Multimodel clustering of social networks in social dampening applying BIG DATA (acquiring knowledge from data)

I N Khaimovich[1,2], V M Ramzaev[1] and V G Chumak[1]

[1]Samara University of Public Administration "International Market Institute", 21, G.S.Aksakova Street, Samara, Russia 443030
[2] Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

e-mail: kovalek68@mail.ru

Abstract. The developed methodology provides a solution to two essential tasks, thereby revealing the gnoseological potential of Big Data technology: social forecasting in the three most significant areas of the information society based on a model which identifies conditions for social resonance; successful implementation of the social dampening procedure based on the use of appropriate management options using multimodal clusterization of social networks based on Big Data technology. The article suggests the tool that helps to increase work efficiency in the sphere of social dampening in the region. The proposed method of regulation may be efficient when it comes to the control of the regional social dampening processes which have variety of forms and broad range of elements and factors, as well as growth dynamics and active transformation of life activities. At the same time using modern products make it possible to evaluate and show changes on a real-time basis which can be useful for local government authorities.

1. Introduction

Let us consider the analysis of the gnoseological potential of Big Data from the standpoint of synergetics, which focuses mainly on unbalanced, disordered systems, which are formed both in natural and social environments, which acquire balance only at certain moments, often extremely fleeting, but no less important and requiring comprehensive thinking.

Certainly, the destabilization of the system, and above all the social system, is extremely interesting from the dialectical positions both as awareness of the reasons, and in terms of exploring the possibility of providing the system with a steady state. However, the pragmatic reality convinces us that numerous ordinary people who are the direct elements of various systems, as well as the political elite, who seek to hold the position of a leading social force possessing certain levers of influence on social dynamics, are still primarily interested in the stability of the system structures. Deviation from the average values of social indicators is clearly perceived as deviance. Under these conditions, the technical capabilities of Big Data allow us to consider them as a tool for indicating the level of deviation of the most diverse social processes from the optimal model for a given society, contributing

to the development of a mechanism for returning to the averaged, balanced form of both a single process and their complex.

In our opinion, in order to conduct a scientific analysis of social processes, it is necessary to reveal the essence of the many-sided deviations that arise in certain situations in society. The analysis carried out in this direction shows the possibility of identifying two main types of changes in social processes. The first type is characterized by a rather slow deviation of social processes from the norms and principles established in society, described by an additive function. In this case, social smoothing, response can be considered as a cognitive impact on each social subject in particular, regardless of its nature: individual or collective. Here, the dominance of the elementary approach to the description of reality manifests itself, in which the quality of a system is determined by the quality of its constituent elements, which are quite accessible for comprehensive analysis using Big Data.

The essence of the second type of changes in social processes is determined by the explosive changes inherent in it and, first of all, by social cataclysms that bring society out of balance. The second type in its essence reveals deviations of the multiplicative type, which are intensified due to the extremely low efficiency of the interaction processes of social subjects. Here, in contrast to the first type, a social explosion occurs in a sharper form, much more concentrated in time and determined in many respects by the disadvantages of intergroup interaction, similar to social resonance. This situation is an example of the manifestation of a systematic approach to reality in social practice, based on the principles of which the quality and, therefore, the sustainability of a social system depends on the links of its constituent elements.

In order to analyze social processes classified in the above manner, ensure their management and bring social deviations to an acceptable standard, the authors propose to introduce into the conceptual apparatus the term “social dampening”, fundamentally new to social and humanitarian knowledge. This concept is intended to denote the desire of society to average in terms of the manifestation of its activity, therefore, in stability. It is expedient to characterize such stable existence of a society as a social equilibrium, a special state of a social system that has developed due to historical continuity, which manifests itself in a whole set of factors of economic, social, ethnic, etc. character.

Deviation from the criteria of social stability is the essence of social balance and leads to a “lifting” of the social equilibrium curve, which necessitates social dampening as a significant manifestation of management processes that can, among other things, solve the problem of cognitive control of consciousness, differing in its mental manifestations.

“Leveling” of quantitative (digital) characteristics should be carried out in indicators acceptable for a given society. Using Big Data as the most sensitive tool for measuring and detailing biogeosocial processes allows for the necessary iterations that cannot be performed using conventional scientific tools, identifying the “dominant that directs the vector of development of a specific phenomenon or process” [1] and the trend of their changes.

The formulation of the problems presented by the authors of the study is also aimed at clarifying the issue related to the need to determine the vector of the social management algorithm itself, namely: if we set a completely obvious goal for us to ensure proper management of social systems, then the following question arises: what (as a social goal) should we control in this way. The use of Big Data technologies provides new opportunities for solving such problems. This, in turn, means the need to determine the factors and criteria for the stable state of structures, and more precisely, the control objects, in the system of social coordinates. The frequency of deviations from these indicators, as well as the ability of their leveling, according to the authors, will be a fundamental point in understanding the quality of social management, which actualizes the need to develop a methodology for assessing the degree of deviation of social processes and the qualitative state of the systems involved in them from their equilibrium states. This method should necessarily include the possibility of both direct and indirect impact on social systems, which, in fact, additionally reveals the above concept of “social dampening” as a fundamentally significant theoretical basis for the formation and subsequent implementation of managerial influences on society, allowing for reduce, by using Big Data, the peak values of social indicators that go beyond the social acceptability. If we ignore them it can not only

partially unbalance the social system, but make it fully socially uncontrollable and prone to social explosions.

At the same time, in addition to social processes leading to a deviation of the environment from the social "normal", which can be regulated by social dampening, it is necessary to consider in greater detail identifying the essence of the process of social resonance. Social resonance is such a state of interacting factors and criteria that characterize the social environment, which leads to an explosive deviation from the "normal". The "removal" of such social aggravations can be interpreted as a process of returning from social resonance to the norm, which seems to the authors an extremely important task of social management, the successful solution of which is determined by the need to take into account the effect of passionate interaction, understood as a kind of "activity manifested in the individual's striving for the goal (often - illusory) and in the capacity for superstressing and sacrifice in order to achieve this goal", at the same time "sacrifice is understood ... as refusal to satisfy immediate needs, sometimes essential to life, for the sake of the dominant social or ideal needs, perceived as a goal "with a predominance of "development needs"[1].

Thus according to the authors, Big Data makes it possible to form a model of "removing" of social deviations, including social resonance, arising from a combination of a wide range of factors, the interaction of which is very difficult to analyze and often invisible to the researcher through the prism of traditional cognitive tools. In its turn, the implementation in practice of this kind of social dampening model let us create a comprehensive understanding of social processes as a neurobiological manifestation of human activity and forms the necessary opportunities to overcome its negative consequences.

Hiding the analysis of additive and multiplicative effects behind the facade of large-scale computer calculations, Big Data makes it possible to ensure that objective decisions are made, especially significant in the management of multifactorial social and natural systems [2,3], which, thanks to the implementation of social dampening procedures, makes it possible to guarantee stability, manageability and predictability of biogeosocial processes.

The carried out research clearly convince that the methodology of this kind must consist of two semantic blocks: the methodological, structuring algorithm for the study of social systems and processes with Big Data technology and management, which determines the possible methods of management influence based on the carried out analysis.

The first methodological block supposes, first of all, isolation, accompanied by analysis and evaluation, in the sphere of the information space of multimodal clusters, the most explosive in terms of their potential to destabilize the social system, taking it out of relative equilibrium. Here, as global clusters, attention is drawn to the three classical spheres of society: political, economic and social. The undeniable scale of these spheres makes it possible to isolate subclusters in each of them with the possibility of further detailing them into groups (including interests) and IP addresses.

This specification is quite possible to carry out on the basis of keywords that are set in accordance with the research task and are used by the Big Data technology to isolate subclusters and elements of their internal structure. Further, the practical implementation of this stage of the proposed methodology will require the parallel development of the "system learning" algorithm, allowing the program to isolate words, terms, concepts, etc. for the subsequent clustering and identification of significant communications.

The next step of the first methodological block is the construction of an algebraic lattice of the number of links of multimodal clustering, which represents a peculiar coordinate system of the specific carried out analysis. It will allow modeling the system of links within the cluster. The increase in the number of links within the cluster (or its subsystem) is the main indicator of the subsequent system out of equilibrium - social resonance.

The greatest interest for the subsequent management impact should present connections that demonstrate sustainable growth. When they are identified, you should:

- precise the topics that ensure the applicability of these links;
- carry out an analysis of the frequency dynamics;

- identify danger zones for social resonance.

The completion of the first methodological block is the differentiation of the resonance rate. This involves the identification of areas of social, informational interactions on which there has been a steady growth; subsequent finding of stable relationships to identify the topic of interaction. At the same time, a sharp increase in the number of links per unit of time by an order of magnitude should be interpreted, within the framework of this methodology, as the main condition for the soonest occurrence of social resonance.

The second methodological block identifies possible options for a practical transition from gnoseological reasoning to direct management impact on the social situation. A management decision point as a manifestation of social dampening can be represented on the graph at the point of transition from a smoothly increasing curve, indicating an increase in the number of social connections (within the identified cluster, subcluster, etc.) to a vertical straight line, representing the ordinal increase in social connections.

In this case, the second methodological block admits the possibility of two management impacts:

- Management of the number of socially significant connections, which primarily involves minimizing them in order to prevent or level social resonance;

- Management of the social environment dissipativity, which provides for an instrumental, informational impact (including through informational attack or subject readdressing of links) on the identified cluster, an impact that does not allow the formation of social relations that could in the long run lead to social resonance.

Otherwise, the coincidence of the “resonance frequencies” in the process of intersubject interaction can cause a global, uncontrolled social resonance.

2. Application of the formal concepts analysis method for social networks

The methodology of social dampening consists of the social resonance definition in social networks groups and the search for control actions to reduce tensions. The methodology of social dampening is based on multimodal clustering of social networks. Clustering is based on the formal concept analysis method (Formal Concept Analysis, FCA) [4–9]. A large amount of structured and unstructured data generates trivial data.

According to the method of formal concepts, we introduce the following definitions:

G is a set of objects, M is a set of attributes, relation $I \subseteq G \times M$ such that $(g,m) \in I$ if and only if, the object *g* has attribute *m*. $K := (G,M,I)$ is called a formal context.

Let us define Galois operators in the following manner for $A \subseteq G, B \subseteq M$:

$$A' \overset{def}{=} \{ m \in M \,|\, g / m \forall g \in A \}, B' \overset{def}{=} \{ g \in G \,|\, g / m \forall m \in B \}.$$

A formal concept is the pair $(A,B): A \subseteq G, B \subseteq M, A' = B$ and $B' = A$, where A is a formal extent, B is a formal intent.

The concepts, ordered by the dependency $(A_1,B_1) \geq (A_2,B_2) \Leftrightarrow A_1 \supseteq A_2 (B_2 \supseteq B_1)$, form a complete lattice, which is called context lattice $\underline{\beta}(G,M,I)$.

The example of social networks context and their context hashtags is shown in Figure 1.

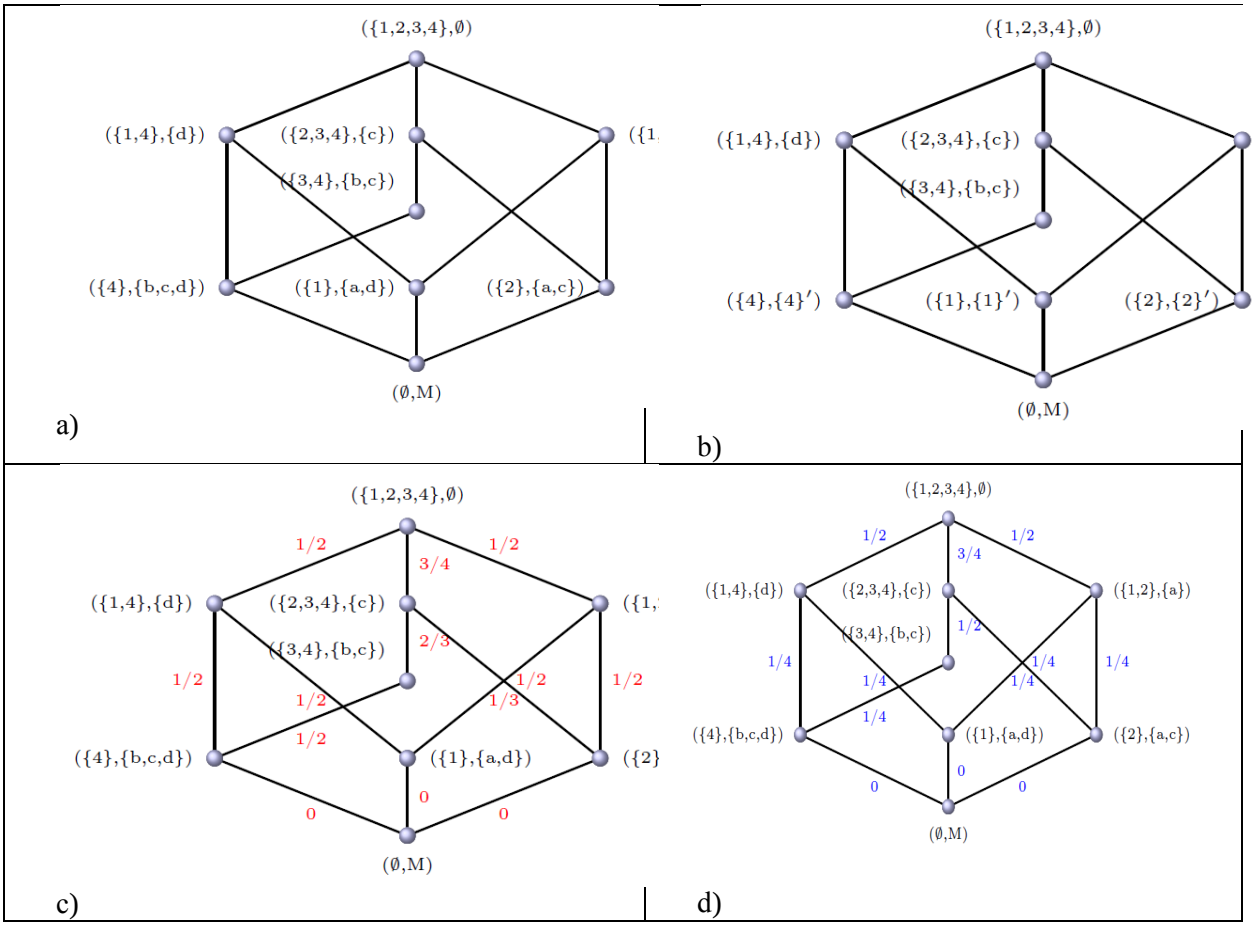

Figure 1. Social networks with context hashtags: A) context hashtags for social networks; B) social hashtags for social networks with implications; C) hashtags with reliability of associative links; D) the example of hashtags with support of association rules.

Let us consider implications in the lattice and define the operation "implication" in the following manner: implication $A \to B$, where $A,B \subseteq M$ exists if $A' \subseteq B'$, i.e. every object that has all attributes from A set also has all attributes from B set. Implications correspond to Armstrong's axioms: reflexivity $(A \to B / A \cup C \to B)$;augmentation$(A \to B / A \cup C \to B)$; pseudo-transitivity $(A \to B, D \cup B \to C / A \to C)$.

Among implications there is the Duquenne-Guigues basis of implications i.e. a minimal number of implications which can help to derive other implications using the Armstrong rules. The basis is sought through methods of machine learning: "object-by-object" algorithm of implication basis construction or through interactive learning procedure. As a result we have a hashtag with implications (Figure 2б).

Implications: $abc \to d, b \to c, cd \to b$ allow to redefine the nodes of hashtags.

Let us consider partial implications or association rules.

Definition 1. $A \xrightarrow[m,n]{} B$ is partial implication (association rule) of context (G,M,I), if $A,B \in M$; has a support of $sup\ p(A \to B) = |(A \cup B)'|/|G|$ has a confidence $conf\ (A \to B) = |(A \cup B)'|/|A'|$.

Further we consider the algorithm identifying association rules:

1) to find all frequent sets of attributes (with support which is not below the assigned);
2) it is enough to find all the frequent closed sets of attributes of the context.

The example of hashtags with reliability of associative links (Figure 1c).

The example of hashtags with support of association rules (Figure 1d).

Good rules with $sup\ p \geq 1/2$ and $min\ \ conf \geq 3/4$ are defined according to the following algorithm:

1. $0 \to c, sup\ p(0 \to c) = conf\,(0 \to c) = 3/4;$
2. $c \to b, sup\ p(c \to b) = 1/2, conf\,(c \to b) = 2/3.$

The use of this clustering method will allow determining the interest groups, with an increase in connections it will be necessary to make management decisions. In terms of information, it will be necessary to combine messages from groups with an increased number of links with calm "stable" groups in which the number of links does not undergo a drastic change.

3. Method of Using Data Mining in Social Dampening Methodology

Formal Concept Analysis (FCA) allows to establish stable links and cluster data (create new knowledge) using Armstrong rules in context lattices. To create a portrait and information model of resonant groups, it is necessary to use BIG DATA technologies [10, 11, 12]. The method of using data mining is as follows:

1. Forming a big data set in hadoop from twitter using the "Samara region" filter which reveals hit counts;
2. Separation of the formed set by various filters associated with the basic factors of resonance deviations;
3. Monitoring of streaming content analysis on filters;
4. Adoption of operational activities in cases of sustained "hits" in hit counts;
5. Development of a program in the Scala language for working with filtering in Big Data field;
6. Debugging and testing the program with a set of practical data;
7. Analysis of the calculation results.

The social network twitter is used to obtain data, since it is an "open" product, its use does not require additional investments, and 50% of Internet users have profiles in this program. Twitter is the second most popular network among users worldwide, second only to Facebook. However, unlike Facebook, which does not provide open access to its data, Twitter provides such access, there are no restrictions on access to server data sets. Users of this social network exchange mostly text information, which is an undoubted advantage in processing. Twitter is not a subject network and most widely reflects public opinion on many issues of interest, so for the formation of groups for analyzing the social resonance in the region, the processing of data from this social network was optimal.

To work with BIG DATA in social networks, it is necessary to use methods of collecting, processing and analyzing data. Data collection is carried out in real time, within a certain geolocation, or within the entire network, using certain patterns. Information of interest for the analysis in this area is: location, date and time, content, content "author" (user), communication between users. Data collection in social networks can be performed using the following tools: Apache Hadoop, Biglnsights (IBM), Cloudera, Hortonworks, Storm. Hortonworks was chosen to carry out research in the field of social dampening. The Twitter Application (apps.twitter.com) was used for work, in which key parameters were defined and refined: API key, API secret, Access token, Access token secret.

To collect data using the Hortonworks, Twitter App, the flume service configuration file was used in the Hortonworks Sandbox virtual machine. After installing the Hortonworks_ Sandbox version 2.3 virtual machine and setting up the flume service, the system is ready to download data from twitter. To view and download the downloaded files, go to the HDFS folder, where we process the data. View of the HDFS file structure in the Hortonworks virtual machine when solving a problem in the sphere of social dampening is shown in Figure 2.

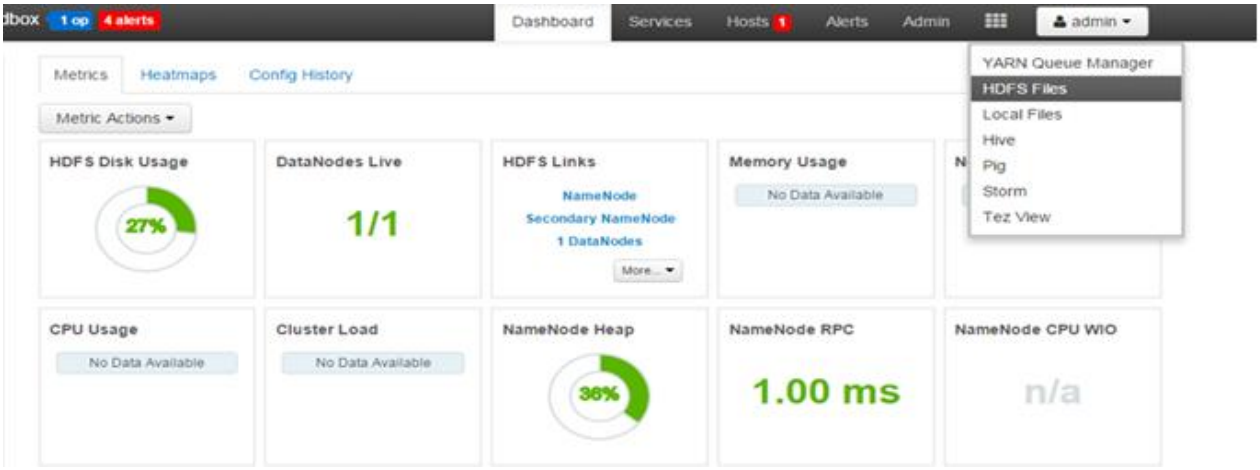

Figure 2. HDFS visualization in Hortoworks when file download to solve social dampening.

The collected data must be structured (i.e., processed) in accordance with the MapReduce paradigm. MapReduce is a framework for performing distributed tasks using a large number of computers forming a cluster.

Using MapReduce allowed us to structure the flow of data from social networks by criteria: fonts, text size, color, link to user profile, location, time, and so on.

To determine the portrait of the respondent, the following types of data are needed: location, text, language, and time. In order to extract only this information, you can use the MapReduce technology built into the Hortonworks Sandbox tool. For data processing, we use the Hive DBMS in the Hadoop environment, which allows performing operations on data and their analysis using SQL-like queries. To do this, create a file for processing and creating the necessary hivedll.sql tables.

Run this file using command: Hive_f hiveddl.sql. Structured data will be presented in the Table 1.

Table 1. Headings to analyze structured XML data in tasks for social dampening.

A	B	C	D	E	F
Data/Time	Time/Zona	language	Text	location	Sentiments

For data analysis the following variables are used. Total amount of twitts (Kol_i) for every location (R) is defined by:

$$Kol_R = \sum_{i=1}^{N} k_i, k_i \in R,$$

where k_i is every next twitt in the considered stream.

Frequency of unique word usage ch(m) is defined from the general variety of L text data:

$$ch(m) = \sum_{i=1}^{N} m_i, m_i \in L.$$

The attitude of every twitt otn (m,rez) may be defined from the thesaurus tez, where the attitude to this word is written up:

$$otn(m, rez) = \begin{cases} 0, m - negative\_meaning \\ 1, m - neural\_meaning \\ 2, m - positive\_meaning \end{cases}.$$

For further work, a dictionary was compiled, consisting of domain filters, to further determine the number of tweets for placement ch (m) and the number of tweets for placement taking into account the relation otn (m, rez). We define a thesaurus taking into account filters by basic factors: salary, unemployment, political developments and housing and utility infrastructure. As a result, we obtain 4 basic factors of resonance deviations.

«Salary» factor P_1 gives the number of twitts in general quantity of text data L:

$$Kol_{otnP_1} = \sum_{i=1}^{N} S_i(S_i \in P_1)/L = 10\%.$$

«Unemployment» factor P_2 gives the number of twitts in general quantity of text data L:

$$Kol_{otnP_2} = \sum_{i=1}^{N} S_i(S_i \in P_2)/L = 9\%.$$

«Political» factor P_3 gives the number of twitts in general quantity of text data L:

$$Kol_{otnP_3} = \sum_{i=1}^{N} S_i(S_i \in P_3)/L = 7\%.$$

«Infrastructure» factor P_4 gives the number of twitts in general quantity of text data L:

$$Kol_{otnP_4} = \sum_{i=1}^{N} S_i(S_i \in P_4)/L = 13\%.$$

4. Results and discussion

As a result, it can be concluded that what factors of deviations are associated with the Samara region. According to Figure 3 it can be seen that the main factor of discontent is the housing sector.

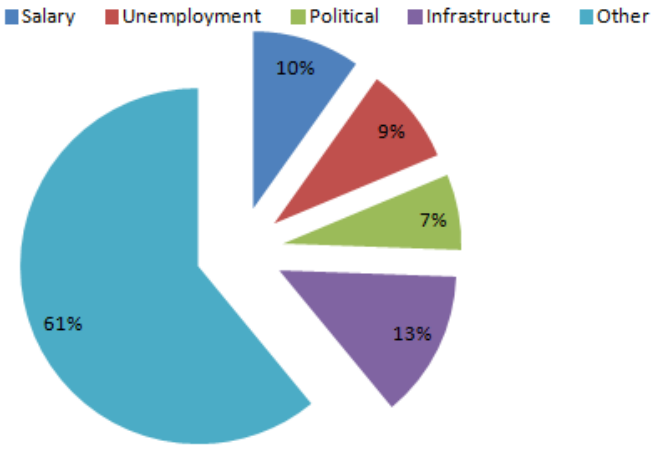

Figure 3. Factors of social dampening in Samara Oblast.

Due to BIG DATA technology, it is possible to store and update data in the “hadoop” file system using the “Samara Oblast” filter (filter1 = {Samara Oblast}). Then, it is necessary to filter this area by the basic factors of social dampening, by installing, for example, the following filters: Filter2 (salary) = {money, ruble *, dollars *, currencies *, crypt *}; Filter3 (unemployment) = {job search, engineer *, worker *, build *}; Filter4 (political developments) = {elections, deputy *, penny *, administrator *}; Filter5 (housing and utility infrastructure) = {garbage collection, pipes *, water *, gas *}.

The set of descriptors by which the Internet discourse will be filtered is determined by the lexical representatives of the concept formed in the world picture by the average Russian-speaking consumer.

To make decisions in the field of social resonance in the region, multimodal clustering of social networks was carried out. A large number of structured and unstructured data of social sites in the considered area can be represented as the next triple (user, group, interest) (Figure 4).

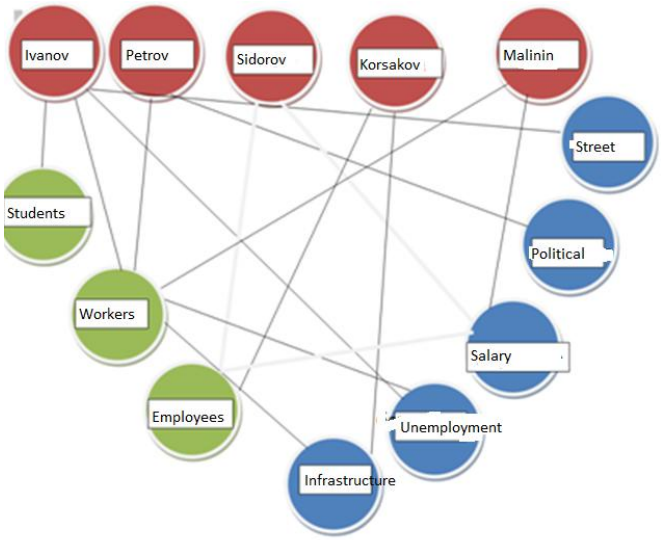

Figure 4. Data for social resonance analysis from Twitter social network as a graph

Using the method of formal concepts, we form a complete lattice, called the context lattice $\underline{\beta}(G,M,I)$. An example of the context of social networks in the field of social dampening and their context hashtags are shown in Table 2 and Figure 5.

Table 2. The example of Social network context and its context hashtags (a - attributes on «salary» filter, b - attributes on «unemployment» filter, c - attributes on «political developments» filter, d - attributes on «housing and utility infrastructure» filter).

	G/M	a	b	c	d
1	pensioners	x			x
2	employees	x		x	
3	workers		x	x	
4	students		x	x	x

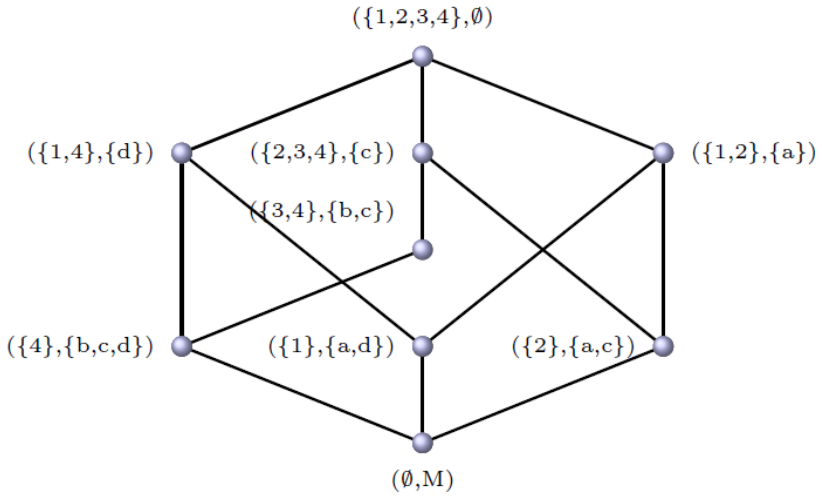

Figure 5. Context hashtags for social network.

The use of this clustering method will give the opportunity to determine interest groups, with an increase in links in which management decisions will be required. But this tool has limitations on use. Users who work with the social network "Twitter" are in the "students" group and partly in the "employees", "workers" groups and only slightly affect the "retirees" group, so for complete management decision making it is necessary to add new groups.

You can get graphs of the user hit counts by filters on data collection time (Figure 6). The data collection time from the Internet is unlimited in BIG DATA technology.

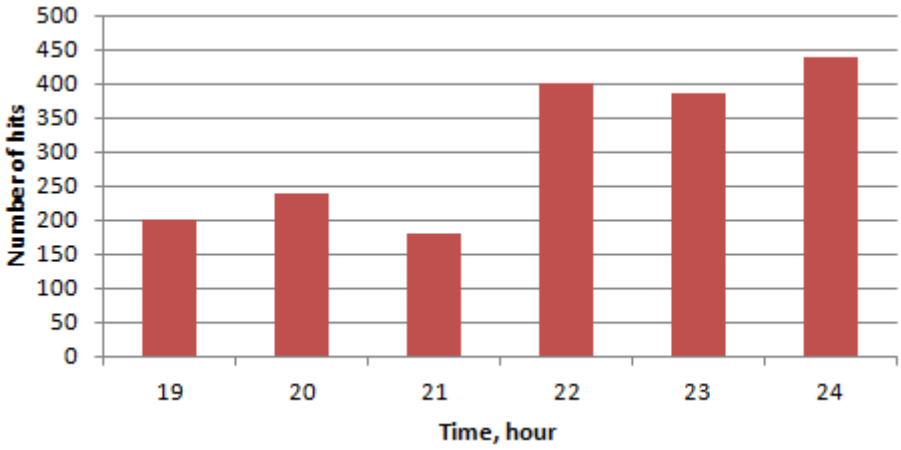

Figure 6. Dependency graph of hit counts using filters from data collection time.

As a result, we obtain a dynamic change of information in real-time from the Internet, which allows monitoring the streaming analysis of unstructured information (In-Memory Data Processing and Stream technology) with minimal investment by filters. To implement this method, a program in the Scala language was written.

After the work of the program, we obtain a dynamic change of parameters in the BIG DATA environment, which allow us to determine social resonance zones in the region, taking into account unstructured information. If steady "bursts" of data on hits counts are detected on charts in accordance with the forms of resonance, management regulation for this type of activity in the region should be implemented.

Thus, a tool has been proposed to increase the effectiveness of work in the field of social dampening in the region. This is the most important task under modern economic conditions, the basis of which is the possibility to make optimal management decisions. The proposed method of regulation can be effective in managing the processes of social dampening of a region, which are characterized

by a variety of forms and a wide range of components and factors, as well as an inherent dynamics of development and active transformation of life activity.

At the same time, the use of modern software and hardware gives the opportunity to make the assessment and visualization of changes in fact in real time, which can be useful for local authorities. It is important to note that social dampening is not a strict limiter of social actions "in amplitude". It only softens social actions, allowing them to manifest themselves in other areas, it does not "break" the social system, but allows it to transform, creating the visibility of smooth compliance with the requirements of the social interaction subjects.

5. References

[1] Bodrov A A, Ramzaev V M 2015 Philosophic aspects of developing new knowledge under data intellectual analysis (BIGDATA) *CEUR Workshop Proceedings* **1490** 338-345
[2] Khaimovich I N, Ramzaev V M and Chumak V G 2016 Use of big data technology in public and municipal management *CEUR Workshop Proceedings* **1638** 864-872
[3] Khaimovich I N, Ramzaev V M and Chumak V G 2015 Challenges of data access in economic research based on Big Data technology *CEUR Workshop Proceedings* **1490** 327-337
[4] Wille R 1982 Restructuring lattice theory: An approach based on hierarchies of concepts *Ordered Sets. NATO Advanced Study Institutes Series (Series C – Mathematical and Physical Sciences)* **83** p 470
[5] Ganter B, Whille R 1996 *Formale Begriffsanalyse* (Springer, Heidelberg) p 540
[6] Ganter B, Whille R 1999 F*ormale Consept Analysis* (Springer) p269
[7] Davey B, Priestly H 1990 *Introduction to Order and Lattices* (Cambridge University Press, Cambridge) p 460
[8] Denecke K, Erne M and Wismath S L 2004 *Galois Connections and Applications* (SpringerScience and Business Media) **565** p 498
[9] Bonacich P 2007 Power and Centrality: A Family of Measures *American Journal of Sociology* **92(5)** 1170-1182
[10] Chumak P V, Ramzaev V M and Khaimovich I N 2015 Models for forecasting the competitive growth of enterprises due to energymodernization *Studies on Russian Economic Development* **26(1)** 49-54
[11] Khaimovich A I, Grechnikov F V 2015 Development of the requirements template for the information support system inthe context of developing new materials involving Big Data *CEUR Workshop Proceedings* **1490** 364-375
[12] Kazanskiy N L, Stepanenko I S, Khaimovich A I, Kravchenko S V, Byzov E V and Moiseev M A 2016 Injectional multilens molding parameters optimization *Computer Optics* **40(2)** 203-214 DOI: 10.18287/ 2412-6179-2016-40-2-203-214