OPEN ACCESS

RESEARCH ARTICLE

The Topic Tracking Based on Modified VSM of Lexical Chain'S Sememe

Jing Ma*, Fei Wu*, Chi Li **, Hengmin Zhu***

*(College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China)

** (College of Mathematics, University of science and technology of China, Hefei, China)

*** (College of Economics and Management, Nanjing University of Posts and Telecommunications, Nanjing, China)

ABSTRACT

Vector Space Model (VSM) has aroused significant research attention in recent years due to its advantage in topic tracking. However, its effectiveness has been restrained by its incapability in revealing same-concept semantic information of different keywords or hidden semantic relations of the text, making the accuracy of topic tracking hardly guaranteed. Confronting these issues with concern, 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: Each lexical chain structure weight is defined as (m + 1)/S, m is the number of the other similar chains, and S is the number of the reports used for extraction of the lexical chains. 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, Vector Space Model, Lexical chain, Sememe

[1] 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. It gathers isolated information scattered in different time and places to demonstrate full details of events and relationships between them ^[1]. Since documents written by natural language can hardly be comprehended by computer, a mathematical representation of the document model is required to be defined to realize document processing by computer. Along with this methodological thinking, several approaches are presented, such as Boolean Model, Vector Space Model and probability Model the conceptual Model, etc., among which 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 ^[2] 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^[3], Starting from mechanism analysis of user modeling based on semantic hierarchy tree, they used domain ontology to accomplish resource description and user modeling. Thereby building a Semantic Vector Space Model. The effort helped add semantic information into VSM, but since the theory and technology research of ontology are not in-depth ^[4], they didn't solve the problem thoroughly. Jin Zhu ^[5],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^[6] 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^[7].

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 include the semantic information and structure information of the text. Finally when applied into Sina Weibo topic tracking, the experiment proved that the method is effective.

[2] BUILDING VECTOR SPACE MODEL BASED ON THE LEXICAL CHAIN'S SEMEME

2.1 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 ^[8]. In the literature ^[9], 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:

- (1) Use the ICTCLAS segmentation tools developed by Chinese academy of sciences to construct the word set with the automatic segmentation of text.
- (2) 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 meets the threshold requirements. Finally insert the word into the current lexical chain or skip it if it does not meet the requirements.
- (3) 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 (2).
- (4) Circulate the operation until the word set is empty.
- Specific process is in Fig. 2.1

Lexical chain build pseudo code is as follows:

- K = 1; / / K's initial value is 1
- LK[] = { }; / / chain's initialization

$$Count = 0$$
:

The Word [] = (W1, W2, W3, ..., Wn); / / participle

Void LexicalChainBuilding(Lk)

 $\{Lk[0] = Word[0]; / / treat the first word as the initial value of lexical chain, k is the lexical chain's serial number.$

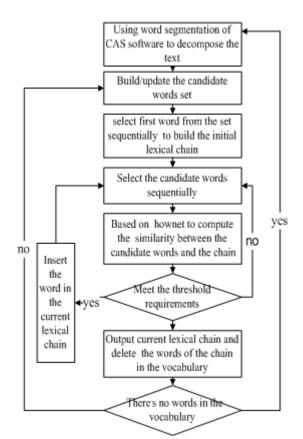


Fig. 2.1 The extraction of lexical chain

For(int i=1; i < = word.length; i++)

{if (SimilaryCompare (word [0], word [i]) > 0.5); / / if the similarity between the two words is greater than 0.5

LK [+ + Count] = word [i]; / / insert to the current chain

Out. Print (LK); / / output current lexical chain

DeleteChainFromWord(Lk); / / delete the words of the chain in the vocabulary

JudgeWordIsEmpty (); / / determine whether the current word set is empty

{if (JudgeWordIsEmpty () = = true)

Break; / / the end

) else

(

K++;

UpdateWord(); / / update the word in the vocabulary

Void LexicalChainBuilding(Lk); / / recursive call lexical chain building program

2.2 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 extracts 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 uses 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.

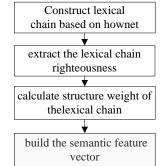


Fig.2.2 the construction of sememe vector space model

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.

[3] 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 training corpus completely reveal the structure characteristics of the subject. Based on this, the topic tracking algorithm is designed as follows:

(1)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).

(2)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".

(3) 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 chains 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).

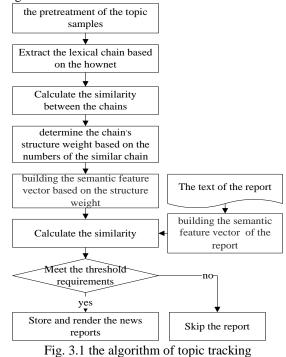
(4)Use the same method to deal with the subsequent reports, then the vector of the reports will eventually be: d = (dw1, dw2, ..., dwn).

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

i

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

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 Fig.3.1:



www.ijera.com

[4] 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, places, eliminate, show, Property, people, monitor, agency, disease, know) B wasps hurt:(place, dead, cure, worm, people, organization, 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:

TH7N9= (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, 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 reports 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: $t = (,0,0,0,0,0,0,0,0,0,0,2,0,3.6\ 2,\ 1.1,\ 3.6,\ 0)$

According to the cosine formula of vector space $\sqrt{145.04 * 25.13}$

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, 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, 4)

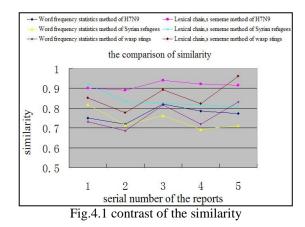
Then we use the words frequency method to construct the vector for the report used in the last experiment: t = (3x, 2, 2, 6, 0, 1, 3, 1, 1, 0, 5, 4, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 1, 1, 3, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)

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 4.1:

Table 4.1 details of the similarity

document ID	The similarity of H7N9-S treatmen prevention		The similarity Syrian refuge	y of the topic of es	The similar wasp stings	ity of the topic of
	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 Fig. 4.1:



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. In order to further verify the superiority of the algorithm, this paper has designed a topic tracking experiment system. The system mainly includes the following three parts: the pretreatment of network reports, solution selection and topic tracking. The solution selection module is organized by the construction of lexical chain's sememe and word frequency statistics. The detail is in Fig. 4.2:

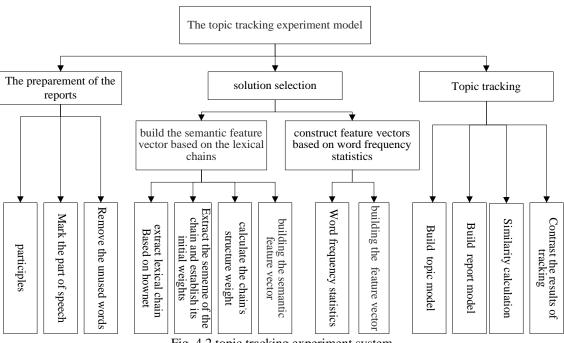


Fig. 4.2 topic tracking experiment system

The TDT established a complete evaluation system which uses the rates of non-response PMiss, the rate of false positives PFA and the loss cost (CDet) Norm of the system to indicate the performance of the system.

The paper downloaded 269, 250, 232 news reports for the three topics, the details is in table 4.2.

The paper set threshold value of 0.5 and 0.6 for two topic tracking, and take H7N9 as experiment to show the process.

	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

Table 4.2 details of the corpus

After importing the data into database, the detail of the initial database is in Fig. 4.3:

After tracking by using word frequency statistics method, the result is shown in Fig. 4.4:

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 Fig. 4.5:

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 4.3:

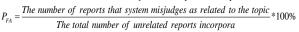
There is complete evaluation system, in which people use misdiagnosis rate PMiss and omissive judgement rate PFA to calculate the overhead of detection CDet, then normalize CDet to loss cost (CDet)Norm, which is the evaluation index of the topic tracking system. The smaller value of (CDet)Norm indicates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

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

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:

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



PMiss and PFA are both as small as possible. The comparison and analysis of the result is in table 4.4, 4.5:

(E) 綱指	(E) 查看(V) 窗[I (¥)										
🗑 备注 🛄	🛾 十六进位 🔣 🛙	日像 🛛 😤 升幕	排序 🏓 降	幕排序 🔧 移除排序								
	tit				passage				url			
				赴方 禽流感 中药 H7N9_								
				早期服用中成药 H7N9								
	3 陇	i台H7N9禽流感:	云南方案公布	资讯频道_凤凰网	2013-4-10 · 记书	者了解到,目前,	机场方面已下发《	昆明机场疾病预防招	浩明所 http://news	.ifeng.com/gundr	ong/detail_2013	04/10
	4 陇	i台H7N9禽流感:	云南方案公布	_云南网	订阅《春城手相	机报》:综合版发	送CCZH到10658000	(5元/月) 娱乐版2	ģ送O http://societ	y.yunnan.cn/htm	nl/2013-04/10/cc	ontent
				医药防治方案_中药配方,								
	6 H	719禽流感中医药	市防治方案发	布+H7N9禽流感,中医药-制	过【慧聪制药工」	业网】记者今天从	省中医药管理局获	悉,《湖南省人感染	H7NS http://info.p	harmacy.hc360.	com/2013/05/10	09434
	7 🕻	央广独家】备战	8H7N9禽流感	多地出台防治方案_中国	广江西:未在动物	物中发现H7N9禽流	感病毒 暂不关闭法	禽交易市场 日前,	江西i http://native	t.cnr.cn/city/201	304/t20130407	51230
		719禽流感中医药						感中医药防治方案,				
		沙小学H7N9禽家						好我校H7N9禽流感				
	10 H	7N9禽流感中医药	胡防治方案公:	前24小时滚动新闻 - 人民	民 H7N9禽流感中[医药防治方案公布	本报讯 4月10日	,济南市中医管理局;	公布了http://www.	people.com.cn/2	4hour/n/2013/04	418/c:
	11 福	建发布人感染H	719禽流感中[医药防治方案 - 新浪新闻	中新网福州4月	12日电 (龙敏)福建	省卫生厅12日正式	发布了《福建省人》	感染 ⊦ http://news	.sina.com.cn/c/2	013-04-13/00442	26811
	12 H	m9禽流感防治应	立急预案 · 百周	之 文库	举报文档 H7N9	會流感防治应急到	「案 乡镇卫生院 专	目人感染H7N9禽流标	感防祥 http://wenk		/982a2ed06f1a	ff00be
	13 福	建发布人感染H	7N9禽流感中[医药防治方案_资讯频道_	戶中新网福州4月	12日电(龙敏)福建	省卫生厅12日正式	发布了《福建省人员	5染H:http://news	ifeng.com/mainla	and/special/h7n9	/conte
	14 益	战H7N9禽流感;	多地出台防治	方案_财经_中国网	浙江:发布人员	影染H7N9禽流感中	医药防治建议方案	方案根据浙江地区的	的特性 http://financ	e.china.com.cn/	news/dfjj/20130	407/1
				业方 禽流感 中药 H7N9_]								
	16 魏	间预防H7N9需制	和感?盘点各	也禽流感预防指南_健康务	质 这一方案提供	了預防、诊治人感	染H7N9常和 5的 多	个 PL 方,称板蓝机	限冲资[http://health	.xinn h.cn/ikzx/	2013/04/09/196	38633
		7N9禽流感的预除			(三)习惯预防 1	.保持手部清洁,	中用正确方法 手	2.11.免手計接給能	af th p:www.	t the post om/k	ndata/1051/010	2405.1
	18 湖	江发布人感染H	7N9禽流感中[医药防治建议方案_资讯》	质 2013-4-6 · 为充	分发挥中医药在除	5治人感染H7N9常家	版明中的作用,省中	医药作 http://news	.ifeng.com/mainla	and/special/h7n9	/conte
	19 #	京中医局公布孙	助H7N9禽流。	感中药处方_新闻_腾讯网	2013-4-5 · 市中	医局昨发布人感染	H7N9禽流感中医到	药预防方案 本报讯 计	普啡 http://news	.qq.com/a/20130	405/000335.htm	3
	20 人	感染H7N9禽流展	感中医药预防;	方案发布—新闻—科学网	方案中发布了西	两个预防H7N9禽流	感的中药处方,但	专家建议市民使用到	顾时中 http://news	.sciencenet.cn/ht	minews/2013/4/	27642
	21 H	m9禽流感的日常	常预防措施_百	度经验	【甘肃卫生厅:	: 按摩迎香穴可预	防H7N9禽流感】甘	肃省《人感染H7N98	套流層 http://jingya	n.baidu.com/art	de/22a299b507	ceef9(
	22 刻	何预防H7N9离影	和感病毒 防治	禽流感的小常识【图】俱	影目前尚不明确.	,根据以往经验及	本次病例流行病学	周查,推测可能为护	带H7 http://health	1.pclady.com.cn/	96/960258.html	
	23 h	n9禽流感_h7n9;	禽流感症状上	7n9诊疗方案_携手健康网	列 h7n9禽流感专题	题介绍h7n9禽流感	预防、治疗知识,内	容包括h7n9禽流感的	定状, http://www.	xsjk.net/topic/h	/n9qlg/	
				中医药预防方案_新闻_應								1
	25 ft	京公布中药处力	5 预防H7N9禽	流感诊疗方案_天极网	记者昨天了解到	到,《北京市中医	管理局人感染H7N9	禽流感中医药预防力	5案(2 http://dh.ye	sky.com/345/34	551845.shtml	

Fig.4.3 the initial database

备注	🛛 十六进位 🔝 图像	分 升幂排序 ≫ 降幂排序 अ 移移	余排序	
	title		passage	url
	19 h7n9禽	流感_h7n9禽流感症状_h7n9诊疗方案_排	\$ h7n9禽流感专题介绍h7n9禽流感预防、治疗知识,内容包括h7n9禽流感症状,	h http://www.xsjk.net/topic/h7n9qlg/
			据了解,针对目前我国个别地区出现人感染H7N9禽流感病例,北京市中医常	
	21 北京公	布中药处方预防H7N9禽流感诊疗方案_	引记者昨天了解到,《北京市中医管理局人感染H7N9禽流感中医药预防方案(2 http://dh.yesky.com/345/34551845.shtml
	22 如何预	防H7N9禽流感?-新闻频道-和讯网	虽然目前国内外尚无针对H7N9禽流感病毒的疫苗,但是我们面对流感病毒t	thttp://news.hexun.com/2013-04-12/153063846.ht
			1.北京市中医管理局组织北京市突发公共卫生事件中医药应急专家委员会专家	
	24 人感染	H7N9禽流感疫情- 搜狐健康	[查看详细] 禽流感治疗方案 西药: 初步确定用达菲进行抗病毒治疗 中药:	http://health.sohu.com/s2013/qlg/
		"流感来袭_禽流感的预防方法	H7N9禽流感突然来袭,并且使人感染H7N9禽流感病毒,是全球首次发现的新	
			1 由于H7N9禽流感病毒是一种新的病原体,从病毒生物学上属于禽源流感病毒	
	27 千龙网]千龙健康频道H7N9禽流感治疗方案	预防并及时治疗各种并发症尤其是医院获得性感染。(据卫生与计划生育委)	员 http://health.qianlong.com/34876/2013/04/07/253
			f 3721健康知识www.jk3721.comH7N9禽流感的症状与治疗方案:H7N9禽流感	
	29 h7n9藩		表 关注禽流感最新资讯,关注最新疾患动态,关注h7n9禽流感,h7n9禽流感如	
	30 H7N9#	;流感的治疗方案_疾病预防_健康指南	用法:每日1剂,水煎服,分2次服用。不能口服者,鼻饲或保留灌肠。(二) http://www.tech-food.com/kndata/1051/0102406.l
			VH7N9禽流感疫情防控方案发布,接下来的问题是,如果有人不幸感染了这种	
			最近两天打开电脑都是H7N9禽流感的消息,而且也有了死亡的案例了,大家	
			》为做到早发现、早报告、早诊断、早隔离、早治疗人感染H7N9禽流感病例,	
		日719禽流感诊疗方案_百度经验	人感染H7N9禽流感是由H7N9亚型禽流感病毒引起的急性呼吸道传染病。自	
	35 能够油	疗H7N9禽流感的有效药物是什么?每经	F 能够治疗H7N9禽流感的有效药 是(2)?	http://www.fbd.com.in/article/2013-00-/7-934
			:我要组织国家人感染H7N9禽流感临床专家组,结合现有所的含疗经验,对	
		:流感来袭_网易网 - 网易新闻	(04-01) 江苏全力救治安徽1例感染H7N9禽流感病例 (04-01) 上海禽流感死亡;	
			虽然H7N9型禽流感暂无疫苗,让人有点担心,产生一点心理波动是正常的。	
			则人民网武汉4月11日电据湖北省卫生厅消息,今日,湖北省卫生厅调整下发	
		病治疗方案库 - 环球医网	中国疾病治疗方案库是中国规模最大的疾病治疗方案系统,针对各种疾病,如	
			台各种微生物感染目标治疗、儿童感染性疾病治疗,专家表示,该《指有》	
	42 西医族	清诊疗-=医康阿,医药健康网=-中医治疗	·三又神经痛之【三又神经兆的中川 2 治 死(王、神多 南、「三又中• 痛	n http://www.ikim.com/pages/Cikipshtm
	43 风湿病	的症状及治疗_治疗风湿病的医院_怎样:	剂 早期怎样治疗风湿病?治疗从湿病的医院那家外,虽然风湿病在我们的生活中	http://health.gansudary.com.cn/fsb/

Fig.4.4 The result of tracking by using word frequency statistics method

六进位 🔣 图像 🛛 🖓 升幕排序 🧏 降幕排序 🕺	● 移标排序	
title	passage	url
31 H7N9禽流感来袭_禽流感的预防方法	H7N9禽流感突然来袭,并且使人感染H7N9禽流感病毒,是全球首次	欠发现的新i http://news.szhk.com/special/H7N9
32 生活小贴士:7招预防感染H7N9禽流感病	毒_财经_腾讯网 由于H7N9禽流感病毒是一种新的病原体,从病毒生物学上属于禽	.狼流感病毒 http://finance.qq.com/a/20130409/002886.htm
33 千龙网千龙健康频道H7N9禽流感治疗;	方案 预防并及时治疗各种并发症尤其是医院获得性感染。(据卫生与计	·划生育委员 http://health.gianlong.com/34876/2013/04/07/25
34 H7N9禽流感的症状与治疗方案_健康新闻	_3721健康知识F 3721健康知识www.jk3721.comH7N9禽流感的症状与治疗方案:H7	N9禽流感在 http://www.jk3721.com/html/jiankangzixun/jianka
35 h7n9禽流感-h7n9禽流感症状-h7n9禽流感	特效药-禽流感洋 关注禽流感最新资讯,关注最新疾患动态,关注h7n9禽流感,h7n	n9禽流感症 http://www.h7n9qinlugan.com/
36 H7N9禽流感的治疗方案_疾病预防_健康排	會南 用法:每日1剂,水煎服,分2次服用。不能口服者,鼻饲或保留	羅肠。(二) http://www.tech-food.com/kndata/1051/0102406
37 H7N9禽流感诊疗方案发布:该不该免费液	治疗_财讯.COM H7N9禽流感疫搐防控方案发布,接下来的问题是,如果有人不幸	感染了这种 http://economy.caixun.com/wkp/20130404-CX03
38 H7N9禽流感现在有治疗方案了吗?快速问	7医生_搜索更多 最近两天打开电脑都是H7N9禽流感的消息,而且也有了死亡的案(例了,大家 http://www.120ask.com/question/41300655.htm
39 卫生部发布《人感染H7N9禽流感疫情防药	空方案》新闻_腾 为做到早发现、早报告、早诊断、早隔离、早治疗人感染H7N9禽%	流感病例,http://news.qq.com/a/20130403/001958.htm
40 人感染H7N9禽流感诊疗方案_百度经验	人感染H7N9禽流感是由H7N9亚型禽流感病毒引起的急性呼吸道传	梁病。自2(http://jingyan.baidu.com/article/fec7a1e532b876
41 康缘药业产品入选H7N9禽流感治疗方案-3	度狐财经 康缘药业(600557.SH)4月12日晚公告称,近日国家卫生和计划经	生育委员会: http://business.sohu.com/20130414/n372638427
42 能够治疗H7N9禽流感的有效药物是什么?	●每经网 能够治疗H7N9禽流感的有效药物是什么?…周三晚间,国家卫计	委发布《人 http://www.nbd.com.cn/articles/2013-04-07/7296
43 人感染H7N9禽流感诊疗方案(2013年第2席		级验,对 http://www.gov.co/gzdt/2013-04/11/coptert_23
44 H7N9禽流感来袭_网易网 - 网易新闻	(04-01) 江苏全力救治安徽1例5 QH 19篇 記录 例 0-01 , 演 會	\$液感列ニート p /new 165 com, peci /c lit anz 15
45 中国出现H7N9禽流感病例_腾讯网-腾讯	新闻 4月9日被诊断为重症肺炎,4月10日晚,经上海市疾病预防控制中	P心检测结身 http://news.qq.com/zt2013/H7N9/index.htm
46 湖北下发H7N9禽流感诊疗新方案 治疗前期	需采集呼吸道标: 人民网武汉4月11日电露湖北省卫生厅消息,今日,湖北省卫生厅	〒调整下发了http://news.10jqka.com.cn/20130411/c53397341
47 中国疾病治疗方案库 - 环球医网	中国疾病治疗方案库是中国规模最大的疾病治疗方案系统,针对各	种疾病,如: http://www.54md.com/treatment/
48 感染性疾病诊疗有了"口袋书"《国家抗微	生物治疗指南》 各种微生物感染目标治疗、儿童感染性疾病治疗,专家表示,(该《指南》 http://www.jkb.com.cn/htmlpage/35/356553.htm
	医治疗 神经内科 三又神经痛之【三又神经痛的中医内治研究】三又神经痛之【三	
50 银眉病治疗的好方法 诊疗银眉病疾病——	- 廊坊新闻网 摸索出一套治疗银屑病的特色病法,或对地解决了传统方法论定的	報題 和的不 http://www.ifeewergn/www.me-9999829/http/
51 支气管哮喘治疗-哮喘最新治疗方法-北京	·癫坊新闻网 摄索出一套治疗银屑病的特色疗法,或为增强了伤伤不去没有的 国济中医医院呼北京国济中医院呼吸疾病。这个时代,此人强烈的支持是少症法	新療医院 首 http://www.gixc.cov
	疾病查询_39健 它是消灭慢性传染病源,预防复发和治疗各种严重并发症的有效;	
53 咽喉部疾病治疗中心」广西总队医院耳鼻	咽喉诊疗中心 这三个面属外因,内因则包括自身健康状况、相关疾病的治疗情况	况等。广 http://www.wjrbh.com/yhzlzx/
54 风湿病的症状及治疗_治疗风湿病的医院_	怎样治疗风湿症早期怎样治疗风湿病?治疗风湿病的医院哪家好?虽说风湿病在我们	门的生活中 [;] http://health.gansudaily.com.cn/fsb/
	查诊断/治疗方药 各类疾病的相关数据,主要整合中医证候、病证关系、中医诊/	

Fig. 4.5 the result of tracking by using lexical chain's sememe method

Table 4.3 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 chains 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

From table 4.4, 4.5, 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.

Table 4.4 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 chains 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
(CDet)Norm	0.29317	0.26168	0.29487	0.28664	0.35475	0.28975

When the threshold is 0.6 the comparison and analysis of the result is in table 4.5:

Table 4.5 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
(CDet)Norm	0.34883	0.27275	0.39832	0.29502	0.40625	0.29900

[5] 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.

References

Journal Papers:

- YANG Yim-ing, CARBONELL J, BROWN R, et al. Learning Approaches for Detecting and Tracking News Event. IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval, 14(4), 1999, 32-43.
- [2] Hu Jiming, Hu Changping. The user modeling based on topic hierarchy tree and semantic vector space model. Journal of intelligence, 32 (8), 2013.8, 838-843.
- [3] Beydoun G, Lopez—Lorca A A, et al. How do we measure and improve the quality of a hierarchical ontology. Journal of Systems and Software, 84 (12),2011, 2363—2373.
- [4] Jin Zhu Lin Hongfei. *Topic tracking and tendency analysis based on HowNet. Journal of intelligence, 24 (5), 2005, 555-561.*
- [5] Gonenc E, Ilyas C. Using Lexical Chains for Keyword Extrac—tion . Information Processing and Management, 43(6), 2007, 1705—1714.
- [6] HowNet[R].*HowNetsHome.Page*.HTTP://WW W.keenage.com.
- [7] J Morris, G Hirst. Lexical Cohesion Computed by Thesauralrelations as all Indieator of the Structure of Text. Computational Linguistics, 17(1), 1991, 21-48.

Books:

- [6] James Allan. Introduction to topic detection and tracking// James Allan,ed. Topic Detection and Tracking Event—based Information Organization. (USA: Kluwer Academic Publishers, 2002).
- [7] Halliday M A K, Hasan R. Cohesion in English. London, (UK: Longman, 1976).