香港城市大學
City University of Hong Kong

Department of Electronic Engineering

PROJECT REPORT

BScIT-2001/02-CS/HWC-CS/HWC-05-BSIT

Study the Use of AI - Web Site Personalization

Student Name: LUI Ho Yan
Student ID: 50195372
Supervisor: Dr CHUN, Andy H W
Assessor: Dr FENG, Jian

Bachelor of Science (Honours) in Information Technology

Contents

1. Abstract

Web Site Personalization is one of the important issues to e-Commence nowadays. A personalized e-business site is more likely to attract and maintain visitors and to build more sales. Personalized sites for employees could improve their productivity by simplifying access to information and applications. Overall customer satisfaction is increased when less time is required to locate desired information, and service is personalized to the customer's needs.

However, the common approaches of personalization engine using now requires a large amount of users active participants (typically filling out a form or following a decision-tree set of questions), which is not convenient to use and usually ignored by most of the users. Instead of explicit asking information from users, it is believed that some implicit users information could be used to perform personalization. Moreover, the users' behaviors analysis in the common personalization engine is not depth and accurate enough.

The main objective of this project is to build a more accurate users' behaviors analysis mechanism, with the help of AI rules, to perform better personalization using implicit information.

2. Objectives

In personalization engine, the major problem/challenge is how to understand a visitor, or how to know the visitors' profiles. The objective of the project is to build a better personalization engine by improving the method of "understanding" users without users active involvements.

To achieve the objective, three major tasks have to perform:

- Investigate what kind of implicit information is meaningful to understand a user
- Decide how could AI rules help to analysis the implicit information
- Decide an advanced approach to create users' profiles based on implicit information with AI rules

Lastly, using the personalization engine to build a personalized EE news web to trial the performance of the personalization engine.

3. Introducing personalization

Personalization on websites is one of major services of most of the e-business websites, two most popular examples of personalization services are amazon.com and garden.com, and there is very common to see "My" version of popular search engines and directories. Personalization has became a key factor to attract visitors in e-businese now. Actually, personalization could also find on the other areas besides websites, such as: software products (like word processing software and digital art applications), and consumer products (like vehicles and cell phones).

In websites, personalization is a process of gathering and storing information about site visitors, analyzing the information, and, based on the analysis, delivering the right information to each visitor at the right time.

Most of the websites personalization services need users' active involvement (like filling out a form or following a decision-tree set of questions) to gather users' information. Some personalization engine would gather users' information by implicit way (like recording click streams) too. The engine analysis the information to decide what may attract or being relevant to the users. Rule-based and filtering techniques are the best-known method. After the analysis, a tailor-made content could be generated to deliver to the users.

Analyzing users' profiles to deliver right content is the most complex and critical part of the personalization engine. The common simple approach is dividing users into different user groups, users in the same group having similar interests or background. The personalization engine would use same group users' behavior on the websites (like what users have clicked to read) to guess what would attract the users of the same group. Most of the personalized websites applied this method.

The aims of the company to perform such complex tasks are to make their websites easier to use for the visitors, in order to increase customer satisfaction to build a good customer relationships. And customer relationship is considered as the key of success in e-business.

4. Background Research

Three major researches have been done to perform this project. They are research of recommendations concept, different methods of performing personalization and the possible factors could reflex users' preference in browsing patterns. The researches are used for deciding the personalization engine.

4.1 Research – Recommendations concept

Doing personalization could be treated as picking up objects from a large group of objects, and the picked objects should be what attract or suitable to the objects users. Therefore, personalization is a process of making recommendations to users. Research of the concept or rules to make right recommendations is necessary to perform good personalization.

In real life, experienced sales usually could make more sales in a shop. That is because, usually, experienced sales understand the users' needs and making suitable recommendations to the users. There are three main concepts to make recommendations.

4.1.1 Recommendations based on popular choices of similar users

It is very common to see that similar users having similar tastes. Therefore, if a user could be grouped into corresponding group based on his/her background, recommendations could be made based on the popular choices of these group users.

This recommendations concept required the understanding of users' background. It is suitable for easy grouping users and trend-related objects.

4.1.2 Recommendations based on the others choices

If users are difficult to be identified or grouped, recommendations could be made by considering his/her real-time behaviors. Connecting the user's behavior to the other user's behavior to predict or guess what would attract the user. For example, in gift shop, most of the customers, who have bought Hello Kitty pencils, would

be interested into Hello Kitty pencil cases too. Therefore, recommendation of Hello Kitty pencil cases is suitable for the users buying Hello Kitty pencils.

This recommendation concept required a fairly large amount of information about general users' behaviors to link up users' different behaviors. It is suitable for new/unknown users.

4.1.2 Recommendations based on the users own preference

This concept is using of users own preference to make recommendations based on users' interests directly. Matching the users' interests with the attributes of the objects, recommend those matching objects.

This recommendation concept required understands of users' interests and the attributes of each object. It is suitable for known users.

4.2 Research - Personalization mechanism

In order to build personalization, research how to perform personalization is necessary

4.2.2 Gathering users' information

The aim of the profiling is to determine visitors' preference. There are two types of method to gather users' information; they are explicit profiling and implicit profiling.

For explicit profiling, visitors are required to provide his/her information to the system by themselves. The method of providing are usually filling forms or answering a set of questions. One example is *MyYahoo.com*, users are asked to customize the websites by themselves.

The advantage of explicit profiling is the information is usually accurate and the implementation is simple. As the information is provided by the users, it must be

true. And the system does not need any further analysis of the users where only providing some forms or questions.

The disadvantage of explicit profiling is the trouble of filling forms or asking questions. This kind of forms or questions are usually long and ignoring, since amount of information needed to understand a visitor is high, therefore, it is inconvenient for the users to do such process before using the personalization services. Moreover, when consider the privacy aspects, most of the visitors are unwilling to provide personal information to the others.

For the implicit profiling, visitors do not need any participle in the profiling process. Whole process of profiling is transparency to the visitors, the visitors browsing patterns (like what users' click, the time between each clicks) would be recorded by the system. The users' information is created by analysis of the users' browsing patterns. Amazon.com has applied this method in its' personalization.

The advantage of implicit profiling is no users' participle. The problem of ignoring or inconvenient of using is eliminated. Moreover, there is no privacy problem. The accuracy of profiling could be increased when more browsing patterns are recorded to analysis, and the visitors' profile is live and responsive to any visitors' interests change.

The disadvantage of implicit profiling is an analysis is needed to convert the users' browsing patterns to meaningful users' information. And a record program also have to be implemented to record the users' browsing patterns, it may affect the system time performance. For new visitors, using implicit profiling, he/she may not achieve a good personalization, since his/her browsing patterns are needed to analysis, which is lack to the system for new visitors.

4.2.3 Analysis of users' information to tailor content

After getting the visitors information, the system could tailor the content due to the users' information or doing recommendations to the visitors. There are two

best-known techniques used in this analysis, they are rule-based and filtering techniques.

Rule-based techniques

Rule-based techniques use a set of predefined rules to customize the contents to the users. This requires the administrator, most likely with the help of a consultant, to figure out the appropriate rules.

Filtering techniques

Filtering techniques employ algorithms to analyze meta data and drive recommendations to tailor the content. The three most common filtering techniques -- simple filtering, content-based filtering, and collaborative filtering.

Simple filtering relies on predefined groups, of visitors to determine what content is delivered to the users. Usually, the content should be popular in the corresponding visitors group. An example of simple filtering is managing access to corporate information. For example, employees identified with the Human Resources department would have personalized Web sites that give them access to information and applications specific to their job.

Content-based filtering works by analyzing the content of the objects to form a representation of the visitor's interests. Normally, the analysis needs to identify a set of key attributes for each object and then fill in the attribute values. To perform personalization, matching the users' information with the attributes to figure out which objects are matched to the users' information. For example, a users having athletic background would match sports news by this filtering.

Collaborative filtering makes recommendation based on the other users' behaviors. It is believed that a visitor behavior could be predicted by other visitors' behaviors. For example, in a newspaper websites, there are two news, news A and B. If most of the visitors would read news B after reading news A, this filtering would tailor news B to visitors that have read news A.

These three filtering techniques are performed on different environment and having different requirement, simple filtering needed grouping of similar visitors, content-based filtering needed analysis of objects contents, and collaborative filtering needed to maintain a database of users' behaviors on each objects. These three filtering techniques could operate together or alone, and each of them having its own strength and weakness.

4.2.4 Comparison of different approaches of personalization in market

The simplest approach of personalization is using rule-based techniques alone. The recommendation are done by predefined rules, there is no-real time operation in the system.

The other simple approach of personalization is using collaborative filtering alone, visitors' profile are not needed. The system only record the click streams of each visitors, and then using these information to predict visitors' interest. This approach needed high visiting rate and huge number of visitor to perform accurate personalization. And it needed real-time operation to perform collaborative filtering.

Commonly, most of the personalization engine would use explicit profiling together with simple filtering. The explicit profiling could provide visitors' background for the system to divide the visitors to different groups to apply simple filtering. This approach has to store each visitor's group and the group preferences. Real-time operation is not needed, but it has to maintain a fairly large database.

The most complicated approach of the personalization is using implicit profiling with content-based filtering techniques. Since it needed a browsing patterns record program, which have to operate in real-time mode. And the process of analyzing browsing patterns to become meaningful visitors' preference is

complicated. And each visitor and object needed a set of attributes to reflex their content/preference for perform content-based filtering. Certainly, it can perform more likely "personal" services, since each visitor having a personal preference.

4.3 Research – Factors reflex visitors' preference in browsing patterns

Since the personalization system of this project using implicit profiling (browsing patterns) to get visitors' information, therefore, it is necessary to decide what kinds of browsing patterns are meaningful to analysis, in order words, deciding what factors could reflex visitors' preference in browsing patterns.

4.3.1 What did the objects visitors read?

It is a direct reflexes of users' profile, the objects that users read could reflex the users' interests directly.

4.3.2 What did the objects visitors not read?

It could reflex what the users do not interest. It could be analysis directly by the objects content. And usually, people would read the monitor from the top to the bottom, therefore, the objects that locate upper than 'read' objects could be treated as 'uninterested' objects.

4.3.3 The sequence of reading.

This is indirect information to show the visitors' preference. Normally, people would scan all of the objects title before clicking on objects to further reading. And the priority of reading should be from the most interest to least. Therefore, the sequence of reading could be used to reflex the users' degree of interests to different objects.

4.3.3 The time spent on reading.

This is indirect information to show the visitors' preference. Sometimes, people would be 'attracted' by an object title to click to read, but after scanning the objects content, he/she may find himself/herself are not really interested to read the objects in details. In the other words, visitors would spend more time on reading objects that attract him/her. Therefore, the time of reading could reflex the users' interests. Certainly, the reading time is also depending on the length and degree of complex of the objects content too.

4.3.4 What is the users' point of view to the news?

Objects may contain more than one main theme, for example, there is news about David Beckham, Manchester United player, who might be transferred to the other club at the end of the season. In this example, a Manchester United fans may feel that this news is about the MU matters, for a Beckham fans, this news is about Beckham only. If the engine updates the Beckham fans profile to become 'interested in' MU matters, the results is not accurate. There is a serious problem for engine that using objects content to update visitors' profile. The users' point of view to the objects should be investigated to achieve more accurate analysis.

These five factors should be considered during the analysis of visitors' preference. In common analysis approach, usually, only what did visitor read would be used to analysis in implicit profiling.

4.4 Research - Conclusion

Making recommendations is the same as perform personalization on the web. The concept of making recommendations could be applied on web site personalization.

Actually, the three filtering techniques are similar to the three concepts of recommendations. Simple filtering make recommendations/personalization based on popular choices of similar users, content-based filtering make recommendations/personalization based on the users own preference, and collaborative filtering make recommendations/personalization based on the others choices. Three filtering techniques could handle three different kinds of situations.

After the research, filtering techniques (including three techniques) and implicit filtering have been chosen to build the personalization engine. Although implicit filtering is complicated and needed extra effort on recording and analyzing browsing patterns, it can prevent the inconvenient of explicit filtering and implicit filtering also can help to handle the change of visitors' preference. Since different filtering techniques having its own strength on different situation, therefore, the engine has applied all of them in order to achieve the best result. Since the website is a news posting web, there is no real-time operation, therefore, it can accept the effect of adding the two real-time operation, collaborative filtering and browsing patterns record program.

5. Methodology

5.1 Website Background – EE Info Net

The web site, EE Info Net, provide news from EE department, what are located in polylink webmail now. Therefore, the design of the web site layout is similar to the polylink webmail system and the users' browsing pattern would be collected automatically. The news is presented by the means likes polylink webmail, a news header would be shown and clicking on it would pop-up the news content. The site also has Admin. level functions, like adding news or simple users account management.

There are six different user groups defined, they are EE, CE, IT, and each program having three different years (i.e. EE year1, EE year2 ...IT year3). Login is required to access the websites.

And there is a set of key attributes for each users and news. For users, this set of attributes reflex the users' profile, for news, it reflex the content of the news. The attributes value is direct proportional to the degree of interest or relevant. The key attributes are *Arts, Music, Religion, Health, Culture, Sports, Science/Technology, Nature and Environment, News and Media, Travel, Academic, Career,* and *Scholarship*.

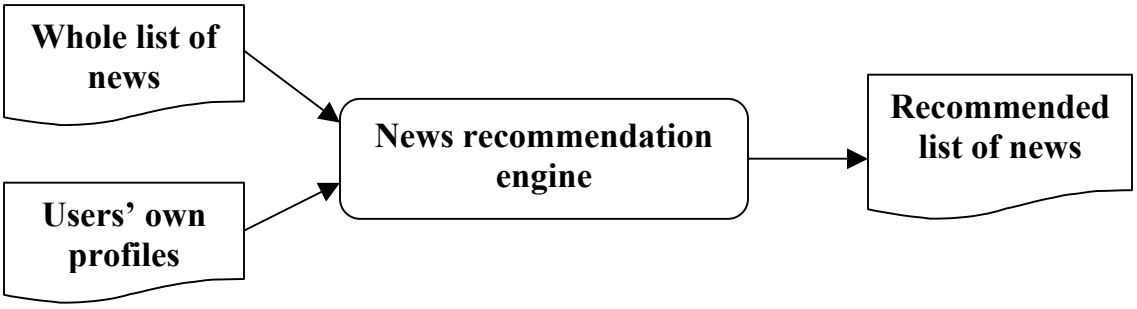
When users browsing on the websites, users' browsing patterns would be recorded by the system. Four aspects would be recorded, they are what users' read, not read, sequence of reading and time spent on reading.

The personalization engine would perform three filtering techniques at different time to perform personalization. They are simple filtering, content-based filtering, and collaborative filtering. After users' logout, the engine will perform users' profiles analysis to update users' profiles.

The personalization engine contains four main components to perform all the works of personalization, they are **news recommendation engine**, **real-time recommendation engine**, **browsing patterns record program**, and **users' profiles decision engine**.

5.2 Details of news recommendation engine

The job of news recommendation engine is to do news recommendations to users from the whole list of news. The recommendations are based on users' own profiles.



User profiles including users' group information and his/her own value of attributes set. In the other words, users' background and users' interests are known, and used to perform recommendation/personalization. Due to these two information are known, the news recommendation engine would use simple filtering and content-based filtering techniques to make recommendation.

Simple filtering would be performed to get a list of group popular news from whole list of news first. Then, the content-based filtering would sort the popular news list to arrange the display order of the news, from top to down, most recommend to least recommend. Because people usually read from top to down, therefore, most recommend or most likely would be read news have put on the top to convenient users.

5.2.1 Simple filtering mechanism

Simple filtering technique is used to make recommendations based on users' background (user group). It would recommend the popular news of the corresponding user group to the user. The display order of the news would be from most recommended to least recommended. Usually, the top 30 news would be chosen to display.

For example, there is a set of news.

Assume there is only five news in list, but in fact, there would be more than 50 news in the list.

	Reading rate of each user group						
	EE yr1	EE yr2	EE yr3	CE yr1	CE yr2	…..	IT yr3
News 1	15	20	2	5	3		50
News 2	33	10	2	12	13		7
News 3	4	19	8	4	6		10
News 4	10	8	40	9	35		5
News 5	11	10	7	16	10		2

Filtering results:

For users in EE yr1 group,

News order: news 2 > 1 > 5 > 4 > 3

For users in CE yr2 group,

News order: news 4 > 2 > 5 > 3 > 1

Therefore, different users in different groups would see different news.

5.2.2 Content-based filtering mechanism

This technique is used to make recommendations based on users' own interests. The user interests/favor could be achieved by users attributes set. And the news content also could be achieved by news attributes set. It would recommend those users interest and news content matched news (from the simple filtering) to the users. The display order of the news would be re-arranged by the level of recommendation.

For example, there is a set of news and users interests.

Assume there is 6 attributes only, in fact, there is 13 attributes in the system.

	Attributes level						
	A	B	C	D	E	F	G
News 1	1	2	0	0	8	0	7
News 2	0	1	1	2	0	9	0
News 3	1	7	0	1	2	0	9
News 4	9	1	0	0	0	7	0
News 5	1	0	9	1	0	0	0

Users attributes level						
A	B	C	D	E	F	G
2.06	5.12	6.94	4.32	0.72	9.74	0.86

In the news attributes level list, only the attributes that level larger than 6 would be concerned to be representative. Therefore, all of the news would have few attributes to represent its content. Then calculate the differences between the news representative attributes level to users' relative attributes to achieve the differences index.

Differences index = Abs[(users' attribute 1 – news attribute 1) + (users' attribute 2 – news attribute 2) +….]

	Differences index
News 1	Abs[(0.72 – 8) + (0.86 – 7)] = 13.42
News 2	Abs[(9.74 – 9)] = 0.74
News 3	Abs[(5.12 – 7) + (0.86 - 9)] = 10.02
News 4	Abs[(2.06 – 9) + (9.74 – 7)] = 4.2
News 5	Abs[(6.94 – 9)] = 2.06

Differences index value is direct proportional to the level of unmatched.

Therefore, results displaying list: News 2 > 5 > 4 > 3 > 1

5.2.3 Additional features

This content-based filtering would consider '***when and where***' situation to provide more advanced recommendations. In different situation, the level of interest to each areas would be changed, for example, people would be more interested in leisure news during holidays and non-office hours.

The system would consider the time when users login, if it is non-office hour, the value of leisure attributes of the users would be slightly increased. Therefore, when performing content-based filtering, leisure news would be placed in font position.

After two filtering techniques, a list of news is achieved. The lists contain all of the popular news of the users' group and the display order is tailor-made to users' own interests, from most interested to least.

The news recommendation engine would operate after user login to achieve a list of recommend news to the users based on users' own profile.

5.3 Details of browsing patterns record program

This program would audit the browsing patterns of the users to record some aspects of patterns for the analysis of users' profiles and real-time news recommendations. It is the component to perform implicit profiling, all of the record process is invisible to the users.

There are four patterns are recorded by the program:

1. **What did visitors read?**
   The program would record all of the read news of the users.

2. **What did visitors not read?**

   As the websites would display 12 news per page, all of the news that in the opened page, which haven't been clicked to read is treated as unread news. The program would record all of the 'unread' news.

3. **Sequence of reading**

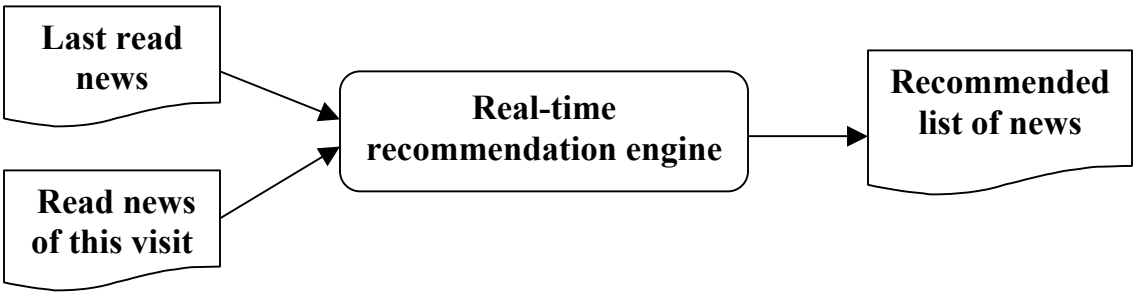
   The sequence of news reading would be collected.

4. **Reading time of news**

   The time interval from the users 'clicks the news to read' to 'close of the pop-up news' or 'open of other news' would be recorded.

5.4 Details of real-time recommendation engine

The job of this engine is to make real-time recommendations to users in response to users real-time behaviors. The recommendations are based on users' behaviors and response in real-time.



User behaviors including last read news and the list of read news of the visit. The real-time recommendation engine would use related recommendation and collaborative filtering techniques to make real-time recommendation. Related recommendation is based on the last read news, and collaborative filtering is based on all of read lists of this visit.

5.4.1 Related recommendation mechanism

This recommendation would recommend the news that having similar content comparing with the last read news.

For example, there is a list of news.

	Attributes level						
	A	B	C	D	E	F	G
News 1	1	2	0	0	8	0	7
News 2	0	1	1	2	0	9	0
News 3	1	7	0	1	2	0	9
News 4	9	1	0	0	0	7	0
News 5	1	0	9	1	0	0	0

Last read attributes level of a user						
A	B	C	D	E	F	G
9	0	2	1	2	8	0

The system would sort out the main attributes from the last read news. Attributes having value larger than 6 is consider as main attributes. In this example, attribute A and F are the main attributes.

Therefore, searching the news list, sorting out all of the news having the main attributes that is the same or partly the same as the last read news. In this example, the system would look for news have attribute A or F with value larger than 6.

Therefore, the recommend news: news 4 and 2

5.4.2 Collaborative filtering mechanism

This filtering technique would compare the users read behavior (what user read), with the other users behavior to find similar users. Using the similar users behavior to predict what would attract the user or read in next.

For example, there is a set of users behavior.

Users	**Read News**
A	1, 2, 8, 10
B	1, 2, 5, 10
C	1, 2, 8
D	3, 4, 5, 10

News that users have read: 1,10

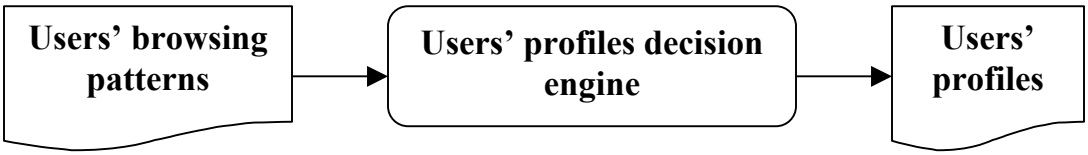
The system would sort out all of the users behavior to find users have read that news too. In this example, users have read news 1 and 10 are sorted out, they are users A and B. They are treated as similar users.

Form the behavior of users A and B, they together have read news 1, 2, 5, 8, and 10. When comparing with the example users now, news 2, 5, and 8 have not read yet. The system would predict the users would interest in news 2, 5 and 8.

Therefore, the recommend news: news 2, 5 and 8

5.5 Details of users' profiles decision engine

It is the most critical part of the personalization engine, since the job of this engine is to understand the users and most of the personalization/recommendation are based on users' profile.



Since this personalization engine only using implicit profiling, the analysis of the users' profiles are based on users' browsing patterns. The browsing patterns are come from the browsing patterns record program. There are four kinds of patterns to analysis. The engine would analysis what user interests and not interests from the browsing patterns and update the users' profiles.

5.5.1 Analysis of users interests

Three browsing patterns are used in this analysis, they are what did users read, sequence of reading and reading time.

There are three steps in this analysis:

5.5.1.1 Reading time concern engine

It would check all of the reading time of the read news to see whether there is any read news having relatively short reading time, for example 3s. If such news existing, this engine would remove this news from read news list and treat it as unread news.

The reason of concerning reading time is because users may make mistakes when choosing news to read. As users decide to read a news by reading the news header only, and the news header only can show partially content of news, therefore, it is very usual for users choosing uninterested news to read. Users usually realize the

mistake after scanning of the news content, he/she would close the pop-up news content window, or read the other news immediately.

Therefore, from the browsing patterns record program, the reading time of such 'mistake' news would be relatively. Concerning the relatively short reading time, the 'mistake' news could be removed from the read news list, as users are not really interested in it, in order to increase the level of precise.

5.5.1.2 Users point of view deciding engine

This engine is used to decide the users point of view to the read news. As news usually having multi-domain attributes, and the users may read a news by only one domain attributes of a news, therefore, we have to consider what attract the users to read the news, in other words, the users' point of view to the news to prevent the incorrect analysis of the news attributes.

The engine would analysis the content of all of 'read' news; pick up all of the news that having multiple major content and at least one of major content is users interest. Analysis those news, if there is any news having a major content attributes that the users do not 'interest', the news could be set as 'unusual' news, and marking the users' uninterested contents attributes contain in the 'unusual' news. The marked content attributes would set to be 'unusual' attributes. Next, search those 'unusual' attributes in the users' read news of this visit, if there is any news having this content attributes as major content, this 'unusual' attributes would be removed from the list, since this content attribute may be users' new favorite only. After the search, the all of the 'unusual attributes' level in the 'unusual' news list would be reduced in order to eliminate the affect in users' profile update. We could get a read news results list, that having considered users' Point Of View.

Example of the operation

Users Profile

Attributes level						
A	B	C	D	E	F	G
2	8	5	2	1	1	9

Read News

	Attributes level						
	A	B	C	D	E	F	G
News 1	0	0	0	0	8	0	7
News 2	8	0	0	0	0	0	0
News 3	7	0	0	0	0	0	9
News 4	0	0	9	0	0	0	0
News 5	0	0	8	8	0	0	0

From the users profile, this user are interested in attributes B and G. And there is a list of read news.

Concerning each news, news 1 domain attributes are E and G, news 2 domain attributes are A, news 3 domain attributes are A and G, news 4 domain attribute is C, and news 5 domain attributes are C and D.

Since user are interested in attributes B and G, ‘unusual’ news is news having multiple domain attributes and at least one of attributes is user interest. Therefore, news 1, 2 and 3 are all ‘unusual’ news. And attributes A, E and G are ‘unusual’ attributes.

When investigate the ‘unusual’ attributes, there are 2 news having attribute A as domain attributes, therefore, attributes could be users’ new interests. Therefore, attribute A and news 3 and 2 are no longer ‘unusual’.

It remains one ‘unusual’ attribute, E. There is only one read news having attribute E as domain, news 1, therefore, the user is believed that reading news 1 because of attributes G, not E. The update value of attribute E of news 1 is reduced by 20%.

Results update Value

	Attributes level						
	A	B	C	D	E	F	G
News 1	0	0	0	0	6.4	0	7
News 2	8	0	0	0	0	0	0
News 3	7	0	0	0	0	0	9
News 4	0	0	9	0	0	0	0
News 5	0	0	8	8	0	0	0

5.5.1.3 Reading sequence concern engine

This engine is used to concern the reading sequence effect. As people usually reading news from most interested to least interested, therefore, the sequence of reading could reflex the different degree of interest of the read news.

The engine would depend on the prior of reading to giving different ‘weight’ to the read news. As the prior of reading going forward, the ‘weight’ would decrease.

Formula of the engine

NEWS_VECTOR[] store the sequence of ‘read’ news

```
For i = 0 to i < n
    WEIGHT_ATTRIBUTE = (NEWS_VECTOR[i].ATTRIBUTE) * (n – i)
    TOTAL _ATTRIBUTE = TOTAL_ATTRIBUTE + WEIGHT_ATTRIBUTE
```

End

AVG_ATTRIBUTE = TOTAL_ATTRIBUTE / (i + (i-1) + (i-1-1)… + 1)

Remark: (n - i) is 'weight index', which is inversely proportion to the sequence of reading.

Example of operation

The 'read' news list

Attribute	A	B	C	D	E	F	G
News 1	7	2	0	0	3	0	1
News 2	0	9	2	0	1	8	0
News 3	7	1	0	0	0	8	0
News 4	0	0	8	8	0	9	2
News 5	0	0	9	0	0	9	1

Sequence of reading, NEWS_VECTOR: 5 > 4 > 3 > 2 > 1
Total number of 'read' news, n = 5

News 5 WEIGHT_ATTRIBUTE = { 0*5, 0*5, 9*5, 0*5, 0*5, 9*5, 1*5 }
News 4 WEIGHT_ATTRIBUTE = { 0*4, 0*4, 8*4, 8*4, 0*4, 9*4, 2*4 }
News 3 WEIGHT_ATTRIBUTE = { 7*3, 1*3, 0*3, 0*3, 0*4, 8*3, 0*3 }
News 2 WEIGHT_ATTRIBUTE = { 0*2, 9*2, 2*2, 0*2, 1*2, 8*2, 0*2 }
News 1 WEIGHT_ATTRIBUTE = { 7*1, 2*1, 0*1, 0*1, 3*1, 8*1, 1*1 }

TOTAL_ATTRIBUTE = { 28, 23, 81, 36, 5, 97, 14 }
AVG_ATTRIBUTE = { 28/ (5+4+3+2+1=15), 23/15, 81/15, 36/15, 5/15,
130/15,14/15 }
= { 1.9, 1.5, 5.4, 2.4, 0.3, 8.7, 0.9 }

From the calculation, the average attribute of read news is achieved. This value reflexes the users interests.

5.5.2 Analysis of users 'uninterests'

This analysis would use the unread news to get what users 'uninterest'.

5.5.2.1 The calculation of unread news:

From the unread news list, slot all of the unread news content attributes, extract the contents level that higher than 7. Reason of only calculating the >7 level is users deciding not read a news is only based on the news header, and the news header only could tell the major attributes of the news. Therefore, only major attributes are used for determine of users 'uninterest'.

Example of operation

Unread news list (showing major attributes only)

Attribute	A	B	C	D	E	F	G
News a	-	-	-	-	8	-	7
News b	-	9	-	-	-	-	-
News c	7	-	-	-	-	-	9
News d	9	-	-	8	-	-	-
News e	-	-	7	-	9	-	-

…….

* The symbol – means the level of this attribute is lower than 7, set to 0

Then, calculate the average attributes value of the news

Average attributes value

Attribute	A	B	C	D	E	F	G
Average value	7+9/5 = 3.2	.8	.4	1.6	3.4	0	3.2

That is the attribute to reflex what did the users do not interest. Level of uninterested is direct proportion to the values.

5.6 Details of users' profiles update

After the analysis of browsing patterns. There is two set of attributes values, what reflex users' 'interest' and 'uninterested'.

Normally, the 'interest' value is more likely to be close to the users' real profiles, since the 'interest' value are analysis by what users read, having users involvement. Therefore, when doing update, the weight of this aspect should be heavier than the 'uninterested' attributes.

Formula and example of operation:

INTEREST_ATTRIBUTE = { 1.9, 1.5, 5.4, 2.4, 0.3, 8.7, 0.9 }
UNINTERESTED_ATTRIBUTE = { 3.2, 1.8, 1.4, 1.6, 3.4, 0, 3.2 }
USER_EXIST_ATTRIBUTE = { 2, 5, 6, 4, 1, 8, 1 }

INTEREST_WEIGHT = 0.2
UNINTERESTED_WEIGHT = -0.1

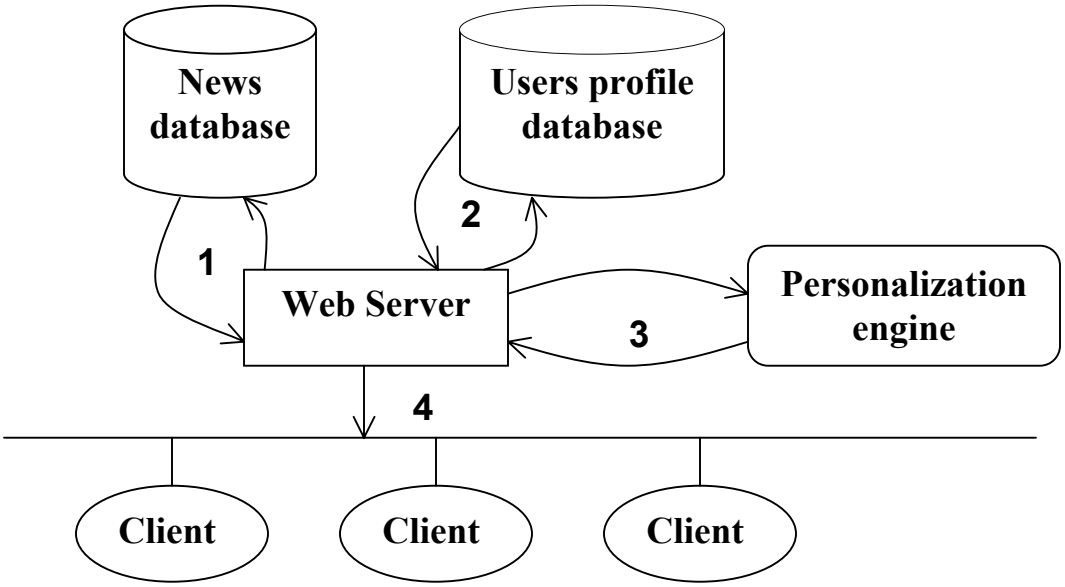
USER_NEW_ATTRIBUTE = USER_EXIST_ATTRIBUTE{}
+ 0.2 * INTEREST_ATTRIBUTE{}
- 0.1 * UNINTERESTED_ATTRIBUTE{}

= { 2, 5, 6, 4, 1, 8, 1 }
+ 0.2 * { 1.9, 1.5, 5.4, 2.4, 0.3, 8.7, 0.9 }
- 0.1 * { 3.2, 1.8, 1.4, 1.6, 3.4, 0, 3.2 }

= { 2.06, 5.12, 6.94, 4.32, 0.72, 9.74, 0.86 }

6. Implementation

6.1 System Architecture



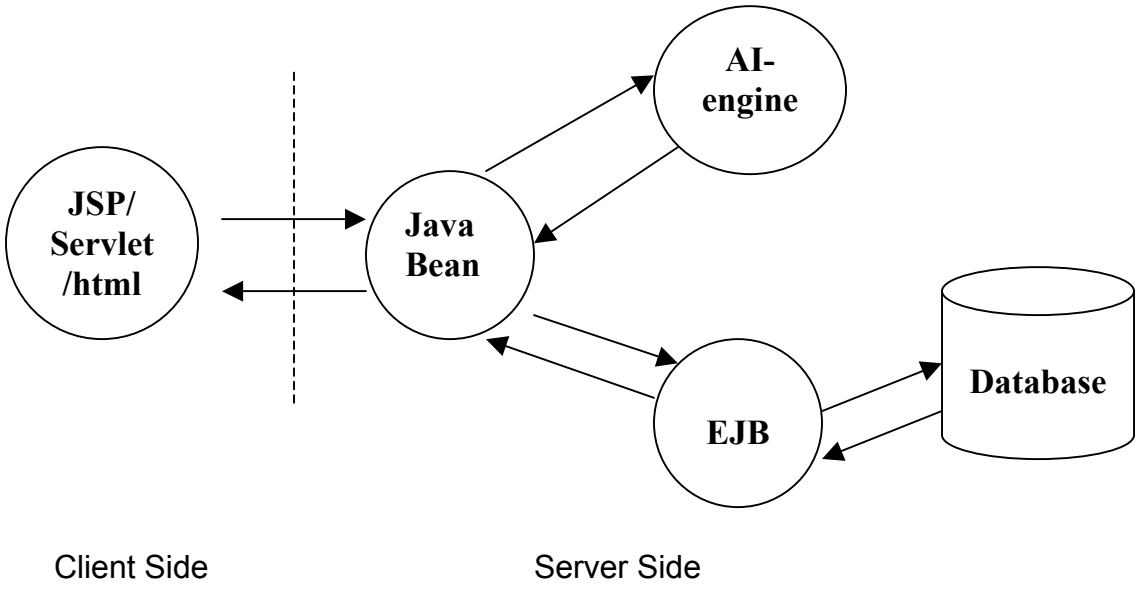
It is a simple architecture diagram of the system. When there is any client request, the web server would handle it. And all of the transmission of the data from database to client or personalization engine is via the web server.

For example, when there is a client login request, the web server would handle it by:

1. Getting all of the news from news database
2. Getting corresponding users profiles from users profile database
3. Passing all of the data to personalization engine to perform personalization service
4. Passing the personalized content to the client

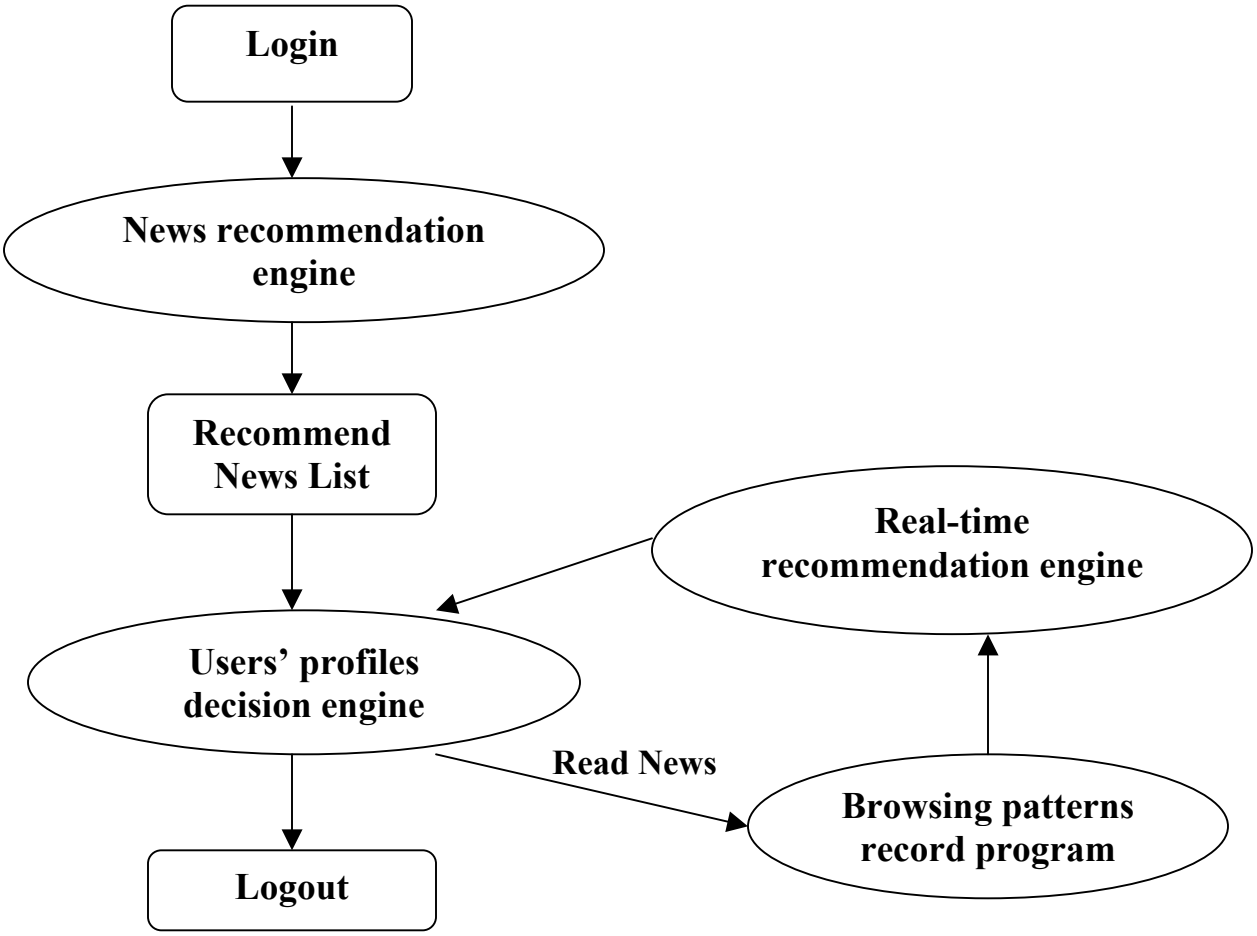
6.2 Platform and decide

As this web site is a typical client/server system, and AI-engine is needed at the server side. In order to making "thin" client, I have chosen J2EE server, which could be perform multi-tier client/server system, to be the platform. And the design consideration is free the client from any business logic. At the client side, JSP/Servlet/html is used for displaying only. At server side, JavaBean and EJB are used to perform business logic, calculation and database access.



Microsoft Access is used as the database. ODBC-JDBC bridge are used for database access. And J2EE Server is used as the Server.

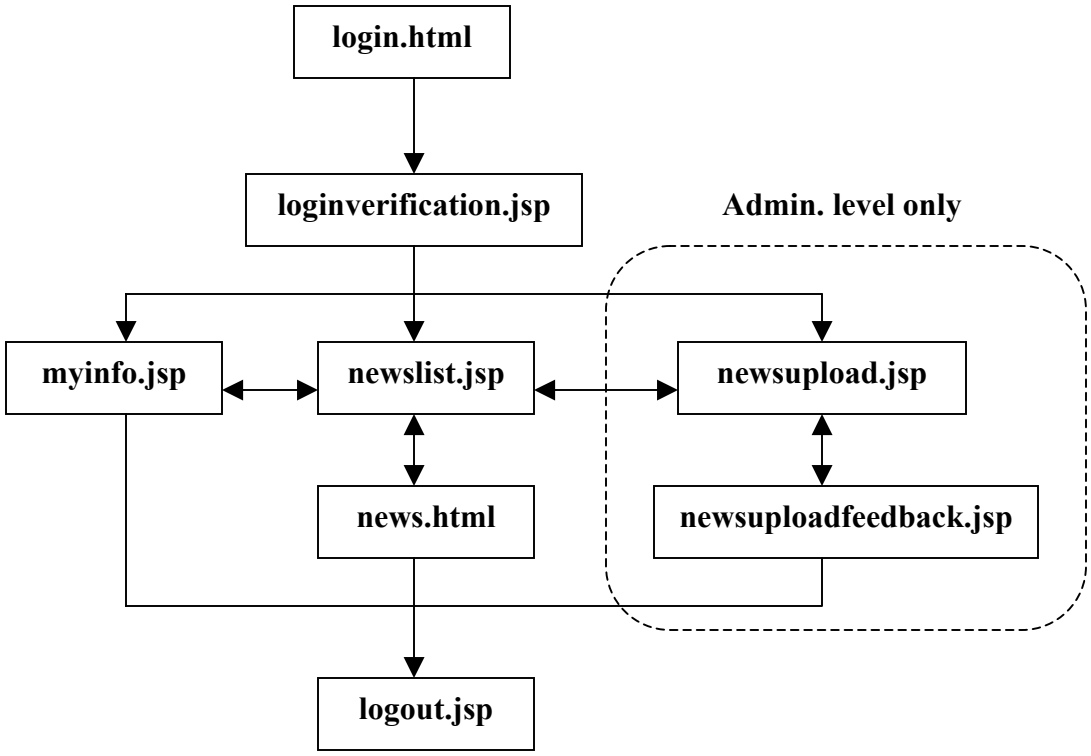
6.3 System structure



There are four main components in the system to perform personalization. They are news recommendation engine, browsing patterns record program, real-time recommendation engine and users' profiles decision engine.

News recommendation engine operates after user login. Browsing patterns record program and real-time recommendation engine operates when user read every news. Users' profiles decision engine operates when users logout.

6.4 Websites main structure



After users login, two functions are provided for normal level users, reading news and viewing his/her own profiles. These two functions are provided by *newslist.jsp* and *myinfo.jsp* respectively. For reading news, *newslist.jsp* would show the header of news, clicking on the header would pop-up news.html to show the content of the news.

For administrator level users, after login, he/she could perform the functions of uploading news to the system. The news should be in html or txt format. This function is provided by *newsupload.jsp* and *newsuploadfeedback.jsp*.

logout.jsp provide logout function for all level of the users to sign out.

6.5 Programs and its functions in the system

There are four major types of program in the system. They are JSP/Servlet/html, JavaBean, EJB and AI-engine.

JSP/Servlet/html is mainly for displaying the objects in the web page.

JavaBean handle users session tracking, data transaction between client & server, transactions logic and database access interface.

EJB handle all of the database access and some transactions logic.

AI-engine is for the personalization services.

JSP/Servlet/html:

File name	Function(s)	Describes
login.html	Asking username and password	
loginverification.jsp	Checking the correctness of login username and password	
myinfo.jsp	Showing the users own profiles	
newslist.jsp	Showing all of the news	
news.html	Pop-up page for displaying news content	
newsupload.jsp	Upload of news in html or txt format	
newsuploadfeedback.jsp	The feedback of news file upload	
logout.jsp	Sign out of users	
Logoutbywindowclose.jsp	Sign out of users	To handle the users close the browser without logout
UploadServlet.java	Upload of the news file	J2EE does not provide upload function. This component is created for file upload.

JavaBean are called or created by JSP file only. It is the interface between JSP and database, and AI-engine. JavaBean handles all of the data transaction between client and server, and ensure the data consistence.

JavaBean:

File name	Function(s)	Describes
JDBFunctionsBean.java	Validation of login username and password.	Function when user login in, called by loginverification.jsp
JUploadBean.java	Database update interface for news upload	Called by newsupload.jsp
JuserSessionBean.java	Session tracking, interface of personalization services, record of browsing patterns	Once users successful login, this JavaBean would be created in seesion scope.

EJB (Enterprise JavaBean) are located at server side, it could be called by JavaBean only. EJB handle all of the database transactions of the system by providing pre-defined functions for the JavaBeans.

EJB:

File name	Function(s)	Describes
DBFunctionsBean.java	Provide defined functions for the JavaBean to call to perform database transaction.	More than 20 functions are provided for JavaBean to handle different transactions.
DBFunctions.java	Remote interface for the JavaBean to call the functions.	Necessary in J2EE architecture.
DBFunctionsHome.java	Home interface for the internal system calls.	Necessary in J2EE architecture.

AI-engine is used for providing personalization services. A set of data passed to the engine then the engines perform personalization/recommendations based on the data. The

AI-engine could be called by the JavaBeans only, and the data are passed from the JavaBeans to the engine.

The AI-engines are implemented by rule-based AI system. And they are served for the personalization engine. Most of the complicated and large amount of data considering processes in the personalization engine would be implemented by the AI-engine.

AI-engine:

File name	Function(s)	Describes
PBengine.java PBsortrules.ilr	Provide content-based filtering technique for the news recommendation engine.	
POVengine.java POVrules.ilr	Provide point of view decision for users' profiles decision engine.	
CFengine.java	Provide collaborative filtering technique for the real-time news recommendations engine	

6.6 Database design

There are totally 10 tables in the database.

Table name	Function
tblAttribute	Store the set of attributes for users profiles and news content
tblDepartment	Store the department name and corresponding key
tblNews	Store the system information of the news, e.g. NewsID
tblNewsContent	Store the set of attributes values of each news to represent the news conten
tblNewsHitInfo	Store the hit rate of the news from different group of users
tblUsers	Store the system information of the users, e.g. UsersID
tblUserFav	Store the set of attributes values of each users to represent the users profiles
tblUserInfo	Store the general information of the users
tblUserReadNewsLog	Store all of the news each user read
tblUserUnReadNewsLog	Store all of the news each user have not read

7. Unit Testing

Test is done to trial the functions and performances of the personalization engine of EE Info Net.

Testing user information

Program: IT

Year: 1

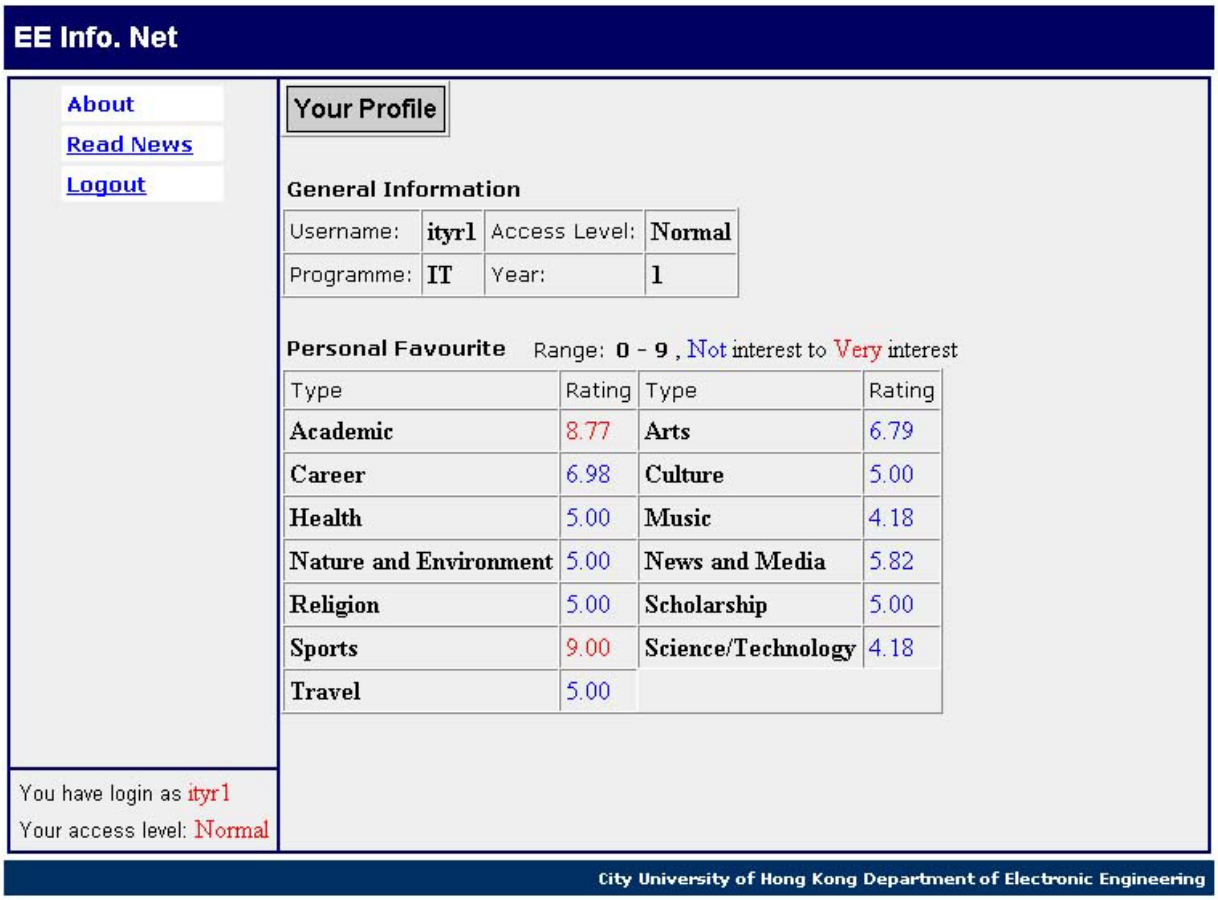
Username: ityr1

Access Level: Normal

Number of visiting of the EE Info Net: 7

Assumption: users read a news means users have click on the news header to pop-up the news content to read.

It is the screen dump of the users "*myinfo.jsp*" page.

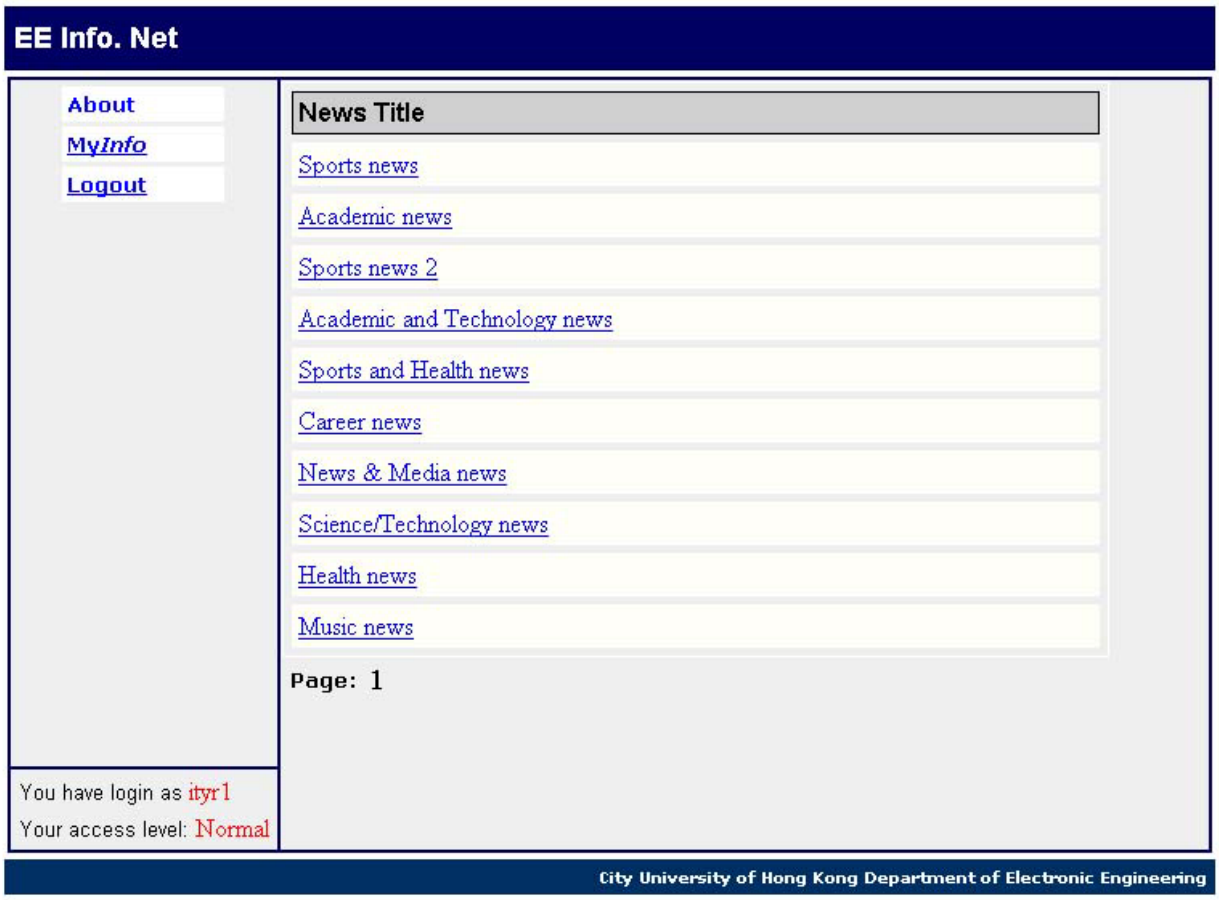


From the screen dump, it shows that user (ityr1), are interested in academic and sports area and not interested in music and science/technology area. Therefore, the personalization engine of the EE Info Net should recommend sports and academic news to the user.

7.1 Test of the news recommendation engine

The news recommendation engine would operate when user just login and sort the news in a personalized order. It would show in page newslist.jsp.

It is the screen dump of users "*newslist.jsp*" page.



From the screen dump, the top 5 news header showed are about sports and academic. And the bottom of the news header is news about music. It shows that the content-based

filtering of news recommendation engine have sorted the news in this order, users most interests news to least interests news.

7.2 Test of the real-time recommendation engine

The real-time recommendation engine would operate when user browsing/reading on the website. The recommendation results would show in page newslist.jsp when users after reading first news of a visit.

It is the screen dump of "*newslist.jsp*" page after user just read news "Sport news".

From the screen dump, three news headers are marked. They are "Sport news 2", "Academic and Technology news", and "Sports and Health news".

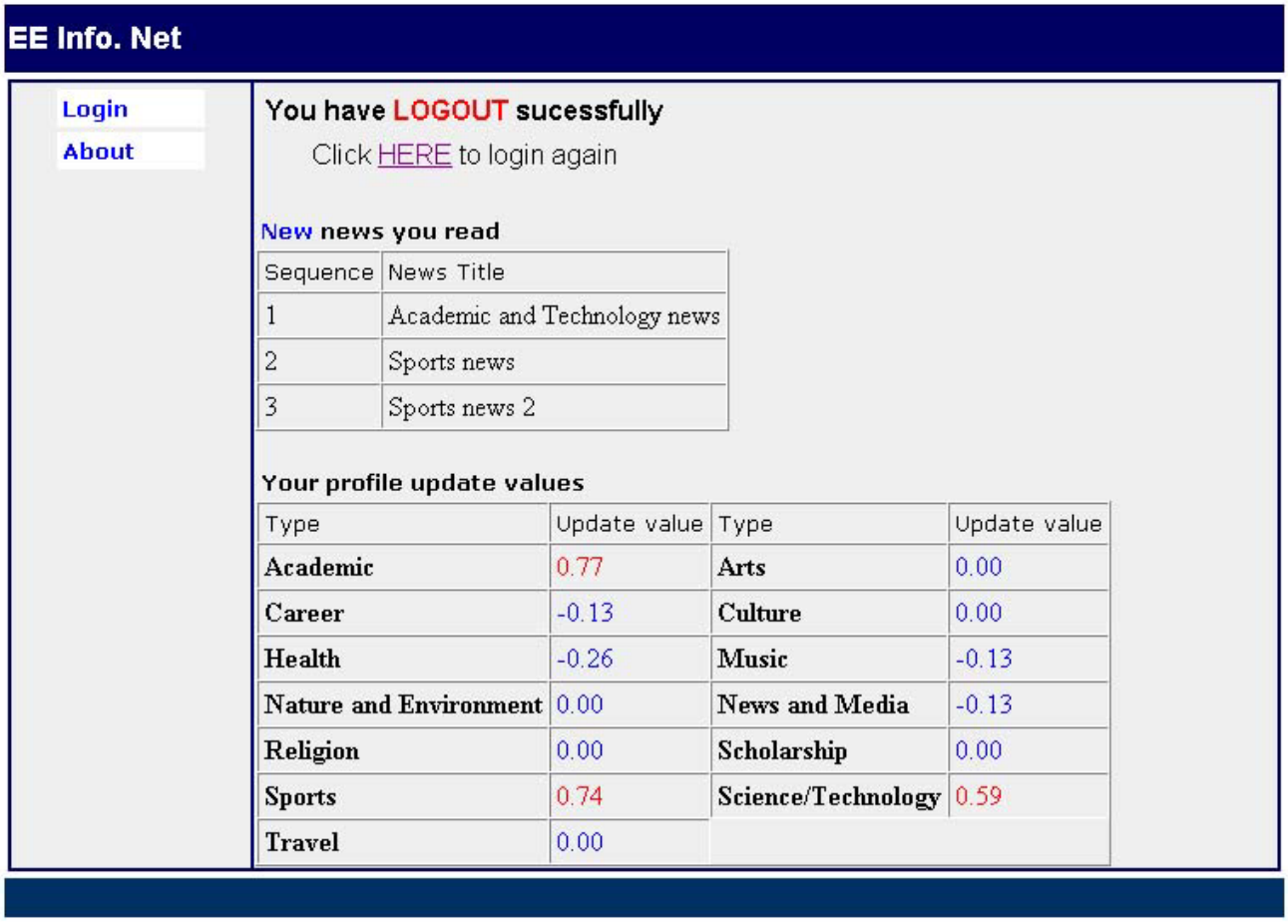
For the “Sport news 2” and “Sports and Health news”, a “**Related**” word is marked. It is because these two news are about sports while the last read news is about sports too. The related recommendation mechanism of the real-time recommendation engine performs these recommendations.

For the “Academic and Technology news” and “Sports news 2”, a “**Similar users pick**” word is marked. It is because these two news are what the other users, who have read “Sports news”, have read too. The collaborative filtering of the real-time recommendation engine does these recommendations. The recommendations are based on the similar users’ behaviors.

7.3 Test of the users' profiles decision engine

The users' profiles decision engine would operate when user logout from the website. The engine would decide users' profiles by the users browsing patterns of this visit.

It is the screen dump of users "*logout.jsp*" page. It showed what user read, the reading sequence and the update value of the users' profile of this visit.



EE Info. Net

Login
About

You have LOGOUT sucessfully
Click HERE to login again

New news you read

Sequence	News Title
1	Academic and Technology news
2	Sports news
3	Sports news 2

Your profile update values

Type	Update value	Type	Update value
Academic	0.77	Arts	0.00
Career	-0.13	Culture	0.00
Health	-0.26	Music	-0.13
Nature and Environment	0.00	News and Media	-0.13
Religion	0.00	Scholarship	0.00
Sports	0.74	Science/Technology	0.59
Travel	0.00		

The screen dump shows that the test user have read 3 news of this visit. They are about sports, academic and technology. From the profile update values, the value of these three attributes are positive, which means that the engine decide the user are interested in these three attributes.

Comparing the update values of academic and science/technology, the values of science/technology is lower than the academic one, user have read the news having these two major attributes. It is because of the users' point of view deciding engine, as from the

existing profiles, user is interested in academic more than science/technology, therefore, the engine operated to decide the user read this “Academic and Science/Technology news” due to the academic attribute more than the science/technology attribute. It shows the operation of the users’ point of view deciding engine of the users’ profiles decisions engine.

The reading sequence concern engine also functioned to set the weight of update of the read news. From this test, user have read two news about sports, however the update value of the sport attribute is still a bit lower than the academic attribute. That is because of the reading sequence concern engine, it would set the “update weighting” of the news according to the reading sequence, the ‘weight’ decrease as the prior of reading going forward, since people usually read news from most interest to least interest. In the other word, the engine operated to decide the user interest in academic more than sports by the reading sequence. It shows the operation of the reading sequence concern engine of the users’ profiles decisions engine.

8. Summary

The objectives of the project are achieved. An advanced personalization engine and a personalized EE news web site are built.

For the personalization engine, AI rules are used to perform the process of "understanding users". With the help of AI rules, the engine could carry out more complicated analysis and more number of factors could be concerned in the analysis. In the other words, the engine could analysis the user profiles in a more accurate manner.

Simple comparison of personalization engine

	Amazon.com	**EE Info Net**
Considered factors	**Real-time engine** 1. Similar users behaviors **Analysis browsing patterns** 1. Read objects	**Real-time engine** 1. Similar users behaviors 2. *"when and where"* situation **Analysis browsing patterns** 1. Read objects *2. Unread objects* *3. Reading sequence* 4. *Reading time*

When comparing with the other personalization engine in the market, it could be said that the performance of the personalization engine created in the project is better than most of the common personalization engine in the market. As user profiles are the most critical factor of success of personalization, and the engine created could perform an accurate user profiling, what leads better personalization.

For the personalized EE news web site, it is more convenient to use when comparing with the other personalized web site. Because most of the personalized web sites required user active involvements for profiling to perform personalization services, what is ignoring and inconvenient to use. However, the EE news web site does not require any user active involvements, the personalization services and profiling are invisible to the users. Users

do not need any extra efforts to achieve the personalization services, what could eliminate the problems of ignoring and inconvenient in most of the personalized web sites.

As web site personalization services is more and more common now, and there is no outstanding breakthrough from the existing personalization techniques. This project may give a new direction of research for improving of the performance of personalization services.

9. Acknowledgments

I would like to thank my supervisor -- Dr. CHUN, Andy H W that giving me a chance to work on this final year project, since then, followed with constant attention to my work, providing suggestions. Special thank to Dr FENG, Jian for her assessment, and helpful suggestions to the project.