Big Data Analysis for Demand Segmentation of Small Business Services by Activity in Region

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

[1]*International Market Institute, Aksakova street, 21, 443030, Samara, Russia*
[2]*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

Annotation

The article suggests a tool for the efficiency improvement in using budget funds in the region in the sphere of small business. This is the most important task in the current economic conditions, in which solution there is a possibility of making effective management decisions. The suggested method of regulation based on the analysis of social networks using BIG DATA technology can be effective in managing various innovative processes of economic development in the region, which are characterized by a variety of forms and a wide range of components and factors, as well as dynamic development and active transformation of life. The use of modern software and hardware from BIG DATA technology allows real time evaluation and visualization of changes

Keywords: competitiveness; territory management; intensive data; mathematical models; BIG DATA technology

1. Introduction

Under modern social and economic conditions the vital task is the state regulation of market economy players, among which one of the most important in the region is small business (SB). The foreign experience shows that without this sector it is impossible to develop economy as far as the economic growth rate depends on it as well as the structure and the quality of up to 40-50% of gross national product.

2. Subject of research

The structure of small and medium-sized enterprises by types of economic activity is varying. As it can be seen in Figure 1, the largest number of enterprises are engaged in trade, repair of motor vehicles, motorcycles, household products and personal items, which is explained by lower barriers to entry these areas of activity.

The following features of SB development management can be distinguished. First, it is necessary to note a wide range of services provided by SB subjects, as well as a huge range of goods sold by them. Secondly, the SB differs significantly in being more mobile in comparison with the large one. By the mobility we mean a continuous change in the market conditions, the closure of old and the emergence of new economic entities, which is explained by the high variability in tastes and preferences of consumers of goods and services of SB entities, i.e. there is a quite active process where some types of activities are substituted by others, determined by a change in consumer demand, which is especially important under the modern conditions of import substitution. Thus, according to the statistics, up to 85% of the new entities of the SB is closed during the first year of its existence. 94 out of the 100 registered small businesses stops operations by the fourth year.

In this regard, the use of traditional methods of public administration, based on the data of monthly, quarterly and annual statistics, does not bring the expected result and does not allow us to identify trends for the development or closing up of certain activities, therefore, often making decisions about financial support and funds allocation for some projects is significantly behind the needs, and in some cases also contradict the changed real market situation by the time the financing begins.

For example, at the present time in the Samara region, state support for small and medium-sized enterprises is being implemented within the framework of the State Program “Development of Entrepreneurship, Trade and Tourism in the Samara Oblast” for 2014 - 2019, approved by the Government of the Samara Oblast Decree No. 699 dated November 29, 2013. Support to businessmen of the Samara Oblast is maintained in different directions and consists of information and consulting services, training, financial assistance, assistance in selling goods and services.

At the same time, it should be noted that, despite a number of measures used by the authorities in the region to manage the development of SB, effective methods for selecting priority directions for the development of SB have not been developed so far, which make it possible to direct budgetary funds to the development and support of entrepreneurs more appropriate [1-4]. The market of small and medium-sized businesses is a quite dynamically changing environment. It is necessary to take this into account in the medium and long term planning and regional authorities should take it as a basis for the support and stimulation of the development of the most high-demand areas of the SB activities and for monitoring the effectiveness of budgetary funds application for programs in this field of entrepreneurship under the changing market conditions.

3. The methods of applying the business intelligence at determining small business segments in the region

This task may be solved with the help of modern information technologies [5,6,7], to which BIG DATA technology refers, directly connected with business intelligence [8,10]. Along with this application of modern BIG DATA technologies provides an opportunity to distinguish the zones – territories of the most active consumption and demand for some or other products and services on the market in real time mode.

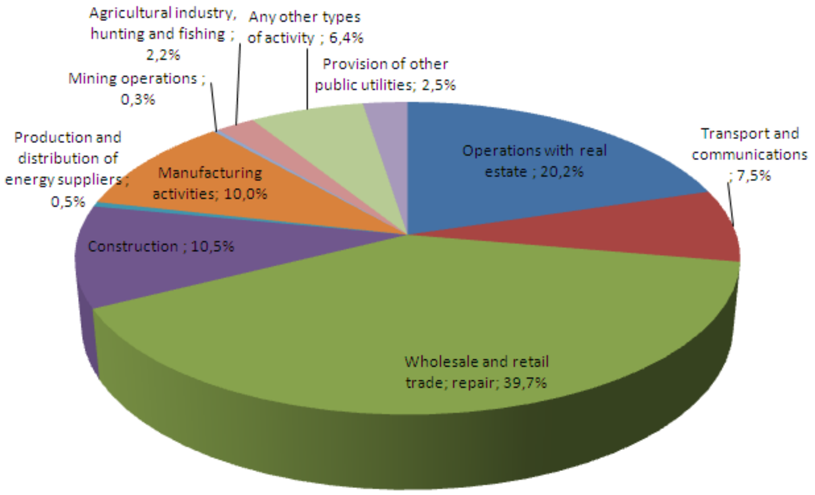

Fig. 1. The structure of small and medium-sized enterprises by types of economic activity at the end of 2014, %.

To manage the development of small and medium-sized businesses in the region the special methodology was worked out based on the BIG DATA technology [9], which consists of the following stages: identification of the role and place of small business in the region; identification of the main types of goods and services, offered by small businesses in the region; creation of consumer profile who uses the services of small business; creation of information model of small business consumer in the region; formation of small business zones in the region; development of guidelines for management decision making.

If the role and place of small business in the region, main types of goods and services, offered by entrepreneurs in the region were analyzed then to create a consumer profile and information model it is necessary to use BIG DATA technology. The method of applying the business intelligence is as follows:

1. Formation of a set of BIG DATA in hadoop from twitter using filter Samara Oblast, showing the hit count;
2. Division of formed set into different filters connected with basic factors of small business;
3. Carrying out monitoring of flow content analysis in filters;
4. Taking quick actions in cases of stable “burst” of hit count;
5. Program development in Scala language to work with filtration in the BIG Data area;
6. Program debugging and testing with a set of practical data;
7. Analysis of computational results.

To receive data we use social network «twitter», as it is “open” product, its application does not require any additional investment, and 50% of Internet users have profiles in this program. Twitter is the second in popularity network among the users in the entire world, come second only to Facebook. However unlike Facebook, which does not make accessible its data, Twitter provides such access, there are no limitations in access to the sets of data in the server. The users of this social network share mainly text messages, and this fact is absolute advantage while processing. Twitter is not a network with a specific focus and more broadly reflects public opinion in many points of interest, that is why the processing of data from this social network was the best possible to form small business zones in the region.

To work with BIG DATA in social networks we used the methods of collecting, processing and analyzing the data. Data collecting is carried out in a real time, within the certain geo location, or within the entire network according to the predefined patterns. Information of interest for analysis in the area of SM is: location, date and time, content, “author” of content (user), links with users. Data collecting may be fulfilled with the help of following tools: Apache Hadoop, Biglnsights (IBM), Cloudera, Hortonworks, Storm. To carry out the research in the field of SB we chose Hortonworks. We used Twitter Application (apps.twitter.com), where the key parameters were defined using API key, API secret, Access token, Access token secret.

For data collecting with Hortonworks, Twitter App we used flume service configuration file in Hortonworks Virtual Machine Sandbox. System is ready to load data from twitter after Hortonworks Virtual Machine Sandbox version 2.3 is installed and flume service is configured. Navigate to HDFS folder in order to view downloaded files for data processing. HDFS view in Hortonworks virtual machine while solving tasks in the area of SB is shown in Fig. 2.

Collected data must be structured (i.e. processed) according to MapReduce paradigm. MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

MapReduce gave ability to structure the data flow from social networks using following criterion: font, text size, color, user profile hyperlink, location, date and others.

In order to define user profile for SB in our research we need data of following types: location, text, language and date. We used MapReduce to retrieve only required data in Hortonworks Sandbox tool. For data processing in Hadoop environment we chose Hive DB that gives ability to operate with the data and apply analysis via SQL-like queries. For this we created sql-script hivedll.sql for necessary tables creation. File contents is shown below:

```
// twitter table identifiers
CREATE EXTERNAL TABLE tweets_raw (
id BIGINT,
created_at STRING,
source STRING,
favorited BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text:STRING,
usr:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
usr STRUCT< screen_name:STRING,  name:STRING, friends_count:INT, followers_count:INT, statuses_count:INT,
verified:BOOLEAN, utc_offset:STRING, -- was INT but nulls are strings time_zone:STRING>,
in_reply_to_screen_name STRING,
yearint,
monthint,
dayint,
hourint
)
CREATE EXTERNAL TABLE time_zone_map (
time_zone string,
country string,
notes string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/data/time_zone_map';
…
create table tweets_sentiment stored as orc as select
id,
case
when sum( polarity ) > 0 then 'positive'
when sum( polarity ) < 0 then 'negative'
else 'neutral' end as sentiment
from l3 group by id;
 -- put everything back together and re-number sentiment
CREATE TABLE tweetsbi
STORED AS ORC
AS
SELECT
t.*,
cases.sentiment
when 'positive' then 2
when 'neutral' then 1
when 'negative' then 0
end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.id = s.id.
```

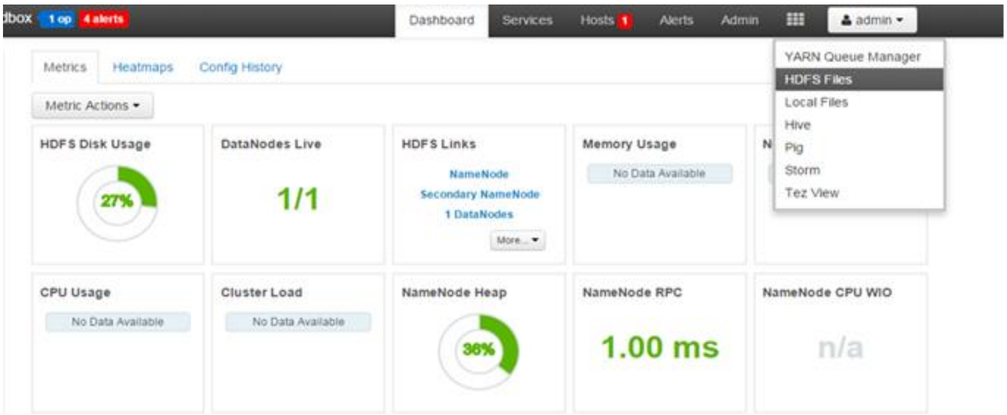

Fig. 2. HDFS view in Hortonworks when downloading files while solving tasks in the area of SB.

Execute script using command: Hive_f hiveddl.sql. Structured data are placed in Table 1.

Table 1. Column headers for structured data analysis in tasks for SB.

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

The following metrics are used for data analysis. The total number of twits (Kol_i) for every location (R) is defined:

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

Where k_i is the every following twit from processing thread.

The unique word frequency ch(m) is defined from total collection L of text data:

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

Relationship of every twit otn (m,rez) can be defined from thesaurus tez, where relationship to every word is set:

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

For further work we created a dictionary with filters of SB domain in order to identify the number of twits by location ch(m) and number of twits by location with respect of relationship otn (m,rez) hereafter. We define thesaurus taking into account filter with base metrics of small and medium-sized businesses: food, clothes, entertainment and kids. In conclusion we got 4 base metrics of medium-sized business.

Metric «food» P_1 is calculated as number of twits in overall text data L:

$$Kol_{otnP_1} = \frac{\sum_{i=1}^{N} S_i(S_i \in P_1)}{L} = 9\%.$$

Metric «clothes» P_2 is calculated as number of twits in overall text data L:

$$Kol_{otnP_2} = \frac{\sum_{i=1}^{N} S_i(S_i \in P_2)}{L} = 8\%.$$

Metric «entertainment» P_3 is calculated as number of twits in overall text data L:

$$Kol_{otnP_3} = \frac{\sum_{i=1}^{N} S_i(S_i \in P_3)}{L} = 6\%.$$

Metric «kids» P_4 is calculated as number of twits in overall text data L:

$$Kol_{otnP_4} = \frac{\sum_{i=1}^{N} S_i(S_i \in P_4)}{L} = 12\%.$$

4. Results and discussion

As a result it is possible to conclude what area of SB is especially in high demand in Samara Oblast. According to the Figure 3 it is apparent that the main strategy of SB promotion for authorities must be connected with opening of centers for children.

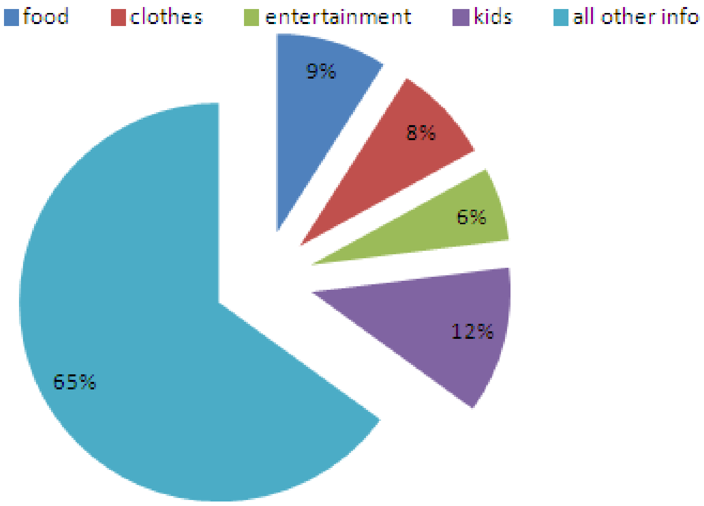

Fig. 3. Metrics of small business in Samara Oblast.

Due to BIG DATA technology it is possible to distribute and update data in «hadoop» file system using filter “Samara Oblast” (filter1= {Samara Oblast}). Then it is necessary to filter this area on base metrics of small and medium-sized businesses, setting up for example the following metrics: Filter2 (food) = {cafe, bar, restaurant, cuisine*, beer*, meat, fish, tavern}; Filter3 (clothes)= {coat, jacket, dres*, skir*, jacke*, bra*, stuf*}; Filter4 (entertainment)= {night club, concert, session, hangout}; Filter5 (kids) = {kindergarten, baby-club, club}.

A set of descriptors for filtering the Internet discourse will be determined by the lexical representatives of the concept formed in the world building of the average Russian-speaking consumer of services. The main in the sphere of concepts "Food" is the micro-situation "Cooking", which includes the following cognitive and propositional structure: Subject - Predicate of cooking (how it is cooked) - Object of cooking - The property of the cooking object - Method of cooking – Premises - Kitchenware – Appliances - Devices - Affair- Substance - Food / Dish – Food quality / Dish quality. In the situation of Internet communication, only the structure elements relevant to the user are being explicated, the lexical interpretation of which let us draw a conclusion about the needs of the residents of a particular district of Samara city. Building –up of block of descriptors on Filter3 (clothes); Filter4 (entertainment); Filter5 (kids) may be fulfilled according to the lexical and semantic fields “clothes”, “fashion”, associative and semantic field “leisure”; concept “childhood”.

For making decision in the area of SB it is necessary to create multimodal clusterization of social networks. The clusterization is based on the method of Formal Concept Analysis (FCA). A large number of structured and unstructured data generate trivial data. For example, the data of social websites in the SB area may be submitted in the form of following three items (user, group, interest) (Fig. 4).

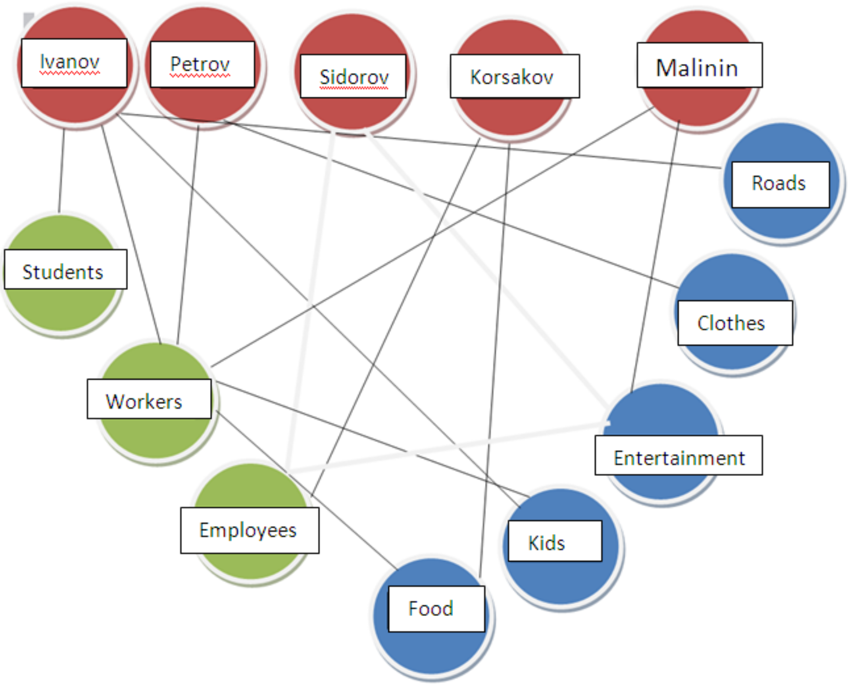

Fig. 4. SB data from social network «twitter» as a graph.

By the method of formal notions it is necessary to introduce the following definitions: G - set of objects, M – feature set, the dependence of $I \subseteq G \times M$ such that $(g,m) \in I$ when and only when the object g posses the feature m; K $:= (G, M, I)$ is called formal context.

Galois operator is defined in the following manner: for $A \subseteq G, B \subseteq M$ $A' \overset{def}{=} \{m \in M \mid g/m \forall g \in A\}$, $B' \overset{def}{=} \{g \in G \mid g/m \forall m \in B\}$, where A is the formal volume, B is the formal content.

Formal notion is the pair $(A,B): A \subseteq G, B \subseteq M, A' = B$ and $B' = A$.,

Notions ordered by ratio $(A_1, B_1) \geq (A_2, B_2) \Leftrightarrow A_1 \supseteq A_2 (B_2 \supseteq B_1)$, from the complete lattice, called a context lattice $\underline{\beta}(G, M, I)$. The example of social network context in the SB area and their context lattice are shown in Table 2 and in Figure5.

Table 2. The example of SB data context from social network (a is the attributes of “food” filter, b is the attributes of “kids” filter, c is the attributes of “entertainment” filter, d is the attributes of “clothes” filter).

	G/M	a	b	c	d
1	Pensioners	x			x
2	Employees	x		x	
3	Workers		x	x	
4	Students		x	x	x

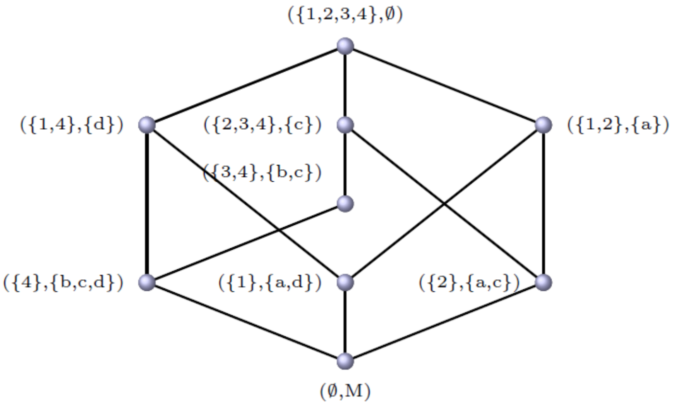

Fig. 5. Context lattice for social network.

The use of this clusterization method permits to define the groups of interest, with increase of connections where it is required to make managerial decisions. But this tool has restrictions of use. Users who work with social network Twitter are in the group of “students” and partly in groups of “employees” and “workers” and slightly in group of “pensioners”, that is why it is necessary to add field marketing research in these groups in order to take management decision.

There is ability to get correlation between number of user requests with respect to filters and date and time of data collecting [9]. Time of data collecting from Internet using Big Data is not limited.

As a result we get dynamic change of information in real time from Internet, which allows conduct monitoring of continuous analysis of unstructured information by filters with minimal investments (In-Memory Data Processing and Stream technology). For the purpose of this method implementation we coded a program using Scala language:

```
val file = spark.textFile(“hdfs://… “)
val errors=file.filter(line=>line.contains(“Samara Oblast“))
//count all the data
errors.count()
//count data mentioning Filter
errors.filter(line=>line. contains(“meat“)).count()
//Fetch the filter as an array of string
errors.filter(line=>line. contains(“food“)).collect()
```

After program execution we got dynamic change of parameters in BIG DATA environment, that allow to identify zone of SM business in geo region taking into account unstructured information. In case of consistent “peaks” detected in hit counts in accordance with forms of business there should be supporting investment program take place for development of small and medium sized businesses with a focus on certain business activity in target area.

In conclusion we suggested a tool for increasing of budget funds usage effectiveness in geo region. This is the most important challenge in modern economic reality, the solution for which is based on opportunity to take management decision in most optimal way. Suggested approach of regulation can be efficient in innovative process management in developing of region economy typical of lots of forms and wide range of factors, as well as dynamic progression and active transformation of daily living.

Using of modern software and hardware allows conducting evaluation and visualization of changes in almost real time that can be useful not only to local region governments but also to businesses in a way of design and implementation of investment projects.

References

[1] Drovyannikov VI, Khaymovich IN. Development of control pattern to manage the competitive improvement of social cluster of the region. Fundamental Studies 2015; 7(4): 822–827.

[2] Drovyannikov VI, Khaymovich IN. Simulation modelling of social cluster administration in Any Logic system. Fundamental Studies 2015; 8(2): 361–366.

[3] Ramzaev VM, Kukolnikova EA, Khaymovich IN. Development of a model for the functioning of production active elements in regional management. Bulletin of SSEU 2014; 12: 87–99.

[4] Ramzaev VM, Khaymovich IN. Integrated model of control over economic development of the region based on competitiveness improvement of the enterprises. Modern Issues of Science and Education 2014: 6: 136 p.

[5] Ramzaev VM, Khaymovich IN, Chumak VG. Forecasting model of competitive growth for enterprises with energy modernization. Forecasting problems 2015; 1: 67–75.

[6] Bonacich P. Power and Centrality: A Family of Measures. American Journal of Sociology 2007; 92(5): 1170–1182.

[7] Chumak PV, Ramzaev VM, Khaimovich IN. Models for forecasting the competitive growth of enterprises due to energy modernization. Studies on Russian Economic Development 2015; 26(1): 49–54.

[8] Chumak VG, Ramzaev VM, Khaimovich IN. Challenges of Data Access in Economic Research based on Big Data Technology. CEUR Workshop Proceedings 2015; 1490; 327–337.

[9] Chumak VG, Ramzaev VM, Khaimovich IN. Use of Big Data technology in public and municipal management. CEUR Workshop Proceedings 2016; 1638: 864–872.

[10] Grechnikov FV, Khaimovich AI. Development of the requirements template for the information support system in the context of developing new materials involving Big Data. CEUR Workshop Proceedings 2015; 1490: 364–375.