The Topic Tracking Based on Semantic Similarity of Sememe's Lexical Chain

Jing Ma, Fei Wu
College of Economics and Management
Nanjing University of Aeronautics and Astronautics
Nanjing, China
Email: {search,wufei}@nuaa.edu.cn

Chi Li
College of Mathematics
University of Science and Technology of China
HeFei, China
Email: lichs@mail.ustc.edu.cn

***Abstract*—In method of Semantic similarity calculating, the major is based on VSM(Vector Space Model).It has aroused significant research attention in recent years due to its advantage in topic tracking. In this paper a modified VSM, namely Semantic Vector Space Model, is put forward. To establish the model, numerous lexical chains based on HowNet are first built, then sememes of the lexical chains are extracted as characteristics of feature vectors. Afterwards, initial weight and structural weight of the characteristics are calculated to construct the Semantic Vector Space Model, encompassing both semantic and structural information. The initial weight is collected from word frequency, while the structure weight is obtained from a designed calculation method. Finally, the model is applied in web news topic tracking with satisfactory experimental results, conforming the method to be effective and desirable.**

Keywords-Topic tracking; Semantic Similarity;Vector Space Model; Lexical chain; Sememe

I. INTRODUCTION

Topic tracking is a method that mainly works to get the topic model on the basis of training corpus and then track the follow-up reports related to the topic. Vector Space Model (shorten as VSM, proposed by G. Salton, A. Wong, and C. S. Yang in the late 1960s) appears to be most popular and successfully applied in the famous SMART system. After that, the model and its related technologies, including selection of items, weight strategy and queuing optimization, had been widely used [1] in text classification, automatic index, information retrieval and many other fields, making it the mainstream model in topic tracking.

One of VSM's advantages is its knowledge representation. A document is transformed into a space vector, the document's operation is thus converted to the vector's mathematical operation, reducing the complexity of the problem. The semantic information of the text, however, is ignored by this method, which means the accuracy cannot be guaranteed. A proper solution here is to use external semantic knowledge to improve Vector Space Model. For example: Hu Jiming [2], Starting from mechanism analysis of user modeling Model. The effort helped add semantic information into VSM, but since the theory and technology research of ontology are not in-depth [3], they didn't solve the problem thoroughly. Jin Zhu [4], made full use of the ontology are not in-depth [3], they didn't solve the problem thoroughly. Jin Zhu [4], made full use of the external semantic resources—HowNet, to realize effective topic tracking and classify subject position on the basis of the information retrieval technology. Although she had considered the semantic meaning of the text, the structure information was neglected.

Lexical chain, put forward by Halliday and Hasan [5] first in 1976, is a kind of external behavior of the continuity of semantic relations between words, it has a corresponding relationship with the structure of the text, providing important clues of the structure and theme [6]. From what has been discussed above, the paper will introduce HowNet and lexical chains in the process of building model, constructing lexical chains based on HowNet. Then it will build a sentimatic vector space model of the topic based on sememe of the lexical chains, which inclued the semantic information and structure information of the text.

II. BUILDING VECTOR SPACE MODEL BASED ON THE LEXICAL CHAIN'S SEMEME.

A. *The extraction of the lexical chain based on HowNet*

HowNet is a commonsense knowledge base which describes the concept represented by Chinese and English words. It reveals the relationship between concepts and attribute of the concepts [7].In the literature [8], Morris and Hirst first introduced Lexical Chain concept, which is constructed to split the text to get the information of text structure. The lexical chain constructed in this paper is based on the semantic similarity, it also contains semantic information and structure information of the text. The lexical chain building steps are as below:

a) Use the ICTCLAS segmentation tools developed by Chinese academy of sciences to construct the word set with the automatic segmentation of text.

b) Select the first word from the set sequentially to build the initial lexical chain. Then select the candidate words sequentially. After that, compute the similarity between the candidate words and the chain if it Meet the threshold requirements. Finally insert the word into the current lexical chain or skip it if it does not meet the requirements.

c) Output current lexical chain and delete the words of the chain in the vocabulary,if the word set is empty then the process is accomplished. If not, switch to operation b).

d) Circulate the operation until the word set is empty.

B. *Building vector space model based on the lexical chain's sememe.*

Since this paper constructs lexical chain based on semantic similarity of words, semantic information of each word in the lexical chains is similar. Based on this, the paper extract the representative sememe from each lexical chain as characteristics of feature vectors. This paper use word frequency as initial weight of the characteristics and the structure weight is obtained from the designed calculation method. Finally, it use the

structure weight to adjust the initial weight of the characteristics to construct the semantic vector space model of the topic. T = (L1 LW1, L2, LW2, L3, LW3;..., Ln, LWn)

Ln represent the sememe of the chain, LWn represent the weight of it. Below is the specific process.

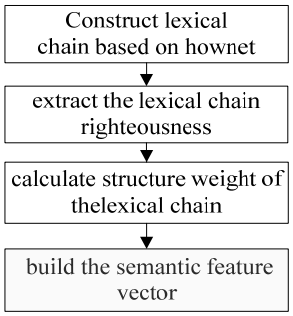

Figure 1. The construction of sememe vector space model

In this way, the vector will not only reduce the dimension of vector space, but also include the semantic and the structural information of the text.

III. THE DESIGN ABOUT THE ALGORITHM OF TOPIC TRACKING

Since our chosen corpus is for a specific topic, we took the TF (word frequency statistics) method to get the initial weights of feature, and lexical chains extracted from all the training corpus completely reveal the structure characteristics of the subject. Based on this, the topic tracking algorithm is designed as follows:

a) Extract the lexical chain and the sememe of it after doing word segmentation, part-of-speech tagging, and removing duplicate words of the topic training samples. Then use the sememe as characteristics of the VSM to constitute a semantic vector. The initial weights of the sememe is the sum of the weight of all the key words in the chain .The initial space vector of the topic is: T = (TW1, TW2,....., TWn).

b) Use the sememe to calculate the similarity between lexical chains. Set a threshold value and define the two lexical chains to be similar when the degree of similarity between lexical chains is greater than the threshold. Count the sum of the other chains which are similar with the current one and define it as “m”.

c) Each lexical chain structure weight is defined as TW=(m + 1)/S, m is the number of the other chains that are similar with the current chain, S is the number of the reports used for extraction of lexical chains. The final weight of each feature of the topic is the product of the initial weight and the structure weight of the lexical chain that has the feature, it is defined as tw = Tw * (m + 1)/S, thus the final vector of the topic is: T = (tw1, tw2,....., twn)..

d) Use the same method to deal with the subsequent reports, then the vector of the reports will eventually be: d = (dw1, dw2,....., dwn).

The paper take the cosine formula of the vector to compute the similarity between the topic and the follow-up reports. The formula is as follows:

sim(T,d)

=cos<T,d>

$$= \sum_{i,j=1,}^{i,j=n.} tw_i * dw_j \Big/ \sqrt{\sum_{i=1}^{i=n} tw_i * tw_i} \sqrt{\sum_{j=1}^{j=n} dw_j * dw_j}$$

T is for the subject; D is for later reports; Twi represent the weight of the ith feature of the topic; dwj represent the weight of the jth feature of the subsequent reports.

e) For each subsequent reports, use the similarity model described above to compute the similarity between the topic and later reports: sim (T, d), when the similarity is greater than the threshold, define them as similar. The specific process is shown in figure 2:

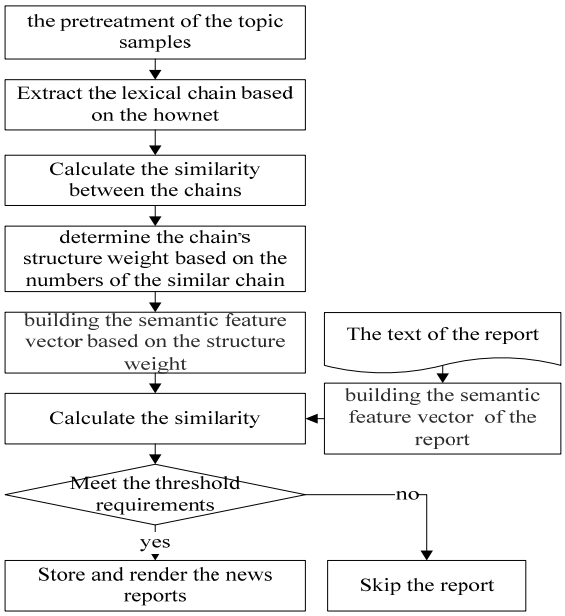

Figure 2. The Algorithm of Topic Tracking

IV. EXPERIMENTS AND RESULTS

This article selects three topics—the H7N9 treatment of bird flu, Syrian refugees, wasp stings—to do the experiments. Based on the operation above,three topic righteousness original feature vectors space are obtained as follows:

a) ***The treatment of H7N9*** (H7N9, bird flu, adjust, cure, published, drug, laces, eliminate, show, disease , know, Property, people, monitor, agency)

b) ***wasps hurt*** (place,dead,cure,worm, organization, people, against, time, damage, parts,tell, bad thing, using, eliminate, check, understand, form, work on)

c) ***Syria's refugee*** (represent, countries, struggle realize,agency, people,rescue, phenomenon (difficult) avoid ,enter,appear, records,situation, increase)

Then after the calculating the vectors are as follows:

- T H7N9 =5.1, 4.4, 0.2, 8.1, 0.8, 4.1, 0.7, 0.7, 0.3, 0.3, 3.6, 1.5, 0.4, 0.9, 0.3, 0.2）
- T wasp stings =(3.2, 1,3,17.2, 0.6, 3.2, 0.6, 2,0.6, 1.2, 1.4, 1,0.8, 0.4, 0.6, 0.4, 0.4, 0.4, 0.4).
- T Syria =(1.6, 15.4, 1.4, 1.6, 15.4, 13.2, 1,3,2.2, 0.4, 0.4, 0.4, 0.8, 0.6)

The paper then selects 5 similar report for each of the topic by domain experts to calculate similarity. Take the topic about H7N9 as example.

- *After processing, the characteristic vector space of one of the five reports is :*

(H7N9,bird flu,0,heal,published,drugs, 0,0,0,0,0,0,0,0).

- *After calculating, the feature vectors of the report is:*

t = (,0,0,0,0,0,0,0,0,0,2,0,3.6 2, 1.1, 3.6, 0).

According to the cosine formula of vector space model, the similarity is :

$$63.8/\sqrt{145.94*35.13}=89\%.$$

To verify the effectiveness of method, this paper uses the traditional vector space model to do an experiment as comparison. Still take the topic of H7N9 as an example. The characteristic vector space constructed by word frequency statistics is: (method 53, H7N9 51, bird flu 44, injection 39, people 32, infection 32, diagnosis and treatment 30,pharmaceutical 24, cases 19, company 18, country 18, varieties 17, flu 15, detection 15, control 15, Chinese medicine 15, prevent 13, virus 13, prevention and cure 12, health 11, kang yuan 11, lab 10, income 9, recommended 9, committee 8, products 8, published 8, diagnosis 7, medicine 7, Capsule 7, patients 6, traditional Chinese medicine 6, drugs 6, Selected 6, sales 6, program 6, control 6, think 5, use 5, the ministry of health 5, expert 4, Business 4,hospital 4, detoxification 4, contact 4, printing 4, state of an illness 4, samples 4, children 4, agency 4, Chinese patent drugs 4.

Then the vector is:

T=（53, 51, 44, 39, 32, 32, 30, 26, 24, 19, 18, 18, 17, 15, 15, 15, 15, 13, 13, 12, 11, 11, 10, 9, 9, 8, 8, 8, 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4）.

Then We use the words frequency method to construct the vector for the report used in the last experiment :

t=(3, 2, 2, 6, 0, 1, 3, 1, 1, 0, 5, 4, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 1, 1, 3, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0).

The similarity is: $1052/\sqrt{16118*131}=72.4\%$, obviously it is lower than the similarity calculated based on the lexical chain's sememe space vector. The similarity of three topics is in table I:

TABLE I DETAILS OF THE SIMILARITY

document ID	The similarity of the topic of H7N9·S treatment and prevention		The similarity of the topic of Syrian refugees		The similarity of the topic of wasp stings	
	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method
1	0.750	0.900	0.819	0.922	0.732	0.851
2	0.720	0.890	0.715	0.832	0.685	0.776
3	0.821	0.940	0.762	0.824	0.816	0.892
4	0.785	0.922	0.692	0.816	0.720	0.822
5	0.772	0.915	0.710	0.806	0.830	0.961

Label data in the coordinate system and connect the point with a straight line, we get figure 3:

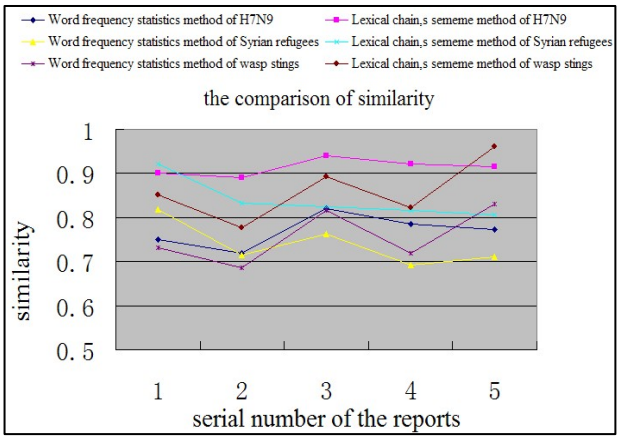

Figure 3 Cotrast of The Similarity

The serial number of the reports is on horizontal axis and the similarity is on the vertical. The picture shows that the similarity of new method is higher.

The paper downloaded 269,250,232 news reports for the three topics,the details is in table Ⅱ and set threshold value of 0.5 and 0.6 for two topic tracking, and take H7N9 as experiment to show the process.

TABLE Ⅱ DETAILS OF THE CORPUS METHOD

	the topic of H7N9·S treatment and prevention	the topic of Syrian refugees	the topic of wasp stings
total	269	250	232
related	49	40	32
unrelated	220	210	200

After tracking by using word frequency statistics, the result is shown in Figure4

Figure 4 The Result of Tracking by Using Word Frequency Statistics Method

There is a total of 43 records, including 39 related to the topic and 4 unrelated .After tracking by using lexical chain's sememe method, the result is shown in Figure 5:

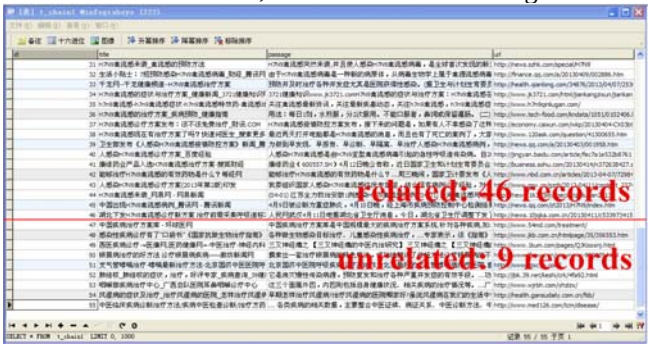

Figure 5 The Result of Tracking by Using Lexical Chain's Sememe Method

There is a total of 55 records, including 46 related to the topic and 9 unrelated . The detail of three topic's tracking result is in table Ⅲ.

TABLE Ⅲ THE DETAIL OF THREE TOPIC'S TRACKING RESULT(T=0.5)

	the topic of H7N9·S treatment and prevention		the topic of Syrian refugees		the topic of wasp stings	
	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method
total	43	55	34	44	26	37
related	39	46	31	36	23	29
unrelated	4	9	3	8	3	8

There is complete evaluation system, in which people use misdiagnosis rate PMiss and omissive judgement rate PFA to calculate the overhead of detection CDet, then normalize CDet to loss cost (CDet)Norm, which is the evaluation index of the topic tracking system. The smaller value of (CDet)Norm indicates the better system performance. The formulas are as follows:

$$C_{\mathrm{Det}} = C_{Miss} * P_{Miss} * P_{target} + C_{FA} * P_{FA} * P_{non-target}$$

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{Miss} * P_{target}, C_{FA} * P_{non-target})}$$

When the threshold is 0.6 the comparison and analysis of the result is in table V:

TABLE Ⅳ COMPARISON AND ANALYSIS OF THE RESULT(T=0.5)

	the topic of H7N9·S treatment and prevention		the topic of Syrian refugees		the topic of wasp stings	
	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method
P_{Miss}	0.20408	0.06122	0.22500	0.10000	0.28125	0.09375
P_{FA}	0.01818	0.04091	0.01426	0.03809	0.01500	0.04000
$(C_{Det})_{Norm}$	0.29317	0.26168	0.29487	0.28664	0.35475	0.28975

TABLE V COMPARISON AND ANALYSIS OF THE RESULT(T=0.6)

	the topic of H7N9·S treatment and prevention		the topic of Syrian refugees		the topic of wasp stings	
	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method	Word frequency statistics method	Lexical chain·s sememe method
P_{Miss}	0.32653	0.18367	0.37500	0.22500	0.40625	0.25000
P_{FA}	0.00455	0.01818	0.00476	0.01429	0.00000	0.01000
$(C_{Det})_{Norm}$	0.34883	0.27275	0.39832	0.29502	0.40625	0.29900

V. CONCLUSIONS

The paper extracts the lexical chains based on the external semantic resource—HowNet, then it takes the sememe of the chain as the feature to build the original feature vector. The weight of the feature is determined by the method of the word frequency statistics combined with the structure weight of the lexical chain, the semantic information and structure information of the text are also fully considered. In the topic tracking experiment system, the loss cost of the improved model is smaller, improving the efficiency of topic tracking.

CMiss=1, CFA=0.1, Ptarget=0.02, Pnon-target=1-Ptarget.

PMiss and PFA are both as small as possible. Their formulas are as follows:

$$P_{Miss} = \frac{The\ number\ of\ related\ reports\ that\ system\ does\ not\ recognize}{The\ total\ number\ of\ related\ reports\ in\ corpora} * 100\%$$

$$P_{FA} = \frac{The\ number\ of\ reports\ that\ system\ misjudges\ as\ related\ to\ the\ topic}{The\ total\ number\ of\ unrelated\ reports\ incorpora} * 100\%$$

PMiss and PFA are both as small as possible. The comparison and analysis of the result is in table IV, V:

From table IV, V, one can indicated that the non-response rates of the lexical chain's sememe is lower than the rates of the word frequency statistics method, but the rate of false positives is higher than that based on word frequency statistics method. Above all, the wastage of the approach based on the lexical chain's sememe is lower than the loss cost based on word frequency statistics method, proving the topic tracking algorithm based on the lexical chain's sememe to be effective.

ACKNOWLEDGEMENT

Footnotes: This paper is supported by the project supported by the National Natural Science Foundation of China (No.71373123)

REFERENCES

[1] YANG Yim-ing, CARBONELL J, BROWN R, et al. Learning Approaches for Detecting and Tracking News Events[J]. IEEE Intelligent Systems : Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32-43.

[2] Hu Jiming ,Hu Changping. The user modeling based on topic hierarchy tree and semantic vector space model [J]. Journal of intelligence, 2013.8, 32 (8) : 838-843.

[3] Beydoun G,Lopez—Lorca A A,et a1. How do we measure and improve the quality of a hierarchical ontology?[J]. Journal of Systems and Software,2011,84 (12): 2363—2373.

[4] Jin Zhu Lin Hongfei. Topic tracking and tendency analysis based on HowNet [J]. Journal of intelligence, 2005, 24 (5) : 555-561.

[5] Halliday M A K, Hasan R. Cohesion in English. London, UK: Longman, 1976.

[6] Gonenc E,Ilyas C. Using Lexical Chains for Keyword Extrac—tion[J]. Information Processing and Management,2007,43(6): 1705—1714.

[7] HowNet[R].HowNetsHome.Page.HTTP://WWW.keenage.com.

[8] J Morris, G Hirst. Lexical Cohesion Computed by Thesauralrelations as all Indieator of the Structure of Text[J]. Computational Linguistics, 1991, 17(1): 21-48.