The 1th IEEE International Conference on Data Stream Mining & Processing
23-27 August 2016, Lviv, Ukraine

Big Data in Information Analytical System «NEWSCAPE»

Prof. Hulianytskyi Leonid, Dr. Malyshko Sergiy
Glushkov Institute of Cybernetics of NAS of Ukraine, 40 Glushkov ave., Kyiv, Ukraine, 03187,
lh_dar@hotmail.com
smsmsm@admpr.com.ua

Abstract — **This paper describes the main features of information analytical system «NEWSCAPE» and corresponding online service. «NEWSCAPE» is a productive information technology, which accumulates a variety of mathematical methods and algorithms, focused on the monitoring and multivariate analysis of media (information space).**

Keywords — information analytical system, information space, online service, big data

I. Introduction

The rapid development of global electronic communications networks, coming from general computerization processes, has led to emerging of a massive information space. Just a few decades ago, the actual task lied in finding the right information, while now the main problem is to screen unnecessary data [1]. Nowadays such tasks as searching for patterns and relationships, tracking trends and tendencies, forecasting increase and decrease of the quantitative parameters of media objects are in the forefront of information space analysis. Developed information-analytical system (IAS) allows to see "simple and clear" in a large, complex and chaotic information space.

The relevance and practical importance of the study is growing together with the expansion of information space. At the moment, there does not exist any single standardized information technology for media monitoring and analysis.

The main purpose of the development of proposed IAS «NEWSCAPE» is to create a modern and efficient information technology and corresponding online service, which accumulate a variety of mathematical methods and algorithms, focused on monitoring and multivariate analysis of information space. Some of the ideas, methods and algorithms, used in the IAS, are of original authorship [2].

Because of the great diversity and specificity of this kind of software, reviewing existing media monitoring systems is beyond the scope of this paper. We would like to note an actual news integration technology «InfoStream» as an example of such kind of product [3, 4].

The paper discusses the IAS «NEWSCAPE» development and usage. Established IAS is based on the modern scientific and technological advancements of the Institute of Cybernetics. The system has been developed and implemented by the combined efforts of the team of highly qualified specialists in the field of mathematical modeling and IT. It is an up-to-date hardware and software system that performs basic functions of media monitoring.

Note that the IAS «NEWSCAPE» continues to evolve and improve in line with the needs of modern practices and requirements. Authors pursue to ensure the effectiveness of the product for a wide range of users. The analytical capabilities of the system, based on modern and original mathematical apparatus, are also growing.

II. General information about the system

IAS «NEWSCAPE» is suggested for usage either in the stand-alone mode, or as a component of intellectual decision support systems (DSS) for complex and conflict situations. The developed system provides analytics with process of developing effective solutions on various levels, from the operational tasks of the situational center to the strategic planning issues.

This IAS can also be of use in a current informational war - both for controlling and planning “our” activities and identifying and analyzing activities of the "enemies".

Programmatically implemented algorithmic tools of the developed IAS are able to:

- provide continuous automated monitoring of the information space, that includes about a thousand of highest-rated sites, main national and regional internet sources, leading print media and television channels (video monitoring);
- accumulate a variety of disparate databases into a single information system;
- store monitoring results for subsequent analysis in a structured way;
- use modern methods of quantitative and qualitative informational analysis;
- support effective teamwork between operators and analysts of different levels.

There are different approaches to the definition of the term "information space". In practice, this concept is often considered a general term, which does not require further explanations. However, many researchers believe that the modern meaning of "information space" is a result of the evolution of the conceptual scheme of area distinction in the overall geopolitical space. These areas could be considered separate spaces with their borders, structure, resources and features of interaction between subjects of social relations [5].

Due to increasing level of computerization, term "information space" received a modern synonym - media space, i.e. space created by electronic communications. In our case, term "information space" means a set of electronic versions of traditional media, online resources (websites, portals) and "new media" (social networks, blogs).

Media analysis process in developed IAS is based on the analysis of categories, formed by users and containing various structural units of interest - objects, persons, processes, events etc.

Architecture and functional content of the IAS was designed to achieve next goals:

- prompt reception of all necessary and accurate information;
- tracking key trends in media space;
- conducting multivariate analysis of objects, persons, events, trends;
- improvement of the large-scale systems and subsystems management;
- simulation and exploration of the complex processes;
- prediction of the possible scenarios of development processes.

Thanks to the created web-based application, you can access proposed IAS from anywhere and with any terminal device (computer, tablet or smartphone), if you have an internet connection.

Since IAS «NEWSCAPE» is based on the economical and socio-political information, it could be used as an effective tool for analysis and decision support in the central and local government institutions (Verkhovna Rada and its committees, Presidential Administration, Cabinet of Ministers of Ukraine), as well as in large companies, organizations, political parties.

Systems of this kind are needed in various media holdings, large corporations and associations that actively promote their goods or services in the information space and want to know an objective picture of their policy success.

Original mathematical methods and algorithms, which are available to users, include:

- analysis of the media objects characteristics, such as number of references, citation index, media activity index, information activity index, information dynamics rating, regional distribution, distribution of media types, estimation of the aggressiveness level, identifying signs of manipulations in the texts;
- content analysis – quantitative, differential and frequency analysis, as well as an analysis of the related words and parts of speech in sentences;
- categorical analysis of objects, which may act as politicians, political, political parties, regions of Ukraine, countries of the world, state capitals, international organizations, industries, food, etc.;
- morphological analysis, able to work with nouns, pronouns, adjectives and adverbs, verb forms, non-dictionary objects and other parts of speech;
- analysis of the quantitative characteristics of the text - sentences, words, nouns, unique nouns, lemmas, word forms – with an ability to calculate numerous important statistical indicators.

III. Monitoring of information space

The basic functionality of the IAS «NEWSCAPE» consists in continuous automated monitoring of information space that currently includes about a thousand of highest-rated sites, main national and regional internet sources, leading print media and television channels (video monitoring). Due to development of the specialized search subsystem, user is able to form complex queries to the IAS and to monitor their results for the media object of interest.

Search subsystem with a standard text query syntax provides initial quantitative results as well as time chart, showing the number of found relevant news and citation index.

Let us introduce the following indicators for the various basic quantitative search results:

S – total number of the units of information in the system for the given period;

N – information blocks, found upon request;

C – citation index, which is equal to the quantity of the found words;

M – media activity rate, equal to $N / S \times 100\%$;

P – media presence index of the given object, equal to C / N.

Additionally, the system allows to obtain basic quantitative characteristics on a daily basis or splitted by the regions of Ukraine. Sometimes, data about main sources of information is useful for the analysis.

In addition to basic indicators, the system «NEWSCAPE» should be able to provide composite indicators of the given information object during selected time period.

Such indicators include scale, uniqueness, regularity, influence and frequency of event and its dynamics index.

The scale of the event is calculated by aggregating and weighting quantitative indicators over the time period. It is directly proportional to the number of object references.

The dynamics index of event is defined over period and is directly proportional to the "steepness of peaks" and the speed of growth or decline of event scale.

The uniqueness of event is calculated by analyzing the historical "jumps and peaks" of the object reference numbers.

The influence of event is an indicator that determines the degree of influence of this event on the other monitoring objects. We consider the relationships "Objects - sources" and set "Objects - receivers."

The regularity of event is an indicator that determines the degree of repetition of similar events.

IV. Search results analysis

IAS «NEWSCAPE» provides easy-to-use interface for quick navigation through results of the search query. The IAS highlights all matches with the search object inside the given text and supports them with brief and detailed forms of display or additional characteristics of the data source. Analyst is provided with ability to conduct a quick and meaningful analysis of the selected information and review major publications on the object for more details.

To identify the "subtle" features of the media objects, additional processing with the specialized IAS subsystems is required. The user can save search results as a final report (RTF format), or as a special file for content analysis subsystem. A detailed description of other IAS «NEWSCAPE» subsystems is beyond the scope of this publication.

Content analysis is one of the most effective subsystems of «NEWSCAPE». Traditionally, content analysis is defined as the quantitative analysis of texts and text arrays, which allows further meaningful interpretations of detected numerical patterns. Content analysis is used to study sources that are invariant in structure or essence, but exist as unstructured or unorganized texts.

The philosophical meaning of content analysis as a research technique lies in moving from the text set to the abstract model content of the text. In that sense, content analysis is a research nomothetic procedure used in the scope ideographic methods.

There exist two basic types of content analysis: quantitative and qualitative.

Text analysis in «NEWSCAPE» can be divided into three types:

- "General Analysis", designed to compute statistics for all nouns and non-dictionary words;
- "Analysis of non-dictionary words", designed to compute quantitative characteristics of words outside the dictionary;
- "Analysis of a set of words", designed to compute statistics of a word set, created by the user.

Let us review some development steps for mathematical methods, used to determine presence of the certain categories during media analysis.

Category is a set of elements that characterizes selected analysis object from some point of view. It could be defined in a descriptive, analytical or algorithmic way.

The day (minimal possible unit of time) is a basic period of analysis. Others aggregated periods (weeks, decades, months) are computed on its base.

Below, the following notation is used:

K – category;

B = *B(t)* – separate block;

t – period;

D_t – total quantity of units in the period *t*;

c – any element;

w(c, B) – number of references of element *c* in the block *B*;

A – set of relevant (filtered) units;

L – number of terms.

Then we set or create a category *K* and look through the text set D_t, available in the current period *t*.

The definition of the category is introduced as follows:

$$K = K_1 \cup K_2 \cup K_3,$$

where K_1 – set of elements, which unambiguously belong to the category *K*,

K_2 – set of terms or word combinations $H_s,\ s = 1,...,L$ (*L* – number of terms), which unambiguously belong to the category *K*,

K_3 – set of elements, which could potentially belong to category *K*.

In general, condition $K_1 \cap K_3 = \varnothing$ should be satisfied. However, cases with $K_1 \cap K_2 \neq \varnothing,\ K_2 \cap K_3 \neq \varnothing$ are allowed.

The following mining algorithm is proposed: after word processing, only significant text elements should be considered.

We denote the current processed block as *B*.

The occurrence frequency of any word in the processed block *B* is calculated by the formula:

$$\tau(c,B) = \frac{w(c,B)}{\|B\|}, \qquad (1)$$

where $\|B\|$ – cardinality of the set *B*.

To determine membership of arbitrary term H_s in block *B*, special characteristic function $\chi(H_s,B)$ is defined.

Normalized level of presence (relevance) of category K in the block B will also be called an index of presence for the corresponding block. Let us define this concept, using a frequency (1):

$$\begin{aligned} f(K,B) &\equiv \alpha_1 \sum_{c\in K^1} \frac{w(c,B)}{\|B\|} + \alpha_2 \sum_{s=1}^{L} \frac{\chi(H_s,B)}{\|B\|} + \alpha_3 \sum_{c\in K^3} \frac{w(c,B)}{\|B\|} = \\ &= \alpha_1 \sum_{c\in K^1} \tau(c,B) + \alpha_2 \sum_{s=1}^{L} \frac{\chi(H_s,B)}{\|B\|} + \alpha_3 \sum_{c\in K^3} \tau(c,B), \end{aligned} \qquad (2)$$

where the normalized values of weighting coefficients $\alpha_1 \geq \alpha_2 \geq \alpha_3$, $\alpha_1 + \alpha_2 + \alpha_3 = 1$ could be identified by experts.

For simplicity, let us introduce special functions:

$$f_1(K,B) \equiv \sum_{c \in K^1} w(c,B),$$
$$f_2(K,B) \equiv \sum_{H_s \in K^2} \chi(H_s,B),$$
$$f_3(K,B) \equiv \sum_{c \in K^3} w(c,B).$$

Then the (2) can be written as:

$$f(K,B) \equiv \alpha_1 \frac{f_1(K,B)}{\|B\|} + \alpha_2 \frac{f_2(K,B)}{\|B\|} + \alpha_3 \frac{f_3(K,B)}{\|B\|} =$$
$$= \frac{1}{\|B\|} \sum_{i=1}^{3} \alpha_i f_i(K,B).$$

General index of the presence (normalized "presence") category *K* for the period *t* in media is computed with (2):

$$F_t(K) \equiv \frac{1}{\|D_t\|} \sum_{B \in A \subseteq D_t} f(K,B),$$

where A – set of relevant units, $D_t = \{B\}$ – set of all present or specially selected units for the period *t*, $A \subseteq D_t$,

$$\|D_t\| = \sum_{B \in D_t} \|B\|.$$

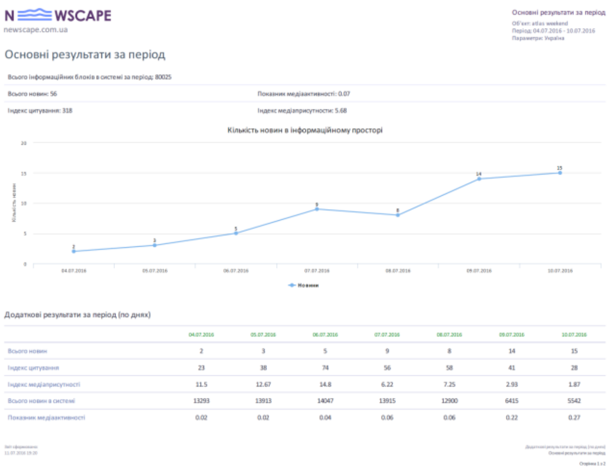
V. Conclusions

As for now, huge amount of information is already accumulated. Significant number of specialized software tools, able to work with it, are developed and implemented as well. At the same time, it is important to note that there does not exist any application or system, which covers that wide range of media analysis tools or offers so advanced software and algorithmic tools as IAS «NEWSCAPE» today in Ukraine.

Since 2010, the system has accumulated more than 14 millions of informational units, more than 100 thousands of printed articles and thousands of video news stories. Economical, social and political information, structured by regions and districts of Ukraine, is also available in a system. There exists a possibility to use data on legal entities, companies and private persons [6], aggregated in the system.

In the future, the system would be developed in the following areas: aggregation of more and more data from diverse sources (expert opinions, television, radio, social media, advertising), standardization and connection of additional data and knowledge bases, improvement of existing and development of new mathematical methods, creation of analytical tools and information technologies for big data. This would provide IAS with more complex mathematical models and methods and improve adequacy and quality of its estimates, predictions and analytical reports for the decision-makers.

For example, Figure 1 shows a general view of the main report from «NEWSCAPE».



N≋WSCAPE
newscape.com.ua

Основні результати за період
Об'єкт: atlas weekend
Період: 04.07.2016 - 10.07.2016
Параметри: Україна

Основні результати за період

Всього інформаційних блоків в системі за період: 80025

Всього новин: 56 — Показник медіаактивності: 0.07

Індекс цитування: 318 — Індекс медіаприсутності: 5.68

Кількість новин в інформаційному просторі

Додаткові результати за період (по днях)

	04.07.2016	05.07.2016	06.07.2016	07.07.2016	08.07.2016	09.07.2016	10.07.2016
Всього новин	2	3	5	9	8	14	15
Індекс цитування	23	38	74	56	58	41	28
Індекс медіаприсутності	11.5	12.67	14.8	6.22	7.25	2.93	1.87
Всього новин в системі	13293	13913	14047	13915	12900	6415	5542
Показник медіаактивності	0.02	0.02	0.04	0.06	0.06	0.22	0.27

Звіт сформовано
11.07.2016 19:20

Додаткові результати за період (по днях)
Основні результати за період

Сторінка 1 з 2

Figure 1. Main report from «NEWSCAPE»

References

[1] V.Mayer-Schönberger and K.Cukier, *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[2] L.F.Hulianytskyi, I.V.Sergienko, S.O.Malyshko. *"About software tools support decision making in problems of group selection", Contol mashines and systems*, no.5, pp. 90-97, 1993. (in Russian)

[3] D.Lande, *The quest for knowledge on the Internet. Professional work*. Moscov: Dialektika, 2005. (in Russian)

[4] D.Lande, A.A.Snarskii, I.V. Bezsudnov, Internetika: *Navigation in complex networks: models and algorithms*. Moscov: Librokom, 2009. (in Russian)

[5] A.V.Manoilo, *State information policy in special circumstances*. Moscov: MIFA, 2003. (in Russian)

[6] L.B.Baran, V.V.Vishnevsyi, K.D.Huliaev, L.F.Hulianytskyi et.a. *Electronic parliament Ukraine: the pilot project. Scientific publications*. S.Dovgyi, Eds. Kiev: Yuston, 2015. (in Ukrainian)