BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **15**, No 2

Sofia • 2015 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2015-0027

Time-Aware and Grey Incidence Theory Based User Interest Modeling for Document Recommendation

Shulin Cheng[1], *Yuejun Liu*[2]

[1] *School of Computer and Information, Anqing Normal University, Anqing 246133, Anhui, China*
[2] *School of Software Engineering, Anyang Normal University, Anyang 455000, Henan, China*
Emails: chengshL@aqtc.edu.cn lyj@aynu.edu.cn

Abstract: *Document recommendation involves the recommendation of documents similar to those that a user has preferred in the past. The Vector Space Model* (*VSM*) *is commonly adopted to denote the document objects and user interests. The user interests are extracted from the documents that a user has browsed. The interest degree of the user is calculated using the TF-IDF method, but the time factor is not considered. The recent documents that a user has browsed embody much more his/her interests. This study proposes a time-aware and grey incidence theory based user interest model to improve document recommendation. First, the time-aware user interest model is proposed based on the analysis of the user interests, document objects and user interest knowledge table. Second, a coefficient vector model of the user interest degree is designed using the grey incidence theory to differentiate the main from the minor user interests. The time-aware and grey incidence theory based user interest model is then exploited to produce document recommendations. Finally, the experiment and evaluation metrics are studied. The results show that the model proposed outperforms other related models and recommends more accurate documents to the users.*

Keywords: *Document recommendation, time-aware, grey incidence theory, vector space model, personalized information service.*

1. Introduction

The rapidly increasing amount of network information overwhelms the users and renders who find the desired information more difficult. The personalized information recommendation service based on interests [1] is a better solution to this problem. Some distinguished recommendation models and methods [2-4] have been proposed, such as rule-based, collaborative filtering, content-based and hybrid recommendation. Among these methods, collaborative filtering recommendation

performs most efficiently in recommending movies, music and books. Content-based recommendation performs well in recommending news and documents. The user interests are important in content-based recommendation, which is modeled on interests. Document recommendation is a widely used classical content-based recommendation.

A user tends to browse information that he/she is interested in [5] and reads such information for a relatively long average browsing time. His/her browsing behaviors and information reflect his/her interests and background knowledge, which are often used to model his/her interests. In a given period of time, the users may be interested in many topics from various fields. According to our experience, two factors are very important to user interests. One is the time factor. The recent information browsed by a user is more representative of his/her interests. In recommending documents, the recent information should be given much more consideration than the old information. The other factor is the interest importance. The importance of each interest varies for every user. Interests with high importance are considered main interests, whereas interests with low importance are considered minor or incidental interests. As such, when a document is recommended to a user, the interest importance should also be taken into account. In this study we address the issues of document recommendation from the perspectives of these two factors, which can be utilized to substantially improve the performance of recommendation.

1.1. Related work

Document recommendation is widely used in the fields of information retrieval [6], which generates recommendations using keywords. Document recommendation belongs to content-based recommendation. In the field of personalized recommendation, content-based recommendation, which performs well in recommending items that contain textual information, is described through keywords and incorporates a user profile. A user profile contains information about the user's tastes, preferences and needs [4]. Both items (i.e., herein a document is represented as an item) and the user profile are mostly described through the Vector Space Model (VSM) and contain a group of keywords, respectively. The keywords of the items and the user profile are used to characterize the main contents of the items and the user interests that are automatically extracted from such items. The method of content-based recommendation computes the similarity between a user profile vector and an item vector and produces top-N recommendations in a descending order of similarity [2, 8]. Content-based recommendation is applied not only in news and document recommendations [9] but also in social recommendations [3, 4, 8]. To describe documents and extract user interests, the user's browsing behaviors and information, such as click and browsing time, must be analyzed [5]. The calculation methods of VSM weight mainly include TF and TF-IDF, which are described in detail in [2, 4]. TF-IDF is better than TF, which represents a generic usage in VSM, but it is sometimes still not accurate enough because it fails to consider the real interest degree of the users. For example, suppose that a user has browsed 100 valid documents, and 10 terms denoting

his/her interests are obtained, then the weights of the user's 10 interests in VSM can be calculated using TF-IDF in these 100 documents. These values of weights represent the relevance between the user interest terms and documents and involves discrimination among the documents in the corpus, but do not exactly indicate the importance degree of each interest term. The TF-IDF value alone is insufficient in denoting the user interest weight. Each user's interest term is related to the documents and affects the user interest degree of the documents. If the total interest degree of the documents is obtained, rather than the association degree of each term in the document, and the weight of each interest of the users (i.e., the interest term's weight in the user interest vector) is computed by modeling between the interest terms' TF-IDF values and the users' total interest degree of the documents, then we can divide the user's main interests in a descending order of the terms' weights.

Exploiting the time factor has proven to be an efficient approach to improve the performance of recommendation [10], as shown in, for example, the Netflix Prize competition. Time-aware recommender systems are indeed receiving increasing attention. Time is regarded as a kind of context, and the user interests depend on the time context [11]. User interests for items tend to change over time, and modeling of the user interests should consider the time factor [12]. Tracking the user interests over time is significant in making timely recommendations [13]. Applying the time factor in a recommender system mainly focuses on the collaborative filtering recommendation [10, 12, 13]. To our knowledge, no study has been yet conducted on the application of the time factor in document recommendation. That is why we consider the time factor to improve the performance of document recommendation.

This study is built on the prior work of content-based recommendation. We consider particularly the user interest importance and the time factor to improve the performance of document recommendation.

1.2. Organization

The rest of the paper is organized as follows: Section 2 provides the related definitions and representations of the user interests and document objects. Section 3 describes the user interest knowledge table. Section 4 demonstrates the time-aware user interest model. Section 5 presents the model of the coefficient vector of the user interest degree, which is designed, using the grey incidence theory. Section 6 presents the recommendation model based on time-aware and grey incidence theory based user interest model. Section 7 compares the experimental evaluation of the proposed model and that of other relevant models. Section 8 makes some concluding remarks and suggests directions for future research.

2. User interests and document objects

2.1. Obtaining interests

The user interests are the basis of recommendations and can be obtained using two main methods: the explicit and implicit methods. The explicit method requires the users to interact with the system, such as submitting his/her hobbies or feedback in

the form of evaluations or ratings of items (i.e., in this case, web pages or documents). Given the interference of the normal browsing behaviors, the explicit method has been rarely used. The implicit method, which is currently the most widely used method, automatically elicits user's interests from his/her browsing behaviors and contents, including web addresses, time of opening and closing pages, and behaviors of saving or printing web pages. The information is saved in web log files. By analyzing these log files and the features of the documents the user has browsed, his/her interests can be extracted.

2.2. Interest representation

The user's interests are represented by several structures, such as keywords [2], VSM [14, 15], semantic ontology [16] and user-item rating matrix [17]. Each of these has its own strength and weakness. In this study, we use VSM, which was the first to be applied in information retrieval [15]. The key point of VSM is to calculate the weights of the terms. The generic methods of calculating the weight of the terms include TF, TF-IDF and the user input. However, they have some obvious disadvantages, as described in [14], such as low accuracy. Here, the calculation method of VSM is improved by considering the contributions of the time factor and the user interest importance, as described in Sections 4 and 5, respectively. To facilitate the computation of the weight of the terms, *n* feature terms are first extracted from the documents that a user has browsed using Stanford Word Segmenter [27]. Stop words are removed from these feature terms, and the rest are used to denote the user's interests, which are defined below.

Definition 2.1. Let $I = \{I_1, I_2, I_3, ..., I_n\}$, I_i represents the *i*-th interest of the user, $\mathrm{NUIV}=(\mathrm{UIV}, \mathrm{IW})$ is the user interest weighted vector tuple, UIV is the time-aware user interest vector, $\mathrm{UIV} = (\mathrm{uiv}(I_1), \mathrm{uiv}(I_2), \mathrm{uiv}(I_3), ..., \mathrm{uiv}(I_n))$; $\mathrm{uiv}(I_i)$ represents the *i*-th interest degree; IW is the coefficient vector of the user interest degree, $\mathrm{IW} = (\mathrm{iw}_1, \mathrm{iw}_2, \mathrm{iw}_3, ..., \mathrm{iw}_n)$, iw_i denotes *i*-th interest importance degree coefficient, which is computed using the method discussed in Section 5 and normalized as $\sum_{i=1}^{n} \mathrm{iw}_i = 1$.

2.3. Document object representation

The document object, namely, the web page, is represented by a two-tuple consisting of a document vector and the user interest degree of document. It is defined as follows:

Definition 2.2. Let D_j be the document object. $D_j = \langle d_j, \mathrm{pid}_j \rangle$, d_j denotes the document vector, pid_j denotes the user total interest degree of the document, $d_j = (u(d_{j1}), u(d_{j2}), ..., u(d_{jm}))$ and $u(d_{jp})$ represents the association degree between the *j*-th document and the *p*-th feature term.

3. User interest knowledge table

3.1. Structure of the user interest knowledge table

Definition 3.1. User interest knowledge table *S*. Let U be the user interest domain, which is a non-empty finite set of interest terms, T be the cluster of equivalence relations in U, which consists of interest attributes, V be the value domain of the equivalence relations, and *f* be a mapping function, which is denoted by $f: U \times T \to V$, then $S = \langle U, T, V, f \rangle$.

The user interest knowledge table is represented by a two-dimensional information table as shown in Table 1. The document objects and interest attributes are listed in the rows and columns, respectively. The last column denotes the user interest degree of the document.

Table 1. User interest knowledge table

Document	I_1	I_2	I_3	...	I_n	PID
d_1	$u(d_{11})$	$u(d_{12})$	$u(d_{13})$	...	$u(d_{1n})$	pid_1
d_2	$u(d_{21})$	$u(d_{22})$	$u(d_{23})$	...	$u(d_{2n})$	pid_2
...	...	...	...	...	...	...
d_m	$u(d_{m1})$	$u(d_{m2})$	$u(d_{m3})$	...	$u(d_{mn})$	pid_m

In Table 1 the row of d_i is the *i*-th document object, the column of I_j is the *j*-th user's interest, $u(d_{ij})$ represents the association degree between the *j*-th interest and *i*-th document, and pid_i represents the user interest degree of document d_i.

Definition 3.2. Document set *D*. D_i represents an efficient document object. D consists of the efficient documents that the user has browsed in a period of time and is defined as $D = \{D_i \mid i = 1, 2, 3, \cdots, m\}$.

3.2. Association degree between the document and user interests

Assume that the user has *n* interests, which are some feature terms derived from *m* documents that he/she has browsed in a period of time. The user's *n* interests are a subset of the collection of all the terms in these documents. For each document, we can compute the TF-IDF [18] value of a term in a corpus, denoted by $u(d_{ij})$ in Table 1. This is also called the association degree between the document and user interest term. TF-IDF is a classical algorithm used to compute the relevance of a term in a document. The formula of computing the degree of user interest I_j in a document d_i is denoted as follows:

$$u(d_{ij}) = \text{TF\_IDF}_{ij} = \frac{f_{ij}}{\max\{f_{i1}, f_{i2}, ..., f_{i,|V|}\}} \times \lg \frac{m}{df_j}. \quad (1)$$

The details of the parameters are referred to in [18]. TF-IDF can efficiently reduce the weights of the terms with high frequencies in Thesaurus and emphasize the weights of the important terms.

3.3. User interest degree of the document

The user interest degree of the document refers to the total interest degree of the document that the user is interested in. In this study the implicit method is used to elicit the user interests of the document from his/her browsing behaviors. The main user browsing behaviors include saving, printing, collecting, reading and paging. The most expressive behaviors among them are saving, printing and reading. We assume that the behaviors of both saving and printing mean that the user is interested in the documents with the maximum interest degree and pid_i is given the max reading time, MaxTime (i.e., an empirically statistical value), which means that the user interest degree of the document equals 1. Otherwise, the user interest degree of the document is measured by the average reading time of the document. As such, the user interest degree of the document is defined as

$$\text{pid}_i = \begin{cases} \text{MaxTime} & \text{when saved or printed,} \\ \dfrac{1}{n_i}\sum_{j=1}^{n_i} t_{ij} & \text{else average reading time,} \end{cases} \tag{2}$$

where pid_i represents the user interest degree of document d_i, n_i denotes the total page number of document d_i, and t_{ij} is the reading time of the *j*-th page of document d_i. To eliminate the noise data of a very short reading time, such as the time of opening the documents and closing them right after that, a threshold t_ρ is set. It is the lower bound of the reading time and is a statistical value. If $t_{ij} < t_\rho$, then the reading time equals 0. If the user saves or prints the document, then pid_i is set to MaxTime. Otherwise, pid_i equals the average reading time. We take the assumption that the user has browsed *m* documents. According to (2), we can calculate the *m* values of pid_i, which in the last are normalized to produce the user real interest degrees of the documents.

4. Time-aware user interests modeling

In traditional VSM, the value of the elements is computed using the TF-IDF method. From Definition 2.1, the value of an element in the vector of $\text{UIV} = (\text{uiv}(I_1), \text{uiv}(I_2), \text{uiv}(I_3), ..., \text{uiv}(I_n))$ is determined by TF-IDF. As a simple example, if one user has browsed m documents in a period of time, as shown in Table 1, $\text{uiv}(I_j)$ equals the summation of $u(d_{ij})$, that is, $\text{uiv}(I_j) = \sum_{i=1}^{m} u(d_{ij})$. However, the recent documents that a user has browsed mostly embody his/her interests, and the documents that he/she has browsed relatively long ago represent less his/her interests. Thus, when computing the value of $\text{uiv}(I_j)$, we should consider the influence of time. That is to say, the element of $u(d_{ij})$ should be

multiplied by a time function $f(t_i)$ [19]. The recent data are then assigned a greater level of importance, and the time function is designed as a decay function

(3) $$f(t_i) = e^{-\lambda t_i},$$

Where: t_i is the time of browsing the document d_i, i = 1, 2, …, m; t_1 represents the latest time; t_m denotes the oldest time, λ is the decay rate defined as $\lambda = \frac{1}{T_0}$; T_0 is the half-life parameter;

(4) $$u'(d_{ij}) = u(d_{ij}) * f(t_i).$$

The new time-aware user interest knowledge table is given as Table 2, which is used to calculate the value of each element of UIV . Table 2 is similar to Table 1, but the last of column of PID is omitted, given that it is not used in modeling the user interests.

Table 2. Time-aware user interest knowledge table

Document	I_1	I_2	I_3	…	I_n
d_1	$u'(d_{11})$	$u'(d_{12})$	$u'(d_{13})$	…	$u'(d_{1n})$
d_2	$u'(d_{21})$	$u'(d_{22})$	$u'(d_{23})$	…	$u'(d_{2n})$
…	…	…	…	…	…
d_m	$u'(d_{m1})$	$u'(d_{m2})$	$u'(d_{m3})$	…	$u'(d_{mn})$

Thus, each degree of interest I_j can be calculated in line with Table 2:

(5) $$\text{uiv}(I_j) = \sum_{i=1}^{m} u'(d_{ij}) = \sum_{i=1}^{m} u(d_{ij}) * f(t_i).$$

5. Modeling the coefficient vector of the user interest degree

From Definition 2.1, IW is the coefficient vector of the user interest degree. Each element of IW represents one interest of the importance degree coefficient. In fact, the interests with a high value of coefficients are the main interests; otherwise, they are minor interests. In this section the coefficient vector of the user interest degree is modeled. First, the meaning of the user interest contribution, which characterizes the user interest importance, is explained and analyzed. Second, the basic grey incidence theory is introduced and two main incidence degree models of calculating the User Interest Contribution Degree (UICD) are proposed. Lastly, the coefficient vector is modeled according to UICD.

5.1. Analysis of UICD

In most of the cases, when the users browse documents, they are inclined to browse the documents they are interested in. The degrees of the user interest in documents are different, which means that the user interests vary in each document. The document can be characterized by a set of terms. These terms determine the user interest degree of the document. Suppose that we obtain the user interest knowledge

table S in a period of time, in which the user interest I_j is regarded as a factor sequence (real-valued data set), and PID, the user interest degree of the document, is regarded as a reference sequence (real-valued data set). An intrinsic relationship holds between I_j and PID. In other words, the user interest degree pid_i of document d_i is determined by the multiple factor sequences I_j, and the contribution of each I_j to pid_i is different. For each d_i, the TF-IDF value of I_j, that is, $u(d_{ij})$ in S, is computed by (1), and the user's interest degree pid_i is computed by (2). Thus, in a period of time, table S consists of n factor sequences I_j and one reference sequence PID. Each I_j has some contribution to PID, that is, the UICD of I_j.

Now suppose that the users have browsed *n* documents, which form the user interest knowledge table S, in which the contribution of I_j to PID indicates the user interest contribution. The computation of each interest contribution degree of every user is converted into calculation of the association degree of the factor sequence to the reference sequence. Thus, the association degree is the importance degree of user interest I_j in the set of I. We find that they are different geometric curves (i.e., more strictly, zigzagged lines) based on the geometric characteristics of the factor and reference sequences. Calculating the importance degree of user interest I_j entails calculating the curve correlation between them, and grey incidence analysis theory can efficiently solve this kind of a problem [20, 21].

5.2. Basic grey incidence theory

The grey system theory, which was proposed by the well-known Chinese professor Mr. Deng Julong in the 1980-ies [22], has been widely applied [23, 24]. Grey incidence analysis theory is one of the important branches of the grey system theory. Its basic task is to analyze and determine the impact of factors based on the geometrical nearness or similarity between sequences in a micro or macro perspective. The principle of the grey incidence theory is to study the similarity between sequences according to the development trends of the geometric curves corresponding to the sequences. The observation values of each sequence are connected one by one and form a zigzagged line. The association degree measure model is built to calculate the grey association degree between the zigzagged lines in accordance with the geometric characteristics of the sequences.

Let X be the feature sequence set, X_i be the feature/factor sequence, and X_0 be the reference sequence, denoted as

$$X = \{X_i \mid i \in I = \{1,\ 2,\ 3,..., n\},\ X_i = (x_i(i),\ x_i(2),..., x_i(m))\},$$

$$X_0 = (x_0(1),\ x_0(2),..., x_0(m)),$$

where X_i, X_0 consist of *m* data points, respectively.

First, the grey association coefficient $r(x_i(k),\ x_0(k))$ [22] of the data point pair $(x_i(k),\ x_0(k))$ is calculated, denoted as

$$r(x_i(k), x_0(k)) = \frac{\min_i \min_k \Delta_i(k) + \zeta \max_i \max_k \Delta_i(k)}{\Delta_i(k) + \zeta \max_i \max_k \Delta_i(k)}, \tag{6}$$

where $\Delta_i(k) = | x_0(k) - x_i(k) |, k = 1, 2, 3, ..., n,$ ζ is the discrimination coefficient, and $\zeta \in [0, 1]$ empirically equals some value from 0.1 up to 0.5. The grey association degree between the factor and reference sequences is denoted as

$$r(X_i, X_0) = \frac{1}{m} \sum_{k=1}^{m} r(x_i(k), x_0(k)), \tag{7}$$

where $r(X_i, X_0)$ is better in some cases, but sometimes it is not so and has serious flaws [24, 25] because it is influenced by the absolute differences in minimum and maximum. For instance, if the extremum occurs in sequences, the grey association coefficient is affected, since the grey association degree is. Moreover, $r(x_i(k), x_0(k))$ is also affected by ζ, which is an empirical value. For the same question, $r(x_i(k), x_0(k))$ varies with different ζ. However, the case of $r_1 < r_2$ and $r_1 > r_2$ may possibly co-occur according to different ζ. Thus, it must be improved in practical specific applications.

Given its development through many years, the grey incidence theory covers two incidence degrees: the Grey Closeness Incidence Degree (GCID) and the Grey Similarity Incidence Degree (GSID) [20, 21, 24], which have been improved by some scholars. We apply the two incidence degrees in analyzing and calculating the UICD.

5.3. GCID

The proximity between the factor and reference sequences is generally called the closeness or nearness of the sequences in a grey system. In the user interest knowledge table S in our study, the factor sequences are I_j representing the user's interests, and the reference sequence is PID denoting the users' interest degree of the document. Thus, the GCID in our research is the proximity between the user interest sequence and the interest degree sequence. The two kinds of sequences are denoted by two zigzagged lines according to their geometric features. Each zigzagged line and coordinate axes form an area. The absolute value of the algebraic difference between the two areas of the two zigzagged lines showing a close relationship between them. The smaller the difference, the greater the proximity is, and vice versa. Before calculating GCID, some related notations are given below.

X_i: the *i*-th user interest sequence, that is, the factor sequence, a zigzagged line in geometry, $X_i = (x_i(1), x_i(2), ..., x_i(m))$ corresponding to I_i.

X_0: the users' interest degree sequence, that is, the reference sequence, also a zigzagged line in geometry, $X_0 = (x_0(1), x_0(2), ..., x_0(m))$, whose meaning is the

same as PID, which is not directly used to facilitate reasoning and obtain a similar representation to X_i.

S_{xi}: the area surrounded by the axes and a zigzagged line X_i.

S_{x0}: the area surrounded by the axes and a zigzagged line X_0.

Definition 4.1. Let ρ_i be the GCID of the sequences, X_i be the user interest factor sequence, PID be the user interest degree of the reference sequence, S_i be the absolute difference value of the area of X_i subtracted from the area of PID, then ρ_i is denoted as

$$\rho_i = \frac{1}{1+S_i}, \tag{8}$$

$$S_i = |S_{xi} - S_{x0}| = \left|\sum_{k=1}^{m}(x_i(k) - x_0(k)) - \frac{1}{2}(x_i(1) - x_0(1) + x_i(m) - x_0(m))\right|. \tag{9}$$

The calculation of the area includes three situations in accordance with the shapes and trends of X_i and X_0: (1) the zigzagged lines of X_i and X_0 represent the same trends; that is, they are either increasing or decreasing and do not intersect; (2) the zigzagged lines of X_i and X_0 are vibrating and do not intersect; (3) the zigzagged lines of X_i and X_0 are vibrating and intersect. The area in the three cases can be calculated using (9) according to the related mathematical knowledge. The proof of (9) is given in [21]. Thus, the final calculation formula of GCID is represented in line with Equations (8) and (9) and is given by

$$\rho_i = \frac{1}{1+\left|\sum_{k=1}^{m}(x_i(k) - x_0(k)) - \frac{1}{2}(x_i(1) - x_0(1) + x_i(m) - x_0(m))\right|}, \tag{10}$$

ρ_i has three characteristics of the incidence degree: normativity, pair symmetry and nearness [21].

5.4. GSID

GCID is defined by Equations (8) and (10) from the visual angle of proximity, whereas GSID is defined by the shape similarity of the two zigzagged lines representing the users' interest sequence and the interest degree sequence. The similarity degree describes the geometric similarity between the user interest sequence zigzagged line and the interest degree sequence zigzagged line. According to [21], the formula of GSID is given by

$$\varepsilon_i = \frac{1}{1+\left|\sum_{k=2}^{m-1}(x_i'(k) - x_0'(k)) + \frac{1}{2}(x_i'(m) - x_0'(m))\right|}, \tag{11}$$

where $x_i'(k) = x_i(k) - x_i(1)$, $x_0'(k) = x_0(k) - x_0(1)$. Similar to GCID, GSID has three characteristics: normativity, pair symmetry and nearness [21].

5.5. UICD

Analyzing the intrinsic nature of the user's browsing behaviors, we conclude that the user interest degree pid_i of document d_i is determined by the contents of d_i and that the contents are characterized by n terms, each of which has some contribution to pid_i. The contribution magnitude is related not only to the closeness but also to the similarity of the zigzagged lines between I_i and pid_i. Representing UICD by only either GCID or GSID is not appropriate. For instance, assume that there three factors which are X_0, X_1 and X_2. X_1 and X_2 denote the factor sequences and X_0 represents the reference sequence. We then obtain S_{x1}, S_{x2}, and S_{x0} according to Subsection 5.3. Thus, $S_1 = |S_{x1} - S_{x0}|$ and $S_2 = |S_{x2} - S_{x0}|$. Possibly, $S_1 = S_2$, that is, GCID ρ_1 equals GCID ρ_2, but X_1 and X_2 show different geometries. Consequently in this case, GSID and GCID, and not only GCID, should be taken into account. We observe that the less the area difference and the higher the similarity is, the greater the contribution of I_i to pid_i is, that is, a greater UICD. In view of GCID and GSID, UICD is denoted as

$$\mathrm{iw}_i = \alpha\rho_i + \beta\varepsilon_i, \tag{12}$$

where α and β are the weight coefficients ranging from 0 up to 1 and $\alpha + \beta = 1$, iw_i belongs to (0, 1] corresponding to I_i, forming the contribution degree vector, that is, the user interest importance degree vector being $\mathrm{IW} = (\mathrm{iw}_1, \mathrm{iw}_2, ..., \mathrm{iw}_n)$.

5.6. Coefficient vector of the user interest degree

The coefficient vector of the user interest degree can be measured by UICD. A group of UICDs is calculated by Equation (12) and normalized and ranked in a descending order, denoted by

$$\mathrm{IW} = (\mathrm{iw}_1 / \sum \mathrm{iw}_j, \mathrm{iw}_2 / \sum \mathrm{iw}_j, ..., \mathrm{iw}_n / \sum \mathrm{iw}_j) = (\mathrm{iw}_1', \mathrm{iw}_2', ..., \mathrm{iw}_j, ..., \mathrm{iw}_n'), \tag{13}$$

where $\mathrm{iw}'_j \geq \mathrm{iw}'_{j+1}$. The user's main interests are selected according to the principle of top-*K* or setting a threshold λ.

6. Personalized document recommendation

6.1 Recommendation based on Conventional VSM

In the content-based recommender system of the conventional VSM (CVSM), the user interest vector is built through the simple summation of all the document vectors that the user has browsed. In other words, the summation of the TF-IDF values is regarded as the final vector element of the user interests.

Before new documents are recommended, the similarities between the user interest vector and the document vectors are calculated and ranked in a descending order. The top-*K* relevant documents are recommended to the users. Let $\mathrm{UIV}_C = (\mathrm{uiv}_C(I_1), \mathrm{uiv}_C(I_2), \mathrm{uiv}_C(I_3), ..., \mathrm{uiv}_C(I_n))$ be the user interest vector in

CVSM, $\text{uiv}_C(I_j) = \sum_{i=1}^{m} u(d_{ij})$, then the similarity between UIV_C and d_i is denoted as

$$\text{Sim}_C(\text{UIV}_C, d_i) = \frac{\text{UIV}_C \cdot d_i}{\| \text{UIV}_C \| \times \| d_i \|}. \tag{14}$$

6.2. Recommendation based on the coefficient vector model

Obviously, CVSM, which uses TF-IDF to compute the vector weights, is not sufficient to produce good recommendation. The fact is that each interest is not always in the same position for the users in a period of time. That is to say, some interests are important, others are not so important. As such, the interest importance degree varies, which can be seen from the coefficient vector model of the user interest degree. In the coefficient vector, the high coefficient values correspond to major interests, and the low coefficient values correspond to minor interests. Thus, before new documents are recommended, the user interest importance degree must be considered. For convenience we call the coefficient vector model of the user interest degree UCVM.

Let $\text{UIV}_M = (\text{uiv}_M(I_1), \text{uiv}_M(I_2), ..., \text{uiv}_M(I_n)))$ be the user interest vector incorporating the user interest importance degree in UCVM, $\text{uiv}_M(I_j) = \sum_{i=1}^{m} u(d_{ij}) \times \text{iw}_j$, iw_j represents the user's *j*-th interest importance degree from the coefficient vector of the user interest degree, so that the similarity between UIV_M and d_i is denoted as follows

$$\text{Sim}_M(\text{UIV}_M, d_i) = \frac{\text{UIV}_M \cdot d_i}{\| \text{UIV}_M \| \times \| d_i \|}. \tag{15}$$

6.3. Recommendation based on the time-aware and coefficient vector model

The aforementioned analysis shows that the user interests are related not only to the importance degree of interest but also to the time factor. Thus, before new documents are recommended, these two aspects should be taken into account at the same time. For convenience, we call the time-aware and coefficient vector model of the user interest degree TCVM. According to Definition 2.1, the user interest vector is $\text{UIV} = (\text{uiv}(I_1), \text{uiv}(I_2), \text{uiv}(I_3), ..., \text{uiv}(I_n))$, where $\text{uiv}(I_j) = \sum_{i=1}^{m} u(d_{ij}) \times f(t_i)$ and incorporating the user interest importance degree, $\text{uiv}(I_j)$ is converted into $\text{uiv}_{\text{TM}}(I_j) = \text{iw}_j \times \sum_{i=1}^{m} u(d_{ij}) \times f(t_i)$. Let UIV_{TM} represents UIV, the similarity between UIV_{TM} and d_i is denoted by

$$\text{Sim}_{\text{TM}}(\text{UIV}_{\text{TM}}, d_i) = \frac{\text{UIV}_{\text{TM}} \cdot d_i}{\| \text{UIV}_{\text{TM}} \| \times \| d_i \|}. \tag{16}$$

7. Experiment

7.1. Trial system and dataset

To validate our proposed model TCVM and the related models discussed in Section 6, we developed a trial system called Information Recommendation System Based on User Interest Models (IR-UIM), which consists of a client side and server side. The client side is responsible for keeping track of the users' browsing behaviors, including saving, printing and average browsing time. The server side is responsible for crawling information, storing information, generating recommendations and delivering them to the client side. The programs of the three models, CVSM, UCVM and TCVM, are designed in the server side. The users read the information from the client side. The crawler downloads the information from some specific web sites for storage in the database. Before formally using IR-UIM, 1000 specific pages in the latest month are downloaded as basic information.

We selected 15 students as users of IR-UIM and divided them into three groups, with five users per group. These three groups corresponded to the three models – CVSM, UCVM and TCVM, and they used the IR-UIM, which recorded the browsing logs in real time. At first, each user browsed the information randomly. Then, after some browsing behaviors were captured, IR-UIM recommended information based on the corresponding recommendation model to each group of users is obtained. The crawler grabbed the 200 latest web pages every day. The experiment lasted for a month. The collected 6000 web pages and the former 1000 pages were both stored in the database. Each student corresponded to a user who was required to use the IR-UIM at least once a day. The IR-UIM recommended 40 web pages to the user every time according to his/her interests.

7.2. Evaluation metrics

In the experiment, in order to compare the performance of the three models, we used the evaluation metrics of precision and recall, which are the most popular metrics for evaluating recommender systems [26]. They are defined as follows:

$$P = \frac{N_A}{N_A + N_B}, \tag{17}$$

$$R = \frac{N_A}{N_A + N_C}, \tag{18}$$

where P denotes precision and R represents recall. The meanings of N_A, N_B, and N_C are described in Table 3. Precision represents the probability of a recommended item being relevant. Recall represents the probability of a relevant item being recommended.

Table 3. The meanings of N_A, N_B, N_C and N_D

Recommendation	Relevant	Not Relevant
Recommended	N_A	N_B
Not Recommended	N_C	N_D

We also define another novel evaluation indicator, called the Ordinal Hit Precision (OHP), which is also used to measure the performances of the recommendation. OHP is used to measure the precision of the users reading the pages according to the recommended order of the system.

$$\text{OHP} = \frac{1}{\phi(K)} \sum_{j=1}^{K} r_{K-j+1}, \quad r_j \in \varphi, \tag{19}$$

where K is the number of top-K recommended documents, $\phi(K) = \sum_{i=1}^{K} r_i$, r_i is the ordinal of the recommended documents, and φ is the set of recommended documents that a user has browsed. For example, assuming that we recommend 4 pages ranked by 1, 2, 3, 4. However, the user is interested in pages 1 and 3 and reads them. We can then compute the ordinal position importance as OHP, that is, OHP = (4+2)/10 = 0.6.

7.3. Results and evaluation

We present our experimental results for applying the three models in recommending pages. The results are shown in the following three figures, which correspond to the average precision, average recall and average OHP of the three groups. For instance, each value of the average precision equals the mean precision value of the five users in each group in one day, as do the average recall and average OHP.

For the first time that the users apply IR-UIM, the IR-UIM randomly recommends 40 web pages to them in order to capture their interests. In the following succeeding days, the users' interests become rich according to their usage of IR-UIM. The changing laws are shown in Figs 1, 2 and 3. In Figs 1 and 2, at the beginning the trend lines of the average precisions and average recalls of CVSM, UCVM and TCVM are low. However, they increase quickly in the next days. About seven days later, they tend to stabilize. TCVM is always over UCVM, although the difference between them is small. The lines of UCVM and TCVM are very adjacent. Within a month of the experiment, the users' interests begin to change slightly. In Fig. 3, at the beginning of the usage, given the lack of sufficient information about the users' interests, the recommended pages are not consistent with their interests, so that the users read the web pages randomly and OHPs are low. However, the lines of OHPs are stable towards the end.

As a whole, UCVM and TCVM are much better than CVSM, in which only slight fluctuation occurs. However, TCVM is better than UCVM. For average precision, UCVM is higher than CVSM by about 16%, and TCVM is higher than UCVM by about 5% after the lines stabilize. In terms of the average recall, UCVM is higher than CVSM by about 19%, and TCVM is higher than UCVM by about 4% after the lines stabilize. In terms of the average OHP, UCVM is higher than CVSM by about 21%, and TCVM is higher than UCVM by about 5% after the lines stabilize. The reasons are obvious. The importance degree of interests is considered in UCVM, which tends to find out the more relevant pages for users than CVSM. The time factor and the interest importance degree are both taken into account in TCVM. Hence, TCVM is the best among the three models. As Fig. 3 shows, most

of the users of UCVM and TCVM browsed the pages in the recommended order, whereas only relatively few of the users of CVSM browsed them. Some of the recommended pages may have been ranked in the wrong page order. For example, the users always empirically like to read the top rank pages. If one page, in which the user is interested is ranked 2, but CVSM ranked it 5, and the user is not interested in the page that is ranked 2, the OHP is low.

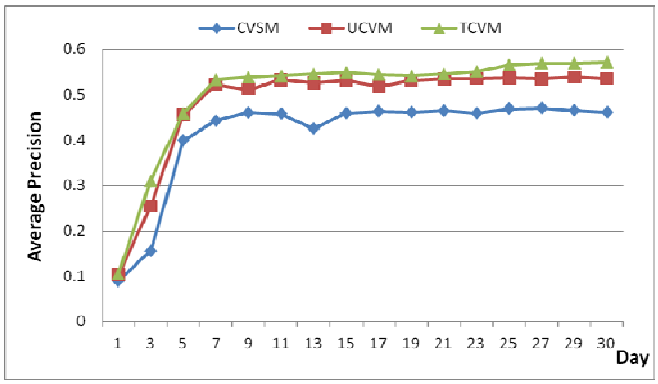

Fig. 1. Average precisions of CVSM, UCVM and TCVM used by groups 1, 2 and 3, respectively

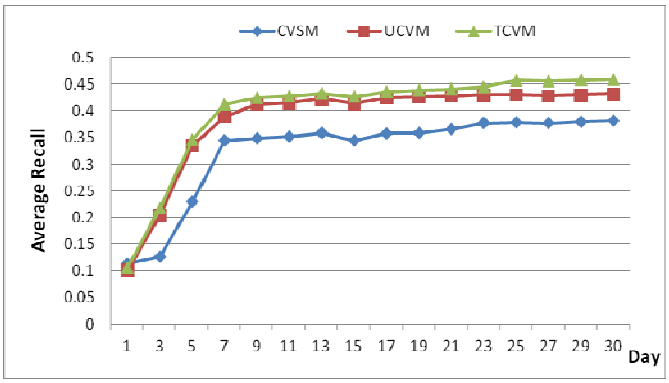

Fig. 2. Average recalls of CVSM, UCVM and TCVM used by groups 1, 2 and 3, respectively

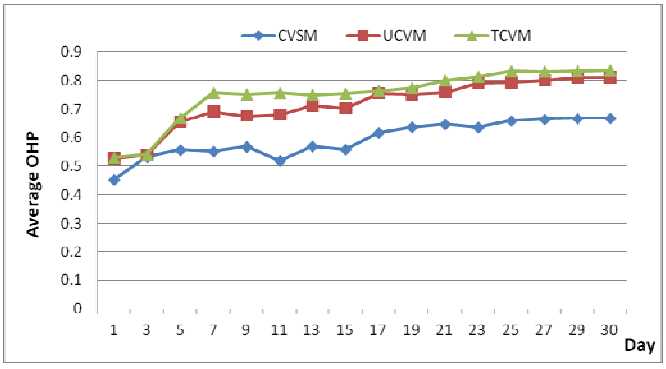

Fig. 3. Average OHPs of CVSM, UCVM and TCVM used by groups 1, 2, and 3, respectively

8. Conclusions

In this paper we presented a time-aware and grey incidence theory-based user interest model, called TCVM, for document recommendation with consideration of the time factor and the user interest importance degree. The time factor emphasizes that the recent information a user has browsed is more related to his/her interests. The user interest importance embodies the major interests and minor interests. Experimental evaluations were conducted for the three related models: CVSM (i.e., denoting conventional VSM), UCVM (i.e., considering interest importance degree) and TCVM. The results indicate that our proposed model TCVM outperforms UCVM and CVSM and provides more accurate recommendations. In conclusion, we state that the TCVM substantially improves the performance by considering the time factor and the interest importance degree and that it is the best among the three models in terms of the three evaluation metrics of precision, recall and OHP. In future studies, we will further analyze three aspects of TCVM: (1) the influence of the time span (i.e., appropriate magnitude of the time span for TCVM), (2) finding out whether or not other time functions are better for TCVM, (3) the application of TCVM concept in other recommender systems, such as collaborative filtering and hybrid recommendations.

***Acknowledgements*:** This work was supported in part by a grant from the Science and Technology Projects of Henan Province (132400410249). We also thank the students who volunteered in the experiment.

References

1. Resnick, P., H. R. Varian. Recommender Systems. – Communications of the ACM, Vol. **40**, 1997, No 3, pp. 56-58.
2. Adomavicius, G., A. Tuzhilin. Toward 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**, 2005, No 6, pp. 734-749.
3. Zhou, X, Y. Xu, Y. Li et al. The State-of-the-Art in Personalized Recommender Systems for Social Networking. – Artificial Intelligence Review, Vol. **37**, 2012, No 2, pp. 119-132.
4. De Campos, L. M., J. M. Fernández-Luna, J. F. Huete et al. Combining Content-Based and Collaborative Recommendations: A Hybrid Approach Based on Bayesian Networks. – International Journal of Approximate Reasoning, Vol. **51**, 2010, No 7, pp. 785-799.
5. Liu, J., P. Dolan, E. R. Pedersen. Personalized News Recommendation Based on Click Behaviour. – In: Proc. of 15th International Conference on Intelligent User Interfaces (IUI), February 2010, pp. 31-40.
6. Lew, M. S., N. Sebe, C. Djeraba et al. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. – ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), Vol. **2**, 2006, No 1, pp. 1-19.
7. Yu, K., A. Schwaighofer, V. Tresp. Collaborative Ensemble Learning: Combining Collaborative and Content-Based Information Filtering via Hierarchical Bayes. – In: Proc. of 19th Conference on Uncertainty in Artificial Intelligence (UAI), August 2002, pp. 616-623.
8. Wang, Z., L. Sun, W. Zhu et al. Joint Social and Content Recommendation for User-Generated Videos in Online Social Network. – IEEE Transactions on Multimedia, Vol. **15**, 2013, No 3, pp. 698-709.

9. Gong, S. Learning User Interest Model for Content-Based Filtering in Personalized Recommendation System. – JDCTA: International Journal of Digital Content Technology and its Applications, Vol. **6**, 2012, No 11, pp. 155-162.
10. Campos, P. G., F. Díez, I. Cantador. Time-Aware Recommender Systems: A Comprehensive Survey and Analysis of Existing Evaluation Protocols. – User Modeling and User-Adapted Interaction, Vol. **24**, 2014, No 1-2, pp. 67-119.
11. Baltrunas, L., X. Amatriain. Towards Time-Dependant Recommendation Based on Implicit Feedback. – In: Proc. of Workshop on Context-Aware Recommender Systems (CARS'09), October 2009.
12. Koren, Y. Collaborative Filtering with Temporal Dynamics. – Communications of the ACM, Vol. **53**, 2010, No 4, pp. 89-97.
13. Liu, N. N., M. Zhao, E. Xiang et al. Online Evolutionary Collaborative Filtering. – In: Proc. of 4th ACM Conference on Recommender Systems (RecSys'10), ACM, September 2010, pp. 95-102.
14. Musto, C. Enhanced Vector Space Models for Content-Based Recommender Systems. – In: Proc. of 4th ACM Conference on Recommender Systems (RecSys'10), September 2010, pp. 361-364.
15. Raghavan, V. V., S. K. M. Wong. A Critical Analysis of Vector Space Model for Information Retrieval. – Journal of the American Society for Information Science, Vol. **37**, 1986, No 5, pp. 279-287.
16. Skillen, K. L., L. Chen, C. D. Nugent et al. Ontological User Modelling and Semantic Rule-Based Reasoning for Personalisation of Help-on-Demand Services in Pervasive Environments. – Future Generation Computer Systems, Vol. **34**, 2014, pp. 97-109.
17. Sarwar, B., G. Karypis, J. Konstan et al. Item-Based Collaborative Filtering Recommendation Algorithms. – In: Proc. of 10th International Conference on World Wide Web (WWW), April 2001, pp. 285-295.
18. Zhang, W., T. Yoshida, X. Tang. A Comparative Study of TF*IDF, LSI and Multi-Words for Text Classification. – Expert Systems with Applications, Vol. **38**, 2011, No 3, pp. 2758-2765.
19. Ding, Y., X. Li. Time Weight Collaborative Filtering. – In: Proc. of 14th ACM International Conference on Information and Knowledge Management (CIKM), ACM, October 2005, pp. 485-492.
20. Liu, S. F., Z. G. Fang, Y. Lin. Study on a New Definition of Degree of Grey Incidence. – Journal of Grey System, Vol. **9**, 2006, No 2, pp. 115-122.
21. Liu, S. F., N. M. Xie, J. Forrest. Novel Models of Grey Relational Analysis Based on Visual Angle of Similarity and Nearness. – Grey Systems: Theory and Application, Vol. **1**, 2011, No 1, pp. 8-18.
22. Deng, J. L. Introduction to Grey System Theory. – The Journal of Grey System, Vol. **1**, 1989, No 1, pp. 1-24.
23. Xiao, X., W. Z. Lin. Application of Protein Grey Incidence Degree Measure to Predict Protein Quaternary Structural Types. – Amino Acids, Vol. **37**, 2009, No 4, pp. 741-749.
24. Yaoguo, D., L. Sifeng, L. Bin et al. Improvement on Degree of Grey Slope Incidence. – Engineering Science, Vol. **2004**, No 3, pp. 41-44.
25. Guoliang, Z. S. Z. Comparison between Computation Models of Grey Interconnect Degree and Analysis on Their Shortages. – Systems Engineering, Vol. **14**, 1996, No 3, pp. 45-49.
26. Hernández del Olmo, F., E. Gaudioso. Evaluation of Recommender Systems: A New Approach. –Expert Systems with Applications, Vol. **35**, 2008, No 3, pp. 790-804.
27. **http://nlp.stanford.edu/software/segmenter.shtml**