A MATCH METHOD BASED ON LATENT SEMANTIC ANALYSIS FOR EARTHQUAKE HAZARD EMERGENCY PLAN

Ding Sun[a], Shuhe Zhao [a*], Zhaohua Zhang[a], Xinjie Shi[a]

[a] School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing 210023, China

Commission IV, WG IV/4

Commission III, WG III/8

Commission V, WG V/4

KEY WORDS: Earthquake, Plan, Latent Semantic Analysis (LSA), Keywords Extraction, Match, Vector Space

ABSTRACT:

The structure of the emergency plan on earthquake is complex, and it’s difficult for decision maker to make a decision in a short time. To solve the problem, this paper presents a match method based on Latent Semantic Analysis (LSA). After the word segmentation preprocessing of emergency plan, we carry out keywords extraction according to the part-of-speech and the frequency of words. Then through LSA, we map the documents and query information to the semantic space, and calculate the correlation of documents and queries by the relation between vectors. The experiments results indicate that the LSA can improve the accuracy of emergency plan retrieval efficiently.

1. INTRODUCTION

Once the earthquake happened, the consequences will be very serious, and emergency plans are the important ways to organize the disaster relief quickly. As it is indicated, the decision-making time could be reduced by 60% if we adopt emergency decision-making method based on plans and the results improved significantly. Therefore, it has been an important issue that how to match the best plan to deal with the emergency from pre-plan database fast and accurately. At present researches on emergency plan matching are mainly case-based distance method (Chen, 2008), and the Jaccard coefficient (Nayak, 2007).

Plans are mostly represented as documents, so the plan matching actually can be converted to the documents matching. In the matching process of documents, the method based on vector space model ignores the correlation between words occurring in the documents, and because this method compute the document similarity according to the number of common words, it is impossible to distinguish semantic ambiguity of natural language. Paper used Latent Semantic Analysis (LSA), which assumes that words that are close in meaning will occur in similar pieces of text. We think that the synonyms have similar semantic structures, and the polysemous words in different meanings must have different semantic structures. The semantic structure of words is reflected in the connection between the frequencies appearing in the documents. We can extract and quantize the semantic structure through statistical methods, then elimination the effects of synonyms and polysemous words, thus improving the accuracy of documents matching.

2. METHOD

The first step of the method is the text pre-processing on the set of documents. Firstly use segmentation tools to each document to word segmentation and part-of-speech tagging, and calculate the frequencies of words. And then extract the keywords. The result of keyword extraction will directly influence the effect of presentations of documents. After the keyword extraction, the Term-Document matrix can be generated based on its results. The next step is singular value decomposition (SVD) of the Term-Document matrix. Finally the documents are expressed as Low dimensional in the semantic space. With the results of SVD, we can also get the vectors representing queries of users in the semantic space. It is easy to calculate the similarities between documents and queries by calculating the angle between the vectors.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-137-2017 | © Authors 2017. CC BY 4.0 License.

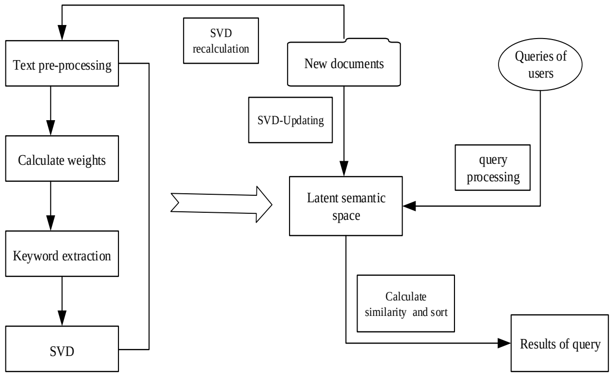

Figure 1. The match method based on LSA

2.1 Keywords extraction

This paper used an open-source word segmentation tool called Institute of Computing Technology, Chinese Lexical Analysis System (ICTLAS). After word segmentation and part-of-speech tagging, we need to get rid of some redundant information from the documents, including punctuation and stop words which did not contribute to the documents. Stops words include function words, prepositions, articles, interjections, conjunctions, pronouns, and so on. After the above processing, we should extract keywords which will be used to generate the Term-Document matrix according to their weight values. We choose the following two factors as the feature to calculate the weight values of keywords:

1) TF-IDF: In general, if a word appears in a document of high frequency, and rarely appears in other documents, we think this word has a very good ability to categories. TF-IDF is one of the standard effectively reflect whether the word could reflect the themes; the formula is as follows:

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}*\log\frac{|D|}{|\{t\in d|d\in D\}|+1} \tag{1}$$

where TF= Term Frequency
$|D|$=numbers of all documents
$|\{t\in d|d\in D\}|$ =numbers of documents that contain the word

2) Part of speech (POS): Emergency plan is a scheme of administrative significance, and it is official document. We think that nouns can reflect the themes of the plan better. Use the following formula, calculate the POS factor type of word i :

$$\mathrm{type}=\begin{cases}1,\mathrm{noun}\\0.8,\mathrm{verb}\\0.6,\mathrm{other}\end{cases} \tag{2}$$

All the weight value calculation of keyword is based on the two factors described above, the weight value of word i is calculated as follows:

$$\mathrm{weight}_{ij}=\mathrm{TF}-\mathrm{IDF}_{ij}*\mathrm{type}_{\mathrm{ij}} \tag{3}$$

We can figure out the weight value of word i in document j according to above formula. Then select the n words with the larger weight values as keywords, and add them into the set of keywords of all documents. Finally we can get the Term-Document matrix which represents the set of all documents. The number n can be set as a constant value according to the need, or we can figure out it to scale with the number of words in documents.

2.2 Latent semantic analysis

LSA is a technique used in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms with "semantic vector spaces". The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. It is not traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, morphologies, or the like, and it takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways (Reidy, 2012). Each document can be seen as a point in the space with dimensions of words. A document containing the semantic, its distribution in the space is not absolutely random, but it obeys some kind of semantic structure. This semantic structure is hidden in the text, and has a potential effect on the appearance of words and the composition of documents. However, because of the factors such as the arbitrariness of word using and the uncertainty of the document topic, the semantic structure is often buried in the "noise". In order to achieve the purpose of information extraction and removal noise, LSA uses SVD to rank lowering. The representation of document in LSA is different from it with high dimensions in traditional vector space model (VSM), but is projected in potential semantic space with low dimension. The method reduces the problem and makes the sparse data became no longer sparse, to rendering out some potential semantic structure.

According to the result of keywords extraction, we can represent the set of documents as a Term-Document matrix. The Term-Document matrix is a sparse matrix in which each row stands for a unique word and each column stands for a document. We can use the weight values figured out in keywords extraction as the elements of the matrix. And the size of the matrix is m by n, m is the number of extracting keywords, n is the number of documents.

Next, LSA applies SVD to the matrix. This is a form of factor analysis, or more properly the mathematical generalization of which factor analysis is a special case. In SVD, a rectangular matrix is decomposed into the product of three other matrices. One component matrix $\mathbf{U}$ describes the original row entities as vectors of derived orthogonal factor values, another $\mathbf{V^T}$ describes the original column entities in the same way, and the third matrix $\mathbf{\Sigma}$ is a diagonal matrix containing scaling values such that when the three components are matrix multiplied, the original matrix is reconstructed. There is a mathematical proof that any matrix can be so decomposed perfectly, using no more factors than the smallest dimension of the original matrix. When fewer than the necessary number of factors are used, the reconstructed matrix is a least-squares best fit. The formula of SVD is as follows:

$$\mathbf{X} = \mathbf{U\Sigma V^T} \tag{4}$$

The values on the diagonal of $\mathbf{\Sigma}$ are also called singular values, in order from largest to smallest, corresponding to the importance of the columns of $\mathbf{U}$ and the rows of $\mathbf{V^T}$. After the singular value decomposition, smaller singular values and the corresponding columns and rows of the matrix $\mathbf{U}$ and $\mathbf{V^T}$ are removed for the purpose of removing noise. To retain the maximum k singular value is equivalent to retain the most important semantic information and removed the useless information. And the number k is called dimension reduction factor.

When we choose the largest k singular values, and their corresponding $\mathbf{U}$ and $\mathbf{V^T}$, and they are matrix multiplied, we can get an approximate matrix of the original matrix :

$$\mathbf{X_k} = \mathbf{U_k \Sigma_k V_k^T} \tag{5}$$

The value of reduction factor k is directly related to the efficiency of semantic space model. If the value of k is too small, it will make some useful information lost. But if in contrast the value is too large, it will increase computational cost and could not remove the “noisy”. According to different sets of documents and processing requirements, the best values of k are not the same. When select the value of k, if $\mathbf{\Sigma}$=diag(σ_1, $\sigma_2 \cdots$, σ_n), and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r = \cdots = \sigma_n = 0$, we can let k meet the contribution rate inequality:

$$\sum_1^k \boldsymbol{\sigma_i} / \sum_1^r \boldsymbol{\sigma_i} \geq \boldsymbol{\theta} \tag{6}$$

where $\boldsymbol{\theta}$ = threshold of raw information

After a truncated singular value decomposition, you can do the following things:

1) Calculates the similarity between documents:

Comparison of vector $\mathbf{\Sigma_k}\boldsymbol{D_i}$ and vector $\mathbf{\Sigma_k}\boldsymbol{D_j}$ (Such as cosine of angle) shows similarity of the document i and j in the low-dimensional space. ($\boldsymbol{D_i}$ and $\boldsymbol{D_j}$ are the column i and j of matrix $\mathbf{V_k^T}$)

2) To retrieve documents according to the requests of user:

The requests of user can be words or documents. After preprocessing for user queries, generate a query vector q according to word frequency information, and express it in a semantic space with k dimension. The expression of the query vector q in the semantic space as follows:

$$\boldsymbol{\hat{q}} = \mathbf{\Sigma}_{\boldsymbol{k}}^{\mathbf{-1}} \mathbf{U_k^T} \mathrm{q} \tag{7}$$

So that you can figure out the similarity between the queries and documents, and sort the documents according to similarity, and then select related documents according to a threshold.

2.3 The renewal of latent semantic space

When new texts or words are added, latent semantic space can be updated in two ways that could be chosen by users according to their needs.

1) When the amount of newly added text is less than the original text, it is too costly to regenerate Term-Document Matrix and execute SVD. In this case, we adopt the SVD-Updating algorithm (Berry, 1995) to update latent semantic space fast and approximately.

Define A which is composed of p document vectors as a collection of new additions. Each document vector is still computed from the frequency and part of speech in the original latent semantic space in the document. A is a sparse matrix whose size is m by p. Adding A to the back of the approximate matrix named $\mathrm{X_k}$ whose size is m by n and rank is k, we get the Y matrix(Y =($\mathrm{X_k}$|A)) whose size is m*(n+p).

Define Z as a matrix whose size is k*(k + p) （Z=(Σ_k|$\mathrm{U_k^T}$A)） . By the singular value decomposition (SVD) of Z, we can get the result Z= $\mathrm{U_Z \Sigma_Z V_Z^T}$; according to the following formula, the singular value decomposition formula of matrix Y can be obtained:

$$\mathbf{U_Y}=\mathbf{U_k}\mathbf{U_Z} \tag{8}$$

$$\mathbf{\Sigma_k}=\mathbf{\Sigma}_Z \tag{9}$$

$$\mathbf{V_Y}=\begin{Bmatrix}\mathbf{V_k} & 0\\ 0 & \mathbf{I_p}\end{Bmatrix}\mathbf{V_Z} \tag{10}$$

2) When the amount of newly added text is more than the original text, the correlation of word vectors varies greatly in the new and old latent semantic spaces, what could ensure the accuracy of new latent semantic space is only SVD recalculation so far.

3. EXPERIMENTAL RESULTS

We experiment on the corpus published by the People's Daily. In the corpus of the People's Daily, there are more than 3000 news articles published in January 1998, and 800 of them were selected after screening. After filtering Stop Words, each document extracts keywords in accordance with the 20% of al all words, generating 8000 * 800 Word-Text Matrix, executing SVD to the matrix and generating the low dimensional semantic space. And the dimension reduction factor k was set as 100.

When the number of relevant documents is set to a smaller threshold, such as 10, the results based on LSA are similar to those based on the traditional vector space approach, and the average accuracy of both methods is more than 80%. However, when the number of relevant documents is set to a bigger threshold, the accuracy of traditional vector space decreases sharply, while LSA still maintains good results. When the threshold is set to 40, the retrieval accuracy of the traditional vector space is less than 50%, while the LSA can still maintain more than 70%. The results of experimental show that the retrieval method based on LSA is better than the method based on traditional VSM.

4. CONCLUSIONS

In this paper, LSA is introduced to realize the accurate representation of emergency plan in semantic space and to improve the accuracy of retrieval of documents. The TF-IDF and POS are chosen as the two factors to calculate the weight value of the words. The keywords are extracted according to the calculation results of weight, and the Term-Document Matrix is generated; after applying SVD to the matrix, we will eventually map the documents of plans to low dimensional semantic space. Not only could it retain the effective information, but also eliminate the noise. According to the singular value decomposition results, the user query is mapped to the same low semantic space to achieve retrieval of plans after completing the similarity calculation by the angle between the vectors. Compared with the traditional vector space model, this method can eliminate the influence of synonyms and polysemous words to a certain extent and improve the accuracy of the plan retrieval.

ACKNOWLEDGEMENTS

The research work was supported by the National Key R&D Program of China (No.:2016YFB0502503). It also was supported by the college students’ innovative program of Nanjing University (No.: X201710284009).

REFERENCES

Berry M W, Dumais S T, O'Brien G W, 1995. Using linear algebra for intelligent information retrieval. Society for Industrial and Applied Mathematics, pp.573-595.

Chen Y, 2008. A case-based distance method for screening in multiple-criteria decision aid. The International Journal of Management Science, 36(3), pp. 373-383.

Nayak, R, 2007, *Facilitating and Improving the Use of Web Services with Data Mining*. In: Taniar, D. (ed.) *Research and Trends in Data Mining Technologies and Application*, pp. 309–327.

Reidy P, 2012. An Introduction to Latent Semantic Analysis. Discourse Processes, 25(2), pp.259-284.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-137-2017 | © Authors 2017. CC BY 4.0 License.