http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (5): 796-801, 2008
ISSN 1812-5638

An Ontology Based Approach for Chinese Web Texts Classification

G.Y. Wei, G.X. Wu, Y.Y. Gu and Y. Ling
Zhejiang Gongshang University, Hangzhou, 310018, People's Republic of China

Abstract: The world wide web is a vast resource of information and services that continues to grow rapidly. Developing an automatic classifier, which has ability of classifying documents into appropriate categories predefined in the topic structure based on document contents is a crucial task. Traditional methods of documents classification need characteristic abstraction and classifier training. The work of collecting trainable text terms is laborious and time-consuming. In order to solve the problem, this study proposes an ontology based approach to improve the efficiency and effectiveness of Chinese web documents classification and retrieval. First, the approach establishes an ontology model based on knowledge base. Second, it creates ontology for each subclass of the classification system. It uses RDFS to convert knowledge into ontology and to define the relations among ontology. Finally, web documents classification is performed automatically using the ontology relevance calculating algorithm. Present experiments show that the accuracy of ontology based approach is very close to most classical methods includes Support Vector Machines, K-Nearest Neighbor and Latent Semantic Analysis. Additionally, ontology based algorithm is more stable and robust and can obtain better recalling rate than other three methods.

Key words: Web document, classification, ontology, algorithm

INTRODUCTION

The world wide web is a vast resource of information and services that continues to grow rapidly. It is difficult to find useful information without rich clues. Search engines are currently popular searching tools on the Web and they deal with retrieving documents from a document collection in response to a user's query. Although it is convenient to use search engines for users, it is still hard to formulate queries precisely in many situations. For example, users might not be familiar with the topic of interest; the topic of interest is too vague to formulate as a query; the words used to describe a topic do not appear in the documents of the topic; additionally, synonymy is also a serious problem for a query (Chen *et al*., 2006). Powerful search engines have been developed to aid in locating unfamiliar documents by category, contents, or subject. Unfortunately, queries often return inconsistent results, with document referrals that meet the search criteria but are of no interest to the user. While it may not be currently feasible to extract in the full meaning of a web document, intelligent algorithm and software are needed to extract features from the words or structure of a web document and employ them to classify and categorize the documents.

Today, there are two main approaches for web documents automated classification: supervised learning and rule-based classification. The supervised learning methods require model training. Naive Bayesian (McCallum and Kamal, 1998) and Bayesian network (Domingos and Pazzani, 1997) models are two examples of the supervised learning methods studied. Implementing the learning algorithms is straightforward; however, collecting training records is usually costly and time-consuming. Bayesian network classifiers require fewer training and can achieve better performance than the naive Bayesian classifier. However, unlike most supervised learning methods, the training process was not fully automated. The system must interact with human experts to construct the semantic Bayesian network during the training process. Another major disadvantage of supervised learning methods is the lack of flexibility and generalizability. Rule-based classification methods use a completely different approach and do not require training data. Such methods typically have two stages. In the first stage, documents are translated to an intermediate representation by either a grouping table lookup or keyword matching. Then, grouping consists of running queries (Sniegoski, 2004). In the second stage, a set of rules is used to map the intermediate groups to final categories. A major advantage of rule-based classification methods is their simplicity. The

Corresponding Author: G.Y. Wei, College of Computer Science and Information Engineering, Zhejiang Gongshang University, Xuezheng Street 18#, Hangzhou, Zhejiang, People's Republic of China
Tel: +86-571-2800-8318 Fax: +86-571-2800-8303

classification rules and intermediate groups can be constructed using a top-down approach. The white box nature of these methods makes system maintenance and fine tuning easy for system designers and users. In addition, these methods are flexible: adding new categories or changing definitions can be achieved relatively easily by switching the inference rules. A major problem with the rule-based classification methods is that they cannot handle terms that are not included in the groups.

This study proposes an ontology based approach to improve the efficiency and effectiveness of web documents classification and retrieval. We establish an ontology model based on knowledge base and use RDFs convert knowledge into ontology and define the relations among ontology. Web documents classification is performed automatically using the ontology relevance calculating algorithm.

RELATED WORKS

A wide range of statistical and machine learning techniques has been applied to text categorization, including multivariate regression models, nearest neighbor classifiers (Tahir and Smith, 2006), probabilistic Bayesian models (Lam and Low, 1997), decision trees, neural networks (Zhang, 2000), symbolic rule learning (Ehrig and Maedche, 2003) and SVMs (Support Vector Machines) (Harris *et al*., 1999; Burges, 1998). These approaches all depend on having some initial labeled training data from which category models are learned. Once category models are trained, new terms can be added with little or no additional human effort. Among them, the most classical approaches for Chinese text classification are SVMs, KNN (K-Nearest Neighbor) (Li *et al*., 2004) and LSA (Latent Semantic Analysis) (Kan, 2004).

SVMs are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. They can also be considered a special case of Tikhonov regularization. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers. Viewing the input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, we construct two parallel hyperplanes, one on each side of the separating one, which are pushed up against the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes. The hope is that, the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the classifier will be.

The KNN algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

LSA is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA can use a term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to terms and whose columns correspond to documents, typically stemmed words that appear in the documents. A typical example of the weighting of the elements of the matrix is TF-IDF (term frequency-inverse document frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up-weighted to reflect their relative importance. This matrix is also common to standard semantic models, though it is not necessarily explicitly expressed as a matrix, since the mathematical properties of

matrices are not always used. LSA transforms the occurrence matrix into a relation between the terms and some concepts and a relation between those concepts and the documents. Thus the terms and documents are now indirectly related through the concepts.

KNN and SVM have been reported as the top performing methods for English text classification (Ehrig and Maedche, 2003). To solve information heterogeneity problems, (Ngamnij, 2003) proposed a metadata dictionary as an assistant mechanism for solving semantic heterogeneity based on domain ontology. It introduced an XML-based data model to manipulate and express the metadata dictionary contents (Martin and Ecklund, 1999) uses general and intuitive knowledge representation languages for indexing the content of web documents and representing knowledge within them. The retrieval of precise information is supported by languages designed to represent semantic content. The use of Conceptual Graphs and simpler notational variants that enhance knowledge readability is advocated.

ONTOLOGY BASED CLASSIFICATION

Knowledge base system: In this study, HowNet (Dong and Dong, 2003) is used as a knowledge base to construct ontology. It is an electronic world knowledge system. And it can represent inter-concept relations and inter-attribute relations of the concepts as connoting in lexicons of the Chinese and their English equivalents. Its applied effectiveness has been world-widely acknowledged. It defines knowledge as a system encompassing the varied relations amongst concepts or attributes of concepts. The relation includes: Hypernym-Hyponym, Synonym, Part-Whole, material-product, attribute-host, etc. HowNet constructs a graph structure of its knowledge base from the inter-concept relations and inter-attribute relations, which is the fundamental difference between HowNet and other tree-structure lexical databases. The design of HowNet is based on its ontological view of the objective world. All physical and non-physical matters undergo a continual process of motions and changes in a specific space and time. The motions and changes are usually reflected by a change in state that in turn, is manifested by a change in the value of some attributes. The way we understand attribute is that any object necessarily carries a set of attributes. Similarities and differences between objects are determined by the attributes they each carry. There can be no objects without attributes. For instance, human beings are attached with natural attributes such as ethnic group, color, gender, age, ability to think, ability to use language as well as social attributes such as nationality, class origin, job, wealth, etc.

Transform knowledge into ontology: Ontology is a set of concepts and their relations. In this study, ontology is transformed from HowNet knowledge base and is represented using RDF and RDFs format.

- Ontology definition

$$\text{Ontology } O = (\text{meta\_info, Concept, Relation, Rule}) \quad (1)$$

In formula (1), meta_info represents the Meta data of O. It includes the name of O, creator, date, etc. Concept represents the set of concepts. Relation represents the set of relations. Rule represents the set of rules.

- Ontology representation

As a general-purpose knowledge representation tool, RDFs provides a syntactic model and semantic structure to define machine executable ontology or metadata schemas. Additionally, its model supports structures inter-operation across heterogeneous resource communities. In this study, RDFs is used to transform HowNet into ontology. The words of HowNet are transformed into the classes of ontology. The Hypernym-Hyponym relation of HowNet is transformed into SubClassOf expression. The attributes of classes is represented with relations including Synonym, Part-Whole and others. Partial codes for defining the relation of ontology using RDFS is following.

```
1: <rdf: RDF
2: xmlns="http://localhost:8080/MyOntology#"
3: xmlns:OT="andOT;"
4: xmlns:rdf="andrdf;"
5: xmlns:rdfs="andrdfs;"
6: xmlns:a="anda;"
7: >
8: <rdf:Property rdf:ID="PartOf"
9: OT:comment="This is a relation that specifies that the
first Concept is a part of the second Concept">
10: </rdf:Property>
11: <rdfs:Class rdf:ID="10"
12: <rfds:subClassOf rdf:resource="#31" >
13: <a: PartOf rdf:resource="#15" >
14: </rdfs:Class>
```

CLASSIFICATION ALGORITHM

Create ontology: Before create ontology, categories must be defined. The existed categories can be used, for example, the directory of Yahoo. The process of creating ontology for every category includes three steps. The ontology creating algorithm is described as Algorithm 1.

Algorithm 1 ontology create

1: Input: dict_Ont and C
2: Output: T_Ont
3: Step 1: Search C or synonyms of C in dict_Ont;
4: Step 2: Transform dict_Ont to a graph by let edges represent Relations of dict_Ont and vertices represent Concepts of dict_Ont.
5: Step 3: Set C and synonyms of C as the center of the graph, then search N nodes neighbor to it. Set these N nodes and their relations as T_Ont.
In the Algorithm 1, dict_On denotes the ontology transformed from Hownet. C denotes a category. T_Ont denotes the ontology of category C.

Calculate relevance: After ontology of category is created, the next work is to calculate the relevance of the ontology. In the ontology model, the relevance is used to measure the relevant degree between the nodes in T_Ont and C. we use relevance score to denote the relevant degree. The Fig. 1 showed an instance of T_Ont.

In Fig. 1, Center denotes C and C's synonyms. C, locates in a relation network, is not an isolated concept in the ontology model. In this relation network, multiple concepts (class) connect to C, but their relevance score are different. If concept C_1 and C have the relation of Synonym and concept C_2 and C have the relation of PartWhole, C_1's relevance score is bigger than C_2. So, the relations affect the relevance score. In order to calculate the relevance score, we divide the relations into four types and use R_1, R_2, R_3 R_4 to denote them. R is Synonym and Instance Of. If the weight of R_1 is w_{R_1}, then w_{R_1} is used to weigh the relevant degree between two concepts which have the relation of R_1. R_2 is Part-Whole relation and its weight is assigned to value w_{R_2}. R_3 is Relation to relation and its weight is assigned to value w_{R_3}. R_4 is subClass Of and others relations, its weight is assigned to value w_{R_4}. It supposes that the expression $1>w_{R_1}, w_{R_3}, w_{R_4}$ is granted. The distance between the concepts and C also affect the relevance score. If the distance is longer, the relevance score is smaller. With consideration of the relations and distance, we can use formula (2) to calculate the relevance score of every node in T_Ont.

$$\text{Sim}(t,C) = \frac{\alpha}{\text{W length}(t,C) + \alpha} \quad (2)$$

where, function Slim(t, C) is defined to compute the relevance score. W length (t, C) is a function computing shortest distance between t and C. α represents a parameter that can be used to adjust relevance score. Function W length (t, C) takes the relations as undirected edges and set a value L for every edge. The parameter L represents the distance between two connected nodes, $L_{r_i} = 1-w_{R_i}$. It indicates if two connected nodes have a relation R_i, the distance L_{R_i} between them is $1-w_{R_i}$.

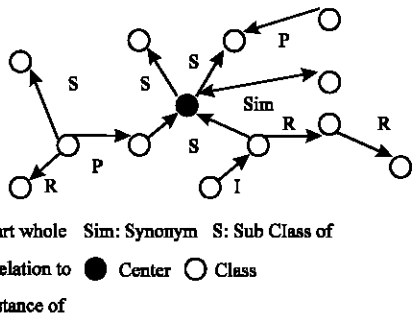

Fig. 1: An instance of T_Ont

The ontology based classification algorithm is described as Algorithm 2.

Algorithm 2 textclassify

1: Input: ontology of every category and W_T which is a new text.
2: Output: which category do *W_T* belong to?
3: int Classify(W_T)
4: {
5: double max_sim=0;
6: int index= -1;
7: for (i=0; i<amount_of_category; i++)
8: {
9: Concept_Set=Search_concept(i,W_T);
10: int number[N];
11: number = Concept_frequency(Concept_Set, W_T);
12: $\text{Sim}(W\_T,i) = \sum_{c \in W\_T \cap O} f_{cW\_T} w_{cO}$
13: if (Sim(W_T, i)> max_sim)
14: {
15: max_sim =Sim(W_T, i);
16: index = i;
17: }
18: }
19: return index;
20: }

EXPERIMENTS

To evaluate the performance of ontology based classification algorithm, we use same test data set do experiments with SVM, KNN, LSA (TF-IDF) and ontology based approach separately. The training text terms are taken from CNLP platform (http://www.nlp.org.cn/). There are ten categories as shown in Table 1. We use 1380

Table 1: Documents number of each category

Category name		Document No.
政治	Politics	250
环境	Environment	100
计算机	Computer	100
交通	Traffic	100
教育	Education	110
经济	Economy	160
军事	Military	120
体育	Gym	220
医药	Medicine	100
艺术	Art	120
Total		1380

Table 2: Comparison of precision rate (SVM, KNN, LSA and ontology based)

Category name		SVM	KNN	LSA (TF-IDF)	Ontology based approach
政治	Politics	0.55	0.64	0.65	0.63
环境	Environment	0.75	0.70	0.86	0.95
计算机	Computer	0.96	0.98	0.96	0.82
交通	Traffic	0.90	0.95	0.94	0.97
教育	Education	0.88	0.90	0.92	0.77
经济	Economy	0.60	0.51	0.80	0.77
军事	Military	0.86	0.79	0.63	0.80
体育	Gym	0.78	0.92	0.81	0.84
医药	Medicine	0.91	0.96	0.85	0.81
艺术	Art	0.82	0.85	0.82	0.83
Average precision rate		0.801	0.820	0.824	0.819

texts for training and 1380 texts to test. We use precision rate and recall rate to measure the performances of these methods. Recall rate and Precision rate for a category are defined as formula (3) and (4) separately.

$$\text{Recall rate} = \alpha/\beta \quad (3)$$

$$\text{Precision rate} = \alpha/\gamma \quad (4)$$

where, α is the number of documents which are classified into category correctly; β is the number of documents of category in the testing data and γ_c is the number of documents which are classified into category.

The experiments results of SVM, KNN and LSA (TF-IDF) approaches are shown in Table 2 and 3. As Table 1 shows SVM approach gets average precision rate 80.1% and the average recall rate 68.3%; 2) KNN approach gets average precision rate 82% and the average recall rate 69.1% and Table 3 shows LSA(TF-IDF) approach gets average precision rate 82.4% and the average recall rate 73.8%.

The experiment results of ontology based approach are also shown in Table 2 and 3. The initial weight of R_1, R_2, R_3, R_4 is 0.9, 0.8, 0.7, 0.6 and the ontology of every category includes 350 concepts (class). The results of experiment of average precision rate 81.9% is and average recall rate is 75.8%.

Table 3: Comparison of recall rate (SVM, KNN, LSA and ontology based)

Category name		SVM	KNN	LSA (TF-IDF)	Ontology based approach
政治	Politics	0.92	0.97	0.95	0.89
环境	Environment	0.50	0.47	0.52	0.57
计算机	Computer	0.60	0.52	0.75	0.89
交通	Traffic	0.75	0.82	0.80	0.70
教育	Education	0.79	0.76	0.85	0.96
经济	Economy	0.60	0.86	0.57	0.53
军事	Military	0.51	0.37	0.50	0.44
体育	Gym	0.81	0.91	0.83	0.98
医药	Medicine	0.65	0.48	0.81	0.79
艺术	Art	0.70	0.75	0.80	0.83
Average recall rate		0.683	0.691	0.738	0.758

PERFORMANCE EVALUATION

Comparing with other three approaches, as shown in Table 2, we can find out that precision rate is improved by ontology based method. In category 2 environment and category 4 traffic, ontology based method achieves higher precision rate than others methods. In category 3 computer, category 5 education and category 9 medicine, ontology based method obtains lowest precision rate. In others categories, its precision rate is close to the average of other three methods. As illustrated in Table 2, ontology based method achieve highest average precision rate among above four methods and its precision rate is most stable.

As shown in Table 3, we can also find out that recall rate improved by ontology based method. In category 2, 3, 5, 8 and 10, ontology based method achieves highest recall rate. In others categories, ontology based method obtains a little lower recall rate than (very close to) other three methods. As shown in Table 3, ontology based method can achieve highest average recall rate among above four methods.

CONCLUSION AND FUTURE WORK

In this study, we propose an ontology based approach to classify web documents. It supports text retrieval by keywords automatically without training texts. The approach constructs ontology based a Chinese knowledge base named as HowNet and creates ontology for each subclass of the classification system. It uses RDFS to transform knowledge into ontology and represent the relations among ontology. An ontology relevance calculating algorithm can classify web documents automatically. Prest experiments show that the accuracy of ontology based approach is very close to classical methods includes Support Vector Machines, K-Nearest Neighbor and Latent Semantic Analysis (include TF-IDF). Comparably, Prest algorithms are more stable and robust and can obtain better recalling rate

than above three methods. According to Chinese character documents, the performance of Prest ontology based algorithms is better than other three algorithms.

In the future, we plan to design a relevant feedback method, which will be used to maintain and training ontology with experts' domain knowledge automatically. In addition, we plan to compare the performance of document retrieval with more others approaches.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation of China under Grants 60673179, Natural Science Foundation of Zhejiang Province of China under Grant number Z106727, Grand Science Project of Zhejiang Province of China under Grant number 2007C13068 and Science Fund for Distinguished Young Scholars of Zhejiang Gongshang University under Grant numbers Q07-03.

REFERENCES

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., 2: 121-167.

Chen, C.M., H.M. Lee and C.C. Tan, 2006. An intelligent web-page classifier with fair feature-subset selection. Eng. Applied Artif. Int., 19: 967-978.

Domingos, P. and M. Pazzani, 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn., 29: 103-107.

Dong, Z.D. and Q. Dong, 2003. HowNet-a hybrid language and knowledge resource. Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, October 26-29, pp: 820 824-10.1109/NLPKE.2003.1276017.

Ehrig, M. and A. Maedche, 2003. Ontology-focused crawling of Web documents. Proceedings of the ACM Symposium on Applied Computing, March 9-12, Melbourne, Florida. ACM, New York, pp: 1174-1178.

Harris, D., D.H. Wu and N.V. Vladimir, 1999. Support vector machines for spam categorization. IEEE. Trans. Neural Network, 10: 1048-1054.

Kan, M.Y., 2004. Web page classification without the web page. Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers and Posters Table of Contents. May 17-20, ACM, New York, USA., pp: 262-263.

Lam, W. and K.F. Low, 1997. Automatic document classification based on probabilistic reasoning: Model and performance analysis. Proceedings of IEEE International Conference on Systems, Man and Cybernetics, October 12-15, Orlando, FL, USA., pp: 2719-2723.

Li, B.L., Q. Lu and S.W. Yu, 2004. An adaptive k-nearest neighbor text categorization strategy. ACM Trans. Asian Lang. Inform. Process., 3: 215-226.

Martin, P. and P. Eklund, 1999. Embedding knowledge in Web documents. Comput. Network, 31: 1403-1419.

McCallum, A. and N. Kamal, 1998. A comparison of event models for naive bayes text classification. Proceedings of AAAI-98 Workshop on Learning for Text Categorization. July 26-27, AAAI Press, Madison, Wisconsin, USA., pp: 41-48.

Ngamnij, A., 2003. A semantic information gathering approach for heterogeneous information sources on WWW. J. Inform. Sci., 29: 357-374.

Sniegoski, C., 2004. Automated syndromic classification of chief complaint records. J. Hopkins APL Tech. Digest, 25: 68-74.

Tahir, M.A. and J. Smith, 2006. Improving nearest neighbor classifier using tabu search and ensemble distance metrics. Proceedings of 6th International Conference on Data Mining (ICDM). December 18-22, IEEE Computer Society, Hong Kong, China, pp: 1086-1090.

Zhang, G.P., 2000. Neural networks for classification: A survey. IEEE. Trans. Syst. Man CY. C., 30: 451-462.