Web page categorization in process of focused crawling

Emil Gatial, Ladislav Hluchý, Martin Šeleng, Michal Laclavík and Zoltán Balogh
Institute of Informatics,
Slovak Academy of Sciences, Dubravska cesta 9, 845 07 Bratislava, Slovakia
e-mail: emil.gatial@savba.sk

Abstract

This article describes content evaluation algorithm based on neural network's neuron implementation (TLU). Designed algorithm processes HTML page text and evaluates the rate of domain membership according to the most pertinent keywords from given domain. Such algorithm is part of implementation focused crawling tool used for domain specific traversing of Internet. In the closing part provides brief algorithm validation on pages written in English and Slovak language as well as proposes possible algorithm improvements.
Keywords: neural network, TLU, web crawler, domain.

1. Introduction

A crawling process is characterized by periodical requesting and downloading of Internet documents. In many cases only small portion of acquired documents are convenient for further processing. Identification of features characterizing the set of convenient documents plays the key role in the process of document separation. Sometimes the features can be obviously identified (i.e. specific keywords, tags, atc.), but more likely the features can't be exactly identified because of the tremendous heterogeneity of Internet documents. The tasks of feature identification and document categorization fall into the domains of Information Retrieval (IR) and Artificial Intelligence (AI). The methods like decision trees [1], rule learning [2], artificial neural networks (ANN) [3] [4], K-nearest neighbor algorithm [5], support vector machines (SVM) [6], and Naive Bayes methods [7]. The most recently explored methods regards the hierarchical categorization investigated in the work of Koller and Sahami [8].
Documents can be filtered using specific properties contained and well observed

In this paper the single domain document membership categorization is explored using the NN approach. The paper put accent on the evaluation of categorization that is used during the web crawling process downloading the pages with similar content. Such approach of processing is known as the focused (sometimes known as topical) crawling is introduced by [9]. The paper proposes new ideas for feature selection and self-learning process.
First, the paper explains the basics of neural network approach and feature selection problem. Next the integration of ANN categorization algorithm with the web crawler is described. The close of this paper evaluates of described categorization.

2. Categorization using artificial neural network

The cornerstone of NN is artificial neuron that is an abstraction of biological neurons. It usually receives many inputs, called input vector, (analogy of dendrites) and the sum of them forms output (analogy of synapse). The sums [1.1] of inputs are weighted and the sum is passed to the *activation function*. The original artificial neuron is the Threshold Logic Unit (TLU) first proposed by

McCulloch and Pitts [10]. Perceptron employs a *threshold* or *step function* taking on the values 1 or 0 only. The schema of such artificial neuron is depicted on the figure 1. The categorization using the ANN approach is based of the evaluation of feature frequencies. A perceptron is a linear threshold classier that separates examples with a hyperplane. It is perhaps the simplest learning model that is used standalone.

$$a = \sum_{ki \in K} N_{ki} \cdot w_{ki} \qquad [1.1]$$

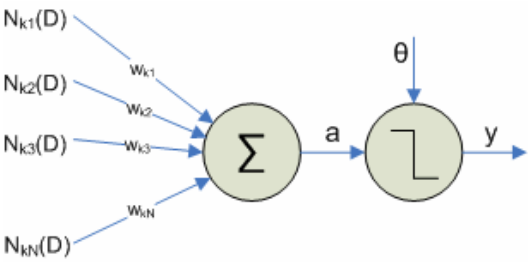

Figure 1 Neuron abstraction using TLU.

2. 1 Feature selection

The feature can be possibly represented using the different words contained by document collection. Even the small document collection can contain large amount of features therefore the reduction of feature set is necessary to train artificial neuron. The behavior of artificial neuron is sensitive to the size and quality of reduced feature set.

First steps of reduction are filtering the stop words and use of stemming algorithms. The two approaches can be applied on the resulting words for feature selection: wrapper approach and the filter approach [11]. The filter approach selects the feature set using the preprocessing methods by evaluation of correlation coefficient, mutual information and odds ratio. The wrapper approach selects initial set and measures the performance of the artificial neuron. In the paper the wrapper method is used and partially examined.

For the simplicity the initial set of features for wrapper approach is selected by supervising person. The application of combination of filter and wrapper approaches is planned for the future research.

2. 2 Learning the TLU

The TLU training process covers weights modification according to the preprocessed training set. The proposed algorithm is using the *perceptron learning rule* [12]. This kind of training is known as supervised training.

The training process repeats until the process exceeds the specified number of iterations or the TLU is fully trained. The provided experiments showed that the training process depends not only on the size of selected features, but on the quality of features. This fact is used for identification of new features.

2. 3 Integration of focused crawler

The crawler periodically passes the downloaded data to the HTML parser (generally, ANN approach can be applied to any document that can be converted to textual document) that extracts text. The frequencies of features are processed from the text. Such frequencies represent the input vector N_k and its weights w_k. According to this vector the output is evaluated and the web crawler decides if the page will be stored or not.

3. Testing and Validation

Following tests are performed on the web pages written in English and Slovak language. The aim of provided tests has to prove that ANN approach is suitable method for focused web crawling.

The two tests were realized. The test no. 1 provides reference measurement that evaluates the performance of TLU trained on the training set that is a subset of evaluated HTML pages. The test no. 2 provides evaluation of TLU performance

trained on the training set that is subset only of the part of HTML pages. The distinction between the two tests lies in the different training set. While test no. 1 provides only reference evaluation according to the subset of all categorized document, the test no. 2 is made with the TLU trained with the training set that is disjunctive with to evaluating training set. The difference of the tests should show the correlation of behavior of TLU trained on the conjunctive and disjunctive train sets. In the practice this means the degradation of TLU performance when the categorization is used on the set that originates from different web site.

In order to evaluate a decision results we first define a contingency matrix representing the possible outcomes of the classification as shown in Table 1. Several measures and the corresponding names used in the IR community are specified in the Table 2.

Table 1 Contingency table for binary classification

	Class Positive (C+)	**Class Negative (C-)**
Assigned positive (A+)	a True Positives	b False Positives
Assigned negative (A-)	c False Negatives	d True Negatives

Table 2 Efficiency measures for binary classification defined in the Artificial Intelligence community

IR	
recall	$\frac{a}{a+c}$
precision	$\frac{a}{a+b}$
fallout	$\frac{a}{a+d}$
error rate	$\frac{b+c}{a+b+c+d}$
F1 measure	$\frac{2a}{2a+b+c}$

Many applications require the recall parameter closing to the value of 1 that means the algorithm refuses the minimal number of web pages that should be accepted (False Negatives).

3. 1 Evaluation of training set in Slovak language

Training set for the test no. 1 consists of the web pages from the following web sites: www.ejobs.sk, www.praca.sk, www.ponuky.sk, www.profesia.sk.

The test no. 2 is trained on the training set consists of first three sites. Evaluation is processed on the site www.profesia.sk.

Table 3 Contingency table of test no. 1

	Class Positive (C+)	**Class Negative (C-)**
Assigned positive (A+)	124	4
Assigned negative (A-)	1	121

Table 4 Efficiency measures for test no. 1

IR	
recall	0.992
precision	0.96875
fallout	0.5061
error rate	0.02
F1 measure	0.9802371

Table 5 Contingency table of test no. 2

	Class Positive (C+)	**Class Negative (C-)**
Assigned positive (A+)	120	2
Assigned negative (A-)	5	123

Table 6 Efficiency measures for test no. 2

IR	
recall	0.96
precision	0.9836066
fallout	0.493827
error rate	0.028
F1 measure	0.9716599

3. 2 Evaluation of training set in English language

Training set for the test no. 1 consists of the web pages from the following web sites: www.jobbankusa.com, www.jobserve.com, www.myjobsearch.com.

The test no. 2 is trained on the training set consists of first three sites. Evaluation is processed on the site www.jobbankusa.com.

Table 7 Contingency table of test no. 1

	Class Positive (C+)	Class Negative (C-)
Assigned positive (A+)	124	12
Assigned negative (A-)	1	113

Table 8 Efficiency measures for test no. 1

IR	
recall	0.992
precision	0.9117647
fallout	0.5232
error rate	0.052
F1 measure	0.95019156

Table 9 Contingency table of test no. 2

	Class Positive (C+)	Class Negative (C-)
Assigned positive (A+)	119	17
Assigned negative (A-)	6	108

Table 10 Efficiency measures for test no. 2

IR	
recall	0.952
precision	0.875
fallout	0.52423
error rate	0.092
F1 measure	0.9118

4. Conclusion

The artificial neural networks can be used in broad field of applications. One of the usages is shown on the web page categorization application for the focused crawling process. The ANN offers following two very important properties for categorization: fast input vector processing and learning possibility that allows dynamic categorization behavior.

Realized experiments showed that the quality of the selected feature is very important for learning and categorization process. Quality of feature usually reflects the correlation to specific domain of categorization. On this idea the wrapper method of feature selection can be improved. The features can be dynamically rearranged to achieve more appropriate performance of TLU. The rearrangement process is the task for the future research.

Evaluated system employs only single TLU for categorization, but ANN are usually formed into complex networks that can provide much better results, moreover create hierarchical categorization [13] of Internet documents.

Acknowledgements

This work is supported by projects int.eu.grid EU 6FP RI-031857, NAZOU SPVV 1025/2004, RAPORT APVT-51-024604, VEGA No. 2/6103/6, VEGA 2/7098/27.

References

[1] Moulinier I and Ganascia JG (1996) Applying an existing machine learning algorithm to text categorization. In: S Wermer, E Riloff and G Scheler Eds., Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, Springer Verlag, Heidelberg, Germany, pp. 343-354.

[2] Apte C, Damerau F and Weiss SM (1994), Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 3(12):233-251.

[3] Ng HT, Goh WB and Low KL (1997), Feature selection, perceptron learning, and a usability case study for text categorization. In: Belkin Nicholas, Desai NarasimhaluA and Willett Peter Eds., Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, July 1997, pp. 67-73.

[4] Wiener E, Pedersen JO and Weigend AS (1995) A neural network approach to topic spotting. In: Proceedings of SDAIR'95, pp. 317-332.

[5] Yang Y and Pedersen JO (1997), A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997). Morgan Kaufmann Publishers, San Francisco, CA, July 1997.

[6] Joachims T (1997), Text categorization with support vector machines: Learning with many relevant features. Technical Report LS-8 Report 23, University of Dortmund, 1997.

[7] McCallum A and Nigam K (1998) A comparison of event models for naive Bayes text classification. In: Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05. AAAI Press, San Francisco, CA, July 1998, pp. 41-48.

[8] Koller D and Sahami M (1997), Hierarchically classifying documents using very few words. In: ICML 1997: Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, pp. 170-178.

[9] Chakrabarti S., M. Van Den Berg, Dom B.(1999), Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery, In: Proceedings of the 8th International WWW Conference, Toronto, Canada, http://www8.org/w8-papers/5a-search-query/crawling/index.html

[10] McCulloch, W. and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 7:115 - 133.

[11] John G, Kohavi R and Pfleger K (1994), Irrelevant features and the subset selection problem. In: Machine Learning: Proceedings of the Eleventh International Conference, Morgan Kaufman Publishers, San Francisco, CA, pp. 121-129.

[12] Gallant, S. I. (1990), Perceptron-based learning algorithms, IEEE Transactions on Neural Networks, 1, 179-191.

[13] Miguel e. Ruiz and Padmini Srinivasan (2002), Hierarchical Text Categorization Using Neural Networks, In: Information Retrieval 5, Kluwer Academic Publishers, pp. 87-118, 2002.