Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 2189–2196

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

The method of multidimensional approach to text summarization

Piotr Janaszkiewicz[a], Przemysław Różewski[a*]

[a]*Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Żołnierska 49, 71-210 Szczecin, Poland*

Abstract

Nowadays, the amount of different type of data for analysis is growing at an alarming rate. Moreover, the size and quantity of textual materials make the extraction of specific information from them more and more complicated and sometimes impossible. The use of external data sources, Linked Data and Online Analytical Processing (OLAP) presents a powerful solution for this task. This study discusses the approach of creating an OLAP interface based on Linked Data and SPARQL and linking them to a summarized text document, which will increase the number of easy accessible information from text.

Keywords: OLAP Queries; Linked data; SPARQL; RDF data cube; Mapping

1. Introduction

According to [1] only 30% of corporate data is transactional data that can be analyzed using traditional OLAP systems. The other data are mainly non-additive data, whose processing and analysis is difficult. OLAP (Online Analytical Processing) is the basic tool for data analysis in business. One of the methods of analyzing texts of great length is summarizing documents (Text summarization). The purpose of summarizing documents is to prepare a shorter version of the source text, ensuring the meaning and main components of the original [2]. A well-prepared summary can significantly reduce the cognitive effort needed to absorb a large amount of text.

In this paper, the main objective is to find a way of presenting a summary of a text document in a way that allows multi-dimensional integration with the user. The proposed method uses external, internet knowledge database (DBpedia, WikiData) [3] to analyze the document and create semantic relationship between the concepts in document. In the first phase of discussed method, the concept are defined. In the next step, based on the internet knowledge database, the semantic depth is added to each concept. We understand the semantic depth as the possibility of referring to a given concept by an element that is its successor in a semantic graph, e.g. we can refer to a race car through a car,

* Corresponding author. *E-mail address:* prozewski@wi.zut.edu.pl

10.1016/j.procs.2019.09.393

a vehicle, or technical object. The user can create different dimensions (usually in form of hierarchies) to see the data based on the semantic depth of each concept. For example, one dimension can be related to geographic and second to technical object. Because we know the semantic depth of each concept for every dimension instance different summary will be produce in order to generate multi-dimensional text data.

2. Online Analytical Processing

In the past the business data had been used mostly for operational data processing [4]. The main application ware sales order management, invoicing, or magazine inventory. Due to changing business orientation the database systems provided various mechanisms for deducing new information from the facts contained in the database. The analytical functionality starts to be important part of database systems.

In the new approach the data should usually be presented in a summarized form, without no standard access path, with very varied methods of selection and formatting. In the result the information to be presented is dynamic. In literature we called this data model as the multidimensional data model [16].

The multidimensional data model consists of number of dimensions. The dimensions usually form hierarchies. The hierarchies enable the interactive change of detail level (granularity) of the information presented. In more complex models the hierarchies can branch. As a result, the multidimensional data model can be represent as a cube, extension of spreadsheet idea (multidimensional tables, dimensions indexed by database values). Moreover, the one can cutting and projecting on the cross-section surface (slice and dice), change of detail level: drill-down and roll-up) and turning (pivot): changes the visible dimensions on the cube.

When aggregating measures, it is important to take into account various rules of aggregation, e.g. Sales amount is usually summed. Temperature or price will rather be averaged. The analytical database stores as a rule only aggregated data. To see the detail data (drill-through) it is necessary to fetch it from data warehouse or operational database.

3. Existing Approaches to multidimensional text data

In this section, we review other strategies to presenting multi-dimensional text data [5].

- DocCube - DocCube treats several document facts as dimension. Multidimensional visualization provides the user with the opportunity to learn the connections between documents. URL links give us direct access to the exploration of the content of the text. At any time, the user can have direct access via a link to documents associated with the selected dimension values [6].
- XML-OLAP - All documents represent facts data and dimensional data. In XML-OLAP query result returns text cube. Text cube contains words, paragraphs or clusters [7].
- Document Cube - Keywords as author surname, publication date or title are uses like multidimensional data. Document cube proposes to link every document with keywords and other similar materials. This way possible to hierarchical navigate. The query result is text cube where cells consist of keywords to the relevant text [8].
- Topic Cube - OLAP must support roll-up and drill-down. The main idea is to use a hierarchical topic tree as a hierarchy for the dimension of the text. This structure allows the user to drill down along this tree and discover the content of text documents to display different levels of topics. The first level in the tree contains the details of the issues, the second level is more general types, and the last level includes the aggregation of all topics. The authors proposed a thematic scope that calculates the probability that the document contains a topic. These measures allow users to know which topic dominates in the document collection [9].
- Tube - The model adopts a cube-like concept for relational databases where cells contain keywords, and keywords are links with a text value [10].
- R-cube - Users provide a list of keywords, and then documents and facts related to the selected context are downloaded. Each paper describes the facts chosen according to their occurrence frequency [11].

Analyzing the above methods, it is easy to notice that the text is not summarized but just linked between each other. Hierarchies are created only on the basis of the ketwords appearing in the document and the frequency of occurrence of words.

4. Method for multi-dimensional text data generation

In this study, the goal is to find a way to representation text document summarization in a multidimensional manner enabling integration with the user. We use external data sources (DBpedia, WikiData) for analysis document. The concept of method is on Fig. 1. Moreover, the detailed research problems, that we are trying to solve, are below.

- Development of an algorithm of exploration and integration of summary results of text documents and the test documents themselves.
- Creation of a multi-faceted data set generated based on the proposed integration algorithms.
- Development of a visualization and search method for the created set.
- Text analysis using different families of automatic text summation algorithms.
- Automatic construction of the semantic model for each of the concepts and relations between them.
- Create PivotTables to compile and compare text.

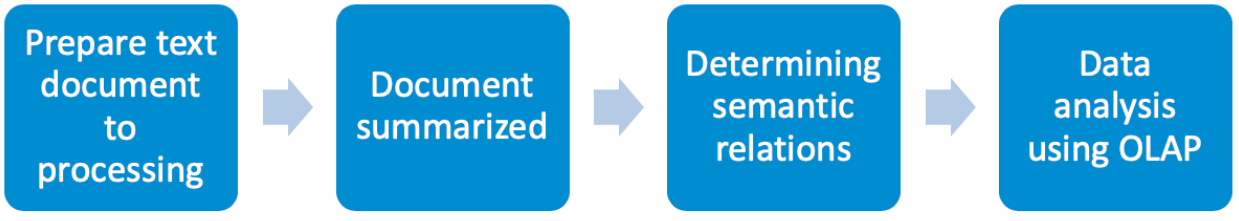

Fig. 1. Concept of the multi-dimensional text data generation method

The multi-dimensional text data generation method consists of four steps (see fig. 1). The first step is the process of preparing a text document, which, for example, includes dividing the text into sentences or words. After receiving a set of text data in the next step, we submit them to the summarization of which can be optional done either at the beginning or just before the presentation of the final results. For summarization, we use methods mainly based on statistical approach. At the next step we go to determine the semantic relationships between the words, sentences. The ItemLinks structure, from WikiData, is attached to each object. At the end of this step, the entire relational model is moved to the database, which allows you to change this relationship into the OLAP model at a later stage. ItemLinks is a reflection of the subclass relationship occurring in WikiData meaning that all instances of this item are instances of another class. The last step is to generate a multi-faceted model in the form of OLAP to analyze the obtained results.

The primary project purpose concerned uses single English document where all final data that the user gets come from this document.

4.1. Technical aspects

Technically speaking the input document was updated and divided from one single word and sentences. This operation created two lists. The first list is word list second list consists of sentences. In the list of words, a stop word operation was carried out thanks to which we got rid of inappropriate words. Then the words were tagged (specific parts of speech) and only those left from the group of nouns and adjectives were left. Finally, we check whether a given the word is in a given sentence if so, we create a list with the following structure [word, sentence] where the word is subjected to lemmatization. The above description is presented in the Fig. 2. The summarization was made using methods based on a statistical approach without training sets. For this purpose, ready-made algorithms were used: Luhn [12], LSA (Latent semantic analysis) [13], TextRunk [14] SumBasic [15].

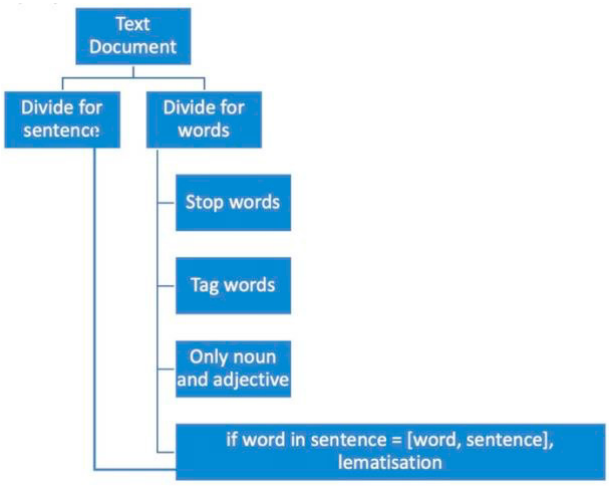

Fig. 2. Text Preparing.

Noun and adjective create relation graph for word. First, we ask DBpedia for ID. The ID links word with the WikiData. Next, we download a relation graph (Item Links for Subclass of) for word. For this moment the word is represented by the relation graph (Item Links). The example of the relation graph for concept business can be see of fig. 3.

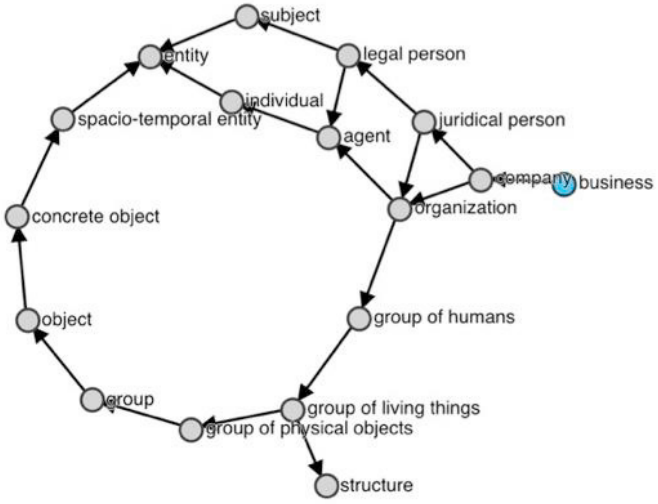

Fig. 3 . Relation Graph for concept 'business'

At the end the relation matrix among word and sentence are computed, where numbers represent word occurring based on the graphs. Results can be interpreted as OLAP model. This relation helps us create a pivoting table.

5. Case studies

5.1. Case study: Summarization before OLAP

The experiment was carried on one themed English document: "Long-term economic analysis November 2018" [10] Document consist of 90 pages. For the analysis document, we use Python and external library: NLTK, SPARQLWrapper, Mysql-Connector, Sumy, Pandas, PyPDF2 and also OWL/RDF - SPARQL EndPoint and Mysql - relation database. Concept of first experiment we can see on Fig. 4.

Fig. 4. First Experiment pipeline

The first experiment result is OLAP structure - partly show Fig. 5

Object		absence	abstract object	academic discipline	action	activity
	Word Count	1	18	1	8	5
absence	1	0	1	0	1	1
abstract object	18	1	17	1	15	7
academic discipline	1	0	1	0	1	1
action	8	1	15	1	4	4
activity	5	1	7	1	4	1
addition	1	1	1	0	1	1
additive function	1	0	1	1	2	1
additive object	1	0	1	1	2	1
agent	3	1	3	0	3	2
anatomical entity	1	0	1	0	1	0
anatomical structure	1	0	1	0	1	0
animal feed	1	1	1	0	1	1
antisymmetric tensor	1	0	1	1	2	1
applied physics	1	1	1	0	1	1
applied science	1	1	1	0	1	1
aptitude	2	0	3	0	3	3
artificial entity	13	1	20	1	16	8
artificial physical object	8	1	12	1	12	5
authority	1	0	1	0	1	1
baryonic matter	1	1	8	1	4	3
base material	1	1	8	1	4	3
behavior	3	1	2	0	1	1
binary relation	2	0	11	1	9	2
biological component	1	0	1	0	1	0
boundary	1	0	2	0	1	1
cardinal measurement scale	1	0	1	1	2	1
category of being	14	1	17	1	15	7

Object		absence	abstract object	academic discipline	action	activity
	Word Count	1	18	1	8	5
absence	1	0	4	0	5	3
abstract object	18	4	126	5	67	29
academic discipline	1	0	5	0	2	1
action	8	5	67	2	30	16
activity	5	3	29	1	16	6
addition	1	1	4	0	4	3
additive function	1	0	5	1	4	1
additive object	1	0	5	1	4	1
agent	3	1	11	0	7	4
anatomical entity	1	0	2	0	1	0
anatomical structure	1	0	2	0	1	0
animal feed	1	1	5	0	5	3
antisymmetric tensor	1	0	5	1	4	1
applied physics	1	1	4	0	4	2
applied science	1	1	4	0	5	3
aptitude	2	0	7	0	3	3
artificial entity	13	3	99	4	50	18
artificial physical object	8	3	34	1	28	13
authority	1	0	4	0	1	1
baryonic matter	1	1	28	1	11	5
base material	1	1	28	1	11	5
behavior	3	2	11	0	8	4
binary relation	2	0	26	2	15	3
biological component	1	0	2	0	1	0
boundary	1	0	4	0	1	1
cardinal measurement scale	1	0	5	1	4	1
category of being	14	4	92	3	51	25

Fig. 5. First Experiment result for unique sentences in left and repeat sentences in right.

As we can see columns and rows represents the relation graph (Item Links). Intersection represents single sum sentences (not repeat) for two Item Links data. The sentence must possess one word belonging to the Item Links from the row and column. On the right we see this same rows and columns, but now we have repetitive sentences at the intersection. At this point, it should be noted that there is a possibility of different words belonging to the same sentence, which causes repetition.

A detailed review of the sentences at the intersection of the structure shown Fig. 6. for the two selected ItemLinks: ‘Action’ and ‘Artificial psychical object’.

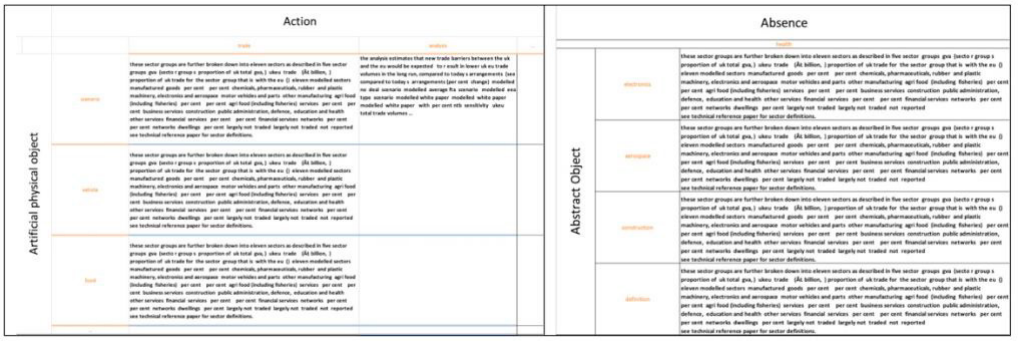

Fig. 6. First Experiment Results - Intersection detailed.

The results of the first experiment demonstrated that it is possible to achieve a complex OLAP model. At this point, we need to pay attention to the fact that we use a limited group of sentences, it is connected with summarization that is done at the beginning of the process, which accelerates the whole process of creating the OLAP model. As you can see thanks to the ItemLinks relationship with words, you can navigate the document hierarchy (roll-up and drill-down). Thanks to the relation between sentences and a word in a sentence, there is the possibility of slice and dice, thanks to which we have access to a specific dice. In connection with a structured hierarchy rigidly connected to the data, it is possible to use the pivot method without losing important information.

5.2. Case study: Summarization after OLAP

The second experiment is very similar with this difference that the summarization stage is performed just before presenting the final results to the user. Above concept, we can see Fig. 7.

Fig. 7. Second Experiment Concept

Model OLAP is show Fig. 8 for sentences not repeat in left and sentences repeat in right.

Object	Word Count	(meta)class	absence	abstract being	abstract object	academic discipline
	1	3	1	153	2	
(meta)class	1	0	0	0	1	0
absence	3	0	0	0	10	0
abstract being	1	0	0	0	1	0
abstract object	153	1	10	1	407	10
academic discipline	2	0	0	0	10	0
academic institution	1	0	0	0	1	0
act	5	0	2	0	17	0
action	28	1	3	1	329	7
activity	37	1	5	1	180	3
addition	2	0	1	0	5	0
additive function	1	0	0	0	8	3
additive object	1	0	0	0	8	3
agent	30	0	7	0	141	2
agricultural building	1	0	1	0	1	0
alphanumeric character	1	0	0	0	3	0
analysis	1	0	0	0	3	0
anatomical entity	2	0	0	0	3	0
anatomical region	1	0	0	0	1	0
anatomical structure	2	0	0	0	3	0
animal behavior	3	0	0	0	3	0
animal feed	1	0	4	0	31	1
animal stable	1	0	1	0	1	0
antisymmetric term	3	0	0	0	23	3
applied physics	1	0	1	0	2	0
applied science	3	0	1	0	8	0
aptitude	5	1	1	0	51	2
architectural structure	4	0	2	0	5	0
art	1	0	0	0	1	0

Object	Word Count	(meta)class	absence	abstract being	abstract object	academic discipline
	1	3	1	153	2	
(meta)class	1	0	0	0	3	0
absence	3	0	0	0	23	0
abstract bei	1	0	0	0	1	0
abstract obj	153	3	23	1	1844	29
academic di	2	0	0	0	29	0
academic ins	1	0	0	0	2	0
act	5	0	2	0	23	0
action	28	1	9	1	752	10
activity	37	1	8	1	405	3
addition	2	0	1	0	14	0
additive fun	1	0	0	0	19	3
additive obj	1	0	0	0	19	3
agent	30	0	8	0	326	2
agricultural	1	0	1	0	2	0
alphanumeri	1	0	0	0	4	0
analysis	1	0	0	0	5	0
anatomical e	2	0	0	0	5	0
anatomical r	1	0	0	0	1	0
anatomical s	2	0	0	0	5	0
animal beha	3	0	0	0	16	0
animal feed	1	0	4	0	74	1
animal stabl	1	0	1	0	2	0
antisymmeti	3	0	0	0	45	3
applied phys	1	0	1	0	5	0
applied scier	3	0	1	0	18	0
aptitude	5	1	1	0	106	2
architectural	4	0	2	0	13	0
art	1	0	0	0	3	0
artificial anir	1	0	1	0	2	0
artificial enti	49	3	13	1	1176	12

Fig. 8. Second Experiment result for unique sentences in left and repeat sentences in right.

A detailed review of the sentences at the intersection of the structure shown Fig.8 for the two selected ItemLinks we can see Fig. 9.

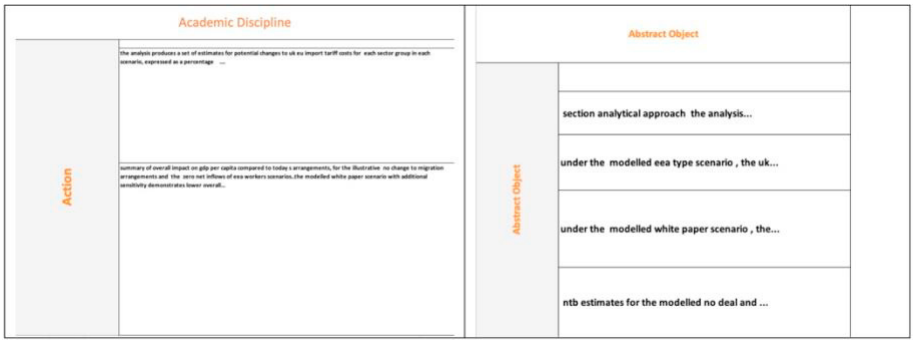

Fig. 9. Second Experiment Result - Intersection Detailed.

The results of the second experiment show that it is possible to achieve a complex OLAP model. At this point, we need to note that we use all the sentences in the text is related to the summarization of which is done at the end of the process what makes the OLAP model more accurate, but at the expense of creating it approx.100 times longer than in the first experiment.

6. Conclusion

As you can see, it is possible to present text data in the form of an OLAP model using a hierarchical tree that enables roll-up and drill-down created based on external data sources. The main problem encountered during the experiment is the long-term processing of non-summarized text, which is the main problem at the moment when the user needs processed data in a relatively short time. Other problems to be solved were, for example, the issue of not finding a subclass of relation in WikiData for a given the word. The solution adopted for this problem is to create a relationship directly with the main root "entity."

The summarization was based on statistical algorithms based on the frequency of occurrence of words in the text, which causes that the summated sentences are repeated, i.e., there is the possibility of the same sentences occurring in different methods of summarization. The use of summary before displaying the results to the user solves this problem by removing duplicate sentences and merging the summarized sentences into one text. Unfortunately, this solution is burdened with a long processing time at the initial stages of the proposed solution. It is worth paying attention to one more aspect here. A ‘subclass of’ sematic relationship was used to create a relationship graph. It was used because it best reflected the relationships built in WikiData to solve the problem. In the future, other sematic relationship will be used, such as ‘the instance of’.

Future work include the method extension about the multi-criteria selection of automatic summarizing algorithms. We can predict that the algorithms multi-criteria selection [17] will increase the accuracy of summarizing in comparison to the use of individual algorithms. In addition, the method will provide the possibility of multi-criteria parameterization of the results obtained [18].

References

[1] Ravat, F., Teste, O. (2007).” Olap aggregation function for textual data warehouse.” *International conference on enterprise information systems* 151–156.
[2] Tseng, F., Lin, W. (2006).” D-tree: A multi-dimensional indexing structure for constructing document warehouses.” *Journal of Information Science and Engineering*, 819–842.
[3] Zaveri, A., Kontokostas., D., Hellmann., S., Umbrich., J., (2017).” Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO” *Semantic Web*, **vol. 9**, no. 1 77-129
[4] Chauduri, S., Dayal, U., (1997).” An overview of data warehousing and OLAP technology.” *SIGMOD Record*, **(26)1**, 65-74

[5] Mustapha B, Youcef O, Sabine L, Yulia S. (2017).” Textual aggregation approaches in OLAP context: A survey.” *International Journal of Information Management* **37** 684–692.
[6] Ravat, F., Teste, O., Tournier, R. (2008).” An aggregation function for textual document olap.” *Data warehousing and knowledge discovery* 55–64.
[7] Park, B., Han, H., Song, I. (2005).” Xml-olap: A multidimensional analysis framework for xml warehouses in data warehousing and knowledge discovery. Berlin, Heidelberg.” *Springer* 32–42.
[8] Tseng, F., Chou, A. (2006).” The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence.” *Decision Support Systems* 727–744.
[9] Zhang, D., Zhai, C., Han, J. (2009).” Topic cube: Topic modeling for olap on multidimensional text databases.” Proceedings of the SIAM *International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA* 1124–1135.
[10] Lauw, H., Lim, E., Pang, H. (1998).” Tube (text-cube) for discovering documentary evidence of associations among entities.” *2007 ACM symposium on applied computing* 259–284.
[11] Perez, J., Berlanga, R., Aramburu, M. (2008a).” Integrating data warehouses with web data: A survey.” *Knowledge and Data Engineering* 940–955.
[10] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/760484/28_November_EU_Exit_-_Long-term_economic_analysis__1_.pdf
[12] Luhn, H. P. (1958).” The automatic creation of literature abstracts.” *IBM Journal of Research and Development* **2(2)** 159–165.
[13] Gong Y., Liu X. (2001).” Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis.” *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States*19-25
[14] ERKAN G., RADEV D.R. (2004) “LexRank: graph-based lexical centrality as salience in text summarization” *Journal of Artificial Intelligence Research* **vol. 22**, no. 1 457–479.
[15] Yih W., Goodman J., Vanderwende L., Suzuki H. (2007)” Multi-Document Summarization by Maximizing Informative Content-Words”. *In Proceedings of IJCAI-07 (The 20th International Joint Conference on Artificial Intelligence)* 1776 - 1782.
[16] Anindya Datta, Helen Thomas, (1999),The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses, *Decision Support Systems*, Volume 27, Issue 3, 1999, 289-301
[17] Wątróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., & Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107-124.
[18] Karczmarczyk, A., Jankowski, J., & Wątróbski, J. (2018). Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one, 13(12), e0209372.