Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (13): 2556-2561, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2556.2561

A Computing Method of Semantic Similarity Based on Distributed Fuzzy Relation

[1]Chang Baoxian, [3]Jiang Yi, [2]Zhu Xiaomei and [1]Chen Wei Wei
[1]College of Sciences,
[2]College of Electronic and Information Engineering, Nanjing University of Technology, Nanjing, China
[3]School of Information Technology, Yangzhou University, Yangzhou, China

Abstract: This study presents using RDF to construct domain concept hierarchical model, considering the inaccuracy and incomplete between the concepts based on the existing method of semantic similarity. The data gathering of RDF model comes from the domain pages automatically and then uses the distributed method to compute therelative-membership grade, constructs the fuzzy concept hierarchical model and at last computes the semantic similarity by using Dijkstra shortest path algorithm, the experiment shows that the method this study presents have more accurate and performance relative to existing methods.

Key words: RDF, dijkstra shortest path algorithm, fuzzy concept hierarchy model, semantic similarity

INTRODUCTION

Because of the rapid development of the Internet technology, the information resources on the Internet is growing exponentially, many different types of information are appeared, such as text, Web services, XML documents, graphics, images, sounds, videos and so on. People hope to obtain the required information quickly and effectively. So, the requirement of the information automatic processing intelligently becomes more and more urgent, one of the core problem of these automatic processing is the computing similarity. Such as the lexical/conceptual similarity calculation of word sense disambiguation, word error correction, ontology mapping, ontology merging; the sentences similarity calculation of automatic question answering system, information filtering technology, automatic text summarization, machine translation; the documents similarity calculation of document automatic classification and clustering technology; the matching degree of Web service discovery and combination. These information objects could be efficiently similarity calculation or not affects the accuracy of information processing directly.

In this study, we research the existing semantic similarity calculation method, construct domain concept hierarchical model by RDF model considering the imprecision and incompleteness between conceptions, the data of RDF model automatically generated from crawling and extracting from the domain pages and then use the distributed method to compute the relative-membership grade, construct the fuzzy concept hierarchy model and finally use the Dijkstra shortest path algorithm to compute the semantic similarity between concepts, the semantic similarity experiment proves that this calculation method, has higher accuracy and efficiency compared with the existing method.

RELATIVE RESEARCH ON SEMANTIC SIMILARITY CALCULATION

Dekang presents a broad definition of similarity in Lin (1998) the intuition tells us, the similarity between the objects A and B and the similarities and differences between them, the more common two objects have, then the similarity is greater, while the more difference between the two objects, the similarity less. When two objects A and B are the same object, reached the maximum similarity. When A and B are unrelated or independent, the similarity is minimal.

Dekang believes that any two words similarity depends on their commonality and differences and then gives the definition of the equation (Lin, 1998) from the perspective of the information theory:

$$Sim(A,B) = \frac{Log_p(Common(A,B))}{Log_p(Discription(A,B))}$$

where, in numerator means the quantity of information which describes the commonality of A and B, denominator means the quantity of information which describes the whole A and B.

According to the differences of similarity computing technology (Goldstone and Son, 2004) they can be divided into: The method based on name identification; the method based on statistics; the method based on synonymy thesaurus and the method based on the graph structure.

Corresponding Author: Chang Baoxian, College of Sciences, Nanjing University of Technology, Nanjing, China

Method based on name identification: The method based on name identification is the common similarity computing method, it uses the grammar-drive technology and morphological similarity to find the similarity between the strings of term representation. It mainly reflects the similarity degree of two elements of linguistics and structure. The method can be subdivided into equal measure and similarity measure.

The method has the small amount of calculation and it's simple, but the problem is when the two terms have the same connotation with different external form, this method will not be measuring semantic similarity correctly. Such as (chief, leadership).

Method based on statistics: The method based on statistics, such as in reference (Chatterjee, 2001), calculates the similarity with the relativity of words, selects a set of feature words firstly, then calculates the correlation between the set and each word (generally uses the frequency of the words appearing actually in the context in the large-scale corpus to measure), so the feature vector for each word can be got and then use the similarity between these vectors (usually use the cosine of the angle between the vectors to compute) as the similarity between two words. The following assumption is, the words are semantic similar between, the context should be similar. This method is more objective, reflects the similarities and differences in semantics and pragmatics comprehensively, but it is dependent on the training corpus, large amount of calculation and complex, in addition, disturbed by data sparse and noise greatly, sometimes appears obvious errors.

Method based on synonymy thesaurus: The method based on synonymy thesaurus, semantic distance between words can also be computed according to the synonym dictionary (Agirre and Rigau, 1995), in synonym dictionary all words are organized in a hierarchy dendrogram. In a dendrogram, there is only one path between any two nodes, so the path length can be measured as the distance between two words. Leacock and Chodorow (1998) uses WordNet to calculate the semantic similarity as follows, he not only considers the shortest path length (C1, C2) between nodes, but also considers other factors, such as limit the times of IS-A links, the depth D of the classification tree. The methods based on synonymy thesaurus is simple and effective, more intuitive and easy to understand, but the result of this method is influenced by subjective consciousness and sometimes it cannot reflect the objective facts accurately, this method reflects the similarities and differences accurately between the semantic aspects of words, but considers less on the pragmatic characteristics between the words:

$$\mathrm{Sim}(C_1,C_2) = -\log\left(\frac{\mathrm{length}(C_1,C_2)}{2D}\right)$$

Method based on the graph structure: The method based on the graph structure (Tversky, 1977) uses the knowledge of concept graph to calculate the concept similarity. Assuming two concept graph G1, G2, their intersection is . The similarity S of concept graph can be defined as two parts, concept similarity Sc and relation similarity Sr, as following Eq.:

$$S_c = \frac{2N(G_c)}{N(G_1)+N(G_2)}$$

wherein, N (G_c), N (G_1), N (G_2) represent the quantity of nodes in G_1, G_2, G_c, respectively:

$$S_r = \frac{2E(G_c)}{E_{G_c}(G_1)+E_{G_c}(G_2)}$$

Wherein E (G_c) represents the quantity of edges of graph G_c; E_{gc} (G_1), E_{gc} (G_2) represent the quantity of edges which one of the point of the edge connects with G_c at least in G_1, G_1.

The calculation method based on the graph structure focuses on the form, ignores the semantic characteristics of concept, emphasizing the similarity of form (structure).

SEMANTIC SIMILARITY CALCULATION BASED ON DISTRIBUTED FUZZY RELATION

This study presents the calculation method of semantic similarity based on distributed fuzzy relations, firstly uses open source search tool Nutch to search from the topic pages and filters pages content according to the theme of judging formula, then uses the NLP tool to deal with the page content, gets the three tuple and stores in HBase distributed, constructs domain concept hierarchy model according to the predicate IS-A relationship and calculates the degree of membership between the conceptual relationship according to the number of HBase data, at last uses the shortest path algorithm to compute the semantic similarity between concepts.

Nutch vertical search: Nutch (2009) is a Web search engine which is an open source code provided by Apache software foundation, we join the discrimination method of topic relevance in the use of Nutch search engine, use a theme correlation algorithm (Page *et al*., 1998) based on

vector space model, get the pages set we need by filtering the result of Nutch search engine. The process is divided into the following two parts:

- **Page analyzing module:** When spiders crawl a hypertext page, firstly analysis the marked structure, search grab information from its mark feature information and access URL links point to other hypertexts, continue to access the URL links through the cycle steps, collect the information of the page
- **Page filter module:** This module is responsible for extracting information on the webpage, extraction is firstly to filter out some script identifier and some useless advertising information, at the same time the layout format information is recorded and judge the topic relevance

The model uses vector space model, the traditional method is analyzing webpage content, extracting text in the webpage and then set the weight of the word according to frequency and the location, at last calculating the theme correlation. Due to the current webpage mostly adopts some optimization aiming at search engine, therefore, judges the webpage if there is a <meta> label, if is further to judge if there is a <keyword> tag, judges the relevance of the webpage quickly through metadata tag , if there are no <meta> tags, processes with the general method.

To calculate the weight of each word in the page, firstly counts the keyword frequency and calculates the frequency ratio, uses the highest frequency of the words as a benchmark, the frequency is presents as $X_i = 1$, X_i of other words, as each word in the set has a different identity, if one words appears in the title, it is much more important than, the word appears in the text, so quantifies the location information and adds to the weight of keywords, one webpage is divided into 4 parts: title(<title><head>) (B1), keywords (<foot><h><big><l><u>) (B2), link words <a> (B3), others(B4). First 3 parts are discriminated by HTML tag, the weight Wij of term Ki in page Dj and the theme relevance page D Sim (D) (Zhang *et al*., 2012) can be calculated by the equation as follows:

$$W_{ij} = \frac{\sum_{t=1}^{4} N_{B_t ij} \times W_{B_t}}{\max\left(\sum_{t=1}^{4} N_{B_t ij} \times W_{B_t}\right)}$$

$$Sim(D) = Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{N} W_1 k \times W_2 k}{\left(\sum_{k=1}^{N} (W_1 2k)\right)\left(\sum_{k=1}^{N} (W_2 2k)\right)}$$

Wherein, N_{btij} is the weight of term K_i in page Dj as to the page Bt, $k_i \in$ {terms in page Di}()(∑_()^)/(∑_()^( )_Dd If the theme relevance Sim(D) is greater than or equal tod, it means the page is relevant to the theme, otherwise is irrelevant, abandon the page, the threshold d is 60%.

Domain concept hierarchical model based on RDF: According to the results above, we use the tool Natural Language Processing (NLP) (Baeza-Yates, 2004) to process the relevant pages, get the following three tuple (subject, predicate, object).

We store the three tuple into HBase, in HBase:

Value = Map(TableName, RowKey, ColumnKey, Timestamp)

The data in every column is stored in column family.

The differences of three tables are the different data. The row-key of the table SPO is (subject, predicate), the Object value is stored in Column family; the row-key of the table POS is (predicate, object), the subject value is stored in Column family: The row-key of the table OSP is (object, subject), the predicate value is stored in Column Family.

We select the basic concepts in domain ontology, the obtained three tuple SPO, filter according to the concept, merged with the same semantic concept of the same depth through the P IS-A relation, the concept hierarchical model is finally obtained, as shown below (Fig. 1).

Relative-membership grade calculation: Definition 1: The relative-membership grade of concept A, for any one concept A in the concept hierarchical model, its relative-membership grade is the probability of the concept A belong to the parent node Q, in this study we define the calculation equation is:

$$F(A) = \frac{\sum_{i=1}^{m} ((1|I))_i \{A=Q\}}{\sum_{i=1}^{n} ((1|I))_i \{A=Q\}}$$

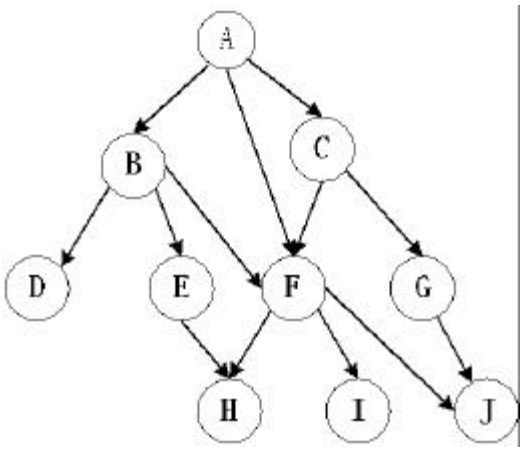

Fig. 1: Concept hierarchical model

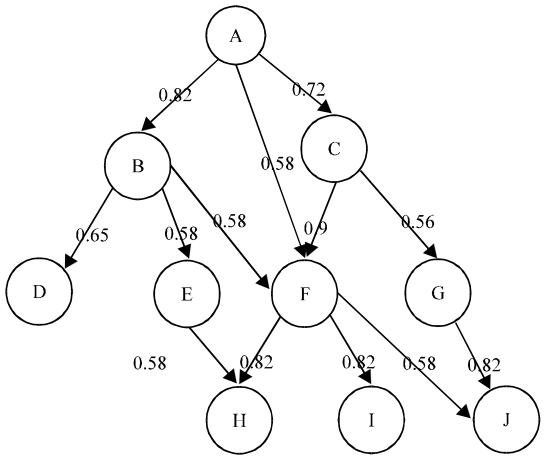

Fig. 2: Fuzzy concept hierarchical model

Wherein $I_I(A) = Q$ means A and Q is a IS-A relation, they satisfy S = A, P = IS-A, O = Q in (S, P, O); $P_i(A) = Q$ means Concept A and Q satisfy relation, S = A, O = Q in (S, P, O); m presents the quantity of the relations satisfy $I_i(A) = Q$, n presents the quantity of the relations satisfy $P_i(A) = Q$.

At last we can get the fuzzy concept hierarchical model as follows (Fig. 2).

Semantic similarity calculation algorithm based on the shortest path: Every two concepts have the relative-membership grade according to the fuzzy concept hierarchical mode above, it can be seen as the weight of the two concepts. So, we can use the optimal path algorithmto calculate the semantic similarity, any concept A and B, their semantic similarity is the shortest semantic distance between them by normalization. This method considers the factors of density and distance, considers the attribute according to the fuzzy relation.

Input: Concept A, concept B, fuzzy concept hierarchical model U.

Step: Initialization: Suppose A is the origin, the path length of A is give to 0 (d[A] = 0), then set the path length of the direct concept Md [M] = w (A, M), set the path length of remaining not direct concept d [v] = ∞.

Choose a concept K which is the nearest concept with A from U, put it into S (The chosen distance is the shortest path length from A to K).

Consider K as a new middle concept, modify every concepts' distance of U; if the distance from A to u (u∈U) through K is shorter than not through K, modify the distance of U, the value is add the distance of K into the edge weight.

Repeat step 2 and 3 until all concepts are in S.

Calculate the shortest path length Sr between A and B normalization:

$$Sim(A,B)=\frac{Sr}{n}$$

Output: The semantic similarity between concept A and B.

Map-reduce calculation: Considering the large scale data in HBase, we use map-reduce to calculate the relative-membership grade and the shortest path distributed, the relative concept and technology of map-reduce please refer to document (Deanand Ghemawat, 2009).

We split data into multiple subdomain U_i and divide into different server, the map work is:

- Calculating the relative-membership grade between parent node and child node in fuzzy concept hierarchical model according to the concept relative-membership grade equation
- Calculating the semantic similarity between A and B in fuzzy concept hierarchical model on this server according to the similarity calculating algorithm

Reduce work is:

- Weighted sum the relative-membership grade from every server, getting the final relative-membership grade between parent and child, the Eq. is:

$$F(A)=\frac{\sum_{i=1}^{m} f_i(A)\times Sum(U_i)}{\sum_{i=1}^{m} Sum(U_i)}$$

Wherein, $f_i(A)$ is the relative-membership grade in subdomain i, ()(?_()^|¡¼¡¼¡½_ _¡½)/(?_()^|¡¼_¡½)D_Dd Weighted sum the similarity between concept A and B, we get the final semantic similarity between A and B, the Eq. is as follows:

$$Sim(A,B)=\frac{\sum_{i=1}^{m} S_i(A,B)\times Sum(U_i)}{\sum_{i=1}^{m} Sum(U_i)}$$

EXPERIMENTS

An experiment from telecom domain is used to test and verify the semantic similarity method this paper presents. Initial URL set is selected from some famous website in telecom domain, seen as Table 1.

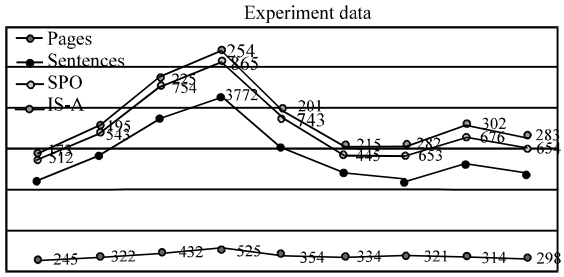

Fig. 3: Experiment test data

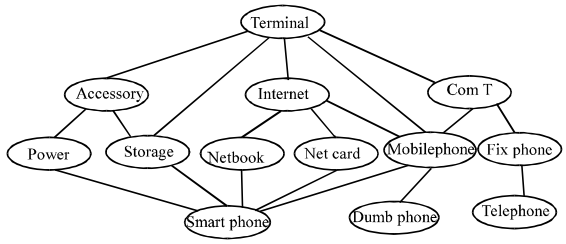

Fig. 4: Concept hierarchical graph (part)

Table 1: Initial URL set of telecom domain

1	http://www.cww.net.cn/	Communication world
2	http://www.chn3g.cn/	3G world
3	http://www.comcw.cn/	Communication operation
4	http://www.c114.net/	No. 1 communication portal
5	http://www.txrjy.com/	Communication home
6	http://www.mc21st.com/	Mobile communication online
7	http://www.cctime.com/	FeiXiang
8	http://www.catr.cn/	Taier
9	http://www.chinacie.com.cn	Communication equipment

Table 2: Part of important concepts in communication domain

Information terminal product	Multimedia communication terminal	IP communication terminal	ISDN Terminal
Telephone accessories	Intercom	Fax	Phone
Satellite radio tracking system	Commercial Computer	Laptop	Home-use PC
Server	Scanner	Storage device	Optical access equipment
Wireless access equipment	HFC access equipment	XDSL access equipment	Ethernet access equipment
SPC exchange	Sim card	Microwave transport	Route
Switch	Telephone	Netbook	……

Table 3: Experiment result

Concept A	Concept B	Semantic similarity
Power	Telephone	0.542
Storage card	Internet terminal	0.350
Smart phone	Accessory	0.417
Terminal	Dumb phone	0.797
Netbook	Telephone	0.373
…	…	…

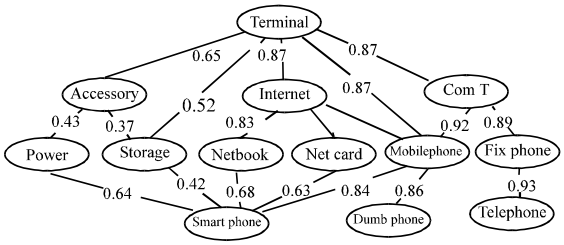

Fig. 5: Fuzzy concept hierarchical graph (part)

For the experiment effect and efficiency, we reduce the search scope, deal with the page only in these websites, grab the pages which contain the terminal equipment and specify the important concepts in the communication domain, we only select the sentence which object or subject are in the concepts set. Part of important concepts is in Table 2; the experiment, we grab 763 pages from 9 websites, the data is seen as Fig. 3.

Figure 4 and 5 is the concept hierarchical graph and fuzzy concept hierarchical graph got from we calculated and processed.

At last we can get the semantic similarity we supposed base on our method as Table 3.

CONCLUSION

The distributed semantic similarity calculation method based on fuzzy relation this study presents, constructs the fuzzy concept hierarchical model by calculating the relative-membership grade between concepts, it considers the inaccuracy and incomplete between the concepts, reflects the relation between concepts objectively, is independent from the domain ontology and expert knowledge. But it has space to promotion on the sample chosen, efficiency and accuracy and these are our next focuses.

ACKNOWLEDGMENT

This study was supported by The National Natural Science Fund under Grant No. 61170201.

REFERENCES

Agirre, E. and G. Rigau, 1995. A proposal for word sense disambiguation using conceptual distance. Proceedings of the International Conference Recent Advances in Natural Language Processing, (RANLP'95), Tzigov Chark, Bulgaria.

Baeza-Yates, R., 2004. Challenges in the interaction of information retrieval and natural language processing. Proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistics, February 15-21, 2004, Seoul, Korea, pp: 445-456.

Chatterjee, N., 2001. A statistical approach for similarity measurement between sentences for EBMT. Proceedings of Symposium Translation Support Systems (STRANS), February 15-17, 2001, IIT, Kanpur.

Dean, J. and S. Ghemawat, 2009. MapReduce: Simplified data processing on large clusters. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/mapreduce-osdi04.pdf

Goldstone, R.L. and J.Y. Son, 2004. Similarity. Psychol. Rev., 100: 254-278.

Leacock, C. and M. Chodorow, 1998. Combining Local Context and Word Net Similarity for Word Sense Identification. In: WordNet: An Electronic Lexical Database, Felbaum, C. (Ed.). MIT Press, Cambridge, MA., pp: 265-283.

Lin, D., 1998. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, July 24-27, 1998, Morgan Kaufmann Publishers Inc., San Francisco, CA., pp: 296-304.

Nutch, 2009. The Java search engine [Z]. http://lucene.apache.org/

Page, L., S. Brin, R. Motwani and T. Winograd, 1998. The PageRank citation ranking bring order to the web. Technical Report, Stanford Digital Library Technologies Project, Standford University, USA.

Tversky, A., 1977. Features of similarity. Psycol. Rev., 84: 327-352.

Zhang, W.L., Y.W. Liu and J. Sun, 2012. Research on vertical search engine based on Nutch. J. Nankai Univ.(Nat. Sci.), 45: 37-44.