Features of Big Text Data Visualization for Managerial Decision Making

E. A. Makarova[1], D. G. Lagerev[1], F.Y. Lozbinev[2]
m4karova.e@yandex.ru | LagerevDG@mail.ru | flozbinev@yandex.ru
[1]Bryansk State Technical University, Bryansk, Russia
[2]RANEPA, Bryansk, Russia

This paper describes text data analysis in the course of managerial decision making. The process of collecting textual data for further analysis as well as the use of visualization in human control over the correctness of data collection is considered in depth. An algorithm modification for creating an "n-gram cloud" visualization is proposed, which can help to make visualization accessible to people with visual impairments. Also, a method of visualization of n-gram vector representation models (word embedding) is proposed. On the basis of the conducted research, a part of a software package was implemented, which is responsible for creating interactive visualizations in a browser and interoperating with them.

Keywords: *visualization, natural language processing, web application accessibility.*

1. Introduction

With the acceleration of scientific and technological progress, as a result, economic growth rates of both global and local markets are rapidly increasing. According to the study [5], the number of mergers and acquisitions in Russia in 2017 increased by 13%. Besides, the number of originated loans is growing. According to the United Credit Bureau, the annual number of loans issued in Russia has increased by 22%, while lending has increased by 53%. In addition to the accelerated capital turnover, growth is also observed in the labor market. In Antal Russia's survey, 27% of employers reported an increase in staff turnover in their companies over the past year [12].

Higher velocity and number of transactions conducted in various spheres of social and economic activity results in greater burden on managers at various levels. This requires either an increase in the decision-making staff or enhancement of information systems supporting managerial decision-making in order to reduce the people's workload. Beside traditional data used in such systems (e.g., credit history and capital for scoring systems used in loan approval), many researchers and manufacturers of technological solutions use unstructured information sources about legal entities and individuals involved in transactions. Examples of such information are data from mass media, social networks, etc.

In addition, some studies have shown that adding analysis of text data from social media to prediction models results in greater accuracy. For example, they help to increase the accuracy of legal entity's bankruptcy prediction [7]. Hence, one of the stages of using managerial decision-making support systems is loading text information into them about an object of socio-economic activity for further use.

Objects of socio-economic relations are widely represented on the Internet both through official websites and in the form of digital reputation, i.e., reviews, news, what appears on the network about them without their direct intervention. However, the amount of such data is constantly growing (due to data duplication, borrowing data from another source, etc.), which requires optimization in terms of speed and cost of their collecting and processing. As small and medium-sized businesses have to contact with ever increasing number of people in the course of their activity, the risk of a transaction with legal entities or individuals unreliable in terms of tax or other laws increases, which may entail long-term consequences, such as costs in public image, etc., and may result in legal entity's bankruptcy.

On the one hand, decisions need to be made faster and faster, their number is growing, which can lead to more errors and risks. This problem can be solved by the integration of data mining systems into DSS by using large volumes of unstructured data available for analysis [10]. On the other hand, the process of collecting and preprocessing these data requires the involvement of serious human and computational resources, which can nullify the economic benefits obtained by adding unstructured data to the process of managerial decision making.

Currently, there are various analytical systems that work not only with structured data but also with unstructured ones, including text data downloaded from social media [11]. In these systems, visualization is rarely used at the stage of collecting and preprocessing big text data. However, collection and preprocessing of data for such systems is still quite time consuming, and there is a significant risk of using irrelevant documents as data sources. Usually one of the two approaches is used: either a fully automatic analysis of collection and preprocessing results (faster) or a fully manual review of a large array of documents (more qualitative). This article discusses a hybrid approach based on vector data visualization that allows adding expert assessment of document relevance at the stage of data collection and preprocessing [18].

2. Extracting information from sources of varying degrees of structuring

Let us consider in more detail the process of collecting and analyzing text information from various sources presented in Figure 1.

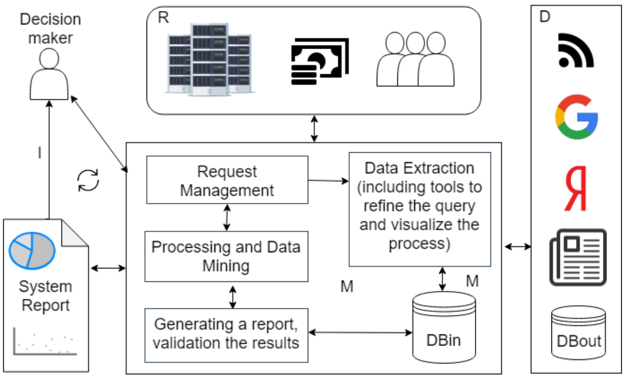
Decision maker
R
D
I
Request Management
Data Extraction (including tools to refine the query and visualize the process)
Processing and Data Mining
M
M
System Report
Generating a report, validation the results
DBin
DBout

Fig. 1. Overview diagram of collecting and processing data

All the processes presented in the diagram are important for the efficient use of unstructured text data for managerial decision making. However, in this paper, the greatest emphasis is placed on the process of collecting information since accuracy and resource consumption of further analysis depends on the quality of the collected data.

In the diagram, DBin is an internal database containing trained models for collecting and analyzing information as well as accumulated information about analysis objects. DBout refers to external databases (structured sources) attached by the user.

Conceptual model for collecting text information is

$$S = < R, M, D; I>,$$

where R is resources (temporary, material, human);

M is information about previous generations;

D is data sent for analysis;

I is amount of information relevant to the task that is available to the system.

$$R = < R_m, R_h, R_p, T>$$

where R_m is money spent on paid services (various APIs and services);

R_p is the number of available experts in the subject area;

R_h is hardware limitations. For some problems, hardware resource limit will only mean speed of calculations and, accordingly, be considered together with T parameter, but for some language processing tasks, for example, when using vector representation of words, the amount of RAM available will be key to the usability of these methods;

T is time spent on collecting information, which, in turn, can be decomposed into the following components:

$$T = T_u + (T_e + T_d) + T_a,$$

where T_u is time spent by the main user of the system;

T_e is time spent by the expert who will check and resolve manually various situations difficult for machine-aided processing;

T_d is time delay between expert's response and continuation of processing (the error to account for the non-round-the-clock availability of the expert);

T_a is time spent on automatic processing.

The task of optimizing data collection consists in reducing R, H, D parameters while increasing I parameter.

It is also assumed that a number of parameters will decrease with each subsequent use of the system due to the training of users and models, accumulation of useful knowledge about the objects of research.

In matters of improving the efficiency of information collection, there are two extremes: to make all the work fully automatic, thereby saving on human resources, or to make process control completely manual. In this paper, an "intermediate" version is considered when an expert is engaged in evaluating the effectiveness of the collection process, but due to the use of various tools, such as visualization, his work time is significantly reduced [**9**].

In addition, the following approaches are used in the developed software package to optimize information collection before analysis:

1) refinement of search queries;
2) ignoring duplicate information;
3) preliminary data analysis, etc.

3. Visualization of big text data for data mining optimization.

Let us consider some features which require human interference for more efficient work and for which various visualization methods have been studied and refined as part of the work on the system [19]. As a data source in this example, we will use web-based media, but the methods being developed are applicable, with some adjustment, to all sources of a similar structure.

When setting up uploading of text documents from a certain source, by which the search is possible, users of the system may encounter the fact that the query does not correspond to the required result, e.g., if the request has turned out to be too "general" or information about homonymous objects is present in the same sources. A way out of this situation may be to view a part of the collected text documents, their brief contents or some metadata. It is time consuming for the user (subject matter expert or employee). Another way to familiarize the user with the downloaded data is to visualize it. In [17], it was already demonstrated that the difference in the content of documents is noticeable in the visualization of an n-gram cloud type (n-gram refers to a word sequence), and it was noted that this method requires further refinement. In the current implementation, visualization has undergone a number of changes, such as combining word weights that have semantic proximity below a certain threshold, excluding "stop words" and words with small weights from visualization.

One of the refinement directions for the visualization method will include its adaptation for use by various groups of people, including those with disabilities. When developing visualizations, it is important to consider all user groups, not only in terms of compliance with international standards but also in terms of the increasing number of potential users. For example, more than 5% of population suffer from various forms of color vision deficiency, which can prevent the user from interacting with the visualization to a full extent [2].

In recent years, the topic of accessible visualization has gained great interest from researchers and software development service providers [14]. For example, Square, Inc. [1] has published an open-source guide to creating accessible data visualizations. Among visualizations they propose there are various types of charts and graphs.

Visualizations related to the analysis of text information are comparatively little studied from this side. Next, we will consider two examples of such visualizations that are important for collecting text data in the described software package.

Classic works devoted to the construction of an n-gram cloud (or "tag cloud", "word cloud") [4, 15], which described algorithms employed by libraries implementing visualization data, could not take into account WCAG recommendations on application adaptation for people with visual impairments since they had appeared before these guidelines were developed.

As part of the software package, a client-server architecture subsystem was implemented as a web application that provides interactive visualizations and implementation of user analysis to data collection process. So, for example, the developed n-gram cloud visualization takes into account WCAG 2.1 recommendations. Therefore, the following restrictions and additions have been introduces to the algorithm:

1) restriction on the contrast of colors
2) exclusion of vertical text orientation [8]
3) setting the minimum and maximum text sizes
4) adding advanced user settings.

Considering that the interface of the existing system was developed as a web application, it will be reasonable to rely on the algorithms used to create and display tag clouds [4] adapting them to the problem being solved and WCAG 2.1 recommendations.

Many ready-made visualization tools do not take into account contrast for different groups of people, including those suffering from visual impairment and color deficiency. However, it should be understood that the purpose of creating a tag cloud is often to effectively illustrate an array of information rather than a detailed analysis [15].

Color contrast according to WCAG 2.1 is:

$$(L_1 + 0.05) / (L_2 + 0.05) \ > \ C_{min}$$

L_1, L_2 are relative brightness of compared colors.

Since all words in visualization will be interactive, the required contrast for them should be calculated as for controls, i.e., $C_{min} = 3$ for n-grams located separately. In addition, contrast for each individual color compared to the background should be equal to $C_{min} = 4,5$ [16]. Calculations show that it is possible to find only 2 colors that will be simultaneously contrastive with the background and between themselves.

Also there appear restrictions on the font size. On the one hand, the minimum size of n-grams should not be less than 16pt [16]. On the other hand, the same standard imposes the condition that all texts on a page can be magnified to 200% maintaining their readability, which constrains the maximum possible font size when displaying a page at the size of 100%. In order to maintain the approximate position of containers in which the text will be when enlarged, CSS Grid technology [3] and slicing floorplan algorithm [4] were used for the interface design.

Besides, the user should be able to add custom settings for colors and sizes of visualization. Let us consider a specific example. In [17], it is described in detail how visual analysis of a part of text documents on a search query allows understanding whether various search entities need to be added or excluded from the query. Figure 2 shows visualization implementation for adjusting data collection for "BMZ" object (AO UK BMZ – Bryansk Machine-Building Plant). Presented figures demonstrate the work with the texts in Russian. User's task is to assess reputation of this legal entity. To do this, it is necessary to collect data on this object. The goal of this visualization is to track whether the context of the request, that implied the search for an enterprise located in the Bryansk region, was transmitted correctly. As the user can see from the visualization, the search settings were incorrect, which resulted in occurrence in the collected data of many documents related to the activity of a similar enterprise in the Republic of Belarus. Exclusion of text documents containing the word "Belarus" from the search results increased significantly the accuracy of the collection by discarding also documents with references to such objects as "Africa", "Chad", etc.

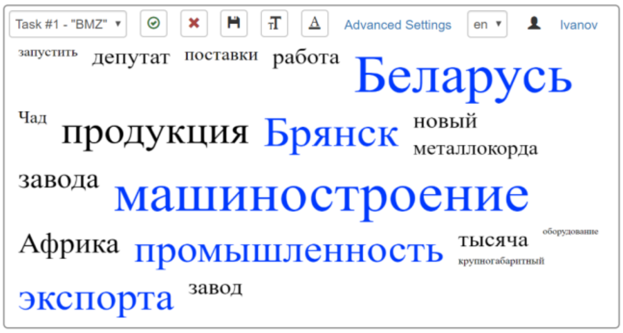

Fig. 2. N-gram cloud visualization of a text document collection in Russian

In addition, since we are talking about displaying in the browser, all elements will have the "tabindex" attribute in ascending order as the significance decreases in the sample and the "aria-label" attribute with the weight of this element to facilitate the perception by people who have vision deficiency and use special programs for reading from the screen.

Typically, to solve a data collection configuration tasks required to view (or use visualization) 20-30 random documents from the collection of documents, depending on the amount of available data. An experiment was conducted on the effect of the method of solving configuration tasks in which five groups of users participated: users of groups 1 and 2 to solve the data collection configuration tasks using the visualization with standard settings, users of groups 3 and 4 used quick skimming of documents, users of group 5 used visualization with user settings, pre-setting time is also included in the final calculation.

The test results for some tasks are presented in table 1. Prior to working with the tasks presented, all users were trained on a test task. Some user groups performed only one group of tasks (for example, analyzing entities associated with "BMZ") at one time, while others immediately started the next task after solving the current task.

On average, time saved using visualization, compared to a quick skimming of texts, varies depending on user's familiarity with the system and ranges from 18 to 42%.

Table 1. Average time spent by the user on one document per document (in seconds).

Task Group	BMZ	BMZ + Bryansk	BMZ + Bryansk +Industry	Ecofrio	Ecofrio + potatoes
Group 1 (one task)	12,5	13	11,5	13	12
Group 2 (three task)	12	13,5	13	14	13
Group 3 (one task)	17,5	19	18,5	16	14
Group 4 (three task)	17	15	14,5	17,5	15
Group 5 (three task)	11	10,5	11	12	11,5

The work on this topic [17] demonstrates how word embedding models [6] pre-trained on different collections of text documents group words differently in terms of their semantic proximity. Also, errors related to the content of the source data occur in models built on word embedding. In the described system, these models are used not only to simplify visualization but also to remove duplicate documents during their further processing. Canvas-based visualization was developed [13] to give the user an opportunity to edit acceptable boundaries of semantic proximity (or cancel n-gram combining if options proposed are unacceptable for the problem being solved). In the center of visualization there is a word position of which in the word embedding model is being explored. Distances from an n-gram are defined so that a two-dimensional vector would be equal to the similarity indicator of this n-gram to the one under study (by default, this value is 0.4). Further, the algorithm selects positions for n-grams in such a way as to ensure the readability of n-grams, including the recommendations described above (no intersections with other elements, a horizontal text of an acceptable size). An example of this visualization for n-grams having maximum semantic affinity with the word "industry" is presented in Figure 3.

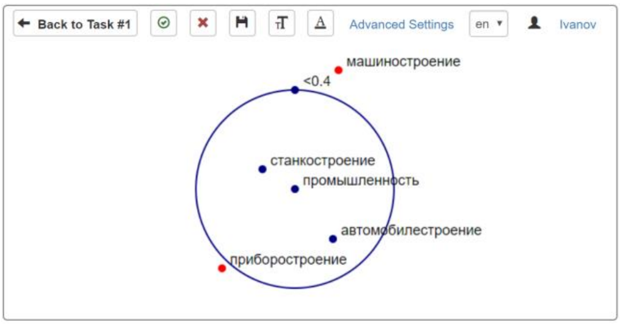

Fig. 3. N-gram based nearest neighbor visualization in the word embedding model

Table 2 demonstrates how using n-gram cloud visualization and applying analysis results to the search query parameters increase the number of relevant documents received during data collection (for 20 random documents from a search sample).

Table 2. Impact of manual adjustment of the request on the number of relevant document search tasks

Number of relevant documents	Object 1 "BMZ"	Object 2 "Isoterm"	Object 3 "Ecofrio"	Object 4 "Spetsstroy"
Before user adjustment	20%	30%	85%	10%
After adjustment	85%	45%	90%	20%

On average, there has been registered an increase in the number of relevant documents by about 24%. The number of relevant documents in the experiment was determined by the method of expert viewing of 20 random documents from a search sample.

4. Conclusion

Adding textual data to analyzed ones in the process of managerial decision making can increase the efficiency. In this paper, special attention is paid to the process of collecting text data from various sources. It is shown that visualization of big text data can significantly reduce time spent on its human processing: time savings compared to skimming of texts is from 18 to 42%, and the number of relevant documents found increases by about 24%. Besides, a part of a software package has been developed, which allows for visualization of text data and models of vector representation of words. When developing visualization algorithms, it is necessary to take into account international standards for creating web applications for people with disabilities, thus making them [applications] accessible to a wide range of users.

In the future, it is planned to continue the study of efficient data collection methods for analysis to support managerial decision-making. In particular, it is planned to study in more detail n-gram vector representation and its use for identifying and deleting duplicate data.

5. References

1. Accessible Colors for Data Visualization. Available by link: https://medium.com/@zachgrosser/accessible-colors-for-data-visualization-2ad64ac4ee7e
2. Causes of Colour Blindness. Available by link: http://www.colourblindawareness.org/colour-blindness/causes-of-colour-blindness/
3. CSS Grid – Table layout is back. Be there and be square. Available by link: https://developers.google.com/web/updates/2017/01/css-grid
4. Kaser O., Lemire D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. Tagging and Metadata for Social Information Organization. A workshop at WWW2007, pp 1086-1087.
5. KPMG presents the results of a survey of Russia's mergers and acquisitions market in 2017. Available by link: https://home.kpmg/ru/en/home/media/press-releases/2018/03/ma-survey-2017.html
6. Kutuzov A, Kutuzov I. (2015) Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the Dialog 2015 Conference, Moscow, Russia
7. Mai F., Mai T., Ling C., Ling M. (2018). Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. European Journal of Operational Research. doi: 10.1016/j.ejor.2018.10.024.
8. Make your information more accessible. National Disability Authority. Available by link:http://nda.ie/Resources/Accessibility-toolkit/Make-your-information-more-accessible/
9. Podvesovskii A.G., Isaev R.A. (2018) Visualization Metaphors for Fuzzy Cognitive Maps. Scientific Visualization, vol. 10, no. 4, pp. 13-29. doi: 10.26583/sv.10.4.02
10. Podvesovskii A.G., Gulakov K.V., Dergachyov K.V., Korostelyov D.A., Lagerev D.G. (2015) The choice of parameters of welding materials on the basis of fuzzy cognitive model with neural network identification of nonlinear dependence. Proceedings of the 2015 International Conference on Mechanical Engineering, Automation and Control Systems (MEACS) (Tomsk, Russia, December 1-4, 2015), IEEE Catalog Number: CFP1561Y-ART, pp. 02-38-NSAP. doi: 10.1109/MEACS.2015.741490
11. Prangnawarat N., Hulpus I., Hayes C. (2015) Event Analysisin Social Media using Clustering of Heterogeneous Information Networks. The 28th International FLAIRS Conference (AAAI Publications) (AAAI)
12. Staff turnover has started to grow. Available by link: https://www.antalrussia.com/news/staff-turnover-has-started-to-grow/
13. The canvas elements. Available by link: https://html.spec.whatwg.org/multipage/canvas.html#the-canvas-element
14. The Future of Data Visualization: Predictions for 2019 and Beyond A. Available by link: https://depictdatastudio.com/the-future-of-data-visualization-predictions-for-2019-and-beyond/
15. Viégas B., Wattenberg M., Feinberg J. (2009) Participatory visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics 15, no. 6, pp. 1137–1144. doi:10.1109/TVCG.2009.17
16. Web Content Accessibility Guidelines (WCAG) 2.1. Available by link: https://www.w3.org/TR/WCAG21/
17. Zakharova A.A., Lagerev D.G., Makarova E.A. (2019) Evaluation of the semantic value of textual information for the development of management decisions. CPT2019 The Conference Proceedings, TzarGrad, Moscow region, Russia
18. Zakharova A.A., Vekhter E.V., Shklyar A.V. (2017) Methods of Solving Problems of Data Analysis Using Analytical Visual Models. Scientific Visualization, vol. 9, no. 4, pp. 78-88. doi: 10.26583/sv.9.4.08
19. Zhao J., Zhao G., Zhao L., Zhao W., (2014). PEARL: An Interactive Visual Analytic Tool for Understanding Personal Emotion Style Derived from Social Media. IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings. doi: 10.1109/VAST.2014.7042496.