Chi-square Classifier for Document Categorization

Mikhail Alexandrov [1], Alexander Gelbukh [1], and George Lozovoi [2]

[1] Center for Computing Research, IPN, Mexico
{dyner, gelbukh}@cic.ipn.mx

[2] Datagistics, Canada
glozovoi@idirect.com

Abstract. The problem of document categorization is considered. The set of domains and the keywords specific for these domains is supposed to be selected beforehand as initial data. We apply the well-known statistical hypothesis test that considers images of documents and domains as normalized vectors. In comparison with existing methods, such approach allows to take into account a random character of initial data. The classifier is developed in the framework of Document Investigator software package.

1 Introduction

Various keyword-based technologies are suggested for document categorization nowadays. In particular, these technologies use the tree of concepts for searching the principal theme of a document [3]; the concentration of keywords to evaluate the contribution of given domains to a document [1]; the probabilities of domain and keywords in Bayes classifiers [4], etc. Practically all technologies use the results of preliminary training. In comparison with all mentioned approaches, the suggested classifier needs more limited information for decision-making and gives a numerical estimation of reliability of results, taking into account the random character of the data used for training and application of the classifier. The classifier tests the statistical hypotheses based on c^{2} -distribution and builds a list of domains relevant to a document with a given probability of documents missing a relevant domain.

2 Initial Data for Classification

Initial data is the list of domains D_j, $j = 1, \ldots, m$; the list of keywords (key-expressions) w_i, $i = 1, \ldots, n$; and the matrix of conditional frequencies $p_{ij} = p(w_i \mid D_j)$ reflecting distribution of the keywords in these domains. Let N_j, $j = 1, \ldots, m$ be the numbers of keyword occurrences in all the documents connected with the domain D_j in full training database, so that $\sum_i p_{ij} = N_j$. We use the term *keyword* to refer any key-expression that can be a single word a word combination. What is more, we take as the same keyword any group with the same stem and meaning. For example, *obli-*

gation, obligations, obligatory, oblige have the same stem *oblig*. Each document can be considered as a vector (x_1,x_2,...,x_n) of keywords, where x_i is the number w_i of keyword occurrences in the document, and $\sum_i x_i = N$ ·

For formal considerations we accept that *the document topic is a direction of the document image vector in the multidimensional space of keywords.* According to this definition all documents reflecting the same topic have parallel or quasi-parallel vectors. Indeed, let us consider a concatenation of *l* copies of the given document *D*. Naturally, it has the same topic as *D*, while its image (lx_1, lx_2, ..., lx_n) is parallel to that of *D*. On the other hand, let us consider a document *D'* that has no relation to the domain under consideration, and attach it to source one *D*. Naturally, the resulting text *D* + *D'* has the same topic with respect to the domain under consideration as *D*, while it has the same image which thus is parallel to that of *D*.

We will consider now the vector of conditional frequencies p_j=(p_{1j}, p_{2j}, ..., p_{nj}), j = 1, 2, ..., *m* mentioned above as the image of a typical document from the domain D_j. Because an operation of normalization does not change vector direction we transform all our frequencies so that: $\sum_i p_{ij} = \sum_i x_i = 1$·

3 The Main Algorithm

3.1 Testing of Hypotheses

The algorithm consists in consequent test of hypotheses about belonging of a document to given domains. Considering image of a document and images of domains as attributes such a test is reduced to the test of uniformity. According to [2] it can be completed by c^2-criterion of uniformity with $n = n$-1 degrees of freedom. The criterion requires a calculation of the series of values c_j^2 as is shown in (1a):

$$c_j^2 = \sum_i \frac{1}{\frac{x_i}{N_j} + \frac{p_{ij}}{N}} \left(x_i - p_{ij}\right)^2 \quad \textbf{(1a)} \qquad c_{kl}^2 = \sum_i \frac{1}{\frac{p_{ik}}{N_l} + \frac{p_{il}}{N_k}} \left(p_{ik} - p_{il}\right)^2 \quad \textbf{(1b)}$$

Then the inequality $c_j^2 \le c_p^2$ is verified for all j = 1,2, ..., *m*, where *p* is the level of hypothesis significance and c_p^2 : $p = P(c^2 > c_p^2)$. All hypotheses for which this inequality holds are accepted. The value *p* is the feasible probability to miss a domain which is in fact relevant for the document.

3.2 Analysis of Errors of Decision

Testing statistical hypothesis we deal with an error to miss a relevant domain. But when several hypotheses are accepted we immediately deal with the other kind of error, which is an analogue of false alarm but for the case of many alternatives. Indeed, because a document is supposed to reflect not more than one domain then among the selected domains only one is relevant. These two kinds of errors prove to be connected by the same way as in the case of two alternatives, namely the less a probability to miss a domain the more hypotheses will be accepted and the more is a number of false domains that appear.

The concrete dependence of these errors is defined by domain distinguishability. The last can be evaluated checking the significance of distinctions between domains on (1b) that is an analogue of (1a) for two domains D_k and D_l. Then it is possible to find two boundary values:

$$p_{max}:\ c^{2}_{p_{max}} = \min_{k,l}\ c^{2}_{kl}\ ; \qquad p_{min}:\ c^{2}_{p_{min}} = \max_{k,l}\ c^{2}_{kl}\ ; \qquad \textbf{(2)}$$

The following considerations follow directly from (2):

1. If $p \quad p_{max}$ then the classifier is able to find not more than 1 relevant domain for a document. In this case the probability of false alarm is 0%.
2. If $p < p_{min}$ then the classifier is able to find at once all domains relevant to a document (or none). In this case the probability of false alarm is (1-1/m)*100%.
3. If $p_{min} \quad p < p_{max}$ then the classifier is able to find k domains, k = 2, ..., m-1 (or none). In this case the probability of false alarm is (1-1/k)*100%.

The typical values for probability to miss a correct domain are 5%-10%. But it would be better to assign this probability taking into account also the mentioned probability of false alarm.

4 Conclusions and Future Work

We suggested a simple keyword-based classifier intended for selection of domains relevant to a text document. The introduction of well-known statistical errors of the first and the second kind allows to set an optimum mode of classifier work. The future development will consist in: a) performing numerous experiments with a real document flow, b) constructing more complex decision rules with respect to accepted hypothesis, c) developing a user-oriented program system.

References

1. Alexandrov, M., Gelbukh, A., and Makagonov, P. *Some keyword-based characteristics for evaluation of thematic structure of multidisciplinary documents.* Proc. of 1st Int. Conf. on Intelligent Text Processing and Computational Linguistics, Mexico City, 2000, pp. 390-401.
2. Cramer, H. *Mathematical methods of statistics*. Cambridge, 1946.
3. Guzman-Arenas, A. *Finding the main themes in a Spanish documents.* Intern. J. of Expert Systems with Applications, 1998, v. 14, N 1/2, pp. 139-148.
4. Mitchel, T. *Machine learning*. New-York, McGraw Hill, 1997.