# WEB PERSONALIZATION USING IMPLICIT INPUT

by

PAN YIH JYH

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

April 2007

## ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to Assoc Prof Dr Tang Enya Kong, my main supervisor, Dr Bali Ranaivo and Dr Chuah Choy Kim, my co-supervisors, for guiding me throughout this research. I really appreciate their patience in helping me in completing this research. It was an inspiring trip having their guidance and teachings throughout this research.

I am grateful to the School of Computer Sciences for offering me the opportunity to work as a Graduate Assistant in the school and funding me for the final two years of my research.

Furthermore, I would love to express my utmost thanks to my friends, Ho Nean Chan and his family, Chan Siew Lin, and Ye Hong Hoe for their concern and assistance that make this thesis possible. I would also like to thank all my other friends who were willing to share their opinion with me and giving me their most sincere advice throughout my journey in doing research.

I wish to express my greatest appreciation to my parents for their endless care and encouragement. Their love has always been my greatest motivation.

Last but not least, I would like to thank God for his infinite love that had strengthened me in times of difficulties throughout my research. This thesis would not have been possible without Him.

## TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# PEMPERSONALISASIAN WEB MELALUI INPUT TERSIRAT

## ABSTRAK

Perkembangan penggunaan Web dalam hidupan harian kami telah menyebabkan lebih banyak kajian dijalankan ke atas konsep *personalisasi*. Kajian-kajian yang dijalankan ke atas konsep tersebut kebanyakannya meggunakan input tersurat, contohnya penilaian pengguna ke atas barangan tertentu, untuk mengetahui kesukaan dan ketidak-sukaan pengguna. Namun, cara sebegini untuk mengumpul maklumat dari pengguna adalah amat membebankan. Oleh sebab itu, pengguna biasanya menjauhkan diri daripada mengemaskini maklumat mereka. Ini telah menyebabkan kehendak sebenar pengguna ini tidak dapat dikesan.

Oleh itu, kajian saya cuba untuk mengatasi masalah ini dengan mengkaji penggunaan input tersirat. Kajian-kajian yangdijalankan selama ini bertumpu kepada mengetahui kesukaan pengguna menggunakan input tersirat, namun kajian untuk memahami ketidak-sukaan pengguna melalui input tersirat jarang dijalankan.

Keadaan ini telah menarik perhatian saya untuk mengkaji kemungkinan penggunaan input tersirat dalam menentukan ketidak-sukaan pengguna. Saya telah membezakan input-input tersirat saya kepada dua kategori, iaitu (a) Petunjuk Minat Positif, dan (b) Petunjuk Minat Negatif.

Satu simulasi laman web e-commerce telah dibina untuk mengumpul input-input yang saya perlukan. Melalui eksperimen ini, saya dapat membuat kesimpulan bahawa input tersirat boleh menjadi petunjuk kepada minat pengguna. Tetapi, tiada petunjuk-petunjuk yang jelas untuk menunjukkan input tersirat juga boleh digunakan untuk mengenalpasti ketidak-sukaan pengguna.

Akhirnya, saya juga mengemukakan beberapa cara untuk menggunakan input tersirat bagi menjangka tahap minat pengguna yang mungkin berguna kepada e-bisnes untuk memahami pengguna tanpa bergantung kepada input tersurat.

# WEB PERSONALIZATION USING IMPLICIT INPUT

## ABSTRACT

The growing importance of the World Wide Web in our lives has intensified the studies on personalization. These studies on personalization generally make use of explicit information, e.g. rating an item to know the interests or disinterests of users. However, this method of obtaining information is intrusive on the users. As a result, users often shy away from updating their likes and dislikes. Consequently, their latest interests are not known.

Hence, my work seeks to look for an alternative way to obtain input from users in a less obtrusive manner, namely implicit input. From my studies, majority of the researches of the use of implicit input are focusing on capturing positive interests of users. However, the disinterests of users are often neglected.

This intrigues me to find out the possibility of using implicit input to capture the disinterests of users as well. Hence, I categorize my selection of implicit input into two groups: (a) positive interest indicators, viz. view, book-mark, add-to-cart, and purchase, and (b) negative interest indicators, viz. skip, delete book-mark, delete from cart.

A simulated online shopping mall is used in my work to observe and gather information from my users. I am able to come to a conclusion that implicit input is indicative of user interests, but there is no clear support to show that implicit input can be suitably used to reflect the disinterests of users.

In the final part of my methodology, I adopted a few strategies of inferring feedback ratings from implicit input, that I embrace could be applied to replace explicit user ratings. As a result, I demonstrate that the list of implicit input studied in my work can be used to generate tangible output, which in turn can be helpful in predicting user interests.

# CHAPTER

# 1

# INTRODUCTION

In the streets and over the media, countless of new products are being advertised everyday, either in the form of tangible products, or services provided. Reaching the potential customers has never been more difficult. Businesses understand that the one size fits all approach has became obsolete. Hence, various business strategies have been developed to resolve this problem, one of which is to exploit the very nature of every consumer's personal needs. The reason for this is simple, to attract more customers, and to enrich them. For instance, there is the sale of vegetarian food in some non-vegetarian based fast food restaurant. Also, a diverse range of perfume products are being produced, in which some are specifically designed for athletes, whilst some others for office workers. Obviously, these products are mostly tailor-made, which is intended to cater for different personal needs. Such effort is usually known as *personalization*.

In like manner, business strategies adopted in the virtual world display a great similarity with those used in the real world. Very often we may come across various personalized contents on the Internet. As an example, we have personalized news in News Feeds in Findory.com[1], personalized recommendations of books in Amazon.com[2], and personalized recommendations of products in Mysimon.com[3]. There is no doubt that

[1] http://www.findory.com
[2] http://www.amazon.com
[3] http://www.mysimon.com

web personalization has attracted more visits from users to the Internet, and it helps to guide users down to the path of their interests more effectively.

Prior to the use of personalization, information is presented to users based on the assumption that the given information is of interest to them. Such approaches had worked in the beginning, but are no longer effective now. This is led by the increasing numbers of users, follow by the diversity of their needs, and the vast amount of information available online. Often, this approach ends up bombarding the users with too much irrelevant information. It makes the process of locating a desired piece of information on the Internet troublesome and tedious. Not surprisingly, this is an undesirable condition, as it discourages users from making any returning visit. It is therefore important that every user is recognized to be different and unique, and the information delivered to them be tailored to their needs. This is what personalization endeavours to achieve.

In general, the process of obtaining personalized information involves a few steps (see Figure 1.1). Although researchers have given their own definitions to the term itself, the definition given by IBM explains it best.

> "Personalization is a process of gathering and storing information about site visitors, analyzing the information, and based on the analysis, delivering right information to each user at the right time"
>
> (IBM High-Volume Web Site Team, 2001)

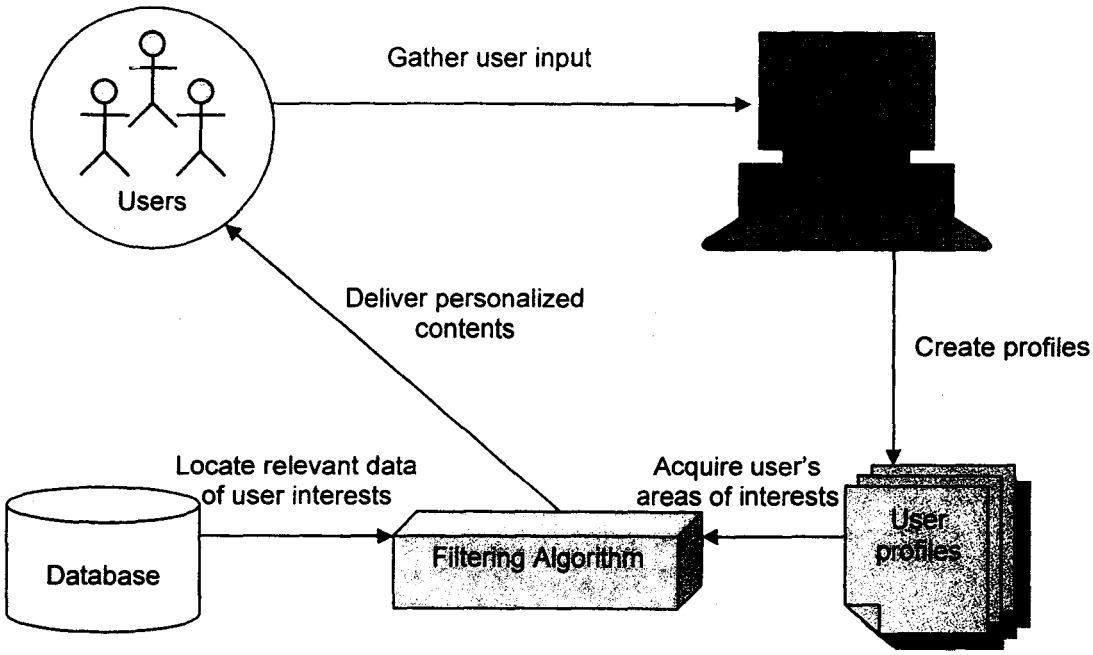


Figure 1.1: General architecture of web personalization

In this thesis, we focus on the process of gathering user input. Current personalization efforts depend mostly on feedbacks provided by users. Information obtained with such approach is known as *explicit input*. Explicit input plays a much passive role in the process. One typical example is to request users to give rating based on the degree of interests they have towards a particular product. One major problem with this approach is that web sites which personalize with such method usually find themselves having information that is already obsolete about the users. This is understandable since most users are reluctant to provide feedbacks as hoped for by Web sites owners, which led to an infrequent updates of user information.

Realizing the limitations of using explicit input, researchers have sought to look at alternative ways, e.g. using *implicit input*. Implicit input refers to information gathered in an unobtrusive manner. For instance, the time a reader takes to go through an article, or the browsing sequence or browsing habits of a given user. Such an approach can benefit any personalization system as every interaction of the users can be utilized to better understand the users' needs (Nichols, 1997). Furthermore, the use of implicit input removes the burden of explicitly providing feedbacks. By this approach, personalized content can be delivered to the users at virtually no cost. Consequently, effort of personalization will be considered as a value added service, which is an important element in attracting returning visits from users. These repeating visits can be particularly valuable for businesses, as it fertilizes the growth in sales.

Despite having various benefits, the use of implicit input has yet to be widely adopted in most current personalization systems. Knowing the potential benefits as mentioned above, it motivates us to have an in depth study of the use of implicit input to better understand user interests.

In this thesis, we report our experiment for studying the credibility of implicit input in determining user interests. Our work concentrates on seven types of implicit input, which we believe can be widely applied in the future.

An e-commerce web site is constructed for this purpose, viz. to collect input from users in a non-intrusive way. At the end of the experiments, we were able to arrive at several interesting results based on the input we gathered. We also deploy several strategies for applying our findings in the implementation of personalization on the Web.

This thesis is organized in a step by step manner, exploring the possible relationships between implicit input and user interests. Beginning with Chapter 2, a general background of studies related to personalization is presented. A detailed discussion of processes involved in personalization, and the use of implicit input are included in Chapter

3. In Chapter 4, we describe our experiment in gathering our desired implicit input. The results from the analysis based on the collected data are reported in Chapter 5. Last but not least, Chapter 6 summarizes the output of this thesis, and some possible future work is also included in the discussions in that chapter.

# CHAPTER

# 2

# WEB PERSONALIZATION: AN OVERVIEW

In this chapter, we present an overview of the concept of Web personalization. We will investigate the origin of the notion of personalization, and give it a working definition that will be adopted throughout this thesis. We will also look at several applications of personalization on the Web, and complete the chapter by summarizing the applicability of personalization.

## 2.1 Background

Before World Wide Web (Web) was first introduced back in the 1990's, the Web was first started as a networked information project. The rationale of the project is simple – to provide a convenient way for people to roam, to browse, and to contribute their information over the digital world. Ultimately, the Web has met its objective over the years, and users around the world can now share their information freely over the Web.

Although the freedom of sharing information may on one hand meet the different needs of various users, it does not guarantee the quality of the information that the users received. Too often when attempting to do a simple search, users will find themselves ending up with results which are usually not up to expectation, if not disappointing. It is not until recently that such problems arise, but in fact it happened far when the Web was

introduced. For this reason, it is only sensible that the idea of adding *personalization* to the Web is introduced (Brazile, 2004).

The notion of personalization is not new, while there are various definitions to the term "personalization". If we follow the definition given by IBM High-Volume Web Site Team (p. 2), one could see personalization more as a user-centric process. On the other hand, personalization has been defined as the effort of "delivering to a group of individuals, relevant information that is retrieved, transformed, and/or deduced from information sources" (Kim, 2002). By incorporating the essence of both definitions mentioned above, we have our working definition on personalization as:

> any effort to learn about user needs/interests, and to deliver relevant information tailored to the needs of the user, in group or individually.

The learning of user needs includes any method of getting to know about users, whereas the delivering of relevant information means looking for information that may be of interest to users, and sending it to them. Hence, in other words, any system that spends the effort in learning about its users, and seeks to deliver information based on its users needs is considered as a personalization system.

## 2.2 Why Personalization?

There is no limit to individuals in sharing their information over the Web. Hence, it is not surprising that users who use the Web will face the vast numbers and varieties of information available online. The number of documents on the Web had increased from just millions to billions over the past few years[4].

[4] http://online.sfsu.edu/~fielden/hist.htm

Furthermore, this is accompanied by the increase in numbers of users who access the World Wide Web as well. In Malaysia alone, a total of 5.7 million Malaysian users accessed the Web in the year 2000[5]. Apparently, we can expect the number to have increased further by the time this thesis is written.

Users will be all at sea if they are left to browse the Web on their own. In such a situation, it has become a burden for users to reach information that they desire in an efficient manner, and it is hard for the content providers to reach their desired audience as well. As a result, the true benefits of information sharing on the Web cannot be reaped.

Hence, personalization comes under the limelight. The main objective of personalization is to deliver information tailored to user needs. From users' perspective, their major concern is to find their desire information within the shortest time possible. A common practice to achieve this is to skim through non-related information. However, a considerable amount of time is normally wasted in this process. Since personalization systems are required to have at least a brief understanding of the users, it is possible to have systems acting on behalf of the users to do the skimming job that save a lot of the users time, and allowing them to concentrate more on their given tasks.

Secondly, making use of personalized contents can be beneficial to content providers as well. On one hand, users' satisfaction can be improved when their needs are met. On the other hand, content providers can also strategize their marketing directions according to what they had learnt about their users through the personalization systems. Generally, both these benefits are not exclusive of each other, and both are essential for the growth of any Web sites or for the increase of revenue for e-commerce Web sites.

[5] http://www.nua.com/surveys/how_many_online/index.html

## 2.3 Customization versus Personalization

The discussion of personalization brings us to another important issue which is: Are both customization and personalization the same thing?

In the context of Web, both terms can be separated not only semantically, but also can be separated from the locus of control for each of them. In most occasions when we visit the "My" version of some well known portals (e.g. MyYahoo, MyMSN, etc.), we will notice that we are allowed to change the layout, or the information that we are more interested (e.g. stocks information, weather forecast, featured news, etc.). Any changes that we made will be instantly recognized by the site. The site would then retain the same layout or same set of information every time we visit that particular website. When this happens, we as the user, have full control over the interaction. We will notice that the scenario itself does not suggest any learning by the systems is involved. This contradicts with our earlier definition on personalization, that there should be learning of user interests involves in a personalization system. Hence, these "My" versions of portals are merely efforts of customization, and not personalization.

## 2.4 Applications of Personalization

### 2.4.1 Recommender Systems

One of the interesting applications of personalization on Web is the recommender systems. In this section, albeit not every existing recommender systems will be included into our discussion, several significant works and systems will be discussed. Our main

purpose for this section is to study how these recommender systems apply the concept of personalization, and where they are commonly used.

To get things started, we have first to understand what a recommender system can perform. Recommender systems differ from other applications through the way it provides personalized content to the users. A recommender system basically delivers personalized content to users in the form of suggestions. A typical example is in which recommender systems can retrieve a list of interesting links that are relevant to a given user, and posting it on the user's first page or the list of links can appear in a separate window in order to attract the attention of that user.

Hence, a given recommender system basically performs several tasks which include:

- learning of user interests
- storing the information learned from users
- filtering relevant contents based on stored information
- suggesting users of the filtered contents

Examples of recommendation systems are commonly found in e-commerce sites, e.g. Amazon.com[6], CDNOW[7], and Moviefinder.com[8] etc.

Schafer et al. (1999) summarized these systems based on their taxonomy of techniques for recommendation, and detailed several important features of recommender systems, which include interfaces used, and the process of finding appropriate recommendations.

Recommender systems are also used in news reading domain as well. WebMate (Chen and Sycara, 1998), Alipes (Widyantoro et al., 1999), and Personal View Agent

[6] www.amazon.com
[7] www.cdnow.com

(PVA) (Chen et al., 2001) are among some systems which suggest interesting news to readers. Although it is not explicitly mentioned of the use of recommender systems in their work, the processes involved in these works do correspond to what a recommender system normally does, where these systems learn about user interests, and based on the knowledge they have, making recommendations to their users of interesting information.

### 2.4.2 Adaptive Hypermedia Systems

Another interesting application of personalization on the Web is the adaptive hypermedia systems. Initially, most adaptive hypermedia systems are commonly used as non web-based systems. However, since the World Wide Web started to grow rapidly after the mid-1990s, many research works had been carried out to put adaptive hypermedia to use in the Web.

According to Brusilovsky (1996), there are three key aspects that we can look for in any adaptive hypermedia systems. First, it has to be a hypermedia system. By hypermedia system, it means any system that allows user to retrieve information of type texts, videos, audios, photographs, or computer graphics for a particular subject. Second, there should be a user model for an adaptive hypermedia system. The user model is usually used for storing data gathered from users. Finally, as an adaptive hypermedia system, the system should be able to utilize user models learned from users, so as to annotate the visible aspects of the system to suit the users.

Hence, in other words, adaptive hypermedia systems can be seen as any hypermedia system that applies the concept of personalization. To have clearer idea of what adaptive hypermedia systems really do, let us take a look at some example systems.

---

[8] www.moviefinder.com

*Anatom-Tutor* (Beaumont, 1994) is a tutoring system for teaching brain anatomy in university. The system has a component that is used for receiving information about its users, in which the collected information is then processed and saved into user models. The user models then allow the system to have an idea of the level of knowledge of its users. This allows Anatom-Tutor to tailor its interface to its users, either by annotating the hypertexts displayed, or by hiding/disposing irrelevant links that are not suitable to the level of knowledge of its users. Through the assistance provided, Anatom-Tutor is able to keep its users focused by avoiding them from getting "lost" in the large information base available.

On the other hand, WebWatcher (Joachims et al., 1995) works as a search assistant. It helps its users to retrieve relevant information over the World Wide Web. WebWatcher helps its users by modifying the page that the users browse. It does so in several ways. A menu bar may be used to both allow users to annotate their search, and/or sought to display suggestions made by the system to the users based on the goals. Also, in each page that the users are visiting, hyperlinks will be highlighted as they are anticipated to be of interests to the users. This is done in WebWatcher by adding a small icon in the shape of an eye. The sizes of the icon represent the confidence level from the system in predicting the relevance of a given hyperlink to the particular user.

Based on the examples, we can clearly see that the use of the notion of personalization in adaptive hypermedia has made things much easier for those who are using the hypermedia systems.

## 2.5 Summary

In this chapter, we have looked at some fundamental aspects of personalization. We have introduced a working definition for the term personalization, and based on it, differentiating the works on personalization from the others. Also, we have introduced several useful applications of personalization that are being used on the Web.

From the discussions, we have seen several advantages of personalization, and the applicability of the concept itself on the Web through two major applications, namely recommender systems and adaptive hypermedia systems. Despite the widely used of the notion of personalization, there remains a certain degree of disagreement as well. In McGovern's (2003) discussion, it is well understood that a simple, well-designed navigation would very much of help to users in locating relevant information, compared with the use of personalization. We do agree that well-structured content is important However, the benefits of personalization should not be written off. As we had argued n earlier in our discussions, we have seen how the use of personalization is able to help users, specifically in a large information space. Nevertheless, there still remain areas of improvement for current personalization approaches, particularly in reducing the cost for obtaining user information.

With that in mind, this thesis focuses more on finding solutions for problems exist in both applications of personalization discussed earlier. We are more concern about the modelling of user interests, as is done in most recommender systems, compared to other characteristics of the users, such as demographics information, level of knowledge, etc., which are typical information for user modelling in most adaptive hypermedia systems. However, this difference does not make both recommender systems and adaptive hypermedia systems mutually exclusive, as they still do share some problems together.

Thus, we anticipate that our work will not only benefit recommender systems, but will also be useful in nurturing adaptive hypermedia systems as well.

In our next chapter, we shall study in detail the processes involved in personalization, and to discuss the core of our work – implicit input.

# CHAPTER

# 3

# WEB PERSONALIZATION: AN INSIGHT

In this chapter, we will look at three major processes involved in personalization, namely *user profiling*, *information filtering*, and *data collection*. Readers are expected to have an insight into several problems that we had identified concerning current personalization approaches upon finishing this chapter. We will cover the main discussion of our work – data collection – in the last section of this chapter. And we will clarify the direction of our work in improving current approaches.

## 3.1 User Profiling

Having a thorough grasp of users' needs is essential for personalization. The *user profiling* process serves this purpose by gathering information about the users. This section will focus more on the output of this process – *user profiles*.

User profiles are a collection of information used to describe a particular user (Adomavicius and Tuzhilin, 1999). This information plays an important role in any personalization system, as it is the only element in personalization that recognizes the differences between users.

Basically, user profiles tell us about who the user is, what he likes (or dislikes), and what his level of knowledge. There are various kinds of information that user profiles can tell. However, the part that intrigues us is the information of the interests of each user.

Since user profiles are used to describe users, it is not surprising that input from users is vital to its construction. Therefore, we will cover two major options for gathering user input, namely *explicit input* and *implicit input* in the last section of this chapter.

### 3.1.1 Static Data and Dynamic Data

In each user profile, data kept can be divided into two main groups: *static* or *dynamic*. Static data, as the name implies, are data that seldom change, and are usually provided by the users themselves, i.e. age, gender, or address. These data basically depict some facts about the given user, e.g. "user X is a male" or "user Y likes to drink beer". Dynamic data on the other hand tells us more about the behaviour of the given user. Such data are usually the product of some analysis process performed on raw data collected from user browsing traits. As an example, dynamic data depicts information such as "when making purchases of more than RM100, user X usually pays with credit card", or "user Y usually buys beers and peanuts together". Both static and dynamic data are widely-used. However, the latter plays a much more dominant role in most user profile representations, which we will see in later sections in this chapter.

### 3.1.2 Representation of User Profiles

There are various kinds of ways to represent a user profile. One can place a set of rules in the profiles, as is done by Adomavicius and Tuzhilin (ibid). Or, as Mobasher et al. (2000) believes that for flexibility sake, it is more reasonable to represent user profiles using pair-wise entries which consist of URLs viewed by users, and a respective weight for each URL that depicts the significance of the URL. On the other hand, both WebMate (Chen and Sycara, 1998) and Alipes (Widyantoro et al., 1999) used keyword vectors for the user profiles representation in their systems. Keyword vectors representation is in

some way similar to Mobasher's pair wise representation. For each keyword, there is also an associated weight value that tells the importance of the particular keyword. This approach is useful in facilitating the process of keeping user profiles up-to-date. Another approach used by both Prestchner et al. (1999) and Chen et al. (2001) in constructing user profiles for their system is the hierarchical form representation. Adapting this representation, categories of information are grouped into a hierarchical structure, which are then used to construct the profiles. Hence, each profile consists of the whole hierarchical structure, and for each category in the hierarchy, a weight value will be attached to it, where the function of the weight is similar to those in keyword vectors representation.

Each way of representing user profiles has its own advantage over the other. However, as Chen et al. (ibid) suggested, most studies focus more on the representation, rather than on the maintenance of the profiles. There is no doubt that proper representation of user profiles is important, but that is not the only criterion that dictates the performance of a user profile. As computing capability has increased over the years, the time required for disseminating user profiles has relatively been shortened. Hence, optimizing the representation of the profiles may not necessarily provide much benefit. Therefore, having an appropriate maintenance mechanism for user profiles has become apparently a more important task to be accomplished.

This issue is proven in PVA (Chen et al., 2001), in which the system adopts the representation uses a hierarchical structure. Different from previous studies, PVA extends the work by Prestchner et al. (1999) by exploiting the characteristics of hierarchical structure, viz. the splitting and merging of nodes, to capture changes in user interests. Benchmarking itself against several other personalization systems, PVA proven that the splitting and merging of nodes in a hierarchical representation helps in identifying both short term and long term interests of the users. Nonetheless, we believe that there is still

room for improvement in the system, specifically in the process of data collection, which we shall cover in section 3.3.

It is understandable that the construction of user profiles is important in any personalization system. However, there is another important area, viz. information filtering that should be considered. Thus, we will cover the discussion of information filtering in the next section.

## 3.2 Information Filtering

Although user profiling is important, understanding the users alone can not bring tangible benefit to the users. There is a need for something that can help users to get rid of irrelevant information that is bombarding them. By incorporating information filtering fitting into the system, users are relieved of the trouble of skimming through every single piece of information in search of their desired information, as it will be taken care of automatically by the filtering process.

Among the many filtering techniques, the three that are most widely-used are: *rule-based filtering*, *content-based filtering*, and *collaborative filtering*. There may be variations of each of these filtering techniques. However, in the following few sections, we will only outline the basic concepts and usage for each of them. Later in our summary, we will point out the approach that is more suitably use for current personalization effort.

### 3.2.1 Rule-based Filtering

Rule-based filtering depends mainly on a set of predefined rules. Generally, these rules comprised of two main parts: the condition part and the action part. This approach may also therefore be known as the if-then rules filtering. As an example, if a user wants to buy a printer, then the system can suggest that he also buys a rim of printing paper. In like manner, whenever a condition is fulfilled, the corresponding action will be triggered.

The derivation of rules requires user data to be analyzed. The user data may refer to transactional data, and usage data such as web logs, demographic information of user, etc.

The process to analyze user data may differ from case to case. In some cases, simple rules are derived. Generally, simple rules can be provided by any party who has a deep understanding of their users, particularly in the purchasing habits of users. This way of forming rules is simple, direct and/or straight-forward. However, the effectiveness and coverage of these rules are always limited, and are often very subjective depending on the group of people who set the rules.

A better way to setting rules is by using data mining techniques. Data mining uses algorithms to extract promising information from user transactions. This information is then used by system owners in the implementation of rules. This latter method has a greater advantage over the previous, because of the possibility of generating a more detailed set of rules which are less biased, which conform more to user trends. Among the more promising and more widely-used data mining techniques are *Apriori* (Agrawal and Srikant, 1994) and frequent pattern growth (FP-growth) (Han et al., 2000).

Applying rule-based filtering in personalization systems is fairly straight-forward. In most cases, user profiles created are in the form of if-then rules (see Figure 3.1). Each user login will be treated as a new session, and respective actions will be taken when

approach would induce users who are interested in computer peripherals to be interested in handheld accessories as well.

Nevertheless, a much more advanced approach would be the use of weighted keyword vectors method, which is commonly applied in information retrieval (IR) community (Balabanovic, 1997). Weighted keyword vectors method is often applied to text documents in IR. Basically, for this approach, items are represented with keyword vectors with their respective weight and computations are then carried out to determine similarity among items. Figure 3. illustrates how content-based filtering interacts with both the collection of information, and the items interested by users in order to produce suggestions to the users.

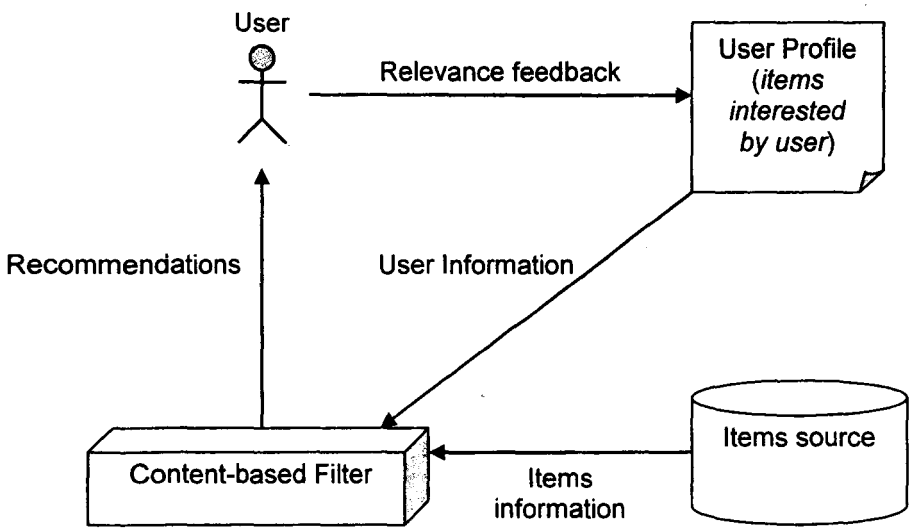


Figure 3.2: General architecture of content-based filtering

To apply content-based filtering to personalization systems, the creation and management of user profiles are very important. For this reason, user profiles are to be represented in the same format as how the items are being represented, as mentioned above. Thus, allows a comparison to be made between the collection of items and the user

profiles in finding the relevant ones. This approach is continually refined by relevance feedback provided by users, which are normally in the form of a rating value. Examples of personalization systems using this approach include *Webmate* (Chen and Sycara, 1998) and *Alipes* (Widyantoro et al., 1999).

### 3.2.3 Collaborative Filtering

The basic idea in collaborative filtering is very similar to that of content-based filtering. However, instead of taking into account similarity between items, collaborative filtering considers similarity between users. As with content-based filtering, finding the similarity between users also varies by how user profiles are being represented. In most cases, user profiles for systems using collaborative filtering are represented using weighted keyword vectors. These profiles rely heavily on user feedback in order to refine their accuracy.

With this approach, users who are perceived to be similar will be grouped together to form a *cluster*, or *neighbourhood*. Given a user, a user profile created from the user will be compared with other user profiles in order to find a suitable group for the user (see Figure 3.). Having done this, collaborative filtering will take information that is shared within the group (e.g. things purchased, news read, etc.) as a basis for making recommendation to the given user. In order to make the recommendation more accurate and comprehensive, systems can also manipulate the recommendation list by ranking them based on the popularity of each item in a particular group.

One challenging task that needs to be tackled in collaborative filtering is the clustering of users. K-means algorithm introduced by Hartigan and Wong (1979) is designated for this purpose. The said algorithm allows system administrators to cluster

users based on any attribute. K-means appears as one of the most promising and popular algorithms.

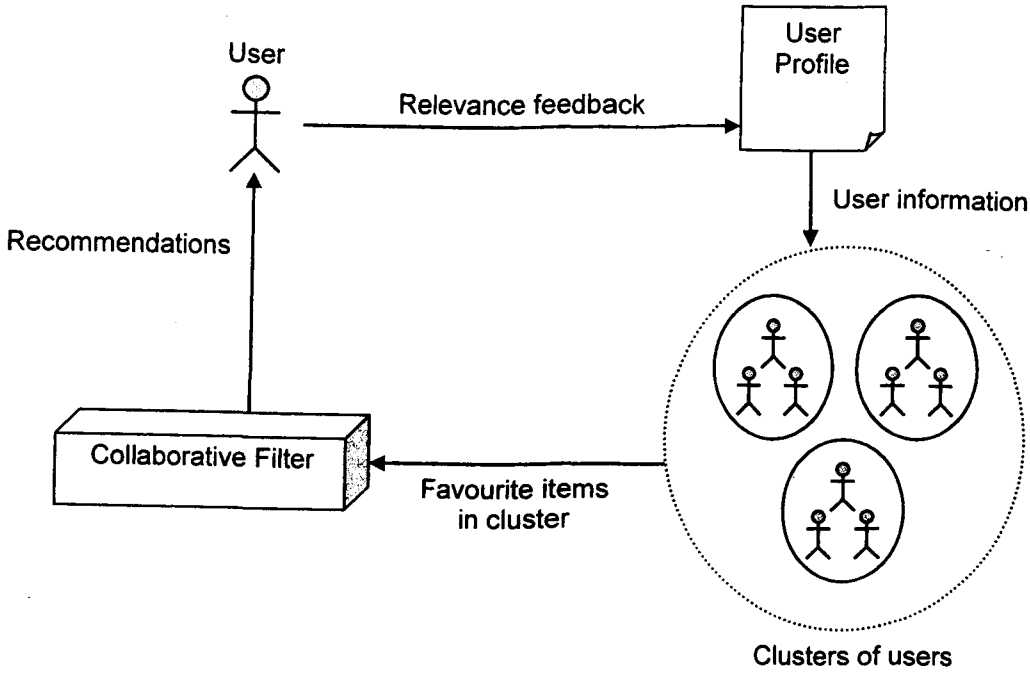


Figure 3.3: General architecture of collaborative filtering

### 3.2.4 Comparison between filtering techniques

Each of the three filtering techniques discussed above has its own strengths and weaknesses. We compare all three filtering techniques below (see Table 3.1).

Table 3.1: Comparison between rule-based filtering, content-based filtering, and collaborative filtering

| | **Rule-based Filtering** | **Content-based Filtering** | **Collaborative Filtering** |
|---|---|---|---|
| **Ease of implementation** | Simple | Difficult | Difficult |
| **Serendipitous Discovery** | None | None | Yes |
| **Adaptability to changes of user interests** | Weak | Strong | Strong |
| **Capturing of users' long term interests** | Ephemeral | Persistent | Persistent |

It is easier to implement and quicker to set up a rule-based filtering system, particularly when there is no data mining process involved. However, the resulting rules are often too common, leading to the lack of accuracy to the actual needs of the users. It is more suitable to be use as an ephemeral strategy to personalize the content, than to capture long term interests of users.

Content-based filtering demonstrates a better way to personalize Web content. It is logical that items preferred by users previously are used as the basis to predict the likeliness of their interests in other items with similar characteristics. This approach can better meet the actual needs of users, and provide recommendation with greater accuracy. However, since only items which users had shown interests are taken into account, it is almost impossible for this approach to exploit other possible areas of interests of the users,