Information systems modeling

UDC 004.04:004.67:004.77

Mykhailo Mozhaiev¹, Pavlo Buslov²

Hon. Prof. M.S. Bokarius Kharkiv Research Institute of Forensic Examinations, Kharkiv, Ukraine
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

METHOD OF MODELING OF A SOCIAL PROFILE
USING BIG DATA STRUCTURE TRANSFORMATION OPTIMIZATION

Abstract. The object of the research are methods and algorithms of optimizing of the Big Data transformation to build a social profile model, the subject of the research are methods of constructing of a social profile. For decision-making person, the problem of scientific methodological and instrumental re-equipment is relevant for the effective fulfillment of a set of managerial tasks and confronting of fundamentally new challenges and threats in society. This task is directly related to the problem of building of a model of the social profile of both the individual and the social group as a whole. Therefore, the problem of optimizing of methods of constructing of a mathematical model of a social profile is certainly relevant. During the research, methods of the mathematical apparatus of graph theory, database theory and the concept of non-relational data stores, Big Data technology, text analytics technologies, parallel data processing methods, methods of neural networks' using, methods of multimedia data analyzing were used. These methods were integrated into the general method, called the method of increasing of the efficiency of constructing of a mathematical model of a social profile. The proposed method improves the adequacy of the social profile model, which will significantly improve and simplify the functioning of information systems for decision-making based on knowledge of the social advantages of certain social groups, which will allow dynamic correction of their behavior. The obtained results of testing the method make it possible to consider it as an effective tool for obtaining of an objective information model of a social portrait of a social group. This is because the correctness of setting and solving of the problem ensured that adequate results were obtained. Unlike the existing ones, the proposed modeling method, which uses an oriented graph, allows to improve significantly the quality and adequacy of this process. Further research should be directed towards the implementation of proposed theoretical developments in real decision-making systems. This will increase the weight of automated decision-making systems for social climate analysis.

Keywords: social profile model; information decision-making systems; graph theory; Big Data technologies; social portrait of a social group.

Introduction

In the context of development of digital economy and informatization of most social processes, the tasks of managing of social and economic systems acquire a new specificity, both in terms of goals and in terms of methods and tools used. For the decision-making person, the problem of scientific, methodological and instrumental re-equipment is relevant for the effective fulfillment of a set of managerial tasks and confronting fundamentally new challenges and threats in society. Mathematical, informational and technical support is needed to enable collection, storage, monitoring, and analysis of large amounts of heterogeneous data generated by all available Internet resources and services (information and news portals, personal blogs, thematic forums, etc.) with the possibility of social profiling (individuals, various groups, social phenomena, events, processes), establishing explicit and hidden relationships between personal and group profiles, as well as identifying causal dependencies in the information field of the society.

Accordingly, new opportunities open for solving very complex and ambiguous problems related to assessment of social tension in the context of individual focus groups, forecasting the consequences of introduction of certain innovations in the life of society, identification and verification of the digital personality (to prevent compromising people by attackers using unauthorized access to personal accounts, counterfeiting accounts using technical means and artificial intelligence methods), developing service-oriented intelligent Internet technologies, and so on.

A special promise is integration of the latest information technologies related to collection and processing of large unstructured data (Big Data), intelligent data analysis (Data Mining) and distributed databases.

1. Research object and its technological audit

Big Data [1], a modern IT industry, includes a range of methods, tools, and technologies for collecting and analyzing of large amounts of different structured and unstructured data in order to achieve efficient results in the context of information distribution across multiple nodes of a computing network and its continuous updating. It is also permissible to apply the said period to the data itself, which are processed using this technology. Initially, the term Big Data was used in the academic environment to solve problems related to growth and diversity of data.

The first Big Data-based solutions appeared in the second half of the zero years and were considered as alternative ones to the classic relational DBMS in Business Intelligence [2].

The technology in question can be applied in cases where there is too much data to be processed by conventional means, in particular a server with a
The purposes and tasks of the research

Within the framework of this research, the need to study and use a number of technological and theoretical tools is revealed: the mathematical apparatus of graph theory, database theory and the concept of non-relational data stores, Big Data technology, text analytics technologies, parallel data processing methods, and methods of neural networks' using, methods for analyzing multimedia data.

The purpose of this article is to improve the methods of automated transformation of unstructured social data from public sources into a structured picture of a person or group of persons by developing algorithms for enumerating of sections, which contain methods of formation of new sections, method of formation of new sections from already built ones, a procedure for eliminating excess elements and a search order.

To achieve this purpose, the research needs to solve the following private tasks:
1. To investigate methods and tools used in automated transformation of unstructured social data.
2. To develop algorithms of sections' enumeration that improve Big Data conversion efficiency.

Research of existing solutions of the problem

Social profiling results based on analysis of public sources of information should have a number of properties, such as explicit structuring, scalability, presentation both in the form of tables and images (graphs, diagrams) [13-18]. This is necessary to enable them to be used in social research applications. Currently, the most relevant areas of application of social profiles are marketing and law enforcement [9-12]. The first group includes collaborative filtering tasks, the solution of which allows improving the quality of advertising campaigns by taking into account the interests of the target audience and predicting its reaction. As a result, it allows us to optimize the costs of marketing research (surveys' conducting, target groups' forming, etc.) and to promote goods and services purposefully, as well as to increase customer loyalty. In the field of law enforcement, the task of building of social profiles intersects with open source intelligence (OSINT). Law enforcement agencies use personal and group profiles to assess the level of social tension and possible reactions of society to provocative actions, to create the "risk groups," and to collect the files for investigative and judicial measures.

Research methods

Within the framework of this research, the need to study and use of the following number of technological and theoretical tools is revealed: the mathematical apparatus of graph theory, database theory and the concept of non-relational data stores, Big Data technology, text analytics technologies, parallel data processing methods, methods of neural networks' using, methods of multimedia data analyzing. Modern developments in the developed countries of the world in the field of information, processing and transmission...
have led to study of oriented information networks, which have wide practical applications.

These networks can often be mathematically described as an oriented graph. We call (s, t) a network when it is necessary to distinguish it from a network with another source and drain. Let us denote it by G(X, G). In this case

\[ X = s \cup t \cup \{x_i\}_{i=1}^{N} \]

is a set of vertices of the network, G – the reflection of X in X, which generates the arcs of the network.

The vertex of the orgraph will be called as terminal in one of two cases:
- No arc enters it;
- Not a single arc comes out of it. In the sense of this definition, the network vertices s and t are terminal.

The arcs and vertices of network G are called network elements; we denote their set as E.

For network G, we introduce the ratio of the continuity of elements in this way, if i is the vertex of the network, than \( \sigma(i) \) is the set of arcs \& from it, and if i is an arc, than \( \sigma(i) \) is the vertex into which this arc enters.

Let it be \( i \in \sigma^{-1}(j) \), if \( j \in \sigma(i) \). In other words, \( \sigma^{-1}(j) \) it is a plurality of elements, for which element j is a successor. Let us note that the continuity ratio of \( \sigma \) elements can be entered on any oriented graph without loops. A network G with a plurality of elements E, for which the ratio \( \sigma \) is given, is denoted through \( G(E, \sigma) \).

Let us consider that the network elements are set by numbers.

We call the correct numbering of elements (by analogy with the correct numbering of vertices) such one that if \( j \in \sigma(i) \), than \( i < j \), where \( i, j \) are the numbers of the network elements.

Let’s note that the elements of the finite orgraph G \( G(E, \sigma) \) without loops can always be correctly numbered. Indeed, it is possible to construct a one-to-one mapping (of graph \( G(E, \sigma) \)) to graph \( F(X, G) \), in which the set of vertices \( X \) corresponds to the set \( E \) of elements of graph G, and the relation of the contiguity of vertices \( G \) is the ratio of the continuity of elements \( \sigma \) on graph G. It is known that on graphs of the form \( (X, G) \), for vertices it is always possible to set the correct numeration [13-20]. Therefore, it can be set on the set \( E \) of elements of the graph \( G(E, \sigma) \), which follows from the mutual uniqueness of the mapping \( \psi \). Let us call with a mixed \( (s, t) \) section or a mixed terminal section such a set of \( E \)-elements of the network that if they are removed from it, then there will be no terminal path.

The element \( l \) is called as redundant in section \( P \) if the set of elements \( \{P \setminus l\} \) is also a section [21].

A section that does not contain redundant features is minimal, and a section that contains redundant features is redundant. In an excess section view, there is always a feature that can be deleted without violating a specific section property. Let us note that from one excess section we can sometimes form several minimal sections, except various excess vertices.

As a source section, we take a set of elements

\[ P_0 = \sigma(s) \].

Clearly, there can be no redundant elements in it. In the original section, the elements are arranged in ascending order.

From each section \( P = \{l_1, l_2, \ldots, l_n\} \), which was formed, we can create \( l \leq n \) new sections.

If \( t \in \sigma(l_j) \), then

\[ P_t = \left(\left(P \setminus l_j\right) \cup \sigma(l_j)\right) \setminus I_t \],

(1)

where \( I_t \) – is the set of redundant elements in \( P_t \).

In each newly constructed section, we will tidy up the elements by number growth, and replace the elements with the same numbers with others. All formed sections are entered in the general list of sections in lexicographic order by growth of numbers of composite elements. Let us note that when creating cuts with \( P \), replacing various elements \( l_j \) with \( \sigma(l_j) \) the same minimum section can be formed several times, which should be listed only for the first time.

Due to the correct numbering of the network elements, each newly constructed section is placed in a lexicographically ordered list after the section on which it is built.

Let us consider a way to exclude redundant elements from each \( P \)-built section

\[ \tilde{P} = \left(P \setminus l_j\right) \cup \sigma(l_j) \].

(2)

In the section, will \( \tilde{P} \) some element \( r_j \) be redundant if all \( (s, r_j) \) paths or all \( (r_j, t) \) paths pass through some element \( r_k \in P \). But for the accepted method of sections’ creating on \( (s, r_j) \) paths of similar elements there can not be \( r_k \) in section \( \tilde{P} \). Therefore, the search for redundant elements should be done, considering only segments of \( (r_j, t) \) paths.

We will put in order the elements by number growth in the generated section \( \tilde{P} \). Then, to check for redundancy \( r_j \) concerning overshoot we can take into account the pieces of \( (r_j, t) \) paths, on which the number of the last element is no more than the number of the last element in the section \( \tilde{P} \).

Let us denote as

\[ \tilde{P} = \{r_1, r_2, \ldots, r_m\} \]

a section, that is formed from a section

\[ P = \{l_1, l_2, \ldots, l_n\} \]

by replacing \( l_i \) with \( \sigma(l_i) \). The procedure of excluding of redundant elements is the following.

The viewed sections are selected from the list of constructed sections in lexicographic order by their component numbers. When we view a section from it, new sections are generated by replacing those elements \( l_i \) for which \( \sigma(l_i) \) do not have a t drain.
Let us consider how to convert a section list to a network. In a formalized form, this task can be formulated as follows.

There is a list $\mu$, consisting of $M$ rows $S_1, \ldots, S_M$, which has the following features.

1. The row $S_i = (a_{i1}, \ldots, a_{im})$ contains $m_i$ non-repeating alphabet characters $A = \{a_i\}_{i=1}^N$.
2. Any character permutation in a row is allowed.
3. There are no $S_i, S_j \in \mu$, such that $S_i \preceq S_j$.
4. There is at least one character pair $a_i, a_j$, for which no row $S_k \in \mu$ exists, such that $a_i \in S_k$, $a_j \in S_k$. In this case, we will say that the symbols $a_i$ and $a_j$ are incompatible.

We need to build the minimum orthograph $R(\Gamma, F)$ by the number of vertices, which has such properties.

In fact, this means that we need to define the order of characters in the list rows in order to combine the maximum number of matching characters in the rows and the incompatible characters would be in one row.

Let’s select the three characters $a_i, a_j, a_k$ such, that symbol $a_i$ belongs to two rows at the same time $S_p \cap S_q$.

Let us introduce such concepts of form as direct form:

$$f = a_i(a_j, a_k),$$

and inverse form

$$\bar{f} = (a_j, a_k)a_i.$$

The forms $f_1 \land f_2$ we call as equivalent if $f_1 = a_i(a_j, a_k), a = a_i(a_k, a_j)$. The forms $f_1 \land f_2$ we call as alternative to each other if $f_1 = \bar{f}_2$ or $f_2 = \bar{f}_1$. Let’s talk that $f_i$ contradicts $f_j$, if the general system of inequalities formed by forms $f_i \land f_j$, is non-essential. It is obviously, that the form $f = a_i(a_j, a_k)$ can only conflict with the following forms:

$$f_1 = a_j(a_i, a_k); f_2 = a_k(a_i, a_j);$$

$$f_3 = (a_j, a_i)a_k; f_4 = (a_k, a_i),$$

where $\ell \in \{1, N\}$.

Let's compose the initial system of forms $\Sigma = \Sigma_0$, that contains straight forms for all such triples of characters. Let us number the forms of this system (let them be $Q$ pieces) and compose a matrix of contradictions of forms.

5. Research results

We will change the system of forms $\Sigma$, replacing, where it is necessary, the forms that contain it, with alternative ones to receive a system $\Sigma = \Sigma^*$ containing a minimum number of conflicting forms.

The value $\delta_i > 0$ is the indicator of the need to replace the form $f_i$ with $f_i^*$.

We introduce auxiliary vectors $D^+, D^- \subset \Delta$ into consideration, in which, when replacing the form $f_i$ with $f_i^*$ i-row $i$ and i-column of matrix $V$, is changed, as well as some components of the vectors $D^+, D^- \subset \Delta$ according to the following rule:

1. We replace $v_{ij}$ with $-v_{ij}$ in the i-column of $V$ matrix, $j = 1, \ldots, Q$, at the same time

$$d_i^+(\text{new}) = d_i^-(\text{old}), \quad d_i^-(\text{new}) = d_i^+(\text{old}), \quad \delta_i(\text{new}) = -\delta_i(\text{old}).$$

2. We replace $v_{ki}$ with $-v_{ki}$ in the i-column of $V$ matrix, $k = 1, \ldots, Q$.

Doing the same for those $f_i$, in which $\delta_i > 0$, we gradually reduce the total number of contradictions in the system $\Sigma$, until it turns out that all $\delta_i \leq 0$, $i = 1, \ldots, Q$, and get a system of forms $\Sigma = \Sigma^*$ with a minimum number of contradictions. It will give a partial (or complete) order on the set of characters $A$, which will allow combining the maximum number of characters in rows to construct a graph of the necessary form.

6. SWOT analysis of research results

**Strengths.** The strengths of this research are that a comparative analysis of methods of personalized social data processing was carried out. Besides, the compilation of algorithms of analyzing of heterogeneous text and multimedia information, storage and presentation models of initial, intermediate analyzed and resulting data were made. It was established that the most promising is the use of the mathematical apparatus of graph theory, database theory and the concept of non-relational data stores, Big Data technology, text analytics technologies, parallel data processing methods, methods of neural networks' using, methods of multimedia data analyzing.

As a result, an oriented orthograph was built, which is a mathematical model of an oriented information system.

**Weaknesses.** The weaknesses of this research are related to the fact that the result is only a model of an oriented information system, which does not fully reflect the entire variety of existing and promising systems of processing of personal data and building of a social profile of both individuals and social groups of various differentiations.

**Opportunities.** The additional possibilities when using the above results in systems, as well as identification and verification of a digital image of a person are to optimize the required time and to improve the quality of such systems.

**Threats.** The options proposed in this paper are theoretical. Practical processing and justification prior to possible implementation solutions are required.
Conclusions

1. To improve the efficiency of automated conversion of unstructured social data, it is desirable to use OSINT technologies and supplement them using methods and algorithms to optimize Big Data transformation by building oriented graphs.

The rationale for using such an apparatus is the correct application of the theory of processing of large amounts of data and the theory of oriented graphs. The result of implementation of the algorithm is the ability to build an oriented graph, which is a mathematical model of oriented information networks that are quite often used in practice.

2. Thus, we have gained a new opportunity to further improve the functioning of computer information networks by optimizing of data structures by converting the information list into a corresponding oriented graph.

3. Further research should be directed towards the implementation of proposed theoretical developments in real decision-making systems. This will increase the weight of automated decision-making systems for analyzing of the social climate of society.

REFERENCES


Received (Надійшла) 28.10.2020
Accepted for publication (Прийнята до друку) 27.01.2021
Метод моделювання соціального профілю
з використанням оптимізації перетворення структур Big Data

М. О. Можаєв, П. В. Буслов

Анотація. Об’єктом дослідження є методи та алгоритми оптимізації перетворення Big Data для побудови моделі соціального профілю, предметом дослідження – методи побудови соціального профілю. Для особи, яка приймає рішення, актуальна проблема наукою методологічного інструментального переоснащення для ефективного виконання комплексу управлінських завдань і противостання принципово новим викликом і загрозам у суспільстві. Це завдання безпосередньо пов’язана з проблемою побудови моделі соціального профілю як індивідуума, так і соціальної групи в цілому. Тому задача оптимізації методів побудови математичної моделі соціального профілю є безумовно актуальною. В ході дослідження використовувалися методи математичного апарату теорії графів, теорії баз даних і концепції нереляційних сховищ даних, технології Big Data, технологій текстової аналітики, методів паралельної обробки даних, методів застосування нейронних мереж, методів аналізу мультимедіа даних. Дани методи були інтегровані в загальний метод, названий методом покращення ефективності побудови математичної моделі соціального профілю. Запропонований метод використовує результати проаналізованого граф дозволяє суттєво покращити якість та адекватність цього процесу. Запропонований метод дозволяє покращити адекватність моделі соціального профілю, що дозволить здійснювати динамічну корекцію систем приняття рішень для аналізу соціального клімату суспільства.

Ключові слова: модель соціального профілю; інформаційні системи приняття рішень; теорія графів; технології Big Data; соціальний портрет соціальної групи.

Метод моделювання соціального профілю
с використанням оптимізації перетворення структур Big Data

М. О. Можаєв, П. В. Буслов

Анотація. Об’єктом дослідження є методи та алгоритми оптимізації перетворення Big Data для побудови моделі соціального профілю, предметом дослідження – методи побудови соціального профілю. Для особи, яка приймає рішення, актуальна проблема наукою методологічного інструментального переоснащення для ефективного виконання комплексу управлінських завдань і противостання принципово новим викликом і загрозам у суспільстві. Це завдання безпосередньо пов’язана з проблемою побудови моделі соціального профілю як індивідуума, так і соціальної групи в цілому. Тому задача оптимізації методів побудови математичної моделі соціального профілю є безумовно актуальною. В ході дослідження використовувалися методи математичного апарату теорії графів, теорії баз даних і концепції нереляційних сховищ даних, технології Big Data, технологій текстової аналітики, методів паралельної обробки даних, методів застосування нейронних мереж, методів аналізу мультимедіа даних. Дани методи були інтегровані в загальний метод, названий методом покращення ефективності побудови математичної моделі соціального профілю. Запропонований метод використовує результати проаналізованого граф дозволяє суттєво покращити якість та адекватність цього процесу. Запропонований метод дозволяє покращити адекватність моделі соціального профілю, що дозволить здійснювати динамічну корекцію систем приняття рішень для аналізу соціального клімату суспільства.

Ключові слова: модель соціального профілю; інформаційні системи приняття рішень; теорія графів; технології Big Data; соціальний портрет соціальної групи.