THE METHOD OF A TWO-LEVEL TEXT-MEANING SIMILARITY APPROXIMATION OF THE CUSTOMERS' OPINIONS

Nina Rizun [1], Paweł Kapłanski [2] and Yurii Taranenko [3]

[1] Gdansk University of Technology, Department of Applied Informatics in Management, Faculty of Management and Economics, nina.rizun@zie.pg.gda.pl
[2] Gdansk University of Technology, Department of Applied Informatics in Management, Faculty of Management and Economics, pawel.kaplanski@zie.pg.gda.pl
[3] Alfred Nobel University, Dnipropetrovs'k, Department of Applied Linguistics and Methods of Teaching Foreign Languages, taranen@rambler.ru

Abstract

The method of two-level text-meaning similarity approximation, consisting in the implementation of the classification of the stages of text opinions of customers and identifying their rank quality level was developed. Proposed and proved the significance of major hypotheses, put as the basis of the developed methodology, notably about the significance of suggestions about the existence of analogies between mathematical bases of the theory of Latent Semantic Analysis, based on the analysis of semantic relationship between the variables and degree of participation of the document or term in the corresponding concept of the document data, and instruments of the theory of Social Network Analysis, directed at revealing the features of objects on the basis of information about structure and strength of their interaction. The Contextual Cluster Structure, as well as Quantitative Ranking evaluation for interpreting the quality level of estimated customers' opinion has formed.

Key words: text-meaning, Latent Semantic Analysis, Social Network Analysis.

Introduction

Large number of opinions is easily accessible nowadays. It is desirable to understand their properties as they potentially contain valuable business information. Opinion mining (or sentiment analysis) tries to extract this valuable information using complex and Semantic Analysis Algorithms.

There exist a few ways of texts semantic analysis, which can be divided into the following groups [1]: linguistic analysis and statistical analysis.

The first group is oriented at defining the sense by the semantic structure of the text and includes lexical, morphological and syntactic analysis. At the present time, there are no established approaches to realization of the task of sematic analysis of text information. It is mostly caused by the exceptional complexity of the problem and by the insufficiency of the studies in the scientific direction of artificial intelligence systems creation.

The second group, as a rule, includes frequency analysis with its variations. The analysis consists of counting the number of words repetitions in a text and applying the results for particular objectives. Different options of various realizations of words calculation and the further results processing create a wide specter of methods and algorithms, suggested in this class.

From the other hand, clustering problem solving, especially in the context of its application for the analysis of text messages, is fundamentally ambiguous, and there are several reasons [2-5].

Firstly, there does not exist a clustering quality criterion, which is definitely the best. There is a number of quite reasonable criteria, as well as a number of algorithms that do not have clearly expressed criterion, but perform reasonable clustering "by construction". All of them can give different results.

Secondly, the number of clusters is usually not known in advance and are set in accordance with some subjective criteria.

Thirdly, the clustering result depends strongly on the metric ρ, the choice of which is typically subjective and also determined by the expert

1 Theoretical Justification

One of the most effective statistical approaches is the Latent Semantic Analysis (or the Latent Semantic Indexing). Latent Semantic Analysis (LSA) – is a method for discovering hidden concepts in document data. Each document and term (word) is then expressed as a vector with elements corresponding to these concepts. Each element in a vector gives the degree of participation of the document or term in the corresponding concept. The goal is not to describe the concepts verbally, but to be able to represent the documents and terms in a unified way for exposing document-document, document-term, and term-term similarities or semantic relationship, which is otherwise hidden [6, 7].

The method of Latent Semantic Analysis enables the extraction of context-dependent meanings of words with the help of statistical processing of large sets of text data. This method is based on the principles of analyzing the main components, which are applied in the creation of artificial neural networks. The totality of all contexts, in which the word is found, sets a number of reciprocal restrictions that allow defining the similarity of words' meanings and words' sets between each other. It allows to model separate cognitive and psycho-linguistic processes of a man.

There exist the following practical tasks with the application of LSA [8-11]:

– document comparison (their clustering and classification);

– search of similar documents in different languages – after the analysis of translated documents base;

– search of documents on the selected terms.

The Latent Semantic Analysis deals well with the problem of synonymy, but partially with the problem of polysemy, because every word is defined by one point in the space. This analysis also enables conduction of documents' automatic categorization, based on their similarity of conceptual content. Independency on language is also one of the advantages of LSA (since it is a mathematical approach). The disadvantage of the method is the decrease of the speed of calculation with the increase of output data volume (for instance, in the SVD-decomposition).

The analysis of networks and networked systems, however, has a long tradition in economics, and an even longer history of graph theory in discrete mathematics [12-14]. From the late 1990s onwards, research on social networks has branched onto a number of fields, and has been generally carried out under the umbrella term of complex networks, a new emerging area in which networks are studied in several domains, using data from a wide variety of sources. The classes of networks studied include computer, biological, financial, medical, physical, and transportation networks, among many others.

In a graph, we call a unit – whether an individual, a company, an indicator – a vertex or a node. A tie between two nodes indicates presence of the relationship connecting them. Absence of a tie indicates absence of the relationship. A tie with a direction is called an arc, and a tie without a direction is called an edge. One could also note the value or volume of flow as the weight of a tie and thus obtain a network that would then be a weighted digraph.

A primary use of graph theory in social network analysis is to identify "important" actors. Centrality and prestige concepts seek to quantify graph theoretic ideas about an individual actor's prominence within a network by summarizing structural relations among the g nodes. Group-level indexes of centralization and prestige assess the dispersion or inequality among all actors' prominences. The three most widely used centrality measures are degree, closeness, and betweenness [15-17].

The goal of research using social network analysis (SNA) mainly is to understand the general properties of the data, which is represented with using the networks, often by analyzing large datasets collected with the aid of technology. The data is often abstracted at the level at which the networks are treated as large graphs, often with little or no concern on whether the nodes (actors) represent people, companies, indicators, or other entities. Such an abstraction is possible because in many ways the problems addressed in complex network research are similar across different domains. Relevant problems include understanding of the structure of the networks (i.e.,

by identifying underlying properties of the link and edge structures), the evolution of such structures (i.e., how the networks change over time), and how information propagates within the networks.

The **objective** of this work consists in developing the method of a two-level text-meaning similarity approximation, which contains the realization of the stages of Contextual Clustering of customers' text opinions and Quantitative Rank identification and aims to fill the scientific *gaps* in the area of the absence the universal method of the contextual and structural clustering of the textual messages taking into account the semantical and statistical specificity at the same time. The methodology is based on justifying the hypothesis of existing analogy between mathematical foundations of the Latent Semantic Analysis and the theory of Social Network Analysis.

The first chapter introduce the theoretical justification of the Contextual Clustering problem. The second chapter contain the research methodology fragments, which presupposed the data source and usage of variables describing, development of the algorithm of the text-meaning discovering on the basis of mathematical tools of the LSA and SNA theories instruments and development of the algorithm of contextual clustering of customers' text opinions and quantitative rank identification. The third chapter proposed the discussion about the experimental results and finding. The conclusions and further research direction placed in the fifth part of the paper.

2 The Research Methodology Fragments

The two-level method of the text-meaning similarity approximation, suggested by the authors, lies in the basis of the following hypotheses:

Hypothesis 1. The process of identification of customers' text-opinions about the used product (service) can be realized as a sequence of the following stages: division of the examined documents into Clusters by means of defining Contextual Interdependencies between the set of documents and the Quantitative Rank identification.

Hypothesis 2. It is mathematically and methodologically justified (significant) to draw an analogy between mathematical bases of the theory of Latent Semantic Analysis, based on the analysis of *semantic* relationship between the variables and degree of participation of the document or term in the corresponding concept of the document data, and instruments of the theory of Social Network Analysis, directed at revealing the features of objects on the basis of information about *structure* and strength of their interaction.

The suggested method of identification and clustering of customers' opinions with the possibility of rank evaluation of qualitative level of products' evaluation, based on the above-mentioned hypotheses, includes the following stages (fig. 1).

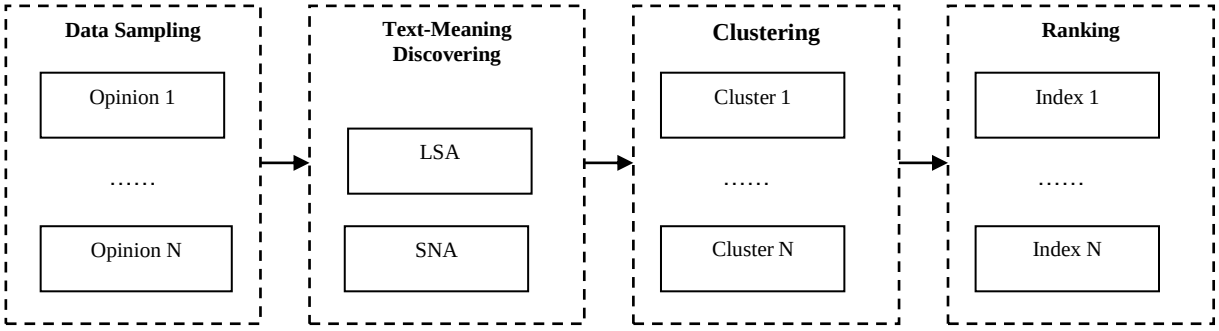

Fig. 1. The Stages of the Method of Identification and Clustering of Customers' Opinions

2.1 Data Source and Usage of Variables

As the experimental data for the research and approbation of the developed method the authors have used opinions of customers of the Starbucks Coffee Houses, received from the official page of Starbucks Coffee Company. Because of the fact that the developed methodology is aimed at the evaluation of text information, quantitative evaluations of Starbucks service quality were removed from the processed texts.

2.2 Development of the Algorithm of the Text-Meaning Discovering on the Basis of Mathematical Tools of the LSA and SNA Theories Instruments

The objectives of this stage are:

– development of the algorithm of the text-meaning discovering for receiving the quantitative measure of customers' text-opinions similarity;

– confirming the significance of Hypothesis 2 about the existence of analogy between mathematical and methodological bases of LSA and SNA in the process of developing the algorithm of the two-level text-meaning similarity approximation.

For this purpose realization we will briefly examine the mathematical instruments, used as the basis of the suggested algorithm of customers' text opinions classification, and then we will analyze the results of their application for a particular example.

The phases of the developed Algorithm of Customers' Opinions Text-Meaning Discovering are presented in the figure 2.

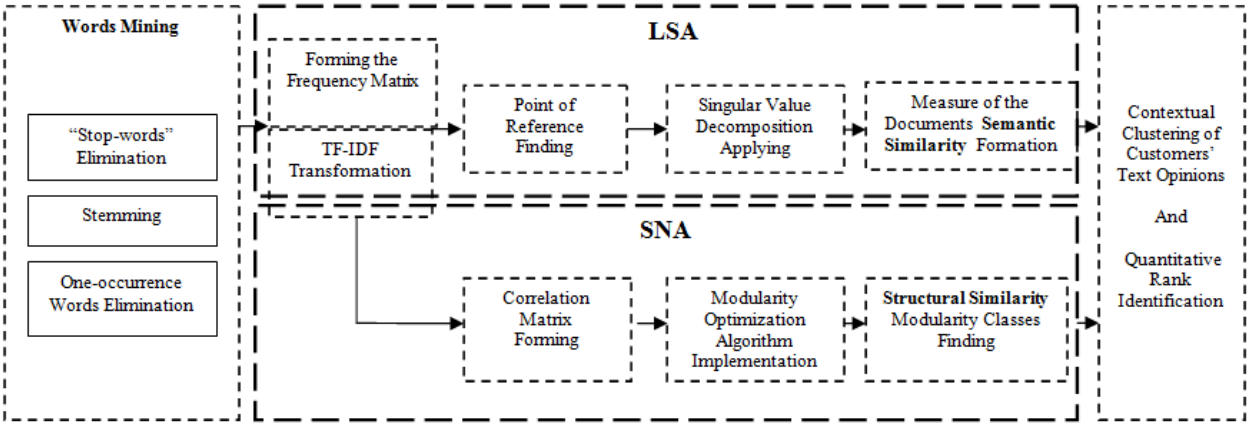

Fig. 2. The Phases of the Algorithm of Customers' Opinions Text-Meaning Discovering

The phase *Words Mining* is the preparative for realizing the Algorithm of Classification of Customers' Text Opinions and presents sequential procedures of removing words, which do not have any sense load (prepositions, pronouns etc.); removing words that are found only once in the whole document; stemming (finding the word's stem – for instance, the Porter's rule).

On the basis on the Words Mining phase results two levels of the Text-meaning Discovering must be implemented: the level of the *Latent Semantic Analysis* and the level of the *Social Network Analysis* text-meaning discovering implementation.

2.2.1 Finding the Semantic of Customers' Text-Opinions Similarity (Level of the Latent Semantic Analysis)

As the scientific background for the level of customers' text-opinions documents *similarity* evaluation we propose to use the phenomena of *semantic similarity between* the documents, which could be revealed via using the *Latent Semantic Analysis* instruments. In this paragraph we proposed authors version of using this instruments for finding the semantic of customers' text-opinions similarity.

1. The phase of *Forming the Frequency Matrix FR* contains the process of building the matrix *relative frequency* of the w-*th* word occurrence in document *t:*

$$tf(w,t) = \frac{k(w,t)}{df}, \tag{1}$$

where $k(w,L_t)$ – the number of w-*th* word occurrences in the text *t*; *df* – the total number of words in the text of *t*.

2. The phase of *Frequency Matrix TF-IDF Transformation* is the process of the obtaining the statistical measure F_{w_i} used for evaluation of the significance of the word in the context of the document, which is part of the list of analyzed documents:

$$F_{w_i} = TF \times IDF = tf \cdot log_2 \cdot D/df \quad (2)$$

tf – words frequency (in the document);

D – total number of documents in the collection.

3. The phase of finding the *Point of Reference* in the scale of the quantitative measure of customers' text-opinions similarity determination can serve to coordinate the term with the highest total weight:

$$F_{w_i}^{PR} = max\{\sum_{i=1}^{n} F_{w_i}\}. \quad (3)$$

The high weight of this term among the customer's text-opinions collection indicates the presence of at least one cluster, the center of which is this centroid term (term group).

4. The phase of the *Singular Value Decomposition* (SVD-transformation) is the process which allows reflecting basic structure of the different dependencies that are present in the original matrix [18]. The mathematical basis of the method is as follows:

Formally let *A* be the $m \times n$ words-document matrix of a collection of documents. Each column of *A* corresponds to a document. The values of the matrix elements $A[i,j]$ represent the frequency identifications F_{w_i} of the term occurrence w_i in the document t_j: $A[i,j] = F_{w_i}$. The dimensions of *A*, *m* and *n*, correspond to the number of words and documents, respectively, in the collection.

Observe that $B = A^T A$ is the document-document matrix. If documents *i* and *j* have *b* words in common, then $B[i,j] = b$. On the other hand, $C = AA^T$ is the word-word matrix. If terms *i* and *j* occur together in c documents then $C[i,j] = c$. Clearly, both *B* and *C* are square and symmetric; *B* is an $m \times m$ matrix, whereas *C* is an $n \times n$ matrix. Now, we perform an Singular Value Decomposition on *A* using matrices *B* and *C* as described in the previous section:

$$A = S\Sigma U^T, \quad (4)$$

where *S* is the matrix of the eigenvectors of *B*, *U* is the matrix of the eigenvectors of *C*, and Σ is the diagonal matrix of the singular values obtained as square roots of the eigenvalues of *B*.

In LSA we ignore these small singular values and replace them by 0. Let us say that we only keep *k* singular values in Σ. Then Σ will be all zeros except the first k entries along its diagonal. As such, we can reduce matrix Σ into Σ_k which is an $k \times k$ matrix containing only the *k* singular values that we keep, and also reduce S and U^T, into S_k and U_k^T, to have *k* columns and rows, respectively. Of course, all these matrix parts that we throw out would have been zeroed anyway by the zeros in Σ. Matrix A is now approximated by:

$$A_k = S_k \Sigma_k U_k^T \quad (5)$$

Observe that, since S_k, Σ_k and U_k^T are $m \times k$, $k \times k$, and $k \times n$ matrices, their product, A_k is again an $m \times n$ matrix. Intuitively, the k remaining ingredients of the eigenvectors in S and U correspond to k "hidden concepts" where the terms and documents participate. The words and documents have now a new representation in words of these hidden concepts. Namely, the words are represented by the indexed vectors-models of the $m \times k$ matrix $S_k \Sigma_k$, whereas the documents by the column vectors the $k \times n$ matrix $\Sigma_k U_k^T$ (coordinates of the words and the documents in the common k-dimension space – the so-called hypothesis space).

5. The stage of the *Measure of the Documents Similarity* formation.

According to the authors, as an instrument of the documents' similarity degree identification, it is advisable to use the *reference dimensional coordinate* $dist_{t_i}$ – the distance between the indexed vectors-models of the documents and coordinates of a *point of reference* in the scale of the customer's text-opinions collection closeness determination.

Taking into account the heuristics introduced in consideration with determining the relative dimensional coordinates $dist_{t_i}$, it is encouraged to use the following author's concepts of standard distance metrics:

– Euclid's measure $dist_{t_i} = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$, where x – vector of the document, y – point of reference words vector;

– Cosine of the edge between the vectors: $dist_{t_i} = \cos\theta = \frac{x \cdot y}{\|x\| \cdot \|y\|}$, where $x \cdot y$ – scalar product of the vectors, $\|x\|$ и $\|y\|$ – quota of the vectors, which are calculated by the formulas:

$$\|x\| = \sqrt{\sum_{i=1i}^{n} x_i^2}, \ \|y\| = \sqrt{\sum_{i=1i}^{n} y_i^2} \quad (6)$$

Then, the *measure of similarity* $K_{i+1,i}$ between pairs of documents is justified to consider the difference between the values of their relative spatial coordinates $dist_{t_i}$:

$$K_{i+1,i} = dist_{i+1} - dist_i. \quad (7)$$

While documents shall be sorted in ascending order of values $dist_{t_i}$.

As a result of this phase of the algorithm of the text-meaning discovering we can receive the vector of the quantitative measure of customers' text-opinions *semantic similarity* $K_{i+1,i}$.

2.2.2 Finding the Structural of Customers' Text-Opinions Similarity (Level of the Social Network Analysis)

As the scientific background for the level of customers' text-opinions documents *similarity* evaluation we propose to use the phenomena of *structural similarity between frequency characteristics of the common words*, found in the documents. In this paragraph we proposed authors version of using the of Social Network Analysis instruments for finding the structural of customers' text-opinions similarity.

1. The phase of the *Correlation Matrix* $W = \{r_{ij}\}$ forming – as a measure of the dependencies between the analyzed of customers' text-opinions r_{ij} on the basis of *TF-IDF* transformed frequency matrix. This matrix is the basis for applying the modularity optimization algorithm of the SNA theory.

2. The phase *Modularity Optimization* algorithm implementation.

For realizing this algorithm we need to presented the dependencies between the analyzed of customers' text-opinions as a graph, in which the *nodes* are the text-opinions documents, and *edges* – correlation dependencies between those documents with weight r_{ij}.

Modularity is a metric that was proposed by Newman and Girvan in reference [19, 20]. It quantifies the quality of a community assignment by measuring how much more dense the connections are within communities compared to what they would be in a particular type of random network. One of the mathematical definitions of modularity was proposed in the original paper on the Louvain method [18]:

$$Q = \frac{1}{2m}\sum_{i,j}\left[r_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{8}$$

Here, $k_i = \sum_j r_{ij}$ is the total the weighted degree of node *i* (weighted-degree of a node is the sum of the weights of all links attached to node i); $m = \frac{1}{2}\sum_{i,j} r_{ij}$ is the total link weight in the network overall.

The Kronecker delta $\delta(c_i, c_j)$ is 1 when nodes *i* and *j* are assigned to the same community and 0 otherwise. Consider one of the terms in the sum. Remember that k_i is the total the weighted degree of node *i*, $\frac{k_j}{2m}$ – the average fraction of this weight that would be assigned to node j, if node i assigned its link weight randomly to other nodes in proportion to their own link weights). Then, $A_{ij} - \frac{k_i k_j}{2m}$ measures how strongly nodes *i* and *j* are in the real network, compared

to how strongly connected we would expect them to be in a random network of the type described above.

In our case, we can transform the classical definition of the Modularity metric as the following:

Molularity is number of edges, which are the correlation relationships r_{ij} *between the documents on the bases of the co-occurrence of the words, falling within text-opinion documents with common structure (clusters) minus the expected number in an equivalent documents network with edges placed at random*.

The problem of modularity optimization can be thought of as assigning communities such that the elements of the sum that contribute are as positive as possible. Larger modularity *Q* indicates better communities. *Q*~=0.3-0.7 indicates good partitions.

As a result of this phase of the algorithm of the text-meaning discovering we can receive the vector of the modularity classes *Q* – quantitative measure of customers' text-opinions *structural* similarity modularity classes *Q*.

2.3 Development of the Algorithm of Contextual Clustering of Customers' Text Opinions and Quantitative Rank Identification

The objective of this stage is to confirm significance of *Hypothesis 1* about the possibility to identify the opinions of customers by realization of the procedure of dividing the analyzed documents into clusters and subsequent rank identification of the qualitative evaluation level with the help of applying the instruments of the LSA and SNA analysis.

For realization of this objective the following phase's implementation are proposed:

1. Partition of the customers' text-opinions on the contextual clusters via applying the k-means algorithms for clustering of the vector of the quantitative measure of customers' text-opinions *contextual* similarity $K_{i+1,i}$.
2. Qualitative *interpretation* of the $K_{i+1,i}$ ranges, which was adopted as a *measure* of the *relevance* to the main revealed context of the customers' text-opinions set.
3. Comparing the results of the *contextual* clusters (LSA implementation) and *structural* modularity classes (SNA implementation) for making the final decision.

3 Experimental Results and Finding

As mentioned above, we will consider testing the developed algorithm on the example of analysis of the opinions of 32 customers of the Starbucks Coffee Houses, obtained from the official web-page. A fragment of the analyzed list of opinions is presented in the table 1:

Table 1. The Fragment of the Analyzing Documents List

Documents ID	Opinion
Doc_0	Coffee quality is not good. Coffees are good with caramel, chocolate, cream... but if you order an espresso is like to drink dirty water.
Doc_1	I wanna rate Starbucks, and the people that work there. I had only good experience with them/ Very polite and friendly stuff. I can say only good about them and their coffee;)
Doc_2	The only one advantage is the wi-fi connection...
Doc_3	I want to rate just one specific store, it's in Reading PA, I dont remember street name, but the building number is 2113 if I'm not wrong. I came there on Easter Sunday, just to get coffee, so what I want to mention that if to compare with other stores across the country (I'm traveling a lot), this one was very clean, well-organized, sugar and dairy station had all verity of things. Personnel was very polite and friendly. Good team work. Well done guys!!!!. Wish you all the best!!!
Doc_4	Always had good experiences with them… really quick and friendly staff.. plus I am a coffee addict so I love that there is a Starbucks everywhere I go..
…	**…**

Fragment of the experimental results of the *Frequency Matrix FR* obtaining is the following (totally 270 words are selected):

Table 2. The Fragment of the Frequency Matrix FR

	Doc_0	**Doc_1**	**Doc_2**	**Doc_3**	**Doc_4**	**Doc_5**	**Doc_6**	**Doc_7**	…
about	0	1	0	0	0	0	0	0	…
accent	0	0	0	0	0	0	0	0	…
acknowledg	0	0	0	0	0	0	0	0	…
addict	0	0	0	0	1	0	0	0	…
after	0	0	0	0	0	0	1	0	…
again	0	0	0	0	0	0	0	0	…
all	0	0	0	2	0	1	0	0	…
almost	0	0	0	0	0	0	0	0	…
alreadi	0	0	0	0	0	0	0	0	…
alway	0	0	0	0	1	0	0	0	…
amex	0	0	0	0	0	0	0	0	…
amount	0	0	0	0	0	0	0	0	…
…	…								

Fragment of the experimental results *Frequency Matrix TF-IDF Transformation* and total weight obtaining are the following (table 3):

Table 3. The fragment of the Frequency Matrix TF-IDF Transformation

	Doc_0	**Doc_1**	**Doc_2**	**Doc_3**	**Doc_4**	**Doc_5**	**Doc_6**	**Doc_7**	**…**	$F_{w_i}^{PR}$
love	0	0	0	0	0,14	0	0	0	…	1,16
servic	0	0	0	0	0	0,05	0	0	…	1,14
excel	0	0	0	0	0	0	0	0	…	0,94
alway	0	0	0	0	0,19	0	0	0	…	0,87
one	0	0	0,46	0,08	0	0	0	0,11	…	0,87
onli	0	0,2	0,46	0	0	0	0	0	…	0,83
the	0	0	0,36	0	0	0,07	0	0,04	…	0,81
good	0,18	0,18	0	0,04	0,13	0,04	0,07	0	…	0,77
your	0	0	0	0	0	0	0	0	…	0,76
coffe	0,09	0,05	0	0,02	0,06	0	0	0,03	…	0,72
nice	0	0	0	0	0	0	0	0	…	0,72
product	0	0	0	0	0	0	0	0	…	0,72
…	…									

The word "*love*" is used as the *Point of Reference*, which at the same time is determine the main context of the documents set. Suppose we set k = 2, i.e. we will consider only the first two singular values. Then we have:

$$\Sigma_2 = \begin{bmatrix} 1{,}15 & 0 \\ 0 & 0{,}98 \end{bmatrix}$$

$$S_k = \begin{array}{l} \text{about} \\ \text{accent} \\ \text{acknowledg} \\ \text{addict} \\ \text{after} \\ \text{again} \\ \text{all} \\ \text{almost} \\ \text{alreadi} \\ \text{alway} \\ \text{am} \\ \text{amex} \\ \text{amount} \\ \dots \end{array} \begin{pmatrix} -0{,}0086 & 0{,}0057 \\ -0{,}0011 & -0{,}0004 \\ -0{,}0023 & 0{,}0019 \\ -0{,}0305 & 0{,}0171 \\ -0{,}0041 & 0{,}0025 \\ -0{,}0015 & 0{,}001 \\ -0{,}0056 & 0{,}0028 \\ -0{,}0034 & 0{,}001 \\ -0{,}0011 & 0{,}0006 \\ -0{,}2459 & 0{,}3619 \\ -0{,}006 & -0{,}0062 \\ -0{,}0039 & 0{,}0021 \\ -0{,}0333 & 0{,}0815 \\ \dots & \dots \end{pmatrix}$$

$$U_k^T = \begin{array}{c} \begin{array}{cccccccc} \text{Doc\_0} & \text{Doc\_1} & \text{Doc\_2} & \text{Doc\_3} & \text{Doc\_4} & \text{Doc\_5} & \text{Doc\_6} & \text{Doc\_7} \end{array} \\ \begin{pmatrix} 0{,}02 & 0{,}06 & 0{,}03 & 0{,}02 & 0{,}06 & 0{,}04 & 0{,}02 & 0{,}01 & \dots \\ -0{,}02 & -0{,}09 & 0{,}02 & 0 & -0{,}09 & 0{,}01 & 0{,}01 & 0{,}01 & \dots \end{pmatrix} \end{array}$$

The words in the concept space are represented by the row vectors of S_2 whereas the documents by the column vectors of U_2^T. In fact we scale the (two) coordinates of these vectors by multiplying with the corresponding singular values of Σ_2 and thus represent the terms by the row vectors of $S_2\Sigma_2$ and the documents by the column vectors of $\Sigma_2 U_2^T$.

Calculation of the *measure of semantic similarity* was conducted with the use of the Cosine measures. The examples of the experimental results of the measure of similarity $K_{i+1,i}$ calculation are presented in the table 4 and fragmentary in the figure 3:

Table 4. The Example of the Experimental Results of the Measure of Similarity $K_{i+1,i}$ Calculation

Documents	$K_{i+1,i}$	**Common (key) words**
1,4	0,038	experi, coffee, good
14, 4	0,049	coffee, love
12, 14	0,094	starbuck, love, coffe
26, 12	0,139	coffe, friend
4, 26	0,285	alway, friend, coffe, love, starbuck
26, 16	0,371	chocol, me, love time, starbuck
12, 3	0,482	starbuck, from
26, 25	0,560	coffe, amout, veri, shop

14, 25	0,63	coffe, shop, well, starbuck
31, 25	0,682	starbuck, like
17, 3	0,740	starbuck, been, all, like
23, 26	0,750	starbuck, customer,servic
26, 28	0,760	servic, coffe
14, 22	0,790	coffe, fan, shop
8, 12	0,810	starbuck, food
12, 25	0,820	friend, coffe, like, starbuck
25, 27	0,850	starbuck, love,servic, coffe
21, 4	0,850	coffe, shop
14, 19	0,900	starbuck, coffe, shop, like
11, 22	1,100	Starbuck, not
10, 24	1,193	starbuck, staff, coffe, , nothing
24, 11	1,207	starbuck, not, or, me, make, custom, day, becaus, unfortun, went, off, one, want, buy, from, coffe, all, back, veri, time
24, 9	1,220	starbuck, not, good
11, 10	1,239	starbuck, store, time, been, all, like, custom, contact, or, went, me, becaus, anyth, who, back, be, refus, out, said, anoth, two, three, day
24, 23	1,310	servic, starbuck
28, 10	1,320	starbuck, becaus
0, 20	1,500	coffe, good, but
21, 23	1,540	order, coffe, like, caramel, not, good
6, 7	1,610	say, like, not, good, put, sticker
20, 18	1,750	starbuck, wait, servic
30, 6	1,857	not, after
18, 29	1,908	starbuck, locat, seem, their, than, not, or, visit, week, those, someth, get, from
5, 15	1,948	want, one, not, good, countri, travel, veri, all, thing
13, 7	1,986	card, say, becaus, all
7, 0	1,993	coffe, like, good
2, 8	2,000	one, starbuck, day

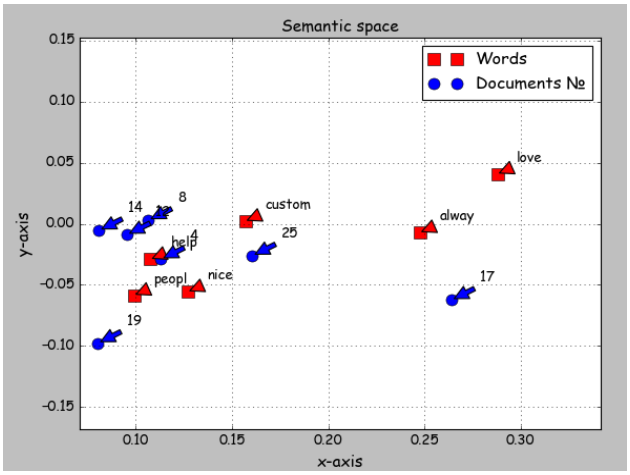

Fig. 3. The Fragment of the LSA Tools Implementation Experimental Results

Source: results of the experiments

For receiving the quantitative measure of customers' text-opinions *structural* similarity modularity classes Q the stage of modularity optimization of the **SNA** theory algorithm was conducted. The matrix of correlation relationships $W = \{r_{ij}\}$ between the documents on the bases of the co-occurrence of the words was obtained (Table 5).

Table 5. The Fragment of the Experimental Results of Correlation Matrix W

	Doc_0	**Doc_1**	**Doc_2**	**Doc_3**	**Doc_4**	**Doc_5**	**Doc_6**	**Doc_7**	**Doc_8**	**Doc_9**	**Doc_10**
Doc_0	1,000										
Doc_1	0,305	1,000									
Doc_2	-0,018	0,324	1,000								
Doc_3	0,123	0,236	0,194	1,000							
Doc_4	0,270	0,425	-0,017	0,089	1,000						
Doc_5	0,063	0,063	-0,027	0,067	-0,005	1,000					
Doc_6	0,251	0,104	-0,018	0,029	0,090	0,078	1,000				
Doc_7	0,214	0,080	0,241	0,237	0,068	-0,016	0,210	1,000			
Doc_8	0,013	0,073	-0,020	0,122	0,035	-0,072	-0,049	0,079	1,000		
Doc_9	0,153	0,127	0,043	-0,039	0,027	-0,046	0,072	0,118	0,018	1,000	
Doc_10	0,091	0,040	-0,027	0,070	0,008	0,017	0,000	0,139	0,063	0,204	1,000

On the basis of this matrix the Modularity optimixzation algorithm was inmlamented. As a result – the 4 Modularity classes Q as quantitative measures of customers' text-opinions *structural* similarity modularity classes with the Modularity coefficient *0,61* was received (Table 6, figure 4).

Table 6. The Example of the Experimental Results of the Modularity Optimization Algorithm Realization

Documents	**Weighted Degree** k_i	**Modularity Class**
Doc_1	3,208856	0
Doc_12	4,236518	0
Doc_14	4,296015	0
Doc_15	4,181601	0
Doc_16	3,73195	0
Doc_17	1,238412	0
Doc_19	4,662027	0
Doc_25	1,068756	0
Doc_26	3,109359	0
Doc_27	1,028323	0
Doc_3	2,166251	0
Doc_31	4,404249	0
Doc_4	4,289584	0
Doc_21	2,757831	1
Doc_22	1,4592	1
Doc_23	0,751695	1
Doc_28	3,001274	1
Doc_29	0,845228	1
Doc_8	1,044076	1
Doc_10	1,84973	2
Doc_11	2,756099	2
Doc_2	1,971137	2
Doc_24	3,676981	2
Doc_9	2,201011	2
Doc_0	3,193837	3
Doc_13	1,05275	3
Doc_18	3,599364	3
Doc_20	1,857333	3
Doc_30	2,057538	3

Doc_5	1,550596	3
Doc_6	1,358176	3
Doc_7	3,964580	3

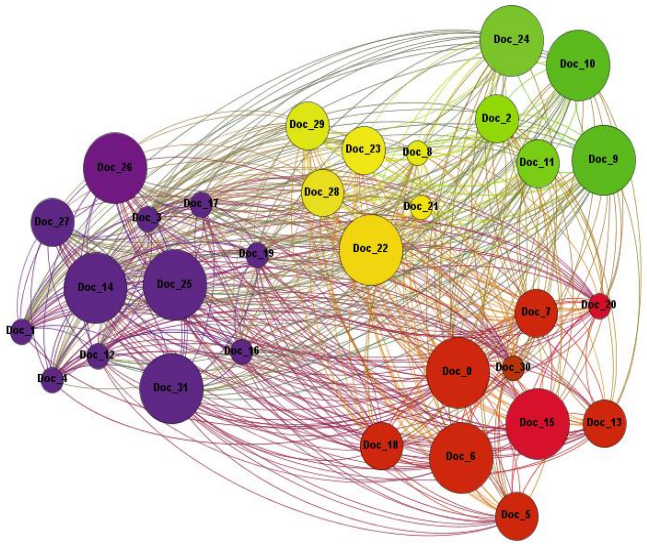

Fig. 4. The Results of the Modularity Optimization Algorithm Realization
Source: results of the experiments

On the bases of the data, presented in the table IV, the classical procedure of the k-means clustering algorithms was realized. As a result, were received:

– absolute and relative ranges of the *Measure of Similarity* with their qualitative *interpretation* (table 7):

Table 7. Ranges of the Similarity Measure

Cluster	Measure of Similarity Range	Relative Rank of the Evaluation	Qualitative Interpretation
1	[0,00; 0,5]	[0,76; 1]	Very good
2	[0,51; 1,0]	[0,51; 0,75]	Good
3	[1,10; 1,5]	[0,26; 0,5]	Satisfactorily
4	[1,51; 2,0]	[0;0,25]	Bad

– the lists of the customers' text-opinions, belonging to *contextual clusters* and *structural modularity classes* (table 8):

Table 8. The Example of the Experimental Results of the Contextual Cluster and Structural Modularity Classes

Clusters	Contextual Clusters	Clusters	Structural Modularity Classes
1	1, 4, 12, 14, 16, 26	-	-
2	3, 15, 17, 19, 23, 25, 27, 31, ***8, 21, 22, 28***	2	1, 3, 4, 12, 14, 15, 16, 17, 19, 25, 26, 27, 31
3	8, 10, 11, 24, ***22, 23, 28***	3	2, 9, 10, 11, 24
4	0, 2, 5, 6, 7, 9, 13,18, 20, 29, 30, ***8, 21, 23***	4	0, 5, 6, 7, 13,18, 20, 30
-	-	5	8, 21, 22, 23, 28, 29

Comparing the results of the data of the Table VIII leads to the following remarks:

1) Belonging the documents of the Structural Modularity Classes two smaller Contextual clusters 1 and 2 allows:

– on the one hand, to establish the fact the presence of more subtle differences *contextual* and *structural* differences between the opininons, describing the qualitative interpretation of "Good";
– on the other hand, the presence of distinction between the *contextual* and the *structural* specifics of the analyzed text-opinions, particularly pronounced when comparing the documents of the different size.

2) The presence of the documents number 8, 21, 22, 28, 23, 21 simultaneously in several clusters may indicate too broad semantic context, or what is the same, the lack of dipole content compliance customers's opinion the key theme of the analyzed documents – "Evaluation the quality coffee at Starbucks network". A brief visual analysis of the context of the selected document allows to confirm this assumption – their content is really blurred. For the most cases authors express their, not related to the topic of the survey, emotions, gives examples of incidents from their own life and only casually refer to the main topic, including a reference to the previous authors' opinion. In that regard, authors proposed to move these documents in a separate cluster with the title of "*Does not correspond to the topic*".

3) Some differences in the composition of the main documents clusters of can also be explained as the differences in the mathematical basis of the used methods, and the blurring of of the content of some customer opininons.

In this connection, the authors propose the following concept of the *adaptive algorithm* of comparing the results of the *contextual* clusters and *structural* modularity classes for making the final decision:

– identification the common clusters between two results;

– isolation of the documents simultaneously belonging to several clusters in the separate cluster;

– adjustment of the obtained clusters and their qualitative interpretation based on the comparison results.

As an example of this *adaptive algorithm* realisation could be the following identification and clustering of customers' opinions results (Table 9):

Table 9. The Example of the Clustering of Customers' Opinions Results

Clusters	Qualitative Interpretation	Structural Modularity Classes
1	Very good	1, 4, 12, 14, 16, 26
2	Good	3, 15, 17, 19, 23, 25, 27, 31
3	Satisfactorily	2, 9, 10, 11, 24
4	Bad	0, 5, 6, 7, 13,18, 20, 30
5	Does not correspond to the topic	8, 21, 22, 23, 28, 29

As software tools for the technical realization of the *two-level text-meaning similarity approximation* method in this paper proposed the usage of the developed by authors latent-

semantic analysis' software (based on the Python language) and standard SNA applications (e.g.Gephi) [21, 22].

4 Conclusions and Further Research Direction

Research contribution

Thus, the method of two-level text-meaning similarity approximation, consisting in the implementation of the classification of the stages of *Contextual Clustering* of customers' text opinions and *Quantitative Rank* identification, have developed. The authors research contribution is to use this method for the express evaluation of customer opinion groups with the following main purpose:

– identification of the qualitative opinion's structure;

– optimization the opinion's search process – extracting the keywords for each clusters;

– cleaning the opinion's documents set via finding the meaingless (with connectinon on the evaluation context) opinion's.

The study limitation was not enough the sample size and studing only the specifice group of the textual opinions of the Starbucks Coffee Houses customers. That is why the future research presuppose to continue the studying the different types of the textual messages - both in terms of subject and size of the analyzed text

5 Practical Implementation

Large number of opinions is easily accessible nowadays. It is desirable to understand their properties as they potentially contain valuable information. Ask Data Anything (ADA) is a system developed by Cognitum that is using a combination of formal logic and statistical analysis to extract dimensions from the data and to expose the dimensions through a natural query language based interface. Currently we are implementing the method of a two-level text-meaning similarity approximation to embed it in the ADA system for the Customers' Opinions mining purposes as a part of joint collaboration with Cognitum company [23-25].

REFERENCES

[1] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. *Indexing by Latent Semantic Analysis*. 1990, № 41(6), pp. 391-407.

[2] Jonathan I. Maletic, Naveen Valluri. *Automatic Software Clustering via Latent Semantic Analysis*. 14th IEEE ASE'99, Cocoa Beach FL, Oct. 12-15th, pp. 251-254

[3] Jon Rune Paulsen, Ramampiaro H. *Combining Latent Semantic Indexing and Clustering to Retrieve and Cluster Biomedical Information*: A 2-step Approach. NIK-2009 conference.

[4] Jing L., Ng M. K., Yang X., Huang J. Z. *A text clustering system based on k-means type subspace clustering and ontology*. International Journal of Intelligent Technology, 1(2): 91–103, 2006.

[5] Roussinov D., Leon Zhao J. *Text Clustering and Summary Techniques for CRM Message Management*. [Online]. Available: https://personal.cis.strath.ac.uk/dmitri.roussinov/Lim-Paper.pdf

[6] Řehůřek, R. *Subspace tracking for latent semantic analysis*. *Advances in Information Retrieval*, 2011, pp. 289–300.

[7] Pedersen, T. Duluth: *Word Sense Induction Applied to Web Page Clustering* : Proceedings of the 7th inter. workshop Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM-2013), 2013, pp. 202–206.

[8] Jurgens D. *The S-Space Package: An Open Source Package for Word Space Models*. Proceedings ACLDemos '10. Proceedings of the ACL System Demonstrations, 2010, pp. 30–35.

[9] Řehůřek R., Sojka. *Software Framework for Topic Modelling with Large Corpora*. Proceedings of the LREC 2010 workshop. New Challenges for NLP Frameworks, 2010, pp. 45–50.

[10] Hofmann T. *Probabilistic Latent Semantic Indexing*. Proceedings of the twenty-second annual inter. SIGIR conf. Research and Development in Information Retrieval, 1999, pp. 50–57.

[11] Roger B. Bradford. *An empirical study of required dimensionality for large-scale latent semantic indexing applications*. Proceedings of the 17th ACM conf. / B. Roger Bradford // Information and Knowledge Management, 2008. – pp. 153–162.

[12] Ahuja R. K., Magnanti Thomas L., Orlin, J. B. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ. 1993.

[13] Bollobas B. *Modern Graph Theory*. Springer, 1998.

[14] West, D. *Introduction to Graph Theory*. Prentice Hall. 1996

[15] Freeman L. C. *Centrality in social networks: Conceptual clarification. Social Networks*, 1979, 1 (3), pp. 223–258.

[16] Freeman L. C. *Visualizing social networks*. Journal of Social Structure, 2000, 1 (1).

[17] Freeman L. C. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver: Empirical Press. 2004

[18] Thomo A. *Latent Semantic Analysis* [Online]. Available: http://www.engr.uvic.ca/~seng474/svd.pdf

[19] M.E.J. Newman, M. Girvan. *Finding and evaluating community structure in networks*. Phys. Rev. E. 69: 026113, 2004.

[20] M.E.J. Newman, C. Moore. *Finding community structure in very large networks*. Phys. Rev. E 70, 066111, 2004.

[21] Kapłanski P., Rizun N., Taranenko Y., Seganti A. *Text-mining Similarity Approximation Operators for Opinion Mining in BI tools*. Chapter: Proceeding of the 11th Scientific Congerence "Internet in the Information Society-2016", Publisher: University of Dąbrowa Górnicza, pp.121-141.

[22] Rizun N., Kapłanski P., Taranenko Y. *Development and Research of the Text Messages Semantic Clustering Methodology*. 2016, Third European Network Intelligence Conference, Publisher: ENIC, 2016.33, pp.180-187

[23] Kapłanski P., Weichbroth P., *Cognitum Ontorion: Knowledge Representation and Reasoning System*, in Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015., 2015. doi: 10.15439/2015F17 pp. 177–184. [Online]. Available: http://dx.doi.org/10.15439/2015F17

[24] Kapłanski P. *Controlled English interface for knowledge bases*, Studia Informatica, vol. 32, no. 2A, pp. 485–494, 2011

[25] Wroblewska A., Kaplanski P., Zarzycki P., Lugowska I., *Semantic Rules Representation in Controlled Natural Language in FluentEditor*, in Human System Interaction (HSI), 2013 The 6th International Conference on. IEEE, 2013, pp. 90–96.