

Wikipedia カテゴリより抽出されたコンセプトクラスに基づく情報推薦モデル

An Information Recommendation Model Based on Concept Classes

Extracted from Wikipedia Categories

陳 健 シュティフ ロマン 金 群

Jian Chen, Roman Y. Shtykh, and Qun Jin

早稲田大学大学院人間科学研究科

Graduate School of Human Sciences, Waseda University, Japan

本研究では、Wikipedia のカテゴリ情報から抽出されたコンセプトクラスに基づく情報推薦モデルを提案する。まず、Wikipedia からコンセプト構造とコンセプトを表現するインデックス情報を抽出し、それに基づいて、利用者の情報アクセス行動のデータをコンセプトクラスごとに収集するとともに、ショート、ミディアム、ロングといった期間に分けてそれらのデータを解析し、各コンセプトの期間ごとの確率を算出し、利用者に情報を推薦するためのモデルを構築する。

Abstract: *In this study, we present an information recommendation model based on a set of concept classes that are extracted from Wikipedia categories and pages. The indices of all the pages are organized so that they represent concepts. Using this information, data representing the users' access behavior are collected and categorized according to the concept classes. The proposed model is then established by analyzing the preprocessed data in terms of short, medium, long periods, and calculating the probabilities corresponding to each concept.*

1. Introduction

Although there are a number of specialized encyclopedias in our bookshelves, since 2001 more and more people are becoming increasingly accustomed to using Wikipedia to find knowledge. Furthermore, recently Wikipedia articles become more and more often referred by scientific paper. Owing to its high reliability, Wikipedia can also be considered as a resource for information recommendation.

In this paper, we propose an information recommendation model that utilizes Wikipedia data. Firstly, we build a set of concept classes that are extracted from Wikipedia categories and pages. When users select some of the results that are recommended by the proposed system, the access data are collected by a unit of one day for each user. Using the collected data, the reuse probability of each concept is estimated in terms of short, medium, and long periods. If a concept belongs to more than two periods, it is classified

as a concept of remarkable category. If a concept is accessed just occasionally and its probability is low, it is classified as a concept of exceptional category. When users use the proposed system next time, the system will gradually adapt to the transition of users' selection among recommendation results, and recommend web pages according to the concept probability estimated by the proposed model.

2. Related Work

Nowadays, a plethora of useful resources can be discovered from Wikipedia. The report by Kashihana et al. [1] 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. [2] 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 their pages as an alternative for creation of concept classes and their representative indices, which are extracted from Wikipedia categories and their pages.

And for the above reasons, mining Wikipedia attracts many researchers. For instance, Mihalcea and Csomai [3] 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 outgoing 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 [4]. Obviously, the level of representativeness of a term used in a title, headline and text of an article differs. The keywords that occur in 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 in [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 for recommendation, it is not sufficient, as users' needs and behaviors must be considered in information recommendation. Analyzing user access logs is a general approach for user-centric recommendation. A new document representation model [6] and identifying relevant websites from user activity [7] are based on implicit user feedback to achieve better results in organizing web documents by clustering and labeling. However, using implicit user feedback has the following problem: although there is a relation between the users' implicit feedback, there is also a possibility that chance of implicit users' feedback can impair the relation between web documents selected by users. Despite this problem, the mining of implicit users' feedback enables us to realize personalized information recommendation. In our work we focused on the implicit feedback coming from the same user, and do not consider interrelation of implicit feedback of different users.

When considering a practical use, processing of information in recommender systems can be very complex. Some approaches like SUGGEST 3.0 [8] use on-line processing modules, but some like Dynamic Link Generation [9] and LinkSelector [10] use both on-line and off-line modules to achieve higher performance compared to those recommender systems that have only an on-line module.

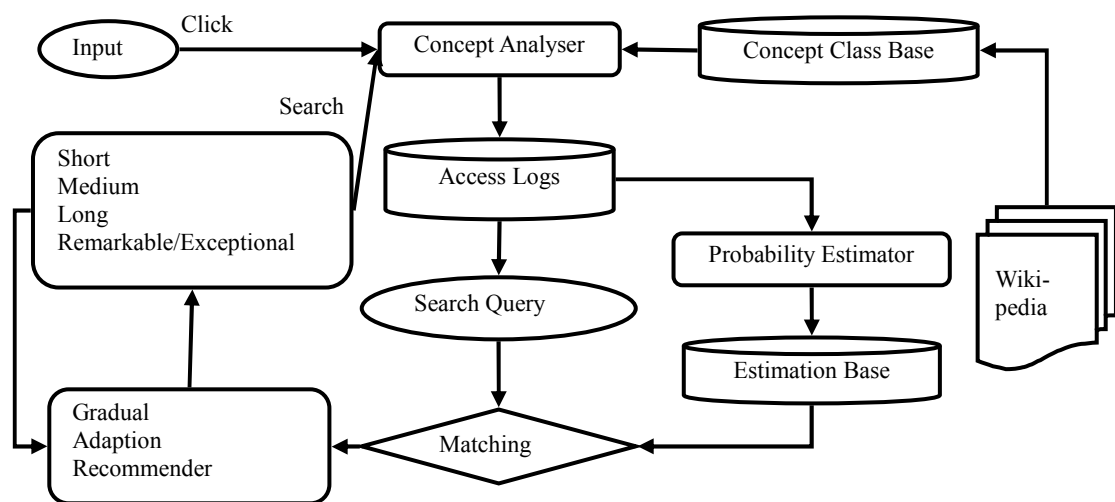


Figure 1: Gradual Adaption Recommendation Model

3. Overview of the Proposed System

Based on the above investigations, we propose an information recommender system that consists of Concept Analyser, Probability Estimator and Gradual Adaption Recommender, (GAR) as shown in Figure 1.

The Concept Analyzer is used to analyze each user's access behavior, and record their 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 behavior and return recommendation results to gradually adapt to the transition of users' interests.

The major features of our proposed model are described as follows.

- (i) We divide users' interests into three terms of short, medium, long periods, and by remarkable, exceptional categories - the first one pays a great attention to users' access behavior at current moment, and the second one focuses on haphazard user access.
- (ii) This model is an adaptive one. It can adapt to a transition of users' information access behaviors.
- (iii) In the model, training is not needed.

Figure 1 shows the basic constitution of our proposed model (GAR). This model consists of three phases, namely pre-processing, access logs analysis, and gradual adaptive recommendation.

4. Concept Class and Index Construction

Wikipedia¹ has 12 major categories. In each subcategory, there are pages and their sub-categories. Figure 2 shows the structure of subcategory "Encyclopedia" in Wikipedia. It has not only its subcategories, but also its pages.

Wikipedia category is regarded as a concept class in our proposed system. Because its pages can represent the subcategory, they are used to create index of subcategories. Figure 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.



Figure 2: The Structure of Subcategory

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 will be created from all of the pages as shown by the "Index 1" in Figure 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,

- (i) The words that are used in title or headline of a page ought to have high weight.
- (ii) A high frequency content-representative words are also important for a page.
- (iii) Embedded links are also given the high 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 our proposed system. After creation of a set of concept classes and their 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, 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 will be discussed in the next section.

¹ http://en.wikipedia.org/wiki/Portal:Contents/Categorical_index

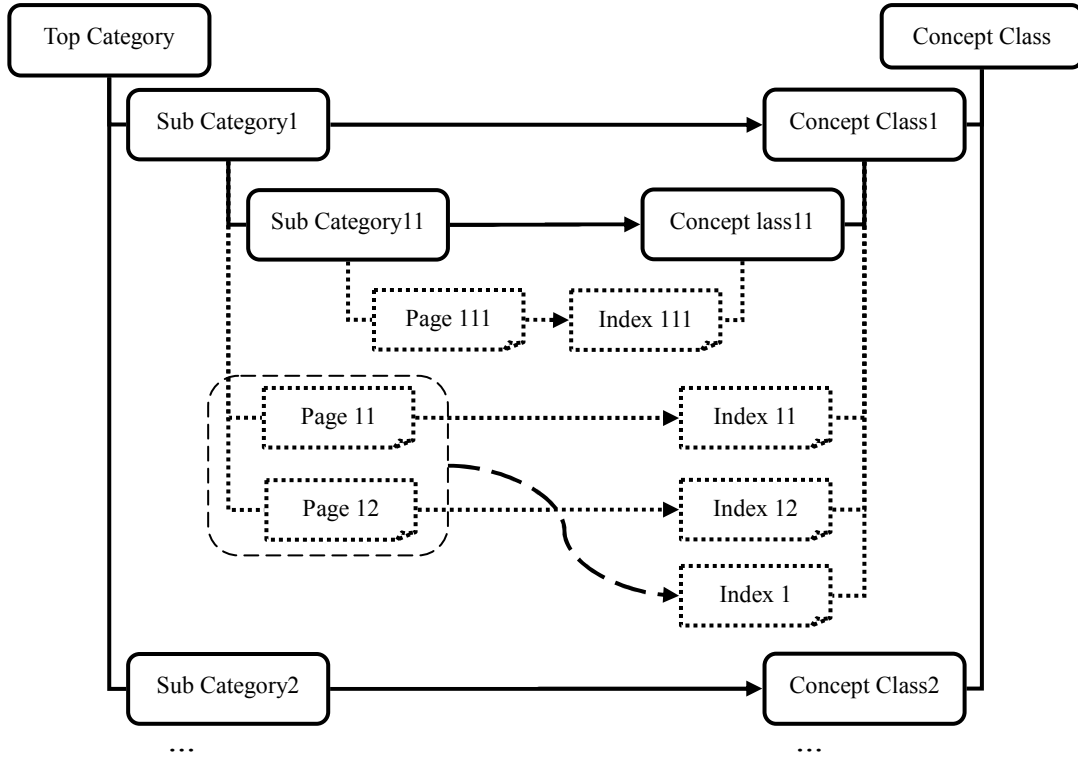


Figure 3: Concept Class and Index Extraction

5. Concept-Based Recommendation

After extracting concept classes from Wikipedia, there are three steps we need to complete for a recommendation.

5.1. Prior Probability Calculation

The prior probability of a keyword for a concept class m can be calculated as follows.

$$\theta'_m = \frac{\sum_{j=0} f(w_{mj}, k_{mj})}{\sum_{i=0} \sum_{j=0} f(w_{ij}, k_{ij})} \quad (1)$$

Here, w_{ij} is the weight of keyword k_{ij} , and k_{ij} is the number of the keyword in concept class i . Using this formula, 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.

Because the prior probability is based on the concept classes, these results can be used for every user.

5.2. User Access Log Analyzer

When users use the proposed system, their access behaviors are recorded and analyzed by the system. Users' access logs consist of their identification data, search keywords, the sample number of selected page links, and access date. When a user selects a link, this behavior is regarded as 1 sample of access.

The weight of keyword is also used to measure the access 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 need to be divided into each keyword by its weight as follows.

$$S_m = \frac{\sum_{j=0} f(w_{mj}, k_{mj})}{\sum_{i=0} \sum_{j=0} f(w_{ij}, k_{ij})} \times 1 \quad (2)$$

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

5.3. Posterior Probability Estimator

We define access data sample as follows. If a link/page that belongs to a concept class D_m is selected, we use d_i to describe the number of click times of D_m , and

d_f to describe the number of click times that concept D_m is not clicked (i.e. other concept classes are clicked).

For the prior probability distribution of D_m , the sample number that belongs to D_m is α_t , the sample number not belonging to D_m is α_f . Based on Formula 1, we can obtain the following equations:

$$\begin{aligned}\alpha_t &= \sum_{j=0} f(w_{mj}, k_{mj}) \\ \alpha_t + \alpha_f &= \sum_{i=0} \sum_{j=0} f(w_{ij}, k_{ij})\end{aligned}\quad (3)$$

By 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:

$$\begin{aligned}P(D_{m+1} = t | \mathcal{D}) &= \int P(D_{m+1} = t, \theta | \mathcal{D}) d\theta \\ &= \int P(D_{m+1} = t | \theta, \mathcal{D}) p(\theta | \mathcal{D}) d\theta \\ &= \int \theta p(\theta | \mathcal{D}) d\theta\end{aligned}\quad (4)$$

Here, \mathcal{D} is a data collection which consists of (D_1, D_2, \dots, D_m) , and is used to describe the likelihood estimation. The integral calculation of Full Bayesian Estimation is very complicated. Generally, it needs the following premises to make it calculable.

- (i) Each sample in \mathcal{D} is independent from each other, and satisfies *iid* (independent and identically distributed) assumption;
- (ii) Given 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 [11]:

$$\begin{aligned}P(D_{m+1} = t | \mathcal{D}) &= \int \theta p(\theta | \mathcal{D}) d\theta \\ &= \frac{\Gamma(d_t + \alpha_t + d_f + \alpha_f)}{\Gamma(d_t + \alpha_t) \Gamma(d_f + \alpha_f)} \int \theta \theta^{d_t + \alpha_t - 1} (1 - \theta)^{d_f + \alpha_f - 1} d\theta \\ &= \frac{d_t + \alpha_t}{d_t + d_f + \alpha_t + \alpha_f}\end{aligned}\quad (5)$$

Substituting Formula 3 for Formula 5, the posterior probability of concept class D_m can be calculated. According to this formula, if the number of the current samples is small, prior distribution has a big impact on the result. On the contrary, if the number of the current samples is big, prior distribution has a little impact on the result.

5.4. Information Recommendation

After creating Estimation Base, the system can start the recommendation for users. The GAR (Gradual Adaption Recommender) is a real-time model with an on-line component and a batch component (Figure 1).

When a user sends a search query to the system, GAR will check if there is a remarkable concept from Estimation Base. If a remarkable concept exists, GAR will return links of the remarkable concept and put them to the top of web page, then 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 which has high probability in each period.

If no remarkable concept is found, GAR will check if an exceptional concept exists. If an exceptional concept exists, GAR will choose links of the concept, then choose links from each period and return the result in a random manner. Like in the previous case, these links belong to the concept which has high probability in each period.

If both remarkable and exceptional concepts do not exist, GAR will choose same number of links from each period respectively. These links belong to the concept 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 is their hot concept or which concept is the concept they almost forgot.

If a user makes a decision on a link and click it, the concept, 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.

If the selected link belongs to the short period, the link number of the short period will be doubled. At the same way, the links number of other terms will be reduced to half. If the link of the short period is clicked continuously, the link number of the short period will be increased to the maximum number, and number of the other links of each period will be reduced to the minimum number.

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.

6. Conclusion and Future Works

In this study, we have proposed a gradual adaption recommendation model based on concept classes extracted from Wikipedia and used for estimation of user information access behaviors 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.

As for the future work, our plan is organized in two steps. Step one is to do a simulation with some different patterns for the short, medium and long periods to find more reasonable ones. The second step is to implement a fully runnable recommender system based on GAR. A set of concept classes extracted from Wikipedia will be used in the system. When a user interacts with the system, the GAR will get the results from an external search engine, and re-rank the results by the analyzed data of user access logs, then return the re-ranked results to the user. Pages selected by the user will be classified according to the available concept classes and recorded as logs for the future analysis.

Moreover, we will 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 our proposed approach with other related recommendation models.

7. References

- [1] M. Kashihana, S. Takeshi, Y. Endo, R. Doi. "Evaluation of Wikipedia (in Japanese)". March. 2008.
- [2] 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.
- [3] R. Mihalcea, A. Csomai. "Wikify! Linking Documents to Encyclopedic Knowledge", CIKM'07, Lisboa, Portugal, November. 2007, pp. 233-241.
- [4] Y. Wang, H. Wang, H. Zhu, Y. Yu. "Natural Language Processing and Information Systems", Springer Berlin / Heidelberg, August. 2007, Vol. Volume 4592/2007.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] L-w Zhang, H. Guo, *Introduction to Bayesian Networks (in Chinese)*, Science Press, 2006.