# Comparing Natural Language Identification Methods based on Markov Processes*

Peter Vojtek and Mária Bieliková

{pvojtek, bielik}@fiit.stuba.sk
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology

**Abstract.** We discover and experiment with categorization-based methods to natural language identification. Two approaches to language identification based on Markov processes are compared, both methods treat the incoming text on the character level. We performed series of experiments with the aim to make certain of high precision in language identification task of selected methods and also with the objective to compare them against themselves. Experimental evaluation was based on large-scaled Multilingual Reuters Corpus with various European and Slavic languages. Our research results showed that both methods are comparable in the task of natural language identification achieving recall as high as 99,75%.

## 1 Introduction

Natural language identification is the process of automated labeling textual documents by their language (e.g. this paper should be labeled as written in English). Although exact definition of the term *natural language* is not formed, the term covers languages used by humans for common communication (like Slovak or English), as a opposite of artificial languages (e.g. C++, Java).

Exploration of automated language identification is usually motivated by simplifying document preprocessing and organization of information, this is also the case of our research, which is involved in a project affiliating methods and tools for acquisition, organization and maintenance of information and knowledge in an environment of heterogeneous information resources[1].

As many language identification approaches exists (see survey by Cole et al. [1]), we point out our main demands with the aim to determine the proper identification method:

- *efficient* – capable to process large number of documents in real-time
- *language independent* – process text quantitatively in contrast to methods based on language specific features

* This work was partially supported by the State programme of research and development "Establishing of Information Society" under the contract No. 1025/04.

[1] Project NAZOU – http://nazou.fiit.stuba.sk/

– *document format independent* – identify language directly from text of document and not rely on meta-information bound with this document (which can be missing or incorrect)

Enlisted demands can be fulfilled by language identification method based on statistical modeling of text. A text modeling technique used by selected identification method should not make use of whole words or even sentences, rather putting stress on the lower level of granularity, hence chains of characters of text should be regarded. According to the mentioned requirements, two techniques satisfy our demands: Markov processes and the N–gram analysis. While we realized experiments with N–gram methods in our previous work [2], in this paper we explore, compare and improve two language identification methods based on Markov processes designed by Dunning [3] (Statistical identification of language) and Teahan [4] (Text classification and segmentation using minimum cross entropy). In the rest of the paper we will refer to these methods by their author's name.

The major contributions of this work are (1) theoretical and experimental comparison of two concurrent Markov processes based language identification methods using large-scaled Reuters Corpora, and (2) enhancement of the process of evaluating the best matching language in the identification phase by normalization by document length, which extends the scenarios of use of both methods.

The rest of the paper is structured as follows. Overview of related work is proposed in Section 2. Identification methods based on Markov processes are explained in Section 3. Proposal of additional castigation of categorization methods using normalization is in Section 4. After that, we report out experimental results aimed at comparison of the language identification methods in Section 5. Finally, Section 6 concludes the paper and points out some issues requiring further work.

## 2 Related Work

One of the simplest approaches to language identification is based on common words and unique combinations of characters [5]. This approach works quite well for large documents, but fails when the incoming textual information in getting smaller (e.g. document containing only one sentence).

Another way of language identification is to use N–grams. On of the most cited method is designed by Cavnar and Trenkle [6], based on list of the most frequently observed N–grams (i.e. sequences of characters), variable N–gram length is used. Suzuki et al. [7] discovers a methods based on N–grams capable to identify language and character encoding together. We experimented with this method using Slavic languages and character encodings in [2].

Many other language identification methods are derived from universal categorization methods, e.g. Naive Bayes, Support Vector Machines [8] or k-Nearest Neigbour [9]. Survey by Aas and Eikvil contains overview of these categorization methods, tools and linguistic corpora [10]. The drawback of these methods is

that the text is usually represented as a bag-of-words and language specific preprocessing as stop-word removal or stemming is necessary, another disadvantage is that the feature space is usually large and must be reduced, although feature space reduction methods based on Information Gain, Principal Component Analysis [11] or Collaborative Filtering [12] are already well explored.

## 3 Language Identification Methods based on Markov Processes

Both language identification methods use the well known supervised learning schema [13]. Statistical model is created for each language in the learning phase. Each language model is constructed from pre-selected training text. Then identification phase can be proceeded, documents to be identified are passed and language tags are assigned to them. The best-fitting language model for each document is determined by an evaluation function.

While the Markov processes theory serves as the basis for both language identification methods, we shortly describe this theory first. Further reading on probabilistic modeling of text can be found in [14].

### 3.1 Markov Processes as Text Modeling Tool

Stochastic process is called the first order Markov process if its state $c_k$ in time $k$ depends only on previous state $c_{k-1}$ in time $k-1$ (Formula 1).

$$P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-1}). \tag{1}$$

In general, $n$-th order Markov process is described in Formula 2.

$$P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-n}, \ldots, c_{k-1}). \tag{2}$$

The character sequence $c_{k-n}, \ldots, c_{k-1}$ is named as Markov process $prefix$ (also the term *context* is used), $c_k$ is usually named $suffix$.

### 3.2 Dunning's Language Identification Method

**Learning Phase – Creating Models of Language Categories** A training text document (representative of particular language) is processed as a stream of characters. This stream is divided into Markov processes with length $k$ characters ($k$ is the order of Markov process). Each unique Markov process is stored together with information about its number of occurrences. After processing the whole document, all Markov processes counts are converted into probabilities using Formula 3 ($k$-th order Markov processes).

$$p(w_1 \ldots w_{k+1}) = \frac{T(w_1 \ldots w_{k+1}) + 1}{T(w_1 \ldots w_k) + |A|} \tag{3}$$

where $|A|$ is the size of an alphabet, $T(w_1 \dots w_k)$ is number of occurrences of Markov process prefix, $T(w_1 \dots w_{k+1})$ is number of occurrences of the whole Markov process and $p(w_1 \dots w_{k+1})$ is the computed probability.

As an example, processing the text "abracadabra" into Markov processes of order $k = 1$ is in Table 1, $c$ is the number a particular Markov process occurred and $p$ is the probability computed using Formula 3.

**Table 1.** Processing the text "abracadabra" into 1st order Markov processes.

| Order k = 1 | | |
|---|---|---|
| Predictions | c | p |
| a → b | 2 | 2/5 |
| → c | 1 | 1/5 |
| → d | 1 | 1/5 |
| b → r | 2 | 2/2 |
| c → a | 1 | 1/1 |
| d → a | 1 | 1/1 |
| r → a | 2 | 2/2 |

**Identification Phase** Language category models (i.e. persistently stored Markov processes bound with their occurrence probabilities) are created for each language in the learning phase. In the identification step, evaluation function is applied to the input text for each language model (Formula 4) and the best matching language model is determined.

$$log\ p = \sum_{w_1 \dots w_{k+1} \in S} T(w_1 \dots w_{k+1}) log\ p(w_{k+1}|w_1 \dots w_k) \tag{4}$$

where $T(w_1 \dots w_{k+1})$ are the number of occurrences of all Markov processes present in the text and $p(w_{k+1}|w_1 \dots w_k)$ is the probability stored in a particular model for each Markov process. While the model can handle only already observed Markov processes, yet unobserved processes on the input are skipped in this phase. Logarithm scaling is used due to avoiding problems of numeric underflow.

Overall probabilities computed for each language category by evaluation function are compared between themselves and the model with a result closest to zero is the best fitting.

### 3.3 Teahan's Language Identification Method

Although the supervised learning schema and Markov processes are also used in this language identification methods, the process of creating the language models and evaluating the best fitting model differs. While this identification method is

more sophisticated and complex, detailed description of the method is beyond the scope of this paper. Deeper explanation and discussion can be found in [4] and [15].

**Learning phase** Dunning's identification method stores Markov processes only of particular length $k$, language models adopted in this method use various length of Markov processes together. Theoretically, this approach brings smoother modeling of a text.

At first, all Markov processes and their counts are extracted from training text. Dunning's language identification method extracts only Markov processes of exact order $k$, Teahan's approach takes into account also all lower orders Markov processes $k-1$, $k-2$, ..., 0 and $-1$ (Note that Markov process of order 0 is the distribution of separate characters in a text and Markov process of order -1 is the estimated distribution of all characters that did not appeared in the training text).

Next, Markov processes counts are converted into probabilities (Formula 5). While different orders of Markov processes are used, "escape" probability mechanism is involved, providing switching from higher orders of Markov processes to lower. Escape probabilities are important when Markov process of length $k$ occurs on input and this process cannot be found in the highest order model table (length $k$). In this case, order of model is decreased to $k-1$, actual Markov process on input is also shortened and this overstepping between different process orders is count in with relevant escape probability.

$$e = \frac{t}{n+t} \text{ and } p(\phi) = \frac{c(\phi)}{n+t} \tag{5}$$

$c(\phi)$ is the number of times a particular prefix of Markov process was followed by the character $\phi$, $n$ is the number of all tokens that have followed and $t$ is number of unique characters that have followed. $e$ is the escape probability and $p(\phi)$ is probability for particular character.

Processing of the text "abracadabra" using method based on Teahan's method is displayed in Table 2, involving Markov processes of orders 1, 0 and -1.

**Identification phase** Models of selected languages are already created in the learning phase. Document written in yet unknown language is processed as a stream of characters and Markov processes of all lengths from $k$ to 0 are extracted. For each model of language and set of all Markov processes present in input text, a cross entropy is computed (6). The language model which has the value $H(M)$ closer to zero is chosen as the best fitting and input document is labeled as written in this language.

$$H(M) = -\sum p_M(w_1, ..., w_m)\ log\ p_M(w_1, ..., w_m) \tag{6}$$

The probability for character model of length $k$ is determined using Formula 7.

**Table 2.** Processing text "abracadabra" using Teahan's method.

| Order k = 1 | | | Order k = 0 | | | Order k = -1 | | |
|---|---|---|---|---|---|---|---|---|
| Predictions | c | p | Predictions | c | p | Predictions | c | p |
| a → b | 2 | $\frac{2}{7}$ | → a | 5 | $\frac{5}{16}$ | → A | 1 | $\frac{1}{\|A\|}$ |
| → c | 1 | $\frac{1}{7}$ | → b | 2 | $\frac{2}{16}$ | | | |
| → d | 1 | $\frac{1}{7}$ | → c | 1 | $\frac{1}{16}$ | | | |
| → Esc | 3 | $\frac{3}{7}$ | → d | 1 | $\frac{1}{16}$ | | | |
| | | | → r | 2 | $\frac{2}{16}$ | | | |
| b → r | 2 | $\frac{2}{3}$ | → Esc | 5 | $\frac{5}{16}$ | | | |
| → Esc | 1 | $\frac{1}{3}$ | | | | | | |
| c → a | 1 | $\frac{1}{2}$ | | | | | | |
| → Esc | 1 | $\frac{1}{2}$ | | | | | | |
| d → a | 1 | $\frac{1}{2}$ | | | | | | |
| → Esc | 1 | $\frac{1}{2}$ | | | | | | |
| r → a | 2 | $\frac{2}{3}$ | | | | | | |
| → Esc | 1 | $\frac{1}{3}$ | | | | | | |

$$p_M(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} p'(w_i | w_{i-k} \dots w_{i-1}) \tag{7}$$

$p'$ gives the probability returned by model of order $k$.

Although the escape mechanism (described in the learning phase of this method) helps deal better with already observed Markov processes (or their sub-processes), when yet unobserved Markov processes is present on input and its first character does not matches first character of any highest level Markov process in a language model, the escape mechanism cannot be applied and the actual Markov process must be skipped.

## 4 Normalization of the Evaluation Function by Document Size

In some cases, we are not aimed at identification of many languages, but only of one exact language – e.g. we have a set of documents written in many languages and we want to filter out only those written in Slovak (note that we even may not exactly know which languages are present in the document set, thus we cannot create models for all languages). In the current state, both language identification methods are not designed to deal with this problem, while they always assign a language label to the input document in the identification phase – when only Slovak language model will be created in the learning phase, all documents from the document set will be labeled as Slovak.

We can deal with this problem by involving normalization of evaluation function by document text length, which enables us to divide the document-space explicitly into two sub-spaces: a sub-space containing documents written in Slovak language and a sub-space where are non-Slovak document. Evaluation functions of both language identification methods are normalized using Formula 8.

$$F(language\ model,\ input\ text)_{norm} = \frac{F(language\ model, input\ text)}{\#_{chars}(input\ text)} \quad (8)$$

$\#_{chars}(input\ text)$ is the number of characters in input text. Note that different approaches to normalization are known (Overview of alternative approaches to normalization is in work of Singhal et al. [16]). In Formula 8, normalization by the number of characters in a text is used while character encoding independence is achieved.

## 5 Experimental Evaluation

The main goal of our experiments is to determine, if more precision modeling of a text involved by the identification method proposed by Teahan (described in Section 3.3) brings better results in language identification when compared with results of the method proposed by Dunning (Section 3.2). Second experiment investigates, if the normalization of an evaluation function (in both language identification methods) allows to separate the state space between models of languages.

### 5.1 Language Identification

Comparison of the language identification methods was performed using eight European languages from the Multilingual Reuters Corpus[2] – Danish, German, Spanish, French, Italian, Norwegian, Portuguese and Swedish. Many models were created for each language with the aim to compare the identification methods in various conditions. Different granularity of modeling of the text was achieved by using 1st, 2nd, 3rd and 4th Markov process orders were (larger orders of Markov processes were not used due to memory limitations), the amount of learning text varied from 25 kB to 200 kB.

After the learning phase was accomplished, 2 000 testing documents for each language were passed and the ability of the language identification methods to correctly label the testing document was measured. Average size of the testing documents in the corpus is 1,2 kB, Figure 1 displays the histogram of the testing documents.

Results of the language identification in Table 3 and 4 contains averaged values for all languages.

[2] Reuters Corpora – http://trec.nist.gov/data/reuters/reuters.html

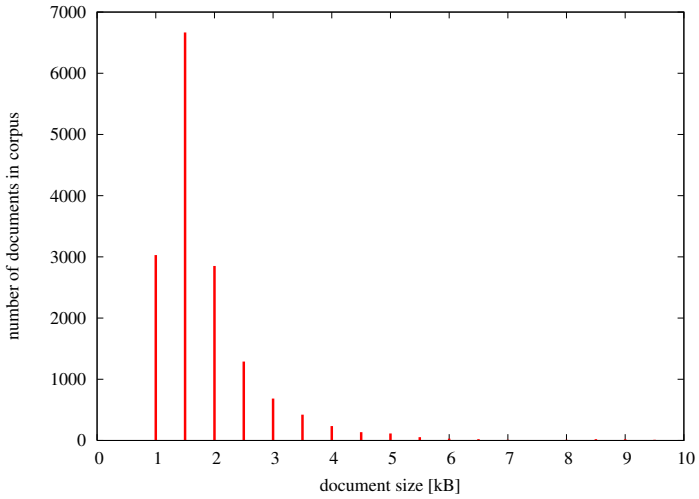


**Fig. 1.** Document size distribution in the Multilingual Reuters Corpus.

**Table 3.** Language identification method proposed by Dunning, *Recall* values.

| Markov process order | training text length [kB] / *Recall* [%] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *25* | *50* | *75* | *100* | *125* | *150* | *175* | *200* |
| 1st | 97.09 | 97.95 | 98.66 | 98.83 | 98.88 | 98.96 | 99.04 | 99.20 |
| 2nd | 97.86 | 99.14 | 99.43 | 99.50 | 99.52 | 99.56 | 99.65 | 99.62 |
| 3rd | 98.05 | 99.13 | 99.48 | 99.55 | 99.56 | 99.59 | 99.65 | 99.67 |
| 4th | 97.67 | 8.83 | 99.08 | 99.28 | 99.36 | 99.61 | 99.63 | 99.70 |

**Table 4.** Language identification method proposed by Teahan, *Recall* values.

| Markov process order | training text length [kB] / *Recall* [%] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *25* | *50* | *75* | *100* | *125* | *150* | *175* | *200* |
| 1st | 97.22 | 98.20 | 98.74 | 98.88 | 99.00 | 99.15 | 99.37 | 99.38 |
| 2nd | 97.92 | 99.07 | 99.30 | 99.45 | 99.56 | 99.64 | 99.71 | 99.75 |
| 3rd | 89.16 | 97.52 | 98.85 | 99.12 | 99.43 | 99.50 | 99.64 | 99.67 |
| 4th | 64.79 | 62.60 | 62.94 | 62.01 | 70.37 | 77.47 | 81.03 | 84.20 |

Comparison of the results in Tables 3 and 4 shows that both methods can deal very well when language models consisting of 1st, 2nd and 3rd Markov process order are used. Although the highest value of recall (99.75%) was achieved using the more sophisticated identification method proposed by Teahan, results shows that this method performs significantly worse when 4th of Markov processes are used. This degradation is present when the language model must treat yet unobserved Markov processes, which harms the Teahan's method more significantly. Unfortunately, adopting some mechanism to avoid this degradation will make already very complicated identification method even more complicated.

### 5.2 Normalization of Evaluation Function

The aim of this experiment is to determine, if the involvement of the normalization can clearly distinguish between languages, even when very similar languages are taken into account. If this hypothesis turns to be true, it will be possible to decide explicitly, where the boundary between languages lies, enabling us to avoid of incorrect labeling of documents written in not learned languages (as described in Section 4). The procedure is the same for both methods – only one language model is created (Slovak language) in the learning phase. Novels in Slovak, Czech and Polish language are evaluated in the identification phase.

Fig. 2 shows results for Dunning's identification method, Teahan's methods is evaluated in Fig. 3. Averaged values of the evaluation function are enhanced by standard mean value, where $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$. 95,4% of all documents of particular size should fall into the interval (assuming the Gaussian distribution).

The Y axis displays the normalized value of evaluation function applied in the identification phase. While only model of the Slovak language was involved, testing documents written in Slovak language naturally score best.

The results are similar for both methods – when documents smaller than 1 000 bytes are processed, documents written in Czech language are in many cases incorrectly labeled as Slovak. This is caused by the fact that Slovak and Czech languages are very similar, such a problem does not occurs when documents in Polish language are passed (e.g. value of $normalized\ evaluation\ function = 4.5$ safely divides Slovak and Polish documents in Fig. 3). The conclusion of this experiment is that when documents written in very similar languages are expected on input, the use of explicit division of state space should be carefully considered.

## 6 Conclusion and Further Work

We explored and compared two language identification methods based on Markov processes in this paper. Although method proposed by Teahan [4] is more complex than Dunning's identification method [3], our experiments based on Reuters Corpora and novels in Slavic languages showed that both methods treat the language identification task in similar way, achieving recall as high as 99,75%. We

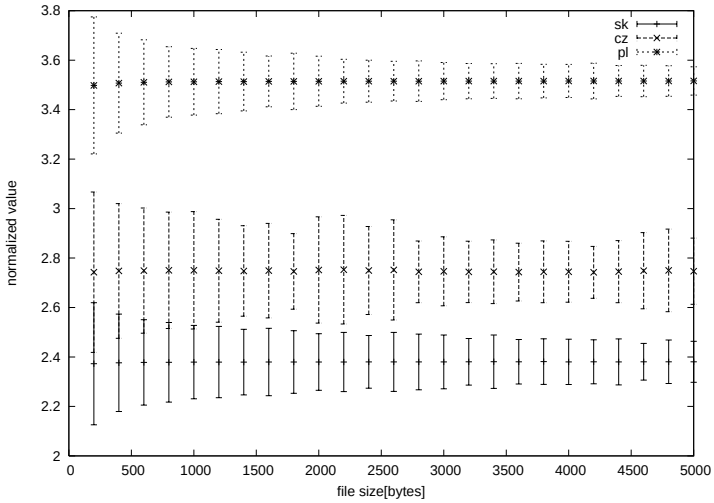


**Fig. 2.** Normalized evaluation function, Dunning's identification method.

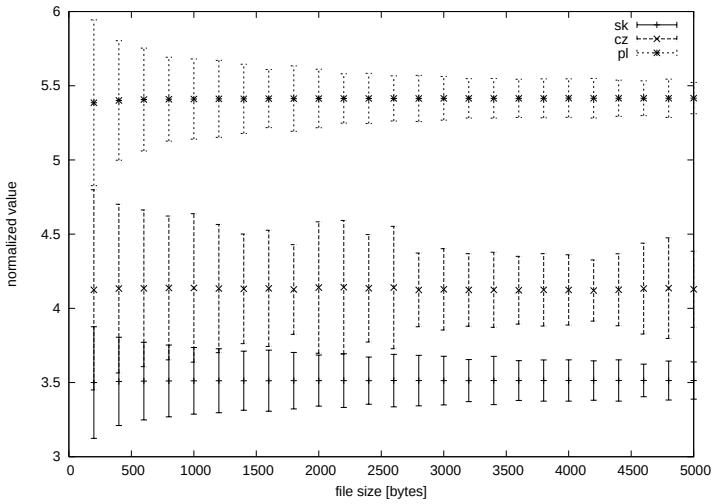


**Fig. 3.** Normalized evaluation function, Teahan's identification method.

improved the identification methods by involving normalization of evaluation function in the identification phase, enhancing the area of application of both methods.

Thanks to satisfactory results, Markov processes based identification methods served as the basis for a software tool incorporated into larger project affiliating tools for acquisition, organization and maintenance of information and knowledge in an environment of heterogeneous information resources [17]. This research project is experimentally evaluated in the domain of job-offers, our language identification tool serves in following ways – language identification of job-offers, document categorization [18] and semantic annotation [19].

Further work should focus on exploring the impact of character level text modeling (e.g. Markov processes, N–grams) in the task of general categorization, as a opposite to traditional bag-of-words representation. Already accomplished experiments include: subject classification [6], authorship categorization[4], genetic sequences classification [3] and we executed some preliminary research in categorization of job-offers [20] in Slovak language.

## References


1. Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V.: Survey of the state of the art in human language technology (1995)
2. Vojtek, P.: Natural Language Identification in the World Wide Web. In Bieliková, M., ed.: IIT.SRC 2006: Student Research Conference, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava (2006) 153–159
3. Dunning, T.: Statistical identification of language. Techical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University (1994)
4. Teahan, W.J.: Text classification and segmentation using minimum cross entropy. In: Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur", Paris, FR (2000)
5. Grefenstette, G.: Comparing two language identification schemes. In: Proceedings of JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data. (1995)
6. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US (1994) 161–175
7. Suzuki, I., Mikami, Y., Ohsato, A., Chubachi, Y.: A language and character set determination method based on n-gram statistics. ACM Transactions on Asian Language Information Processing (TALIP) **1**(3) (2002) 269–278
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3) (1995) 273–297
9. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: 22nd Annual International SIGIR, Berkley (1999) 42–49
10. Aas, K., Eikvil, L.: Text categorisation: A survey (1999)
11. Zhang, R., Shepherd, M., Duffy, J., Watters, C.: Automatic web page categorization using principal component analysis. hicss **0** (2007) 73a
12. Song, Y., Zhou, D., Huang, J., Councill, I.G., Zha, H., Giles, C.L.: Boosting the feature space: Text classification for unstructured data on the web. In: ICDM '06:

Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2006) 1064–1069
13. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer (2006)
14. Baldi, P., Frasconi, P., Smyth, P.: Modeling the Internet and the Web: Probabilistic Methods and Algorithms. Wiley (2003)
15. Teahan, W.J., Harper, D.J.: Combining ppm models using a text mining approach. In: DCC '01: Proceedings of the Data Compression Conference (DCC '01), Washington, DC, USA, IEEE Computer Society (2001) 153
16. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Research and Development in Information Retrieval. (1996) 21–29
17. Návrat, P., Bieliková, M., Rozinajová, V.: Acquiring, organising and presenting information and knowledge from the web. In: Proc. of Int. Conf. on Computer Systems and Technologies - CompSysTechŠ06, Varna, Bulgaria (2006)
18. Gatial, E., Balogh, Z., Laclavík, M., Ciglan, M., Hluchý, L.: Focused web crawling mechanism based on page relevance. In Vojtáš, P., ed.: Proc. of ITAT 05, Workshop on Theory and Practice of IT, Račková dolina, Slovakia (2005)
19. Laclavík, M., Šeleng, M., Gatial, E., Balogh, Z., Hluchý, L.: Ontology based text annotation ontea. In et.al., Y.K., ed.: Proc. of 16th European-Japanese Conf. on Information Modelling and Knowledge Bases, EJC 06, Paris, FR (2006)
20. Vojtek, P.: Improving Text Categorization Based on Markov Processes. In Bieliková, M., ed.: IIT.SRC 2007: Student Research Conference, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava (2007) 217–224