Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Tools for Texts Monitoring and Analysis Aimed at the Field of Social Disasters, Catastrophes, and Terrorism *

Svetlana Cojocaru, Mircea Petic and Grigorii Horos

Abstract

Throughout its life mankind faces various disasters and catastrophes: natural, technogenic, social. One of the important sources of information for these situations is the huge volume of unstructured data available in the global information networks. In this study, we describe a tool set that includes methods for extracting relevant texts from the networks, their classification and analysis. Two stages are described: preparatory and processing. The first one deals with patterns (texts and keywords) creation, during the second phase news texts are processed using database and controlled vocabulary.

Keywords: computational linguistic resources, linguistic markers.

1 Introduction

We live in an era that is increasingly affected by various natural disasters (earthquakes, forest fires, floods), together with technogenic disasters (explosions, leakage of toxic substances) or caused by individuals, such as terrorism.

A characteristic feature of contemporary society is the special role of information networks, where different signals related to disasters that may be produced or already have been produced, may occur promptly.

* This research is performed in the frame of a grant (SPS G4877) supported by the NATO international Science for Peace and Security Program.

This source of information can be used to prevent and mitigate the social consequences of disasters. It consists from a large volume of data, accessible on global information networks: mass-media, social networks, blogs, etc. A social disaster is usually followed by a huge amount of data generated in the form of news, discussion, expression of views etc. This information is quite difficult to process due to its unstructured form. In order to make right decisions in appropriate time, special analytical tools are needed that would give decision-makers support for a second opinion. Currently, such means practically do not exist, with the exception of guidance from a narrow range of applications (e.g., forest fire monitoring and forecasting). Achieving their implementation could significantly improve modelling and social disaster mitigation.

Our goal is to elaborate a tool which will collect and classify tweets and news texts related to disasters topics.

This article is carried out within a project [1] that aims creating a set of tools, methods and algorithms to detect different nature of social disasters. The main idea is to monitor information sources from network in three neighboring countries (Ukraine, Romania and Republic of Moldova), to find pertinent texts in four languages: Ukrainian, Russian, Romanian and English, to process them at Situation Analytical Center established in Kiev to suggest the appropriate resilience scenarios to decision makers. At this step of research, we intend to process a large number of Romanian text data that is available on the Internet and is stored in an unstructured manner.

In the following sections, we will describe methods of texts sources processing that permit extracting markers of social disasters. To achieve this, we will gather an amount of data, with the intention to generate a lexicon afterwards, which comprises relevant words with reference to technological, social or natural disasters. Subsequently, based on these markers, we will form a lexicon that serves as basis for our future system of identification and classification of texts from the Internet.

The article consists of several sections. Section 2 introduces several related works that describe the state of the art in our research field.

Section 3 describes the general approach, emphasizing two processing stages. Section 4 presents the collection of articles obtained manually. Section 5 describes the stages of lexicon of markers creation. Sections 6 and 7 are concerned with automation of texts collection and streamline of their processing. The article ends with conclusions and some directions for future works.

2 State of the Art

There are many papers that analyse the topic of social disasters warning and decision making. Social media is considered to be a quick information propagation tool for being informed about recent human kind catastrophes [2]. That is why it is frequently used for disaster monitoring and mitigation [3]. The social networks give the possibility to share the information concerning the damage of disaster [4]. Another idea is the context-aware social networking module for interaction [5], which offers the possibility for posting and locates the place of a disaster in the social network [6]. Moreover there are attempts to elaborate applications for smart phones that would be capable of the disaster response [7].

However, the proposed solutions should also take into account the regional constraints. In [3, 8], the specifics of the problem regarding Romania, Ukraine and the Republic of Moldova is described. We can mention especially the paper [3], where a basis of Romanian Controlled Vocabulary is established. This resource was elaborated being based on a number of professional controlled vocabularies, e.g. those of The International Press Telecommunications Council [9], specialized authorities from Romania, USA and authors experience.

Another example of a professional vocabulary is presented in [10]. It consists of sub vocabularies divided in 17 categories. Every category has several authors representing the authorities that are responsible for the terminology.

Republic of Moldova has also such a Service of Civil Protection and Exceptional Situation classifier [11]. It consists of situations descriptions that are possible in the Republic of Moldova. For our research this resource serves as a base for social disaster classification.

On the other hand the problem of continuous completion of the Controlled Vocabulary is a permanent task. The keywords that should be included in this vocabulary are taken from social networks, blogs, news online articles etc [12], disaster related keywords being of big importance for emergency system [4].

3 Methodology

One of the first steps of the project is on-going monitoring of the information networks or news sites or social networks used with the aim of finding texts containing any signals about something that has occurred or is about to occur somewhere.

In selecting the relevant information, we can point out relatively distinct approaches for different sources of information: news sites and social networks. In both cases, processing algorithms will take into account the date of publication, because we need to operate with fresh data. Also, it is necessary to exclude various promotional posts and other information having a character of noise, in relation to our topics [13].

At the first glance, it would seem that processing of social networks has an additional key that would allow us to select the posts easily, related to a particular topic – hashtags. They can assist in following chain of posts concerning the given topic. But, as it was mentioned in [14], it is impossible to establish a priori and form a controlled vocabulary of these hashtags, because they operate with a specific lexicon, often unpredictable. Many times hashtags didnt contain any words related directly to disaster subjects, labelling the corresponding event by some toponyms, expressions or some other proper nouns. As examples can serve the well-known hashtag #jesuischarlie or #colectiv related to the fire in a night club in Bucharest, which took dozens of young lives. Therefore we will treat social networks in a manner different from news sites. In this case the controlled vocabulary is formed from two parts: a static one, consisting of keywords selected apriori from different sources and a dynamic part, which comprises relevant hashtags.

The method of their extraction will be described below. To create the above-described tools, we will accomplish the following steps:

Stage I. Preparatory phase.

- Manual Romanian texts collection and their (manual) classification.
- Elaboration of a lexicon of markers based on these texts and other available sources.
- Automated enriching of the vocabulary by flexing and derivation.
- Creating the database with classified texts.

Stage II. Processing phase.

- Automated news texts selection.
- Texts filtering according to lexicon of markers.
- Texts analysis.

In the following sections we will present the approaches that we consider to accomplish these steps.

4 Online News Articles

We used a number of sites from the Republic of Moldova and Romania and managed to collect 616 relevant texts in Romanian. Collected texts were pre-processed and manually classified. 616 news articles were divided into 10 categories of disasters which are present in the Service of Civil Protection and Exceptional Situation classifier: railway, air and cars accidents, fire, earthquake, hurricanes, radioactive sub-stances, attacks, flood and diseases. In case we have a text that can be considered as part of two or more categories, it is assigned to those categories. These 10 categories covering 84% of cases are present in the classifier [11]. The other 16% of the situations are not so frequent, therefore we omitted them. These are related to mass loss of wildlife, vegetation destruction on a vast territory, considerable change of atmosphere transparency, etc.

They contain 142840 words, from several news sites (see details in Table 1). All texts were lemmatized and annotated with the part of speech tagger. This fact will help us in the identification of those words that refer to social disasters. The next step was to attach, if necessary, a part of sentence, in order to avoid ambiguities and classify them. Finally, we established 10 groups of texts that refer to a specific social situation, with its own set of lexical markers.

Table 1. Statistics on collected articles

Nr.	Category name	Number of articles	Number of words	Article average nr. of words
1	Hurricanes	5	3380	676
2	Earthquake	6	2412	402
3	Radioactive contamination	10	2406	241
4	Diseases	12	5060	422
5	Railway accidents	40	10784	270
6	Air accidents	100	25675	257
7	Cars accidents	100	18005	180
8	Attacks	100	27236	272
9	Flood	103	23922	232
10	Fire	140	24960	178
	Total	616	143840	233

When processing this collection we also applied the disambiguation methods. This is necessary because the words can have multiple meanings occurring in different contexts. For example, the word "bomb" can have a meaning "an explosive weapon detonated by impact" or the meaning of "something very cool/good". Another example would be the word "incendiary", which means both something causing (or designed to cause) fires or something arousing to action (for example, an incendiary speech).

5 Lexicon of Markers

As it was written above, the obtained collection of annotated texts serves as a source for lexicon of markers. Those words that correspond to social disaster topic were manually selected and added to this linguistic resource. In our research, as it was written in [14], we also use some other sources: the site of Service of Civil Protection and Exceptional Situation and Romanian Controlled Vocabulary, developed in [4, 15].

These three sources constitute the main part of lexicon of markers. At the moment we have a lexicon of markers containing more than 350 lemma words. There is a number of samples from this collection: accidenta (injure), alarma (alarm), alerta (alert), avaria (failure), bombarda (bomb), deraia (derailing), detona (detonate), distruge (destroy), evacua (evacuate), exploda (explode), inunda (flooding), nenoroci (perish), ucide (kill), vătăma (hurt), pustii (devastate).

This lexicon of markers needs to be permanently populated by other keywords. The usual way is to increase the number of news articles and to select new words. On the other way we may use the internal linguistic mechanisms to increase the number of words from the lexicon starting with the existent set. That is why a useful component of the system would consist in automatic completion of the lexicon of markers.

The tools we use for text monitoring and analysis operate with word stems. As the Romanian language belongs to the class of inflectional ones, the process of word forming or derivation of a number of vowel or consonant alternations may occur, generating new stems. For example, there are three different stems for Romanian verb "a dărăpăna" (to run-down): dărapăn, dărăpăn, dărapen.

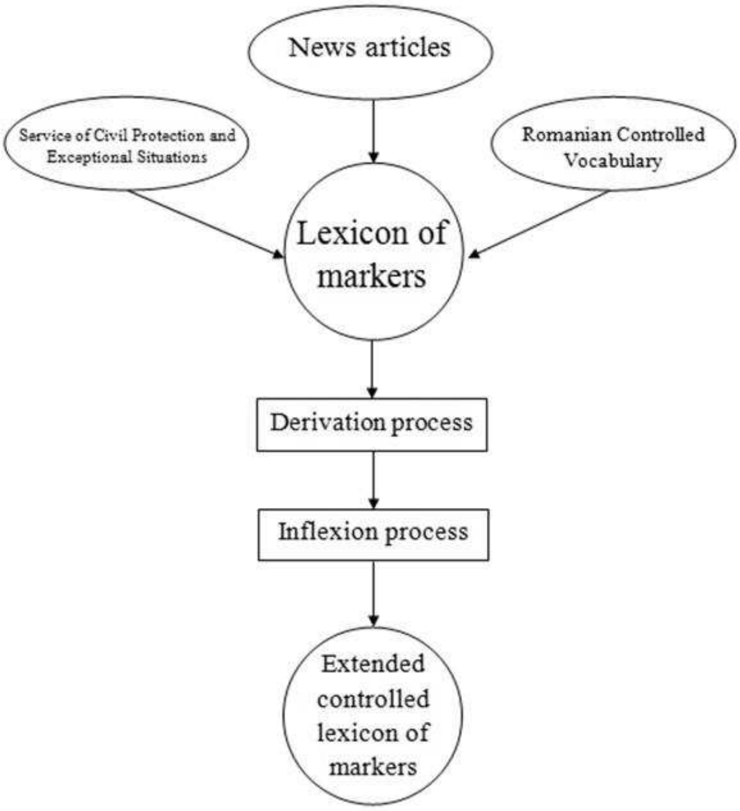

News articles
Service of Civil Protection and Exceptional Situations
Romanian Controlled Vocabulary
Lexicon of markers
Derivation process
Inflexion process
Extended controlled lexicon of markers

Figure 1. The process of extended controlled lexicon of marker creation

Therefore, for each word it is necessary to have all the possible stems. We use in-house elaborated tool to inflect the selected keywords (it has a general purpose, also applicable in our case). The tool is based on grammar rules with scattered contexts, and word-forms generation is reduced to the corresponding grammar rules interpretation [16]. The inflexion is based on the knowledge about the morphological group of a given word. An algorithm for calculating morphological group of an arbitrary word was developed [17]. Obviously, for each ending we can establish a correspondence with morphological characteristics of the word-form, thus obtaining a possibility of morphological annotation. If in the case of inflection, this does not change the meaning of the obtained words, and the problem is solved automatically [18], in the case of derivation the process of automation is more complicated. Our goal consists in realizing automatic derivation without use of semantic elements in the process of derivation. The only information was concerning the character representation, and in some cases, the part of the speech.

As we described in [19] there are four algorithms: affix substitution, derivatives projection, formal derivation rules and derivational constraints. A few affixes form the overwhelming majority of derivatives: 12 of 41 prefixes formed 88.2% of all derivatives with prefixes, analogously, 52 of the 420 suffixes formed 87.7% of all derivatives with suffixes.

Even if we apply these four algorithms a step of validation is needed. Our approach was based on the Internet filtration and manual validation of the generated words [9]. The method shows that we can increase the vocabulary by approximately 15%, with the accuracy of 89%.

The processes described above refer to the static part of the lexicon. In the case of the dynamic one its completion is performed at processing phase, when tweets containing hashtags are analysed. If a tweet is identified as relevant to disaster topics, its hashtags $H = \{h_1, h_2, ..., h_n\}$ are extracted. It is necessary to determine which of them belongs to the field of interest. For this purpose, all of them are initially followed during m subsequent steps and corresponding tweets are analysed. In case if they refer to the same disasters topic, the hashtags identical to those contained in the set H are extracted and included in the dynamic part of the controlled vocabulary.

6 Automated Texts Collection

For the processing phase a Crawler-based [20] application service has been elaborated. It inspects various news websites, downloads and extracts the text of this news, and stores it in the database. Crawler is written using Node.js and Request library. Since each site is unique by its structure, plug-ins were written for each of them, taking into account its specific design. Fig. 2 presents the application interface with the result of extracting the text from ProTV news site.

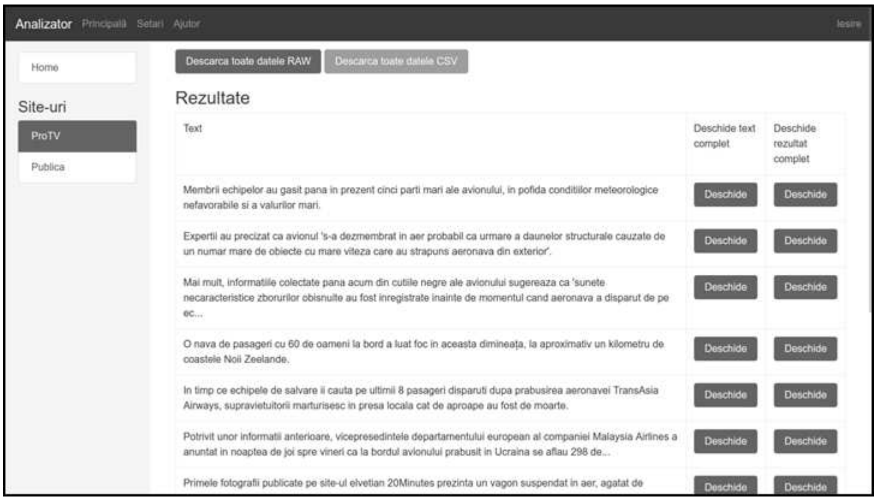
Articles are extracted as follows: RSS-feed is downloaded, and then news is filtered in accordance with the lexicon of markers. Our approach considers the lexicon markers as classifier attributes for the process of classification in those 10 categories. The process of classification is done with j48tree classifier implemented in Weka programme. This approach gives us approximately 76% of accuracy.

After the filtration the full text of news is extracted because RSS-feed is just a part of the article. The process of text extraction is the following: the corresponding page on the website ProTV in HTML format is downloaded, the page with pure text is extracted, then it is cleansed from the rest of HTML tags that remain and only the final result is stored in the database.

7 How to Streamline the Process

In order to automate the process of text collecting referring to disasters, we started with Natural kit for NodeJS. It consists of different natural language tools, including two classifiers, Naive Bayes and logistic regression. Using the established 10 categories we populated a database with collected texts, presenting them in corresponding format. The idea is to have an amount of texts classified in different topics.

The given text is analysed by comparing with the classified texts



Rezultate

Text	Deschide text complet	Deschide rezultat complet
Membrii echipelor au gasit pana in prezent cinci parti mari ale avionului, in pofida conditiilor meteorologice nefavorabile si a valurilor mari.	Deschide	Deschide
Expertii au precizat ca avionul 's-a dezmembrat in aer probabil ca urmare a daunelor structurale cauzate de un numar mare de obiecte cu mare viteza care au strapuns aeronava din exterior'.	Deschide	Deschide
Mai mult, informatiile colectate pana acum din cutiile negre ale avionului sugereaza ca 'sunete necaracteristice zborurilor obisnuite au fost inregistrate inainte de momentul cand aeronava a disparut de pe ec...	Deschide	Deschide
O nava de pasageri cu 60 de oameni la bord a luat foc in aceasta dimineața, la aproximativ un kilometru de coastele Noii Zeelande.	Deschide	Deschide
In timp ce echipele de salvare ii cauta pe ultimii 8 pasageri disparuti dupa prabusirea aeronavei TransAsia Airways, supravietuitorii marturisesc in presa locala cat de aproape au fost de moarte.	Deschide	Deschide
Potrivit unor informatii anterioare, vicepresedintele departamentului european al companiei Malaysia Airlines a anuntat in noaptea de joi spre vineri ca la bordul avionului prabusit in Ucraina se aflau 298 de...	Deschide	Deschide
Primele fotografii publicate pe site-ul elvetian 20Minutes prezinta un vagon suspendat in aer, agatat de	Deschide	Deschide

Figure 2. The application interface

from the database, formed by the collected texts. The tool gives a similarity score to a certain text which can offer a statistic idea of how close the analysed text is to a selected topic.

The below example presents the results of processing of the following text: Experţii au precizat că avionul 's-a dezmembrat in aer probabil ca urmare a daunelor structurale cauzate de un număr mare de obiecte cu mare viteză, care au străpuns aeronava din exterior'. (in engl. *Experts have said that the plane 'was dismembered in the air probably due to structural damage caused by a large number of objects with high-speed that have penetrated the aircraft from the outside'.*)

```
   {label:'Air accidents',value:9.46168293230331e-40}
   {label:'Railway accidents',value:1.671323536752323e-40}
   {label:'Radioactive contamination',
value:9.51103968862598e-45}
   {label:'Fire',value:5.94988550817385e-45}
   {label:'Cars accidents',value:1.232757059054971e-45}
```

```
  {label:'Flood',value:6.389181886502171e-49}
node analyze.js 80,83s user 0,04s system 100%
cpu 1:20,87 total
```

One can see that the highest similarity score is achieved in the first case related to “Air accidents”, which corresponds to the meaning of the processed text.

One of the observations is that if the database grows by n, then the processing time is growing by $2n$. At the moment the processing time is quite high, e.g. the processing time in the example above is more than one minute. The research direction must be taken to streamline the process. Our goal is to increase the processing speed of texts. One approach is to optimize the information in the database.

We performed the following experiments on a sub-collection of texts (45 news articles, consisting of 11257 words, referring to railway, air and car accidents). The same procedure was applied: annotation at sentence and word levels, providing morphological information using UAIC Romanian Part of Speech Tagger [11]. Based on the obtained results, we got 2659 unique lemmas. In addition, extracting only those which have the frequency more than one, and part of speech noun, verb, adjective and adverb, we obtained 1093 different lemmas. So, the procedure showed how to reduce the number of susceptible words for markers and to optimize the processing time.

On the other hand, in order to reduce processing time, changes were made on the contents of the database, excluding from the collected texts those words or even sentences that are not directly related to the subject of disaster. For example, the following text: ”Two persons were killed and seven others injured on Monday evening in a bus explosion in the Armenian capital, Yerevan, announced the Ministry for Emergency Situations of the Caucasian Republic, AFP informs” can be reduced to a short form, namely: “Two persons were killed and seven others injured in a bus explosion”. The remainder is sufficient to serve as a pattern for analysis and classification of new extracted texts, but the processing time will be significantly decreased. These minimizing of database records must be made with great accuracy, not to lose substantial information. Obviously, the cuts are operated in patterns only, not in the news, where, for example, place of the event and source of information can be important for mitigation scenarios.

8 Conclusions and Further Work

Our experience has shown that the proposed tool provides acceptable results. In order to obtain a better classification it is necessary to increase the number of collected texts, especially those related to the topics of hurricanes, earthquake, radioactive contamination and diseases. Despite the optimization measures, the processing time tends to increase with the expansion of the database and we decided to develop a distributed processing algorithm using the 64 node cluster from the Institute of Mathematics and Computer Science.

References

[1] NATO Science for Peace and Security Program. [Online]. Available: http://science.iasa.kpi.ua/sps.

[2] N. O. Hodas, G. V. Steeg, S. Chikkagoudar, J. Harrison, E. Bell, C. D. Corley, "Disentangling the Lexicons of Disaster Response in Twitter," in *WWW 2015 Companion*, Florence, Italy, May 1822, 2015, pp. 1201–1204.

[3] H. N. Teodorescu, "Using analytics and social media for monitoring and mitigation of social disasters," *Procedia Engineering*, vol. 107C, pp. 325–334, 2015.

[4] A. S. Gowri, R. Kavitha, "Tweet Alert: Effective Utilization of Social Networks for Emergency Alert and Disaster Management System," *International Research Journal of Engineering and Technology*, vol. 2, no. 8, 2015, pp. 1065–1070.

[5] R. Simionescu, "Hybrid POS Tagger," in *Proceedings of Language Resources and Tools with Industrial Applications Workshop* (Eurolan 2011 Summer School), Cluj-Napoca, Romania, 2011, pp. 21–28.

[6] R. Rana, I. Kristiansson, I. Hallberg and K. Synnes, “An Architecture for Mobile Social Networking Applications,” in *First International Conference on Computational Intelligence Communication Systems and Networks*, CICSYN ’09, 2009, pp. 241–246.

[7] B. Brownlee and Y. Liang, *Mobile Ad Hoc Networks: An Evaluation of Smart phone Technologies,* Kingston (ONTARIO), Canada: Royal Military College of Canada, 2011, 40 p.

[8] N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha and V. V. Savastiyanov, “Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters,” in *Improving Disaster Resilience and Mitigation – IT Means and Tools*, Springer, 2014, pp. 119–134.

[9] J. Hjelm, *Why IPTV: Interactivity, Technologies, Services,* John Wiley & Sons, 2008, 370 p.

[10] U.S. Department of Health & Human Services. Disaster Information Management Research Center. Disaster Glossaries, [Online]. Available: https://disaster.nlm.nih.gov/dimrc/glossaries.html

[11] Service of Civil Protection and Exceptional Situation classifier, [Online]. Available: http://www.dse.md/ro/clasificator

[12] N. D. Pankratova and V.O. Dozirtsiv, “Application of methods for text analysis of the emotional tone to identify social disasters,” in *System analysis and information technology: 18-th International conference SAIT 2016*, Kyiv, Ukraine, May 30 – June 2, 2016, pp. 38.

[13] C. Bolea, “Vocabulary, Synonyms and Sentiments of Hazard-related Posts on Social Networks. An analysis for Romanian messages,” in *Proceedings of IEEE Conference SPED 2015*, Bucharest, Oct. 2015, pp. 48–53.

[14] M. Petic, S. Cojocaru and V. Gîsca, “Exploring list of markers in automatic unstructured text data processing,” in *Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iaşi, Romania, 26-27 November 2015, pp. 125–136.

[15] H. N. Teodorescu - private mail

[16] A. Alhazov, E. Boian and S. Cojocaru, “Modelling Inflections in Romanian Language by P Systems with String Replication,” in *Proceedings of the 10th Workshop on Membrane Computing*, WMC10, Curtea de Arges, Romania, August 24 27, 2009, pp. 116 –128.

[17] S. Cojocaru and E. Boian, “Determination of infexional group using P systems,” *Computer Science Journal of Moldova*, vol. 18, no. 1, pp. 70–81, 2010.

[18] M. Petic and S. Cojocaru, “Vocabulary enriching for text analysis,” in *System analysis and information technology: 17-th International conference SAIT 2015*, Kyiv, Ukraine, June 2225, 2015. pp. 37–38.

[19] M. Petic, V. Gîsca and O. Palade, “Multilingual mechanisms in computational derivational morphology,” in *Proceedings of Workshop on Language Resources and Tools with Industrial Applications LRTIA-2011*, Cluj-Napoca, Romania, pp. 29–38.

[20] M. Najork and J. L. Wiener, “Breadth-first crawling yields high-quality pages,” in *Proceedings of the Tenth Conference on World Wide Web*, Hong Kong, May 2001, Elsevier Science, pp. 114–118.

Svetlana Cojocaru, Mircea Petic,
Grigore Horoş

Received May 15, 2016

Institute of Mathematics and Computer Science
Address: 5, Academiei street, Chisinau, Republic of Moldova, MD 2028
Phone: (373 22) 72-59-82
E–mail: svetlana.cojocaru@math.md, mircea.petic@math.md,
grigori.horos@math.md