Computer Science and Information Systems 18(2):419–439 https://doi.org/10.2298/CSIS200120003Y

Text Recommendation Based on Time Series and Multi-label Information

Yi Yin[1,2*], Dan Feng[1], Zhan Shi[1] and Lin Ouyang[2]

[1] Wuhan National Laboratory for Optoelectronics,
Huazhong University of Science and Technology;
Key Laboratory of Data Storage System, Ministry of Education,
Huazhong University of Science and Technology,
430074 Wuhan, Hubei, China
[*] yinyi@wust.edu.cn, {dfeng, zshi}@hust.edu.cn
[2] School of Computer Science and Technology,
Wuhan University of Science and Technology,
430073 Wuhan, Hubei, China
ouyanglin@wust.edu.cn

Abstract. One of the key functions of the method of text recommendation is to build a correlation analysis to all the text collection. At present, most of the text recommendation methods use the citation network, but less to consider the internal relations, which has become a challenge and an opportunity for the research of text recommendation. Therefore, we propose a new method to ameliorate the above problem based on the time series in this paper. We specify a certain text collection according to the interests of users and integrate the varied label values of the text, then we build the correlation coefficient between text and its related text with the differential analysis, finally the similarity degree of the text is calculated out by using the improved cosine similarity correlation matrix to promote a recommendation of similar text. Our experiments indicate that we are able to ensure the quality of text, with an improvement of accuracy by 8.63% as well as an improvement of recall rate by 5.25%.

Keywords: time series; label value; correlation coefficient; similarity degree

1. Introduction

Nowadays it's becoming a dogma that more and more dates have been connected together as a colossal directed network with a common case of cited node [1, 2]. In the directed network, the latest data squint to achieve their own functions in view of the original data. One of the most typical applications is the processing of text data, for instance, the new text customarily needs to cite the original ones to achieve its own research and innovation.

That may easily lead to some ineluctable similarity of these mutually referred texts, the association analysis of these similarity texts has become one of the main ideals of the text recommendation to meet the needs of different users. Many previous text

[*] Corresponding author

recommendations have been used to take the citation relationship as the main approach to achieve their function, which means, if the text 1 cited the text 2, they are related. However, these methods are too simple to get the useful information [3]. Therefore, more and more researches of text recommendation have been indulged in exploring the label value on this basis, a variety of recommendation methods have been put forward according to the multi-attributes of the text [4, 5], like the keyword, research field as well as citation times and so on.

Still, we should pay attention to these ubiquitous problems hidden in the application of the label value in the text recommendation. (1) The label value is quite abstract, that is to say, there are still unfathomed symptoms, the user will not get the accurate text easily through these tags and he is more likely to capture the text in vain. (2) The label value analysis of the existing methods of text recommendation usually adopts an extensive weighted average, which ignores the individual differences. That model built with this method cannot reflect the different characteristics of the global data. (3) Generally speaking, there have been two accesses to the label value for the user, one way is based on the explicit formulation of the user himself, the other way is to utilize the helpful information of text to capture the label value covertly. However, both methods get the problem of how to extract information and give a standard measurement, which results in a waste of the user's time and a lower expectation of the recommendation system [6]. At the meantime, these two problems should be the main reason of people being not satisfied with many existing methods of text recommendation. (4) Most of the existing text recommendation methods only use a single scoring model extensively to mark the label value one by one and they would default to be independent of each other, which might be ignoring of these multitudinous label values.

In order to solve the problems mentioned above, we propose a new text recommendation method based on the time series, which is combined with the multi-label value. This method imports the time series into the label value. What the time series denotes was a set of statistical data which can be arranged in order of time. In this paper, the statistical index of time series is exactly the varied of label value of the text, the value of label value is directly related to the length of time, and the label value can usually be obtained through continuous summary in the time series. When the time series is introduced into the label value, the label value can be dynamically extracted and created. Thus, we can capture a set of ideal recommendable texts with time-validity.

So far, there have been some researches on the applying of the time series in text analysis. ZR Jiang et al. had applied the time series to subject analysis. S Bjork et al. [7] analysed the pattern of innovation cycle time of the economy knowledge that the French Nobel Prize winner from 1930 to 2005 had kept to base on the time series. A Lercher et al. [8] gave a research of correlation rule of text with the time series analysis, etc. However, there has been always a pity that scarcely any researcher had taken the combination of time series and label value into consideration in the area of text recommendation.

There are several benefits for the combined utilization of time series and multi-label value in text recommendation method: (1) the text data is one kind of fairly detailed data, and the thoughts attached on text data always have limitations. In order to give a response to the specific needs of users, it is always necessary to carry out a full range of multi-angle analysis of the text, so here may be a large number of label values that can

meet the needs of different users [9], as label value is one of the attributes of the text itself. (2) For the text recommendation, the time of a text is a key factor that cannot be ignored by the user. This is because the information content of the text is timeliness. For example, the timeliness of a news text is very strong; its time interval needs to be accurate to day or even less, as for academic texts, although their timeliness is not strong, it is still necessary to meet the need of users according to the time series. The reason of a text being arranged in the order of time is that the latter text is based on the development of society and science, and especially the previous text. (3) There is citation relationship between a text and its reference, which takes time of each text as the premise, that is to say, the recent text can cite the earlier text well but not on the contrary as data are time ordered. If we ignore the text sequence, the research such as text citation times and references relationship will be out of the question, the time series is the prerequisite, the citation of a text is the need to be guaranteed with very strong time series.

Based on the introduction above, the main contributions of our work in this paper include the following aspects: (1) Quantify the multi-label value of the text, which is more flexible than the previous method of obtaining information through the subject of the text and citation times monolithically. (2) Establish a combination model of a time series with multi-label information to capture the correlation coefficient between texts by the method of differential cryptanalysis, and carry out the similarities between texts [10]. (3) Consider the evaluation value of a text to the user, the practicability of the proposed method is verified on a large-scale real data set.

Our ground work in this paper is elaborated as follow. We introduce the related work in the second chapter and describe the concrete form of a time series and process of the label value in the third chapter. The basic definition and the assignment of label value, along with the computer realization method are mentioned to obtain the correlation coefficient of a text by a difference equation in the chapter four, while the fifth chapter demonstrates our method of text recommendation, then we conclude our work in section 6.

2. Current Practice and Research

At present, some text recommendation algorithms based on time series usually extract some label information in the text shared by the users, and then generate the algorithm by making the label information connected through statistical learning method. When extracting label information, the algorithm generally pays attention to the importance of a part of label information in common knowledge. For example, focusing on keyword label indicates that the algorithm wants to recommend more relevant text to users, or focusing on the authoritative label indicates that the algorithm wants to recommend more classic text to users, etc. Different label information and models reflect different emphases of the algorithm. The analysis of the cited text usually use the graph model to represent the relationship between texts, and a large amount of text recommendation researches have adopted the method of extracting label value.

Gupta S et al. came up with a recommendation method based on the theme and the core idea of the text [11], in their method, the user had to provide a full text (including

title, abstract, text and reference) for extracting the core idea. Similarly, Tellez E S et al. introduced an independent framework to recommend the useful text [12], in which users also had to enter a full academic literature text to generate some different information, and then submitted it to the existing network information resources to realize the recommendation. Mäntylä M V et al. took the cited text abstract, introduction and conclusion to obtain better recommendation results [13], however, these methods not only actually increased the user's burden, but also could not be able to provide extra information on the basis of a section of the user's interested information in the real environment. Caruccio L et al. studied the problem of how to use references to recommend based on the user's query without an additional reference list [14], they designed a non-parametric probability model, and calculated the correlation between the two texts by using the reference information. There were some other studies focusing on improving the topic similarities of the cited texts. Huang S et al. found out that in the topic clustering of citation, using references could effectively avoid the "drifting" [15]. Harman D et al. had verified the different extraction methods of references, which would have an influence on the quality of information retrieval [16]. All of these researches are effective methods for text recommendation.

At the same time, there also have been some researches on the single role of a time series in the text. For example, the generation of timing [17], timing based clustering and classification [18], information retrieval for future [19], etc.

According to the above analysis, a large number of recommendation algorithms based on shared text have the problems of independence of label information of default text or simplification of calculation method. That is to say, the same method is used in the quantitative calculation of label. Such a label extraction method is not comprehensive in the expression of text content.

3. Time Series and Multi-Label Value

This chapter does a quantitative analysis of label value. Each access of the user is determined by a variety of label value. Therefore, the co-analysis of multi-label value denotes the relationship between the texts, which has benefit of finding the similar text to recommend to the user.

3.1. Defined Variable

We supposed u as an arbitrary user, and text set as $\mathbf{X} = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$. The whole label value of $\mathbf{X}$ can be denoted by $\mathbf{g}(\mathbf{X})$,

$$\mathbf{g}(\mathbf{X}) = \begin{pmatrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \vdots \\ g_m(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} g_1(x_1) & g_1(x_2) & \cdots & g_1(x_n) \\ g_2(x_1) & g_2(x_2) & \cdots & g_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ g_m(x_1) & g_m(x_2) & \cdots & g_m(x_n) \end{pmatrix}$$

Here, there were a number of m of label values for each x_i, label value of x_i $\mathbf{g}(x_i)$ should be $\mathbf{g}(x_i) = (g_1(x_i), g_2(x_i), \cdots, g_m(x_i))$. If x_i cited a set of text $\mathbf{x}_i$, we would like to

mark this text set as $\mathbf{x}_i = \{x_i^1, \quad x_i^2, \cdots, x_i^l\}$, The whole label value of $\mathbf{x}_i$ can be denoted by $\mathbf{g}(\mathbf{x}_i)$,

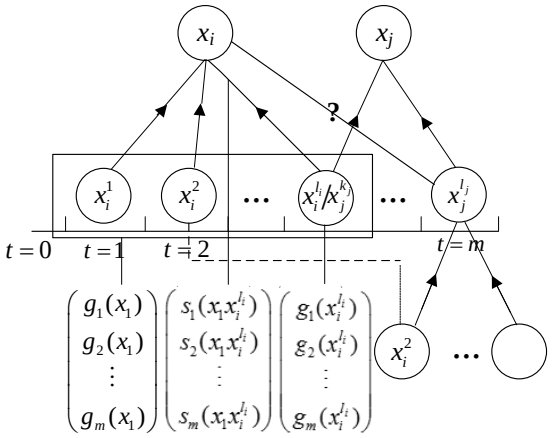
$$\mathbf{g}(\mathbf{x}_i) = \begin{pmatrix} g_1(\mathbf{x}_i) \\ g_2(\mathbf{x}_i) \\ \vdots \\ g_m(\mathbf{x}_i) \end{pmatrix} = \begin{pmatrix} g_1(x_i^1) & g_1(x_i^2) & \cdots & g_1(x_i^l) \\ g_2(x_i^1) & g_2(x_i^2) & \cdots & g_2(x_i^l) \\ \vdots & \vdots & \ddots & \vdots \\ g_m(x_i^1) & g_m(x_i^2) & \cdots & g_m(x_i^l) \end{pmatrix}.$$

As x_i cite this set of text $\mathbf{x}_i$, $\mathbf{x}_i = \{x_i^1, x_i^2, \cdots, x_i^1\}$, the correlation coefficient of x_i and $\mathbf{x}_i$ should be,

$$\mathbf{s}(x_i\mathbf{x}_i) = \begin{pmatrix} \mathbf{s}_1(x_i\mathbf{x}_i) \\ \mathbf{s}_2(x_i\mathbf{x}_i) \\ \vdots \\ \mathbf{s}_m(x_i\mathbf{x}_i) \end{pmatrix} = \begin{pmatrix} s_1(x_i x_i^1) & s_1(x_i x_i^2) & \cdots & s_1(x_i x_i^l) \\ s_2(x_i x_i^1) & s_2(x_i x_i^2) & \cdots & s_2(x_i x_i^l) \\ \vdots & \vdots & \ddots & \vdots \\ s_m(x_i x_i^1) & s_m(x_i x_i^2) & \cdots & s_m(x_i x_i^l) \end{pmatrix}$$

The first equation represents a set of importance relationships between x_i and $\mathbf{x}_i$, and the second equation represents the correlation coefficient which every text had referred by x_i.

Figure 1 showed the basic frame structure of a text reference and the relevant parameters.



Tags / Text	time	keyword parameter	authoritative parameter	subject classification	cited degree parameter
x_1					
x_2					
⋮					
x_n					

Fig. 1. Citation structure of text.

Figure 1 includes the citation correlation, label value, a time series, the correlation coefficient between the text and their correlation diagram. As we can see from this figure, for the text x_i and x_j, they have the same cited text x_i^l or x_j^k , and we can estimate

the similarity of x_i and x_j according to the comparison of two correlation coefficients between $\begin{pmatrix} s_1(x_i x_i^l) \\ s_2(x_i x_i^l) \\ \vdots \\ s_m(x_i x_i^l) \end{pmatrix}$ and $\begin{pmatrix} s_1(x_j x_j^k) \\ s_2(x_j x_j^k) \\ \vdots \\ s_m(x_j x_j^k) \end{pmatrix}$.

Among the variants defined in this paper, if the text $\mathbf{x}_i$ is cited by x_i, which is also the text in set $\mathbf{X}$, that means $\mathbf{x}_i \subset \mathbf{X}$. In this paper, we adopted these two different recording ways to make it easier to discuss the text, for example, the text x_i^1 in the figure 2, it was not only cited by x_i, but also cited by x_j, so x_i^l could be marked as x_j^k. The text x_i^2 can be cited by x_j^k as well as x_i, $\{x_i, x_j, x_i^1, x_i^2, x_i^l, x_j^k\}$ all came from the set of $\mathbf{X}$.

3.2. Definition and Exposition of Time Series

The role of a time series in text recommendation is very important [20]. For text recommendation, there are two kinds of time, one is the published time of the text, the other is the user' access date to the text. The text published time usually reflects the existence of the text; the access time reflects the need dateline of the user for text information. In this paper, the text denoted to be the published time. If the label information comes from text, the label is a kind of special word information. If the label information comes from other statistics information, the label is not word information. Therefore, label information cannot be a part of word.

Here, t is used to represent the time parameter. For the time parameter t, there are several properties. First, as we have mentioned in chapter 1, all the texts $\mathbf{x}_i$ are cited by x_i arranged in chronological order of time t. Second, in view of the fact that text information is of timeliness, this paper makes a restriction on the publication time, which limits its minimum number of years. In this paper, we had to restrict the publication time of the text by limiting t to the minimum year in view of the timeliness of text. There will be a difference between the publication time and the minimum number of years, the minimum year denoted the initial time, let the text of the initial time or before as 0, then for all the text after this time node could be: the time parameter t = publication time – the initial time. For instance, when $\mathbf{x}_i$ is cited by x_i, the time parameter of text $\mathbf{x}_i$ can be arranged from small to large. If there is time span between text x_i^{l-1} and x_i^l, the time span of these two texts is also retained. Third, we can use the limitation of the publication time to avoid the multiple citation, and to reduce the amount of calculation. For example, x_i cites x_i^l, and x_i^l cites the other texts, after many such citations, some published time of papers here are less than the minimum year, so we can discard them. In other word, there is no need to calculate the correlation coefficient between these discard texts with x_i^l. If the text is sorted according to different time series, it may have an impact on the final classification result. However, this paper assumes that they are arranged according to the chronological order of natural years, forming a unified sorting order. This paper only discusses the correlation coefficients arranged according to this chronological order. Fourth, the correlation coefficient and label values are functions of time parameters, which are defined the similarity of the texts by using the difference. As for the definition and physical significance of difference

equations, the change would reflect on timeline, and we would be able to calculate the correlation coefficient of x_i and $\mathbf{x}_i$ with the difference equations. Fifth, for the texts in $\mathbf{x}_i$, there may be a number of the same time value. To solve this problem, we can do a secondary arrangement according to the value of keywords, from large to small. Which means, when $t(x_i^l) - t(x_i^{l-1}) = 0$ and $t(x_i^l) = t(x_i^{l-1}) \neq 0$, the sequence of x_i^{l-1} and x_i^l can be arranged from large to small according to one of the element value in $\mathbf{g}(\mathbf{x}_i)$. Sixth, consideration of the timeliness of the text is necessary, text information would generally show a process of decay with the increase of time, for instance, the near-term text would be more important than the older ones.

Last but not the least, texts has a citation relationship, and the citation relationship had to take time as a prerequisite, that means, the near time text can cite the earlier text well but not on the contrary. Therefore, text citation should to be guaranteed with a strong time order.

3.3. Establish the Correlation Coefficient

Based on above definition, we discussed the value of correlation coefficient $\mathbf{s}(x_i\mathbf{x}^i)$ between x_i and $\mathbf{x}_i$. Assumed that the time series t as all the cited text $\mathbf{x}_i$ of x_i, $t = \{0, 1, 2, \cdots, m\}$. That means, t from 1 to m. When $t = 0$, the corresponding text denoted the most distal one of $\mathbf{x}_i$, on the contrary, when $t = m$, the corresponding text denoted the most proximal one of $\mathbf{x}_i$. For the correlation coefficient, since it represents the relationship between x_i and $\mathbf{x}_i$, we can establish a connection between them. At a given time, the influence degree of x_i^l to x_i is denoted as $\begin{pmatrix} g_1(x_i^l) \\ g_2(x_i^l) \\ \vdots \\ g_m(x_i^l) \end{pmatrix}' \cdot \begin{pmatrix} s_1(x_i x_i^l) \\ s_2(x_i x_i^l) \\ \vdots \\ s_m(x_i x_i^l) \end{pmatrix}$. Since the importance of text exhibits attenuation over time, and the whole process is discrete, therefore, the influence of $\mathbf{x}_i$ on x_i is $\begin{pmatrix} g_1(\mathbf{x}_i) \\ g_2(\mathbf{x}_i) \\ \vdots \\ g_m(\mathbf{x}_i) \end{pmatrix}' \cdot \begin{pmatrix} \mathbf{s}_1(x_i\mathbf{x}_i) \\ \mathbf{s}_2(x_i\mathbf{x}_i) \\ \vdots \\ \mathbf{s}_m(x_i\mathbf{x}_i) \end{pmatrix}$. The connection between x_i and $\mathbf{x}_i$ is the superposition of the influence of $\mathbf{x}_i$ on x_i in the whole time $T = \{1, 2, \cdots, \tau\}$. Then, the correlation of x_i and $\mathbf{x}_i$ could be:

$$\begin{pmatrix} g_1(x_i) \\ g_2(x_i) \\ \vdots \\ g_m(x_i) \end{pmatrix} - \begin{pmatrix} g_1(\mathbf{x}_i) \\ g_2(\mathbf{x}_i) \\ \vdots \\ g_m(\mathbf{x}_i) \end{pmatrix}' \cdot \begin{pmatrix} \mathbf{s}_1(x_i\mathbf{x}_i) \\ \mathbf{s}_2(x_i\mathbf{x}_i) \\ \vdots \\ \mathbf{s}_m(x_i\mathbf{x}_i) \end{pmatrix} = \mathbf{0} \quad (1)$$

Which can be converted to:

$$\begin{cases} g_1(x_i) - \sum_{t=1}^{l} g_1(x_i^t) \cdot s_1(x_i x_i^t) = 0 \\ g_2(x_i) - \sum_{t=1}^{l} g_2(x_i^t) \cdot s_2(x_i x_i^t) = 0 \\ \vdots \\ g_m(x_i) - \sum_{t=1}^{l} g_m(x_i^t) \cdot s_m(x_i x_i^t) = 0 \end{cases} \tag{2}$$

Each element in the equation set (2) is arranged in chronological order. We build a frame diagram of discrete system according to the equation set (2), which is shown in figure2.

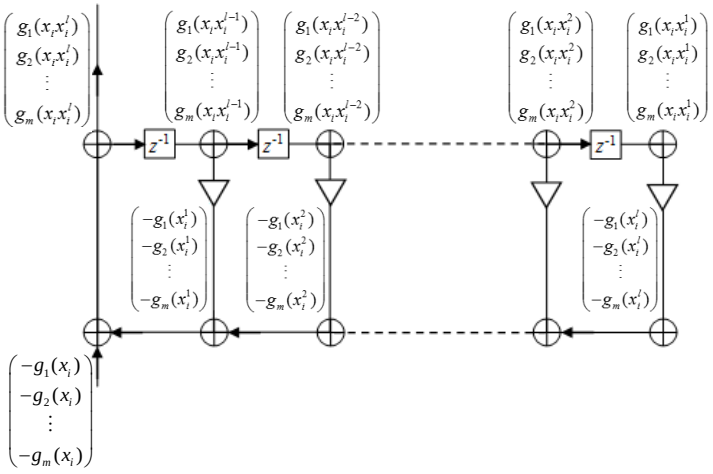


Fig. 2. System chart of the correlation function.

It shows the whole situation of text association, which is composed of n simple subsystems connected in a loop way. If the final association between texts is regarded as a steady state, figure 2 shows the state diagram of the whole association state. From this graph, the correlation equation can be established, that is the state equation, and then get the correlation coefficient between different texts by solving the state equation.

As we can see in figure2, the equation set (2) can be described as difference equation set, it was actually a linear time-invariant system and a zero state response system. By solving the difference equation, we carried out all of the correlation coefficient of x_i and $\{x_i^1, x_i^2, \cdots, x_i^l\}$. There were $m \times l$ solutions of this equation set:

$$\begin{pmatrix} s_1(x_i x_i^1) - D_1(1) \\ s_1(x_i x_i^2) - D_1(2) \\ \vdots \\ s_1(x_i x_i^l) - D_1(l) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1l} \\ a_{11}^2 & a_{12}^2 & \cdots & a_{1l}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{11}^l & a_{12}^l & \cdots & a_{1l}^l \end{pmatrix} \begin{pmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1l} \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} s_2(x_i x_i^1) - D_2(1) \\ s_2(x_i x_i^2) - D_2(2) \\ \vdots \\ s_2(x_i x_i^l) - D_2(l) \end{pmatrix} = \begin{pmatrix} a_{21} & a_{22} & \cdots & a_{2l} \\ a_{21}^2 & a_{22}^2 & \cdots & a_{2l}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{21}^l & a_{22}^l & \cdots & a_{2l}^l \end{pmatrix} \begin{pmatrix} C_{21} \\ C_{22} \\ \vdots \\ C_{2l} \end{pmatrix} \tag{4}$$

$$\vdots$$

$$\begin{pmatrix} s_m(x_i x_i^1) - D_{m1}(1) \\ s_m(x_i x_i^2) - D_{m2}(2) \\ \vdots \\ s_m(x_i x_i^l) - D_{ml}(l) \end{pmatrix} = \begin{pmatrix} a_{m1} & a_{m2} & \cdots & a_{ml} \\ a_{m1}^2 & a_{m2}^2 & \cdots & a_{ml}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}^l & a_{m2}^l & \cdots & a_{ml}^l \end{pmatrix} \begin{pmatrix} C_{m1} \\ C_{m2} \\ \vdots \\ C_{ml} \end{pmatrix} \tag{5}$$

Then we simplified the above function as:

$$\begin{pmatrix} \mathbf{s}_1(x_i \mathbf{x}^i) \\ \mathbf{s}_2(x_i \mathbf{x}^i) \\ \vdots \\ \mathbf{s}_m(x_i \mathbf{x}^i) \end{pmatrix} - \begin{pmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_m \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_m \end{pmatrix}' \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_m \end{pmatrix} \tag{6}$$

Related to the equation set (2), $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)'$ could be able to denote as the solutions of $\mathbf{s}(x_i \mathbf{x}^i)$, and $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m)'$ denoted the characteristic root of each equation, and $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_m)'$ denoted the coefficient of the characteristic root of each equation.

Here, the initial conditions of equation set (2) should be

$$\begin{cases} g_1(x_i^{-t}) \cdot s_1(x_i x_i^{-t}) = 0 \\ g_2(x_i^{-t}) \cdot s_2(x_i x_i^{-t}) = 0 \\ \vdots \\ g_m(x_i^{-t}) \cdot s_m(x_i x_i^{-t}) = 0 \end{cases} \tag{7}$$

Therefore, the values of **D**, **V** and **C** can be carried out, and then the correlation coefficient $\mathbf{s}(x_i \mathbf{x}^i)$ was obtained.

In the same way, we could also get the correlation coefficient between x_j and $\mathbf{x}_j$, so that all the text **X** and their cited text can eventually be carried out as a value of correlation coefficient.

3.4. Comparison of Correlation Coefficient

If the user *u* wants to obtain a series of texts related to the x_i, it is necessary to compare the correlation coefficient of the text that related to the x_i. For example, in Figure 1, $\{x_i^1, x_i^2, \cdots, x_i^l\}$ are directly related to x_i, and x_j is indirectly related to x_i. Both the similarity of x_i and $\{x_i^1, x_i^2, \cdots, x_i^l\}$and the similarity of x_i and x_j, we had to compare their correlation coefficient. As the correlation coefficients are actually *m* vectors that composed with a set of label value, so we had to find out a method that was able to compare the vectors. We eventually came up with an improved cosine similarity to restrain the impact resulted from the *m* label value.

To obtain the similarity of x_i and x_j, we adopted an improved cosine similarity formula as below:

$$sim(\mathbf{s}(x_i x_k^i), \mathbf{s}(x_j x_j^1)) = \frac{\begin{pmatrix} s_1(x_i x_i^k) \\ s_2(x_i x_i^k) \\ \vdots \\ s_m(x_i x_i^k) \end{pmatrix}' \begin{pmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \ddots & \\ & & & \rho_m \end{pmatrix} \begin{pmatrix} s_1(x_j x_j^1) \\ s_2(x_j x_j^1) \\ \vdots \\ s_m(x_j x_j^1) \end{pmatrix}}{\left(\begin{pmatrix} s_1(x_i x_i^k) \\ s_2(x_i x_i^k) \\ \vdots \\ s_m(x_i x_i^k) \end{pmatrix}' \begin{pmatrix} s_1(x_i x_i^k) \\ s_2(x_i x_i^k) \\ \vdots \\ s_m(x_i x_i^k) \end{pmatrix} \right)^{\frac{1}{2}} \cdot \left(\begin{pmatrix} s_1(x_j x_j^1) \\ s_2(x_j x_j^1) \\ \vdots \\ s_m(x_j x_j^1) \end{pmatrix}' \begin{pmatrix} s_1(x_j x_j^1) \\ s_2(x_j x_j^1) \\ \vdots \\ s_m(x_j x_j^1) \end{pmatrix} \right)^{\frac{1}{2}}} \quad (8)$$

Here, $\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$was applied to subtly adjust $\mathbf{s}(x_i x_i^l)$ and $\mathbf{s}(x_j x_j^1)$ into a more adaptive similarity. Parameter $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$ is an important role, it controlled the rate of the m label values in the whole recommendation process, which denoted the restraint degree of label value.

Through this method, we can obtained a series of similar texts related to x_i, so we are able to offer the texts with high similarity to the user.

4. Label Value

For the research of text recommendation, there have been some relatively standard text libraries at home and abroad, we can compare different recommendation methods and system performance in the common text library. And we can also adjust the function parameters, allowing users more flexible access to the text label information [21].

4.1. Selection of Label Value

We have selected a part of texts from Web of Science† (WOS) to test our method in this paper. There were 31393 texts in the area of recommendation system according to the searching keywords “recommendation system” (RS). Web of Science is a product of Thomson Scientific in the United States, which is omnibus multi-disciplinary core journal citation index database and includes three famous citation databases (SCI, SSCI and A&HCI) and two major chemical information database (CCR&IC), as well as the three of SCIE, CPCI-S and CPCI-SSH Citation Database. In virtue of the recommendation method, the date from the search platform of ISI Web of Knowledge have been classified to five major kinds and 151 secondary classifications based on the subject category, and we have obtained an universal network file format (*.net) date set

† http://isiknowledge.com/

to do the citation analysis by the analytical tool of Web of Science [22]. Thus, we could not only effectively reveal the relationship between the internal connection and the texts (which provides the intrinsic link of scientific research), but also could find out the opportune ones to recommend to the user through a multi-tier analytical approach with the help of the provided analytical tool.

The main contents of label value were shown below. We could capture lots of meaningful label value by searching the keyword “text mining” on the WOS.

As can be seen in the figure3, these contents, like the same keyword, similar keywords, author name, journal name, publication time, number of citations, other text references, etc. can be selected as label value. These label value were the key factors in the composition of the text that is needed for an appropriate extraction in text recommendation.

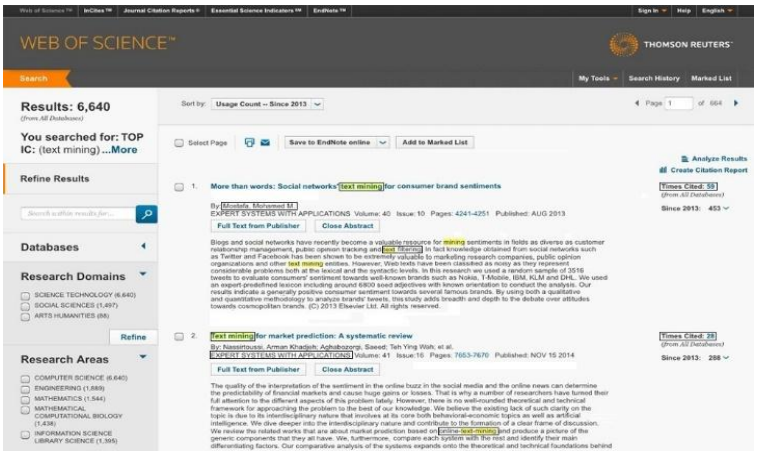


Fig. 3. Example of label value.

4.2. Classification of Label Value

For all of the text, we choose four kinds of label value as the valuable parameters and establish the weight for each type of label values. These four kinds of label value respectively named {keyword, the subject category, text authority, cited degree}, which could reflect the value of a text to the user as a four-dimensional vector. This paper considers the different needs of users, and the needs of users are comprehensive from the perspective of the paper data. Therefore, this article combines the type description method of statistical data to quantify the label information. The theoretical basis for the selection of the label content of the paper is that its selection method is based on statistical principles, and combined with the "statistical data type description" method to select the label.

(1) Keywords. It is the most important criteria for users to obtain the text, as well as the most important label in the method of text recommendation [23]. For the search of texts, the more detailed keywords are, the more accurate the results are. However, for

the text recommendation, less keywords may lead to a large number of irrelevant texts and an unstable accuracy at mean time, while too many keywords may result in missing some texts and a reduced recall rate. Therefore, the choice of the type and the number of keywords would have a direct impact on the text mining. In general, the role of secondary keywords is more important. We specified that there must be at least one keyword, whether it was a primary or secondary keyword, it must be presented in the title, summary, or keywords columns. In our work, we divided the keywords into three types according to the contents of the text: primary keywords, secondary keywords and non- keywords. As shown in Figure3, "text mining" was the main keyword, but for the first two texts in Figure3, "bi-text mining" and "discovery knowledge" are the secondary keywords. It is easy to see that there might be a completely different recommendation if we ignore the impact of secondary keywords. For example, although the keyword like "web mining" was totally different with the keyword "text mining", they can express some similar meaning. We converted the keyword into numeric variable, which is generally covered by the title, abstract, author's name, address and so on.

(2) Subject category. Since "WOS" has covered almost all the fields of discipline, the retrieved texts according to one certain keyword may belong to more than one subject category. For instance, if we use "text mining" as the keyword to search, besides of the computer science [24], the retrieved texts can be distributed in linguistics, psychology, Communication Sciences and the related subjects, and so on. What's more, it could be divided into a subject which has little to do with "text mining", like physics and neuroscience, if there is some blend of text mining related algorithm and the subject itself. So we need to label the subject of a text with three different types: original disciplines, related disciplines and other disciplines. Subject categories can be obtained directly from the publishing organization to which the text belongs, and it is easily distinguishable.

(3) Text authority. It is generally believed that there is a very important criterion to judge whether the authority of scientific research texts is in the level of publication of journals or conferences. There are many methods for the classification of publications at home and abroad. For the experimental data of text, we classified the texts in "WOS" into five kinds according to the authority of the texts, which is based on the commonly used classification methods of the academic journals in China, namely core A, core B, core C, regular D and the others.

(4) Cited time. The number of citations per text is easily available. The data of this parameter should be bigger. But if the difference between the two texts is small, the importance of these two texts is hard to determine. Moreover, the number of citations is also relative value, in which there is no uniform unit. For numerical variable, the date should be processed with the discrete way according to width and frequency or another settled method, as the cited times can be processed basing on the times rank 0, 1~10, 11~15 and >15. We could also use Gaussian distribution to set out a mean of the cited number, and then divided the cited degree according to the variance. However, we should mention that there was not enough theoretical foundation for this method. In this paper, we used the cited degree to distinguish the cited times, and divided the cited degree into four grades basing on the cited times: strong the citation intensity, medium citation intensity, week citation intensity and no citation intensity.

4.3. Classification of label value

As mentioned above, the label value of texts can be divided into four categories, but these four types of label information need to be converted to numerical data. This section discussed the assignment of the four categories of label value. As we have defined the label value {keyword, subject category, text authority, citation degree} as $\{\mathbf{g}_1(\mathbf{X}), \mathbf{g}_2(\mathbf{X}), \mathbf{g}_3(\mathbf{X}), \mathbf{g}_4(\mathbf{X})\}$.

(1) Assignment of keywords. For the importance of keywords, we uses tf-idf method to assign this label value. Using r_i to express the key word of x_i, when the number of keywords is R, the label value of the keyword is expressed as $\mathbf{g}_1(\mathbf{X}) = \sum_R \mathrm{tf}(r_i) \times \mathrm{idf}(r_i)$, and the label value of the keyword between 0 and 1. Here, we assume that there is only one primary keyword in the text. The purpose of this processing is to find the recommendable text in a wider scope.

(2) Assignment of subject category. For the subject category, in general, most texts have gotten a clear distinction. We could use the "rule of thumb"[‡] to assign this parameter, which means, we could define the connection of two variables as the correlation strength. If the value of correlation strength equaled 0-0.05, it means non-correlation; if the value of correlation strength equaled 0.05-0.25, it means week correlation; if the value of correlation strength equaled 025-0.60, it means medium correlation; if the value of correlation strength equaled 0.6-1, it means strong correlation. According to the definition of the subject category, referring to the thumb rule, the text is related to the subject, the intensity of which is as follows: the original disciplines = 0.60; the related disciplines = 0.25; and other disciplines = 0.05. Under a special circumstance, some texts will appear in two or even three categories in the same time, then we should accumulate all of the related categories strength.

(3) Assignment of authority parameter. There was long-hold dogma that the definition of authoritative parameters had a strict standard and with no ambiguity. Therefore, we had directly assigned it with accurate numerical data. We also use the thumb rule to do the assignment. Based on the connection of the text and its authority of affiliated institutions, the correlation strength should be assigned as, core A = 1, core B = 0.65, core C = 0.25, general D = 0.05 and other = 0. What needs to be considered is that the authority of the text would change over time, that means, for two texts x_i and x_{i+1}, even though they had the same authority parameters, when the published time of x_i was earlier than x_{i+1}, then generally came to a authority value comparison as $x_i \geq x_{i+1}$. Here, let the original authority parameter x_i to be $\tilde{\mu}^i$, after adjusting the time parameter, its authoritative value should be $\mathbf{g}_3(\mathbf{X}) = \tilde{\mathbf{g}}_3(\mathbf{X}) \cdot t(\mathbf{X})$, here $t(\mathbf{X}) = t/(t+1)$.

(4) Assignment of cited time. The four levels of citations can also correspond to the strength of the four categories of thumb rules. Assuming that the four levels of texts are denoted as {Π1, Π2, Π3, Π4}, they denoted respectively the {strong citation intensity, medium citation intensity, week citation intensity and non-citation intensity}. After disposing the grade of Π4, the correlation coefficient of three remaining categories can be adjusted according to their mutual citation intensity. Specifically, according to the law of the famous economics budget allocation, namely the law of 60:30:10, we could

[‡] https://en.wikipedia.org/wiki/Rule_of_thumb

distribute the correlation strength according to this law and then adjust the citation correlation strength basing on the citation condition. The number of the strong correlation text accounts for 10% of the total number, and the number of the medium correlation text accounts for about 30% of the total number, and the number of the weak correlation text is 60%.

The citation intensity is adjusted as follows: when P1 refers to P2, P2 is adjusted to a strong citation intensity from the medium citation intensity, when P1 refers to P3, P3 is adjusted by the weak citation intensity to the medium citation intensity; when P2 refers to P3, then the weak correlation strength value of P3, then the weak correlation strength value of P3 is multiplied by P2. In this way, the citation intensity is assigned according to this law, and then the existing citation intensity is adjusted and assigned according to the citation situation of the text.

5. Experiment

In this chapter, we first calculate the correlation coefficient of the text through the experiment, and then analyze the influence of our methods on the text recommendation. At last, we compare the difference between our method and other methods in text recommendation. Here, the method in this paper is abbreviated as TSLI method.

5.1. Data Pre-Processing

Firstly, basing on the introduction in section 4.3, we use the tf-idf method to obtain the similarity of similar keywords. To save the calculation time, tf(r_i) should be limited to the extraction of titles and abstracts for each text x_i, at the same time, we obtained the value of keywords $R = 4$. The value of each R would be discussed in the next section. Moreover, we should calculate the values of $\mathbf{g}_2(\mathbf{X})$, $\mathbf{g}_3(\mathbf{X})$, $\mathbf{g}_4(\mathbf{X})$ with the method mentioned in section 4.3. Among the more than 30 thousands selected texts, there was a total of 26993 ones with no correlation, and these texts would not be considered in the TSLI method. For the rest texts, after being arranged in the descending order of citation degree, the rest qualified texts could be divided into strong, medium and weak correlation according to the 60:30:10 rules and their citation degrees were shown in figure 4.

As shown in figure 4, the degree of the correlation has been subject to a long tailed distribution, which indicted that some week correlation texts might also meet the preferences of users, so it was necessary to take these three kinds of correlation degrees into consideration all time. Therefore, we carried out the label values of more than 4 thousands texts to obtain the value they have brought to the user through the correlation analysis. Basing on the above condition, we randomly selected a text x_i, and we obtained the label value of its 31 cited text's original value and normalization value. The result of this experiment is shown in figure 5. This figure is the set of four kinds of label value of $\{\mathbf{g}_1(\mathbf{X}), \mathbf{g}_2(\mathbf{X}), \mathbf{g}_3(\mathbf{X}), \mathbf{g}_4(\mathbf{X})\}$. Figure 6 showed the label value of $\mathbf{x}_i$. As all the four vectors $\{\mathbf{g}_1(\mathbf{X}),\mathbf{g}_2(\mathbf{X}),\mathbf{g}_3(\mathbf{X}),\mathbf{g}_4(\mathbf{X})\}$ have different metrics and units, and that would

always lead to an impact on the results of data analysis. In order to eliminate the dimensional effect between the indexes, it is necessary to standardize the data and make

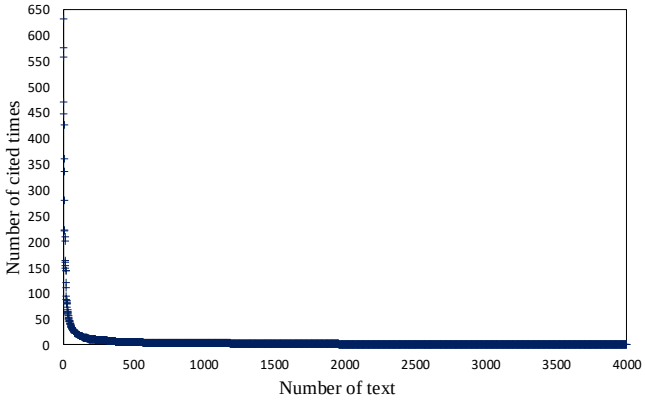


Fig. 4. The citation times of the three kinds of efficient citation degree

the comparability between the data indexes. After the original label information being standardized by the data, the indexes are in the same order of magnitude, so that the whole citation network can be operated and compared comprehensively.

In this paper, we used the Z-score standardization method. We have given the mean and standard deviation of the original data, and carried out the standardization of the data.

The processed data are in accordance with the standard normal distribution, with mean = 0 and variance = 1. The transformation function is:

$$\mathbf{g}^{*}(\mathbf{x}_i)=\frac{\mathbf{g}(\mathbf{x}_i)-\mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)}$$

μ is the sample mean, σ is the sample standard deviation. The converted $\mathbf{x}_i$ is shown in figure 6.

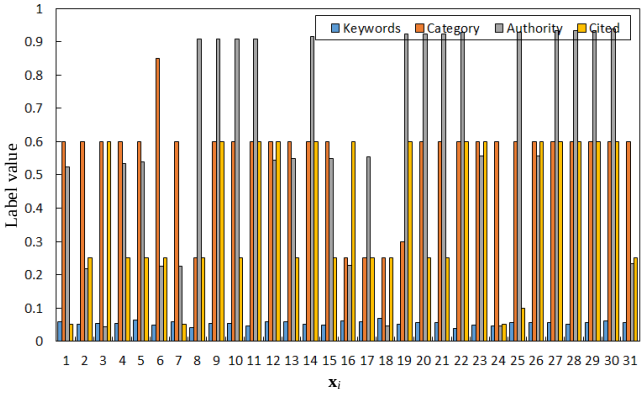


Fig. 5. The label value of text x_i

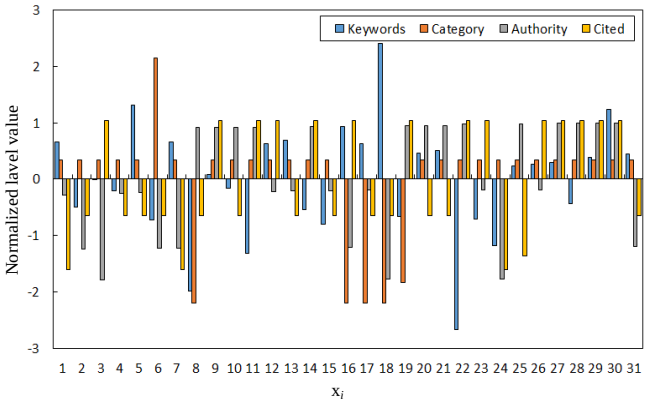


Fig. 6. Standardization of label value of text x_i

The reason why we had used the Z-score instead of using other standard methods is that the standardized data were subject to the standard normal distribution, after using this method for normalization, the vast majority of samples would be concentrated near to the average, which was conducive to the selection of samples. So with the same method, we could also select the text $\mathbf{x}_j$ referred by x_j. In our experiment, we selected out 16 qualified texts from the whole cited texts of x_j, these 16 texts contained some texts in $\mathbf{x}_i$ at the same time. The label value of $\mathbf{x}_j$ is shown in figure 7.

In this way, we finally found out all of the correlation coefficients between the text and its references. After we obtaining the values of $\mathbf{g}(\mathbf{x}_i)$ and $\mathbf{g}(\mathbf{x}_j)$, we could subsequently calculate the values of $\mathbf{s}(x_i\mathbf{x}_i)$ and $\mathbf{s}(x_j\mathbf{x}_j)$ according to the formulas (6) and (8).

5.2. Calculation of Similarity

We have discussed the method of obtaining correlation coefficient above, but there should be uncertain parameters also needed to be discussed, which included the number of the keywords R, the constraint parameter of label value $\boldsymbol{\alpha}$ and the number of the recommendable texts k.

First, consider the case where the number of keywords R and the parameter $\boldsymbol{\alpha}$ affect each of the importance coefficients. The influence of the keyword is more important than other factors in the text recommendation method, which can be seen from most of the pretreatment process. In this paper, we had defined the influence degree of those four types of label value $\{\mathbf{g}_1(\mathbf{X}), \mathbf{g}_2(\mathbf{X}), \mathbf{g}_3(\mathbf{X}), \mathbf{g}_4(\mathbf{X})\}$ as $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, which was exactly the parameter $\boldsymbol{\alpha}$ in formula (8), and with $\alpha_1 > \max\{\alpha_2, \alpha_3, \alpha_4\}$. Besides of the α_1, we unified the constraint parameters of the other three parameter as $\alpha_2 = \alpha_3 = \alpha_4$, and then we converted $\{\alpha_1 \cdot \mathbf{g}_1(\mathbf{X}), \alpha_2 \cdot \mathbf{g}_2(\mathbf{X}), \alpha_3 \cdot \mathbf{g}_3(\mathbf{X}), \alpha_4 \cdot \mathbf{g}_4(\mathbf{X})\}$ into$\{\alpha_1 \cdot \mathbf{g}_1(\mathbf{X}), \alpha_2 \cdot \mathbf{g}_2(\mathbf{X}), \alpha_2 \cdot \mathbf{g}_3(\mathbf{X}), \alpha_2 \cdot \mathbf{g}_4(\mathbf{X})\}$ and assuming $\alpha_1 + \alpha_2 = 1$. This assumption is to highlight the key words in the label value of the significant position, which is also in accordance with the existing traditional recommend- ation method. In addition, as for the influence of

keywords, the number of keyword R is also a non-negligible parameter, it would directly affect the assignment of keywords.

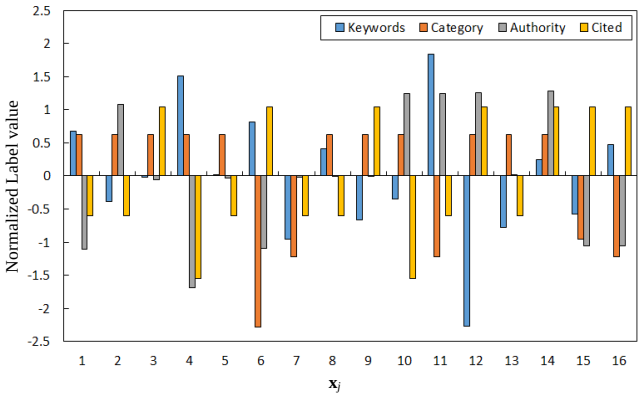


Fig. 7. Standardization of label value of text x_j

Next, we compared the correlation coefficient of x_i and its references x_i^l with the correlation coefficient of x_j and its references $\mathbf{x}_{j.}$ Then we were able to determine the value of R and α based on the cosine similarity calculated with formula (8). We respectively valued (α_1, α_2) as {(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4)}, R as {2, 3, 4, 5, 6} to do the test and obtain the similarity value, the results were shown in figure 8. The value of (α_1, α_2) is {(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4)}, and the value of R is {2, 3, 4, 5, 6}, which is the empirical value to verify the validity of the experiment. Considering that the values of (α_1, α_2) and R have been defined in the theory of the previous chapter, only these three parameters need to be taken as the values in the paper.

Figure 8 showed the effects of the different parameters (α_1, α_2) and R on the sim value. When R=2 and R=5, for any value of (α_1, α_2), the sim value seemed more decentralized than the other three distribution results. When (α_1, α_2) are valued as (0.9, 0.1) and (0.8, 0.2), their sim values also appeared decentralized, which indicated that both of the R=2 and R=5 were not the high discernible parameter in determining the recommendable text. As for R=7, the change of standard deviation became fairly obvious corresponding to the different value of (α_1, α_2). These results manifested that once the value of (α_1, α_2) or R do not completely adapt to certain types of samples, the recommendation method would be very unstable. As shown in the figure, when $R = 4$, no matter what the value of (α_1, α_2) was, the cosine similarity value has changed significantly, moreover, the change of standard deviation was still obvious when compared with other values of R. When (α_1, α_2) = (0.6, 0.4), which showed that the obtained cosine similarity values were concentrated, all of the standard deviation of different R value could be differentiated obviously. Therefore, in this paper we valued (α_1, α_2, R) as (0.6, 0.4, 4) to do the recommendation.

The similarity measure in text recommendation refers to calculating the similarity between texts. The larger the similarity value, the smaller the difference of text. There are many methods to calculate the similarity, cosine similarity is a mature method. For many different texts to calculate the similarity between them, a good way is to map the

labels in these texts to the vector space, form the mapping relationship between the labels and the vector data, and judge the similarity of the text by calculating the difference value of one or more different vectors.

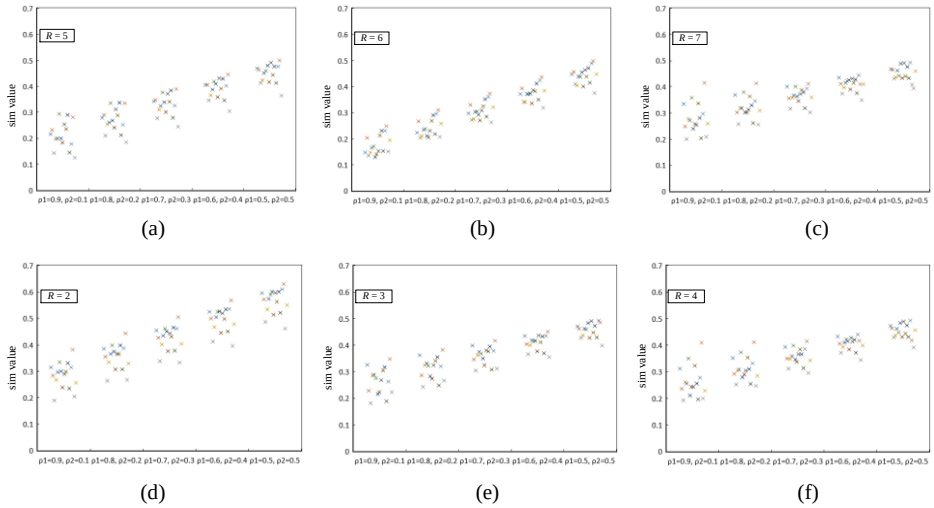


Fig. 8. The influence of different (α1, α2) and Ron the sim value.

5.3. Comparison of Methods

In order to verify the correlation between the recommended texts, we compared our method with other five text recommendation methods. Firstly, we defined:

$$\text{Precision}: P@k = \frac{R_C@k}{R_T} \times 100\%$$

$$\text{Recall}: R@k = \frac{R_C@k}{k_T} \times 100\%$$

$$\text{F-measure}: F1@k = \frac{P@k \times R@k \times 2}{P@k + R@k} \times 100\%$$

Here, R_C is the number of recommendable texts, R_T is the total number of texts obtained by the user, k_T is the number of recommended texts.

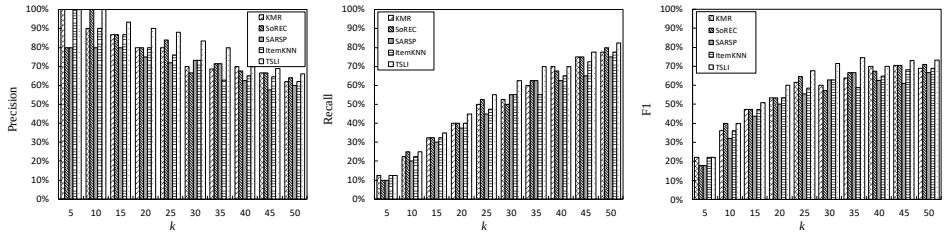
The KMR [25] method is a keyword matching method that can only recommend articles with similar keywords to users by comparing the keywords of different articles.

The SoREC [26] method uses the shared user feature space to combine the social relations with the score information. By combining the two pieces of information, the SoREC identifies the users who are similar in the score and have social relations to make recommendations.

The SARSP [27] method divides users with similar interests into one class, and then the users in this class recommend each other.

The ItemKNN [28] method is to use the item's content / attributes as a vector to find a similar relationship between the users to realize the recommendation process

All this adds up to a true that our method got the most significant advantages. As mentioned before we defined the parameters as: α_1=0.6, α_2=0.4, R=4. The result is shown in figure 9, there were 5-50 texts in this experiment, our TSLI method has showed a better precision rate and recall rate than other methods. At mean time, for the average precision rate and average recall rate, there are respectively an improvement of 8.63% and 5.25%, moreover, their maximum appreciation have come to 12.76% and 7.25%. These results indicated that TSLI was able to carry out a better recommendation result for different number of texts.



(a) (b) (c)

Fig. 9. Comparison of accuracy rate, recall rate and F1 value for different quantity of text.

6. Conclusion

The correlation analysis of these similarity texts has become one of the main ideal of the text recommendation to meet the needs of different users. When the user has captured a certain text of his interest, he would like to get a series of similar texts related to it rather than searching for a large number of texts. We always capture the content of the text information basing on a number of label values, such as the research area, keywords, cited time, etc.

However, these data are very abstract, it is not easy for users to obtain the text they want through this information, and the user needs the text of its label information to be no overlap.

Basing on the information mentioned above, to model the recommendation process better and to reveal the potential influences of the text correlation on the result of the recommendation, this paper analyzed the similarity between texts and elaborated the recommendation processes from three aspects. (1) Defined and classified the label value of text. Each type of label information is introduced into the timing relationship and assigned. (2) Using the difference method to arrange the label value of a text xi and its cited text in a chronological order, then a set of correlation coefficients is formed for the text xi and the label information for each text is cited by x_i. (3) After setting the parameters of correlation coefficient, the consistent text was recommendable to the user basing on the comparison of the citation relationship of the text. Our experimental have effectively testified the quality of the recommended text by our reliable method.

Acknowledgment. This work was partly supported by National Basic Research 973 Program of China under Grant No. 2011CB302301. And Natural Science Fund of Hubei Province under Grant No. 22013CFA131.

References

1. Zhou J, Zeng A, Fan Y, et al. Identifying Important Scholars via Directed Scientific Collaboration Networks. Scientometrics, Vol. 114, No. 3, 1327-1343. (2018)
2. Ren Z M, Mariani M S, Zhang Y C, et al. Randomizing Growing Networks with a Time-Respecting Null Model. Physical Review E, Vol. 97, No. 5, 052311. (2018)
3. Liu X, Zhang J, Guo C. Full-text Citation Analysis: A New Method to Enhance Scholarly Networks. Journal of the Association for Information Science and Technology, Vol. 64, No. 9, 1852-1863. (2013)
4. Deng S, Huang L, Xu G, et al. On Deep Learning for Trust-Aware Recommendations in Social Networks. IEEE Transactions on Neural Networks & Learning Systems, Vol. 28, No. 5, 1164-1177. (2017)
5. Gang, L., Hanwen, Z. (2020) "An Ontology Constructing Technology Oriented on Massive Social Security Policy Documents", Cognitive Systems Research, 60, pp. 97-105.
6. Shen X L, Li Y J, Sun Y. Wearable Health Information Systems Intermittent Discontinuance: A Revised Expectation-Disconfirmation Model. Industrial Management & Data Systems, Vol. 118, No. 3, 506-523. (2018)
7. Bjork S, Offer A, Söderberg G. Time Series Citation Data: the Nobel Prize in Economics. Scientometrics, Vol. 98, No. 1, 185-196. (2014)
8. Shen X L, Li Y J, Sun Y. Wearable Health Information Systems Intermittent Discontinuance: A Revised Expectation-Disconfirmation Model. Industrial Management & Data Systems, Vol. 118, No. 3, 506-523. (2018)
9. Suominen H. Guest Editorial: Text Mining and Information Analysis of Health Documents. Artificial Intelligence in Medicine, Vol. 61, No. 3, 127-130. (2014)
10. Xue, Q., Zhu, Y., & Wang, J. (2019). Joint Distribution Estimation and Naïve Bayes Classification under Local Differential Privacy. IEEE transactions on emerging topics in computing, 1.
11. Gupta S, Varma V. Scientific Article Recommendation by Using Distributed Representations of Text and Graph. Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 1267-1268. (2017)
12. Tellez E S, Moctezuma D, Miranda-Jiménez S, et al. An Automated Text Categorization Framework Based on Hyperparameter Optimization. Knowledge-Based Systems, Vol. 149, 110-123. (2018)
13. Mäntylä M V, Graziotin D, Kuutila M. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. Computer Science Review, Vol. 27, 16-32. (2018)
14. Caruccio L, Deufemia V, Esposito S, et al. Combining Collaborative Filtering and Semantic-Based Techniques to Recommend Components for Mashup Design. Computational Intelligence for Semantic Knowledge Management. Springer, Cham, 25-37. (2020)
15. Huang S, Yu Y, Xue G R, et al. TSSP: Multi-Features Based Reinforcement Algorithm to Find Related Papers. Web Intelligence & Agent Systems, Vol. 4, No. 3, 271-287. (2006)
16. Harman D. Information Retrieval: The Early Years. Foundations and Trends® in Information Retrieval, Vol. 13, No. 5, 425-577. (2019)
17. Adams B, Phung D, Venkatesh S. Social Reader: Towards Browsing the Social Web. Multimedia tools and Applications, Vol. 69, No. 3, 951-990. (2014)

18. Wrixon A, Belov A, Keller M, et al. Static Timing Analysis with Improved Accuracy and Efficiency: U.S. Patent Application 10/002,225. 2018-6-19. (2018)
19. Thorne J, Vlachos A. Automated Fact Checking: Task Formulations, Methods and Future Directions. Arxiv Preprint Arxiv:1806.07687, (2018)
20. Ling Wu, Chi-Hua Chen*, Qishan Zhang, "A Mobile Positioning Method Based on Deep Learning Techniques," Electronics, 8, no. 1, Article ID 59, January 2019.
21. Sun, S., Kadoch, M., Gong, L., & Rong, B. (2015). Integrating network function virtualization with SDR and SDN for 4G/5G networks. IEEE Network, 29(3), 54-59.
22. Fabisiak, L. 2018. "Web Service Usability Analysis Based on User Preferences," Journal of Organizational and End User Computing (30:4), pp. 1-13.
23. Bi, Zhongqin; Dou, Shuming; Liu, Zhe; Li, Yongbin. A Recommendations Model with Multiaspect Awareness and Hierarchical User-Product Attention Mechanisms. Computer Science and Information Systems, 2020, 17(3), pp. 849-865.
24. Lv, Zhihan, Dongliang Chen, Ranran Lou, and Qingjun Wang. "Intelligent edge computing based on machine learning for smart city." Future Generation Computer Systems (2020).
25. Zhao W, Wu R, Liu H. Paper Recommendation Based on the Knowledge Gap between a Researcher's Background Knowledge and Research Target. Information Processing & Management, Vol. 52, No. 5, 976-88. (2016)
26. Du M, Vidal J M, Markovsky B. SOREC: A Semantic Content-Based Recommendation System for Parsimonious Sociology Theory Construction. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 138-144. (2019)
27. Asabere N Y, Xia F, Meng Q, et al. Scholarly Paper Recommendation Based on Social Awareness and Folksonomy. International Journal of Parallel Emergent & Distributed Systems, Vol. 30, No. 3, 211-232. (2015)
28. Wang X, Sheng Y, Deng H, et al. Top-N-Targets-Balanced Recommendation Based on Attentional Sequence-to-Sequence Learning. IEEE Access, Vol. 7, 120262-120272. (2019)

Yi Yin, born in 1983, Ph. D., His research interests include machine learning, recommendation system, data mining and neural networks.

Dan Feng, born in 1970. Ph.D., professor, Ph.D. supervisor. Her research interests include computer system architecture, parallel processing, fault tolerance theory, disk array architecture, and computer storage system.

Zhan Shi, born in 1976. Ph. D., associate researcher. His research interests include massive information storage, storage service and storage management.

Lin Ouyang, born in 1974, Ph. D., His research interests include distributed system, image processing and neural networks.

Received: January 20, 2020; Accepted: December 20, 2020