG(req)$ is the set of customers satisfying the certain requirement of customers' DPP; req_{min} is the minimal value of req ; req_{max} is the maximal value of req ; and U is the set of all customers.

Equation (7.1) is used to determine a Web customer's DPP on an internet shop. A higher $dpp(u_k)$ implies a higher possibility that customer k would like to purchase a product on the internet shop. Equation (7.2) shows the method of classifying customers based on grading their degree of potential purchase.

7.4 Personalized E-Marketing Planning

After a Web customer's degree of potential purchase on an internet service provider has been estimated, a personalized marketing plan for e-commerce becomes feasible. An online service company can develop a Web-marketing strategy based on its own circumstances and the customer modeling. As far as an internet shop is concerned, its business objectives are: making profit, occupying the market, attracting customers, and expanding the company. Therefore, a series of e-marketing and promotion plans based on the customer modeling can be drawn up in order to maximize profit and increase marketing impact. Table 7.3 gives a brief example of an e-marketing and promotion strategy based on the proposed Web-user modeling for an internet shop. The intention of the personalized plan is either to increase sales or to promote the internet shop itself by increasing the Web customer's browsing time on the Web site.

Table 7.3 A sample personalized e-marketing and promotion plan

Customer Set	E-marketing and promotion plan
Customer with high degree of potential purchase	Emphasis on leading the customer to the payment page; directly providing product details such as price and special features.
Customer with relatively high degree of potential purchase	Emphasis on showing positive comments on products; providing on-sale information.
Customer with relatively low degree of potential purchase	Providing on-sale information; promoting the Web site by showing entertaining products.
Customer with low degree of potential purchase	Emphasis on promoting the Web site by showing entertaining products.

7.5 Evaluation Criteria

In order to evaluate the performance of the proposed work, some evaluation criteria are presented in this section. According to the structure of the proposed framework, it is obvious that the accuracy of user classification and the effectiveness of the personalized e-marketing planning determine the overall performance of the proposed system.

The degree of potential purchase only describes a Web customer's likely behavior in an internet shop, which means its accuracy can never be confirmed because the future statistical data of customers can never be obtained. Therefore, this work proposes a method to indirectly evaluate the performance of user classification. If no personalized e-marketing plan is applied, 20% of customers with the highest and lowest DPP are tracked and analyzed in their respective time intervals. A simple equation is given to represent the accuracy of the user classification:

$$A_u = \sqrt{p_{high} \times (1 - p_{low})} \quad (7.3)$$

where A_u indicates the accuracy of the user classification; p_{high} is the purchase rate of the customers with high DPP; and p_{low} is the purchase rate of the customers with low DPP.

In terms of the personalized marketing planning, its performance can be assessed by using the fusion of component comparison. The customers' navigational data recorded in two consecutive time slices, with and without personalized marketing planning, is employed to make a comparison. The effectiveness of personalized marketing planning over normal marketing planning is given by

$$E_{pm} = \eta_6 \left(\frac{N_p}{N_{np}} - 1 \right) + \eta_7 \left(\frac{I_p}{I_{np}} - 1 \right) + \eta_8 \left(\frac{T_p}{T_{np}} - 1 \right) \quad (7.4)$$

where E_{pm} represents the effectiveness of the personalized marketing planning; N_p is the total number of customers who purchase any product during the personalized marketing

time slice; N_{np} is the total number of customers who purchase any product during the normal marketing time slice; I_p is the total number of items sold during the personalized marketing time slice; I_{np} is the total number of items sold during the normal marketing time slice; T_p is the overall browsing time by all customers during the personalized marketing time slice; T_{np} is the overall browsing time by all customers during the normal marketing time slice; and η_6 , η_7 and η_8 are factors of importance corresponding to customer, product, and brand implanting respectively, which are determined by the company's long term and short term strategies.

7.6 Experimental Results

The proposed framework is a type of e-commerce model combined with Web user classification. Extensive experiment results and evaluation results can only be achieved by collaborating with an industrial company. Limited by practical conditions, this research only presents some experiment results in a simulation environment. Assume the Web site <http://www.ebay.ca> is an internet shop embedded with the proposed user classification module. The following experiment results show the performance of the proposed method through monitoring a user's buying activities.

First, the user's general interest model is acquired through the work introduced in Chapter 4. The user's degrees of interest on some categories are shown in Table 7.4.

Table 7.4 A user's partial degrees of interest

Category	Degree of interest
Electronics	0.43
Tennis	0.61
Computer	0.49
Politics	0.23

By using the internet monitor program, the user's navigational data has been recorded. It shows that the user has spent 5 hours in 3 days browsing projector items that belong to the electronics category. Therefore, the user's local browsing model related to the electronics category can be determined, as shown in Figure 7.7.

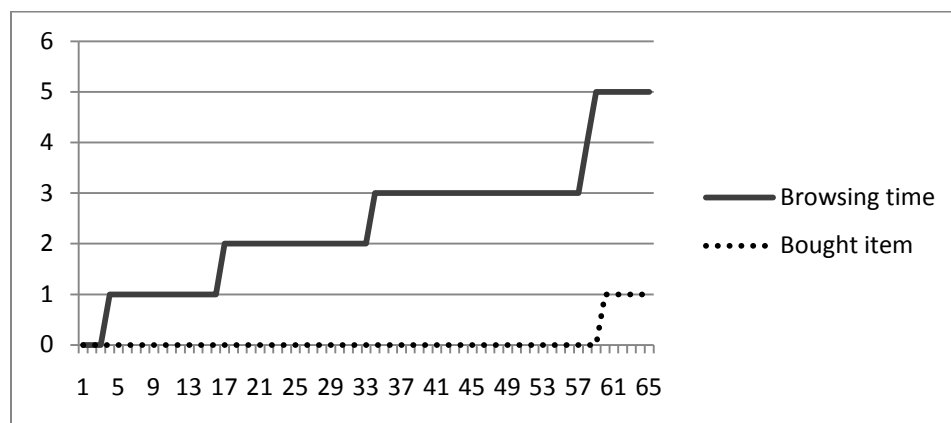


Figure 7.7 A customer's local browsing model

Applying Equation (7.2), the user's degree of potential purchase is obtained, as shown in Figure 7.8.

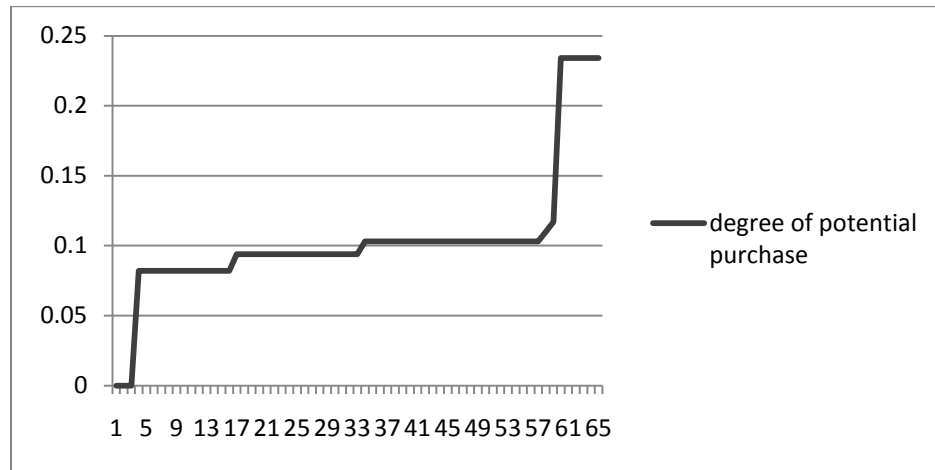


Figure 7.8 A customer's degree of potential purchase for electronics

Though the experiment results for the proposed personalized marketing strategy is unable to be provided due to limited experiment conditions, it is still shown that a Web customer is able to be effectively classified only if a group of deliberate thresholds matching the internet shop are determined.

7.7 Summary

A framework to classify a Web customer based on the customer's general interest model and local browsing model has been presented. A customer's interest model is built based

on the work presented in Chapter 4. By monitoring the user's navigational behavior at an internet shop, the user's local browsing model represented by browsing time and the number of items bought is identified. A Web user classification method is proposed, which aims to identify more potential customers for an online service company. A sample of a personalized e-marketing strategy is also presented in this work. In the experiment, a customer's navigational data on *eBay* is examined and analyzed using the proposed user classification method. The experiment results show that the proposed method can classify customers. The proposed model can be potentially integrated with personalized online services. However, future work is necessary to extend testing of the proposed approach. A commercial Web service would be ideal for examining the performance of the proposed model.

Chapter 8

Conclusions and Future Research

Web personalization is a process that provides customized online services using user profiles. Many research topics such as Web page classification, user modeling, and system integration are related to Web personalization. The overall goals of conducting personalization within online services are to satisfy a user's specific needs, save a user's time and effort, and improve the performance of systems. Due to the variety of online services, the approaches to Web personalization vary in different circumstances. However, most personalized Web systems more or less share the same structure and components, because they all need to determine a user's needs, classify products, and perform the process of matching products with their needs.

8.1 Main Contributions of the Dissertation

The overall contribution of this dissertation is the development of three Web personalization systems: a personalized search system, a personalized Web news recommender, and a personalized e-marketing system. Contributions are made in the

research areas of Web page classification, user modeling, and system integration, as summarized below.

A framework of Web page classification based on three different features is presented in this dissertation. In the preprocessing step, terms determined according to an ontology database are weighted using a TF/IDF method. Then, the proposed classification approach uses directory service, meta information, and content information for classifying Web pages separately. A fusion method is used to re-classify Web pages based on the results given by the three separate classifiers. In the experiments, the proposed method is applied to automatically classify Web pages into five categories. The experimental results show that the proposed method can effectively classify some types of Web pages, such as news and shopping Web pages. However, the performance of the proposed classification method is decreased when the Web pages contain heavy noisy information. Nevertheless, the proposed classification method can be used for the online services that require unsupervised Web page categorization.

A method to building and updating a user's models of interests and preferences based on the user's Web navigational data is developed. Using the proposed Web page classification method, Web pages viewed by a user are mapped to categories based on certain topics. A user's models of interests and preferences are updated using the Naïve Bayes model. In the experiments, the proposed method is utilized to determine what category topics a user may be interested in, and what information sources the user prefers

to retrieve information from. The experimental results show that a user's model of interests and preferences can be effectively built by the proposed method.

A personalized Web search system using both Web content-based and behavior-based modeling approaches is proposed. For the purpose of this research, a user's search context model is constructed according to the user's click-through during a search session. Through analyzing a user's search context information, the proposed system determines how the user's interest and preference models influence his or her search behavior. Then, by computing the ranking score of a Web page based on the user models, a re-ranking algorithm is developed. The experimental results illustrate the effectiveness of the proposed personalized search approach.

A hybrid recommendation system utilizing both content-based and collaborative filtering for recommending Web news is presented. The proposed Web page classification method is applied to classify collected Web news from various Web sites. By calculating the similarity between a user model and the content of a news item, the probability of recommending the news to the user is determined. In order to improve the performance of the system, a method of determining trusted users is utilized. In the process, test information is sent to users, which is used to identify trusted users by their responses. Therefore, this work proposes a solution of integrating user modeling, content-based filtering, and collaborative filtering for a personalized recommendation system. The experimental results demonstrate the effectiveness of the hybrid recommender system for recommending Web news.

A framework to classify a Web customer based on the customer's general interest model and local browsing model is proposed. A customer's interest model is built by using the Web page classification and user modeling approaches developed in this dissertation. By monitoring the user's navigational behavior at an internet shop, the user's local browsing model represented by browsing time and the number of items purchased is identified. A Web customer classification method is proposed, which aims to attract more potential customers for the e-commerce company. A sample personalized e-marketing strategy is also presented. In the experiment, a customer's navigational data on eBay is examined and analyzed using the proposed customer classification method. The experiment results show that the proposed method can classify the customers correctly. The proposed model can be potentially integrated with the personalized online service systems.

8.2 Future Research

There are some limitations observed in this research. Further study is necessary to improve the proposed approaches.

In the approaches of Web page classification and user modeling, more categories and classes should be considered to describe Web pages and user interests. Since the Internet is a very large repository of diverse data and information, the performance of the Web page classification approach can only be thoroughly investigated when a more

comprehensive category system is used. For instance, if a category is not included in the category index, the Web pages relevant to this category have to be classified into other categories, which affects the precision and recall rates. For the same reason, the category structure of the user modeling method needs to be extended in the future research.

In the user modeling approach, other updating algorithms can be investigated to compare with the Naïve Bayes algorithm. Although the calculation of a user's models is subjective, different updating algorithms can be compared with respect to the computational burdens, time to boundary, robustness, etc. In an environment with high scalabilities of users and data, an algorithm that can estimate user models efficiently and effectively is necessary.

Experiments with more participants are needed to demonstrate the performance of the personalized Web search, news recommender and e-marketing systems. To a system whose performance is meant to be evaluated by users, the number of participants is one of the most important experimental factors. The experimental results can be more convincing if more users test and evaluate the systems.

In order to investigate the practical value of the personalized e-marketing model, it is necessary to integrate the model into a pilot or commercial system. The purposes of any e-commerce model are to serve customers better and increase profit for companies. If an internet shop can test the proposed personalized e-marketing model, advantages and limitations of the model can be thoroughly investigated.

References

- Adam, N.R., V. Atluri, E. Bertino, and E. Ferrari (2002). A content-based authorization model for digital libraries. *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 296-315.
- Adomavicius, G. and A. Tuzhilin (2005a). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749.
- Adomavicius, G. and A. Tuzhilin (2005b). Personalization technologies: A process-oriented perspective. *Communications of the ACM*, vol. 48, no. 10, pp. 83-90.
- Agichtein, E. (2006). Web information extraction and user modeling: Towards closing the gap. *IEEE Data Engineering Bulletin*, vol. 29, no. 4, pp. 37-44.
- Aggarwal, C.C., S.C. Gates, and P.S. Yu (2004). On using partial supervision for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 245-255.
- Albers, S. and M. Clement (2007). Analyzing the success drivers of e-business companies. *IEEE Transactions on Engineering Management*, vol. 54, no. 2, pp. 301-314.
- Allen, R.B. (1990). User models: Theory, method and practice. *International Journal of Man-machine Studies*, vol. 32, no. 5, pp. 511-543.
- Ardissono, L., A. Goy, G. Petrone, and M. Segnan (2005). A multi-agent infrastructure for developing personalized web-based systems. *ACM Transactions on Internet Technology*, vol. 5, no. 1, pp. 47-69.

- Aspray, W. and P.E. Ceruzzi (2008). *The Internet and American Business*. Cambridge, MA: MIT Press.
- Avancini, H., L. Candela, and U. Straccia (2007). Recommenders in a personalized, collaborative digital library environment. *Journal of Intelligent Information Systems*, vol. 28, no. 3, pp. 253-283.
- Baeza-Yates, R. and B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Harlow, England: Addison-Wesley.
- Bernstein, M.S., D. Tan, G. Smith, M. Czerwinski, and E. Horvitz (2010). Personalization via friendsourcing. *ACM Transactions on Computer-Human Interaction*, vol. 17, no. 2, 6:1-6:28.
- Bhatia, M.P.S. and A. Kumar (2009). Contextual paradigm for ad hoc retrieval of user-centric web data. *IET Software*, vol. 3, no. 4, pp. 264-275.
- Biancalana, C. and A. Micarelli (2009). Social tagging in query expansion: A new way for personalized Web search. *Proceedings of the 2009 International Conference on Computational Science and Engineering*, vol. 4, pp. 1060-1065.
- Bobadilla, J., F. Serradilla, and J. Bernal (2010). A new collaborative filtering metric that improves the behavior of recommender systems. *Knowledge-Based Systems*, vol. 23, no. 6, pp. 520-528.
- Bomhardt, C. (2004). NewsRec, A SVM-driven personal recommendation system for news Websites. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Beijing, China, pp. 545-548.
- Boutemedjet, S. and D. Ziou (2008). A graphical model for context-aware visual content

- recommendation. *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 52-62.
- Box, G.E.P., W.G. Hunter, and J.S. Hunter (1978). *Statistics for Experimenters*. John Wiley & Sons.
- Brin, S. and L. Page (1998). The anatomy of a large-scale hyper textual Web search engine. *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.
- Burke. R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331-370.
- Capra, R.G. and M.A. Perez-Quinones (2005). Using Web search engines to find and refind information. *Computer*, vol. 38, no. 10, pp. 36-42.
- Chaffey, D. (2009). *Internet Marketing: Strategy, Implementation and Practice*. Harlow, UK; New York: Financial Times/Prentice Hall.
- Chan, M. and H. Chen (2007). Incorporating Web analysis into neural networks: An example in Hopfield net searching. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 352-358.
- Changchien, S.W., C. Lee, and Y. Hsu (2004). On-line personalized sales promotion in electronic commerce. *Expert Systems with Applications*, vol. 27, no. 1, pp. 35-52.
- Chen, C., Y. Chen, and C. Liu (2007). Learning performance assessment approach using Web-based learning portfolios for E-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 37, no. 6, pp. 1349-1359.
- Chen, W., Zhang, L., Chen, C., & Bu, J. (2009). A hybrid phonic Web news recommender system for pervasive access. *Proceedings of the 2009 International Conference on Communications and Mobile Computing*, Kunming, China, pp. 122-

- Cheung, K., K. Tsui, and J. Liu (2004). Extended latent class models for collaborative recommendation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 34, no. 1, pp. 143-148.
- Chirita, P.A., C. Firan, and W. Nejdl (2006). Summarizing local context to personalize global Web search. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, pp. 287-296.
- Chirita, P.A., W. Nejdl, R. Paiu, and C. Kohlschutter (2005). Using ODP metadata to personalize search. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, pp. 178-185.
- Cho, J., K. Kwon, and Y. Park (2007). Collaborative filtering using dual information sources. *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 30-38.
- Choi, B. and X. Peng (2004). Dynamic and hierarchical classification of Web pages. *Online Information Review*, vol. 28, no. 2, pp. 139-147.
- Choi, D.H. and B.S. Ahn (2009). Eliciting customer preferences for products from navigation behavior on the Web: A multicriteria decision approach with implicit feedback. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 4, pp. 880-889.
- Choi, S.H., Y.-S Jeong, and M.K. Jeong (2010). A hybrid recommendation method with reduced data for large-scale application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 5, pp. 557-566.

- Claypool, M., D. Brown, P. Le, and M. Waseda (2001). Inferring user interest. *IEEE Internet Computing*, vol. 5, no. 6, pp. 32-39.
- Cohen, S., J. Ferreira, A. Horne, B. Kibbee, H. Mistlebauer, and A. Smith (2000). MyLibrary: Personalized electronic services in the Cornell University library. *D-Lib Magazine*, vol. 6, no. 4, pp. 5-7.
- Comer, D. (2007). *The Internet Book: Everything You Need to Know About Computer Networking and How the Internet Works*. Upper Saddle River, NJ: Pearson Prentice Hall.
- Conesa, J., V.C. Storey, and V. Sugumaran (2008). Improving web-query processing through semantic knowledge. *Data & Knowledge Engineering*, vol. 66, no. 1, pp. 18-34.
- Daoud, M., L. Tamine-Lechani, and M. Boughanem (2008). Learning user interests for a session-based personalized search. *Proceedings of the 2nd International Symposium on Information Interaction in Context*, London, UK, pp. 57-64.
- Di Giacomo, M., D. Mahoney, J. Bollen, A. Monroy-Hernandez, and C.M. Ruiz-Meraz (2000). MyLibrary, a personalization service for digital library environments. *Proceedings of the Joint DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, pp. 18-20.
- Eirinaki, M. and M. Vazirgiznnis (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology*, vol. 3, no. 1, pp. 1-27.
- Embleton, K. and H. Heinrich (2008). Searching to find. *Searcher*, vol. 16, no. 2, pp. 22-46.

- Frias-Martinez, E., S.Y. Chen, and X. Liu (2006). Data mining approaches to user modeling for adaptive hypermedia: Survey and future direction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 36, no. 6, pp. 734-749.
- Gao, Y. and Y. Li (2007). An intelligent fuzzy-based recommendation system for consumer electronic products. *Expert Systems with Applications*, vol. 33, no. 1, pp. 230-240.
- Gerstel, O., S., Kuttan, E.S. Laber, R. Matichin, D. Peleg, A.A. Pessoa, and C. Souza (2007). Reducing human interactions in Web directory searches. *ACM Transactions on Information Systems*, vol. 25, no. 4, pp. 20-28.
- Ghosh, J.K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis, Theory and Methods*, New York: Springer Science+Business Media.
- Godoy, D., S. Schiaffino, and A. Amandi (2010). Integrating user modeling approaches into a framework for recommender agents. *Internet Research*, vol. 20, no. 1, pp. 29-54.
- Goy, A., L. Ardissono, and G. Petrone (2007). Personalization in e-commerce applications. *The Adaptive Web, LNCS*, vol. 4321, pp. 485-520, Springer-Verlag Berlin Heidelberg.
- Gunduz, S. and M.T. Ozsu (2003). A user interest model for Web page navigation. *Proceedings of the International Workshop on Data Mining for Actionable Knowledge*, Seoul, Korea, pp. 46-57.
- Guo, Y., K. Ramamohanarao, and L.A.F. Park (2007). Personalized PageRank for Web

- page prediction based on access time-length and frequency. *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, California, USA, pp. 687-690.
- Ha, S.H. (2006). Digital content recommender on the Internet. *IEEE Intelligent Systems*, vol. 21, no. 2, pp. 70-77.
- Haug, W. (2006). Population censuses on the Internet. *IUSSP General Population Conference*, Salvador de Bahia, Brazil, pp18-24.
- Haveliwala, T.H. (2003). Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784-796.
- Health, T. (2008). How will we interact with the Web of data? *IEEE Internet Computing*, vol. 12, no. 5, pp. 88-91.
- Herlocker, J.L., J.A. Konstan, L.G. Terveen, and J.T. Riedl (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5-53.
- Hsieh, S.M., S.J. Huang, C.C. Hsu, and H.C. Chang (2004). Personal documents recommendation system based on data mining techniques. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Beijing, China, pp. 51-57.
- Hurley, N.J., M.P. O'Mahony, and G.C.M. Silvestre (2007). Attacking recommender systems: A cost-benefit analysis. *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 64-68.
- Hwang, H., A. Balmin, B. Reinwald, and E. Nijkamp (2010). BinRank: Scaling dynamic

- authority-based search using materialized SubGraphs. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 8, pp. 1176-1190.
- Isa, D., V.P. Kallimani, and L.H. Lee (2009). Using the self organizing map for clustering of text documents. *Expert Systems with Applications*, vol. 36, no. 5, pp. 9584-9591.
- Jarvelin, K. and J. Kekalainen (2000). IR evaluation methods for retrieving highly relevant documents. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, pp. 41-48.
- Jiang, T. and A. Tuzhilin (2009). Improving personalization solutions through optimal segmentation of customer bases. *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 3, pp. 305-320.
- Jin, X., R. Li, X. Shen, and R. Bie (2007). Automatic Web pages categorization with ReliefF and hidden naive Bayes. *Proceedings of the 2007 ACM Symposium on Applied Computing*, Seoul, Korea, pp. 617-621.
- Kalaignanam, K., T. Kushwaha, and P. Varadarajan (2008). Marketing operations efficiency and the Internet: An organizing framework. *Journal of Business Research*, vol. 61, no. 4, pp. 300-308.
- Kan, M.K. and H.O.N. Thi (2005). Fast Webpage classification using URL features. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany, pp. 325-326.
- Kim, H., P. Howland, and H. Park (2005). Dimension reduction in text classification with

- support vector machines. *Journal of Machine Learning Research*, vol. 6, pp. 37-53.
- Konstan, J.A. (2004). Introduction to recommender systems: Algorithms and evaluation. *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 1-4.
- Koraljka, G. and L. Marianne (2009). Automated classification of Web pages in hierarchical browsing. *Journal of Documentation*, vol. 65, no. 6, pp. 901-925.
- Kotsiantis, S. and P. Pintelas (2005). Logiboost of simple Bayesian classifier. *Informatica*, vol. 29, no. 1, pp. 53-59.
- Lan, M., C.L. Tan, J. Su, and Y. Lu (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721-735.
- Lancieri, L. and N. Durand (2006). Internet user behavior: Compared study of the access traces and application to the discovery of communities. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 36, no. 1, pp. 208-219.
- Laroche, M. (2010). New developments in modeling Internet consumer behavior: Introduction to the special issue. *Journal of Business Research*, vol. 63, no. 9-10, pp. 915-918.
- Lee, H.J., and S.J. Park (2007). MONERS: A news recommender for the mobile web. *Expert Systems with Applications*, vol. 32, no. 1, pp. 143-150.
- Lee, W., C. Liu, and C. Lu (2002). Intelligent agent-based systems for personalized recommendations in Internet commerce. *Expert Systems with Applications*, vol. 22, no. 4, pp. 275-284.

- Leung, K., W. Ng, and D. Lee (2008). Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1501-1517.
- Li, L., W. Chu, J. Langford, and R.E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, USA, pp. 661-670.
- Li, Q., R. Lau, T. Shih, and F. Li (2008). Technology supports for distributed and collaborative learning over the Internet. *ACM Transactions on Internet Technology*, vol. 8, no. 2, pp. 10:1-10:24.
- Liang, T., Y. Yang, D. Chen, and Y. Ku (2008). A semantic-expansion approach to personalized knowledge recommendation. *Decision Support Systems*, vol. 45, no. 3, pp. 401-412.
- Liu, F., C. Yu, and W. Meng (2004). Personalized Web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 28-40.
- Liu, H. and V. Keselj (2007). Combined mining of Web server logs and Web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering*, vol. 61, no. 2, pp. 304-330.
- Long, M.M, and T. Tellefsen (2007). Internet integration into the industrial selling process: A step-by-step approach. *Industrial Marketing Management*, vol. 36, no. 5, pp. 676-689.
- Luxenburger, J., S. Elbassuoni, and G. Weikum (2008). Matching task profiles and user

- needs in personalized Web search. *Proceeding of the 17th ACM International Conference on Information and Knowledge Management*, Napa Valley, USA, pp. 689-698.
- Ma, Z., G. Pant, and O. Sheng (2007). Interest-based personalized search. *ACM Transactions on Information Systems*, vol. 25, no. 1, 5:1-5:38.
- Meiss, M.R., F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani (2008). Ranking Web sites with real user traffic. *Proceedings of the International Conference on Web Search and Web Data Mining*, Palo Alto, USA, pp. 65-76.
- Miller, B.N., J.A. Konstan, and J. Riedl (2004). PocketLens: Toward a personal recommender system. *ACM Transactions on Information Systems*, vol. 22, no. 3, pp. 437-476.
- Miller, G.A. (2009). WordNet – about us. *WordNet*, Princeton University, <http://wordnet.princeton.edu>. (Accessed in March, 2011.)
- Mittal, A., A. Kassim, and T. Tan (2007). *Bayesian Network Technologies: Applications and Graphical Models*. Hershey, PA: IGI Pub.
- Mobasher, B., R. Burke, R. Bhaumik, and J.J. Sandvig (2007). Attacks and remedies in collaborative recommendation. *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 56-63.
- Mulpuru, S., C. Johnson, and S. Wright (2007). Which personalization tools work for ecommerce – and why. Forrester Research Report for eBusiness, Channel & Product Management Professionals, http://www.forrester.com/rb/Research/which_personalization_tools_work_for_ecommerce_%26%238212%3B/q/id/44345/t/2. (Accessed in March, 2011.)

- Ng, W., L. Deng, and D.L. Lee (2007). Mining user preferences using spy voting for search engine personalization. *ACM Transaction on Information Technology*, vol. 7, no. 4, 19-1:19:27.
- Page, L., S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the Web. Technical Report, Department of Computer Science, Stanford University.
- Papastathopoulou, P. and G.J. Avlonitis (2009). Classifying enterprises on the basis of WWW use: A behavioural approach. *Internet Research*, vol. 19, no. 3, pp. 332-347.
- Pathak, A., S. Chakrabarti, and M. Gupta (2008). Index design for dynamic personalized PageRank. *Proceedings of the 24th IEEE International Conference on Data Engineering*, Cancun, Mexico, pp. 1489-1491.
- Pera, M. S., & Ng, Y. (2008). Utilizing phrase-similarity measures for detecting and clustering informative RSS news articles. *Integrated Computer-Aided Engineering*, vol. 15, no. 4, pp. 331-350.
- Qiu, F. and J. Cho (2006). Automatic identification of user interest for personalized search. *Proceedings of the 15th International Conference on World Wide Web*, pp. 727-736.
- Radlinski, F. and T. Joachims (2007). Search engines that learn from implicit feedback. *Computer*, vol. 40, no. 8 pp. 34-40.
- Rangarajan, S.K., V.V. Phoha, K.S. Balagani, R.R. Selmic, and S.S. Iyengar (2004). Adaptive neural network clustering of Web users. *Computer*, vol. 37, no. 4, pp. 34-40.

- Ras, M. and S.V. Bussel (2007), Web archiving user survey. Research Report, National Library of the Netherlands, http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf. (Accessed in March, 2011.)
- Ridley, D.D. (2009). *Information Retrieval*, Hoboken, N.J.: Wiley.
- Ricardo, B.Y. and R.N. Berthier (1999). *Modern Information Retrieval*. New York: Addison-Wesley/ACM Press.
- Robnik-Sikonja, M. and I. Kononenko (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, vol. 53, no. 1-2 pp. 23-69.
- Roland, T. and P.K. Kannan (2003). E-service: A new paradigm for business in the electronic environment. *Communications of the ACM*, vol. 46, no. 6, pp. 37-42.
- Ru, Y. and E. Horowitz (2007). Automated classification of HTML forms on e-commerce Web sites. *Online Information Review*, vol. 31, no. 4, pp. 451-466.
- Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, vol. 24, no. 5, pp. 513-523.
- Santos, E., H. Nguyen, Q. Zhao, and H. Wang (2003). User modeling for intent prediction in information analysis. *Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomics Society*, pp. 1034-1038.
- Scott, M.L. (2005). *Dewey Decimal Classification: A Study Manual and Number Building Guide*. Westport, C.T.: Libraries unlimited.
- Shen, X., B. Tan, and C. Zhai (2005). Implicit user modeling for personalized search. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, pp. 824-831.

- Shi, X. and C. Yang (2007). Mining related queries from Web search engine query logs using an improved association rule mining model. *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1871-1883.
- Shuen, A.A. (2008). *Web 2.0: A Strategy Guide: Business Thinking and Strategies Behind Successful Web 2.0 Implementations*. Cambridge: O'Reilly Media.
- Smyth, B. (2007). A community-based approach to personalizing Web search. *Computer*, vol. 40, no. 8, pp. 42-50.
- Speretta, M. and S. Gauch (2005). Personalized search based on user search histories. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 622-628.
- Stamou, S. and A. Ntoulas (2009). Search personalization through query and page topical analysis. *User Modeling and User-Adapted Interaction*, vol. 19, no. 1-2, pp. 5-33.
- Statistics Canada (2010). Canadian Internet use survey. Available online: <http://www.statcan.gc.ca/daily-quotidien/100510/dq100510a-eng.htm>. (Accessed in March, 2011.)
- Su, J., B. Wang, and V.S. Tseng (2008). Effective ranking and recommendation on Web page retrieval by integrating association mining and PageRank. *Proceedings of the 2008 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia, vol. 3, pp. 455-458.
- Sugiyama, K., K. Hatano, M. Yoshikawa, and S. Uemura (2004). Adaptive Web search based on user profile constructed without any effort from users. *Proceedings of the 13th International Conference on World Wide Web*, New York, USA, pp. 675-684.

- Tan, B., X. Shen, and C. Zhai (2006). Mining long-term search history to improve search accuracy. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, USA, pp. 718-723.
- Teevan, J., S.T. Dumais, and E. Horvitz (2010). Potential for personalization. *ACM Transactions on Computer-Human Interaction*, vol. 17, no. 1, 4:1-4:31.
- Theng, Y.L., E. Duncker, and N. Mohd-Nasir (1999). Design guidelines and user-centred digital libraries. *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pp. 167-183.
- Trestian, I., Ranjan, S., Kuzmanovic, A., & Nucci, A. (2010). Googling the Internet: Profiling Internet endpoints via the World Wide Web. *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, pp. 666-679.
- Tyler, S.K. and J. Teevan (2010). Large scale query log analysis of re-finding. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, New York, USA, pp. 191-200.
- Van Den Heuvel, W.-J. and M.P. Papazoglou (2010). Toward business transaction management in smart service networks. *IEEE Internet Computing*, vol. 14, no. 4, pp. 71-75.
- Varadarajan, R. and M.S. Yadav (2009). Marketing strategy in an Internet-enabled enviroment: A retrospective on the first ten years of JIM and a prospective on the next ten years. *Journal of Interactive Marketing*, vol. 23, no. 1, pp. 11-22.
- Von Borstel, F.D. and J.L. Gordillo (2010). Model-based development of virtual laboratories for robotics over the Internet. *IEEE Transactions on Systems, Man, and*

- Cybernetics, Part A*, vol. 40, no. 3, pp. 623-634.
- Wang, F. and H. Shao (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications*, vol. 27, no. 3, pp. 365-377.
- Wang, T.Y. and H.M. Chiang (2007). Fuzzy support vector machine for multi-class text categorization. *Information Processing & Management*, vol. 43, no. 4, pp. 914-929.
- Wang, Y. and F. Makedon (2004). Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, Stanford, California, pp. 497-498.
- Wei, Y., L. Moreau, and N.R. Jennings (2005). A market-based approach to recommender systems. *ACM Transactions on Information Systems*, vol. 23, no. 3, pp. 227-266.
- Wen, H., L. Fang, and L. Guan (2008a). Automatic Web page classification using various features. *Proceedings of the 9th Pacific Rim Conference on Multimedia*, Tainan, Taiwan, pp. 368-376.
- Wen, H., L. Fang, and L. Guan (2008b). Modelling an individual's Web search interests by utilizing navigational data. *Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing*, Cairns, Australia, pp. 691-695.
- Wen, H., L. Fang, and L. Guan, (2009). A multi-agent based automatic Web recommendation model. *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, Baoding, Hebei, China, pp. 1482-1487.

- Wen, H., L. Fang, and L. Guan (2010). Classifying customers using navigational history for developing personalized Web services. *Proceedings of the 7th International Conference on Service Systems and Service Management*, Tokyo, Japan, pp. 1-6.
- Weng, S., B. Lin, and W. Chen (2009). Using contextual information and multidimensional approach for recommendation. *Expert Systems with Applications*, vol. 36, no. 2, pp. 1268-1279.
- Wu, D., I. Im, M. Tremaine, K. Instone, and M. Turoff (2003). A framework for classifying personalization scheme used on e-commerce websites. *Proceedings of the 36th Hawaii International Conference on System Sciences*, Big Island, Hawaii, USA, 12 pages.
- Xue, G., J. Han, Y. Yu, and Q. Yang (2009). User language model for collaborative personalized search. *ACM Transactions on Information Systems*, vol. 27, no. 2, 11:1-11:28.
- Yang, H.C and C.H. Lee (2004). A text mining approach on automatic generation of Web directories and hierarchies. *Expert Systems with Applications*, vol. 24, no. 4, pp. 645-663.
- Yong, S.K., B.J. Yum, J. Song, and M.K. Su (2005). Development of a recommendation system based on navigational and behavioral patterns of customers in e-commerce sites. *Expert Systems with Applications*, vol. 28, no. 2, pp. 381-393.
- Yu, B. and Z.B. Xu (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems*, vol. 21, no. 4, pp. 355-362.

- Yu, K., A. Schwaighofer, V. Tresp, X. Xu, and H.P. Kriegel (2004). Probabilistic memory-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 56-69.
- Zhu, T., R. Greiner, and G. Haubl (2003). Learning a model of a Web user's interests. *Proceedings of User Modeling*, Johnstown, PA, pp. 22-26.
- Zigoris, P. and Y. Zhang (2006). Bayesian adaptive user profiling with explicit & implicit feedback. *Proceedings of the 15th ACM Internatioanl Conference on Information and Knowledge Management*, Arlingto, VA, pp. 397-404.
- Zo, H. and K. Ramamurthy (2009). Consumer selection of e-commerce Websites in a B2C environment: A discrete decision choice model. *IEEE Transactions on System, Man and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 4, pp. 819-839.