A New Method for Conceptual Classification of Multi-label Texts in Web Mining Based on Ontology

Mahnaz Khani, Hamid Reza Naji*, and Mohammad Malakooti

Department of Computer Engineering,
Islamic Azad University
Dubai, UAE
mahnaz_khani2000@yahoo.com, hamidnaji@ieee.org,
malakooti@iau.ae

Abstract. This paper presents a new inductive learning method for conceptual classification of multi-label texts in web mining based on ontology through Term Space Reduction (TSR) and through using mutual information measure. Laboratory results show the presented method has high precision in compare to existing methods of SVM, Find Similar, Naïve Bayes Nets, and Decision Trees. It should be noted that break–even point is used in micro–averaging for appropriate classification of data complex entitled "Reuters–21578 Apte Split".

Keywords: Ontology, TSR, Conceptual Classification, Web Mining.

1 Introduction

Primarily, web pages show textual data with no semantic interpretation adaptability. Therefore, processing according to keyword-based methods has been turned into one of major problems of web. Working with websites will turn much more difficult without appropriate semantic knowledge on them i.e. websites. Vivid and clear-cut data semantic display is associated by theories of domain (for example, ontology). Using ontology is considered as one of main methods in semantic web. Recently, ontology has been turned as one of the most important subjects in knowledge, management and e-commerce engineering. It should be noted that ontology is pillar of knowledge which provides official display of specific domains. At this study, an inductive learning method has been presented for conceptual classification of Multi-label texts for web mining based on ontology through using Term Space Reduction (TSR) and also using mutual information (MI).For, Term Space Reduction (TSR) may increase efficacy and performance averagely equal to or less than 5% [1]. Recall and Precision is criterion of evaluation for the proposed method [2]. If a term is classified inside a category, that term is positive towards that category, otherwise, that term is negative towards that category. At this method, micro–averaging is used for evaluation of proposed Recall and Precision method. If some terms are turned positive towards category, based on used ontology, the term which is nearer to C category

* Corresponding author.

V.V. Das, E. Ariwa, and S.B. Rahayu (Eds.): SPIT 2011, LNICST 62, pp. 43–48, 2012.

semantically is considered positive (correct) while the rest terms towards C category are considered negative (incorrect). Principally, at this stage, significance of ontology will be taken into consideration, causing terms to be classified accurately and precisely in correct categories. The Laboratory results show that the presented method has high precision average than existing methods of SVM, Find Similar, Naïve Bayes, Bayes Nets, and Decision Trees.

2 Presenting a New Method for Conceptual Sectioning of Text for Web Mining Based on Ontology

At this part, a new inductive learning method has been presented for classification of Multi-label text based on Term Space Reduction (TSR) and Mutual Information (MI) measure through using ontology. The difference of new proposed method with method presented by[3] is as follows: Depending on type of classification which may increase efficacy and performance averagely less than or equal to five percent, in this method, we use ontology and Term Space Reduction (TSR) [1]. Similar to method [3], we use mutual information (MI) measure for selection of term.The main stages of the method include as follows:

1. The specified Stop Words are removed from set of series of documents [2].such as a, an, the, that.
2. Root of words is specified through the application of Porter algorithm and terms are reduced according to their roots form. [4] (For example, "Compute", "Computing" , "Computer" are reduced to "Compute" and "Walker", "Walking" and "Walks" is reduced to "Walk".
3. The terms which occur less than five times at set of series of test are removed [4][5][6].because, the word which occurs only some terms, it is not reliable statistically.
4. The terms which have been used only in one document are removed [7].
5. MI size of remaining terms is obtained and 300 terms, which their size are more than remaining terms, are used for testing categories in one category [2].

$$MI(t_i, C) = P(t_i, C)\log_2 \frac{P(t_i, c)}{P(t_i)P(c)}. \tag{1}$$

$$P(t_i, C) = \frac{N_c(t_i)}{N_C} \quad , \quad P(t_i) = \frac{N(t_i)}{N_C} \quad P_C = \frac{N_C}{N}.$$

$N_C(t_i)$ denotes the number of occurrences of term t_i in category C, N_C denotes the number of occurrences of all terms in category C,N(t_i) denotes the number of occurrences of term t_i in the collection, N denotes the number of occurrences of all terms in the collection After specifying 300 terms, which enjoys the highest size of

MI in category, a K × N matrix is considered. K is number of terms and N is the number of documents in each category. This matrix is document descriptor matrix and shows binary [1, 0] weight of terms in documents. If term t_i has existed in di document, the amount one is specified, otherwise, the zero amount is displayed. Then, cosine similarity measurement [3] is used for constructing "S" document similarity matrix.

$$S_{(i,j)} = \frac{A(i)A(j)}{\|A(i)\| \times \|A(j)\|}. \tag{2}$$

$S_{(i,\ j)}$): shows similarity degree between document *di* and document *dj* and $S(i, j) \in [0,1]$

A(i), A(j) :ith and jth column vectors of the document descriptor matrix A.

We can obtain the term-document relevance matrix R by applying the inner product of the document descriptor matrix A to the document-similarity matrix S, shown as follows:

$$R=A.S \tag{3}$$

Then, R matrix is multiplied in $\bar{1}$ vector and $\overline{V_c}$ vector is obtained for C category.

$$\overline{V_c} = R.\bar{1}\ \overline{V_c} = R.\bar{1}\bar{1} = [1,1,...,1]^T \tag{4}$$

$\overline{V_c}$ vector is normalized through the application of average weight. At this method, each vector element is divided into total elements of vector, aimed at obtaining its normal. In fact, ith term weight in c category is obtained through the application of $\overline{V_c}$ into W_{C_i} as follows:

$$W_{C_i} = w_{c_i} \times \log_2 \frac{|c|}{cf_i}. \tag{5}$$

W_{C_i} : denotes the refined weight of the ith term in the refined category descriptor vector $\overline{w_c}$. $|c|$: denotes the number of categories. cf_i: denotes the number of category descriptor vectors containing term t_i.

This refinement reduces the weights of the terms that appear in most of the categories and increases the weights of the terms that only appear in a few categories. Assume that the document descriptor vector of a testing document d_{new} is $\overline{d}_{new}$. We can then apply the inner product to calculate the relevance score Score(c, d_{new}) of category c with respect to the testing document d_{new} as follows:

$$Score(c, d_{new}) = \overline{d}_{new}.\overline{w}_c. \tag{6}$$

In other words, we choose the maximum relevance score L among them. If the relevance score between a category and the testing document divided by L is not less than a predefined threshold value , where $\lambda \in [0,1]$, then the document is classified into that category.

1) Using Ontology: If rank of some terms to C category is turned positive, according to the used ontology, the term, which is nearer to C category semantically, is classified positive as correct (TPi) while the rest terms to C category, as categorized positive, will be considered as incorrect. Principally, significance of ontology is specified at this stage and will cause categorization of terms in correct categories with more precision and accuracy. Because, when ontology is used, the number of categorized positive documents are turned zero incorrectly, causing singularity precision with various threshold limit between zero and one. This procedure is shown in the flowchart of Figure 1.

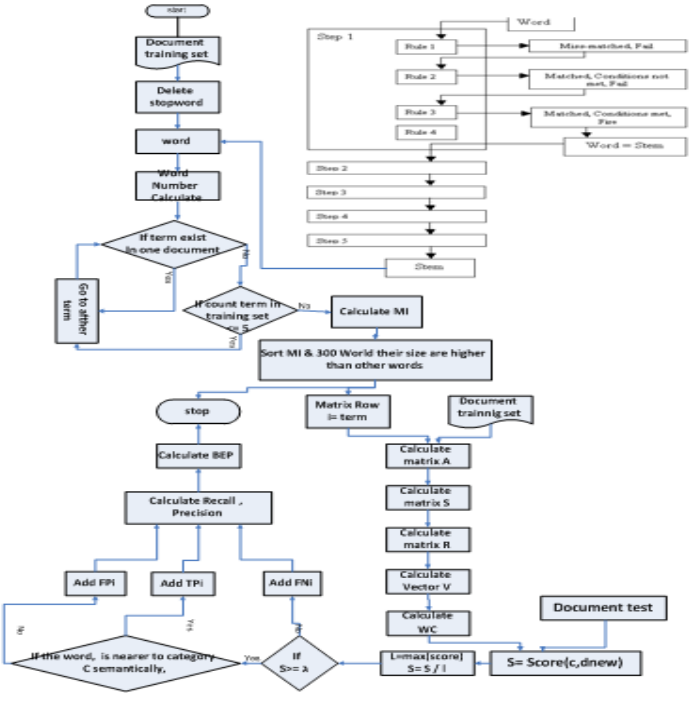

Fig. 1. Flowchart of the new proposed method

3 Implementation and Comparison of Methods

For implementation of proposed method for categorization of Multi-label text, set of 10-category data of "Reuters–21578 &Apte Split" and Delphi 7.0 version and SQL

Server 2005 are used through the application of XP Windows. Table1 shows results of six various algorithms in 10 normal categories appropriately. The presented new proposed method shows better results than other methods, indicating average 94.5 percent for 10 normal categories. After it, SVMs has shown better results, indicating 2.5 percent less than our proposed method and contains average 92 percent for 10 normal categories. The authenticity and accuracy of Decision Tree stands at 3.6 percent less than SVM, indicating average 88.4 percent for 10 normal categories. Bayes Nets has the efficacies for improvement of naïve Bayes as it is expected, but its privileges are rather partial. All advance learning algorithms increase efficacy and performance as much as 15 to 20 percent in comparison with development of searching of Rocchio (Find Similar) type. It should be noted that inductive learning method based on ontology and SVMs show satisfactory and best results in categorization and produce the best results for this set of test series.

Table 1. Breakeven performance for Reuters-21578 Aptè split 10 categories

Method Category	Find similar Rocchio 1971	Naïve Bayes Lewis 1994	Bayes Nets Sahami 1996	Decision trees Chinkering, 1997	Linear SVM Vapnik 1995	Multilabel Method Chang, Chen 2006	The proposed method With ontology
Earn	92. 9 %	95. 9 %	95. 8 %	97. 8 %	98. 0%	97.5%	60%
Acq	64. 7 %	87. 8 %	88. 3 %	89. 7 %	93. 6%	95.1%	100%
Money-fx	46. 7 %	56. 6 %	58. 8 %	66. 2 %	74. 5%	79.2%	100%
Grain	67. 5 %	78. 8 %	81. 4 %	85. 0 %	94. 6%	84.7%	100%
Crude	70. 1 %	79. 5 %	79. 6 %	85. 0 %	88. 9%	84.4%	100%
Trade	65. 1 %	63. 9 %	69. 0 %	72. 5 %	75. 9	85%	100%
Interest	63. 4 %	64. 9 %	71. 3 %	67. 1 %	77. 7%	81%	100%
Ship	49. 2 %	85. 4 %	84. 4 %	74. 2 %	85. 6%	85.4%	93.64%
Wheat	68. 9%	69. 7%	82. 7%	92. 5%	91. 8%	79.8%	93.64%
corn	48. 2 %	65. 3 %	76. 4 %	91. 8 %	90. 3%	78.2%	96.36%
Average	64. 6 %	81. 5 %	85 %	88. 4 %	92. 0%	91.3%	94.5%

In our implementation singularity is observed for all precision points. As the result show for our proposed method based on ontology, the average separation point stood at approx. 94.5 percent through 100 percent of training data in 8categories and 10 percent of training data in two "earn" and "acq" categories. The results show that if 100 percent of data are used in these two categories, the average separation point will be exceeded. It should be noted that threshold limit used in our proposed model is based on (R–cut) [8].Figure 2 Shows ROC curve for grain category and SVM privileges are observed in the length of Recall – Precision space.

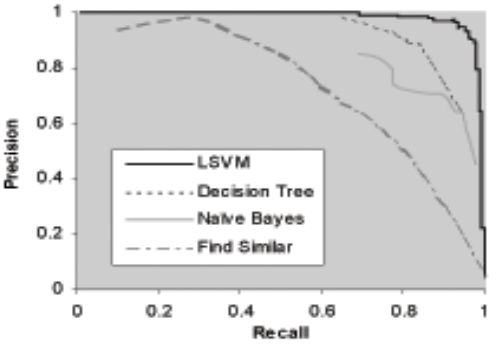

Fig. 2. ROC curve for grain category

4 Conclusion and Future Research

Laboratory results shown in Table1 indicate that the presented method has high average precision than methods of SVM, Find Similar, Naïve Bayes, Bayes Nets and Decision Tree. The issue of paralleling and improvement of performance up to proposed method can be operational. Regards to the capabilities of ontology, effective steps can be carried out on subject of training and synchronization of ontology based on obtained feedbacks, with the aim of producing best results. Also, ontology can be used in Term Space Reduction (TSR) of text with the aim of obtaining better and certain results.

References

1. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14th International Conference on Machine Learning, Nashville, USA, pp. 412–420 (1997)
2. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Survey 34(1), 1–47 (2002)
3. Chang, Y.-C., Chen, S.-M., Liau, C.-J.: A New Inductive Learning Method for Multilabel Text Categorization. In: Ali, M., Dapoigny, R. (eds.) IEA/AIE 2006. LNCS (LNAI), vol. 4031, pp. 1249–1258. Springer, Heidelberg (2006)
4. Sever, H., Gorur, A., Tolun, M.R.: Text Categorization with ILA. In: Yazıcı, A., Şener, C. (eds.) ISCIS 2003. LNCS, vol. 2869, pp. 300–307. Springer, Heidelberg (2003)
5. Baker, L.D., McCallum, A.K.: Distributional clustering of words for text classification. In: SIGIR 1998, 21st ACM Int. Conference on Research and Development in Information Retrieval (Melbourne, AU), pp. 96–103 (1998)
6. Cohen, W.W.: Learning to classify English text with ILP methods. In: De Raedt, L. (ed.) Advances in Inductive Logic Programming, pp. 124–143. IOS Press, Amsterdam (1995)
7. Dumais, S., Platt, J., Heckerman, D.: Inductive Learning Algorithms and Representations for Text Categorization (1995)
8. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1 (1999)