Construction of thesaurus in the field of car patent

Tingting Mao[1], Xueqiang Lv[1], Kehui Liu[2]

[1](Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, 100101, China)

[2](Beijing Research Center of Urban System Engineering, Beijing, 100089, China)

Keywords: car patent, thesaurus, equivalent relationship, correlative relationship, hierarchical relationship.

Abstract. Automatic construction of the thesaurus in the car patent field is an important semantic tool of car patent organization and retrieval. Not only the traditional method of constructing thesaurus is time-consuming and laborious, but also the speed of updating the thesaurus is slow. This paper puts forward equivalent relationship features weighting, correlative relationship features weighting and rank coefficient calculation three algorithms to respectively identify equivalent relationship, correlative relationship and hierarchical relationship, so as to achieve automatic construction of the thesaurus in the car patent field. This method brings new ideas to construct the thesaurus.

Introduction

Now, the car is people's important traffic tool and researches for the car grow with each passing day. With the people's attention to the car, more and more car patents appear. How to effectively organize and retrieve these patents is a more urgent problem to resolve.The thesaurus of car patent is a collection of car patent knowledge and it is an important tool for the car patent database construction and retrieve.It can effectively identify whether patent belongs to the car related field and improve the users' retrieval efficiency. At present, not only the speed of hand-built thesaurus is slow, but also the disunity of different staff's experience and knowledge causes different construction standards. The most important is that the speed of updating the thesaurus is slow and it is not conducive to the application. Therefore, the research for automatically building the thesaurus in the car patent field has great practical significance.

At present, the research for the thesaurus is mainly focused on thesaurus conversion[1] and comprehensive upgrade and maintenance of existing thesaurus[2]. However, in terms of automatically building the thesaurus, the main methods include transformation based on WordNet[3], concept space[4], integration of existing vocabulary.These automatic construction methods are either built out the specific use environment, or only identify the correlative relationship between words. This paper applies the related knowledge of statistics and equivalent relationship features weighting, correlative relationship features weighting and rank coefficient calculation on the basis of computer-related technology to automatically identify equivalent, correlative and hierarchical relationship between words in the car patent field, according to the three relationships to automatically build the thesaurus in the related field of car patent.

Construction technology of thesaurus in the car patent field

The thesaurus in the car patent field is composed by some relationships between vocabulary in the related field of car patent. Its specific performance is equivalent, correlative and hierarchical relationship between vocabulary.

Equivalent relationship identification. Equivalent relationship between words in this paper refers to words that have the same meaning such as bicycle and bike, words that have the opposite meaning such as input shaft and output shaft, words that have the similar meaning such as protective layer and insulating layer.

Because synonyms in this paper are terms in the patent, most of these terms are combination

words that patent authors use. However, the vocabulary of existing knowledge table is limited and calculation method of using existing knowledge table is obviously not feasible. In addition, every patent author has differences in the style of using words, so taking advantage of pattern matching to calculate will be limited. Therefore, this paper carried out co-occurrence analysis and literal similarity to identify synonyms.

1)Co-occurrence analysis

If two words are synonymous, the words that appear around them are often same. The words that are far away from the current word have low association degree with the current word. According to the above theory, use the words around the current word to represent the current word and use them to conduct a gradient co-occurrence window weighting. Specific weighting method is formula (1).

$$\mathrm{wei}(t_i, t_j) = \sum_{n=1}^{k} \begin{cases} \max\left\{10 - \mathrm{dis}(t_i, t_j)_n, 1\right\}, & \mathrm{dis}(t_i, t_j)_n < 16 \\ 0, & \mathrm{dis}(t_i, t_j)_n \geq 16 \end{cases} \quad (1)$$

In the formula (1), t_i is the current word; t_j is the around words of t_i; k represents the total number of current word; $\mathrm{wei}(t_i, t_j)$ represents weight in the vector constructed by word t_j and target t_i; $\mathrm{dis}(t_i, t_j)_n$ represents the number of non-stop word that appears between word t_j and target t_i in the same section.

Target words vector is constructed by formula (1) and this paper takes advantage of cosine distance to measure the semantic similarity distance between two words. The calculation method of cosine distance is formula (2).

$$W(T_1, T_2) = \frac{\sum_{i=1}^{k}(W_{1i} \times W_{2i})}{\sqrt{(\sum_{i=1}^{k} {W_{1i}}^2) \times (\sum_{i=1}^{k} {W_{2i}}^2)}} \quad (2)$$

In the formula (2), T_1 and T_2 are two different words; K represents the dimension of feature vector; W_{1i} represents the value of the i dimension in the feature vector of word T_1; W_{2i} represents the value of the i dimension in the feature vector of word T_2; $W(T_1, T_2)$ is the distance between T_1 and T_2.

2) Literal similarity

For different words, if they are more similar in literal, the association degree between words is greater. Therefore, this paper uses the literal similarity algorithm that is mentioned in the literature [5].Specific calculation method is formula (3).

$$S(T_i, T_j) = \frac{2 \times L_{ij}}{L_i + L_j} \quad (3)$$

In the formula (3), T_i and T_j are differents words; $S(T_i, T_j)$ represents the literal similarity between words; L_{ij} represents the number of same word in T_i and T_j; L_i and T_j are respectively the length of T_i and T_j.

3) Equivalent relationship features weighting

In order to improve the effect of identifying equivalent relationship between words, this paper puts forward co-occurrence analysis and literal similarity to identify equivalent relationship between words. Calculation method is formula (4).

$$\mathrm{Sim}(T_i, T_j) = \alpha \cdot W(T_i, T_j) + \beta \cdot S(T_i, T_j) \quad (4)$$

In the formula (4), T_i and T_j are two different words; $\mathrm{Sim}(T_i, T_j)$ that is calculated by formula (2) represents similarity between two words; $S(T_i, T_j)$ that is calculated by formula (3) represents literal similarity between two words; α and β are weight factors and $\alpha+\beta=1$.

Correlative relationship identification. Correlative relationship mainly calculate the correlation degree between words. The algorithms of calculating the correlation degree have mutual information, Dice measure etc[6]. This paper uses Dice measure and lexical semantic distance to calculate association between words. Dice measure method is formula (5). Although it can efficiently overcome zero-probability event and low-frequency phenomenon, but it cannot correctly

reflect correlative relationship of two words that donot appear in the same document. The semantic distance between words just makes up for the shortage of Dice measure in this aspect.

$$\mathrm{Dice}(T_i, T_j) = 2F(T_i, T_j)/(F(T_i) + F(T_j)) \quad (5)$$

In the formula (5), T_i and T_j are two different words; $\mathrm{Dice}(T_i, T_j)$ represents Dice measurement value of T_i and T_j; $F(T_i, T_j)$ represents the number of articles in which T_i and T_j appear together; $F(T_i)$ represents the number of articles in which T_i appears; $F(T_j)$ represents the number of articles in which T_j appears.

1) Lexical semantic vector construction

For the term T that appears in the patent, it can use others terms that appear with T at the same time to replace. So, this paper puts forward formula (6) to construct corresponding semantic vector.

$$\mathrm{wei}(T, T_i) = \sum_{n=1}^{n=k} F(T_i)_n \quad (6)$$

In the formula (6), T_i is the term that is extracted from patent literatures; $\mathrm{wei}(T, T_i)$ represents the weight value in the vector that is constructed by T and T_i; k represents the total number of articles; $F(T_i)_n$ represents the number of T_i that appears in the n article.

2) Correlative relationship features weighting

This paper takes advantage of complementary characteristic of cosine distance and Dice measure and puts forward formula (7) to identify association degree between terms and chooses top 20 words.

$$\mathrm{Rel}(T_i, T_j) = \gamma \cdot W(T_i, T_j) + \delta \cdot \mathrm{Dice}(T_i, T_j) \quad (7)$$

In the formula (7), T_i and T_j are different words; $\mathrm{Rel}(T_i, T_j)$ represents correlation degree between words; $W(T_i, T_j)$ calculated by formula (2) represents semantic weight between words; $\mathrm{Dice}(T_i, T_j)$ calculated by formula (5) represents Dice measurement weight; γ and δ are weight factors and $\gamma+\delta=1$.

Hierarchical relationship identification. At present, hierarchical relationship identification is mainly based on literal similarity [13] and cluster [7]. The former cannot reflect actual application words, and cannot identify rank relationship words that donot have literal similarity. Therefore, this paper chooses rank relationship based on cluster to identify.

1) Hierarchical cluster algorithm

According to the different methods of calculating distance between cluster, hierarchical cluster algorithm can be divided into the following three methods:

Single connectivity algorithm takes the distance of the most similar two samples between cluster as two clusters' distance.Obviously, it might take dissimilar clusters into a cluster.

Full connectivity algorithm is in contrast with single connectivity algorithm. It takes the distance of the most dissimilar two samples as two clusters' distance. It overcomes the shortcomings of single connectivity algorithm.

Average connectivity algorithm takes the average distance of two samples between cluster as two clusters' distance.

2) Rank coefficient calculation

At present, Du Huiping[14] etc ever have mentioned that it can take word frequency and word length as factors of considering hyponymy. The possibility of the higher frequency as hypernums is greater and the possibility of the longer length as hyponyms is greater. This paper also thinks that the possibility of the greater inverse document frequency as hypernums is greater. According to the above factors, this paper quantifies the level of vocabulary such as formula (8).

$$\mathrm{H(Ti)} = \frac{\mathrm{Freq(Ti)} * \mathrm{idf(Ti)}}{\mathrm{len(Ti)}} \quad (8)$$

$$\mathrm{idf(Ti)} = \log(N/n_i) \quad (9)$$

In the formula (8) and (9), $\mathrm{H(Ti)}$ represents rank coefficient of vocabulary; $\mathrm{Freq}(T_i)$ represents total frequency of T_i; $\mathrm{len(Ti)}$ represents the length of T_i; $\mathrm{idf(Ti)}$ represents inverse document frequency of vocabulary; N represents total number of document; n_i represents the number of document in which appears Ti.

Data and experimental results

The terms in construction thesaurus are extracted from 1246 patent literatures. And it uses 3140 terms that appear more than 40 times to conduct synonyms, correlative relationship and hierarchical relationship identification. And through the identification results of three relationships construct thesaurus.

Equivalent relationship identification experiment. This paper uses formula (4) to calculate the equivalent relationship between two words. The experimental results are shown in Table 1.

Table 1 Synonym identification results

Patent number	Extracted synonyms pair	Correct synonyms pair	Correct rate (%)
1246	467	251	53.75%

Through the results, we can see that the method that is used by this paper has a certain effect, but since the case of the words that are not synonyms but accompany with each other is more, leading that the experimantal result is not enough high.

Correlative relationship identification experiment. This paper uses formula (7) and formula (5) to calculate the correlative relationship between words. As follows, this paper uses PWM to compare the method that is put forward by this paper with only Dice measure to calculate correlation degree. For example Table 2.

Table 2 Correlation degree generation results

Serial number	Dice and consine distance		Dice measure	
	word	Correlation degree	word	Correlation degree
1	Voltage waveform	0.66257	Voltage waveform	0.5
2	Alternating current power	0.63059	Power supply	0.44444
3	Signal wave	0.63059	Air temperature sensor	0.4
4	Charge and discharge efficiency	0.63059	Powered device	0.4
5	Peak voltage	0.63059	Current detection value	0.4
6	Induction motor	0.63059	Solid electrolyte plate	0.4
7	Command generating unit	0.63059	Power control device	0.4
8	Sine wave function	0.63059	Stator coil	0.4
9	Power control device	0.63059	Alternating current power	0.4
10	Motor drive device	0.63059	Current control system	0.4

From the Table 2, we can see that the method that merges Dice measure and cosine distance is better than the methods that only uses Dice measure and generated correlation results are basically

reasonable.

Hierarchical relationship identification experiment. This paper uses single connectivity cluster, average connectivity cluster and full connectivity cluster three algorithms to conduct rank cluster experiments and each algorithm chooses a plurality of thresholds. It selects the best set as the experimental results through artificial scoring .The experimental results are shown in Table 3:

Table 3 Rank identification results comparision

Cluster algorithm	Threshold	Word cluster number	Single word cluster number	The largest cluster word number
Single connectivity	0.55	527	321	1774
	0.6	743	480	454
	0.65	955	639	111
	0.7	1172	838	75
Average connectivity	0.55	861	456	65
	0.6	1015	619	64
	0.65	1183	782	64
	0.7	1342	945	64
Full connectivity	0.55	1051	576	64
	0.6	1196	734	64
	0.65	1344	889	64
	0.7	1484	1043	63

After artificial comparative analysis, this paper selects the corresponding experimental results of average connectivity algorithm and threshold as 0.65.

Thesaurus evaluation

This paper counts number of entrance words, formal theme words, genus item, subitem, parameter, unrelated number. The results are shown in Table 4:

Table 4 the thesaurus of car patent field vocabulary performance parameters

Total number	Informal theme number	Formal theme number	Genus item	subitem	Parameter item	Unrelated number
3140	251	2889	1957	1957	13353	841

In the study of library and information science, some scholars[8] use entrance rate,correlation ratio,accessibility and word family scale four index to evaluate concept relationship control performance of thesaurus.

$$\text{Entrance rate} = \frac{\text{number of entrance word}}{\text{total number of formal theme word}} = \frac{251}{2889} = 0.086$$

$$\text{Correlation ratio} = \frac{\text{number of theme word having parameter item}}{\text{total number of formal theme word}} = \frac{2889-841}{2889} = 0.709$$

$$\text{Reference degree} = \frac{\text{number of genus item plus sub item plus parameter item}}{\text{total number of formal theme word}} = \frac{1957+1957+13353}{2889} = 5.977$$

$$\text{Word family scale} = \frac{\text{number of theme word}}{\text{number of the first word in the family}} = \frac{2889}{1183} = 2.442$$

Through above parameters of car patent field thesaurus, we can see that the entrance rate of thesaurus constructed by this paper is lower.The main reason is that a part of synonym cannot be correctly identified. However, for association ratio and reference degree, the main reason is that correlative relationship redundancy is more and it has error correlative relationship.

Conclusion

This paper puts forward the equivalent relationship weight method of co-occurrence analysis and literal similarity features , correlative relationship weight method of imerging Dice measure and semantic distance between words and hierarchical relationship method of imerging word length,

word frequency and word inverse document frequency. It achieves automatic construction of thesaurus in the car patent related field. For the constructed thesaurus, though the constructed thesaurus has low entrance and association ratio,reference degree are higher, but the speed of the thesaurus constructed by hand is fast and it is easy to update and application is convenient and construction standard is unified and it provides new ideas to automatically construct thesaurus.

Acknowledgements

The research work was supported by National Natural Science Foundation of China under Grants No. 61271304 and Beijing Natural Science Foundation of Class B Key Project under Grants No. KZ201311232037.

References

[1] Chang Chun. Ontology construction and transformation in agricultural information management [D]. PhD thesis of Chinese agricultural science academy, 2004.

[2] Ceng Jianxun, Chang Chun. Thesaurus compilation and application in the network era [J]. Library information work, 2009.

[3] Zhang Li, Li Jingjiao, Hu Minghan etc. Chinese WordNet research and implementation [J]. Northeastern University Journal, 2003.

[4] Chang Chun. Ontology construction and transformation in agricultural information management [D]. PhD thesis of Chinese agricultural science academy, 2004.

[5] Lu Yong, Zhang Chengzhi, Hou Hanqing. Multi strategy Chinese synonym automatic extraction research based on encyclopedia resources [J]. Chinese library journal, 2010, 36(185):056-062.

[6] Zhong Yunyun, Hou Hanqing, Du Huiping. Construction of e-government thesaurus and its applied research [J]. China index, 2008, 6(2).

[7] Du Huiping, Zhong Yunyun. Nature language thesaurus’s automatic construction research [M]. Nanjing: southeast university press, 2009:72-72.113-114.

[8] Juan, Sun Aili, Wang Haixiong etc. International LIS Studies Thesaurus performance evaluation [J]. Modern intelligence 2011.5.5.