Modeling the Customer's Contextual Expectations Based on Latent Semantic Analysis Algorithms

Nina Rizun[1], Katarzyna Ossowska[1] and Yurii Taranenko [2]

[1] Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland
{nina.rizun, katarzyna.ossowska}@zie.pg.gda.pl
[2] Alfred Nobel University, Naberezhna Lenina Str., 18, 49000, Dnipro, Ukraine
taranen@rambler.ru

Abstract. This paper presents the approach of modeling the system of customer's contextual expectations. The novelty consists in systematic usage of the different technic. The first technic – the theory of Benefits Language as an instrument of forming Concept of Benefits for studied product. The second one – the combination of advantages the probabilistic Latent Dirichlet Allocation (LDA) and Linear Algebra based Latent Semantic Analysis (LSA) methods as an instrument of textual data retrieval. The verification of the proposed approach for specifies type of product – films – was conducted. The main research plan was realized: Contextual Summary using the LDA-based algorithm was formed; the Contextual Frameworks using LSA-based approach were performed; the Manually Created Contextual Expectations Dictionary was built. The results of case study, based of the Polish-language film reviews corpora analysis, allowed to make the conclusions about the possibility to use proposed approach for building the system of Customer's Contextual Expectations.

Keywords: Benefits Language; Latent Semantic Analysis; Latent Dirichlet Allocation; Customer; Contextual Expectations

1 Introduction

Nowadays, in the age of Internet, access to open data detects the huge possibilities for information retrieval. More and more often we hear about the concept of open data which is unrestricted access, in addition to reuse and analysis by external institutions, organizations and people. It's such information that can be freely processed, add another data (so-called remix) and then published. More and more data are available in text format (such as reviews on books, movies, etc.). Algorithms of Latent Semantic Relations Analysis are one of the important tools for extraction and recognition of significant facts from textual data sets. Another aspect of research is to find ways and means of using the information obtained by Semantic tools applying in order to maximize the expected benefit. In this area, one of the modern tools for formulating the concept of benefits is the Benefits Language [1-5]. One of the conditions for the formation of the concept of benefits for studied product is the collection and processing of information about the client's expectations. This process often requires additional time

and financial costs. Therefore, one of the ways to obtain the "maximum benefit" of using the Benefits language is to develop a methodology of building the system of customer's contextual expectations via systematic usage of benefit's theory and algorithms of semantic analysis of textual opinions of open Internet pages'.

2 Related Research

Benefits language is a very well-known technique in door-to-door selling. It's a technique that present product in a way that client would like to buy it. Benefits language is based on F-A-B rule (Feature-Advantage-Benefit). It means, that first are shown features of a product (colour, material, shape etc.) then advantages which are result of those features and in the end – benefits. Benefits are profits which the customers will reach when they will buy and then will be using the product [1, 2].

The purpose of the Semantic Relationships Analysis is to extract "semantic structure" of the collection of information flow and automatically expands term into the underlying topic. Significant progress on the problem of presenting and analysing the data have been made by researchers in the field of information retrieval (IR) [6-8]. The basic methodology proposed by IR researchers for text collection – reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts.

The vector model [9-13] of text representation is one of the first methods used to solve latent semantic relations revealing the topic modeling problems. Initially, this model was used in topic detection tasks by extracting events from the information flow [10, 14]. The representation of the corpora in this case realized with the help of vectors models form, in which each word is weighted according to the chosen weight function [15, 16]. For solving the problem of finding the similarity of documents (terms) from the point of view of the relation to the same topic. The most appropriate metric is cosine measure of the edge between the vectors [9, 10].

Latent Semantic Analysis (LSA) is a theory and method for extracting context-dependent word meanings by statistical processing of large sets of text data [18]. The LSA also works with a vector representation of the "bag of words" text units. The text corpora are represented as a numeric matrix Word-Document, the rows of which correspond to words, and columns to text units – documents.

The revealing the Latent Semantic Relationships via LSA between words/documents usually applying the singular decomposition [15, 16]. According to the theorem on singular decomposition (SVD), any real rectangular matrix can be decomposed into a product of three matrices [9-11].

The significant *shortcomings* of LSA method are: probabilities for each topic and the document distributed uniformly, which does not correspond to the actual characteristics of the collections of documents [13, 15-16]; increasing the size of the analyzed documents significantly reduces the quality of recognition of hidden relations [14, 17].

Another group Latent Semantic Relationships studying is the Probabilistic thematic modeling. It is a set of algorithms that allow analyzing words in textual corpuses and extract from them topics, links between topics [19-13]. Latent Dirichlet Allocation (LDA) is a generative model that explains the results of observations using implicit

groups, which allows one to explain why some parts of the data are similar. It was proposed by David Blei [20 21].

The main *drawback* of the LDA is the lack of convincing linguistic justifications. From the point of view of texts, the assumption of the Dirichlet distribution is not justified. Additionally, in the process of the assigning the topics to documents usually LDA use the maximal form possible (not always very high) level of probability of a documents belonging to the topic.

The main *purpose* of this paper to contribute the new methodological approach to the theory of modeling the system of customer's contextual expectations via systematic usage of:

– technics of Benefits Language as an instrument of forming Concept of Benefits for studied product,

– advantages of the algorithms of Latent Semantic Relations Analysis of textual customer's opinions of open Internet pages' as a data retrieval instrument for building the Concept of Benefits.

For demonstration the of basic workability of the author's approach realization as a sample for *case study* the Polish-language film reviews corpora (FRCS) from the filmweb.pl is used.

3 Methodology

On this paper the following author's definitions will be used:

1. *Corpora* (films reviews corpora sample, FRCS) is a collection of the textual Documents.
2. *Term* is a word after preprocessing.
3. *Latent Semantic/Probabilistic topics* is a basic unit of Latent Semantic Relations, received by *LSA/LDA* approach.
4. *Subjectively Positive* (*CFSP*) and *Subjectively Negative (CFSN) Corpora Samples* is a result of the classification of the FRCS on the basis of information on the subjective assessment of films by the reviewers (measured by 10-point scale). For this purpose the following heuristic was adopted: to consider the CFSP, if the subjective review's assessment is more than 7 points, and CFSN – if it is equal or less 4 points.
5. *Contextual Summary* (CS) is a set of Latent Probabilistic Topics (LPT), described the main context of the *Corpora.*
6. *Contextual Framework* (CF) is a set of main Latent Semantic Topics (LST) with keywords, considered in the *Contextual Summary*.
7. *Contextual Bigram* (CB) is a combination of (keyword (noun) and its contextually close term (adjective), which describe the particular topic.
8. *Contextual Expectations Dictionary* (CED) is a set of CB with the numeric indicators of the level of their positivity/negativity with respect to particular topic.
9. *Hierarchical Semantic Corpora* (HSC) is a structure of the clustered paragraphs of the FRCS, which relate to a particular Topic form CF.

3.1 Novelty and Research Plan

With *aim* to develop the system of customer's contextual expectations the following scientific research question was raised: *Is it possible to build the system of customer's contextual expectations via systematic usage of Benefit's Theory and Algorithms of Semantic Analysis of Textual Opinions?*

For finding the answers for this question the following main concepts were formulated:

Concept 1. Taking into account the specificity of chosen case study and due to the peculiarity of the requirements of film's review writing [11], each paragraph of such document could be identified as an indivisible contextual unit, which characterized by a particular Latent topic and should be analyzed separately.

Concept 2. The quality of Contextual Frameworks for the *CFSP* and *CFSN* building, could be improved due to the systematic usage: LDA-method recognizing of the Latent Probabilistic Topics within the Samples; LSA-method of Latent Semantic Relations recognition within the set of Latent topics.

Concept 3: The system of customer's contextual expectations can be interpreted as a CED after transformation using the special expectation's labels "*Should have"*, "*Could have"* and "*Won't have"*.

On the bases of this concepts, the following research plan was developed: *to demonstrate the possibility of building the system of customer's contextual expectations via the realizing the following steps*:

1. Formation of the *Contextual Summary* for CFSP and CFSN using the LDA-based algorithm and Paragraph-oriented Approach Concept.
2. Building the *Contextual Frameworks* for CFSP and CFSN using LSA-based approach.
3. Creating a *HSC* on the bases of Contextual Frameworks for CFSP and CFSN.
4. Manually Creating the *CED*.
5. Formation the system of *Customer's Contextual Expectations* using the Benefits Language concept.

The size of *case study* sample for research plan realization is 3000 films reviews (1500 *Subjectively Positive* and 1500 *Subjectively Negative*). All texts were presented in .txt format files and processed using the Python-based programing tools.

3.2 Latent Semantic Analysis Results

Step 1. Formation of the List of the Latent Probabilistic Topics for CFSP and CFSN using the LDA-based algorithm and Paragraph-oriented Approach Concept

This step presupposes the realizing the following stages: text preprocessing [11]; text preparation; LDA-modeling.

The stage of *LDA-modeling* gives the possibility to receive the *Contextual Summary* of Latent Probabilistic Topics with information about most probable (significant) words and assign these topics with maximal probability to particular paragraphs. For obtaining

the optimal combination – Number of topics / terms in topic main – the values of Perplexity [19] were analyzed. The optimum value of the Perplexity achieved in the point, when further changes in the parameters do not lead to its significant decrease.

The structure of the *Contextual Summary*, as a *CS results*, is presented in Table 1. The recall rate as the ratio of the number of topically recognized paragraphs (probability of belonging the paragraph of topic >0,7) to the total number of paragraphs is within 90-95%.

Table 1. The Structure of the Contextual Summary

CFSP		**CFSN**	
Number of paragraphs	9900	Number of paragraphs	10650
Number of topics	26730	Number of topics in paragraph	33015
Average Number of topics in paragraph	2.7	Average Number of topics in paragraph	3.1
Average Number of terms in Topic	6.8	Average Number of terms in Topic	7.3
Average Perplexity Value	102	Average Perplexity Value	99

Step 2. Building the CF for CFSP and CFSN using LSA-based approach

For recognition of Latent Semantic Relations within the *CS* and recognition *CF* via semantic clustering of the CS elements, the following stages are presupposed: text preprocessing; creating the Term-Document Matrix; SVD process; identifying the hidden semantic connection within the CS; LSA clustering of CS elements / terms in the semantic dimension [10].

After realization of first three stages, based on the matrices of cosine distances between the vectors of CS elements (topics) and terms, the LSA clustering step have been realized. As a method the k-means clustering [12, 21] had been chosen. For finding the optimal number of clusters, the following *combination* of two factors was used [9]: the degree of the boundaries fuzziness between the clusters; the closeness of the similarity measures values within a cluster of the CS elements/terms similarity within the cluster

The optimal combination for these two parameters, as a *CS results*, was reached for 5 clusters for *CFSP* and for 4 clusters for *CFSN*. On the bases of this information, for each cluster of both CS the *Contextual Frameworks* were obtained (Tables 2-3). For this: the list of keywords was identified (KW); the weights for keywords were formed (W); the Contextual Labels of each of the CF were specified.

Table 2. Contextual Frameworks for CFSP

"Hero"		**"Director"**		**"Script"**		**"Plot"**		**"Spectator"**	
KW	**W**	**KW**	**W**	**KW**	**W**	**KW**	**W**	**KW**	**W**
hero	1.0	director	1.0	script	1.0	plot	1.0	spectator	1.0
playing	0.8	creator	0.8	history	0.8	hero	0.8	fan	0.6
person	0.6	stage	0.6	writer	0.6	action	0.6	watch	0.8
actor	0.4	drama	0.4	picture	0.4	film	0.4	interest	0.4
main	0.2	effect	0.2	layer	0.2	history	0.2	film	0.2

Table 3. Contextual Frameworks for CFSN

"Hero"		"Actor"		"Creator"		"Plot"	
KW	**W**	**KW**	**W**	**KW**	**W**	**KW**	**W**
hero	1.0	actor	1.0	writer	1.0	plot	1.0
spectator	0.8	character	0.8	director	0.8	history	0.8
climate	0.6	picture	0.6	film	0.01	stage	0.6
person	0.4	role	0.4	spectator	0.6	script	0.4
fan	0.2	history	0.2	action	0.2	scenarist	0.2

Step 3. Creating a HSC on the bases of Contextual Frameworks for CFSP and CFSN

After receiving the Contextual Frameworks for CFSP and CFSN the basic algorithm for *HSC* building aw developed. The main idea of this algorithm – the process of paragraph selection, semantically close to particular CF (as a set of keywords).

Step 4. Manually Creating of Film Reviews' Corpora-based Contextual Expectations Dictionary

The algorithm of the Manually Creating CED presupposed the retaliation of the following stages:

1. Forming a set CB (is implemented using Python NLTK functions `nltk.bigrams(...))`.
2. Estimation of weights of CB. Assumes the definition of the absolute weight of bigrams, estimated by the frequency of occurrence of this bigram on the CFSP and CFSN.
3. Correction of Weights of CB using Relevance Frequency. In this paper, in addition to the frequency of use of the C of positive or negative tonality, the parameter is used to reverse the frequency of using this term in negative or positive utterances – RF (Relevance Frequency) [24]. The basic meaning of the RF measure is that the weight of the word is calculated on the basis of information about the distribution of this bigrams in the texts of the collection and takes into account the belonging of the collection texts to certain classes (positive, negative):

$$RF_S = \log_2\left(2 + \frac{a}{\max(1,b)}\right) \tag{1}$$

where a – the number of paragraphs related to category S (CFSP or CFSN) and containing this bigram; b – the number of paragraphs not related to category S and containing this bigram as well.

4. Forming the rules for Polarity Scores building: formulating the Polarity limits for scaling the Weights and there interpreting as an expectation's labels "*Should have"*, "*Could have"* and "*Won't have"*.
5. Analysis of the structure of the received elements of the CED for the positive and negative parts of the Corpora.
6. Based on this algorithm, as an CS result, the *Corpora-based Hierarchical* CED was obtained. Generally, 797 CB for CFSP and 429 for CFSN were formed.

3.3 Defining the System of Contextual Expectations

Corpora-based Hierarchical CED is a main source for realizing the step of Contextual Expectation System formulation. As identified in step 4, CED contains the positive, negative and neutral features of opinions about specific product "films" and concerns a separate topic. To create a demo example, the authors modeled the situation of having information on only four (identical) topics for each (CFSP and CFSN) group of CED: *Hero*, *Actor*, *Director* and *Script*. The algorithm of Contextual Expectation System formulation presented in Figure 1.

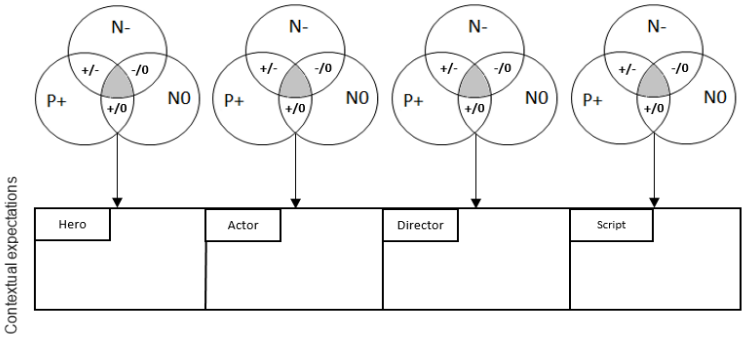

Fig. 1. Algorithm of Contextual Customer's Expectation System Forming

Legend: **P+** – positive, **N-** – negative; **N0** – neutral; **+/-** - positive and negative; **-/0** – negative and neutral; **+/0** – positive and neutral; ▒ positive and negative and neutral

For demonstration this algorithm realization, the data of CED for keyword "*Hero"* was used (Tables 4-5). The process of the Contextual Expectation System builds on the following concept: Based on MoSCoW technique [1-4], for prioritization customer requirements, authors grouped features in three categories: "*Should have*" – represents a high priority item that should be included in the solution, if possible, "*Could have*" – describes a requirement that is perceived as desirable, but not necessarily (it will be included if time and resources allow), "*Won`t have*" – represents a requirement that, with the consent of stakeholders, will not be implemented in a given release.

Table 4. The fragment of Polarity of Hero's Positive` Features

Birgams		Polarity scores	Polarity
hero	amazing	1.00	Positive
hero	beautiful	0.83	Positive
		…	
hero	solid	0.33	Neutral
hero	wakes	0.33	Neutral
		…	
It is not a	hero	-1.00	Negative

Table 5. The fragment of Polarity of Hero's Negative` Features

Birgams		Polarity scores	Polarity
lack of	distinctive	0.83	Negative
is not	valid	0.83	Negative
		…	
angry	hero	0.33	Neutral
demonic	hero	0.17	Neutral
		…	
beautiful	hero	-1.00	Positive

Features from "*Should have*" group are those which were indicated as positive, that why movie should have hero with this features to increase number of positive opinions. "*Could h*ave" features are neutral one. In authors opinion, those are the feature than movie can have but is not necessary and it would not have influence on opinion while negative will have bad influence and those features should not occur. In our *CS* example, we didn't have features that was, at the same time, positive and negative, positive and neutral, negative and neutral, positive and negative and neutral that`s why common space between those kind of features is empty (Figure 2).

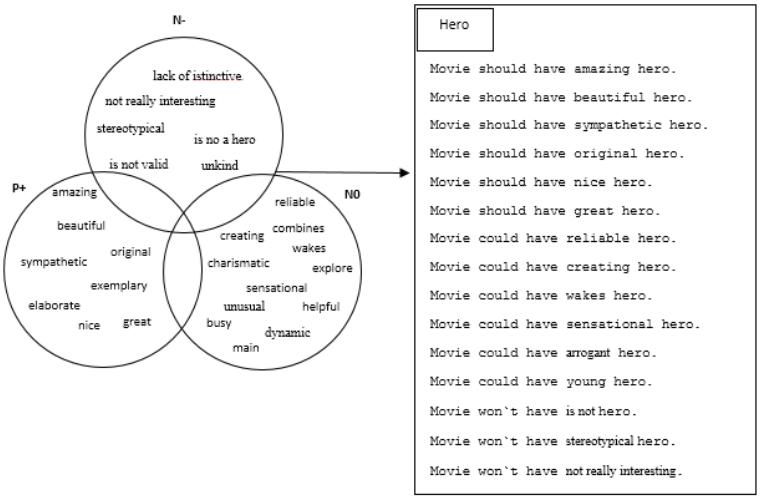

Fig. 2. The Example of the Algorithm of Contextual Expectation System Forming Realization

The Last step of Contextual Expectation System Forming – creating advantages and benefits from features. According to the benefits language and rule F-A-B first we create advantages which are results of features, then, in the end, benefits. In the Table 8 we propose the example of Features for Benefit: "*The spectator is more likely to watch the movie*". But for real situation the benefits should be different (Table 6).

Table 6. The Example of Features-Advantages-Benefits about "Hero"

"Hero"	
Feature	**Advantage**
Amazing	Arouses admiration
Beautiful	Slightly
Sympathetic	It arouses positive feelings
Original	Arouses curiosity
Elaborate	Professional approach to the viewer
Nice	It arouses positive feelings
Great	Arouses admiration

3.4 Conclusions

In this paper authors contribute the new scientific approach for the system of customer's contextual expectations building. Verifying the approach on the case study of Polish-language Film Reviews Corpus analyzing helps to authors to find the answer to the main research question: the systematic usage of Benefit's Theory and combination of the Algorithms of Semantic Analysis of Textual Opinions is the real scientific instrument of building the system of customer's contextual expectations on the bases of textual opinions of open Internet pages'.

Prospects for future research in terms of the use of Semantic Analysis tools consist in the possibility of using the Concept of Benefits for studied product to create a Methodology for: measuring the degree of positivity / negativity of the user's opinions on the basis of an assessment of the measure of closeness of their opinions to ideal; recognizing the structure of the Customer's expectations expressions formulations.

4 Acknowledgments

The research results, presented in the paper, are supported by the Polish National Centre for Research and Development (NCBiR) under Grant No. PBS3/B3/35/2015, the project "Structuring and classification of Internet contents with the prediction of its dynamics".

References

1. Ossowska K., Szewc L., Weichbroth P., Garnik I., Sikorski M.: Exploring ontological approach for user requirements elicitation in design of online virtual agents// Information Systems: Development, Research, Applications, Education/ ed. Stanisław Wrycza : Springer International Publishing, 2016, s.40-55
2. Ossowska K., Szewc L., Orłowski C. The principles of model building concepts which are applied to the design patterns for Smart Cities (April. 2017), Intelligent Information and Database Systems. DOI: 10.1007/978-3-319-54430-4.

3. Ossowska K. Design modern technologies for older people by using expert systems containing benefits language (May 2017), Zeszyty Naukowe Politechniki Poznańskiej, Organizacja i Zarządzanie
4. Ossowska, C.Orłowski, Projektowanie systemów informatycznych z wykorzystaniem języka korzyści (December 2016), Projektowanie i realizacja systemów informatycznych zarządzania. Wybrane aspekty
5. Ossowska K., Czaja A, Model of personalization website using benefits language (Jun 2017)
6. Baeza-Yates R., Ribeiro-Neto B. (2011) Modern Information Retrieval. Addison-Wesley, Wokingham, UK, 1999. Second edition
7. Furnas G.W., Deerwester, S., Dumais S.T., Landauer T.K., Harshman R.A., Streeter L.A., Lochbaum K.E. (1998) Information retrieval using a singular value decomposition model of latent semantic structure. In Proc. ACM SIGIR Conf., s. 465-480, ACM, New York
8. Gerard Salton, Michael J. (1983) McGill Introduction to modern information retrieval. New York McGraw-Hill - McGraw-Hill computer science series, XV, 448 p
9. Rizun N., Kapłanski P., Taranenko Y. (2016) Development and Research of the Text Messages Semantic Clustering Methodology. 2016, Third European Network Intelligence Conference, Publisher: ENIC, # 33, pp.180-187
10. Rizun N., Kapłanski P., Taranenko Y. (2016) Method of a Two-Level Text-Meaning Similarity Approximation of the Customers' Opinions. Economic Studies – Scientific Papers. University of Economics in Katowice, Nr. 296/2016, pp.64-85.
11. Rizun N., Taranenko Y. (2017) Development of the Algorithm of Polish Language Film Reviews Preprocessing. Proceeding of the 2nd International Conference on Information Technologies in Management, Publisher: Rocznik Naukowy Wydziału Zarządzania WSM, http://www.wsmciechanow.edu.pl/rocznik-naukowy/ (in print).
12. Kapłanski P., Rizun N., Taranenko Y., Seganti A. (2016) Text-mining Similarity Approximation Operators for Opinion Mining in BI tools. Chapter: Proceeding of the 11th Scientific Conference "Internet in the Information Society-2016", Publisher: University of Dąbrowa Górnicza, pp.121-141
13. Salton G., Wong A., Yang C. S. (1975) A Vector Space Model for Automatic Indexing, Communications of the ACM, Vol. 18, Nr. 11, s. 613-620
14. Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988) Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285
15. Deerwester S., Susan T. Dumais, Harshman R. (1990) Indexing by Latent Semantic Analysis. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf
16. Eden L. (2007) Matrix Methods in Data Mining and Pattern Recognition, SIAM.
17. Ed. Charu Aggarwal, Cheng Xiang Zhai, (2012) Mining Text Data (Springer).
18. Bahl L., Baker J., Jelinek E., & Mercer R. (1977) Perplexity – a measure of the difficulty of speech recognition tasks. In Program, 94th Meeting of the Acoustical Society of America, volume 62, page S63.
19. Blei D., Ng A., Jordan M. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 3: pp. 993–1022.
20. Blei, David M. (2012) Introduction to Probabilistic Topic Models. Comm. ACM 55 (4), April, 2012: pp. 77-84
21. Daud Ali, Li Juanzi, Zhou Lizhu, Muhammad Faqir (2010) Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. pp. 280-301.
22. David M. Blei. Topic modeling. http: //www.cs.princeton.edu/~blei/topicmodeling.html