Web Document Classification Based on Rough Set

Qiguo Duan, Duoqian Miao, and Min Chen

Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
The Key Laboratory of "Embedded System and Service Computing", Ministry of Education, Shanghai 201804, China
dqgcn@126.com, miaoduoqian@163.com, tomatocm@163.com

Abstract. For traditional way of Web document representation in Vector Space Model, zero-valued similarity problem between vectors occurs frequently, which decreases classificatory quality when defining the relation between Web documents. In this paper, a novel Web document representation and classification approach based on rough set is proposed. Firstly, TF*IDF weighting scheme is used to assign weight values for Web document's vector. The weights of those terms which do not occur in a Web document are considered missing information. Then rough set for incomplete information is introduced to supplement loss and expand Web document representation. Through generating tolerance classes in both term space and Web document space, the missing information of Web document can be complemented by incorporating the corresponding weights of terms in tolerance classes, which extends the essential information to Web document. Finally, Web document classification algorithm is implemented. Experimental results show that the performance of the classification is greatly improved.

Keywords: Rough sets, Web document classification, Web mining.

1 Introduction

With the rapid growth of information on the World Wide Web, automatic classification of Web documents has become important for effective retrieval. As one of the essential techniques for Web mining, Web document classification has been studied extensively [1], [2], [3]. Nowadays, many Web document classification methods are based on the Vector Space Model (VSM), which is a widely used data model for text mining. In VSM, a Web document is represented as a term vector. Term weights, contained in each term vector, are assigned by weighting schemes. Traditionally, the weights of those terms which do not occur in the Web document are assigned zero value. A single Web document is usually represented by relatively few terms, thereby the Web document vector is characteristic of high dimension and sparseness, which results in zero-valued similarity between vectors. This problem would decrease classificatory quality because the relation

A. An et al. (Eds.): RSFDGrC 2007, LNAI 4482, pp. 240–247, 2007.

between Web documents is defined by measuring distance of the corresponding vectors.

In this paper, a rough set approach to Web document representation and classification is proposed. Instead of assigning zero to the weights of those terms are absent in a Web document, these weights are considered missing information. Thus Web document is represented as an incomplete term vector firstly. Then through generating tolerance classes in both term space and Web document space, the missing information of Web page can be complemented by incorporating the corresponding weights of terms in tolerance classes. Only using a little heuristics knowledge, the zero-valued similarity problem can be avoided through complementing the potential missing information and therefore the classification performance can be improved.

The rest of the paper is organized as follows. Section 2 describes the weighting scheme briefly. Section 3 introduces the extended rough set for incomplete information. Section 4 presents the novel approach to Web document representation and classification in detail. Section 5 reports and discusses the experimental results and section 6 concludes the paper.

2 Weighting Scheme

In VSM, each Web document is viewed as a bag of terms and represented by a term vector. In this paper, we apply the popular TF*IDF (Term Frequency times Inverse Document Frequency) weighting scheme to assign weight values for Web document's vector. The standard TF*IDF is defined as follows:

$$w_{ij} = tf_{ij} \times \log(N/df_i) \ . \tag{1}$$

where tf_{ij} is the frequency of the term t_i in Web document d_j; df_i is number of Web documents in which term t_i occurs; N is the total number of Web documents. Normalization by vector's length is applied to all vectors:

$$w^*_{ij} = w_{ij} / \sqrt{\sum_{t_k \in d_i} (w_{ik})^2} \ . \tag{2}$$

Assume that there are N Web documents and n different terms in a set of Web document. Using TF*IDF, each Web document is represented by an n-dimensional term vector. The N Web documents in the set can be represented by an $N \times n$ matrix, $DW = [w^{'}_{ij}]$, where $w^{'}_{ij} = w^*_{ij}$, if the term t_j occurs in the Web document d_j; otherwise, $w^{'}_{ij} = 0$. Together with decision attributes, i.e., the class label of Web documents, the matrix can be considered as a decision table. According to the weight computation, if the term t_j is absent in the Web document d_i, $w^{'}_{ij}$ is equal to zero. This way of assigning the weights to absent terms brings zero-valued similarity problem between vectors. In this paper, as an extended rough set, tolerance rough set is preferred to avoid zero-valued similarity through complementing the incomplete information of Web documents.

3 Extended Rough Set for Incomplete Information

Rough set theory, introduced by Pawlak, is a formal mathematical tool to deal with incomplete or imprecise information [4]. It has been successful in many applications [9] [10]. The classical rough set theory is based on equivalence relation that divides the universe of objects into disjoint classes [4]. By relaxing the equivalence relation to a tolerance relation, where transitivity property is not required, a generalized tolerance space is introduced below [5], [6], [7], [8].

Let $S = (U, A, V, f)$ be an information system, where U is a nonempty finite set of objects called universe of discourse, A is a nonempty finite set of conditional attributes; and for every a$\in A$, such that $f: U \rightarrow V_a$, where V_a is called the value set of attribute a.

Definition 1. *If some of the precise attribute values in an information system are not known, i.e., missing or known partially, then such a system is called an incomplete information system. Otherwise the system is called a complete information system.*

Definition 2. *Let $S = (U, A, V, f)$ be an incomplete information system and the sign * denote null value, a tolerance relation T is defined as:*

$$T(B) = \{(x, y) \in U \times U | \forall b \in B, b(x) = b(y) \vee b(x) = * \vee b(y) = *\} . \quad (3)$$

where $B \subseteq A$. Obviously, T is reflexive and symmetric, but not transferable. Let $I_B(x) = \{y \in U | (x, y) \in T(B)\}$, and then $I_B(x)$ is called the tolerance class of the object x with respect to the set $B \subseteq A$.

Definition 3. *Let $S = (U, A, V, f)$ be an incomplete information system, $X \subseteq U$, $B \subseteq A$, the upper approximation and lower approximation of X with regard to attribute set B under the tolerance relation T can be defined as:*

$$U_B(X) = \{x \in U \,|I_B(x) \cap X \neq \emptyset\} . \quad (4)$$

$$L_B(X) = \{x \in U \,|I_B(x) \subseteq X\} . \quad (5)$$

4 Web Document Representation and Classification

4.1 Web Document Representation

According to Section 3, here we introduce the corresponding concepts in the Web document classification domain.

An incomplete information system for a web page set is represented as $WS = (U, TS \cup \{class\}, f)$, where U is the set of Web documents, each Web document is an object $d \in U$; TS is the set of total terms which occur in the Web document set, *class* is the decision attribute, i.e., the class label of the Web documents. The weights of those terms which do not occur in a Web document are considered missing information and denoted by sign * instead of zero.

In Web document space, the tolerance relation and tolerance class of Web document are defined as:

Definition 4. *For a subset of TS, $B \subseteq TS$, a tolerance relation T(B) on U is defined as:*

$$T(B) = \{(d_x, d_y) \in U \times U | \forall b \in B, |b(d_x) - b(d_y)| \leq \delta \vee b(d_x) = * \vee b(d_y) = *\} . \tag{6}$$

Because weights are real values, the requirement $b(d_x) = b(d_y)$ is too strict. Here it is replaced with $|b(d_x) - b(d_y)| \leq \delta$, where $\delta \in [0, 1]$. Consequently, tolerance class of a Web document d_x with respect to $B \subseteq TS$, $I_B(d_x)$, is the set of Web documents which are indiscernible to d_x, i.e., $I_B(d_x) = \{d_y \in U | (d_x, d_y) \in T(B)\}$.

On the other hand, correlation between terms is valuable for complementing missing information. Thus, the tolerance class of term is also defined in term space. Let $U = \{d_1, \ldots, d_M\}$ be a set of Web documents and $TS = \{t_1, ..., t_N\}$ set of terms for U. The tolerance space of term is defined over a universe of all terms for U.

Definition 5. *Let $f_U(t_i, t_j)$ denotes the number of Web documents in U in which both terms t_i and t_j occurs. The uncertainty function I with regards to co-occurrence threshold θ defined as:*

$$I_\theta(t_i) = \{t_j | f_U(t_i, t_j) \geqslant \theta\} \cup \{t_i\} . \tag{7}$$

Clearly, the above function satisfies conditions of being reflexive: $t_i \in I_\theta(t_j)$ and symmetric: $t_j \in I_\theta(t_i) \Longleftrightarrow t_i \in I_\theta(t_j)$ for any $t_i, t_j \in T$. Thus, $I_\theta(t_i)$ is the tolerance class of term t_i. Tolerance class of terms is generated to capture conceptually related terms into classes. The degree of correlation of terms in tolerance classes can be controlled by varying the threshold θ.

In tolerance space of term, an expanded representation of Web document can be acquired by representing Web document as set of tolerance classes of terms it contains. This can be achieved by simply representing Web document with its upper approximation, e.g., the Web document $d_i \in U$ is represented by:

$$U_R(d_i) = \{t_i \in T | I_\theta(t_i) \cap d_i \neq \emptyset\} . \tag{8}$$

This approach to Web document representation expands Web document because it takes into consideration not only terms actually occurring Web document but also other related terms with similar meanings.

4.2 Missing Weights Complement

The best values of these missing weights are determined by incorporating two parts, i.e., weights of terms in term's tolerance class and corresponding term weight of the most similar vector, which has the same class label in tolerance class of the Web document. Here, the similarity measure between vectors is computed based on the distance:

$$Sim(d_x, d_y) = \frac{1}{1 + \sum_{k=1}^{M} |w_{ik} - w_{jk}|} . \tag{9}$$

After the tolerance classes for both term and Web document are generated, the essential information (i.e., the similarity between Web documents and the correlation between terms) is identified. To complement missing weights of terms in the Web document's vectors, we produce an improved TF*IDF weighting scheme based on the traditional TF*IDF. The improved weighting scheme is defined as below.

$$w_{ij} = \begin{cases} 1 + \log(f_{d_i}(t_j)) \times \log \frac{N}{f_D(t_j)} & \texttt{if } t_j \in d_i, \\ \alpha \times w_{kj} & \texttt{if } t_j \notin U_R(d_i), \\ \alpha \times w_{kj} + \beta \times (\min_{t_n \in d_i \wedge t_n \in I_\theta(t_j)} w_{in}) & \texttt{if } t_j \in U_R(d_i) \wedge t_j \notin d_i. \end{cases} \tag{10}$$

In above formula, w_{kj} is the weight value of corresponding term of the most similarity vector with the same class label in Web document tolerance class; α, $\beta \in [0, 1]$, they adjust the relative impact of relevant terms and Web documents respectively. Here, let parameters α and β be 0.2.

To demonstrate the use of the improved TF*IDF weighting scheme, we detail an example as follows.

Example: Let Web document set be $U=\{d_1,d_2,d_3,d_4,d_5,d_6,d_7\}$, term set be $TS=\{t_1,t_2,t_3,t_4,t_5\}$, $B=TS$, the class label set be $class=\{C1,C2\}$, the frequency data is listed in Table 1.

Table 1. Sample Web document-term frequency array

	t_1	t_2	t_3	t_4	t_5	Class
d_1	0	0	6	8	0	*C1*
d_2	1	3	12	0	9	*C1*
d_3	2	3	0	12	14	*C1*
d_4	0	0	5	4	2	*C2*
d_5	10	9	4	0	3	*C2*
d_6	12	14	2	2	0	*C2*
d_7	11	12	0	4	2	*C2*

Let co-occurrence threshold θ equal 4, tolerance class of each term t_i (i=1,2, ...,5) and upper approximations of the Web document d_j (j=1,2,...,7) can be computed as below:

$I_\theta(t_1) = I_\theta(t_2) = \{t_1, t_2\}$; $I_\theta(t_3) = \{t_3\}$; $I_\theta(t_4) = I_\theta(t_4) = \{t_4, t_5\}$.

$U_B(d_1) = U_B(d_4) = \{t_3, t_4, t_5\}$; $U_B(d_2) = \{t_1, t_2, t_3\}$; $U_B(d_3) = U_B(d_7) = \{t_1, t_2, t_4, t_5\}$; $U_B(d_5) = U_B(d_6) = \{t_1, t_2, t_3, t_4, t_5\}$.

Note that the Web document d_1 and d_4 have different class label. We weigh them with traditional TF*IDF and improved TF*IDF respectively, result is listed in Table 2.

4.3 Web Document Classification

Firstly, terms are extracted from training set of Web documents, and then tolerance classes of Web documents and terms are computed. Secondly, the missing

Table 2. Weight of normal TF*IDF versus of improved TF*IDF

Traditional TF*IDF			Improved TF*IDF		
term	d_1	d_4	term	d_1	d_4
t_1	0	0	t_1	0.067	0.115
t_2	0	0	t_2	0.072	0.118
t_3	0.684	0.636	t_3	0.684	0.636
t_4	0.731	0.600	t_4	0.731	0.600
t_5	0	0.487	t_5	0.091	0.487

weights of incomplete vectors are complemented. Thirdly, the classifier is constructed. Finally, the new Web document is classified into the category where the similarity measure is the highest among all other categories.

The similarities are computed between the new Web document and each category centroid, in which the similarity formula is defined as follows:

$$Dis(d_i, c_j) = \frac{\sum_{k=1}^{M} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{M} w_{ik}^2) \times (\sum_{k=1}^{M} w_{jk}^2)}} . \tag{11}$$

where d_i is the new Web document, c_j is the jth category centroid, M is the term dimension.

5 Experimental Evaluation

5.1 Experimental Data Sets

To evaluate the proposed approach, we use two popular data collections in our experiments. The first one is the WebKB data set [1], which contains 8282 Web documents collected from computer science departments of various universities. The pages were manually classified into the following categories: student, faculty, staff, department, course, project, other (respectively abbreviated here as St, Fa, Sta, De, Co, Pr, Ot). In our experiments, each category is employed.The second collection is the Reuters-21578 [2], which has 21578 documents collected from the Reuters newswire. Of the 135 categories, only the most populous eight categories are used, i.e, acq, corn, crude, earn, grain, interest, money and trade (respectively abbreviated here as Ac, Co, Cr, Ea, Gr, In, Mo, Tr). The construction of each data set for our experiments is done as follows: Firstly, we randomly select 10% of the Web documents from the each category, and put them into test set to evaluate the performance of classifier. Then, the rest are used to create training sets. We extract and select the 100 most frequently occurred keywords from each category. For WebKB data set and Reuters-21578, the total numbers of all distinct keywords are 463 and 689 respectively.

[1] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[2] http://www.research.att.com/ lewis/reuters21578.html

5.2 Performance Measures

To analyze the performance of classification, we adopt the popular F1 measure. F1 measure is combination of *recall* (re) and *precision* (pr), F1=$2.re.pr/(re+pr)$. *Precision* means the rate of documents classified correctly among the result of classifier and *recall* signifies the rate of correct classified documents among them to be classified correctly. The F1 measure which is the harmonic mean of precision and recall is used in this study since it takes into account effects of both quantities.

5.3 Experimental Results and Discussion

The results on WebKB data set are summarized in Table 3. Our approach yields a higher performance compared to the normal VSM for all categories. For example, in student category, our approach yields the F1 values of 75.6%, whereas the normal VSM yields the F1 values of 67.1%.

Table 3. Comparison of classification performance on WebKB

	St	Fa	Sta	De	Co	Pr	Ot	Avg
VSM	0.671	0.613	0.437	0.468	0.635	0.554	0.725	0.586
RS	0.756	0.734	0.633	0.630	0.691	0.712	0.787	0.710

Table 4. Comparison of classification performance on Reuters-21578

	Ac	Co	Cr	Ea	Gr	In	Mo	Tr	Avg
VSM	0.710	0.575	0.644	0.723	0.681	0.637	0.625	0.612	0.651
RS	0.736	0.673	0.727	0.780	0.769	0.740	0.768	0.694	0.736

In Table 3, *avg* shows summarized result which is calculated by averaging the F1 values over all categories. Our approach yields higher average classification performance of 12.4% over the normal VSM.We perform the same experiments on the Reuters-21578. The results are shown in Table 4, in which *avg* also shows summarized result. Our approach yields higher average classification performance of 8.5% over the normal VSM for Reuters-21578.

6 Conclusion

In this paper, a novel approach to Web document representation and classification based on rough set is proposed. For traditional way of Web document representation in the VSM, zero-valued similarity between vectors would decrease classificatory quality. Instead of assigning zero to the weights of those terms are absent in a Web page, these weights are considered missing information. Rough set for incomplete information is applied to discover valuable information, i.e., indiscernibility between Web documents and correlation between terms. Then,

the information is used for expanding representation of Web document to avoid zero-valued similarity. To validate the proposed approach, we compared our approach with the VSM. The experimental results show that the proposed approach yields a considerable improvement of classification performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.60475019) and the Ph.D. programs Foundation of Ministry of Education of China (No.20060247039).

References

1. Michelangelo Ceci, Donato Malerba: Hierarchical Classification of HTML Documents with WebClassII. F. Sebastiani (Ed.): ECIR 2003, LNCS 2633, pages 57-72, 2003.
2. Lawrence Kai Shih, David R. Karger: Using URLs and Table Layout for Web Classification Tasks. WWW2004, pages 193-202, 2004.
3. Sebastiani,F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1-47, March 2002.
4. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic, Dordrecht (1991)
5. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae 27, pages 245-253, 1996.
6. Kryszkiewicz, M.: Rough set approach to incomplete information system. Information Sciences, 112:39-49, 1998.
7. Tu Bao Ho, Ngoc Binh Nguyen: Nonhierarchical Document Clustering based on A Tolerance Tough Set Model. International Journal of Intelligent Systems, Vol. 17, pages 199-212, 2002.
8. Chi Lang Ngo, Hung Son Nguyen: A Tolerance Rough Set Approach to Clustering Web Search Results. In: J.-F. Boulicaut et al. (eds.): PKDD 2004. Springer-Verlag, Berlin Heidelberg, pages 515-517, 2004.
9. D.Q. Miao, L.S. Hou: A comparison of rough set methods and representative inductive learning algorithms, Fundamenta Informaticae, v 59, n 2-3, pages 203-219, 2004.
10. Y.Y.Yao, C.-J.Liau, N.Zhong: Granular computing based on rough sets, quotient space theory, and belief functions. Proceedings of ISMIS03, pages 152-159, 2003.