A New Algorithm for Uncertain Problem of WEB Page Classification

Xiaodan Zhang
Institute of Scientific and Technical Information of China, Beijing, P.R. China, Beijing, China
Email: zhangshenyang@126.com

Abstract—**For solving the uncertain problem in the process of WEB page classification, a general fusion classification model and algorithm are proposed, which based on model theory of information fusion. In the model, the hidden classification information is extracted from the WEB page，pre-processed firstly, then the processed data are input into the fusion mode, which deals with the different data with fusion algorithm, then the final classification results are concluded. An improved Bayesian network is proposed, which can not only solve the uncertain problem during the WEB page classification, but also can reduce the time complexity of the inference. The WEB data of NSTL are adopted in the experiment. The experiment proves the fusion model and fusion algorithm can solve the uncertain problem effectively**

Index Terms—**uncertain problem, WEB page classification, Bayesian Network, NSTL**

I. INTRODUCTION

WEB page classification is the hotpot problem in the information retrieval filed. How to improve the WEB page classification precision is the key problem now.

There are many characteristics in the process of WEB page classification, such as, the predefined type being multiclass usually, the class border existing fuzziness, the WEB page belonging to multi-type possibly, the training sample tagging by hand uncertainly and incompletely possibly, etc., which can lead to the uncertain WEB page classification result. So, the WEB page classification is a process of solving uncertain problem [1-3].

How to solve the uncertain problem is important in the paper.

In the paper, a WEB page classification fusion model and algorithm are proposed, which can solve the uncertain problem in the process of the WEB page classification effectively, and improve the precision and recall rate. In the end, the model and algorithm are realized in NSTL (National Science and Technology Library) system.

II. WEB PAGE CLASSIFICATION FUSION MODEL

The traditional WEB page classification algorithms use text classification algorithm with the only text information of the WEB page. But there is much hidden classification information in the WEB page [Shen et al. (2004)]. The key point is how to make the best of the hidden information (including WEB label, media keyword and multi-media information, etc.) to solve the uncertain problem in the WEB page classification.

A. Information fusion

Information fusion is the process of acquisition and integration kinds of information sources, multi-media information and multi-format information to get complete, precise, timely, effective and general information. The information fusion system can express the features of detection object completely, eliminate the uncertainty of information and improve the reliability of the input data. The technology has been applied in many fields [4-7].

Based on the advantage of fusion mode of information fusion, a WEB page classification fusion model for uncertain problems is built in the followed.

B. WEB page classification fusion model

For simplifying the complexity of WEB page classification, different kind information is extracted and preprocessed firstly, after the processed results deal with, the final classification result is got; the process is showed as Figure 1.

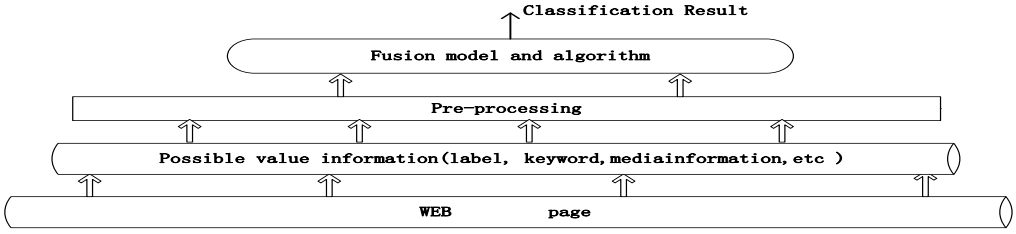


Figure 1. WEB page classification fusion model of uncertainty

From Figure 1, we can see, much value information of the WEB page is extracted firstly, such as text, label, multi-medias, etc. After pre-processed differently, the obtained information is input into the fusion center. The

doi:10.4304/jsw.7.3.526-531

final WEB page classification result outputs after fusion inference of the fusion center.

In the fusion model, the structure disposes much value information of the WEB page, which can get fusion result precisely, and reduce the uncertainty and complexity of the whole system. The system can realize flexible, self-adapting and efficient inference.

III. WEB PAGE CLASSIFICATION FUSION ALGORITHM

The fusion model proposed in the segment 2 can reduce the uncertainty during the classification process. So, the indication and quantitive measure become the basic problem of the algorithm study.

Kinds of uncertain factors in the classification process influence the WEB page precision, so, it is important to select an appropriate uncertain algorithm.

For uncertain problem, there are many algorithms now. Such as Bayesian algorithm, certainty theory, possibility theory, etc. Bayesian algorithm is better than other algorithm because of the stronger probability theory basement, easy problem express ability and strong uncertain inference ability [8-12].

WEB page classification is a process of causal reasoning. Bayesian network is a kind of effective causal network, which can express complex relation of multi-variable. So, Bayesian network is suited to the uncertainty of WEB page classification [13-16].

The uncertain problem has been included in the classification process, and the probability value is used at expressing and measuring uncertainty in the study of uncertainty.

The nodes of the fusion Bayesian network are showed as followed:

Input node, expressing all kinds input parameters of the fusion Bayesian network, which are the information of every parameter, including label and other possible value information for classification in the WEB page.

Output node, expressing all output parameters of the fusion Bayesian network, which are the probability values of the classification result.

Middle situation node, expressing situation transform during the process of inference in the fusion Bayesian network.

If there are n nodes in the WEB page classification fusion system, which include all kinds value information nodes extracted from the WEB page, middle situation node and typal node, etc... Probability value is used in the express of the relation between the nodes. Then the correlation between the nodes can be expressed as Figure 2.

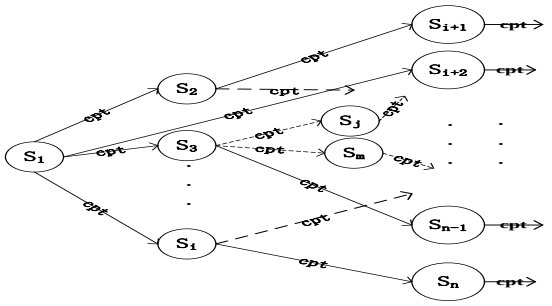


Figure 2. Bayesian network express of WEB page classification

From Figure 2, we can see a classification fusion Bayesian network. Because the WEB page classification can be seen as a process of probability reasoning, the express and solution of uncertain problem can been used in the process of WEB page classification.

The nodes of Bayesian network can be divided into information node, situation node, typal node and other node. In those, information node shows the value information extracted from the WEB page, which includes possible value information for classification, and is the main channel of getting WEB page information. The edge shows correlation between nodes, for example, parameter information causes the change of situation, situation influences parameter, etc... Condition probability value shows correlation of adjacent nodes.

After the structure and condition probability of the node are decided, the WEB page classification inference can be executed to get typal node probability value automatically.

In the paper, Simulated Annealing Algorithm and Maximum Likelihood Estimate Algorithm are select as the structure study algorithm and parameter study algorithm of the fusion Bayesian network respectively.

IV. IMPROVED BAYESIAN NETWORK INFERENCE ALGORITHM

The main disadvantage of the fusion Bayesian network is that inference time is hard to be predicted, which is the NP. But real-time and high efficiency are required for the WEB page classification. How to improve the Bayesian network is the main problem in the uncertain inference process.

Bayesian network is a graph structure with nodes and edges, the nodes show situation in the process of classification, and the edges show the relation between the nodes. When the Bayesian network is determined, the WEB page classification can be seen as a process of multi-information node inferring typal node. Graph search strategy is adopted in searching typal node in the paper.

The basic idea of the Graph search strategy is pattern recognition, and the different search path can be seen as different mode type.

Statistical decision theory is the basement of path classification, which can judge classification feature vector of extracted current path, and decide the

classification of current situation development trend according to the condition probability density function of the classification.

To solve the uncertainty problem, the Bayesian network inference algorithm utilizes kinds of WEB page information. According to the Estimated Condition Probability density $p(x_1, x_2, \cdots, x_k \mid w_i)$, the extracted feature vector of current information is $(x_1, x_2, \cdots, x_k)$, the Bayesian path classification algorithm in formula (1) judges current situation path step by step.

Suppose that the initial node of the Bayesian network is S_0, search threshold value is N, the classification situation node set is S_T. The determined node set is D1, undetermined node set is D2. The steps of WEB page classification inference algorithm (WCIA) are showed as the followed.

Step 1, initial node of S_0 is input into determined node table D1 (D1 is realized with stack), if S_0 belongs to typal situation node set S_T, then the classification result is S_0, and the algorithm ends and quits.

Step 2, if the determined node table D1 is empty, then the algorithm fails and quits. Otherwise, goes on.

Step 3, a node S_n pop pinged from D1, as current classification node, is moved to undetermined node set D2.

Step 4, if S_n belongs to other situation node set S_T, then the inference successes and quits.

Step 5, if S_n doesn't belong to object classification node set S_T, and the depth value of the current node is greater than depth threshold N, then turns to Step 2.

Step 6, according to the next node S_n got from the structure of Bayesian network, all sub node of S_n are generated (if S_n has no sub node, then is put into D2, and turns to Step 2), the node same as D1 table is deleted, and put into D2 in turn.

Step 7, feature extraction and classification judgment for current path. If the number of feature vector is less than sample Volume k of Bayesian network, then turns to Step 2. Otherwise, Bayesian judgment formula (1) is called to judge whether $(x_1, x_2, \cdots, x_k)$ belonging to typal object node set S_T or not. If judgment result belongs to typal object node set S_T, then the algorithm quits successfully. Otherwise, recursive call the algorithm. If the recursive call returns successfully, then turns to step 4. If the recursive calls returns unsuccessfully, then pushes n_f into determining table D2, then turns to Step 2.

Path classification is done when the WCIA algorithm searches the depth value of the node. The theory basement of classification inference algorithm is Bayesian judgment theory, which is on the condition of having multi-judgment object, so as to decide to the possible judgment object according to the current inference development.

When the object judgment node is composed of n possible judgment object, which is expressed as $S_T = \{S_{T1}, S_{T1}, \cdots, S_{Tn}\}$, based on the feature attributes are independent each other, and judgment function $L_{ij}(X)$ is constructed by path classification as formula (1).

$$L_{ij}(X) = \frac{p(x_1, x_2, \cdots, x_k | S_{Ti})}{p(x_1, x_2, \cdots, x_k | S_{Tj})} = \frac{\prod_{i=1}^{k} p(x_i | S_{Ti})}{\prod_{i=1}^{l} p(x_i | S_{Tj})} \quad (1)$$

The WCIA algorithm is similar to the process of graph search typal node. During the process of the program running, for the time and space complexity, the influence of stack operation, floating point arithmetic operation and chain table operation etc. are ignored for the complexity of graph search, and the time consuming operation is search node in the process of search mainly. So, the number of all search nodes is as the measure before the algorithm quits. The premise condition is when the time of study is fully more, the structure and condition probability of Bayesian network is finalized. If the complexity of WCIA algorithm is $f(n)$, the sample volume of Bayesian network classification system is n, the inference result of the algorithm after inference is got. ($p_1, p_2, ..., p_m$) is the probable value of every inference. The worst complexity of the algorithm is proved as the followed.

$$f(n) = p_1 \times 1 \times n + p_2 \times 2 \times n + \cdots\cdots + p_m \times m \times n$$
$$= (p_1 \times 1 + p_2 \times 2 + \cdots + p_m \times m) \times n$$

When $p_i \langle = 1$,

$$f(n) <= (1 + 2 + \cdots + m) \times n$$
$$<= \frac{m(m+1)}{2} n$$

End.

From the above proof we can get that the complexity of graph search inference algorithm is $\frac{m(m+1)}{2} n$, which is polynomial complexity, so the complexity of traditional Bayesian network inference is got, and the efficiency of WEB page classification is improved.

V. EXPERIMENT

JAVA, MYSQL are adopted in the WEB page classification system.

The main design idea of the system is the information fusion model. At first, information extracting and pre-proceeding are executed, then the fusion Bayesian network is built, finally the Bayesian network inference algorithm concludes the classification result.

（1）Information extracting

The useless information in the WEB page (such as advertisement) is filtered, and the hidden value information such as label information, multi-media information, key word etc., is extracted. These functions are realized by software.

（2）Data pre-processing

The extracted multi- information is deal with and pre-proceeded by different algorithm, then input into the Bayesian network. These functions are realized by software.

（3）Parameter discretization

Bayesian network needs discretization data. The different kind information needs discretization before input into the system. Different type parameter uses different value way.

Multi-classification parameter, which is same as logic parameter, only in the situation of the parameter value is more than two.

Real number measure:

When the real number measure is continuous, discretization needs done. In the paper, continuous function is fuzzy, continuous real number value is turn into information in Partition. Probability estimate is as membership function. Then Direct Mapping is done according to the value multi-classification variable.

The steps of real number discretization are as followed.

a. Normalization processing. The relevant formula is showed as followed.

$$x' = \frac{x-a}{\sigma}, x' \in [0,1]$$

b. Refer to Figure 3, the corresponding value of continuous variable discretization is checked.

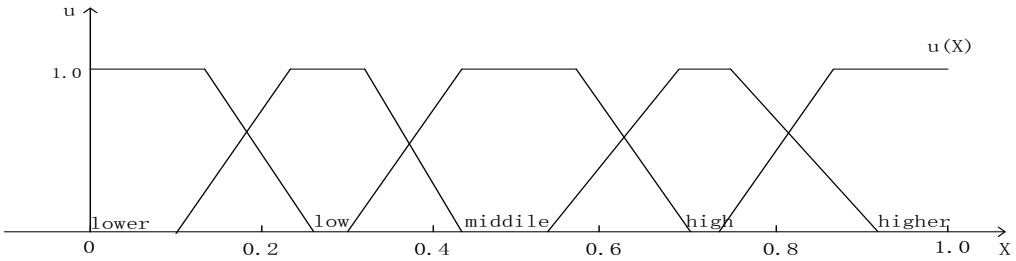


Figure3. Initialization member function

（4）Realization of Bayesian network

In the express of Bayesian network classification, nodes express multi-information and the situation of classification process, the edges express cause relation between nodes. In the Bayesian network design process of WEB page classification, the main mission is determined the meaning of net node and directed edge.

The net node can be divided into two situations, that are the information parameter and the system situation. The directed edge of the net can be divided into the four relations according to the connective nodes.

Parameter- parameter， which expresses the influence of the front to back parameter values.

Parameter-situation, which expresses the influence of the parameter to system situation.

Situation - parameter, which expresses the current situation deciding the parameter value.

Situation–situation, which expresses the transform relation between the situations.

The structure study and parameter study of Bayesian network adopt Simulated Annealing Algorithm and Maximum Likelihood Estimation Algorithm, the improved Bayesian network inference algorithm (WCIA) is as the classification inference algorithm.

（5）Data set, the science WEB pages of NSTL are as the dataset. NSTL (National Science and Technology Library) is as authoritative provider of literature information in our country, the main serve of the literature system is search based on keyword. Now the science WEB resource is bringing into the NSTL system for more service, which includes hotspot science information discovering, science development trend analysis, showing, and science information forecasting, etc.

Climate transformation and energy sources are as the samples in the paper. the WEB site includes England BBC(http://news.bbc.co.uk/weather/hi/climate) ， NOWPublic energy sources (http://www.nowpublic.com/tag/Energy/news) ， England ’Nature’ journal (http://www.nature.com/news/archive/keyword/global+warming.html)， NATION_the globe warming up and climatic change(http://www.thenation.com/section/global-warming-and-climate-change),UPI(http://www.upi.com/Science_News/Resource-Wars/) etc.. The Catalog includes policy, industry, new energy, Low Carbon Economy, climatic change and pollution. The corpuses are the WEB page downloaded from those WEB sites, which are selected and processed, classified into the appointed class partly. The other data are as the test sample.

25000 files are in the dataset, which of every file occupies 1-10 KB at average, about 3000 ten thousand words, which are representative very much, and can be on behalf of corpus setting covering completely.

In the experiment, the ratio of train corpus and test corpus is 2:1. In the classes choose, policy, new energy resources and Low Carbon Economy exit the problem of fuzzy class border, so which is suit to the proof of the fusion model.

The experiment includes the comparisons between the fusion model algorithm and the traditional classification algorithm, one is between the fusion algorithm, the KNN algorithm and SVM algorithm with the same sample, the other is between the fusion algorithm and the KNN algorithm with the same text sample, and between the fusion algorithm and the SVM algorithm with the same image sample. The comparison results are showed as followed.

TABLE 1.
CONTRAST OF THREE ALGORITHMS

Methods\ precise	Recall rate	precision	F1
KNN	80.6%	91.1%	85.5%
SVM	76.9%	92.3%	83.9%
Fusion model	85.4%	95.3%	90.1%

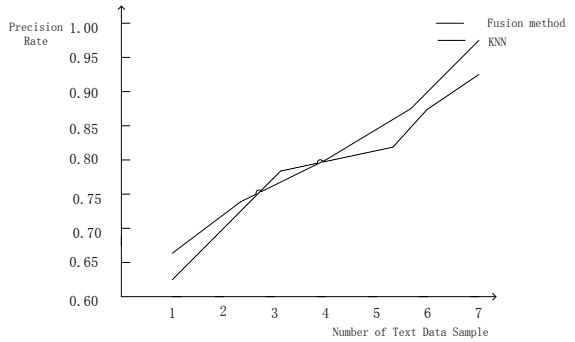


Figure 4. Comparison of the fusion algorithm and KNN algorithm with the same text data set

The Figure 4 is the precision comparison of the fusion algorithm and KNN algorithm with the same text data set. In the figure we can see that the abscissa axis shows the number of data set, the abscissa axis shows the classification precision. The solid curve indicates the growing of the fusion algorithm classification precision, and the dotted curve indicates the growing of the KNN algorithm. From the Figure 4, the precision of fusion algorithm is higher than the KNN. When the numbers of the abscissa are 2.7 and 3.9, the two curves are overlapping. The solid curve is rising reposefully, but the black curve is flexural. And the precision of the two algorithms is higher with the number of the data increasing mainly.

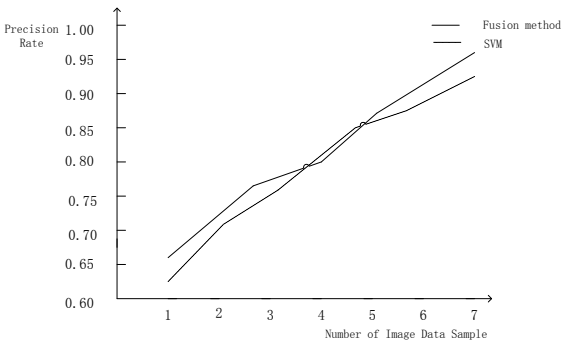


Figure 5. Comparison of the fusion algorithm and SVM algorithm with the same image data set

From Figure 5, we can see, the precision of the fusion algorithm is superior to SVM with the same image date set. The solid curve indicates the growing of the fusion algorithm classification precision, and the dotted curve indicates the growing of the SVM algorithm,

The precision of the fusion algorithm is superior to SVM. When the numbers of the sample are 3.8, 4.1 and 6.7, the two curves are overlapping. With the number of samples is more, the precision of the two algorithms is higher, and the precision of the fusion algorithm is higher than the SVM.

VI. CONCLUSION

In the paper, a classification fusion model and fusion algorithm are proposed to solve the uncertain problem of WEB page classification, and an improved Bayesian network inference algorithm is proposed for solving the NP hard problem. At last, the fusion model and fusion algorithm are realized. The main study content shows as the followed.

Analyzing the uncertain problem in the WEB page classification, and drawing a conclusion that the WEB page classification is a process of solving uncertain problem.

A Bayesian network fusion model and algorithm for uncertain problem are built. The model and the algorithm analyze multi-information in the WEB page, realize the qualitative and quantitative knowledge express of the uncertain information for the uncertain problem, and solve the uncertain problem of WEB classification through the express and the inference algorithm based on the Bayesian network effectively.

To the NP problem in the process of Bayesian network inference, an improved inference algorithm based on graph search theory is proposed, which is proved that the inference time complexity is Polynomial level in the worst condition.

The model and algorithm are realized in the NSTL system. The result expresses the model and algorithm can improve the precision and recall effectively.

In conclusion, the uncertain problem of WEB classification is studied, and the fusion model and algorithm based on Bayesian network is proposed. The

experiment proves that the model and algorithm can solve the uncertain problem in the WEB classification and improve the classification precision and recall rate effectively.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No. 60803050), Project of the 12th five-year-plan scheme of China (No. 2011BAH10B05 and No. 2011BAH10B03).

REFERENCES

[1] Zhang, XD... Text classification based on decision fusion model[C]. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp.266-273). ACM Press, Salvador, Brazil, 2009.

[2] Zhang, B., Chen Y., Fan W., Fox, E. A., Goncalves, M., Cristo, M. & Calado, P.(2006a). Intelligent GP fusion from multiple sources for text classification[C]. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. pp. 477-484). ACM Press, Bremen, Germany.

[3] Zhang Xiao-dan, Zhao Hai. Study on general situation evaluation method of search Strategy[J]. System Simulation journal. 2005.

[4] Suhaila Zainudin. Towards estimating gene network using structure learning[C]. International Conference on Computational Intelligence, 2006:436-441.

[5] Shen, D., J.-T. Sun, Q. Yang, and Z. Chen (2006). A comparison of implicit and explicit links for web page classification[C]. In Proceedings of the 15th International Conference on World Wide Web, New York, NY, pp. 643–650. ACM Press.

[6] Zhang Xiao-dan. A decision layer text classification method[P],china: 2010.12.

[7] Zhang Xiao-dan. A system of file Automated classification[P], 201020200043, China: 2011.5.

[8] Zhang Xiao-dan. A method of file Automated classification[P], 201020200044, China: 2010.12.

[9] Daniel Ball，Optimal decomposition of large scale sensor networks for energy efficient sensing[M], International Conference on Frontiers of Design and Manufacturing(ICFDM'2006), 2006: 167-172

[10] Ouyang Yun， Feature selection integrated Bayesian Network method for remote sensing image classification using spectral and textural information[C] ， 3rd International Symposium on Future Intelligent Earth Observing Satellites，2006：1-3.

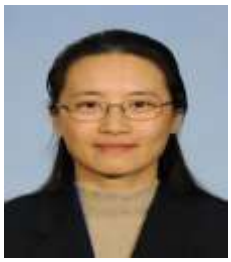
[11] Ruqiang Yan，A Bayesian network approach to energy-aware distributed sensing[C], Sensors, 2004. Proceedings of IEEE，2004：44-48

[12] Zhang, XD...Research on Design & Implementation of Embedded UPS for Web-based Application [J], Journal of Information & Computational Science, 2011. 56-62

[13] Zhang, XD..., Study on WEB page fusion classification model [CA], Advanced Materials Research, 2011: 86-96

[14] Zhang, XD..., Bayesian network model for fault diagnosis of hydropower equipment [J], Dongbei Daxue Xuebao/Journal of Northeastern University, 2006: 559-562

[15] Prasanna Desikan, Mining Temporally Changing Web Usage Graphs[C], Advances in Web Mining and Web Usage Analysis,2006:1-17



Xiaodan Zhang PH.D., College of Information Science and Engineering, Northeastern University. Post doctor, College of Computer, Beijing Institute of Technology. Study fields are information fusion and information mining. Associate professor, Institute of Scientific and Technical Information of China.