A Semantic Matching of Information Segments for Tolerating Error Chinese Words*

Maoyuan Zhang[1, 3], Chunyan Zou[2], Zhengding Lu[1], and Zhigang Wang[1]

[1] Department of Computer Science and Technology,
HuaZhong University of Science and Technology,
430074 Wuhan, P.R. China
zmydragon@163.com
[2] School of Foreign Languages, HuaZhong Normal University,
430079 Wuhan, P.R. China
[3] Schoole of Management, HuaZhong University of Science and Technology,
430074 Wuhan, P.R. China

Abstract. There exist new words and error words in Chinese information of web pages. In this paper, we introduce our definition of semantic similarity between sememes and their theorems. On the base of proving the theorems, the influence of the parameter is analyzed. Moreover, this paper presents a novel definition of the word similarity based on the sememe similarity, which can be used to match the new Chinese words with the existing Chinese words and match the error Chinese words with correct Chinese words. And also, based on the novel word similarity, a matching method of information segments is presented to recognize the category of Chinese web information segments, in which new words and error words occur. In addition, the experiment of the matching methods is presented. Therefore, the novel matching method is an efficient method both in theory and from experimental results.

1 Introduction

A huge amount of web information exists on the Internet, and the information discovery methods of web pages have been researched to gain the valuable information exactly. Semantic matching is a puzzling problem, not only for information discovery but also for natural language processing. For example, an ontology search engine [1] used vector-matching algorithm, a XML document clustering [2] used the matching method to gain the similarity between documents, and a visual speech recognition system [3] used the matching method to distinguish a word from other words.

Since information segments are composed of words, semantic matching of information segments is based on the methods of word semantic matching. Methods of word semantic matching can be categorized into two groups: the word matching methods based on occurrence frequency and those based on conceptual relationship.

* This work was partially supported by National Natural Science Foundation of China under Grant 60403027.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, pp. 48 – 59, 2006.

The former matching methods use the word occurrence frequency in documents or the common characters between words to calculate the similarity between words. Some of the former semantic matching methods [4][5][6] did not take the structural relationship of the lexical taxonomy into consideration.

The latter matching methods use the word conceptual relationship in the semantic network to calculate the similarity between words. As for the semantic network, assume that a lexical taxonomy is structured in a tree like hierarchy with a node for a concept, and Rada et al. [7] has proven that the minimum number of edges separating concepts c1 and c2 is a metric for measuring the conceptual distance between c1 and c2. Their work forms the basis of the relationship-based matching methods. Some semantic matching methods [8][9][10][11], considered the conceptual distance in the hierarchy of the lexical taxonomy, but did not take into account the conceptual depth in the hierarchy of the lexical taxonomy. Furthermore, a word similarity method based on different ontologies [12] considered both the conceptual distance and the conceptual depth in the taxonomy hierarchy, but had to modify and expand its semantic networks when a new word appeared.

There exist some problems for Chinese information in web pages. On the one hand, the information of web pages is likely to change, and new Chinese words may appear with the development of the human knowledge. On the other hand, the information of web pages may include error Chinese words led by contrived factors and natural factors. For instance, someone propagandizing illegal information inserted irrelative Chinese characters into the words to prevent information extraction tools from gaining the meanings of the Chinese words. Hence, facing those problems, the semantic matching methods should be of the adaptive ability to match the new Chinese words with the existing Chinese words, and be of the semantic tolerance ability to gain the correct semantic meanings from the error Chinese words.

The rest of this paper is organized as follows. In Section 2, we introduce our definition of semantic similarity between sememes. In Section 3, we introduce and prove some theorems on the semantic similarity between sememes, and then analyze the influence of the parameter on the base of the theorems. In Section 4, we present a novel definition of the word similarity based on the sememe similarity, which can be used to match the new Chinese words with the existing Chinese words and match the error Chinese words with correct Chinese words. And then, a matching method of information segments is presented to recognize the category of web information segments. In Section 5, the experiment of the matching methods is presented.

2 Semantic Similarity Between Sememes

2.1 Sememe Network

HowNet is an online bilingual common sense hierarchical ontology describing semantic relations between concepts (represented by Chinese and English words) and also describing semantic relations between the attributes of concepts [13]. As the most basic language unit, one or more Chinese sememes form a Chinese word with some certain meanings, one or more Chinese words form a phrase, and one or more words

and phrases form a sentence. Each Chinese sememe or word has its own meaning, from which the meaning of a phrase or a sentence that contains them originates. So there is a tight semantic relationship between Chinese sememes and words.

From medicinal information in web pages, we extracted the Chinese words about medicine, such as a Chinese word "药名", which means "name of medicine" in English. Furthermore, we extracted the sememes from the Chinese words about medicine, such as sememe "药" and "名", which mean "medicine" and "name" respectively in English. Based on the sememes and the Chinese words about medicine, we constructed a HowNet—medicine according to the constructional principle of the HowNet. HowNet—medicine is a special kind of HowNet for the medicinal information, and describes semantic relations between concepts (*represented by Chinese sememes and words*) and also semantic relations between the attributes of concepts. The difference between HowNet—medicine and HowNet lies in that the previous one is based on Chinese sememes and words.

2.2 Similarity Function of Sememes

If we assign "exactly the same" with a value of 1 and "no similarity" as 0, then the interval of similarity is [0,1]. The values of argument of similarity function may cover a large range up to infinity, so it is intuitive that the similarity function is a nonlinear function. Hence, The nonlinearity of the similarity function is taken into account in the derivation of the formula for semantic similarity between sememes.

Sememe Similarity Based on Path Length. Given two sememes $seme_1$ and $seme_2$, we need to find the semantic similarity of them. We can do this by analysis of the knowledge base, as follows: Sememes are associated with concepts in the hierarchical HowNet—medicine. Hence, we can find the first concept in the hierarchical semantic network that subsumes the concepts representing the compared sememes. One direct method for similarity calculation is to find the minimum length of path connecting the two concepts representing the compared sememes.

Definition 1: Suppose $Seme_1$ and $Seme_2$ are two sememes, and the minimum length of path connecting the two concepts representing them is L. Then the sememe similarity based on path length is defined as

$$Sim1(Seme1, Seme2) = f_1(L) = e^{-\alpha L} \ . \tag{1}$$

where constant $\alpha>0$ and $L \in [0,+\infty)$.

Expanded Sememe Similarity. The depth of the sememe is derived by counting the levels from the concept representing the sememe to the top of the lexical hierarchy. Sememes at upper layers of hierarchical semantic nets have more general concepts and less semantic similarity between sememes than sememes at lower layers. This must be taken into account in calculating the sememe similarity. We therefore need to scale down $Sim1(seme_1, seme_2)$ for sememes at upper layers(lower depth) and to scale up $Sim1(seme_1, seme_2)$ for sememes at lower layers(higher depth). Moreover, the similarity is constrained to [0,1].

Definition 2: Function $f_2(h_1, h_2)$ is defined as

$$f_2(h1,h2)=\frac{e^{\beta(h1+h2)/2}-e^{-\beta(h1+h2)/2}}{e^{\beta(h1+h2)/2}+e^{-\beta(h1+h2)/2}}$$, where constant $\beta>0$ and $h1,h2\in[0,+\infty)$.

Definition 3: Suppose $Seme_1$ and $Seme_2$ are two sememes, their depths are h_1 and h_2 respectively, and the minimum length of path connecting the two concepts representing them is L. Then the expanded sememe similarity is defined as

$$\begin{aligned}\mathrm{Sim2}(seme1,seme2)&=f(f_1(L),f_2(h1,h2))\\&=f_1(L)\times f_2(h1,h2).\end{aligned}\tag{2}$$

3 Theorems on the Sememe Similarity

In order to analyze the parameterβ 's influence on the expanded sememe similarity, we will introduce some theorems. Some functions are defined as follows before introducing some theorems.

Definition 4: Function u(h) is defined as $u(h)=\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}$, where the constant $\beta>0$ and the argument $h\in[0,+\infty)$.

Definition 5: Function w(h) is defined as $w(h)=4h/(e^{\beta h}+e^{-\beta h})^2$, where the constant $\beta>0$ and the argument $h\in[0,+\infty)$.

Definition 6: Function v(h) is defined as

$$v(\beta,h)=u(h+d)-u(h)=\frac{e^{\beta(h+d)}-e^{-\beta(h+d)}}{e^{\beta(h+d)}+e^{-\beta(h+d)}}-\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}},$$

where argument $\beta,h\in[0,+\infty)$, and constant $d\geq 0$.

3.1 Theorems and Their Proofs

Based on those function definitions above, we introduce and prove some theorems as follows.

Theorem 1: The range of the function u(h), defined by the Definition 4, is [0,1]. Moreover, if $u(h_o)\geq a$, where a is a constant and h_o is a value of the argument h, then $\beta\geq\frac{1}{2h_o}\ln\frac{1+a}{1-a}$.

Proof: Differentiating u(h) with respect to variable h as follows,

$$\begin{aligned}du/dh&=(1/(e^{\beta h}+e^{-\beta h})^2)[\beta(e^{\beta h}+e^{-\beta h})^2-\beta(e^{\beta h}-e^{-\beta h})^2]\\&=4\beta/(e^{\beta h}+e^{-\beta h})^2\ .\end{aligned}\tag{3}$$

Since β>0, it can be obtained that $du/dh > 0$. That is, the function u(h) is a monotonically increasing function, which is shown in the figure 1.

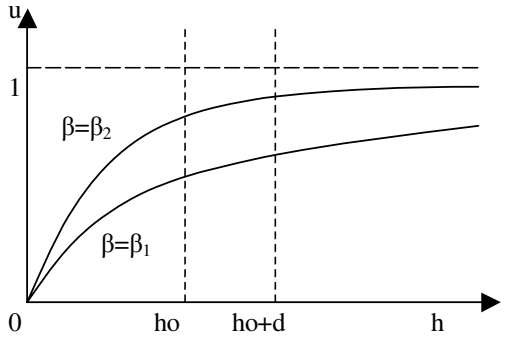

Fig. 1. The curve of function u(h)

According to the above conclusion, the function u(h) has the minimum value 0 when h=0, and has the maximum value 1 when $h \to +\infty$. Hence, the function's range is [0,1].

Differentiating the derivative of function u(h) with respect to variable h as follows,

$$\begin{aligned} d^2u/dh^2 &= (-8\beta/(e^{\beta h}+e^{-\beta h})^3)(\beta e^{\beta h}-\beta e^{-\beta h}) \\ &= -8\beta^2(e^{\beta h}-e^{-\beta h})/(e^{\beta h}+e^{-\beta h})^3 . \end{aligned} \tag{4}$$

Since βh>0, it can be obtained that $e^{\beta h}-e^{-\beta h} \le 0$.

So $d^2u/dh^2 \le 0$, that is to say, the derivative function of u(h) is a monotonically decreasing function.

If $u(h_o) \ge a$, then $(e^{\beta h_o}-e^{-\beta h_o})/(e^{\beta h_o}+e^{-\beta h_o}) \ge a$, that is, $(e^{2\beta h_o}-1)/(e^{2\beta h_o}+1) \ge a$.

Hence, it can be easily proved that $e^{2\beta h_o} \ge (1+a)/(1-a)$, that is to say $\beta \ge \frac{1}{2h_o}\ln\frac{1+a}{1-a}$.

This completes the proof.

Theorem 2: There exist natural number b and c, where b=c/(2β) and 1.54<c<1.55, such that the function w(h) defined by the Definition 5 is monotonically increasing when $h \in [0,b]$ and is monotonically decreasing when $h \in (b,+\infty)$.

Proof: Differentiating function w(h) with respect to variable h as follows,

$$\begin{aligned} dw/dh &= (4/(e^{\beta h}+e^{-\beta h})^4)[(e^{\beta h}+e^{-\beta h})^2 - \\ &\quad 2\beta h(e^{\beta h}+e^{-\beta h})(e^{\beta h}-e^{-\beta h})] \\ &= (4/(e^{\beta h}+e^{-\beta h})^3)[(e^{\beta h}+e^{-\beta h}) - 2\beta h(e^{\beta h}-e^{-\beta h})] \\ &= (4e^{\beta h}/(e^{\beta h}+e^{-\beta h})^3)[(1+e^{-2\beta h}) - 2\beta h(1-e^{-2\beta h})] \\ &= (4e^{\beta h}/(e^{\beta h}+e^{-\beta h})^3)[(1+2\beta h)e^{-2\beta h}+1-2\beta h]. \end{aligned} \tag{5}$$

Let us introduce a new function

$$g(x) = (1+x)e^{-x}+1-x, \text{ where } x \geq 0. \tag{6}$$

Differentiating g(x) with respect to variable x as follows,

$$dg/dx = e^{-x} - (1+x)e^{-x} - 1 = -1 - xe^{-x} < 0$$

So the function g(x) is a monotonically decreasing function.

According to the above conclusion, the function u(h) has the maximum value 2 when x=0, and has the minimum value, a negative infinity, when $x \to +\infty$. Moreover, the equation g(x)=0 has an unique solution.

Suppose the unique solution is c, then g(x)>0 when 0≤x≤c, and g(x)<0 when x>c. When x=1.54, g(x)= 0.0045>0. When x=1.55, g(x)= -0.0088<0. So it can be obtained that 1.54<c<1.55.

From equation (5) and (6), it is easily given that

$$dw/dh = (4e^{\beta h}/(e^{\beta h}+e^{-\beta h})^3) \times g(2\beta h).$$

Hence, when 0≤2βh<c, it is obtained that g(2βh)>0 and then $dw/dh > 0$. Moreover, when 2βh>c, it is obtained that g(2βh)<0 and then $dw/dh < 0$.

Suppose b=c/(2β). Therefore, we can obtain that the function w(h) is monotonically increasing when $h \in [0, b]$ and is monotonically decreasing when $h \in (b, +\infty)$.

This completes the proof.

Theorem 3: As to the function v(β, h), defined by the Definition 6, suppose h_o is a value of the variable h. Then, the function v(β, h) is monotonically decreasing with respect to variable β when $h=h_o$ and $2\beta h_o > c$, where the constant c>1.55.

Proof: Let us introduce the function w(h) from the theorem 2 as follows

$$w(h) = 4h/(e^{\beta h}+e^{-\beta h})^2.$$

Differentiating v(β, h) with respect to variable β as follows

$$\begin{aligned} \partial v/\partial \beta &= [4(h+d)/(e^{\beta(h+d)}+e^{-\beta(h+d)})^2] - [4h/(e^{\beta h}+e^{-\beta h})^2] \\ &= w(h+d) - w(h) \end{aligned}$$

According to the theorem 2, w(h+d)≤w(h), when h≥c/(2β).

So that

$$\partial v / \partial \beta = w(h+d) - w(h) \le 0 \text{, when } h \ge c/(2\beta).$$

Since $2\beta h_o > c$, it is easy to see that $h_o > c/(2\beta)$.

Hence, it is obtained that the function v(β, h) is monotonically decreasing with respect to variable β when $h=h_o$ and $2\beta h_o > c$

This completes the proof.

Corollary 1: Suppose $Seme_1$ and $Seme_2$ are two sememes. Then the range of the similarity function $Sim2(Seme_1, Seme_2) = f(f_1(L), f_2(h_1, h_2))$, which is defined by Definition 3, is [0,1], and the similarity function Sim2 is nonlinear.

Proof: Let $h=(h_1+h_2)/2$, then

$$f_2(h1, h2) = \frac{e^{\beta(h1+h2)/2} - e^{-\beta(h1+h2)/2}}{e^{\beta(h1+h2)/2} + e^{-\beta(h1+h2)/2}} = u(h) \cdot$$

According to the theorem 1, $0 \le u(h) \le 1$, and thus $0 \le f_2(h1, h2) \le 1$.

In addition, $0 \le f_1(l) = e^{-\alpha l} \le 1$.

So that

$$0 \le f(f_1(L), f_2(h1, h2)) = f_1(L) * f_2(h1, h2) \le 1.$$

That is to say, the range of the similarity function $Sim2(seme_1, seme_2)$ is [0,1].

As to the function $Sim2(seme_1, seme_2) = f(f_1(L), f_2(h_1, h_2))$, the domains of argument L, h_1 and h2 are $[0,+\infty)$. If the function is linear, then $Sim2(seme_1, seme_2) \to +\infty$ when $L \to +\infty$. This leads to a contradiction, so the similarity function Sim2 is nonlinear.

This completes the proof.

3.2 Analyses of Parameter β 's Influence

Suppose the maximal depth of the semantic network is h_{max}, then the value of the function u(h) for $h > h_{max}$, is not significant to the similarity calculating, but the value of the function u(h) for $h \in [0, h_{max}]$ has influence to the similarity calculating. Hence, the range of the function u(h) for $h \in [0, h_{max}]$ should be large enough to be close to the interval [0,1], in order to gain better effect of influence. That is to say, the value of $u(h_{max})$ should be large enough to be close to 1.

On one hand, the value of $u(h_{max})$ should be large enough. In order to make $u(h_{max}) > a$, according to the theorem 1, the parameter β should satisfy that $\beta \ge \frac{1}{2h_{max}} \ln \frac{1+a}{1-a}$. For instance, if a =0.95 and h_{max} =10, then β≥0.183.

On the other hand, it is not promised that the larger the value of $u(h_{max})$ is, the better the parameter β 's influence effect is. As shown in the figure 1, suppose $\beta_2 \ge \beta_1$, $2h_o\beta_1 > c$, c>1.55, and d=1. Then the increment of the function u(h) is equals to $v(\beta_1, h_o)$ (shown in the theorem 3) when $\beta=\beta_1$ and h increases from h_o to h_o+d,. And also, the increment of the function u(h) is equals to $v(\beta_2, h_o)$ (shown in the theorem 3) when $\beta=\beta_2$ and h increases from h_o to h_o+d. According to the theorem 3 and the assume $\beta_2 \ge \beta_1$, it can be obtained that $v(\beta_2, h_o) \le v(\beta_1, h_o)$. Therefore, the larger is the value of

β, the smaller is the difference ($u(h_o+d)-u(h_o)$). Hence, an excessively large value of β can decreases the depth's influence on the similarity to some extent.

According to the above conclusion, the parameter β should satisfy that $\beta \geq \frac{1}{2h_{\max}} \ln \frac{1+a}{1-a}$ and should not be set as an excessively large value, to gain better effect of the depth's influence on the similarity.

4 Semantic Matching Method Based on Sememes

4.1 Word Similarity Based on Sememes

Definition 7: Suppose w is a word, and assume $Seme_1$, $Seme_2$, …, $Seme_n$ are the sememes forming the word w. Then the word w can be denoted by the sememe vector, which is defined as

$$SemeV=(Seme_1, Seme_2, \ldots, Seme_n) . \tag{7}$$

Definition 8: Suppose Seme is a sememe, and SemeV is a sememe vector ($Seme_1$, $Seme_2$, …, $Seme_n$). Then the similarity function between a sememe and a sememe vector, is defined as

$$Sim3(Seme, SemeV) = \max_{j=1}^{n} Sim2(Seme, Seme_j) \cdot \tag{8}$$

where the function Sim2 is defined by Definition 3.

Definition 9: Suppose the sememe vector of word w_1 is $SemeV_1=(Seme_{11}, Seme_{12}, \ldots, Seme_{1m})$, and the sememe vector of word w_2 is $SemeV_2=(Seme_{21}, Seme_{22}, \ldots, Seme_{2n})$. Then the word similarity based on sememes is defined as

$$Sim4(w1, w2) = \frac{1}{|SemeV_1|} \sum_{i=1}^{m} Sim3(Seme_{1i}, SemeV_2) \cdot \tag{9}$$

where $|SemeV_1|$ denotes the dimension of the vector $SemeV_1$.

4.2 The Solution of Matching for New Words and Error Words

With the development the human knowledge, the new Chinese words are formed by the sememes just as the existing Chinese words are, so the new Chinese words can also be denoted by the sememe vectors as well as the existing Chinese words. On the other hand, the error Chinese words are formed by the sememes just as the correct Chinese words are, so the error Chinese words can also be denoted by the sememe vectors as well as the correct Chinese words.

HowNet describes semantic relations between concepts (represented by Chinese and English words). Facing the new Chinese words, HowNet cannot give the semantic relation between the new words and the existing words, and the semantic similarity based on words cannot be applied. Hence, in order to calculate the similarity between words, we have to modify the HowNet by adding the new words. Moreover, HowNet cannot give the semantic relations between the error Chinese words and the correct Chinese words.

HowNet—medicine, introduced in Section 2, describes semantic relations between concepts (represented by Chinese sememes and words). Facing new Chinese words, we can extract the sememes forming the words, gain the semantic relations between sememes from HowNet—medicine, and give the semantic similarity between the new Chinese words and the existing Chinese words by applying the word similarity function based on sememes (defined by Definition 9). Facing error Chinese words, in the same way, we can give the semantic similarity between the error Chinese words and the correct Chinese words.

4.3 The Matching Method of Chinese Information Segments in Web Pages

Before recognizing which item an information segment belongs to, we set up a set of information items. For instance, the item set of medicine information includes an item on medicine name, an item on medicine usage and so on. Moreover, we set up a feature word set for every information item, and the feature word set is composed of some feature words to denote the information item. As shown in figure 2, F_1 is a feature set for the information item $item_1$.

The sememe-based semantic matching of Chinese information segments includes four layers. The first layer is named as word-sememe conversion, extracts the sememes from the words in the Chinese information segment and then uses the sememe vectors to denote the words. The second layer calculates the semantic similarity between sememe vectors and the feature words in the feature sets by using Eq.9, and then sums the similarities for the same sememe vector and the same feature set. For instance, m_{11} is the sum of similarities for $SemeV_1$ and set F_1. The third layer compares the sums of the same feature set. For instance, after comparing m_{1j}, m_{2j}, …, m_{ij}, …,m_{pj}, y_j is set as m_{ij} if m_{ij} is the largest. The fourth layer combines all components and output the vector $y=(y_1, y_2, \ldots,y_q)$. And then, we can recognize which item the Chinese information segment belongs to, by judging which component of vector y is the largest.

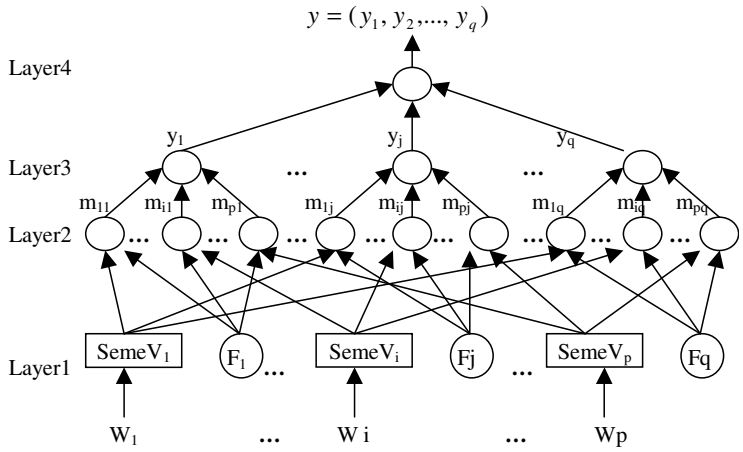

Fig. 2. The sememe-based semantic matching of Chinese information segments

4.4 Web-MIND System

We have successfully designed a monitoring system for Chinese medicine information in web pages, named as Web-MIND system. It can search, extract and analyze the illegal medicine information from the web pages, for the medicine administration to monitor the Chinese medicine information in web pages.

The Web-MIND system is composed of five components, which are search engine, segmentation of Chinese words, classification of web pages, information extraction of web pages, and matching of information segments. The search engine component applies the search engine "google" to search the Chinese web pages about medicine. The segmentation component of Chinese words uses a segmentation method of Chinese words based on language situation, which we have presented in the paper [14], to segment the Chinese sentences into Chinese words. The classification component uses a fuzzy classification method based on feature selection for web pages, which we have presented in the paper [15], to filter out all non-medicine web pages. The information extraction component extracts the information segments about medicine. The matching component of information segments uses the matching method of Chinese web information segments, which we present in this paper, to recognize which category the information segments belong to.

5 Experiments

The Web-MIND system got 930 Chinese web pages on medicine information from the Internet. Some new Chinese words and error Chinese words appearing in these pages did not exist in both HowNet and HowNet—medicine. After the information extraction component of the Web-MIND system extracting the information segments about medicine, the Web-MIND system adopted two methods to recognize which item the information segments belong to, and compared their accuracies. The two methods are a semantic matching method based on words and the semantic matching method based on sememes. Here, parameters were set as α=0.2 and β=0.6.

As shown in table 1, the two methods' accuracies were compared with respect to the category of information segments, which were medicine information about name, about efficacy, about caution, about usage and about manufacturer.

Table 1. The two methods' results

Information segments about	Quantity	Method A		Method B	
		Correct	Accuracy	Correct	Accuracy
Name	930	793	85.3%	830	89.2%
Efficacy	930	774	83.2%	823	88.5%
Caution	760	629	82.8%	668	87.9%
Usage	760	641	84.3%	676	89.1%
Manufacturer	640	544	85.0%	569	88.9%
Total	4020	3381	84.1%	3566	88.7%

On the one hand, the total accuracy of method A was 84.1% while the total accuracy of method B was 88.7%. Hence, the method B was more efficient than the method A with respect to the total accuracy.

On the other hand, the max deviation of the method A's accuracy was equal to 84.1%-82.8%=1.3%, while the max deviation of the method B's accuracy was equal to 88.7%-87.9%=0.8%. As to the information on efficacy and on caution, in which some new words came into being with much probability, the method B was more efficient than the method A as shown in the figure 3. Hence, the method B had more stable accuracies than the method A.

According to the above conclusion, the semantic matching method based on sememes is an efficient matching method to recognize which item the Chinese information segments belong to, especially for the information including some new Chinese words and error Chinese words.

6 Conclusion

This paper presents a novel semantic similarity between sememes, and analyzes the influence of the parameter by introducing and proving some theorems. On the base of semantic similarity between sememes, a semantic matching method of Chinese information segments is presented to recognize which item Chinese web information segments belong to, in which new words and error words occur. The semantic matching method is an efficient matching method, by applied successfully to the Web-MIND system.

References

1. Gao, M., Liu, C., Chen, F., An ontology search engine based on semantic analysis, 3rd International Conference on Information Technology and Applications, Sydney, Australia, (2005)256 – 259
2. Yang, J., Cheung, W.K., Chen, X., Integrating element and term semantics for similarity-based XML document clustering, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, University of Technology of Compiegne, France, IEEE Computer Society Press (2005) 222 – 228.
3. Da, L.G, Facon, J., Borges, D.L, Visual speech recognition: a solution from feature extraction to words classification, Proceeding of Symposium on Computer Graphics and Image Processing, XVI Brazilian (2003) 399 – 405.
4. Shen, H.T., Shu, Y., Yu, B, Efficient semantic-based content search in P2P network, IEEE Transactions on Knowledge and Data Engineering, 16 (7) (2004) 813 – 826.
5. Yi, S., Huang, B., Tatchan Weng, XML application schema matching using similarity measure and relaxation labeling, Information Sciences, 169(1-2) (2005) 27-46.
6. Nakashima, T., Classification of characteristic words of electronic newspaper based on the directed relation, 2001 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, Victoria, B.C., Canada, IEEE Computer Society Press (2001) 591 – 594.
7. Rada, R., Mili, H., Bichnell, E., Blettner, M., Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man, and Cybernetics, 9(1) (1989) 17-30.

8. Cross, V., Fuzzy semantic distance measures between ontological concepts, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the Fuzzy Information, Alberta, Canada, IEEE Computer Society Press (2004).635 – 640.
9. Soo, V., Yang, S., Chen, S., Fu, Y., Ontology acquisition and semantic retrieval from semantic annotated Chinese poetry, Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, Tuscon, AZ, USA, IEEE Computer Society Press (2004) 345 – 346.
10. Vladimir, A.O., Ontology based semantic similarity comparison of documents, 14th International Workshop on Database and Expert Systems Applications (DEXA'03), Prague, Czech Republic (2003) 735-738.
11. Cheng, L., Lu, Z., Wen, K., The exploration and application about amphibolous matching based on semantics, Journal Huazhong University of Science & Technology (Nature Science Edition), 31 (2) (2003) 23-25.
12. Rodriguez, M.A., Egenhofer, M.J., Determining semantic similarity among entity classes from different ontologies, IEEE Transactions on Knowledge and Data Engineering, 15 (2) (2003) 442 – 456.
13. Guan Y., Wang, X., Kong, X., Zhao, J., Quantifying semantic similarity of Chinese words from HowNet, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, China, IEEE Computer Society (2002) 234-239.
14. Zhang, M.Y., Lu, Z.D., Zou, C.Y., A Chinese word segmentation based on language situation in processing ambiguous words, Information Sciences, 162(3--4) (2004) 275-285.
15. Zhang, M.Y., Lu Z.D., A fuzzy classification based on feature selection for web pages, The 2004 IEEE/WIC/ACM International Conference on Web intelligence, Beijing, China, IEEE Computer Society Press (2004) 469-472.