AN INFORMATION SERVICE MODEL FOR REMOTE SENSING EMERGENCY SERVICES

Zhaohua Zhang[a], Shuhe Zhao[a,*] , Xian Li [a], Dianmin Cong[a], Ding Sun[a]

[a] School of Geographic and Oceanographic Sciences, Nanjing University, 163 Xianlin Ave, Qixia District, Nanjing 210023, China

Commission VI, WG VI/4
Commission III, WG III/8
Commission V, WG V/4

KEY WORDS: remote sensing emergency, information service, semantic network, words segmentation, preference analysis, PLSA

ABSTRACT:

This paper presents a method on the semantic access environment, which can solve the problem about how to identify the correct natural disaster emergency knowledge and return to the demanders. The study data is natural disaster knowledge text set. Firstly, based on the remote sensing emergency knowledge database, we utilize the sematic network to extract the key words in the input documents dataset. Then, using the semantic analysis based on words segmentation and PLSA, to establish the sematic access environment to identify the requirement of users and match the emergency knowledge in the database. Finally, the user preference model was established, which could help the system to return the corresponding information to the different users. The results indicate that semantic analysis can dispose the natural disaster knowledge effectively, which will realize diversified information service, enhance the precision of information retrieval and satisfy the requirement of users.

1. INTRODUCTION

China is one of the countries which have the most frequent and severe natural disasters in the world. Due to the enormous data size, valid data cannot be retrieved accurately, which will disturb the disaster analysis.

In recent years, many studies focus on the semantic analysis. For example, Moro *et al* utilized the wikipeadia to analyze the ontological relations between sematic information. Speer *et al* established the semantic network named “Concept Net” in 2013, which mainly describe the hierarchical relation about different vocabularies. Harrington *et al* also measured the distance between vocabulary vectors in semantic space and the distance represents the similarity of the different words.

This paper utilizes the semantic analysis in retrieval and recommendation about emergency data in order to enhance the accuracy and efficiency. The main research contents include three parts: remote sensing emergency semantic network, semantic analysis model and the information service model based on user preference.

2. DATA

The test data mainly includes emergency knowledge documents such as electric power, earthquake, terrorist activities emergency methods. The expert knowledge mainly includes disaster emergency database which is supplied by Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences. The former is the basic data and the latter is utilized to establish the semantic network and extract the key words in emergency

[*] Corresponding author: Dr. Shuhe Zhao, associate professor, E-mail: zhaosh@nju.edu.cn

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-187-2017 | © Authors 2017. CC BY 4.0 License.

knowledge dataset.

3. METHOD

The research method goes as following. First, in order to segment the source documents, we combine the word segmentation system with the emergency knowledge database to extract the key words in those unprocessed documents.

Second, according to accurate matching in remote sensing emergency knowledge, this paper establishes a semantic analysis model. In view of document, topics and words, our research establishes the semantic analysis model to realize the accurate emergency services about multivariate information of time and space and provide precision information to the users.

Third, aiming at recommending the information according to the users' preference, our study utilizes the machine learning algorithm and user searching record to establish the information service model based on user preference which can achieve the result of exact and quick recommendation to satisfy the different users' preference.

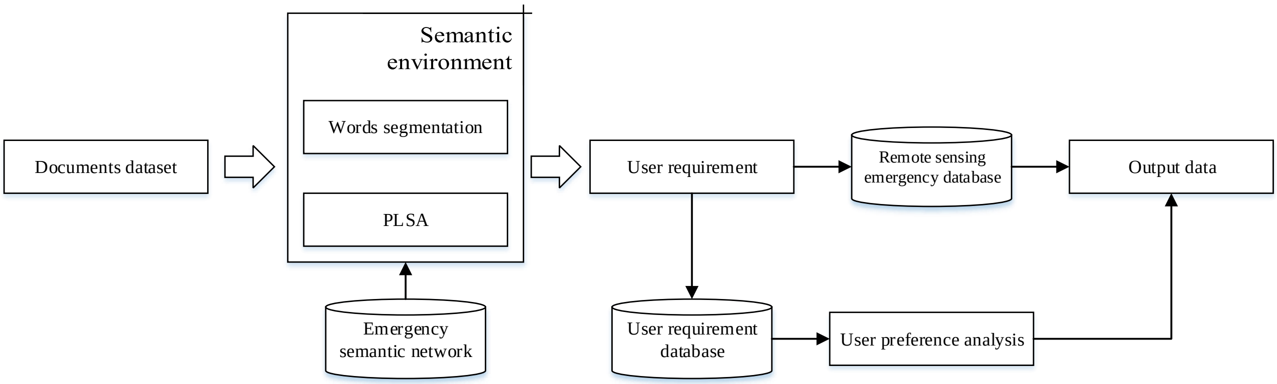

Figure 1. Study flow chart

3.1 Emergency semantic network

Utilize expert knowledge in the remote sensing emergency database to extract the key words and map the words with the documents, we will form a complicated relation network.

3.1.1 Words segmentation: The words segmentation problem should be concerned first. Chinses text is different from the Western language which can be segmented by those symbols such as blank. Besides, there are lots of problems such as ambiguity and unfamiliar words should be considered.

Many researchers studied on the Chinese words segmentation system. We utilize the "ICTCLAS System" which is an open source program designed by the Institute of Computing Chinese Academy of Sciences.

3.2.1 Key words extraction: Key words are the most representative character string unit in texts, such as names, place names, technical terms and so on. They express the main ideas in their articles. Therefore, the result of the key words extraction has an important impact on the text retrieval and classification. We utilize the key words collections provided by disaster emergency knowledge database to count the vocabularies which are similar to those key words in the texts. The main idea of the key words extraction is below:

1. Put the key words collections into the directory of the "ICTCLAS" system.
2. Utilize the updated "ICTCLAS" to segment the documents dataset.
3. Calculate the frequency of the similar words in each document.
4. Merge the frequency of the same words in different documents.

3.2 Semantic analysis model

Here we utilize the Probabilistic Latent Semantic

Analysis(PLSA) to establish the mapping relation between documents, topics and words. PLSA is put forward by T. Hoffmann.

3.2.1 Standardization of semantic weight: Our research measures the frequency of those key words we extract in 3.1.2 according to every documents so that we can obtain a $\mathbf{N} \times \mathbf{M}$ "document - word" matrix N(d, w). This matrix reflects the main content of the document to some degree, but many high-frequency vocabularies also cannot reflect the correct topic of the documents. Therefore, we should standardize the "document - word" matrix in order to feature the key words which can imply the topic of the articles.

This paper utilizes the "tf-idf" formula to standardize the matrix. The formula is below:

$$\mathrm{N(d,w)} = \mathrm{idf(term)}\sqrt{freg} \tag{1}$$

$$\mathrm{idf(term)} = \ln\left(\frac{numDocs}{docFreg+1} + 1.0\right) \tag{2}$$

Where *freg*=the frequency of the key words in the documents.

docFreg=the number of the documents which contain the key words.

numDocs=the total number of the documents.

3.2.2 PLSA: PLSA assumes every article is mixed by the random topics. Different words may correspond to different topics. PLSA structure is below:

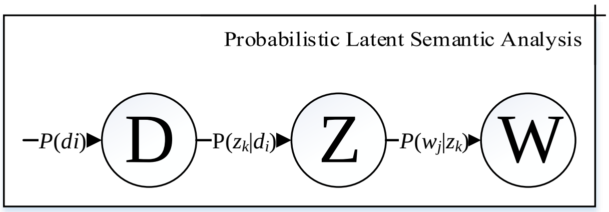

Figure 2. PLSA structure

Where D = documents; Z = topic; W = words.

P(di) = the probability of choosing the documents.

P(z_k|d_i) = the probability of the words in topic z_k existing in document d_i.

P(w_j|z_k) = the probability of the words belonging to the topic z_k.

Every topic follows the multinomial distribution on the words and every document follows the multinomial distribution on the topic. The detailed procedures are below:

a. Select the document d_i in probability *P*(*di*)

b. Select the topic z_k in probability P(z_k|d_i)

c. Select a key word in probability *P*(w_j|z_k)

As the result, we can get the joint distribution about pairs of ($d_{i,}$ w_j):

$$\mathrm{P}(d_i, w_j) = \mathrm{P}(d_i)\mathrm{P}(d_i|w_j) \tag{3}$$

$$\mathrm{P}(w_j|d_i) = \textstyle\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{4}$$

PLSA utilizes Expectation Maximization Algorithm(EM), which is an iterative method to find maximum likelihood, to estimate P(zk|di) and P(wj|zk). The detailed procedures are below:

a. E step: Utilize the estimated value of the latent variable to calculate the maximum likelihood value.

$$P(Z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{I-1}^{K} P(w_j|z_I)P(z_I|d_i)} \tag{5}$$

b. M step: Maximize the maximum likelihood value calculated through E step to estimate the parameter. This paper utilizes formulas (6) and (7) to reestimate the model.

$$P(w_j|z_k) = \frac{\sum_{i-1}^{N} n(d_i|w_j)P(z_k|d_i, w_j)}{\sum_{m-1}^{M}\sum_{i-1}^{N} n(d_i, w_m)P(z_k|d_i, w_m)} \tag{6}$$

$$P(w_j|z_k) = \frac{\sum_{j-1}^{M} n(d_i, w_m)P(Z_k|d_i, w_j)}{n(d_i)} \tag{7}$$

c. The parameter we get in the M step will be utilized in next E step. E and M step will iterate until the result become convergence. This paper utilizes the formula below to check the convergence condition:

$$E(L) = \sum_{i-1}^{N}\sum_{j-1}^{M} n(d_i, w_j)\sum_{k-1}^{K} P(z_k|d_i, w_j)\log_n[P(w_j|z_k)P(z_k|d_i)] \tag{8}$$

This paper utilizes the final result of P(zk|di) and P(wj|zk) to construct two matrixes: U=(P(zk|di))K,I, V=(P(wj|zk))J,K. U represents the probability distribution of latent semantic relations(topics) in the documents and V represents the probability distribution of words in the latent semantic relations.

3.3 User preference model

The model will reserve the machine recognizable information and generate a database. Then, utilize the function to filter the database and select valid information to establish the user sample database in the information service model based on user preference.

In order to cluster the information prepared to be pushed, this paper utilizes the "documents-topics" matrix in semantic analysis to calculate the probability center vector of the different documents. The included angle between different documents vectors will express the discrepancy between documents. The formulas are below:

$$\mathrm{sim}(Z_t, Z_c) = \frac{\sum_k P(Z_t|t)P(Z_k|c)}{\sqrt{\sum_k [P(Z_k|t)^2]}\sqrt{\sum_k [P(Z_k|c)^2]}} \tag{9}$$

$$\mathrm{P}(\mathrm{Z}|c_i) = (\mathrm{P}(Z_1|c_i), \mathrm{P}(Z_2|c_i) \ldots \mathrm{P}(Z_k|c_i))^T \tag{10}$$

Where t = unclassified texts
c = type of the texts
P(Z|ci) = probability center vector

To solve the problem that the information which users prefer is real-time adjusting, our research will establish the information service model mainly on two aspects including selecting the training sample and building the preference based on those training sample. First, our study utilizes the weighted time decay function to filter the sample and select the most representative data during the recent time. Then, utilize the BP neural network algorithm to train the filtered sample data and mine the potential information which can represent the recent users' preferences.

4. RESULTS AND CONCLUSIONS

4.1 Results

The experiment texts we utilized are mainly provided by China Electric Power Research Institute and other texts is searched on the Internet and totally we get 127 documents. Those documents belong to 4 topics including earthquake, power emergency, people activities and terrorist attack. This paper utilizes the key words in the emergency knowledge database to extract the words in those documents and finally we obtain 687 key words.

In order to express the result, due to the huge "topics - words" matrix, this paper only can present parts of the matrix as the result which shows the words with top ten probability in each topics.

"1"	"2"	"3"	"4"
Earthquake shockproof Aftershock Shock resistance Evacuate waggle	Electricity Power grid System Power cut Electric wire Voltage	Olympics Beijing Athlete Game International Equipment	Anti-terrorist Xinjiang Terrorist Security Police Weapon
Intensity scale Construction Communication precaution	tower Transmission Load cable	Crowd Traffic Sport Activity	Riot Patrol Street Government

Table 1. Parts of the "topics - words" matrix

In order to evaluate the result of the PLSA, this paper introduce the precision index and recall index as the evaluation indicators. The formulas are below:

$$precision\ ratio = \frac{correct\ classification\ texts}{reacall\ texts}$$

$$recall\ ratio = \frac{correct\ classification\ texts}{manual\ classification\ texts}$$

Calculate the average precision and recall ration of all categories of texts and the result shows that the precision ration is 82.57% and the recall ration is 78.87%.

Recall texts	109
Correct classification	90
Precision ratio	82.57%
Recall ratio	78.87%

Table 2. Precision and recall ratio

4.2 Conclusions

This paper selected the typical places prone to natural disasters as application demonstration.

First, we utilize the "ICTCLAS" combined with the emergency knowledge database to segment the source documents data so that we can extract the key words from those documents. Then, considering the view of the vocabulary, and text, the system utilizes the PLSA to establish semantic analysis model. Finally, establish the information service based on user preference which can cluster and recommend the preferred information accurately and efficiently.

According to the results, the system will be robust to achieve the function of the emergency requirements in the typical multiple disaster areas. Meanwhile, the system can analysis the requirements of the users accurately and recommend the relevant information, which will realize the rapid respond to the emergency.

ACKNOWLEDGEMENTS

The research work was supported by the National Key R&D Program of China(No.2016YFB0502503).

REFERENCES

Berry, M W., 1995. Using linear algebra for intelligent information retrieval. *Siam Review*, 37(4), pp. 573-595.

Davies, C., 2013. *Reading Geography between the Lines: Extracting Local Place Knowledge from Text*. Springer International Publishing, pp. 320-337.

Harrington, B., 2010. A Semantic Network Approach to Measuring Relatedness. In: *International Conference on Computational Linguistics*, Beijing, China, Posters Volume, pp. 356-364.

Hofmann, T., 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2), pp. 177-196.

Jones, C. B., 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10), pp. 1045-1065.

Klien, E., 2006. Ontology-based discovery of geographic information services—An application in disaster management. *Computers Environment & Urban Systems*, 30(1), pp. 102-123.

Lingpeng, Y., 2004. Document Re-ranking Basedon Global and Local Terms. In: *Proceedings of 3rd ACL SIGHAN Workshop*, Barcelona, Spain, pp. 17-23.

Marrero, M., 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5), pp. 482-489.

Moro, A., 2012. WiSeNet: building a wikipedia-based semantic network with ontologized relations. In: *Conference on Information and Knowledge Management*, ACM, pp. 1672-1676.

Speer, R., 2013. *ConceptNet 5: A Large Semantic Network for Relational Knowledge.* The People's Web Meets NLP. Springer Berlin Heidelberg, pp. 161-176.

Wan, R., 2009. Efficient Probabilistic Latent Semantic Analysis through Parallelization. In: *Asia Information Retrieval Symposium on Information Retrieval Technology*, Springer-Verlag, pp. 432-443.

https://doi.org/10.5194/isprs-archives-XLII-2-W7-187-2017 |