MEST JOURNAL
DOI 10.12709/mest.02.02.01.10

ONE METHOD OF ANALYSIS OF RESEARCH PUBLICATIONS' ELEMENTS

Shakhovska Natalya
Lviv Polytechnic National University, Lviv, Ukraine
Noha Roman
Lviv Polytechnic National University, Lviv, Ukraine

JEL code: **C02, C6, C65, C8, C88**

Summary:
In this paper the method of research publications elements analysis that determins common qualities of research publications and their clustering as an instrument of selecting and sorting out the information about research scholars has been introduced. There is described the method, which analyses the document. The document consists of it's name, keywords, author, main part, and literature. Defining elements of the document is based on the allocation of such text features as: location in the document; location of a paragraph; type of writing; character recognition. The sum of the individual weights of words and sentences tend to be determined after further modification according to specific settings associated with each weight, gives the total weight of the sentence. The algorithm to create a database publishing features is presented. The architecture of scientific scholars forming system is described, as well as the purpose of the database relations. The database is implemented in SQL Server 2005. The business-part of the analyzer for data retrieval and for data finding is described.

Keywords:
research school, clustering, k-means, parsing, decision support system, system architecture, relational database

1. INTRODUCTION

The Research School (the Research Area) is the policy and operational unit supporting research across the whole institution. It embraces five interlinked units. The Research Grants and Contracts team provides expertise and support for the development and submission of research grants and contracts. The Innovation and Enterprise team supports the growing research and enterprise portfolio and manages intellectual property.

The Research Careers Development team is responsible for providing the framework to support researchers at all stages of their career and provides structured support in collaboration with specialist faculty and HR offerings. The

The address of the corresponding author:
Shakhovska Natalya
natalya233@gmail.com

Research Degrees team is the main point of contact for postgraduate research students and their supervisors. The Research Strategy, Information and Governance team develops, implements and monitors the institutional research strategy. It also has responsibility for research governance, policy, management and reporting of research-related management information, and development and dissemination of institutional research publications and other communications.

The remit includes management of institutional submissions to research assessment exercises (e.g. RAE 2008, REF 2014), statutory returns, coordination of research and enterprise-related consultations and management of institutional strategic research investment.

Abstracting is the process of obtaining information of primary importance from one or several sources in order to create a shorter version of it to meet the requirements of some users or tasks (Brandow, Mitze, & Rau, 1995) (Solton, 1979).

Among the segmental items of scientific publications the following ones can be defined: the author, the research institution, the subject, the keywords. It is determining of these 4 elements that gives the possibility of faster content searching as well as text and structured information integrating.

The process of abstracting is divided into three phases: analysis of the source text, identify specific fragments and the formation of appropriate conclusion. Most current work focused around technology developed referencing one document (Brandow, Mitze, & Rau, 1995).

The method involves compiling quotes emphasis on the selection of characteristic fragments (usually sentences). This method of mapping phrase patterns allocated blocks biggest lexical and statistical relevance. Creating a final document in this case is merging aggregation of selected fragments.

Most methods used linear model weights. The basis of the analytical phase of this model is the procedure for appointing weights for each block of text according to characteristics such as the location of the block in the original frequency of appearance in the text, the frequency of use of key sentences, as well as indicators of statistical significance. The sum of individual weights are usually determined after further modifications according to specific configuration parameters associated with each weight, gives the total weight of the entire block of text (Solton, 1979).

2. THE METHOD OF RESEARCH SCHOLAR CLUSTERING

Uploading the data, analysis and selection of the publications elements are the necessary steps to acquire information needed from the content for its further processing. Suppose there is a certain publication P. We will analyse the document, which consists of name T, keywords K, author A, main part M, literature L (Shakhovska & Stakhiv, 2012):

$$D = \{T, K, A, M, L\}$$

Defining elements of the document is based on the allocation of such features text:

- *location in the document;*
- *location of a paragraph (left, right, centered);*
- *type of writing (bold, italic, underline, normal);*
- *character recognition.*

Based on these characteristics formed the basis of the rules of recognition elements of the document (Table 1).

Table 1. The recognition elements of the document parsing

id	type	place	paragraph	alpha	symbols
1	title	BEGIN	{Center;Right}	{Bold}	
2	author	BEGIN	{Center; Left}		{By;©: (C) }
3	keyword	BEGIN			{Keyword;}
4	main	CENTER			
5	literature	END		{Typical, Italic}	

To form essay there is stand out from the main part of the sentence. Bulk, in turn, is divided into fragments by divisions and sections, introduced by the authors. It is believed that the sentences that appear in

the introduction and conclusion, with higher informative value than a sentence with the middle of text.

First of all, we introduce the concept of weight sentence.

The coefficient is defined as the location

$$Location = \frac{1}{n \cdot m},$$

where $n = \overline{1..3}$, $m = \overline{1..3}$ are the places calls to the main part and paragraph respectively. Begin and end of text or paragraph estimated value of 1, the middle is as 3. Coefficient key phrase is determined by entering the sentence U of elements of a set of significant sentences from A membership function:

$$Cuephrase = \mu_A(U),$$

A={"In the end", "Hovewer" …}.

Index of statistical significance is formed on the basis of visiting sentence keywords specified by the author of the article:

$$Statterm = \mu_K(U).$$

The value added is defined as the presence of terms related words sentences that appear in the article's headline to the total number of words in a sentence (words) except for words whose length is less than three characters:

$$Addterm = \frac{word}{words}.$$

So after being allowed to study all the documents necessary to accomplish the following: to exclude a statement that its content has hit the consolidated data repository and perform the final sorting sentences. For the task of bringing to the final ranking factor "information novelty" use the following method:

Let we have two sets of sentences $B = \varnothing$ and $A = \{A_i | i = 1,2,..., N\}$, N is count of sentences in text. For every sentence A_i the usefulness $P(i)_i$ set q_i: $P(i)_i = q_i, i = 1,2,..., N$.

The sentences from set *A* sort Descending $P(i)_i$.

If A_i has the biggest $P(i)_i$, we take it in B. The usefulness for sentences in A set s $P(i) = P(i)/kq_i$, where $k > 0$ – factor clipping similar sentences.

Is A empty? If NOT, go to 1.

After publication abstracting the expected material's "characteristics" have been obtained.

This characteristics have been analyzed and the necessary information has been received the research publication clustering can operate.

Clustering is the automatic partition of the elements of a certain set into groups. It can be achieved by using the k-nearest neighbor algorithm. The k-nearest neighbor algorithm consists of several steps.

Step 1 – Setting up the number of neighbors - k.

Since the features of clustering (author, research institution, subject, keywords) have not been arranged properly the d-isolated points matrix is to be applied:

$$l(X.x, Y.x) = \begin{cases} 1, X.x = Y.x \\ 0, X.x = Y.x \end{cases}$$

$$d(X, X_i) = \sum_i^p l(X.A_i, Y.A_i) + \sum_j^r l(X.D_j, Y.D_j) \sum_t^w l(X.B_t, Y.B_t) + l(X.C, Y.C)$$

where:

p – number of authors of both of the publications,

r – total number of keywords,

w – total number of scientific institutions,

$X.A_i$ - the author with number i for scientific publication X etc.

Step 2 – Determining the k - nearest neighbors for every object.

Object X_i is considered to be the nearest neighbor for X if $d(X_i, X) = \max_i d(X_i, X)$, $i = \overline{1, N}$, where N is the number of publications.

Step 3 – Object X is defined to be of the same type as most of his nearest k neighbors.

If the object is not registered in any of the clusters loose bounds between the object and the clusters are being searched for. If the value of distance between the objects X and X_i is smaller than one third of its maximum number they are loosely bound:

$$d_s(X, X_i) \le \frac{\max d(X, X_i)}{3}.$$

Let us demonstrate how scientific school is the forming.

We have publications P1 and P2. First highlight information about the author, academic institutions, keyword and subject.

We will set P1 and P2 with certain characteristics:

$$P1=\begin{cases} A=a11,a12 \\ B=b1 \\ C=c1 \\ D=d11,d12,d13 \end{cases}$$ and

$$P2=\begin{cases} A=a21,a22 \\ B=b2 \\ C=c2 \\ D=d21,d22,d23 \end{cases}$$

where a11,a12 are authors etc.

Now let us we have P3 and P4. We get similar information and obtain the following:

$$P3=\begin{cases} A=a31,a11 \\ B=b31,b1 \\ C=c3 \\ D=d31,d32,d13,d13 \end{cases}$$ and

$$P4=\begin{cases} A=a41,a22 \\ B=b41,b2 \\ C=c4 \\ D=d41,d42,d43,d22 \end{cases}.$$

We determine the number of common elements for each of the publications.

Publication P3 and P4 have some common characteristics of P1 and P2, and that's *a11* (author), *b1* (research institution) and *d13* (keywords).

We have four sets, broken down by characteristics. Now we can combine multiply P1 .. P4 with common characteristics. Since P1 and P3 and P2 and P4 are joint authors, research institutions and key words, we obtain clusters {P1, P3} and {P2, P4}:

$$P1,P3=\begin{cases} A=a11 \\ B=b1 \\ D=d13 \end{cases}$$ and

$$P2,P4=\begin{cases} A=a22 \\ B=b2 \\ D=d22 \end{cases}.$$

And now we have such schools *Sch*:

$$Sch1=\{P1,P3\}$$ and

$$Sch2=\{P2,P4\}.$$

Having a loose bound allows to use the common quality definition method in the title of the publication.

Suppose there are certain titles C1, C2, C3. For example:

C1 = ”Searching and saving information with the help of the web search engine”

C2 = ”Review and saving files in the file System“

C3 = ”Searching for information in the World Wide Web“

Let the titles be divided into two parts: right and left using the symmetric division in length. Suppose that the left part is more valuable in terms of its informative importance than the right half. The subjects are to be divided right and left and the common qualities are to be picked out. Words such as „and, so” should be ignored. Words with capital letters should not be ignored because there may be cases of it functioning as abbreviation. In addition the endings of the words should be cut-off. The result is

C1l=C3l= “searching, information“,

C1l=C2l= “saving“.

The created between the publications in which more than half of the words in the left column match is called the strong bound. Since C1l and C3l have two common names in their titles there is supposed to be strong relation between the publications P1 and P3. Therefore the titles C1l and C2l have a loose bound in their titles.

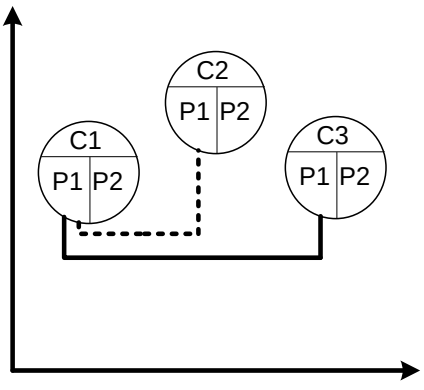

Fig. 1. Introduction of common features in the title of the publication

Such bounds between the titles can be applied for additional bound loading between the publications which in its turn may influence the process of decision making which of the existing

scientific schools the research publication refers to or if it is to be left for creating a new school.

3. APPROBATION

We have a three-tier architecture in the system analysis of publications (figure 2).

It consists of:

- lower level - working with data: processing database of scientific publications;
- level of business logic, where clustering is performed properly scientific publications and predict the development of scientific schools;
- level of presentation of data.

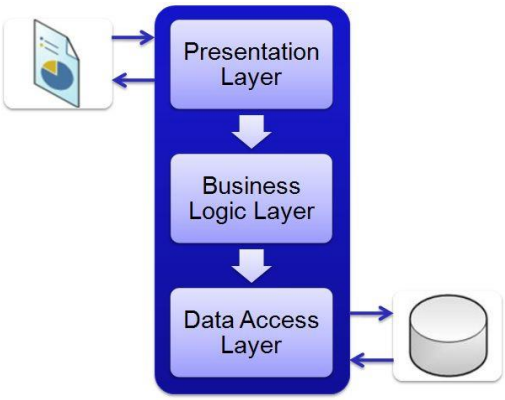

Fig.2 The system architecture.

The first stage of implementation is to design a database (DB). At the stage of system analysis, we found the essence of which exist in the subject area, and identified their attributes. All entities in varying degrees are reflected in the designed database. Let us describe the purpose of the database relations (figure 3).

1) Sentence is contains the information of all sentences in the publication. This entity has the following attributes: SentenceID, WordsWeight (Weight of words), Format (Format), Place (Place), Sum (amount);
2) Keywords (Key words) are contains all the keywords in the text. This entity has the following attributes: WordID, Word (word), Frequency (Frequency), Place (Place), Format (Format), UserWeight (weight user), Sum (amount);
3) Words-Sentence is contains the relationship between words and sentences in the text. This entity has the following attributes: ID, WordID, SentenceID.

Designed database has entity StopWords, which has a purely official character, has the following attributes: ID, Word (figure 3).

And now we can construct an algorithm of filling databases of scientific publications. Input data for the reference to the publication of scientific school is the file containing the publication. In this file we need to define the basic characteristics of publication:

- *Author (s) of publication,*
- *Scientific institutions,.*
- *Topic publication,.*
- *Abstract,.*
- *Keywords,.*
- *Text.*

Algorithm to create a database publishing features involves:

Step 1 Scientific article as a structured text information is divided into sentences and words;

Step 2 There is discarded words that contain at least three characters;

Step 3 Word classification by removing from the total list of words contained in the relation "stop words" and useless words and phrases.

Step 4 The common list of words in a documents forming. We save information about their format and location in the text;

Step 5 The total word list is modified during stemming (the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form). We discard the end of words, remove words from the same database, but increases the value corresponding to the number of occurrences of the word frequency and weight that were previously assigned these words are added. Thus we form relation "Keywords";

One can make his keywords and determine their weight, thus guiding system for the selection of the information related to the keywords.

The relation "stop words" are part of the official languages, i.e., conjunctions and pronouns words and false otherwise.

Summarization algorithm is also based on the notion of loss clause, designed to articles analysis (Radev & McKeown, 1998).

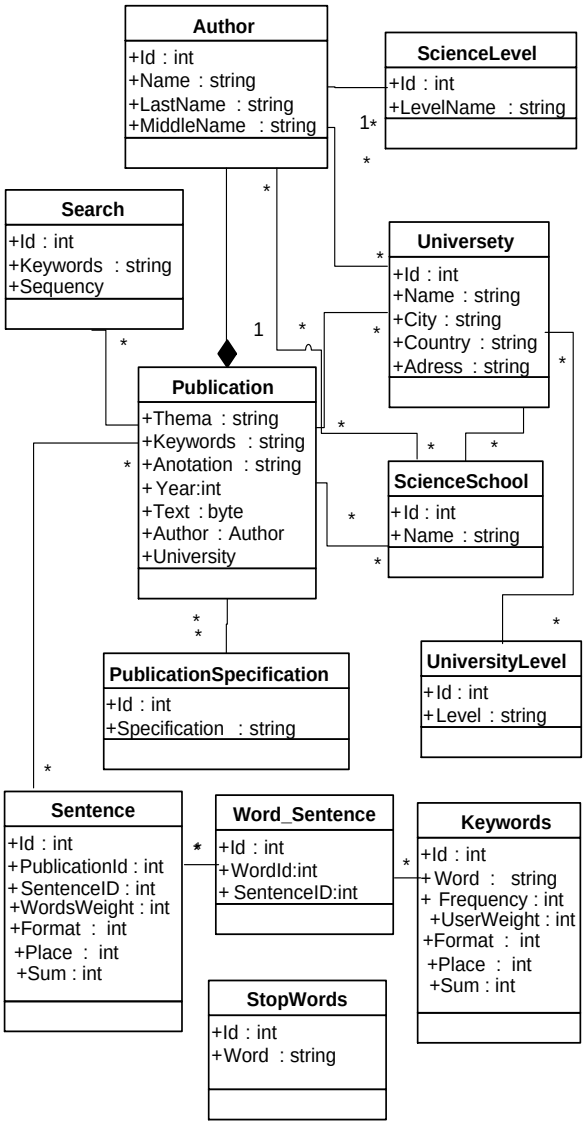

Fig. 3 The database schema

Basis analytic stage in this model is the procedure for the appointment of weights for each block of text according to characteristics such as the location of the block in the original, the frequency of appearance in the text, the frequency of key usage in sentences, as well as indicators of statistical significance. The sum of the individual weights of words and sentences tend to be determined after further modification according to specific settings associated with each weight, gives the total weight of the sentence U (Carbonell & Goldstein, 1998):

$$Weight(U) = WordWeight(U) + Place(U) + Format(U)$$

To form abstract there is out the sentence from main part.

The main part is divided into fragments by the departments and sections, introduced by the authors. Let us the sentence appearing in the introduction and conclusion are higher informative value than a sentence with the middle of the text.

The word weight Q we determine as:

$$Weight(Q) = Frequency(Q) + Place(Q) + Format(Q) + UserWeight(Q)$$

Frequency rate $Frequency(Q)$ is he ratio of occurrence of a word ($word$) to the total number of words ($words$) document. Thus, the estimated importance of words within a single document:

$$Frequency(Q) = \frac{word}{words} \cdot 100$$

The ratio $Place(Q)$ is defined as the location function of a part of the text of the article. Defining elements of the document based on selection of features of text (Ando, 2000):

- *locations in the document,*
- *locations of paragraph (left, right, centered),*
- *type of writing (bold, italic, underline, normal nakeslennya),*
- *recognition character.*

The formatting word ration $Format(Q)$ is determined depending on whether the word highlighted in bold, italic, or underlined. If the word is not formatted, the coefficient is 0 if one format is - 1 if two, then - 2 if three, then - 3.

The indicator $UseWeight(Q)$ is based on evaluation of speech user.

Record Entries parts whose weight exceeds the minimum specified, provides business-level application, shown in Figure 4.

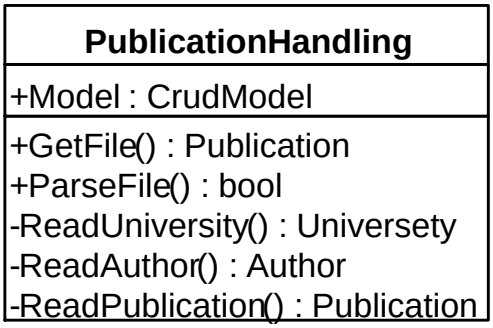

Fig. 4. Business level system to extract information from a work file

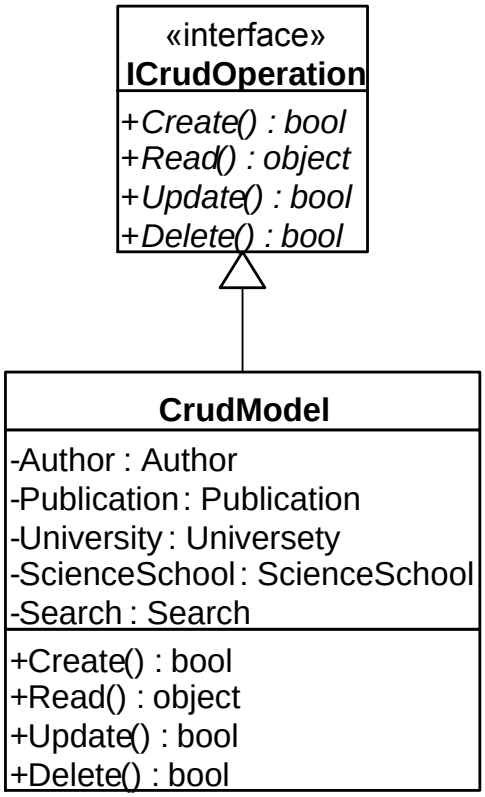

Fig. 5. Data working level

In the next step information is passed to the data level. Since this layer is responsible for working with the database, and there we write the data into the appropriate database table (Fig. 5).
Based on the data there is obtained by clustering scientific publications.

In the database that is shown in Fig. 3 there is a "Search" relation. The information in this relation is the input to predict the development of scientific schools. As mentioned above, another important criterion of scientific research development is to determine the school directly relevant specialized scientific topics and prospects of its development. With record users search for sources of information, it is advisable to develop a data analyzer determination of current topics.
The algorithm of the analyzer is not complicated. When searching for information a user enters data it is looking for, according searches the database for the presence of information sought. In the absence of data in the database, the user query written in the search table. So we will have a record that shows that the publication of this topic is not considered. Also based on the entries in this table, we can search for current ratings.
And finally, most importantly, we can analyze which schools were engaged earlier similar topics, and inform them of the new current topic. Also, the person who is searching, we inform that schools deal with similar topics and / or specific people that you can contact. Analyzer algorithm is presented in Fig. 6.

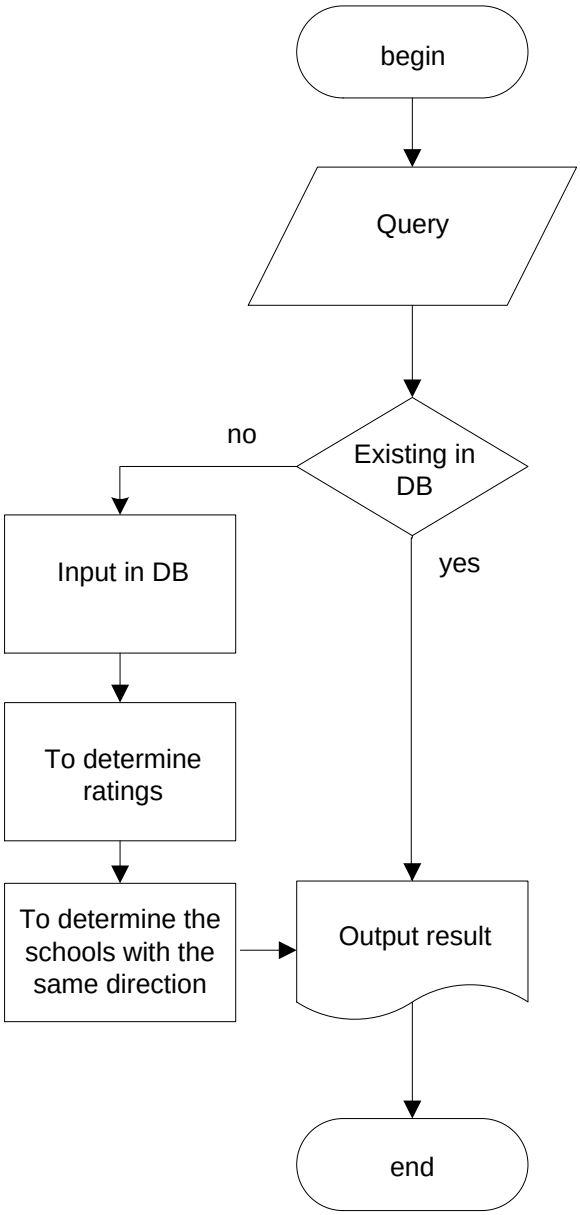

Fig. 6 Business-part of the analyzer for data retrieval

For work of this analyzer the level of the business will match and it will have the following form (Fig. 7).

The database is implemented in SQL Server 2005 database that allows you to use a large

arsenal of ready-made solutions for data analysis and texts.

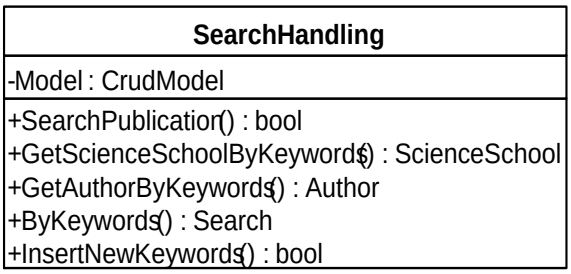

Fig. 7 Business-part of the analyzer for data finding

Structure of database is defined by standards ANSI Z39.19, ISO 2788-1986, ISO 5964-1985, APA Sixth Edition 7.25-2001, APA Sixth Edition 7.24-90. Updating a document in the system is due to its analysis. As the documents differ both in format and content, it provides the ability to replenish empowerment for disassembly.

In module structuring documents transmitted there are tape that indicates the address of the file. Depending on where the file is, it can be a path to a file on the local disk or URL on the Internet.

To take into account effects associated with differences subjective knowledge receiver and transmitter in communication processes, which are the consequences of different amounts of knowledge in software, there is used thesaurus model that relates the semantic properties of information the user with the ability to perceive information.

Updating a document in the system is due to its analysis. As the documents differ both in format and content, it provides the ability to replenish empowerment for disassembly.

In module structuring documents transmitted there are tape that indicates the address of the file. Depending on where the file is, it can be a path to a file on the local disk or URL on the Internet.

To describe the metadata there is recommended to apply a basic set of elements of Dublin Core. The concept of metadata is used in many communities, including governments, libraries, educational institutions and commercial companies. So, it will be easier to integrate the system with other. This document is of a hierarchical structure that contains metadata for all sections regardless of the depth of nesting and all resources such as images and tables.

For approbation we analyzed 208 scientific publications and have result of publication clustering:

Relation database	Count
Lviv Polytechnic National University	42
Kharkiv National University of Radioelectronics	42
Ternopil State University	27
Cloud computing	
National Technical University of Ukraine: KPI	45
National Aerospace University	32
Lviv Polytechnic National University	28

4. CONCLUSIONS

In this paper the method of common quality definition for research publications and their clustering has been introduced. Clustering is used to sort out the information about scientific schools. The method of finding the bound between the research publication and the research school it refers to has been developed. We have a three-tier architecture in the system analysis of publications. The sum of the individual weights of words and sentences tend to be determined after further modification according to specific settings associated with each weight, gives the total weight of the sentence. The algorythm of publication analyzer is built. The database for publication and its elemens storage is designed.

WORKS CITED

Ando R.K. (2000). Multidocument Summarization by Visualizing Topical Content. *Proc. ANLP/NAACL 2000 Workshop on Automatic Summarization* (79-88). Ithaca, NY: Cornell University.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31*(5), 675-685.

Carbonell J. & Goldstein J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Proc. 21st Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, (p. 335-336). NY: ACM Press

Radev D.R. & McKeown K.R. (1998). Generating Natural Language Summaries from Multiple Online Sources. *Computational Linguistics, 24(3)*, 469-500.

Shakhovska, N., & Stakhiv, Z. (2012). Automated system for essay conclusion. *Proc. of 14 International scientific conference SAIT-2012, April 24*, (428). Kyiv: Ukraine.

Solton, D. (1979). *Dinamicheskie Bibliotechno-Informatsionnye Sistemy, Per S Angl.* Mir: Publ. House M.

Received for publication: 15.10.2013
Revision received: 28.11.2013
Accepted for publication: 21.12.2013

How to cite this article?

Style – **APA** *Sixth Edition:*

Shakhovska, N., & Noha, R. (2014, 01 15). One method of analysis of research publications' elements. (Z. Čekerevac, Ed.) *MEST Journal, 2*(1), 94-102. doi:10.12709/mest.02.02.01.10

Style – **Chicago** *Fifteenth Edition:*

Shakhovska, Natalya, and Roman Noha. "One method of analysis of research publications' elements." Edited by Zoran Čekerevac. *MEST Journal* (MESTE) 2, no. 1 (01 2014): 94-102.

Style – **GOST** *Name Sort:*

Shakhovska Natalya and Noha Roman One method of analysis of research publications' elements [Journal] = One method of analysis of research publications' elements // MEST Journal / ed. Čekerevac Zoran. - Belgrade : MESTE, 01 15, 2014. - 1 : Vol. 2. - pp. 94-102. - ISSN 2334-7058 (Online); ISSN 2334-7171.

Style – **Harvard** *Anglia:*

Shakhovska, N. & Noha, R., 2014. One method of analysis of research publications' elements. *MEST Journal,* 15 01, 2(1), pp. 94-102.

Style – **ISO 690** *Numerical Reference:*

One method of analysis of research publications' elements. **Shakhovska, Natalya and Noha, Roman.** [ed.] Zoran Čekerevac. 1, Belgrade : MESTE, 01 15, 2014, MEST Journal, Vol. 2, pp. 94-102. ISSN 2334-7058 (Online); ISSN 2334-7171.