Gradual Adaption Model for Information Recommendation Based on User Access Behavior

Jian Chen
Graduate School of Human Sciences
Waseda University
Tokorozawa, Japan
wecan_chen@fuji.waseda.jp

Roman Y. Shtykh
Graduate School of Human Sciences
Waseda University
Tokorozawa, Japan
roman@akane.waseda.jp

Qun Jin
Faculty of Human Sciences
Waseda University
Tokorozawa, Japan
jin@waseda.jp

Abstract—**In this study, we propose a gradual adaption model for information recommendation. This model is based on a set of concept classes that are extracted from Wikipedia categories and pages. Using the extracted information, data representing the users' information access behavior is collected by a unit of one day for each user, and analyzed in terms of short, medium, long periods, and by remarkable and exceptional categories. The proposed model is then established by analyzing the pre-processed data based on Full Bayesian Estimation. We further present experimental simulation results, and show the operability and effectiveness of the proposed model.**

Keywords-information recommendation; data mining; gradual adaption; Wikipedia

I. INTRODUCTION

Today we are surrounded by a *plethora* of information. Except traditional information in books, a variety of web resources are connected with each other by the Internet. We can use search engines to search such information, but the problem is that we cannot retrieve and perceive all search results that are above a number of thousands.

Efficient use of web resources is an important issue we try to resolve. In this study, we propose an information recommendation model called Gradual Adaption Model (GAM) [1]. In this model, we build a set of concept classes that are extracted from Wikipedia categories and pages.

In fact, more and more people are becoming increasingly accustomed to use Wikipedia to find knowledge since 2001. Furthermore, recently Wikipedia articles become more and more often referred by scientific papers. Owing to its good quality and reliability, Wikipedia can also be considered as a resource for information recommendation. Based on this, we investigated Wikipedia, and found that its category structure can be used for extracting a set of concept classes that are used as a classification criterion in our proposed model.

When users access web pages through the system that is built with the proposed GAM, it classifies these Web pages by concept classes. Such user access data are collected by a unit of one day for each user. Based on the collected data, the reuse probability of each concept class is estimated in terms of short, medium, and long periods by Full Bayesian Estimation. If a concept class belongs to more than two periods, it is classified as a concept class of remarkable category. If a concept class is accessed just occasionally, it means its probability is so low that the concept class almost impossibly appears in the front of recommendation results, it is classified as a concept class of exceptional category. When users access web resources next time, GAM will gradually adapt to the transition of users' selection, and recommend web pages for users, according to the concept class probability that is estimated by GAM.

This paper is organized as follows. In Section II, related works are introduced. Section III gives a brief introduction on Full Bayesian Estimation that we apply in this study. A detailed description on GAM is provided in Section IV, and simulation results are discussed in Section V. Finally, Section VI concludes this study and directs future works.

II. RELATED WORKS

As we know, Wikipedia is an open resource that can be modified by anyone. Therefore, we have to face a number of problems such as its reliability and trustworthiness. Further, we will consider these issues by overviewing the works dedicated to Wikipedia, and discuss several modern information recommendation approaches.

A. Wikipedia Information Resource

Nowadays, *an enormous amount* of useful resources can be discovered from Wikipedia. The report by Kashihana et al. [2] shows: there are more than 2.1 million items of the English version recorded in Wikipedia by January 2008. While the number of English articles in Encyclopaedia Britannica (2008 version) is more than 75 thousand. The report also says that the accuracy of articles in Wikipedia and those in the Encyclopaedia is almost equal.

Today, Wikipedia data set, its content and structure are widely used for extracting metadata for research. Wikipedia was found to have an impressive coverage of contemporary documents. As found by Milne et al. [3] after comparing Wikipedia articles and links with a manually-created professional thesaurus, it is a good source of hierarchical and associative relations, with good coverage and accuracy for many areas. Therefore, we can consider Wikipedia categories and theirs pages as an alternative for creation of concept classes and theirs representative indices, which are extracted from Wikipedia categories and theirs pages.

And for the above reasons, mining Wikipedia attracts many researchers. For instance, Mihalcea and Csomai [4] consider the abundance of links embedded in Wikipedia pages and try to extract keywords automatically from them. In addition to embedded links (that can be further classified as incoming links and outcoming links), section headings, template items of Wikipedia pages are considered as semantic features and used to represent a page. The similarity of two Wikipedia pages sharing these features can be used as a page similarity measure [5]. Obviously, the level of representativeness of a term used in a title, headline and text of an article differs. The keywords that occur in the title, headlines and embedded links are better representatives of pages, therefore, they gain higher-weighted values.

An attempt to find good quality articles of Wikipedia in order to recommend them automatically is described by Thomas and Sheth [5]. The idea is to analyze the change of Wikipedia pages by semantic convergence and estimate if these are good articles. This approach can find good articles in Wikipedia, but, in our opinion, to achieve better results and user satisfaction from recommendations, it is important to consider users' needs and behaviors during the information recommendation process.

B. Information Recommendation

Recently, information recommendation is a focus, and attracts much attention by a lot of users and researchers. The web mining [7, 8] approaches have been extensively used for information recommendation. Generally, web mining has been divided into three main areas: usage mining, structure mining, and content mining [7]. As an additional area, semantic web mining [9] was proposed. The following are the data types that are found in the web and mined by these approaches.

- Content data: The text and multimedia data in web pages. It is the real data that is designed and provided to users of a web site.
- Structure data: The data consist of organization inside a web page, internal and external links, and the web site hierarchy.
- Usage data: The web site access logs data.
- User profile: The information data of users. It includes both of data provided by users and data created by the web site.
- Semantic data: The data describe the structure and definition of a semantic web site.

Although web mining is divided into four areas, but they are associated mostly each other, not exclusively.

We recognize the importance of such web mining as content, structure and user profile. In this study, we focus on WUM (Web Usage Mining). In this area, a new document representation model [8] was presented recently. This model is based on implicit users' feedback to achieve better results in organizing web documents, such as clustering and labeling. This model was experimented on a web site with small vocabulary and specific to certain topics. Identifying Relevant Websites from User Activity [10] is another attempt of organizing web pages. It is also based on implicit users' feedback but faces the following problem – to improve retrieval accuracy. The model needs to spend more time to train the system.

However, using implicit users' feedback has such a problem: although there is a relation between the users' implicit feedback, there is also a possibility that a chanciness of implicit users' feedback can impair the relation between web documents clicked by users. Despite this problem, the mining of implicit users' feedback enables us to realize personalized information recommendation. In our work we focus on the implicit feedback coming from the same user, and do not consider interrelation of implicit feedback of different users.

We also noticed that due to the explosive growth of information on the web, web personalization has gained great momentum both in the research and commercial areas [11]. This fact encourages us to use implicit users' feedback to personalize information recommendation.

Dynamic Link Generation [12] is one of early WUM systems. It consists of off-line and on-line modules. In the off-line module, pre-processor extracts information from user access logs and generates records, then clusters the records to categories. The on-line module is used to classify user session records and identify the top matching categories, then return the links that belong to the identified categories to the user.

SUGGEST 3.0 [13] is another kind of WUM systems, but it has only the on-line component. In SUGGEST3.0, the off-line job, like that in Dynamic Link Generation, was realized in the on-line component dynamically. The aim of SUGGEST 3.0 is to manage large web sites. But the size of access logs used to evaluate the system is small and limited.

LinkSelector [14] is a web mining approach focusing on structure and usage. By this approach, hyperlinks-structural relationships were extracted from existing web sites and theirs access logs. Based on the relationships, a group of hyperlinks was given to users. Using a heuristic approach, users can access the group to find the information they want.

From the related studies we overviewed above, we can see that implicit users' feedback is widely used in recommendation systems. Further, recommender systems which consist of both off-line and on-line modules have higher performance than those which only have on-line module. Moreover, concept grouping is more user-friendly because it is easier to retrieve information from a concept group than from unstructured and not interrelated pool of information.

III. FULL BAYESIAN ESTIMATION

In this study, we use Full Bayesian Estimation that has the learning function for the proposed GAM.

The proposed model analyzes the selected link of web pages, and estimates which concept class it belongs to. One link selection is one data sample. The data sample belongs to each concept class. This is expressed as in Eq. (1).

$$\mathrm{D} = \{D_1, D_2, ..., D_n\} \qquad (1)$$

where D_i (i = 1, 2, …, n) represents an aggregate of access samples of concept class that D consists of. D_1 is an aggregate of access samples of concept class D_1. D_2 , ..., D_m are the same as D_1. They are the aggregate of access samples, and belong to concept classes D_2 …, D_m respectively.

We define data sample that is used in Full Bayesian Estimation as follows. If a link that belongs to a concept class D_m is clicked, we use d_t to describe the number of click times of D_m, and d_f to describe the number of click times that concept class D_m is not clicked (i.e., other concept classes are clicked). For the history logs (not including current day), we use a variable α_t to describe the number of click times of concept class D_m, and α_f to describe the number of click times that concept class D_m is not clicked.

For example, if the whole click times is 6 at current day, and the 2 times belong to concept D_m, it means $d_t = 2$, and $d_f = 4$, then we can calculate according to Eq. (2) [15], and obtain the click probability of the concept D_m is $^{2}\!/_{6}$.

$$\theta^* = \frac{d_t}{d_t + d_f} = \frac{d_t}{d} \qquad (2)$$

Equation (2) is called as Maximum Likelihood Estimation. Because the empirical value is disregarded by Maximum Likelihood Estimation, haphazardness can give a big influence on the estimation result.

But in Full Bayesian Estimation, the join of Prior Distribution (based on the history click samples) and Likelihood Estimation is used to calculate the Posterior Distribution θ. Its expression is described as follows.

$$\begin{aligned} &P(D_{m+1} = t \,|\, \text{Đ}) = \int P(D_{m+1} = t, \theta \,|\, \text{Đ}) d\theta \\ &= \int P(D_{m+1} = t \,|\, \theta, \text{Đ}) p(\theta \,|\, \text{Đ}) d\theta \\ &= \int \theta p(\theta \,|\, \text{Đ}) d\theta \qquad (3) \end{aligned}$$

where Đ is a data collection which consists of (D_1 , D_2 , …, D_m), and is used to describe the Likelihood Estimation. The integral calculation of Full Bayesian Estimation as shown in Eq. (3) is very complicated. Generally, it needs the following premises to make it calculable.

- Each sample in Đ is independent with each other, and satisfies *iid* (independent and identically distributed) assumption;
- About the current click times d_t and d_f , theirs prior distribution satisfies Bate Distribution $B[\alpha_t, \alpha_f]$.

Thus, the Full Bayesian Estimation formula can be expressed as follows [15].

$$
\begin{aligned}
P(D_{m+1}=t\,|\text{Đ}) &= \int \theta p(\theta\,|\text{Đ})d\theta \\
&= \frac{\Gamma(d_t+\alpha_t+d_f+\alpha_f)}{\Gamma(d_t+\alpha_t)\Gamma(d_f+\alpha_f)}\int \theta\theta^{d_t+\alpha_t-1}(1-\theta)^{d_f+\alpha_f-1}d\theta \\
&= \frac{d_t+\alpha_t}{d_t+d_f+\alpha_t+\alpha_f} \qquad (4)
\end{aligned}
$$

According to Eq. (4), if the number of the current samples is small, prior distribution has a big contribution on the result. On the contrary, if the number of the current samples is big, prior distribution has a little contribution on the result.

IV. A RECOMMENDER SYSTEM BASED ON GAM

In this study, we propose an information recommender system that is based on GAM (Gradual Adaptation Model), which consists of Concept Analyzer, Probability Estimator, and Gradual Adaption Recommender, as shown in Fig. 1.

The Concept Analyzer is used to analyze each user's access data representing his/her behaviors, and record the access data into logs. The Probability Estimator is used to estimate reuse probability of concept class for each user. The Gradual Adaption Recommender is used to analyze users' access logs and return recommendation results to gradually adapt to the transition of users' focus of interests.

The major features of the proposed system are described as follows.

- We divide users' interests into three terms of short, medium, long periods, and by remarkable, exceptional categories - which either pay a great attention to users' access behavior at current moment, or focus on casual user access.
- This system is an adaptive one. It can adapt to a transition of users' information access behaviors.
- In the system, training is not needed.

GAM is established based on Full Bayesian Estimation introduced in the previous section for estimation of user information access. This model consists of off-line and on-line components. The off-line component is used to analyze each user's access logs periodically and estimate concept classes' reuse probability for each user. The on-line component is used to analyze users' current access behavior and return recommendation results to gradually adapt to the transition of users' interests.

Fig. 1 shows the basic constitution of the proposed model (GAM), which consists of four phases, namely data pre-processing, access logs analysis, probability

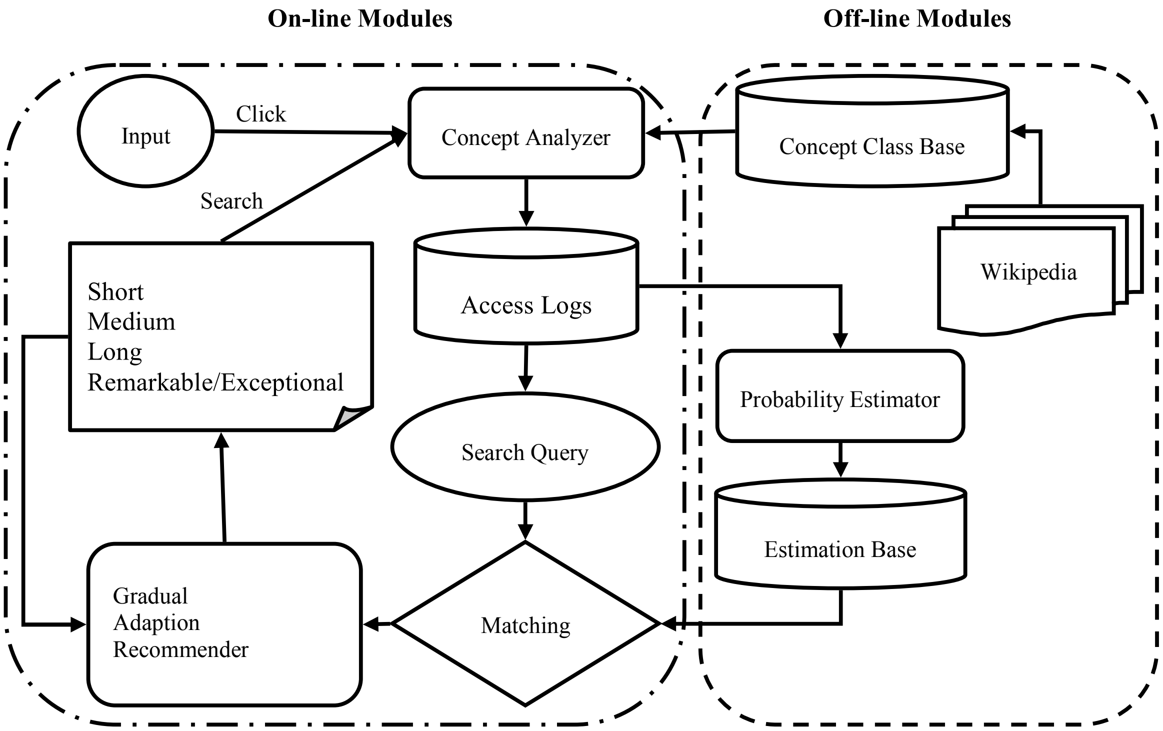
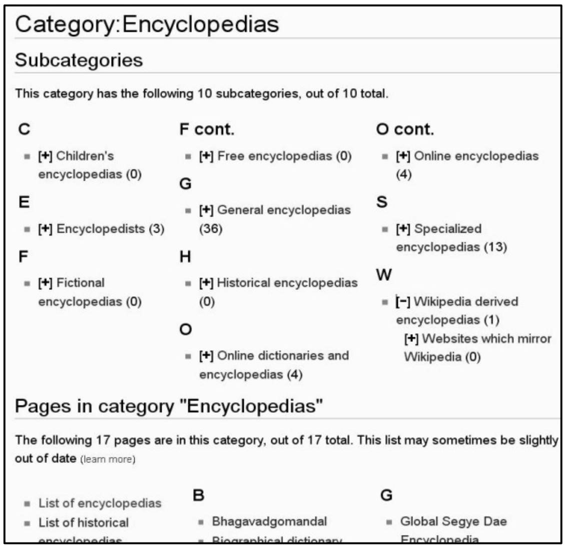


Figure 1. Gradual Adaption Model

estimating, and gradual adaptive recommendation.

A. Data Pre-processing



Category:Encyclopedias

Subcategories

This category has the following 10 subcategories, out of 10 total.

C

- [+] Children's encyclopedias (0)

E

- [+] Encyclopedists (3)

F

- [+] Fictional encyclopedias (0)

F cont.

- [+] Free encyclopedias (0)

G

- [+] General encyclopedias (36)

H

- [+] Historical encyclopedias (0)

O

- [+] Online dictionaries and encyclopedias (4)

O cont.

- [+] Online encyclopedias (4)

S

- [+] Specialized encyclopedias (13)

W

- [−] Wikipedia derived encyclopedias (1)
  [+] Websites which mirror Wikipedia (0)

Pages in category "Encyclopedias"

The following 17 pages are in this category, out of 17 total. This list may sometimes be slightly out of date (learn more)

- List of encyclopedias
- List of historical encyclopedias

B

- Bhagavadgomandal
- Biographical dictionary

G

- Global Segye Dae Encyclopedia

Figure 2. The Structure of Subcategory

At the data pre-processing phase, concept class base is created. Wikipedia [16] has 12 major categories. In each subcategory, there are pages and theirs sub-categories. Fig. 2 shows the structure of subcategory “Encyclopedia” [17] in Wikipedia. It has not only its subcategories, but also its pages.

Wikipedia category is regarded as a concept class in the proposed system. Because its pages can represent the subcategory, they are used to create index data of subcategories. Fig. 3 is the image for how to extract concept classes form Wikipedia categories.

A solid line text box means a category, or a concept class. A dotted line text box means a page, or an index data.

At first, Wikipedia categories are used to create a set of concept classes by one-to-one relationship. Then, all of pages are used to create indices for the categories that they belong to. Especially, if a category owns more than one page, its index data will be created from all of the pages as shown by the “Index 11+12” in Fig. 3.

We know that each word does not have the same importance in a page, and we need to give the weight to the words based on the importance in a page. Obviously,

- The words that are used in the title or headline of a page ought to have higher weight.
- A high frequency content-representative words are also more important for a page.

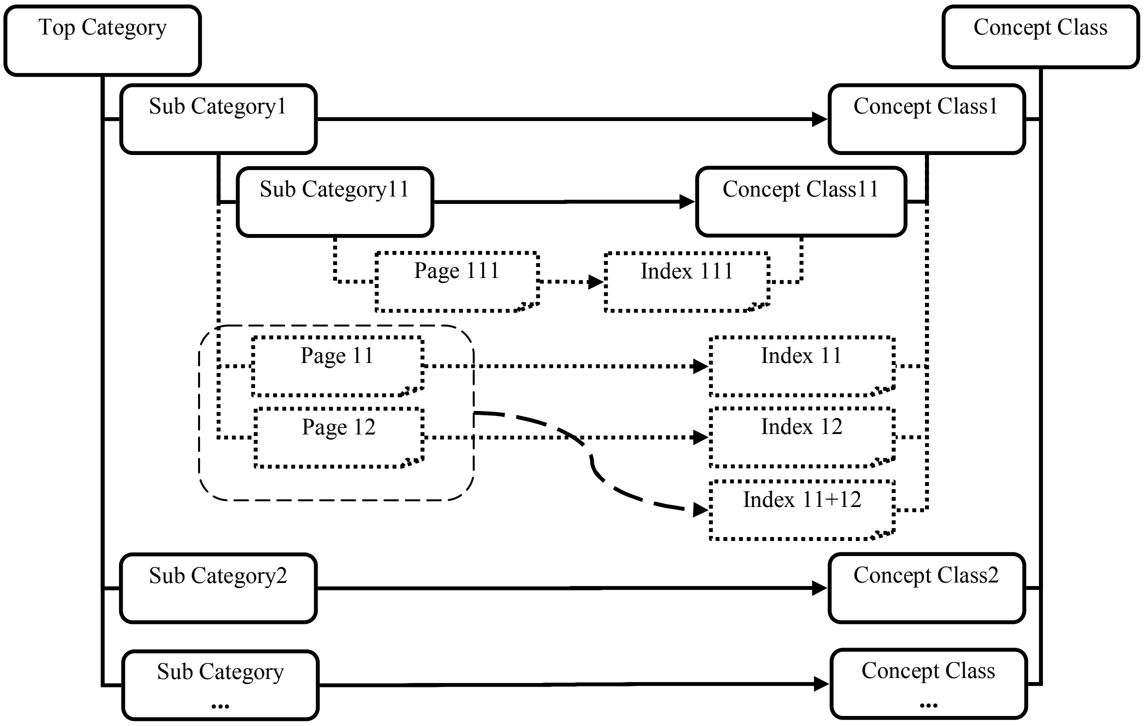


Figure 3. Concept Class and Index Extraction

- Embedded links are also given the higher weight.

Based on the above consideration, the weight is given to the items when creating indices.

Concept class extraction is a pre-processing step of the proposed system. After creation of a set of concept classes and theirs indices is done, the information recommender system can be started.

When users interact with the proposed system and provide feedback information at the first time, the system can use previously prepared index information of pages and the prior probability of concept class to give out the appropriate results to users. After users select some of results, then the access logs of user selections are used to calculate the posterior probability of concept classes. The details on how to record access logs and calculate posterior probability as discussed in Section III.

Each concept class consists of a number of keywords and URLs of web pages. Constitution of concept classes is shown in Fig. 4. It shows there are multiple concept classes in Concept Class Base. Each concept class owns multiple keywords, and some of keywords belong to multiple concept classes. When users access this system, their identifying information, clicked concept classes that include links, keywords, access date, click frequency are recorded in Access Logs. For example, a user browses recommendation results on keyword "culture" and clicks a link belonging to concept class "Art". As shown in Fig. 1, the user query will be sent to Concept Analyzer. After receiving this query, Concept Analyzer will analyze the query, and check the user's identifying information, concept class and keywords.

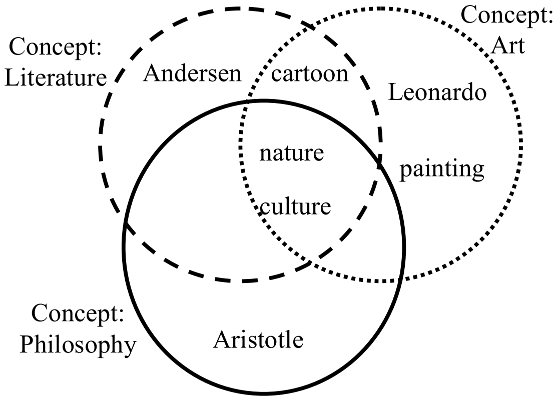


Figure 4. Constitution of Concepts

According to Fig. 4, because the clicked link belongs to concept class "Art", the sample number of concept class "Art" will be increased, and concept class, keywords, user identifying information, access date will be recorded into Access Logs. For the other concept classes (concept classes that were not accessed), there is nothing to do.

Considering the weight of keywords, Eq. (2) can be changed as follows.

$$\theta^* = \frac{\sum_{j=0} f(w_{mj}, k_{mj})}{\sum_{i=0}\sum_{j=0} f(w_{ij}, k_{ij})} \tag{5}$$

where, w_{ij} is the weight of a keyword, and k_{ij} is the selected sample number of the keyword in concept class i. Using Eq. (5), we can calculate the prior probability of each keyword from indices. As to the keyword's weight, if the keyword is in the text body, its weight is set to small values. If it is in the headline, its weight is higher than the former. If it is in an embedded link, its weight is regarded as a value between those that can be given to a headline and a text body.

B. Access Log Analyzing

When users use the proposed system, their access data representing their behaviors are analyzed and recorded by Concept Analyzer.

As the reason described above, the weight of keyword is also used to measure the selected data sample number. If the search keyword is only one, it means the sample number of this keyword increases by 1. If there are a number of search keywords, the sample number of access is divided into each keyword by its weight as follows.

$$s_m = \frac{\sum_{j=0} f(w_{mj}, k_{mj})}{\sum_{i=0}\sum_{j=0} f(w_{ij}, k_{ij})} \tag{6}$$

where i is the number of keywords, and each keyword has j weight types in the selected page. This result is recorded into the Access Logs of proposed system.

User interests are not static. They may change with time and/or environment. Therefore, we try to analyze user access logs by three periods: short, medium and long.

C. Probability Estimating

For the definition of periods, in this study, a fixed period is applied, though dynamic approaches and mechanisms as proposed in [18] can also be considered. To be simplified, in this paper, we assume the short period to 7 days (a week), the medium period to 30 days (a month), and the long period to 90 days (a quarter) (as shown in Fig. 5). All of the three periods start at previous day (-1) and do not include the current day. The short period is designed to reflect temporary interests of users. The medium period is designed for an interest that is affected by some factors, i.e., this interest is relatively stable during a period. The long period is designed for a long-term user interests. In Section V, the different features of these three periods will be shown by simulation.

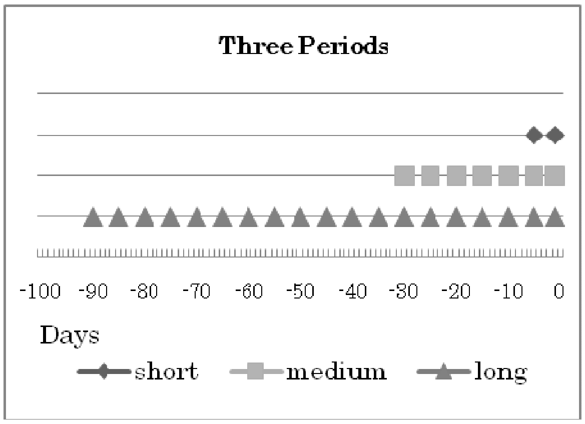


Figure 5. The Definition of Each Period

Except the three periods, we design two specific categories, namely remarkable and exceptional. Remarkable is based on the three periods. If there is a concept class belongs to more than one period, we call such concept class as remarkable concept class. The remarkable concept class means high degree of interest of a user in a particular concept class. There is also another category called exceptional. Exceptional category is an aggregate of a concept class that has a little chance to be clicked by users, but may be useful occasionally in the future for users.

The part located in the right side of Fig. 1 and surrounded by dotted line is the off-line component of system. As an off-line component, it attempts to improve the performance of the system. Probability Estimator is a part of off-line component and used to estimate the probability of concept classes. It is designed as a batch process and runs at a specific time (e.g., at the midnight) of every day.

The probability estimation is based on Eq. (4). For example, if we need to estimate a probability of concept class "Artists" and it is about user A in short period, four data items are necessary. The one pair is the sample number clicked and non-clicked by user A at the current day. The other pair is the summation of sample number clicked and non-clicked by user A in short period. Using these data, the estimator can calculate the probability of concept class "Artists" in short period. The estimator can also calculate the probabilities of other concept classes in the same way. Thus, Estimation Base can be created.

D. *Gradual Adaption Recommendation*

After creating Estimation Base, the system can start

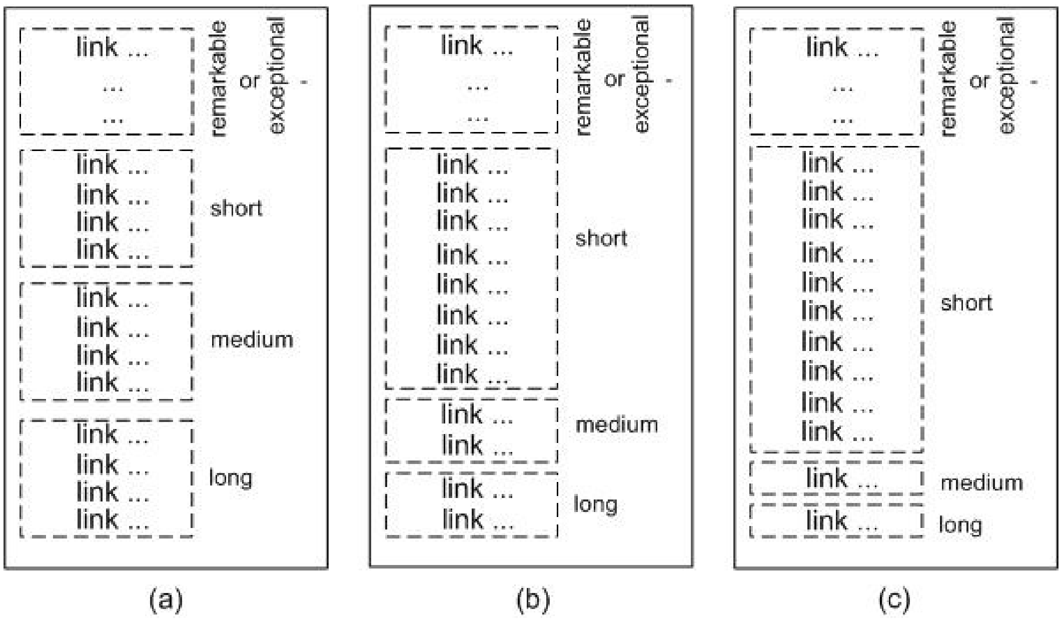


Figure 6. Three periods and two categories

the recommendation for users. The GAR (Gradual Adaption Recommender) is an on-line component. It is shown in the left part of off-line component in Fig. 1, and surrounded by dotted line.

When a user sends a search query to the system, GAR will check if there is a remarkable concept class from Estimation Base. As shown in Fig. 6, if remarkable concept class exists, GAR will return links of remarkable concept class and put them at the top of a recommendation page, choose a certain number of links from each period respectively, and add them below the remarkable links. Of course, these links belong to the concept class which has high probability in each period.

If a remarkable concept class is not found, GAR will check if an exceptional concept class exists. If an exceptional concept class exists, GAR will choose links of the concept class, then choose links from each period, and return the result in a random manner. The same as in the previous case, these links belong to the concept class which has high probability in each period.

If both remarkable and exceptional concept classes do not exist, GAR will choose the same number of links from each period respectively. These links belong to the concept class which has high probability in each of them. Then, these links are returned to a user in a random fashion.

Using the described approach, GAR gives a user a hint about which concept class is their hot concept class or which concept class is the concept class they almost forgot.

Fig. 6 (a) is the first response to a user. If the user makes a decision on a link and click it, the concept class, keyword and period or category information about the link will be sent to the system. Obtaining such information, the system will apperceive the user's demands.

As show in Fig. 6 (b), if the selected link belongs to short period, the links number of short period will be doubled. At the same way, the links number of other terms will be reduced to half.

As show in Fig. 6 (c), if the link of short period is clicked continuously, the links number of short period will be increased to a maximum number, and the number of other links of each period will be reduced to a minimum number. If a link that belongs to the short period is not clicked continuously, and another link that belongs to the other period is clicked - for instance, a link that belongs to the long period is clicked - in this case, the number of recommended links for the short period will be reduced to half, and at the same time, the number of links for the long period will be doubled.

The same things occur in case of other periods, and GAR will apperceive the change and redress the recommendation result. Therefore, GAR can give a high satisfaction rating to users.

V. SIMULATION AND EVALUATION

In order to verify the operability of the proposed GAM, we pre-produced the model. The system was built by open source software: Java, Tomcat, MySQL, and Nekohtml were used. Using them, Concept Analyzer, Probability Estimator, Gradual Adaptive Recommender were built

For the simulation, we consider three basic cases to evaluate the system. The first case is a user who has a long-term interest. In this case, the probability of the interested concept class ought to keep a high rate in long period.

The second case is a user who has a temporary interest. The user access the concept class of temporary interest sometime. In this case, this concept class ought to keep a low rate in the three periods.

The third case is a user who has two interests, and these interests are affected by some factors easily. In the case, there is a possibility that the probability of the relation concept class can change hugely in the short or medium period, but not in the long period.

A. Concept Class Base

We use Wikipedia on DVD Version 0.5 [19] (we refer to it as Wikipedia 0.5 for brevity) as the test data. The lowest categories in Wikipedia 0.5's topic hierarchies are used as the concept classes in the simulation. Based on Wikipedia 0.5, we gained more than 2000 web pages that belong to 180 concept classes. These concept classes were ready in advance, and saved in the Concept Class Base.

B. Setting of Simulation Cases

For case one, the concept class "Philosophical thought movements" is assumed to be used every day, and the assumed number of clicks (a user's accesses) was set to 0 and 40 per day. It means the user is interested in the concept class, and has a long-term interest in the concept class.

For case two, the concept class "Philosophers" is assumed to be used per three days, and the assumed number of clicks was set to 0 and 10. It means the user has little interest in this concept class.

For case three, two concept classes of "Art" and "Artists" are assumed to be used, and the number of clicks is dynamically varying. Most times it is set to 0 to 20, but sometimes it is set to 0 to 10 (likes case two), some other times it is set to 0 to 80 (large than case one).

Obviously, concept classes "Philosophers" and "Philosophical thought movements", and concept classes "Art" and "Artists" are similar respectively in cases described above. It means that similar concept classes have similar keywords. We expect that our model can differentiate the concept classes even if they contain similar keywords, and gain the results as we explained above.

C. *Simulation Results*

We simulated the three test cases during a period of 150 days, and obtained the results. The results are what we expected, showing high adjustability and adaptability of the proposed model.

In the short period, we can see the movement of the concept rate changing frequently. In some days, the probability of concept classes in case three is bigger than case one – for instance, "Art" concept class gets higher probability at 2008/12/12 point , "Artists" concept class gets higher probability at 2008/10/31 point (Fig 7).

In the medium period, the change is becoming smaller. But the probability of concept classes in case two is bigger than case three in some days (Fig 8). The exchange of probability between "Art" and "Artists" is also can be seen at 2008/10/31 and 2008/11/28 point.

In the long period, the change becomes quite stable. There is no big change in the long period (Fig 9).

From the simulation results, we found that the proposed model adapts well to the change of user's interests, as we expected. Thus, if a concept class is used frequently, it ought to have a high probability in the long period. If the concept class is used to a certain extent, it ought to have a quick change in the short or the medium period. If the concept class is used rarely, the rate ought to keep at a low level. This result demonstrates that the proposed model is operable and effective for modeling situations similar to those in the above-mentioned cases.

VI. CONCLUSION

In this study, we have proposed a gradual adaption model (GAM) for estimation of user information access behavior, based on Full Bayesian Estimation with a learning function, in order to solve the uncertainty problem caused by differences in user information access behaviors. A variety of users' information access data are collected and analyzed in terms of short, medium, long periods, and by remarkable and exceptional categories. We have further implemented a prototype system based on the proposed model, designed experimental simulations with three assumed cases to show operability and effectiveness of the model.

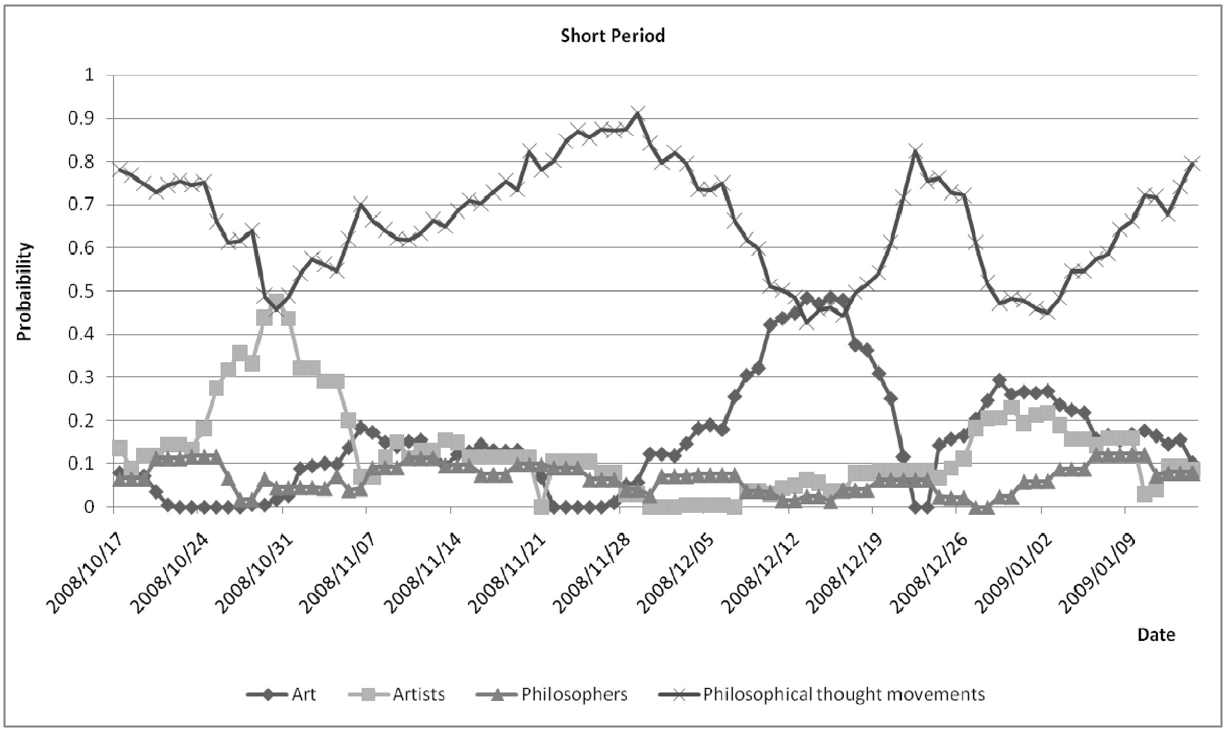


Figure 7. Probability of Concept classes in Short Period

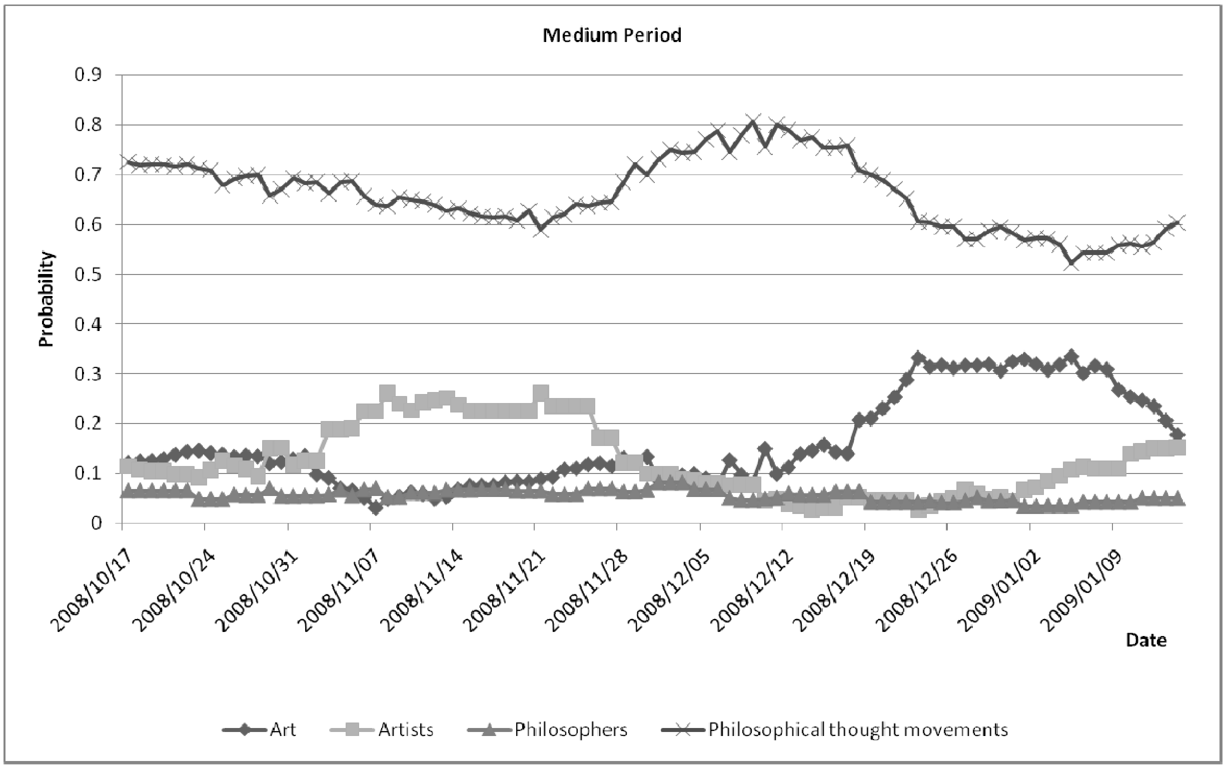


Figure 8. Probability of Concept classes in Medium Period

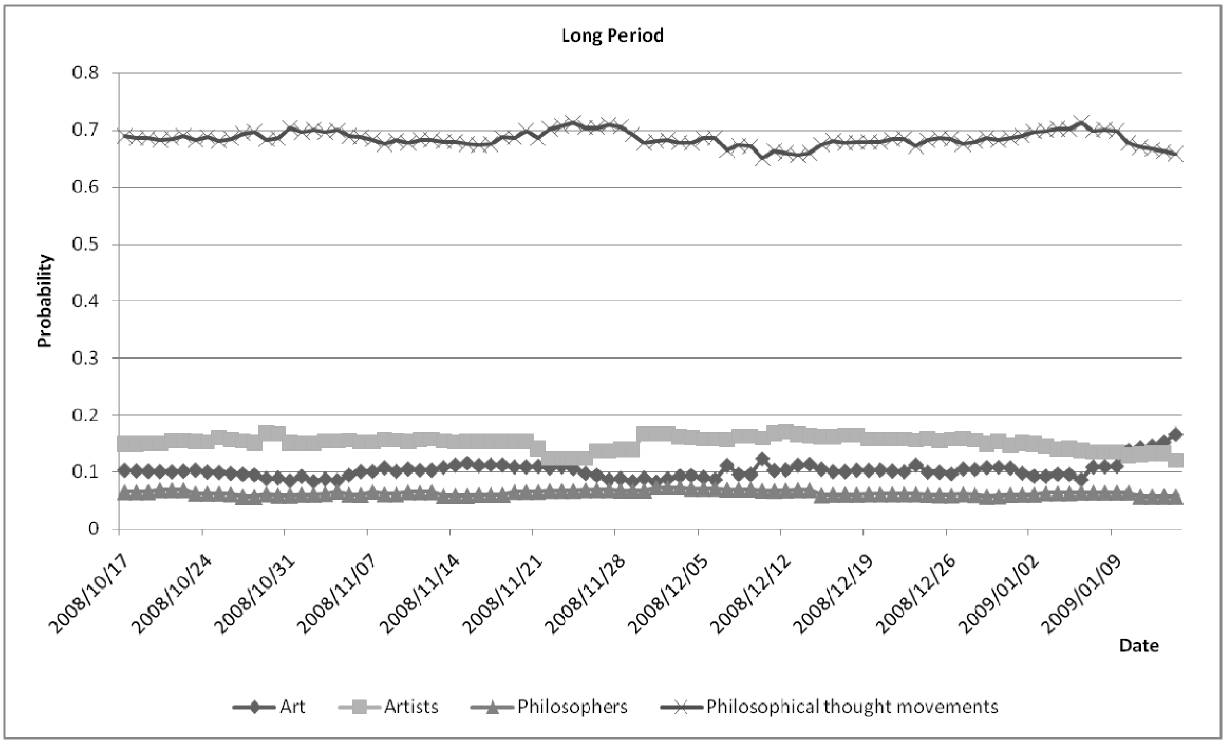


Figure 9. Probability of Concept classes in Long Period

The simulation results have shown that the proposed model can recognize the transition of users' access behaviors (web page selections, in particular) sensitively in the short period. The users' long-term interest is kept a high probability in the long period. The three periods of GAM can correctly distinguish long-term and temporary interest of users. Based on the results, when a user inputs a keyword and selects a link of a concept class that belongs to the long period, GAM can return the links of the concept classes that belong to the long period and match with the input keyword. Because the other concept classes that belong to the short and medium periods are filtered, GAM can help user to find the information that he/she is seeking quickly. Of course, GAM can detect which period is focused by a user, therefore, it can gradually adapt to the transition of users' selection, and provide appropriate information to the user.

As for future works, we will set more different patterns for the short, medium and long periods to find more reasonable ones. Using a dynamic sampling to set the three periods is one of the future works. Moreover, we will implement a fully runnable system, and evaluate the proposed model with users' involvement. We expect such experiment results can give us insights on how to further improve the model. We will also compare the proposed approach with other related recommendation models.

REFERENCES

[1] J. Chen, R. Shtykh, Q. Jin, "Gradual Adation Model for Estimation of User Information Access Behavior," ICSNC '08, pp. 378-383.

[2] M. Kashihana, S. Takeshi, Y. Endo, R. Doi, "Evaluation of Wikipedia (in Japanese)," March 2008.

[3] D. Milne, O. Medelyan, Ian H. Witten, "Mining Domain-Specific Thesauri from Wikipedia: A case study," Proc. IEEE/WIC/ACM International Conference on Web Intelligence (WI' 06), Hong Kong, China, 2006, pp. 442-448.

[4] R. Mihalcea, A. Csomai, "Wikify! Linking Documents to Encyclopedic Knowledge," CIKM'07, Lisboa, Portugal, November 2007, pp. 233-241.

[5] Y. Wang, H. Wang, H. Zhu, Y. Yu, "Natural Language Processing and Information Systems," Springer Berlin / Heidelberg, August 2007, Vol. Volume 4592/2007.

[6] C. Thomas, Amit P. Sheth, "Semantic Convergence of Wikipedia Articles," IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), Silicon Valley, USA, 2007, pp. 600-606.

[7] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," ACM SIGKDD, Vol 1, Issue 2, Jan. 2000, pp. 12–23.

[8] B. Poblete, R. Baeza-Yates, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. WWW2008, Beijing, China, Apr. 2008, pp. 41-48.

[9] G. Stumme, A. Hotho, B. Berendt, "Semantic Web Mining State of the Art and Future Directions," Elsevier Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 4, No. 2, 2006, pp. 124-143.

[10] M. Bilenko, R. W. White, "Mining the Search Trails of Surfing Crowds: Identifying Relevant Websites From User Activity," Proc. WWW2008, Beijing, China, Apr. 2008, pp. 51-60.

[11] M. Eirinaki, M. Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol. 3, No. 1, 2003, pp. 1–27.

[12] T-W. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Proc. WWW1996, Paris, France, May 1996, pp. 1007-1014.

[13] R. Baraglia, F. Silvestri, "An Online Recommender System for Large Web Sites," Proc. IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), Beijing, China, Sep. 2004, pp. 199-205.

[14] X. Fang, O.R. Liu Sheng, "LinkSelector: A Web Mining Approach to Hyperlink Selection for Web Portals," ACM Transactions on Internet Technology, Vol. 4, No. 2, May 2004, pp. 209–237.

[15] L. Zhang, H. Guo, *Introduction to Bayesian Networks (in Chinese)*, Science Press, 2006.

[16] http://en.wikipedia.org/wiki/Portal:Contents/Categorical_index

[17] http://en.wikipedia.org/wiki/Category:EncyclopediEn

[18] J. Chen, R. Shtykh, Q. Jin, "A Web Recommender System Based on Dynamic Sampling of User Information Access Behaviors," submitted to CIT'09, Xiamen, China, Oct. 2009.

[19] http://www.wikipediaondvd.com/nav/art/d/w.html