Algorithm for Generating Fuzzy Rules for WWW Document Classification

Piotr Dziwiński[1] and Danuta Rutkowska[1,2]

[1] Department of Computer Engineering
Faculty of Mechanical Engineering and Computer Science
Czestochowa University of Technology, Poland
dziwinski@kik.pcz.czest.pl
[2] Department of Knowledge Engineering and Computer Intelligence
Academy of Humanities and Economics in Lodz, Poland
drutko@kik.pcz.czest.pl

Abstract. This paper presents the dense areas based algorithm for generating fuzzy rules for classification WWW documents. Description document clusters in the form of fuzzy rules (FR) make possible the presentation of information in the form fuzzy granules. Moreover, each cluster might be described by several fuzzy rules. These fuzzy rules can be used as the knowledge base for searching new information from WWW resources with regard to specific topics and users' requirements.

1 Introduction

An enormous growth of the Internet and significant development of telecommunication techniques enable users to access WWW resources. About 11.5 billion WWW pages have been created until January 2005 [9]. Changeability and large amount of WWW pages is a big challenge for modern crawlers and search engines. They should reflect WWW resources as accurately as possible and also hold information about WWW resources as freshly as possible. Complete crawling entire Web is impossible in reasonable time, no matter which technology is available at the site where the search engines operate. The ideal crawler should be able to recognize relevance and importance of Web pages. The crawlers can order new links extracted from downloaded WWW pages. This can be accomplished by using different methods (e.g. measurement similarity between pages and a current query, amount of the links to point out WWW pages or the most popular Page Rank). Precise description of these methods as well others may be found in [6]. To deal with enormous WWW resources these methods are not enough efficient. One of the possible solutions of this problem can be the use of methods that are based on focused crawling. This approach makes possible to avoid areas of WWW resources that are not relevant to the information requirements of the user. Hersovici et al., in paper [11], proposed the shark-search algorithm of WWW resources. This approach searches WWW resources based on the assumption that relevant WWW pages usually are in relevant neighbourhood. Chakrabarti et al., in paper [4], proposed the focused crawling approach in

L. Rutkowski et al. (Eds.): ICAISC 2006, LNAI 4029, pp. 1111–1119, 2006.

which a crawler selectively discovers pages that are relevant to a pre-defined set of topics. The topics are specified by using exemplary documents. Rungsawang et al. [17] proposed the consecutive crawling to take advantage of experience from earlier crawling process. They build some knowledge base, which are used to produce better result for the next crawling. The classification process plays an important role in the scope of information retrieval. In order to classify WWW pages, different methods are used, such as: k-nearest neighbors algorithm (k-NN), Naive Bayes, support vector machines (SVM), decision trees, neural networks or induction of classification rules. Accurate description of these and other methods may be found in [15], [7], [13], [21]. A classifier can be used to distinguish between relevant and irrelevant WWW pages or resources, to help in a semi-automatic construction of large knowledge bases or to classify unknown Web pages to some predefined categories.

This paper presents an algorithm for fuzzy rules determination, based on dense areas, for WWW documents classification. This paper is organized as follows. Section 1 presents an introduction concerning information retrieval. The next section shortly describes a vector space model for text classification, preprocessing of WWW pages (parsing, stemming, tokenization of WWW pages), and most important methods to calculate importance of tokens of WWW pages and used normalization. A fuzzy inference system, simple fuzzy rules and membership functions are presented in Section 3. An algorithm for determination of the fuzzy rules based on dense areas is presented in Section 4. Experiments and their results are described in Section 5. Final remarks and directions of further works are outlined in Section 6.

2 Preprocessing of WWW Pages

Preprocessing of WWW pages is performed at the first phase. First of all, the WWW pages are parsed. The WWW pages are reduced to an unstructured representation. This can be achieved by retaining the text enclosed between *<html>* and *</html>* tags. The html tags, part of the text between script tags, some characters (e.g. %, #, $), numbers and other elements can be removed. The conversion of strings that encode international characters to standard representation should be done, if necessary. A well-designed parser ought to have a simple rule base (including e.g. html tags, script tags), which can be easily modified.

Then, stemming is applied to WWW pages. The stemming process reduces words to their root form, which become an actual index of terms. One of first method of this kind is Porter stemming algorithm [16] introduced in 1978. A new version of this algorithm may be found in [12].

Next, the stop-words are removed from WWW pages. The next phase is the process of tokenization. In this phase, a list of keywords (terms), in the form of pairs *(term, number of appearance in document)*, is created. If we have a keyword dictionary for some languages, we perform reduction of terms. As a result of the initial process with regard to WWW pages, we obtain a vector space model of terms in the form

$$\boldsymbol{x}_i = [x_{i0}, x_{i1}, \ldots, x_{iK-1}]^T, \quad i = 0, 1, \ldots, N-1$$

where: N – number of WWW pages, K – number of keywords, x_{ij} – number of appearance of keyword j in document i, for $j = 0, 1, \ldots, K-1$.

The vectors of keywords that represent WWW pages reflect a degree of association between keywords and WWW pages in the form of term weights. Most of methods determine these weights based on statistics. One of the simplest methods calculates the term frequency (TF) of WWW pages as follows [3]

$$TF_{ij} = \frac{x_{ij}}{k_i} \tag{1}$$

where k_i – number of keywords in document i.

This method is a very poor way to determine the degree of association between keywords and documents. The term weights can be inversed according to the number of occurrences in different documents. In order to achieve this relationship, we calculate the inverse document frequency by the following formula [3]

$$IDF_j = \log \frac{N}{n_j} \tag{2}$$

where n_j – number of documents in the collection of N documents in which keyword j occurs.

The most popular method determining term weights is the product of TF and IDF, in the form [3]

$$TF_{ij} \cdot IDF_j = \frac{x_{ij}}{k_i} \cdot \log\left(\frac{N}{n_j}\right). \tag{3}$$

WWW documents have different sizes. For a short document, possibility of use a keyword is smaller that for large documents. For large documents, TF achieves large values. In order to neutralize this unfavorable effect, we can use one of normalization methods. The most popular method is the cosine normalization, which gives the following equations, when applied to (1) and (3), respectively

$$Norm \quad TF_{ij} = \frac{\frac{x_{ij}}{k_i}}{\sqrt{\sum_{m=0}^{K-1}\left(\frac{x_{im}}{k_i}\right)^2}} \tag{4}$$

$$Norm \quad TF_{ij} \cdot IDF_j = \frac{\frac{x_{ij}}{k_i} \cdot \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{m=0}^{K-1}\left(\frac{x_{im}}{k_i} \cdot \log\left(\frac{N}{n_m}\right)\right)^2}}. \tag{5}$$

The number of keywords to represent WWW pages can be restricted by selecting those keywords for which the number of occurrences in different documents is bigger than a certain threshold, P, which depends on the number of WWW pages. Some of the keywords, for which the frequency of occurrence is too high, can be removed.

3 Fuzzy Rules and Fuzzy Inference System

The key point of the issue of focused crawling is searching new information from WWW resources related to some specific topic or a set of topics. The information requirements by users, and topic classification of the WWW documents, are uncertain. We cannot describe these needs by use of the classical set theory and Boolean logic. In order to catch this non-crisp information, we should apply the fuzzy logic [23], [24]. In paper [14], Kraft et al., presents different approaches to fuzzy rules construction. In this paper, we are focusing exclusively on fuzzy description of classes of WWW documents that belong to specific topics, in the form of fuzzy rules. These rules can be a part of a fuzzy inference system (FIS). Precise description of FIS can be found e.g. in [20], [18], [19]. Standard FIS consists of the following elements: rule base, fuzzyfication unit, inference unit, defuzzyfication unit. The rules, in the rule base, are of the following form

$$R^{(l)}: \textbf{ IF } x_0 \text{ is } A_0^l \textbf{ AND } x_1 \text{ is } A_1^l \textbf{ AND } \ldots \textbf{ AND } x_{K-1} \text{ is } A_{K-1}^l \quad (6)$$

$$\textbf{THEN } y \text{ is } B^l$$

where: $l = 0, 1, \ldots, L-1$, so L – number of fuzzy rules; $\boldsymbol{x} = [x_0, x_1, \ldots, x_{K-1}]^T$ – linguistic variables that corresponds to inputs of FIS; A_i^l – fuzzy set for input linguistic variable, for $i = 0, 1, \ldots, K-1$, and B^l – fuzzy set that corresponds to the output linguistic variable y.

For classification tasks, we apply the simpler form of the rules, as follows

$$R^{(l)}: \textbf{ IF } x_0 \text{ is } A_0^l \textbf{ AND } x_1 \text{ is } A_1^l \textbf{ AND } \ldots \textbf{ AND } x_{K-1} \text{ is } A_{K-1}^l \quad (7)$$

$$\textbf{THEN } \quad y \in \text{class } \quad \mathbf{c}$$

where c is the number associated with the class corresponding to this rule, $c = 0, 1, \ldots, C-1$, and C – number of classes; see Section 4.

The inference based on the rule of this type determines the rule activation degree. Fuzzy set A_i^l is defined by Gaussian membership function [20]

$$\mu_A(x_i) = \exp\left(\frac{-(x_i - v_i^l)^2}{(\sigma_i^l)^2}\right) \quad (8)$$

or asymmetrical Gaussian membership function

$$\mu_{A_i^l}(x_i) = \begin{cases} exp\left(\frac{-(x_i - v_i^l)^2}{(\sigma_{l_i}^l)^2}\right) & \text{if } x_i < v_i^l \\ exp\left(\frac{-(x_i - v_i^l)^2}{(\sigma_{r_i}^l)^2}\right) & \text{if } x_i > v_i^l \\ 1 & \text{if } x_i = v_i^l \end{cases} \quad (9)$$

where: v_i^l – center, σ^l – width, σ_r^l, σ_l^l – right and left parts of width of the membership functions.

To determine fuzzy rules (8) and (9), it is necessary to calculate the parameters of fuzzy sets A_i^l .

4 Algorithm for Fuzzy Rules Generation

In [22] Tao et al. provide unsupervised fuzzy clustering algorithm to cluster pixels in a color image. This algorithm is based on the subtractive clustering algorithm, proposed in [5], which uses the density function, similar to that expressed by Eq. (10). The "density" is understood as *number of points in the neighborhood*. We modify this algorithm in order to determine fuzzy rules for classification WWW documents. In this way, we obtain the following algorithm.

Let us denote:
$\boldsymbol{x}^c = [\boldsymbol{x}_0^c, \boldsymbol{x}_1^c, \ldots, \boldsymbol{x}_{N^c-1}^c]^T$, where N^c – number of vectors which describe WWW pages that belong to cluster c, for $c = 0, 1, \ldots, C-1$; C – number of classes, $\boldsymbol{x}_i^c = [x_{i0}^c, x_{i1}^c, ..., x_{iK-1}^c]^T$, where K – number of keywords (vector length); this vector includes term weights, $\boldsymbol{d}^c = [d_0^c, d_1^c, ..., d_{N^c-1}^c]^T$ – density vector that corresponds to individual vectors $\boldsymbol{x}_i^c$ which includes information about the number of other vectors in radius r_k^c, $\boldsymbol{D}^l = [D^{c0}, D^{c1}, \ldots, D^{cL^c-1}]^T$ – density vector that corresponds to membership functions $\mu_{A^{cl}}$, where $l = 0, 1, \ldots, L^c - 1$, and L^c – number of generated rules for class c.

The basic function that describes the density of vectors, also called the density function, is expressed as follows

$$d_i^c = \sum_{j=0}^{N^c-1} \prod_{k=0}^{K-1} exp\left(-\frac{(x_{jk}^c - x_{ik}^c)^2}{(r_k^c)^2}\right) - \sum_{l=0}^{L^c-1} D^{cl} \cdot \prod_{k=0}^{K-1} \mu_{A_k^{cl}}(x_{ik}^c) \qquad (10)$$

where membership function $\mu_{A_k^{cl}}$ is defined by (8) and (9).

The algorithm that determines fuzzy rules can be presented in the following steps:

1. Let $L = 0$;
2. Determine the parameter value, r_k^c, depending on the domain of the input values
$$r_k^c = \frac{\max_{j=0,1,\ldots,N^c-1}(x_{jk}^c) - \min_{j=0,1,\ldots,N^c-1}(x_{jk}^c)}{R} \qquad (11)$$
where R – multiple factor domain of input values (this factor influences on the number of fuzzy rules and accuracy of classification).
3. Calculate the density function for each vector x_i^c, by use of Eq. (10).
4. Determine m such that
$$d_m^c = \max_{i=0,1,\ldots,N^c-1}(d_i^c)$$
5. If the density value $d_m^c > 0$, then go to step (6), else stop the algorithm.
6. Set initial values of parameters of membership functions $\mu_{A_k'^c}$
$$v_k'^c = x_{mk}^c, \quad \sigma_k'^c = r_k^c,$$

7. Refresh center and width parameters of membership functions $\mu_{A_k'^c}$, using Eqs. (12), (13) [8]

$$v_k^{'c} = \sqrt{\frac{\sum_{j=0}^{N^c-1} \mu_{A_k'^c}(x_{jk}^c) \cdot x_{jk}^c}{\sum_{j=0}^{N^c-1} \mu_{A_k'^c}(x_{jk}^c)}} \tag{12}$$

$$\sigma_k^{'c} = \sqrt{\frac{\sum_{j=0}^{N^c-1} \mu_{A_k'^c}(x_{jk}^c) \cdot (x_{jk}^c - v_k^{'c})^2}{\sum_{j=0}^{N^c-1} \mu_{A_k'^c}(x_{jk}^c)}} \tag{13}$$

where $k = 0, 1, \ldots, K-1$ and $c = 0, 1, \ldots, C-1$.
Equation (12) is used to calculate the arithmetic weight-mean of the vectors, in radius r_k^c, by means of membership function $\mu_{A_k'^c}$.

8. Refresh the value of density d_m^c using the following formula

$$d_m^c = \sum_{j=0}^{N^c-1} \prod_{k=0}^{K-1} \mu_k^{'c}(x_{jk}^c) - \sum_{l=0}^{L^c-1} D^{cl} \cdot \prod_{k=0}^{K-1} \mu_{A_k^{cl}}(v_k^{'c}) \tag{14}$$

9. Increase the number of rules $L^c = L^c + 1$
Then, add the new membership function, determined in step 7, as follows

$$\mu_{A_k^{cL^c-1}} = \mu_{A_k'^c}$$

and the density value obtained in step 8,

$$D^{cL^c-1} = d_m^c$$

where $k = 0, 1, \ldots, K-1$, and K – number of keywords,

10. Go to step 3.

5 Experimental Results

In this paper, the algorithm for determination of fuzzy rules for WWW document classification, based on dense areas is introduced. The algorithm was tested on the set of 568 abstracts of documents that belong to four classes: Artificial Intelligence (116), Robotics and Vision (92), Systems (202) and Theory (158), available at WWW server [25]. In the initial process, we obtained 3819 keywords.

Experiments have been performed for selected keyword factor values P from 1 to 60. The number of keywords was determined by selecting these keywords for which the number of occurrences in different documents was bigger than threshold P. For each set of keywords, the cosine normalization $TF_{ij} \cdot IDF_j$ was calculated in order to create vectors that represent WWW pages. This set of

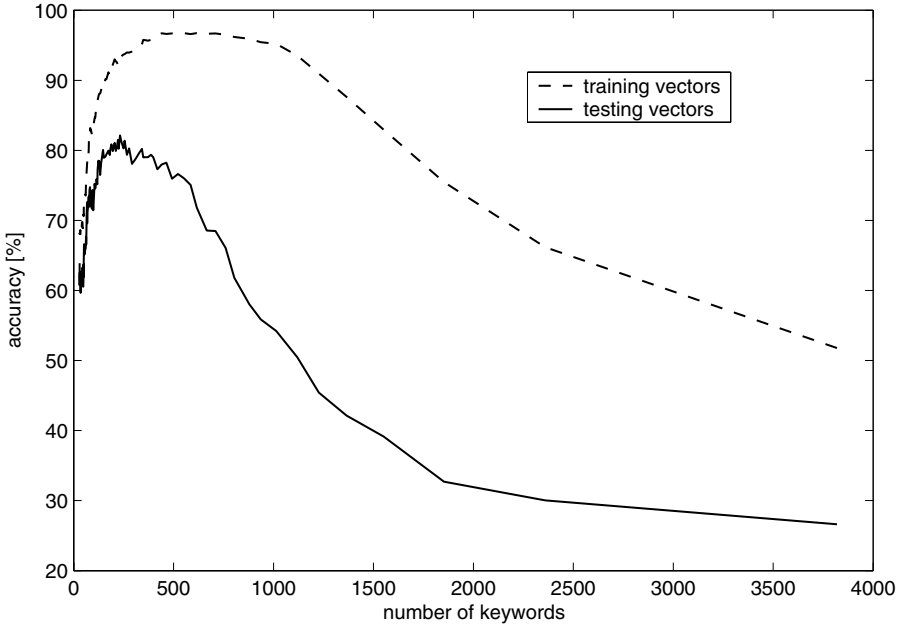

Fig. 1. Average accuracy of determination of fuzzy rules for different number of keywords for a constant value of training and testing vectors

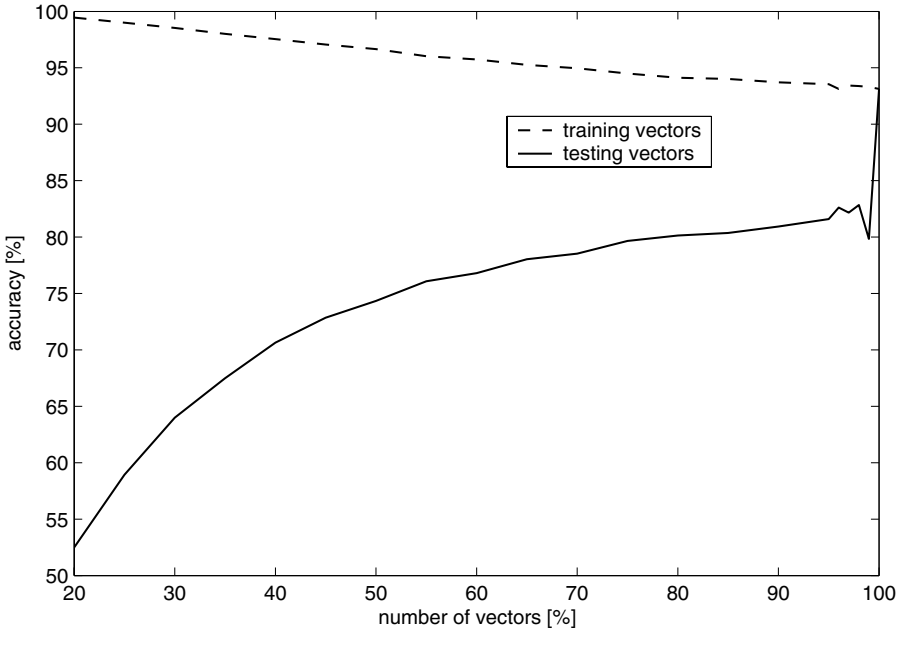

Fig. 2. Average accuracy of determination of fuzzy rules for different number of training and testing vectors for a constant value of keywords

vectors have been divided randomly into a training set (95% of all vectors), and a testing set (5% of all vectors). The vectors from the testing set are not included in the training set. The process of determination of fuzzy rules has been executed 100 times for different training sets. The results are presented in Fig. 1. The best average result for training vectors is 96.75% for 551 keywords and 82.14% for 232 keywords for testing vectors. Then, for 232 keywords (the best set of keywords for the testing vectors set) the experiments have been done 100 times. We determined ability of the fuzzy inference systems to correct classification, depending on different length of the learning and testing sets. This relationship is shown in Fig. 2. The experiments have been performed for other methods of calculation of degree of association between keywords and WWW pages, such as: TF, IDF, normalization TF, $TF \cdot IDF$.

6 Final Remarks

Correctness of the classification weakly depends on methods of calculation of term weights, for the algorithm presented in this paper. The results for the testing vectors are worse than for the training vectors because of huge number of keywords and the fact that some WWW pages contain only a small subset of the keywords. Another reason is that some keywords from the testing sets may not occur in the training sets or may occur very rarely. The effectiveness of the algorithm for generating fuzzy rules is good even for very small training sets (63.9% for 30% of all vectors, 52.5% for 20% of all vectors).

In further work, we plan to apply a two-phase algorithm. The first phase can be used to obtain centers of fuzzy rules in high dimensional keyword space. These centers can further be used for reduction of the keyword space. In this purpose, we plan to employ a method based on the Orthogonal Basis of Centroids [2]. The algorithm of generating the fuzzy rules will be applied once more for vectors in a low dimensional space at the second phase. We expect that this two-phase algorithm may improve the results. Some experiments have already been done, and the improvements observed.

References

1. Baldi P., Frasconi P., Smyth P., Modeling the Internet and the Web, Probabilistic Methods and Algorithms. Wiley. 2003.
2. Berry M. W., Survey of Text Mining, Clustering, Classification, and Retrieval. Springer-Verlag, New York, 2004.
3. Bo-Yeong K., Dae-Won K., Sang-Jo L., Exploiting concept clusters for content-based information retrieval, Information Sciences 170 (2005) 443-462.
4. Chakrabarti S., Van den Berg M., Dom B., Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks 31 (1999) 1623-1640.
5. Chiu S.L., Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Systems 2(3) (1994) 267–278.
6. Cho J., Garcia-Molina H., Page L., Efficient crawling through URL ordering, Computer Networks and ISDN Systems 30 (1998) 161-172.

7. Cortes C., Vapnik V.N., Support vector networks, Machine Learning 20 (1995) 1-25.
8. Euntai K., Minkee P., Seunghwan J., Mignon P., A new approach to fuzzy modeling, IEEE Transaction on Fuzzy Systems, Vol.5, No.3 (1997) 328-337.
9. Gulli A., Signorini A., The Indexable Web is More than 11.5 bilion pages, http://www.cs.uiowa.edu/~asignori/pubs/web-size.
10. Hastie T., Tibshirani R., Friedman J., Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer (2001).
11. Hersovici M., Jacovi M., Maarek Y. S., Pelleg D., Shtalhaim M., Ur S., The shark-search algorithm - an application: tailored web site mapping. In: Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia (1998) 317-326.
12. Jones S.K., Willet P., Readings in Information Retrieval, San Francisco: Morgan Kaufmann (1997).
13. Kłopotek A. M., Intelligent Search Engines. EXIT, Warszawa (2001), (in Polish).
14. Kraft D.H., Martin-Bautista M.J., Chen J., Sanchez D., Rules and fuzzy rules in text concept, extraction and usage, International Journal of Approximate Reasoning 34 (2003) 145-161.
15. Lam W., Ho C.Y., Using a generalized instance set for automatic text categorization. In Proc. SIGIR-98, 21stACM Int. Conf. on Research and Development in Information Retrieval (1998) 81-89.
16. Porter M., An algorithm for suffix stripping, Program, Vol.14, No.3 (1978) 130-137.
17. Rungsawang A., Angkawattanawit N., Learnable topic-specific web crawler, Journal of Network and Computer Application 28 (2005) 97-114.
18. Rutkowska D., Piliński M., Rutkowski L., Neural Networks, Genetic Algorithms and Fuzzy Systems. PWN, Warszawa, 1999, (in Polish).
19. Rutkowska D., Neuro–Fuzzy Architectures and Hybrid Learning. Springer–Verlag, Heidelberg, (2002).
20. Rutkowski L. Artificial Inteligence Methods and Techniques. PWN, Warszawa, 2005, (in Polish).
21. Sebastiani F., Machine learning in automated text categorization, ACM Computing Surveys (2002) 1-47.
22. Tao C.W., Unsupervised fuzzy clustering with multi-center clusters, Fuzzy Sets and Systems, Vol.128 (2002) 305-322.
23. Zadeh L. A., Fuzzy logic = Computing with words, IEEE Transactions of Fuzzy Systems, Vol.4, No.2 (1996).
24. Zadeh L. A., Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 90 (1997) 111-127.
25. Source test data: http://www.cs.rochester.edu/trs.