International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 10 (2014), pp. 121-128
http://dx.doi.org/10.14257/ijmue.2014.9.10.12

Image Retrieval of Semantic Similarity Measure based on Probability-weighted

Qian Wang[1], Chunli Zhang[2] and Lixin Song[2]

[1] *School of Computer Science,*
[2]*School of Electrical and Electronic Engineering,*
Harbin University of Science and Technology
Harbin 150080, China
qiamwamg@163.com

Abstract

For multi-level semantic structure, the asymmetry of similarity between semantic concepts, as well as the different correlation of semantics between the children nodes and father node, this dissertation proposed a novel similarity calculation method of image semantic based on the probabilistic weighting. This method combines the image feature mapping the visual characteristics of the underlying semantic with the domain ontology description to build a tree-like hierarchical semantic model. According to what the posterior probability and conditional probability were gained by Bayesian network learning, and further for those semantic similarity who are based on semantic distance took the weighted processing so as to get the final similarity of image semantic. Moreover, taking the medical image semantics as the experiment object in weighted method can improve the retrieval performance compared with the non-weighted similarity calculation method.

Keywords: *Hierarchical semantic structure; Semantic similarity; Bayesian network; Probability-weighted*

1. Introduction

Effective image similarity calculation method provides efficient retrieval an important protection. In order to make the image retrieval capability reach the level that people can understand, semantic-based image retrieval technology gradually become into a research hotspot, in which the image semantic similarity metrics become one of the key research questions [1].

The traditional semantic-based image retrieval is based on the use of text marked images to achieve an exact match keyword, while the image retrieval should be imprecise retrieval form. form, image retrieval based on tree-like hierarchical semantic model and similarity measure should get attention [2-4], its main use in the semantic model: (1) hierarchical semantic based on keyword [2]; (2) hierarchical semantic that combines keywords with image semantic features mapped from low level visual features [4]; its main use of distance-based semantic similarity measure on semantic similarity measure. There are three tree hierarchical semantic similarity measure methods， one of which is to calculate semantic distance between concepts, and then converted to semantic similarity [5]; one is similarity metrics based on the amount of information [6,7]; another one is a semantic similarity measure method that integrates distance and information content. But all these recent researches ignore the influence of semantic asymmetry [9] and differences between "is-a" relation nodes.

In this paper, support vector machine (SVM) is used to map the low level image features classification into visual semantic, and it is associated with image description domain

ISSN: 1975-0080 IJMUE
Copyright ⓒ 2014 SERSC

ontology to construct multi-level semantic structure of the tree model. Then consider the conditional probability between semantic concepts which are got from Bayesian network learning as the forward and reverse factor and weight the semantic similarity model, which overcomes asymmetry and reflects the similarity measure of children nodes and father node under the condition of different correlation.

2. Multi-level Concept of Semantic Similarity Measures

To achieve semantic image retrieval, in-depth study of semantic similarity calculation is done on the basis of the given image hierarchical semantic description model.

2.1. Tree-based Semantic Similarity Measure

In a tree structure, the formula of semantic similarity between concept i and concept j is [10]:

$$d(i,j)=g(dep(c)\cdot f(l(c_1,c_2)\cdot f(den(c_1,c_2)) . \quad (1)$$

$$s(i,j)=\lambda\cdot d(i,j)+(1-\lambda)\cdot spath(i,j) \quad (2)$$

In which λ is adjustable parameter, g (*) is proportional function, f (*) is inverse proportional function. S is the matrix of similarity between concepts in the semantic tree model, s (i, j) is the similarity between concept i and concept j.

2.2. Multi-level Semantic Description Model

To realize image semantic retrieval, semantic descriptions need to be done to form hierarchical semantic structure from the visual semantics to the high-level semantics layer. In the specific application, ontology to realize image semantic description is used in some current studies to realize tree-like hierarchical semantic structure. In this study, the image low level visual features are mapped to the visual semantic associated with domain ontology concept semantic to constitute a tree-like hierarchical semantic.

Firstly, using binary classification or multi-class classification methods, the image low level features are mapped to visual semantics, and visual concept ontology semantics of these visual features constitute the image of the object. Then combine the domain ontology semantic of image objects and visual concepts to construct a tree-like image hierarchical semantic description model.

2.3. Probability-weighted Semantic Similarity Calculation

The similarity between concepts is asymmetric, and in the practical application of semantic search, the matching also has a direction [11]. In addition, there are also some differences the similarities between siblings semantic concepts. However, traditional similarity calculation method can not reflect this difference. Therefore, a method to fix is needed.

Bayesian network is able to propose a causal relationship between the semantic description of the qualitative and quantitative, and tree-like hierarchical semantic structure can be considered precisely as Bayesian network structure, so Bayesian inference process can be further realized. Semantic similarity between concepts is affected not only by the traditional semantic distance and other factors, but also by the causality between of semantic concepts. So this paper proposes an improved method, using probability to weighting traditional similarity.

 Copyright ⓒ 2014 SERSC

2.3.1. Bayesian Network Learning: The multi hierarchy semantic structure model is taken as the topological structure of Bayesian network, and pick the training set. Under the conditions that node parameters are independent and the data set is with integrity, using maximum likelihood estimation method [12] to obtain conditional probability between the nodes. Based on the conditional probability distribution, add evidence, then the posterior probability of each node in the Bayesian network can be got through reasoning.

The algorithm is as follows, the Bayesian network is made up of n random variables X={X1, X2, ⋯, Xn}, and the and the joint probability distribution is:

$$p(X_1, X_2, \cdots, X_n) = \prod_i p(X_i \mid Pa_i) \tag{3}$$

Pa_i is the parent node of X_i. Make Pa_i^j as the jth of the parent node of X_i, x_i^k is the k-th value of X_i, the network parameters can be expressed as:

$$\theta_{ijk} = p(x_i^k \mid Pa_i^j), \qquad \text{and} \sum_k \theta_{ijk} = 1 \tag{4}$$

The purpose of this learning is to find the maximum parameter learning vector θ appeared on the complete data set. Assuming the distribution of the observations is independent in the case of unknown, the maximum likelihood function of parameter can be expressed as:

$$L_D(\theta) = \log \prod_i \prod_j \prod_k \theta_{ijk}^{N_{ijk}} \tag{5}$$

In which N_{ijk} is the number of x_i^k observations in the dataset occur in the case of Pa_i^j appears.

2.3.2. Weighting Algorithm: In order to weight the similarity matrix S which is obtained based on semantic distance, building weight matrix W with same dimension of matrix S is needed. W is shown as follows:

$$W = \begin{bmatrix} 1 & w_{12} & \cdots & w_{1j} & \cdots & w_{1n} \\ w_{21} & 1 & \cdots & w_{2j} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & & & \vdots \\ w_{i1} & w_{i2} & & 1 & & w_{in} \\ \vdots & \vdots & & & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nj} & \cdots & 1 \end{bmatrix}$$

In which W_{ij} is the weight coefficient between node i and node j, i=1,2,⋯,n, j=1,2,⋯,n.

To obtain W_{ij}, firstly, calculate the weight value of the shortest path from the node i to j to obtain the matrix W_{ij}. Because this path maybe not the only one, the t-th is calculated as follows:

$$w_{ij}^t = \prod_{l=i}^{j} w_{lk} \tag{6}$$

$$\text{s.t.} \quad w_{lk} = P(B_k \mid B_l) \quad and \quad |k - l| = 1$$

In which l and k are the neighbor nodes when node i goes to node j, and w_{ij} is the weight value between the two nodes. $P(B_K|B_l)$ is the probability of B_K when B_l meets the condition, when k is the parent node of l, P is its posteriori probability, while when k is child node of l, P is the conditional probability.

Copyright ⓒ 2014 SERSC

Secondly, when the shortest path is not only, W_{ij} as a final concept has a weight of two internodes:

$$w_{ij} = \max(w_{ij}^{1}, \cdots w_{ij}^{t}, \cdots, w_{ij}^{q}) \quad (7)$$

In which q is the number of the shortest path. Finally, the weighted semantic similarity is as follows:

$$Sim = W \cdot S \quad (8)$$

In which the factor of this matrix $Sim(i, j) = W_{ij} \bullet s(i, j)$.

2.3.3. Analysis of Probability Weighted Similarity Principle: The conditional probability in the Bayesian networks can quantitative said child nodes degree of dependence on the parent node, while the posterior probability can quantitative said the incidence of the parent node reasoned by child nodes. Therefore, consider the conditional probability and posterior probability as distance-based semantic similarity measure forward and reverse weight factor which reflects asymmetry between nodes in the hierarchical model of semantic.

The differences between its semantics can be seen through probabilistic weighting the similarity measure of siblings based on distance. Then confirm the feasibility through the analysis. A hierarchical structure is shown in Fig. 1 in which A~F are the six concepts of multi-level semantic structure. A and B are parent-node semantic, and C、D、E and F are child-node semantic in the next layer.

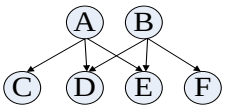

Figure 1. The Example Level Structure

According to the known probability P(A|C)=1, P(B|F)=1,(A|D)+P(B|D)=1,P(A|E)+P(B|E)=1, and assume that P(A|D)≥P(B|D), P(B|E)≥P(A|E). So we can know that:

$$P(A|C) > P(A|D) > 1/2 > P(A|E) \quad (9)$$

When the samples are enough, P(C)=P(D)=P(E)=P(F), and then:

$$P(A|C)P(C) > P(A|D)P(D) \geq P(A|E)P(E) \quad (10)$$

And thus:

$$P(C|A) \geq P(D|A) \geq P(E|A),\ P(A|C)P(D|A) \geq P(A|C)P(E|A) \quad (11)$$

Consider them as weight value of C and D and of C and E, which is expressed as follows:

$$W_{CD} \geq W_{CE} \quad (12)$$

In the same way, P(B|F)P(E|B)≥P(B|F)P(D|B), and

$$W_{FE} \geq W_{FD} \quad (13)$$

As can be seen from the semantic distance measure based on traditional methods, the similarity between its nodes is: S(C,D)=S(C,E), S(F,E)=S(F,D), after the probability weighted according to equation (12) and (13), Sim(C,D)≥Sim(C,E), Sim(F,E)≥Sim(F,D). As can be seen, the influence of the causality between nodes to the similarity can be expressed by the

 Copyright ⓒ 2014 SERSC

probabilistic weighting, and it's feasible to describe the similarity differences between child-parent nodes and child nodes in the same layer.

2.4 Application Example

This paper uses mammography and its clinical diagnosis descriptive semantics from DDSM (Digital Database for Screening Mammography) of University of South Florida as a research instance, to constitute breast calcifications hierarchical semantic structure model.

In the process of constructing hierarchical semantic model, low-level features extracted from mammography are mapped to visual semantics, including mean, variance and energy of gray features, the roughness of texture features, as well as calcification cluster density of shape features. "One-to-one" multi-classification in support vector machine is used to form hierarchical visual semantic such as roughness, gray evenness and calcification cluster density.

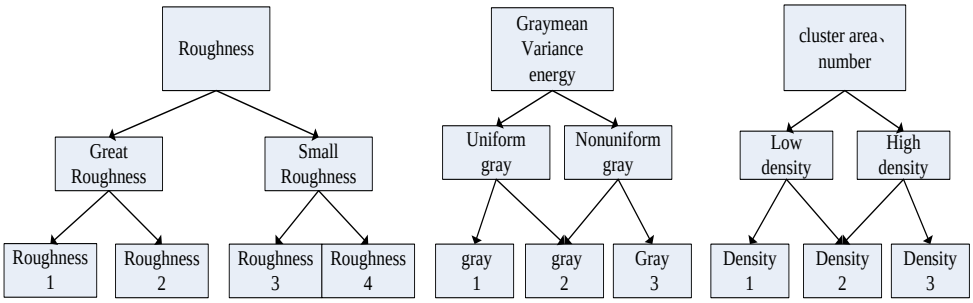

Figure 2. Semantic Extraction of Visual Features

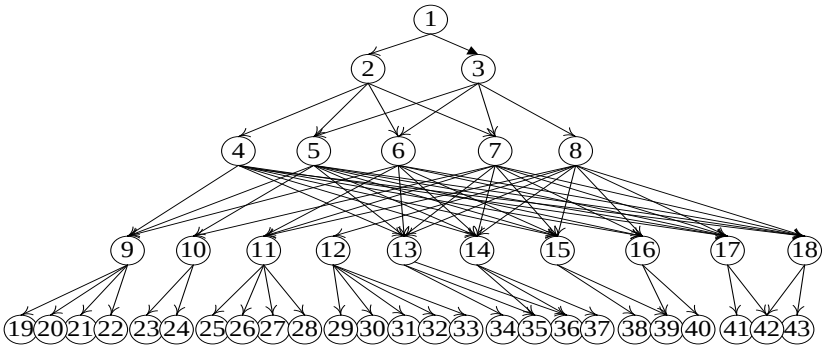

Figure. 3 Multi-level Semantic Structure of Breast Case

Combine the semantic features of image with the domain ontology concept semantic to constitute a multi-level semantic structure model shown in Figure 3, the model and the semantic of each node shown in Table 1.

Use the hierarchical semantic structure model as topology of Bayesian network, select 180 breast cases as the training set, use maximum likelihood estimation method to obtain conditional probability between nodes; On the basis of the conditional probability distribution, add evidence and use of Jtree algorithm [13] to obtain the posterior probability; And then obtain weights between concept nodes according to the weights matrix calculation method mentioned in 2.3.2, weight the similarity based on the traditional semantic distance and get the final degree of similarity between semantic concepts.

Copyright ⓒ 2014 SERSC

Table 1. The Semantic of Each Node

Sementic of the describe case		Visual feature semantics	
Semantic Name	Number	Semantic Name	Number
Calcified lesions	1	Great roughness	13
Benign	2	Small roughness	14
Malignant	3	roughness 1-4	34-37
Category 1-5	4-8	Uniform gray	15
Simple shape	9	Nonuniform gray	16
Scattered distribution	10	gray 1-3	38-40
Intensive distribution	11	Low density	17
Complex shape	12	High density	18
Other semantic	19-33	Density 1-3	41-42

3. Experimental Results

Experimental database is set up by 248 calcification cases. And make the evaluation of performance about the similarity calculation method.

3.1. Experimental Comparison of Weighted Semantic Similarity

To compare changes of weighted semantic similarity, take 4 semantics to calculate similarity in Figure 3. 35, 36, 37 represent three levels of roughness, semantic similarity between 35 and 36 should be greater than it between 35 and 37. 14 is parent node of 35, 36 and 37. The similarity between parent node and child nodes is asymmetric, as shown in Table 2.

Table 2. The Comparison of Semantic Similarity Before and After Weighting

Semantic node		14	35	36	37
Weight value	14	1	0.00125	0.42781	0.67184
	35	0.09999	1	0.44529	0.00007
	36	0.29999	0.29326	1	0.28918
	37	0.66573	0.00121	0.41869	1
Unweighted semantic similarity	14	1	0.73034	0.73034	0.73034
	35	0.73034	1	0.75229	0.75229
	36	0.73034	0.75229	1	0.75229
	37	0.73034	0.75229	0.75229	1
Weighted semantic similarity	14	1	0.00091	0.31244	0.49067
	35	0.07303	1	0.33499	0.00054
	36	0.21910	0.22062	1	0.21755
	37	0.48621	0.00091	0.31498	1

By Table 2, it can be seen using probability-weighted approach more clearly indicate semantic similarity difference and asymmetry between parent node and child nodes at the same level.

 Copyright ⓒ 2014 SERSC

3.2. Comparative Experimental Results of Image Semantic Retrieval

In order to verify the effectiveness of the probability-weighted algorithm in image semantic retrieval, make a comparison between it and the result without weighted. Randomly select three cases of breast to complete the three experiments. The input is image and its corresponding semantics, including visual semantics extracted from image and corresponding descriptive semantics given by doctor. Section 1, the descriptive semantics given by doctor is "benign (Node 2), clustered (node 27), pleomorphic (node 31), category 4 (Node 7) ", the semantic features extracted from the image is "roughness 2 (node 35), (node 40), low density (node 17) "; Section 2, the descriptive semantics given by doctor is "malignant (Node 3), clustered, pleomorphic, category 5 (node 8)", the semantic features extracted from the image is "roughness 3, gray uniformly 3, high density "; Section 3, the descriptive semantics given by doctor is "benign, lucent centered, N / A, category 2 (node 5)", the semantic features extracted from the image is " roughness 2, gray uniformly 3, low density".

Figure 5 is the comparison of weighted and non-weighted similarity measure retrieval results, it contains three experiments, 50 cases in front of each group sorted by descending according to the similarity. As can be seen, the weighted similarity retrieval result, its sequence is: semantics are exactly the same with the retrieval example, one semantic is different but similar, others similarity sorting rules (two or more semantics are different but similar). Therefore, this retrieval method is better to solve hierarchical semantics asymmetry and different relevance between semantic concepts.

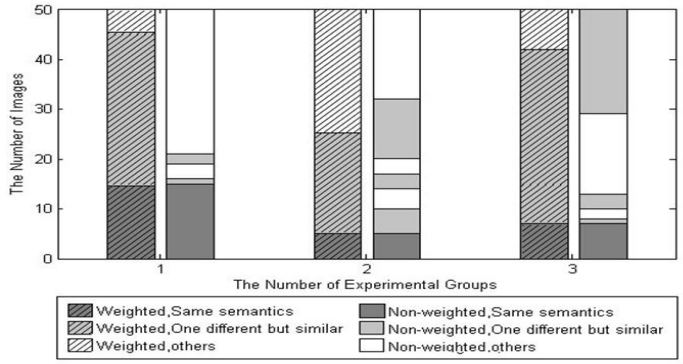

Figure 5. The Comparision of Image Semantic Retrieval Results

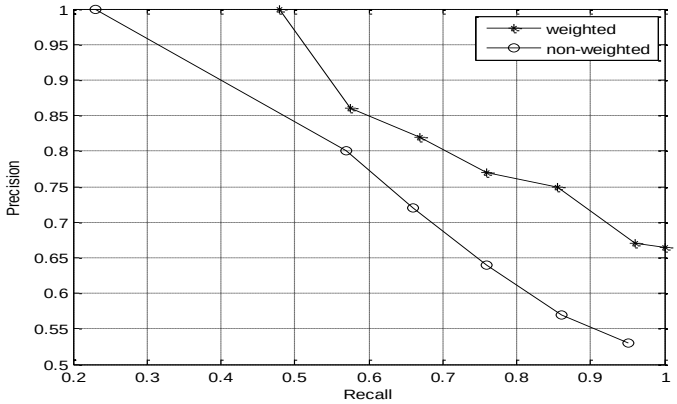

Figure 6. The Comparison of PVR Cures

Copyright ⓒ 2014 SERSC

In order to verify the performance of weighted similarity retrieval, this paper presents the precision and recall curves of weighted and non-weighted retrieval results, as shown in Figure 6. It can be seen the weighted retrieval performance is better than the non-weighted one.

4. Conclusion

This paper proposes a probability-weighted image semantic similarity measure approach based on a hierarchical semantic model of tree structure, which weights semantic similarity measure approach based on the distance and semantic similarity measure approach based on the conditional probability and posterior probability, and then combines them. This approach solves the asymmetric between hierarchical semantic similarities and the different semantic relevance between child nodes and parent node, in order to improve the semantic image retrieval performance significantly. Calculation will increase with the increase of the concept. The next step should consider optimization problems, in order to calculate the similarity more easily.

Acknowledgements

This work is supported by Natural Science Foundation of Heilongjiang province (F200912), Science and technology innovation fund of Harbin (20IORFXXS026).

References

[1] C. H. Chen, S. L. Hsieh and Y. C. Weng, *et al.* "Semantic similarity measure in biomedical domain leverage web search engine", 32nd Annual International Conference of the IEEE EMBS, (**2010**), Buenos Aires, Argentina.
[2] S. Chen, Z. Li and L. Yuan, "An Image Retrieval Method Based on Multi-level Semantic Similarity Measure", J. Northwestern Polytechnical University, vol. 26, (**2008**), pp. 588—591.
[3] G. Fu, T. Sun and X. Jiang, "Image Retrieval Based on Semantic-binding Hierarchical Visual Vocabulary", J. Shanghai Jiaotong University, vol. 45, (**2011**), pp. 154—158.
[4] P. Huang, C. Chun and W. Can, "Using Weighted Image Annotation Improve Web Image Retrieval", J. Zhejiang University (Engineering Science), vol. 43, (**2009**), pp. 2129—2135.
[5] H. A. Nguyen and H. Al-Mubaid, "A Combination-based Semantic Similarity Measure Using Multiple Information Sources", IEEE International Conference on information Reuse and Integration, (**2006**), Waikoloa, USA.
[6] P. Resnik, "Using Information Content to Evaluate Semantic Similarity In a Taxonomy", Proceeding of the 14th International Joint Conference on Artificial Intelligence, (**1995**), Montreal, Canada.
[7] L. Ziyu and H. Lei, "Calculation Research on the Concept of Semantic Similarity Based on Domain Ontology Model", J. Railway Society, vol. 33, (**2011**), pp. 52-57.
[8] H. Shiguo and G. Geng, "Summary of Semantic Similarity Measure Methods", J. Computer Applications and Software, vol. 25, (**2008**), pp. 37-39.
[9] H. Xu, Y. Fang and Y. Feng, "Improved Concept Similarity Calculation Method Based on Semantic Distance in Web Service Matching", J. Computer Applications, vol. 31, (**2011**), pp. 2808-2810.
[10] S. X. Xia, Z. H. Hu and Q. Niu, *et al.* "An Approach of Semantic Similarity Measure between Ontology Concepts Based on Multi Expression Programming", Sixth Web Information Systems and Applications Conference, (**2009**), Xuzhou, China.
[11] H. Zhe and C. Zheng, "Improved Concept Semantic Similarity Computation", J. Computer Engineering and Design, vol. 31, (**2010**), pp. 1121-1124.
[12] S. Russela and P. Norving, "Artificial Intelligence", M. People China Posts and Telecommunications Press (**2004**)
[13] F. V. Jensen and T. D. Nielsen, "Bayesian Networks and Decision Graphs", M. (**2007**).

 Copyright ⓒ 2014 SERSC