The Electronic Library

A cross-language personalized recommendation model in digital libraries

Yuangen Lai Jianxun Zeng

A cross-language personalized recommendation model in digital libraries

Yuangen Lai and Jianxun Zeng
*Institute of Scientific and Technical Information of China,
Beijing, China*

Received 29 August 2011
Revised 31 October 2011
Accepted 28 November 2011

Abstract

Purpose – The purpose of this paper is to develop a cross-language personalized recommendation model based on web log mining, which can recommend academic articles, in different languages, to users according to their demands.

Design/methodology/approach – The proposed model takes advantage of web log data archived in digital libraries and learns user profiles by means of integration analysis of a user's multiple online behaviors. Moreover, keyword translation was carried out to eliminate language dissimilarity between user and item profiles. Finally, article recommendation can be achieved using various existing algorithms.

Findings – The proposed model can recommend articles in different languages to users according to their demands, and the integration analysis of multiple online behaviors can help to better understand a user's interests.

Practical implications – This study has a significant implication for digital libraries in non-English countries, since English is the most popular language in current academic articles and it is a very common phenomenon for users in these countries to obtain literatures presented by more than one language. Furthermore, this approach is also useful for other text-based item recommendation systems.

Originality/value – A lot of research work has been done in the personalized recommendation area, but few works have discussed the recommendation problem under multiple linguistic circumstances. This paper deals with cross-language recommendation and, moreover, the proposed model puts forward an integration analysis method based on multiple online behaviors to understand users' interests, which can provide references for other recommendation systems in the digital age.

Keywords Digital libraries, User studies, Internet, Languages, Programming and algorithm theory, Recommendation system, Cross-language, Web mining, Personalized services

Paper type Research paper

1. Introduction

With the development of network technology, the amount of academic articles available to users increases exponentially in digital libraries. However, information overload has also caused problems more serious for users to obtain literatures that they are interested in. Many methods have already been proposed to help users more easily and quickly reach their desired target, among which recommendation systems have obtained a lot of attention and succeeded in many industries (Liao *et al.*, 2006).

A recommendation system is a class of software which helps users obtain the most suitable products according to their preferences, needs, or tastes (Rashid *et al.*, 2002; Martínez *et al.*, 2007). An entire recommendation system always includes the following steps:

The Electronic Library
Vol. 31 No. 3, 2013
pp. 264-277
© Emerald Group Publishing Limited
0264-0473
DOI 10.1108/EL-08-2011-0126

(1) Gather user's preference information and learn user profiles.

(2) Carry out item recommendation work using relevant algorithms which are commonly classified into content-based, collaborative and hybrid recommendation (Adomavicius and Tuzhilin, 2005).

In text-based item recommendation applications, both user and item profiles are always represented by keyword vectors (Semeraro *et al.*, 2007) and current recommendation systems assumes that all keywords in user profiles are in same language. Up to date, there is little literature concerning recommendation systems under multiple linguistic environments.

If users just want to obtain articles in a single language, current recommendation systems may work well. However, demands for academic articles in multiple languages become greater and greater along with the development of globalization, especially in non-English countries since English is the most popular language in current academic articles. In fact, digital libraries in many countries have already provided articles in multiple languages. Under such circumstances, recommendation systems based on a single language may not be sufficient for they cannot recommend articles in different languages and users have no choice to select items in the language that they are interested in. Furthermore, language chaos may occur in user's keyword vectors if the keywords are extracted from items previously seen or rated by the same user but these items are not in the same language. For example, it is very common for Chinese to get literature in Chinese as well as in English from digital libraries. Assuming a user has viewed two articles in one day (for example, one is in Chinese, and the other is in English) and the user profile is automatically computed by extracting keywords from articles that have been viewed, the profile of this user may have keywords in two languages at least. Obviously, such profiles are incomprehensible for any of the current recommendation systems. Therefore, it is necessary for text-based item recommendation systems of digital libraries to take account of the language factor, in particular those in non-English countries.

To provide recommendations under multi-linguistic environments, we present a framework of a cross-language personalized recommendation system to solve both information overload and the language barrier problem and to help users obtain their desired academic articles more effectively. Since the acquisition of user's preference information is very important for recommendation systems, the method of user profile computation has also been discussed in details in this paper.

The structure of this article is as following. A brief review of recommendation systems has been made in Section 2. In Section 3, a cross-language personalized recommendation model has been presented and key techniques of the model have been discussed. Then, discussions have been made in Section 4 and Section 5 presents conclusions and potential future research directions.

2. Related work

Recommendation systems emerged as an independent research area in the mid-1990s, when a paper on collaborative filtering appeared (Goldberg *et al.*, 1992). Since then, a lot of work has been done in both industry and academia on developing new approaches for recommendation systems because of the abundance of practical applications that help users deal with information overload and provide personalized recommendations to them. Examples of such applications include recommending books, CDs, movies (Miller *et al.* 2003), music (Yoshii *et al.*, 2008), tourisms (Sebastia *et al.*, 2008), and so on.Formally, recommendation systems can be formulated as follows (Adomavicius and Tuzhilin, 2005). Let U be the set of all users and S be the set of all possible items that can be recommended. Traditionally, the recommendation process starts with the specification of the initial set of ratings that is either explicitly provided by users or implicitly inferred by systems. Once these initial ratings are specified, recommendation systems can estimate the rating function R :

$$R : U \times S \rightarrow Ratings$$

for the (user, item) pairs that have not been rated yet by users.

After function R is estimated for the whole $U \times S$ domain, recommendation systems can select the item s'_u with the highest rating (or a set of k highest-rated items) for user u and recommend it(s) to him. More formally:

$$\forall u \in U, s'_u = \arg\max_{s \in S} R(u, s)$$

According to the way that recommendations are made, recommendation systems have been usually classified into three types, content-based, collaborative and hybrid. In content-based recommendation methods, users are recommended items similar to those that they preferred in the past (Basu *et al.*, 1998), while items that people with similar tastes and preferences liked in the past are recommended in collaborative recommendation methods (Sarwar *et al.*, 2000). Furthermore, content-based and collaborative methods can be combined into hybrid approaches through several different ways, which have been demonstrated to be capable of providing more accurate recommendations than pure content-based and collaborative approaches (Pazzani, 1999).

In text-based item recommendation applications, items are always represented by a set of keywords computed by extracting features from their content. For example, a content-based component of the Fab system represents web page content with the 100 most important words (Balabanovic and Shoham, 1997). More formally, let $Content(s)$ be an item profile. Similarly, let $ContentBased\Pr ofile(u)$ be the profile of user u containing his tastes and preferences, which are learned by analyzing the content of items previously seen or rated by user u and are usually computed by keyword analysis techniques from information retrieval (Daniela *et al.*, 2010). That is, $Content(s)$ and $ContentBased\Pr ofile(u)$ are both described by vector of keyword (or its weights). With further analysis, we can found that the underlying ideas of current text-based item recommendation systems are as follows:

- Profiles of both user and item are represented by the same language;
- items previously seen or rated by the same user are in the same language too.

If users only want to get articles in a single language from digital libraries, current text-based item recommendation systems could work well. However, if users need papers in different languages (this phenomenon is very popular for non-English users such as Chinese and Japanese), these two assumptions cannot be met and then recommendation systems based on a single language become powerless. First, these recommendation systems cannot recommend items in one language according to user profiles that are described in another language. Second, language chaos may occur in user profiles based on keywords that have been extracted from items previously seen or rated if these items are in different languages. For example, assuming a user has read two articles about information retrieval (for example, one is in English and the other in Chinese) and user profile is computed by extracting keywords from articles viewed, the keyword vector of this given user would consist of both Chinese and English terms. Obviously, it is difficult to obtain good practical results using such user profiles in recommendation systems. Consequently, users have to suffer from not only language barriers but also information overload problems for obtaining articles in other languages.

To address this issue, it is necessary to develop personalized recommendation systems under multi-linguistic environments. However, the language factor has seldom been considered in current researches about text-based item recommendation, although many studies have been carried out in this area. In this paper, we present a cross-language recommendation model in order to facilitate users obtaining academic articles in their desired language. Compared with cross-language information retrieval (CLIR), which deals with language barrier through providing multi-lingual retrieval results (Xu *et al.*, 2001), we aim at solving language barrier and information overload problem by recommending multi-lingual articles to users according to their preferences.

For simplicity, the recommendation model proposed in following sections will focus on the bilingual problem, which specifically refers to Chinese and English. The analogy of recommendation systems for other languages can be similarly made from this study.

3. Cross-language recommendation model

Before proceeding to model construction, we briefly list our goals for the cross-language recommendation system. First, we expect that our model can recommend articles in multiple languages according to user's preferences. Second, we try to build user profiles that are effective for article recommendation, even if languages of articles viewed by the same user are different. Third, we hope that our model can analyze user's interests in an implicit manner, since it is difficult to acquire explicit information of user's preferences.

3.1 Issues

In this sub-section, key issues for the cross-language recommendation task are discussed and our initial approaches to address these issues are also proposed.

B: *Data acquisition*. As we know, recommendation systems take advantage of users' preferences to help them beat their desired target. Thus, the more detailed a user's preference information is, the more effective recommendation systems can be. But in practice, it is difficult to get explicit rating information (Rodríguez *et al.*, 2010). Fortunately, each operation that users take on the web site of digital libraries has been automatically archived in the database, which provides an opportunity to learn user profiles from historic and current online behaviors and to make predictions about future needs and requirements. In this paper, we would analyze user's preferences on the basis of web log data.

Interest feature extraction. There are two characteristics of web log data archived in databases of digital libraries. First, the data quantity is overwhelming and it is unable to be digested by human analysts. Second, web log data has various types, such as accessing log and retrieval logs, and each type has its own fields and formats. Which type(s) of data should be selected to analyze a user's preferences? How to deal with this wealth data and extract feature from them? We choose multiple types of log data to analyze users' interests and learn user profiles, in which web mining has been used as the analyzing method.

User profile representation. In addition to feature extraction, another issue of user profile computation is its representation. There are many types of representation discussed in the literature (Middleton *et al.*, 2004), among which vector space model (VSM) is the most common method utilized in text-based item recommendation systems. Since our purpose is to achieve personalized recommendation in digital libraries and the candidate items to be recommended mainly refer to academic articles, we select VSM to represent user profiles.

Cross-language recommendation method. As mentioned above, this paper focuses on the bilingual problem and two languages are involved in our model. One is used to describe user profiles (called as native language for convenience), and the other is to represent academic articles (called as target language). Obviously, if both native language and target language are the same, the recommendation problem discussed above would turn into the study under a single language environment. From this perspective, the key step for cross-language recommendation is to eliminate language dissimilarity between user and item profiles, and we will choose translation of user profiles to achieve the goal.

In addition, there are many other issues in personalized recommendation systems, e.g. cold-start (Rodríguez *et al.*, 2010) and scalable problem (Takács *et al.*, 2009). In particular, we focus on the cross-language recommendation thesis and user profile computation in digital libraries.

3.2 Architecture description

Taking the above issues into consideration, we have developed a recommendation framework, as shown in Figure 1. The framework consists of three main phases, feature extraction, keyword translation and article recommendation. For the feature extraction phase, web log data have been input and the user's feature vector has been computed. Then, the vector is separately represented by native language and target language, which is the function of the second phase. Furthermore, recommendation work has been performed in the third phase through combining the feature vectors and corresponding article databases. In short, our main idea can be taken as to add a keyword translation component between user profile computation and item recommendation work, the purpose of which is to make full use of existing recommendation techniques by means of unifying language dissimilarity before item matching.

3.3 Phase 1: feature extraction

The first phase in Figure 1 is feature extraction, which aims to learn user profiles. As mentioned above, we take advantage of web log data to achieve the goal.

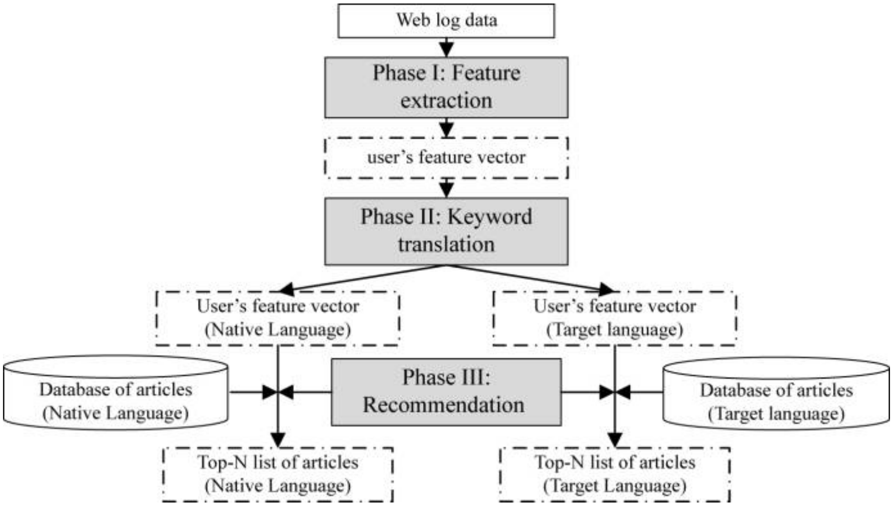
Web log data
Phase I: Feature extraction
user's feature vector
Phase II: Keyword translation
User's feature vector (Native Language)
User's feature vector (Target language)
Database of articles (Native Language)
Phase III: Recommendation
Database of articles (Target language)
Top-N list of articles (Native Language)
Top-N list of articles (Target Language)

Figure 1. Architecture of proposed model

3.3.1 Extraction steps. With the advent of information technology, networks have become the main channel for obtaining academic articles. After accessing the web site of digital libraries (it is always necessary for users to register with the web site at their first visit and simultaneously provide some demographic information), users query documents using search engines, and then acquire the desired articles. The whole process does not need any interference from librarians and is completely different from the traditional pattern.

Certainly, such a change has led to some modifications for acquisition of user's information. On one hand, the accuracy of user's demographic information (e.g. gender, age) that are presented by users themselves in the registration process is unwarranted due to the lack of effective validation and verification mechanism, resulting in the infeasibility to make recommendations on the basis of demographic information. On the other hand, every operation that users take on the web site is archived in the data warehouse of digital libraries, providing an excellent test-bed for us to understand user's preferences. Thus, we attempt to learn user profiles according to these online behaviors.

In summary, steps to obtain articles from digital libraries mainly include logging in, retrieving, browsing and downloading behaviors. Specifically, users can query the literatures, browse the secondary sources and download the full text from the web site of digital libraries. Generally, all behaviors are not absolutely necessary but each of them has its corresponding log data, as shown in Figure 2(a). For the querying behavior, retrieval log data is available, which contains information about querying time, querying sentence etc. In terms of browsing and downloading behaviors, their related information are also automatically logged, which have been named respectively as browsing log data and downloading log data in this paper. The former includes several fields, such as browsing time, user name and ID number of the secondary sources browsed. Similarly, the later consists of operation information and ID number of the downloaded articles. Since both of the ID numbers are unique, it is easy to obtain

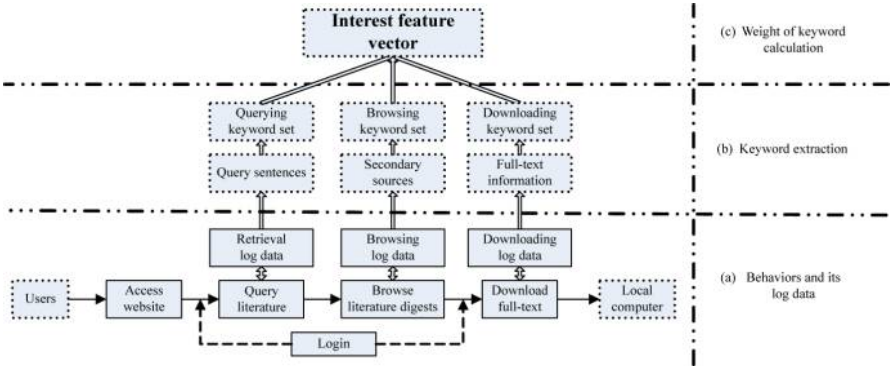

Figure 2.
Steps for obtaining articles and extracting feature

corresponding secondary source and full-text information from background databases of digital libraries.

Although querying, browsing or downloading behaviors can partly reflect user's interests in literatures, we suggest that comprehensive analysis of these three behaviors can help us better understand users' demands. Therefore, an integration method will be discussed in the following, which includes two steps, i.e. keyword extraction and weight calculation.

3.3.2 Keyword extraction. Keyword extraction in Figure 2(b) includes two parts. First, we parse querying sentences, secondary sources and full-text information from respective log data as mentioned above. Second, keywords are separated extracted from the results of the previous part. Since there are significant differences between querying sentence, secondary source and full-text information, different approaches have been adopted to fulfill this task. In respect of querying sentence, regular expression method is a good choice to extract querying terms which are directly thought as keywords in this paper. Turning to secondary source and full-text information, natural language processing and information retrieval technologies (Salton, 1989) are used and the steps are as follows. First, we adopted word segmentation techniques to produce a set of keywords for each source text, typically nouns and noun phrases. Second, the number of occurrences of each keyword is computed and those keywords with high occurrences would be selected to represent user's preferences. In order to learn the dynamic characteristics of user's interests, the occurring time of each behavior is also recorded in the extraction process. Finally, behavior information of each user is recorded in the operation table, in which user's name, behavioral type and occurring time of behaviors, and keywords related to each operation are included.

Table I illustrates how such a extraction works, in which samples of behavior records of two users have been given. The number following keywords in "keyword extracted" column indicates term frequency appeared in the related text information, representing keyword's importance for user's preferences. In particularly, A001 took a querying operation on Jan. 1 2011, and keywords extracted from related querying sentence include "信息检索" (meaning is information retrieval) and "文本挖掘" (meaning is text mining). Similarly, this user took a browsing behavior at the same day, and the extracted keywords include "信息检索" (meaning is information retrieval) and "排序" (meaning is sorting), etc. In this browsing operation, the term frequency of "信息检索"

(meaning is information retrieval) is bigger than that of “排序” (meaning is sorting), i.e. 8 vs 4 respectively, indicating that this user is more interested in “information retrieval” than in “sorting”.

3.3.3 Weight calculation. As mentioned above, querying, browsing or downloading behaviors can reflect a user’s interests to a certain extent as well as to a different degree. In order to comprehensively understand a user’s demands, it is better to adopt weight calculation for different behaviors instead of being simply accumulated or separately considered. Intuitively, keywords repeatedly appeared in all three behaviors can be considered to have more meanings for user’s interests and should be given higher weight than those only appeared in one operation. After weight calculation, the profile of each user is represented by feature vector $V(u)$:

$$V(u) = \{(K_1, \omega_1), (K_2, \omega_2) \cdots, (K_n, \omega_n)\}$$

where Kdenotes keyword, ω is its weight, and n represents the number of keywords.

In order to specify the keyword weights, both behavioral type and occurring time of each behavior have been considered in this paper. For type of behaviors, we argue arbitrarily that downloading behavior is more important for representing user’s preferences than querying and browsing behaviors. For occurring time of behaviors, we suggest that the more recently the behavior happens, the higher weight that related keywords should be set to. That is, a weighted sum for the given keyword is carried out as the following.

$$\omega_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} \tag{1}$$

where ω_i denotes the weight of keyword i, x_{1i} is the score of keyword i in querying operation, x_{2i} is the score in browsing operation, x_{3i} is the score in downloading operation, and α_1, α_2, α_3 are the weight of querying, browsing and downloading behaviors respectively ($\alpha_1 + \alpha_2 + \alpha_3 = 1$). For calculating the score of keyword i in different behaviors, the following model is adopted.

$$x_{ji} = \sum_k t_k f_{ji}^k, j = 1, 2, 3 \tag{2}$$

where k is the number of records including keyword i in operation table for the given user, t_k is the time factor for each operation and f_{ji}^k represents the keyword importance in each operation (In Table I, f_{ji}^k is represented by term frequency). As mentioned above, the more recently operation happens, the more significance it represents for user’s interests. Thus, t_k can be considered as a decreasing function:

User	Type of behaviors	Time of behaviors	Keywords extracted
A001	Querying	1 January 2011	信息检索（1）；文本挖掘（1）
A001	Browsing	1 January 2011	信息检索（8）；排序（4）；…
A001	Downloading	1 January 2011	信息检索（20）；算法（9）；…
A001	Querying	1 January 2011	信息检索（1）
A002	Querying	1 January 2011	信息检索（1）；协同过滤（1）
A002	Querying	1 January 2011	Information retrieval(1);Text mining(1)

Table I.
Example of operation table

$$t_k = \frac{1}{2^{t\_length}} \tag{3}$$

where *t_length* is time interval between occurring time of each behavior and the point of analytic time, which can be represented in months, quarters or years, etc.

Taking user A001 in Table I as an example, the following conditions have been assumed:

- The point of analytic time is 1 January 2011, and the unit of *t_length* is year.
- α_1, α_2, and α_3 are set to 0.3, 0.3 and 0.4.

Then, we can conclude the querying score of keyword "信息检索" (meaning is information retrieval) is 1.5, the browsing score is 8, and the downloading score is 20. Thus, its weight is $1.5{*}0.3 + 8{*}0.3 + 20{*}0.4 = 10.85$. Similarly, the weight for keyword "文本挖掘" (meaning is text mining), "排序" (meaning is sorting) and "算法" (meaning is algorithm) can be calculated in the same way, which is 0.3, 1.2 and 3.6, respectively. As a sequence, the profile of user A001 can be defined as follows.

$$V(A001) = \{(\text{信息检索},10.85),(\text{文本挖掘},0.3),(\text{排序},1.2),(\text{算法},3.6)\}$$

Using the same method, we can get the profile of user A002 from Table I:

$$V(A002) = \{(\text{信息检索},0.3),(\text{协同过滤},0.3),\\ (Information\ \mathrm{Re}\,trieval,0.15),(Text\ Mining,0.15)\}$$

3.4 Phase 11: keyword translation

In this phase, the feature vector derived from the previous phase is converted into two vectors respectively in target language and native language, as shown in Figure 1. That is, the output of this phase is two keyword vectors for each user, both of which describe user's interests. One is in target language, and the other in native language, which are expressed as:

$$V_c(u) = \{(K_{c1}, \omega_{c1}), (K_{c2}, \omega_{c2})\cdots, (K_{cm}, \omega_{cm})\}$$

$$V_f(u) = \{(K_{f1}, \omega_{f1}), (K_{f2}, \omega_{f2})\cdots, (K_{fp}, \omega_{fp})\}$$

where $V_c(u)$ is the vector in native language; $V_f(u)$ is that in target language; K and ω are similar to those of $V(u)$; m and p represent the number of keywords, respectively. The purpose of this conversion is to eliminate language dissimilarity between user profiles and academic articles to be recommended.

In order to translate feature vectors into a different language, there are several different ways, e.g. machine translation (MT), bilingual dictionary (Ballesteros and Croft, 1998) and a statistical model based on parallel corpus (Nie *et al.*, 1999). Since users may use querying keywords or read literatures in different languages, there exist three scenarios for keywords in vector $V(u)$:

1. all in native language;
2. all in target language;
3. hybrid language, i.e. part in native language and others in target language.

From this perspective, keyword translation in our study is bidirectional rather than unidirectional, and we choose bilingual dictionary to achieve the translation work.

As we know, semantic disambiguation is one of the main hurdles for improving translation effectiveness (Ballesteros and Croft, 1998). In our study, the following methods have been adopted to improve the conversion accuracy. First, we argue semantic relationships may occur among the keywords in feature vector of a given user. Therefore, co-occurrence statistical analysis is an optional way to make judgment for translation of keywords. Second, the text information related to user's online behaviors, including querying keywords, secondary sources browsed and full-text information downloaded of the same user, is used as corpus to eliminate keyword ambiguity when features vectors are in the hybrid language scenario. Taking user A001 in Table I as an example, there is a Chinese keyword “排序“ (meaning is sorting) in his feature vector, which has several possible English translations (e.g. ranking, ordering and sorting). If this user has used querying keywords or read documents in English and term “sorting” has higher term occurrence frequency in his behavior information than others, we consider it is the correct translation of this Chinese keyword.

Another problem that may occur after the translation of language hybrid vectors is the repeated occurrence of keywords. Taking user A002 in Table I as an example, both “information retrieval” and “信息检索” (meaning is information retrieval too) simultaneously appear in his feature vector since the user has taken querying behaviors using keywords in both languages. Then, two same keywords may exist in vector $V_c(A002)$ and $V_f(A002)$ after the translation process. We take a merger step to deal with this problem, which would add up all the weight of the same keyword. Thus, the feature vector of user A002 in native language and target language would be defined as follows.

$$V_c(A002)=\{(\text{信息检索},0.45),(\text{协同过滤},0.3),(\text{文本挖掘},0.15)\}$$

$$\begin{aligned}V_f(A002)=\{&(\mathit{Information}\ \mathrm{Re}\mathit{trieval},0.45),(\mathit{Collaborative\ Filtering},0.3),\\&(\mathit{Text\ Mining},0.15)\}\end{aligned}$$

3.5 Phase 111: article recommendation

Based on users' feature vectors in different languages, we can recommend articles in more than one language to them. In particular, we can recommend literature in native language taking advantage of user profiles represented in native language, and suggest items in target language using feature vectors in target language. Since both user profiles and items are in the same language, existing methods in text-based item recommendation systems can be adopted.

3.5.1 Content-based recommendation. As mentioned in the previous section, content-based recommendation is known to work well for recommending texts that informative content descriptors exist. Spontaneously, we can choose it to recommend academic articles in digital libraries. To adopt this algorithm, the work of article's feature representation needs to be completed in advance.

Feature representation begins with the parsing of each article to produce a set of features, typically nouns and noun phrases. Representative feature are sequentially selected from this set of extracted features. Then, the feature selection step has been taken using the popular method of *TF-IDF* (Salton and Buckley, 1988), and each term has been assigned a weight on the basis of its performance. As a consequence, the content of article *s* can be defined as:

$$Content(s) = (w_1, w_2, \cdots, w_k)$$

where w_k indicates the term weight. Since two languages are involved in our study, academic articles in both native language and target language should be represented respectively by vectors of term weights, which are expressed as follows.

$$Content(s_c) = (w_{c1}, w_{c2}, \cdots, w_{ck})$$

$$Content(s_f) = (w_{f1}, w_{f2}, \cdots, w_{fj})$$

Where $Content(s_c)$ is the feature representation of articles in native language; $Content(s_f)$ is that in target language.

Content-based methods recommend items similar to those that users liked in the past (Basu *et al.*, 1998). That is, the Top-N list of best-matching items are recommended according to the comparison between the candidate items and items previously rated by the user, and several techniques such as cosine similarity measure and Bayesian classifier can be used (Pazzani and Billsus, 1997). In this study, the rating function is defined as:

$$R_c(u, s) = score(V_c(u), Content(s_c))$$

$$R_f(u, s) = score(V_f(u), Content(s_f))$$

where $R_c(u, s)$ indicates the correlation between the user's preferences and the content of articles in native language; $R_f(u, s)$ indicates that in target language. Since both user profiles and academic articles are represented by vectors of keyword weights, cosine similarity measure is adopted to calculate the similarity in our study.

3.5.2 Collaborative filtering recommendation. Collaborative filtering recommendation is another algorithm that has been commonly used, based on similarities among the preferences of a group of users that are known as neighbors. According to user profiles in native language (or target language), we can find neighbors to each user, and the proximity measures frequently used include Pearson correlation, constrained Pearson correlation, Jaccard coefficient, etc. (Adomavicius and Tuzhilin, 2005). After neighborhood computing process, the suggestion items can be predicted.

It should be noted there are two special cases in which collaborative filtering method in our study cannot smoothly work. First, it is difficult to find neighbors for a given user. For example, all the similarities between the given user and others are less than the threshold. Second, although there are neighbors that have similar preferences, all of them have viewed articles in a same language. For example, assuming the neighbors of the particular user only viewed items in Chinese, it is impossible to recommend literature in English based on collaborative filtering algorithm. In these cases, the content-based recommendation method discussed in the previous sub-sector is adopted to make the recommendation. That is, academic articles would be recommended on the basis of the similarity between their content and user profiles.

4. Discussion

As more and more academic articles become available, it is hard to get desired papers efficiently from digital libraries due to information overload problems. As a consequence, recommendation systems have obtained a lot of attention and good results have been achieved in many practical applications. But so far, there are few documents concerning recommendation systems under multiple linguistic environments, and all text-based item recommendation systems assume items previously seen or rated by the same user are in same language and profiles of both user and item are represented by the same language too.

However, demands for articles in multiple languages is becoming ever greater with the development of globalization, especially in non-English countries. In order to get the desired literature in foreign languages, users have to suffer from not only a language barrier but also an information overload problem. Thus, it is quite necessary for digital libraries to develop new techniques to help users more effectively reach their target.

In this article, we have presented a cross-language recommendation model aiming at achieving the recommendation of suitable academic articles in multiple languages. We convert the bilingual recommendation problem into the study under a single language environment by means of eliminating language dissimilarity between user and item profiles, and then make full of existing recommendation techniques. Moreover, we have proposed a method to understand a user's interests in an implicit manner. Our proposed model has several advantages. First, it can recommend literature in one language to users according to their profiles described in another language. Second, it can deal with the language chaos phenomenon which may occur in a user's keyword vectors if items previously seen or rated by the same user are not in same language. This phenomenon has not been discussed in existing literatures though it is very common for users in non-English countries. Third, the integration analysis of multiple online behaviors can help to better understand a user's preferences.

As an initial research for recommendation systems under multi-lingual environments, many limitations still exist in our study. First, the method of integration analysis of various online behaviors is some distance from practical application, and its parameters are set out arbitrarily in this paper. The analysis of online behaviors can help to better understand users and their demands, but many questions still need support from other empirical studies since digital libraries have only appeared in recent decades. In reality, a user's anonymous access and the difference of a keyword's number among the behavior information are the two major unsolved issues. Second, the translation of keywords is a key technique for application of cross-language recommendation systems. Though many studies have been done and good results have been achieved in CLIR area (Ballesteros and Croft, 1998), the application effect of their translation methods on recommendation systems should be further studied. Moreover, the limitations of current recommendation systems (such as cold star, sparseness) have not discussed in this paper.

5. Conclusion and future work

On the basis of analysis of user demand for academic articles and current researches on personalized recommendation systems, we argue it is necessary to take account of language factors in recommendation systems, and thus we propose a cross-language recommendation model to achieve the recommendation of articles in multiple languages. Furthermore, the computation of user profiles in an implicit manner has been described in detail. A promising area of this study is digital libraries in non-English countries, since it is very common phenomenon for users in these countries to obtain literatures presented by more than one language. In the near future, we plan to pay more attention to the computation of user profiles in digital libraries and empirical research with large-scale practical data.

References

Adomavicius, G. and Tuzhilin, A. (2005), "Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17 No. 6, pp. 734-749.

Balabanovic, M. and Shoham, Y. (1997), "Fab: content-based, collaborative recommendation", *Communications of the ACM*, Vol. 40 No. 3, pp. 66-72.

Ballesteros, L. and Croft, W.B. (1998), "Resolving ambiguity for cross-language retrieval", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA*, pp. 64-71.

Basu, C., Hirsh, H. and Cohen, W. (1998), "Recommendation as classification: using social and content-based information in recommendation", *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, Menlo Park, CA, USA*, pp. 714-720.

Daniela, G., Silvia, S. and Analía, A. (2010), "Integrating user modeling approaches into a framework for recommender agents", *Internet Research: Electronic Networking Applications and Policy*, Vol. 20 No. 1, pp. 29-54.

Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. (1992), "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, Vol. 35 No. 12, pp. 61-70.

Liao, I-E., Liao, S.-C., Kao, K.-F. and Harn, I.-F. (2006), "A personal ontology model for library recommendation system", *Digital Libraries: Achievements, Challenges and Opportunities*, Vol. 4312, pp. 173-182.

Martínez, L., Pérez, L.G. and Barranco, M. (2007), "A multigranular linguistic content-based recommendation model", *International Journal of Intelligent Systems*, Vol. 22 No. 5, pp. 419-434.

Middleton, S.E., Shadbolt, N.R. and Roure, D.C.D. (2004), "Ontological user profiling in recommender systems", *ACM Transactions on Information Systems*, Vol. 22 No. 1, pp. 54-88.

Miller, B.N., Albert, I., Lam, S.K., Knostan, J.A. and Riedl, J. (2003), "MovieLens unplugged: experiences with an occasionally connected recommender system", *Proceedings of the 8th International Conference on Intelligent User Interfaces, Florida, USA*, pp. 263-266.

Nie, J.-Y., Simard, M., Isabelle, P. and Durand, R. (1999), "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA*, pp. 74-81.

Pazzani, M. and Billsus, D. (1997), "Learning and revising user profiles: the identification of interesting web sites", *Machine Learning*, Vol. 27 No. 3, pp. 313-331.

Pazzani, M.J. (1999), "A framework for collaborative, content-based and demographic filtering", *Artificial Intelligence Review*, Vol. 13 No. 5, pp. 393-408.

Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., Mcnee, S.M., Konstan, J.A. and Riedl, J. (2002), "Getting to know you: learning new user preferences in recommender system", *Proceedings of the 7th International Conference on Intelligent User Interfaces, New York, USA*, pp. 127-134.

Rodríguez, R.M., Espinilla, M., Sánchez, P.J. and Martínez-López, L. (2010), "Using linguistic incomplete preference relations to cold start recommendations", *Internet Research: Electronic Networking Applications and Policy*, Vol. 20 No. 3, pp. 296-315.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Boston, MA.

Salton, G. and Buckley, C. (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 25 No. 5, pp. 322-328.

Sarwar, B.M., Karypis, G., Konstan, J.A. and Riedl, J.T. (2000), "Application of dimensionality reduction in recommender system-a case study", *ACM 2000 KDD Workshop on Web Mining for e-commerce-Challenges and Opportunities, Boston, USA*, pp. 285-295.

Sebastia, L., Garcia, I., Onaindia, E. and Guzman, C. (2008), "e-Tourism: a tourist recommendation and planning application", *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, Washington, USA*, pp. 89-96.

Semeraro, G., Basile, P., Gemmis, M.D. and Lops, P. (2007), "Content-based recommendation services for personalized digital libraries", *Digital Libraries: Research and Development*, Vol. 4877, pp. 77-86.

Takács, G., Pilászy, I., Németh, B. and Tikk, D. (2009), "Scalable collaborative filtering approaches for large recommender systems", *Journal of Machine Learning Research*, Vol. 10, pp. 623-656.

Xu, J.X., Weischedal, R. and Nguyen, C. (2001), "Evaluating a probabilistic model for cross-lingual information retrieval", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA*, pp. 105-110.

Yoshii, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G. (2008), "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16 No. 2, pp. 435-447.

Further reading

Choi, S.-H., Jeong, Y.-S. and Jeong, M.-K. (2010), "A hybrid recommendation method with reduced data for large-scale application", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40 No. 5, pp. 557-566.

About the authors

Yuangen Lai is presently working as Assistant Professor in the Institute of Scientific and Technical Information of China. He obtained his PhD in Management Science from Beijing Institute of Technology. His research areas are knowledge organization, data mining and user research. Yuangen Lai is the corresponding author and can be contacted at: laiyg@istic.ac.cn

Jianxun Zeng is a Professor and the Director of the Information Resource Center in the Institute of Scientific and Technical Information of China. He obtained his PhD in Management Science from Wuhan University. His research areas are knowledge organization, information resource management and knowledge linking.

This article has been cited by:

1. Liang-Chu Chen Department of Information Management, National Defense University, Taipei, Taiwan Ting-Jung Yu Department of Information Management, National Defense University, Taipei, Taiwan Chi-Li Chang Department of Information Management, National Defense University, Taipei, Taiwan . 2015. TMTpedia: a case of extended Wikipedia for the military-based application in Taiwan. *The Electronic Library* **33**:3, 450-467. [Abstract] [Full Text] [PDF]
2. David Gunnarsson Lorentzen. 2014. Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics* **99**:2, 409-445. [CrossRef]