Development and Research of the Text Messages Semantic Clustering Methodology

Nina Rizun
Gdansk University of Technology
Department of Applied Informatics in Management
Faculty of Management and Economics
Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland
Email: nina.rizun@zie.pg.gda.pl

Paweł Kapłanski
Gdansk University of Technology
Department of Applied Informatics in Management
Faculty of Management and Economics
Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland
Email: pawel.kaplanski@zie.pg.gda.pl

Yurii Taranenko
Alfred Nobel University, Dnipropetrovs'k
Department of Applied Linguistics and Methods of Teaching Foreign Languages
Naberezhna Lenina Str., 18, Dnipropetrovs'k
Ukraine
Email: taranen@rambler.ru

Abstract — **The methodology of semantic clustering analysis of customer's text-opinions collection is developed. The author's version of the mathematical models of formalization and practical realization of short textual messages semantic clustering procedure is proposed, based on the customer's text-opinions collection Latent Semantic Analysis knowledge extracting method. An algorithm for semantic clustering of the text-opinions is developed, the distinctive characteristics of which is the introduction of concepts and methods of identification point of reference in the scale of text-opinions collection closeness determination; instrument of the documents' closeness degree identification; measure of similarity between pairs of documents. The version of quantitative evaluation of the clustering results is developed. The concepts of resolving power of the method of semantic clustering and level of the clustering procedure quality are proposed. Analysis of the specific features and the effectiveness level of various distance measures is conducted.**

Keywords — semantic clustering; text-opinions; Latent Semantic Analysis; closeness; measure of similarity

I. Introduction

One of the Business Intelligence objectives is the clustering of text-messages, documents, links, which dedicates to identifying the semantically related texts in the multidimensional space of the information features as well as the definition of cluster centers, which are topical headings (abstracts) [1-5]. The most clustering techniques are practically reduced to the classification problem solution [5-7] and are based on simple vector markup model (Vector Space Model) – classical model in the field of information retrieval.

Within this model a document is described by a vector in which each term used in the document is associated with its importance (weight) in the framework of this document. Importance of the term is based on statistical information about the occurrence of terms in this document and/or another documents. Clusters description is also represented by a vector, and for evaluation of the documents closeness and describing the cluster, a scalar product of vectors of cluster (class) description and the vector of the document are used.

However, this problem can be solved only if we have the vector of the class description, preliminary composed by experts.

The clustering problem is the following: to divide the sample into disjoint subsets called clusters, so that each cluster is composed of objects that are close to the metric *ρ*, and objects of different clusters are differed significantly. At the same time the clusters quantity and their characteristic features are unknown in advance.

One of the perspective directions of automation of the text messages clustering process is Latent Semantic Analysis (LSA), allowing to identify the structure of semantic relationships between words used by the statistical analysis of a large group of documents. Due to this method we can automatically distinguish different shades of meaning of the same word in the context of its use.

II. Objectives

Clustering problem solving, especially in the context of its application for the analysis of text messages, is fundamentally ambiguous, and there are several reasons.

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.

For the improvement of quality of the text messages clustering procedure the following **goal** has been set by authors – development and research of the **methodology of semantic clustering analysis** on the example of the Product Customer Opinions.

To accomplish this goal it is necessary to perform the following research tasks:

1. Formalization of a *mathematical model of short text messages semantic clustering* based on the *LSA* method of extracting knowledge from a text-opinions collection.

2. Development of *clustering algorithm of text-opinions collection* via the improvement of the measure of documents similarity calculating algorithms.

3. Formalization of the algorithms of *quantitative identification* of clustering procedure results.

4. Identification of the *features and the effectiveness of various distance metrics* underlying the improved algorithms for calculating a measure of the documents similarity.

III. Mathematical model of the customer’s text-opinions collection

The expanded and modified version of the semantic clustering of customer’s text-opinions collection model formalization in the form of an algebraic system is proposed.

The general model of semantic clustering (MSCs) is suggested to be submitted in the form of the following system:

$$R = \langle T, M, f(dist_t)\rangle \to \langle C, F\rangle \qquad (1)$$

where:

T – the set of customer’s text-opinions τ;

M – the algorithm of knowledge extraction from customer’s text-opinions collection;

$f(dist_t)$ – clustering algorithm: the projection *T* →*C*, which to any object *t∈T* assigns the label of cluster *c∈C*. Moreover, it is assumed that the set of labels in *C* is defined by experts.

$dist_t$ – the function of the distance used to determine the closeness (belonging) of the document to the cluster;

C – non-empty set of the contextual clusters (the results of the customer’s text-opinions collection τ division);

F – non-empty set of the contextual clusters descriptions (lists of key words and their significance).

For further consideration of texts as objects of clustering and extraction of significant attributes of their classification, the following concepts are introduced and formalized:

1. **Lexical unit (term)** w_i – keyword: forms of the word (stemming results – the procedure of the words bases selection) except for the words that have no semantic load (prepositions, pronouns, etc.).

2. **The frequency identifications** D_{w_i} of the term occurrence w_i in the document t_j*:*

– *relative frequency* of the w-*th* term occurrence in document *t:*

$$TF(w,t) = \frac{k(w,L_t)}{S(L_t)}, \qquad (2)$$

where $k(w,L_t)$ – the number of *w-th* term occurrences in the text *t*; $S(L_t)$ – the total number of terms in the text of *t*.

– *inverse frequency* of the *w-th* term occurrences in a set of text-opinions collection *τ*:

$$IDF(w,\tau) = log_2 \frac{n}{m(w,\tau)}, \qquad (3)$$

where *n* – the total number of the documents in a set τ; $m(w,\tau)$ – the number of the documents in a set *τ*, which contains the term *w*;

– *standardized value* $IDF_N(w,\tau)$:

$$IDF_N(w,\tau) = \frac{log_2 \dfrac{n}{m(w,\tau)}}{log_2 n} = log_n \frac{n}{m(w,\tau)}. \qquad (4)$$

This value, according to [18], determines the amount of information associated with the onset of the event "with a random choice of *τ* opinions found in this document, at least one occurrence of *w-th* term". For computing the value of the correctness of calculation suggest that if $m(w,\tau)=0$, then $IDF_N(w,\tau)=1$;

– *weight (importance)* of the *w-th* term of the text *t* in the text-opinions set *τ* is determined by the normalized at range [0,1] non-negative value:

$$D_{w_i} = TFIDF_N(w,t) = TF(w,t) \bullet IDF_N(w,\tau) \qquad (5)$$

The weight of *w-th* term, calculated this way, is proportional to the frequency of its usage in the document and is inversely proportional to the frequency of the use of the term in other documents in the collection.

3**. Plane model** $Mod_t = \langle \phi_t, D_t \rangle$ **of text-opinion** *t* is offered as a combination of the following elements:

– the one-dimensional vector of the terms $\phi_t = \{w_i\}_{i=1}^{S_t}$, which text-opinions contain, where S_t – the total number of terms in the document *t*;

– the one-dimensional vector of the terms frequency identifications of the document $D_t = (D_{w_i}, D_{w_2}, .., D_{w_{St}})$;

4. The **method of knowledge extraction** from customer’s text-opinions collection *M*.

As a method of knowledge extraction from customer’s text-opinions collection in order to create the multidimensional document’s model proposes to use the tools of LSA [8-16]. This method is based on the idea that the totality of all the contexts in which the terms are occurred and not encountered, defines a set of mutual constraints, which in a large extent allow determining the similarity of meanings of words and sets of words between each other.

The most common version of LSA is based on the *Singular Value Decomposition* (SVD), which allows reflecting basic structure of the different dependencies that are present in the original matrix. The mathematical basis of the method is as follows:

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

Observe that $B = A^TA$ 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 term-term 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 *SVD* on *A* using matrices *B* and *C* as described in the previous section: $A=S\Sigma U^T$, 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:

$$A_k = S_k\Sigma_k U_k^T\,. \qquad (6)$$

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 terms and documents have now a new representation in terms of these hidden concepts. Namely, the terms are represented by the row vectors of the $m\times k$ matrix: $S_k\Sigma_k$, whereas the documents – by the column vectors of the $k\times n$ matrix $\Sigma_k U_k^T$.

Thus, the basic idea of *LSA* is that a matrix A_k containing only the *k* first linearly independent components A and represents the basic structure of the associative dependency of the original matrix, and at the same time does not contain noise.

5. **K-dimensional model** of the text-opinions collection represented by the array of indexed vectors-models of:

– terms $SD_w^k = \left(SD_{w_1}^k, SD_{w_2}^k,..., SD_{w_m}^k\right)$;

– documents $UD_t^k = \left(UD_{t_1}^k, UD_{t_2}^k,..., UD_{t_n}^k\right)$,

in the common *k*-dimension space (the so-called hypothesis space).

6. The function of **distance** underlying the clustering algorithm $f(dist_t)$ and used to determine the closeness (belonging) of the document to the cluster, based on the following assumptions: to measure the degree of closeness between the documents it is proposed to use the standard distance metrics, namely:

a. Euclid distance [18].

b. Hellinger measure [19].

c. Metric of closeness within multinomial model [20].

d. Cosine measure [21].

IV. Semantic clustering algorithm of the customer's text-opinions collection

In the **semantic clustering algorithm of the customers text-opinions collection** $f(dist_t)$ it is proposed to use the author's concept of the standard distance interpretation in conjunction with the LSA method results.

Step 1: The reference point on the scale measuring the closeness degree of the customer's text-opinions collection determination.

Step 2: The closeness degree of the documents identification.

Step 3: The measures of similarity between pairs of documents recognition.

This algorithm is based on the following **heuristics**:

Heuristic 1. As a *point of reference* in the scale of the customer's text-opinions collection closeness determination can be used a coordinate of the term $Centroid^k = \left(SD^1, SD^2,..., SD^k\right)$ (group of terms) with the highest total weight:

$$D_{w_i}^{Centroid} = max\{\sum_{i=1}^{n} D_{w_i}\}. \qquad (7)$$

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).

Heuristic 2. As an instrument of the documents' closeness degree identification should be used 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.

Heuristic 3. 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 $dist_{t_i}$:

$$K_{i+1,i} = dist_{i+1} - dist_i\,. \qquad (8)$$

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

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 distance:

$$dist_{t_i} = Euclidean(t_i, Centroid) = \sqrt{\sum_{i=1}^{k}\left(UD_{t_i}^i - SD^i\right)^2} \qquad (9)$$

Hellinger measure:

$$dist_{t_i} = Hellinger(t_i, Centroid) = 1 - \sqrt{\sum_{i=1}^{k} \left(UD^i_{t_i} \bullet SD^i\right)} \qquad (10)$$

Metric of closeness within multinomial model of text representation:

$$dist_{t_i} = Multinomial(t_i, Centroid) = \prod_{i=1}^{k} (UD^i_{t_i})^{SD^i} \qquad (11)$$

Cosine similarity measure:

$$dist_{t_i} = cos(t_i, Centroid) = \frac{\sum_{i=1}^{k} UD^i_{t_i} \bullet SD^i}{\left\| \sqrt{\sum_{i=1}^{k} (UD^i_{t_i})^2} \right\| \bullet \left\| \sqrt{\sum_{i=1}^{k} (SD^i)^2} \right\|} \qquad (12)$$

Then the cosine of the angle between the documents vectors, which are sorted in ascending order of values $dist_{t_i}$, will be determined according to the relationship:

$$\begin{aligned} & cos((t_{i+1}, Centroid) - (t_i, Centroid)) = \\ & cos(t_{i+1}, Centroid) \cdot cos(t_i, Centroid) \\ & + \sqrt{1 - cos^2(t_{i+1}, Centroid)} \cdot \sqrt{1 - cos^2(t_i, Centroid)} \end{aligned} \qquad (13)$$

By using the non-negative words' weights the cosine measure of similarity between the pairs of documents takes values in the range [0, 1], so for evaluation of the vectors difference the following formula are used:

$$K_{i+1,i} = 1 - cos((t_{i+1}, Centroid) - (t_i, Centroid)). \qquad (14)$$

For the purpose of carrying out a quantitative evaluation of the clustering procedure results level, the authors proposed to use the following identifiers:

1. The **resolving power** of the method of semantic clustering *Resolution* – the ability to select from a customer's text-opinions collection the maximal possible (in the case of the experts given the number of clusters – expected) number of clusters C^n.

2. The **level** of the clustering procedure **quality** introduced by the authors as a combination of two factors in the following concepts:

– the *degree of the boundaries fuzziness* between the clusters $Border_C$, defined as the variation coefficient of distances between the marginal value of the similarity measures $Border\_Dist_{C\,j+1,j} = max(K^C_{j+1}) - max(K^C_j)$ of the documents in the clusters:

$$Border_C = \frac{\sigma(Border\_Dist_{c_j})}{\overline{Border\_Dist_{Cj}}}; \qquad (15)$$

– the *closeness of the similarity measures* values within a cluster Dev_C of the documents similarity within the cluster:

$$Dev_{C_{ji}} = \frac{\sigma_{c_j}}{\overline{K^C_j}} \cdot \qquad (16)$$

Then, the problem of **clustering of the customer's text-opinions collection** can be formalized as follows:

dividing of the set of ***text-opinions*** *models* τ *into disjoint subsets of clusters with maximal resolution and quality level so, that each cluster consisted of documents that are close in relative measure of similarity* $K_{i+1,i}$ *with respect to the reference point on the scale measuring the closeness degree of the customer's text-opinions collection determination, and documents of the different clusters are differed significantly.*

The results of implementation of the methodology of **semantic clustering of the customer's text-opinions collection** are the following:

– the set *C* of the contextual cluster' models of the text-opinions:

$$CL\_Mod_t = \left\langle \phi_t, dist_{t_i}, c \right\rangle, \qquad (17)$$

additionally including: the information about the distance to current document from the reference point $Centroid^k$ and the cluster' label c;

– non-empty set of contextual descriptive clusters:

$$F_c = \left\langle \bigcap_c w_i, \bigcap_c D_{w_i} \right\rangle \qquad (18)$$

namely, lists of key words and their significance.

V. Experiments and Results

As experimental data for research and testing of the methodology developed by the authors, the text-opinions collection of the Starbucks coffee shop network customers were used, which were obtained from the official page (https://www.trustpilot.com/review/www.starbucks.com).

Moreover, due to the fact that the developed methodology is used to estimate the textual information, quantitative evaluation results, which are present in the Starbucks official page, have been removed from the processed text.

For conducting experiments and research the program for linguistic analysis, developed by the authors in Python, was used. The program interface allows the possibility of implementing a flexible research process for analysis clustering algorithm results using the following experimental parameters:

– the variety of measures for determining the relative spatial coordinates (only metric of Euclidean and Cosine measure were chosen for the experiment - experiment modes 1, 2);

– usage of the additional procedure of the exclusion of the words occurring in the text-opinions collection only once OOWO (experiment modes 3, 4, formed as an addition to the modes 1, 2);

– usage of the relative/inverse frequency of the *w-th* term occurrence in document *t* (experiment modes 5, 6, formed as an addition to the modes 3, 2);

As a results of each experiment modes in Table 1 the following values are presented:

– the number of selected clusters;

– the number of text-opinions documents in clusters (in%);

– the marginal value of the similarity measures in the cluster.

It was assumed that the maximum possible (experts given) number of clusters generated as a result of this method, is equal to 5 and includes the following contextual interpretation:

C={*C_1="Very good", C_2="Good", C_3="Satisfactorily", C_4="Bad", C_5="Does not correspond to the topic"*}.

TABLE I. THE RESULTS OF THE CLUSTERING OF THE CUSTOMER'S TEXT-OPINIONS COLLECTION PROCEDURE WITH THE DIFFERENT EXPERIMENTAL MODES

Nr. of Mode	**Experimental Mode**	**Number of selected clusters**	**Number of text-opinions documents in clusters (%)**					**The marginal value of the similarity measures in the cluster**				
			C_1	***C_2***	***C_3***	***C_4***	***C_5***	***C_1***	***C_2***	***C_3***	***C_4***	***C_5***
1	Euclidean	3	-	53,13%	-	31,25%	15,63%	-	0	-	0,02	0,409
3	Euclidean+ OOWO	4	27,27%	27,27%	-	36,36%	9,09%	0,001	0,002	-	0,005	0,223
5	Euclidean+ OOWO + TF_IDF	5	18,75%	25,00%	21,88%	28,13%	6,25%	0	0,002	0,012	0,158	0,528
2	Cosine	4	50,00%	25,00%	-	12,50%	12,50%	0	0,001	-	0,006	0,385
4	Cosine+ OOWO	5	43,75%	21,88%	21,88%	6,25%	6,25%	0	0,001	-	0,093	0,342
6	Cosine+ OOWO + TF_IDF	5	50,00%	12,50%	15,63%	15,63%	6,25%	0	0,005	0,109	0,258	0,632

Results of the experiments indicate the following (Table 2, Figures 4-6): selection of the experiments' parameters affect the resolution of the semantic clustering of the customer's text-opinions collection method as follows:

1. Absence in the clustering algorithm of the procedures of exclusion of words occurring in the text-opinions collection only once and using the inverse frequency of *w-th* term in *t-th* document (experiment modes 1, 2) characterized by (Figure 1):

– high level of terms spread in *k*-dimensional hypothesis space, which is reflected in increasing of *C_5* cluster size, brings together a group of documents with a maximum number of words that occur in the text-collection once and having a significant remoteness from a reference point on proposed scale of measurement and a low degree of similarity between documents within the cluster;

– low semantic density of documents in the *k*-dimensional hypotheses space that causes a decrease of the resolving power of the method – reduction of the resulting number of clusters;

– low quality of the clustering procedure, expressed in: boundaries fuzziness between the clusters (low values of the index $Border_C$) and the absence of significant values of similarity measures within a cluster (high index of variation coefficient of measures of documents similarity within the cluster Dev_C).

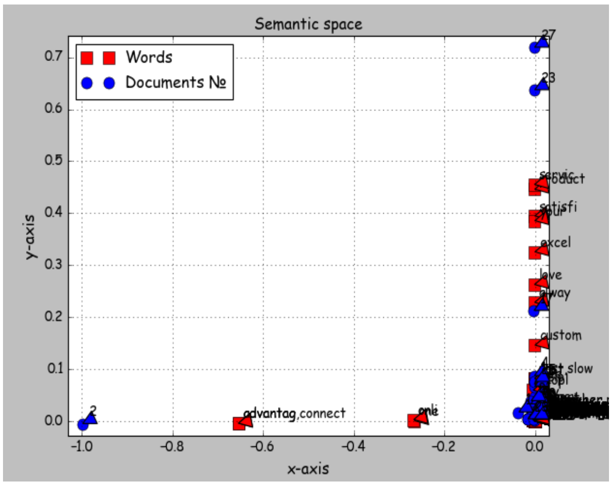

Figure 1. Results of LSA experiment with the mode Euclidean

2. Adding to the clustering procedure the modes of exclusion the words occurring in the text-opinions collection only once (experiments modes 3 and 4) allows (Figure 2):

– to reduce the spread of terms and increase the semantic density of the documents in the *k*-dimensional hypotheses space that causes a decrease of the *C_5* cluster size, brings together a group of documents with a maximum number of words, and improves the quality of clustering procedure (the border between the clusters become more clear, the difference between the similarity measure within a cluster decreases);

– to increase the resolving power of the clustering method, which results in increasing the number of resulting clusters.

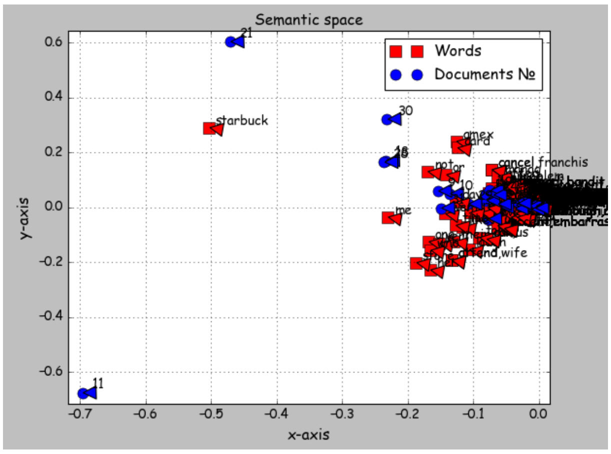

Figure 2. Results of LSA experiment with the mode Euclidean + OOWO

3. Adding to the clustering procedure the mode of using the inverse frequency of *w-th* term of the *t-th* document (experimental modes 5, 6) maximally contributes to the quality and resolving power of the clustering method (Figure 3).

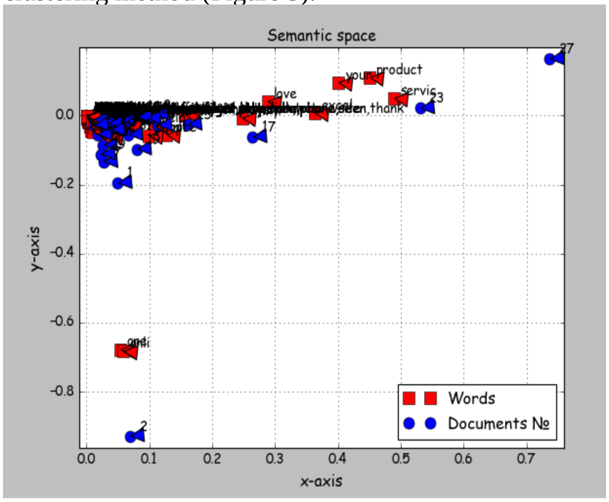

Figure 3. Results of LSA experiment with the mode Euclidean + OOWO + TF_IDF

4. Changing the modes connected with the reference dimensional coordinate (proposed by the authors of the Euclidean or Cosine measure metrics concept) allows to make the following conclusions ((Figure 4):

– Euclidean measure is less suitable for comparison of the documents that are significantly different in size, which is characterized by low resistance of main characteristics of the semantic clustering method to experiment modes changing (low resolving power of the method in Modes 1, 3).

– Cosine measure has a high quality of structural analysis and comparison of the documents, which is characterized by high (relatively to Euclidean measure) semantic clustering procedure quality even in modes 1 and 2.

TABLE II. RESULTS OF THE CLUSTERING PROCEDURE EVALUATION

Nr. Of Mode	**Experimental Mode**	**The level of the clustering procedure quality**		**Resolving power**
		Border$_C$	*Dev*$_C$	
1	Euclidean	0,0250	25%	0,600
3	Euclidean+ OOWE	0,29694	18%	0,800
5	Euclidean+ OOWE + TF_IDF	0,1289	10%	1,000
2	Cosine	0,0800	23%	0,800
4	Cosine+ OOWE	0,1388	15%	0,800
6	Cosine+OOWE + TF_IDF	0,2522	7%	1,000

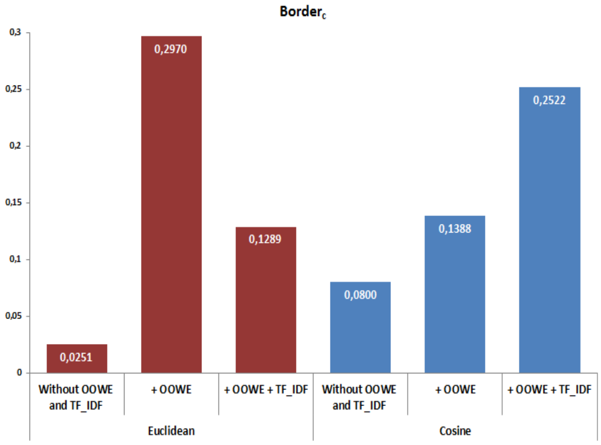

Figure 4. Comparison results of the degree of boundaries fuzziness between the clusters on different experimental modes

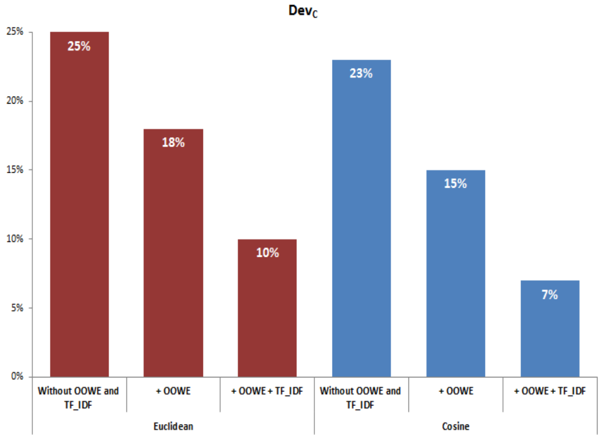

Figure 5. Comparison results of the closeness of similarity measures on different experimental modes

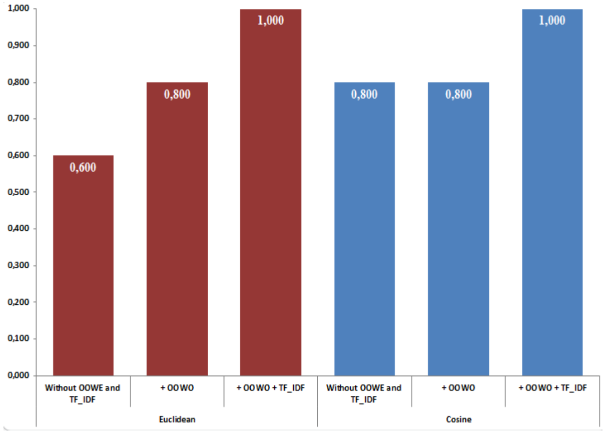

Figure 6. Comparison results of the Resolving power on different experimental modes

VI. Possible experimental errors types and research modes setting procedure

It should also be noted that in spite of rather high experimental results the quality of obtained cluster composition is not totally homogeneous and uniform.

The authors propose the concept of distinguishing two groups of possible clustering errors: a *false choice* and *misses*.

The *false choice* – it is a situation where: the documents contain too much information, not related directly to the products characteristics (emotions, cases of life, etc.); or the author of the document uses specific not official terminology; or document critically small in size (no more than 3 words). Then the model of the document has initially distorted meaning and its getting into any of the clusters has almost equal probability and depends to a large extent on the features of the mathematical method, and to a lesser – of the content of the document itself.

Misses – these are cases connected with specifics of the text collection context (for example, a high degree of terms heterogeneity, used in a opinions) low level of the clustering procedure quality, causing in turn inaccuracies in classifying the documents placed at the border between the two clusters, to either of them.

In this context, the program for linguistic analysis, realized the authors methodology of the semantic clustering presupposed realization of the *following specific setup options*, which enhance the quality and resolving power of the method:

– changing the modes of the experiment (modes 1-6);

– changing in the list of stop-words – words that have no semantic load (prepositions, pronouns);

– visual analysis of the cluster's structure (using the tool "magnifying glass") in order to clarify the boundaries of the clusters;

– analysing of the quantitative characteristics of the method (resolving power and quality for the clustering procedure) and selection of the final version of the deviding textual opinion into the clusters.

VII. Conclusions and Further Research Direction

Thus, the authors have developed and studied the methodology of semantic clustering analysis, which allows increasing the quality of the procedure. At the framework of this research:

1. The author's version of the mathematical models of formalization and practical realization of short textual messages semantic clustering procedure is proposed, based on the customer's text-opinions collection LSA-extracting knowledge method. The basic concepts form that model are formulated.

2. An algorithm for semantic clustering of the customer's text-opinions collection is developed, the distinctive characteristics of which is the introduction of the concepts and methods of identification:

– the *point of reference* in the scale of the customer's text-opinions collection closeness determination can serve to coordinate the term with the highest total weight;

– the instrument of documents' *closeness degree identification* on the bases of 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;

– the *measure of similarity* between pairs of documents is justified to consider the difference between the values of their relative spatial coordinates.

3. The author version of the quantitative evaluation of the clustering procedure results level are purposed, namely indicators:

– *resolving power* of the method of semantic clustering – number of resulted clusters;

– *level* of the clustering procedure *quality* – the combination of the degree of the boundaries fuzziness between the clusters and closeness of the similarity measures values within the cluster.

4. Analysis of the specific features and the effectiveness level of various distance measures are conducted. Dependence of changes the quality of the clustering results correspondently the variety of experimental parameters and modes is identified.

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 methodology of semantic clustering analysis to embed it in the ADA system for the Customers' Opinions mining purposes as a part of joint collaboration with Cognitum company.

References

[1] 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

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

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

[4] D. Dobrowolski, P. Kaplanski, A. Marciniak, and Z. Lojewski, "Semantic OLAP with FluentEditor and Ontorion Semantic Excel Toolchain,"IARIA, vol. SEMAPRO 2015: The Ninth International Conference on Advances in Semantic Processing, 2015. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid =semapro_2015_3_30_30051

[5] P. Kapłanski and P. Weichbroth, "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

[6] P. Kapłanski, "Controlled english interface for knowledge bases," Studia Informatica, vol. 32, no. 2A, pp. 485–494, 2011

[7] A. Wroblewska, P. Kaplanski, P. Zarzycki, and I. Lugowska, "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

[8] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41, 1990, pp. 391-407

[9] A. Dasgupta, R. Kumar, P. Raghavan, and A. Tomkins. Variable latent semantic indexing. In ACM SIGKDD 2005, 2005.

[10] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990

[11] Miles Efron. Eigenvalue-based model selection during latent semantic indexing: Research articles. J. Am. Soc. Inf. Sci. Technol., 56(9): pp 969–988, 2005

[12] Ricardo Olmos , José A. LeónUsing, Inmaculada Escudero, Guillermo Jorge-Botana. "Latent semantic analysis to grade brief summaries: some proposals". Int. J. Cont. Engineering Education and Life-Long Learning, Vol. 21, Nos. 2/3, 2011

[13] Foltz, P. W. "Latent semantic analysis for text-based research", Behavior Research Methods, Instruments and Computers, 1996, Vol. 28, No. 2, pp.197–202

[14] Xuren Wang, Qiuhui Zheng. "Text Emotion Classification Research Based on Improved Latent Semantic Analysis Algorithm". Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)

[15] Dumis S, Fumas G, Landauer T et al. "Using Latent Semantic Analysis to Improve Access to Textual Information". Proceedings of Computer Human Interaction, 1988.217-285

[16] Salton G, Wong A, Yang CS. "A Vector Space Model for Automatic Indexing". Communications of the ACM, 1995,18(11) : pp 613-620.

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

[18] Department of Statistics, Stanford University, Fall, 2008. [Online]. Available: http://www.econ.upf.edu/~michael/stanford/maeb4.pdf

[19] Lee C-H. "Learning inductive rules using hellinger measure". Applied Artificial Intelli-gence, Volume 13, Number 8, 1 December 1999 , pp. 743-762(20)

[20] Han, E., Karypis G., Kumar V. "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification". 16th International Conference on Machine Learn-ing – Denver, 1999. – pp. 41-56

[21] Koller D., Sahami M. Hierarchically classyffying documents using very few words // Koller D., Sahami M., Proc. ICML-97. Nashvilee, 1997, pp.170-176