The role of the Clustering in the Field of Information Retrieval

Kristína Machová, Valentín Maták, Peter Bednár

Department of Cybernetics and Artificial Intelligence,
Technical University of Košice, Letná 9, 04200 Košice, Slovakia
Kristina.Machova@tuke.sk, valentin.matak@centrum.sk, Peter.Bednar@tuke.sk

Abstract: The paper describes possibilities of using clustering methods in information retrieval, particularly on text documents, which can be found on the Internet. The focus is on automatic extraction of information from texts including pre-processing of text documents. The paper presents also results of experiments, which were carried out using the 20NewsGroup collection of documents and Reuters- 21578 collection of documents. These experiments concern with k-means clustering and k-means clustering with controlled initialisation.

Keywords: clustering, machine learning, information retrieval, automatic extraction

1 Introduction

This paper presents some aspects of information retrieval [10] from web pages with the aid of machine learning. Web pages are considered the most common form of information representation. They can be found not only in worldwide Internet but are hidden in various (LAN, MAN, WAN) nets. Since information located on the web pages contains some level of noise, the application of pre-processing methods and selecting suitable representation are necessary. As far as representation is concerned, a suitable weighting text documents is important. The weighted and pre-processed text documents form a suitable input for classification or clustering methods of machine learning. Please use font type Times Roman CE only.

We used clustering machine learning methods. More information about machine learning can be found in [5], [6]. The quality of used clustering method can be measured with the aid of various coefficients calculated from the contingency table, e.g. precision coefficient. Precision π_j can be defined in following way:

$$\pi_j = \frac{TP_j}{TP_j + FP_j},$$

where TP_j (FP_j) is the number of correctly (incorrectly) predicted positive examples of the class c_j. In the frame of this work, we focused on classification of text documents from web pages. The most used methods for document classification are Naïve Bayes classifier, NBCI method [3], and kNN classifier. We performed tests using the kNN classifier (k Nearest Neighbours) [6], which is based on examples. This classifier stores in its memory all training examples – documents.

In our experiments, we focused on text document clustering [7]. We have used the k-means algorithm, which is defined in the following way. Let us assume n objects and k clusters. Each object represents a vector in d-dimensional space. Then each cluster can be represented as a centre of gravity of those objects, which belong to the cluster. One of disadvantages of this method is the risk of falling into a local minimum. This falling depends on the initial random selection of initial examples – documents. Two other disadvantages are the selection of the number of clusters and considering clusters as spheres in multidimensional feature space. The 1st mentioned disadvantage causes some sensitivity on changing coordinates, what is connected with used type of weighting. Better results can be achieved using a modification of the algorithm by employing the incremental actualisation of centres of clusters. There are some possibilities, how to cluster text documents with the aid of the Fuzzy k-means algorithm [9] or neural networks. [8] focuses on the optimisation of structures of neural nets.

2 Text Document Processing

This paper is mainly about information retrieval and information extraction from web pages. The purpose of this work is to transfer retrieved text information to the clusters, which represents domain of user interests. From the point of text document processing, we were interested in the selection suitable type of text documents. Before we discuss the used type of weighting of text documents, we want to mention, that we used the vector representation model. The words have various importance for document representation. That's why some relative value must be defined. This value - weight will represent the sense of the word. Resulting list of indexing terms can be ordered according to their weights – the information can be used while reducing the number of used terms. In this way the weights represent a selective force of the terms. This selective force expresses the ability of a term to find a subset of documents from the whole document corpus. This subset will differ from the subsets found by other terms. The term, which finds all documents from the corpus, has the minimum selective force. The

process of weight definition is called weighting. Various types of weighting have been tested in our work [4]. According this tests, we decided to used Sparck, Jones and Robertson weighting for following experiments with clustering method.

The automatic extraction consists of several steps: lexical analysis – token formation, elimination of words without meaning, lemmatisation and weighting. According to [2], transformation of documents using some standard specification is possible. Lexical analysis was performed in our tests by "Lower case filter" from the library "jbowl2" [1]. The elimination of words without meaning was made with the aid of "Stop words filter" from the library „jbowl2". Lemmatisation (stemming) was carried out by "Stem filter" from the library "jbowl2". Finally, weighting was accomplished by the "index filter" from the library "jbow2".

3 Experiments

In our experiments, 20 News Groups data set was used. 20 News Groups is a simple data set, which is composed from Internet discussion documents. It contains 19953 documents assigned (classified) to only one from twenty categories. Dimension of the lexical profile is 111474. Its advantage is nearly uniform distribution of documents into the categories and implicit classification to only one category. Division of this data set on the training and testing sets was realized by random selection using the proportion 1:1.

3.1 Clustering by K-Means on 20 News Group

In the case of information retrieval from various web pages, no categories are specified to which retrieved documents belong. These categories can be defined with the aid of clustering – a kind of unsupervised learning. Consequently, given documents can be classified to the categories - clusters. We have realized a set of experiments with clustering method *k*-means using documents from 20NewsGroups. Each of ten experiments started with a different initial number from the generator of random numbers (random seed). On the base of generated random numbers, some documents were chosen from the whole corpus. These documents become initial centres of clusters. The number of clusters to be formed was twenty (20 categories exist in the used corpus of documents). Stop-condition of the clustering algorithm was set to maximum number of iteration (50 iterations were used) and to epsilon (set to 0,1) expressing the difference between two subsequent values of error function. The stop-condition was a logical disjunction of both particular conditions. Documents from the corpus were weighted using the SJR weight function.

Figure 1 presents the achieved values of average precision related to individual categories (represented by wider bars) and positive standard deviation of precision according to individual categories (illustrated as narrow bars). Since the random initialisation was made, each cluster was initialised by a randomly selected document from the set of twenty categories. The worst case would be the case when all clusters would be initialised by the same document. The used generator of random numbers has sufficiently big period, so initialisation of not all but only a few clusters by documents of the same category happened in the practice. In the ideal case, the documents of particular categories should separate into individual clusters. The figure illustrates achieved precision with great dispersion – standard deviation oscillates in the interval <10,20> %, therefore the hypothesis about strong dependence on initialisation seems to be strongly supported. The category 9 has relatively high average precision but quite high standard deviation as well. On the other hand, categories 3 and 19 show lower precision and lower standard deviation as well.

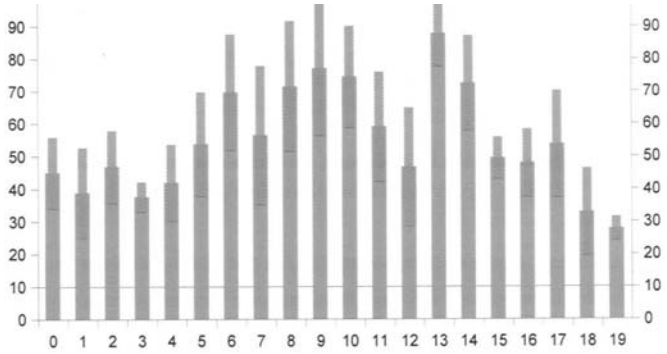

Figure 1
Average precision and standard deviation of the clustering by k-means algorithm according to categories

3.2 Clustering by K-Means with Controlled Initialisation on 20 News Group

This set of experiments is parametrically identical to the previous experiments. The only difference is that now the initialisation is not random but controlled. Our system initialises the **i**-th cluster by an example which represents an average of ten randomly selected examples belonging to the **i**-th category.

Figure 2 presents the achieved values of average precision related to individual categories (represented by wider bars) and positive standard deviation of precision according to individual categories (illustrated as narrow bars). The results have

proven the importance of controlled initialisation for decreasing standard deviation.

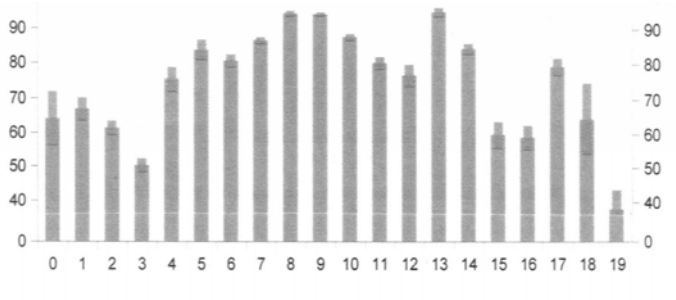

Figure 2
Average precision and standard deviation of the clustering by k-means algorithm with controlled initialisation according to categories

The precise values of average precision and standard deviation of the clustering by k-means algorithm without and with controlled initialisation presented in Figure 1 and Figure 2 can be found in Table 4.

Table 4
Average precision and standard deviation of the clustering by k-means algorithm without and with controlled initialisation

	Clustering without controlled initialisation		Clustering with controlled initialisation	
Category	Precision	Standard deviation	Precision	Standard deviation
1	44,96	10,96	64,09	7,86
2	38,87	13,94	66,91	3,37
3	46,80	11,11	61,40	2,14
4	37,49	04,69	50,40	1,89
5	41,89	11,78	75,44	3,47
6	53,68	16,09	83,81	2,92
7	69,61	17,94	80,69	1,88
8	56,48	21,39	86,62	1,00
9	71,44	20,13	94,41	0,85
10	77,04	20,96	94,20	0,59
11	74,43	15,72	87,62	0,93
12	59,01	16,96	80,23	1,81
13	46,65	18,27	76,72	3,07
14	87,78	10,11	94,94	1,41

15	72,51	14,82	84,35	1,60
16	49,43	06,54	59,76	3,84
17	47,87	10,42	58,98	3,50
18	53,69	16,55	79,22	2,51
19	32,87	13,49	64,35	10,41
20	27,66	03,83	37,89	5,79

Conclusions

The paper presents fundament of clustering and refines it according to requirements of the domain of text document classification. The most important findings and facts, which were deducted from the algorithm implementation and testing are presented in the experimental part. Comparison of results of the experiments with clustering method *k*-means and k-means with controlled initialisation is described.

Clustering is suitable for application in electronic information systems, for library applications, applications for design and realisation of Internet crawling – realisation of structural search, automatic actualisation of catalogues, search for mirror pages and pages located on other URLs after their migration, elimination of very similar search results to a given question, automatic detection of plagiarism and so on.

Some questions remain open, for example the question of cluster labelling and documents clustering with implicitly assigned more than one category (fuzzy *k*-means, cluster overlapping). The presented work deals with only a small part from the domain of information extraction from web-pages using machine learning methods.

Acknowledgement

The work presented in the paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/1060/04 project ”Document classification and annotation for the Semantic web”.

References

[1] Bednár, P.: API Java knižnice HTML Parser: http://sourceforge.net/projects/jbowl

[2] Kolár, J., Samuelis, L., Rajchman, P.: Notes on the Experience of Transforming Distributed Learning Materials into Scorm Standard Specifications. Advanced Distributed Learning. Information & Security. An International Journal. Vol. 14, ProCon Ltd., Sofia, 2004, 81-86, ISSN 1311-1493

[3] Kučera, M., Ježek, K., Hynek, J.: Kategorizace textů metodou NBCI. Katedra informatiky a výpočetní techniky, Západočeská univerzita, Plzeň, 2004

[4] Machová, K., Maták, V., Bednár, P.: Information Extraction from the Web Pages Using Machine Learning Methods. Proc. of the IIS – Information and Intelligent Systems – 16th International Conference, 21-23 September 2005, Varaždin, University of Zagreb, Croatia, 2005, 407-414, ISBN: 953-6071-25-8

[5] Machová, K.: Machine Learning. Principles and algorithms. ELFA s.r.o., 2002, Košice, 117s., ISBN 80 89066 51 8

[6] Mitchell, T. M.: Machine Learning. McGraw-Hill Companies, Inc., Singapore, 1997, 414 ps., ISBN 0-07-042807-7

[7] Muresan, G., Harper, D. J.: Document Clustering and Language Models for System-Mediated Information Access. Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries ECDL'01, Darmstad, September 2001, ISBN 3-540-42537-3

[8] Olej, V., Křupka, J.: A Genetic Method for Optimization Fuzzy Neural Networks Structure. International Symposium on Computational Intelligence, ISCI 2000, Advances in Soft Computing, The State of the Art in Computational Intelligence, A Springer–Verlag Company, Germany, 2000, pp.197-202, ISSN 1615-3871, ISBN 3-7908-1322-2

[9] Song, D., Cao, G., Bruza, P.: Fuzzy K-means clustering in information retrieval, Information Ecology. Distributed Systems Technology Centre, The University of Queensland, QLD 4072, Australia, 28 July 2003

[10] Van Rijsbergen C. J. (1979): *Information Retrieval.* Department of Computing Science, University of Glasgow