DYNAMIC VOCABULARIES FOR STUDYING INTERNET NEWS[4]

Mikhail Alexandrov, Daria Beresneva, Alexander Makarov

Abstract: *Nowadays there are many toolkits (methods and software) for automatic topic identification of documents. However economists, sociologists, politicians need tools not only for topic identification but also for analysis of changes in given topics related to time. In the paper we propose a simple technology, which could help to solve such a problem. For this: selected publications are distributed on sets associated with consequent time intervals, keywords are extracted from each document set, and these keywords are combined to reflect their dynamics. These combinations are named 'dynamic vocabularies'. We present two programs for building dynamic vocabularies and then we demonstrate our technology on real example related to the topic of Euro integration of Ukraine (2013-2014)*

Keywords*: dynamic vocabularies, topic identification, Internet-sociology.*

ACM Classification Keywords: *I.2 Artificial Intelligence.*

Introduction

Dynamic of social-economical and social-political processes is an object of consideration of economists, sociologists and politicians. To study such a dynamics the mentioned specialists have to read many materials circulating on the Internet and distributed in time. These efforts can be essentially reduced when one has initial knowledge (information) about topic(s) under consideration. Just for this case we propose a technology based on so-called 'dynamic vocabularies'. Such vocabularies are keywords lists, which should be extracted from publications and then combined by a special way. The publications are supposed to refer to a certain period of time. The proposed technology is demonstrated on the real example related to protests in Ukraine (2013-2014).

Dynamic vocabularies were considered in [Alexandrov, 2001; Makagonov, 2006]. In this paper we use new tools and propose new realization of dynamic vocabularies.

In section 2 we present tools for building dynamic vocabularies. Section 3 describes the results of the experiments. Section 4 contains conclusions.

Tools

Source of information

Initially several reliable Internet resources are chosen. One should not change these resources in order to keep the quality of information. We also fix the time of analysis, time step, and therefore we obtain a series of intervals on the time axis. Then all the collected documents are distributed between these intervals. Therefore, for *n* intervals we have just *n* corresponding document sets. To find documents one can use any usual search engines (Google, etc.). In our work we use our own crawler, which takes into account the title of topic with its keywords, sources of information, and time period.

Keyword extraction

Keyword lists are built for each document set. To select keywords we use here so-called criterion of specificity.

Definition: The level of specificity of a given word $\boldsymbol{w}$ in a given document corpus is a number $K \geq 1$, which shows how much its frequency in the document corpus $f(\boldsymbol{w})$ exceeds its frequency in the General Lexis $F(\boldsymbol{w})$:

[4] Work done under partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

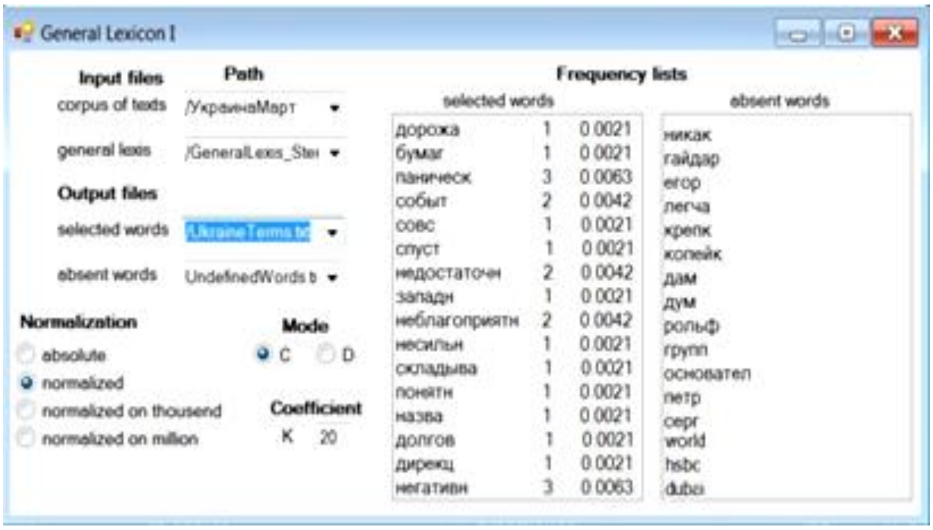
K = *f*(***w***) /*F*(***w***). Speaking ‘General Lexis’ we mean General Lexis of a given language (it is Russian in our case). We use here the Lexis from [Sharoff, http].

Example: Let we have the following data for the word ‘protest’: *f(protest)*=$0{,}8*10^{-6}$, *F(protest)*= $0{,}5*10^{-8}$. Let the critical value *K*=100. It is easy to see that *f*(.) > *K***F*(.). It means that the word ‘protest’ has the level of specificity more than 100 and therefore this word will be selected. The threshold 100 is taken here only as an example.

In the real practice this value is determined experimentally: we should not lose many useful words and from the other hand we should not have the many unnecessary ones. In particularly, in our experiments we used *K*=50.

The criterion of specificity is calculated by the program LexisTerm, which is a free software developed in Peru [Lopez, 2011]. It should note that LexisTerm can use two modes: the corpus mode and the document mode. In the first case the program considers all documents as one large document. The definition presented above refers to this mode. In the corpus mode LexisTerm selects words being specific for the whole corpus but it loses words being specific for individual documents. In the second case the program considers each document independently. In the document mode all specific words are selected.

In our work we use a new version of the mentioned program - LexisTerm-I (International). Unlike LexisTerm, this program: (a) allows to process both English and Russian texts; (b) selects those words that satisfy the criterion of specificity and those that are absent in the General Lexis. A screenshot of the LexisTerm-I interface is shown in Figure 1.



General Lexicon I

Input files — **Path**

corpus of texts: /УкраинаМарт

general lexis: /GeneralLexis_Ster

Output files

selected words: /UkraineTerms.txt

absent words: UndefinedWords.t

Normalization

○ absolute
● normalized
○ normalized on thousend
○ normalized on million

Mode

● C ○ D

Coefficient

K 20

Frequency lists

selected words			absent words
дорожа	1	0.0021	никак
бумаг	1	0.0021	гайдар
паническ	3	0.0063	егор
событ	2	0.0042	легча
совс	1	0.0021	крепк
спуст	1	0.0021	копейк
недостаточн	2	0.0042	дам
западн	1	0.0021	дум
неблагоприятн	2	0.0042	рольф
несильн	1	0.0021	групп
складыва	1	0.0021	основател
понятн	1	0.0021	петр
назва	1	0.0021	серг
долгов	1	0.0021	world
дирекц	1	0.0021	hsbc
негативн	3	0.0063	dubai

Figure 1. Interface of the program LexisTerm-I

Keyword are selected in the corpus mode and not in the document mode, because the individual documents prove to be very subjective with respect to events under consideration. At the final stage all keyword lists are manually corrected by an expert.

Dynamic vocabularies

Keywords are selected on a weekly basis. Using the resulting lists we can combine them in order to form 4 types of dynamic vocabularies. They concern all weeks and each week:

1. the common words for all weeks or for a given part of all weeks,
2. words that belong to the current week and don’t belong the previous week, we name them ‘new words’,
3. words that belong to the current week and the previous week, we name them ‘repeated words’,
4. words that don’t belong to the current week but belong to the previous week, we name them ‘old words’.

Speaking ‘common words’ we mean the words that appear in *m* weekly keyword lists where $m \geq M$ and *M* is a given threshold. The Figure 2 shows the distribution of keywords between weeks.

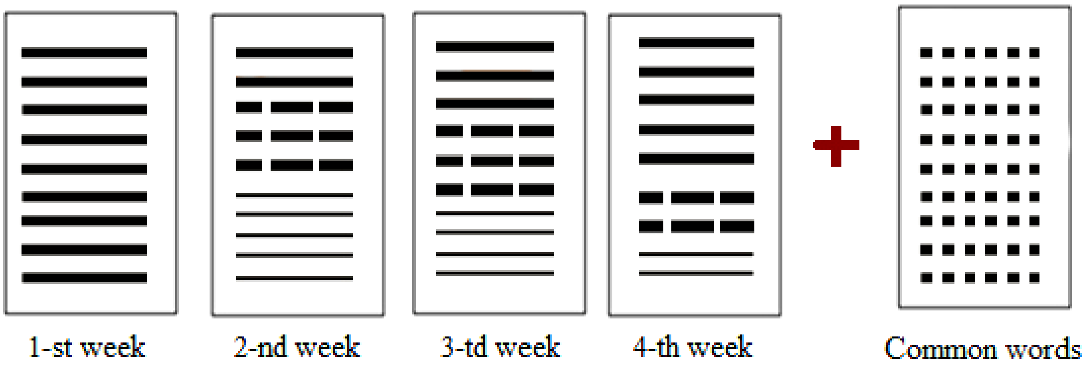
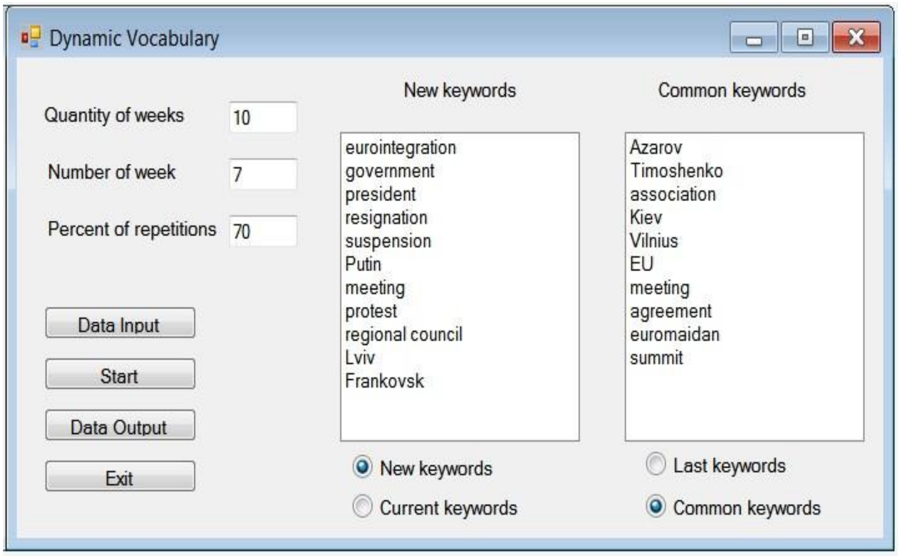


Figure 2. Dynamic vocabularies

Here: thick lines are the new words, dashed thick lines are the repeated words, thin lines are the old words, and dotted thick lines are common words.



Quantity of weeks 10

Number of week 7

Percent of repetitions 70

New keywords
eurointegration
government
president
resignation
suspension
Putin
meeting
protest
regional council
Lviv
Frankovsk

Common keywords
Azarov
Timoshenko
association
Kiev
Vilnius
EU
meeting
agreement
euromaidan
summit

● New keywords
○ Current keywords

○ Last keywords
● Common keywords

Figure 3. Interface of the program DynVoc (keyword lists are translated to English)

Dynamic vocabularies are built automatically using the program DynVoc (Dynamic Vocabularies). The input of the program is a set of weekly keyword lists and the output of the program is a set of dynamic vocabularies. The interface of the program is presented on Figure 3.

Experiment

In the experiment we studied publications related to mass protests in Ukraine. The principal topic of the protests was Ukrainian Eurointegration. The protests lasted approximately 23 weeks (October 2013 – March 2014). We considered only the first 12 weeks (October 2013 – December 2013), which defined the further development of the Ukrainian events. We used 4 popular Russian Internet editions “Arguments and Facts”, “Russia Today”, “RIA Novosti”, “Gazeta.ru”. All papers were divided on weeks. On average we had 15 papers per one week.

Figure 4 shows vocabularies concerning the 7-th week (November 8-15, 2013). By that moment the suspension of the process of European integration had led to the numerous protests in Kiev, Lviv and other Ukrainian cities. The protesters demand the resignation of the Ukrainian Prime Minister Azarov. The Russian President Putin asks the EU to depoliticize the topic of Ukrainian Eurointegration. From other hand the perspectives of integration or partnership with the ex-Soviet countries as Kazakhstan, Armenia etc. are not already discussed.

New	***Repeated***	***Old***	***Common***
State	Azarov	Armenia	association
depoliticized	agreement	EuroSummit	power
Eurointegration	cooperation	Kazakhstan	East
Evromaydan		chancellor	European Union
compensation		conflict	Kiev
credit		Moscow	agreement
Lviv		dishonest	Ukraine
international		required	membership
meeting		partnership	Yanukovych
Regional Council		justice	Azarov
opposition		opposed	Tymoshenko
resignation		ratification	
demanded		Rogozin	
government		Russia	
president		marketing	
premier		speculation	
Suspend		customs	
protest		technology	
Wrongful		criminal	
Putin		economic	
resolution		electric power	
sovereignty			
Frankivsk			

Figure 4. Example of dynamic vocabulary (keyword lists are translated to English)

One who knows the situation in Ukraine may agree that in totally this dynamic vocabulary reflects the contents of protests at that moment in November. Here: Lviv and Frankivsk are the regional centers; Armenia and Kazakhstan are ex-Soviet countries; sovereignty, compensation and credits are related to economical aspects of Eurointegration; President Putin (Russia) says about depoliticization; suspended agreement, protests, Euromaydan and Eurosummit are in one chain; the other chain is President Yanukovich (Ukraine), dishonest activity and criminal environment; the third chain is Mr. Rogozin from the Russian Government, Ukrainian market, and cooperation with Russia; etc.

Note. Common words here are the words that appear in more than 50% of weekly keyword lists. The threshold 50% was assigned by user (one of the authors). So, these words may be absent in some weakly keyword lists. Just for these reason some common words can be included both in the last column and simultaneously in one of the first three columns, see the words Azarov and agreement

Conclusions

The results of completed work are:

- the simple method for studying dynamics of topics;
- software for building dynamic vocabularies;
- experiment with the real data.

The proposed approach 'works' well when user (expert) has an initial knowledge about events, objects, persons related to the topic of interest and he/she wants only to meet with dynamics of this topic. Otherwise it is necessary to use the other ways as document annotation or selection of representative documents from given document sets.

In future we suppose:

- to automate the process of keyword lists correction;
- to include the procedures of visualization to the program DynVoc.

Bibliography

[Alexandrov, 2001] Alexandrov, M: Dynamic domain-oriented dictionaries as a tool for revealing tendencies in development of some scientific and technological disciplines and interaction between them. // Mexican National Council on Sciences and Technologies (CONACyT), Reg.No. 39011-a.

[Lopez, 2011] Lopez, R., Alexandrov, M.:Tejada, J: LexisTerm - The Program for Term Selection by the Criterion of Specificity, // Proc. of 4-th Intern. Conf. on Intelligent Inform and Engineering Systems, ITHEA Publ, 2011, p. 8-15

[Makagonov, P., 2006] Makagonov. P., Figueroa, A., Gelbukh, A. Studying Evolution of a Branch of Knowledge by Constructing and Analyzing its Ontology.// Springer, LNCS, 2006, 10 pp.

[Sharoff, http] http:// www.artint.ru/projects/frqlist.php, General Lexis of Russian

Authors' Information

Mikhail Alexandrov – *Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: MAlexandrov@ mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Daria Beresneva – *M.Sc student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State Research University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia*

e-mail: dejame@ yandex.ru

Major Fields of Scientific Research: mathematical modeling, world economy

Alexander Makarov – *Researcher, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia;*

e-mail: mackarov54@ gmail.com

Major Fields of Scientific Research: data mining, Internet sociology