research school, clustering, text mining

Natalya SHAKHOVSKA[*]*, Roman NOHA*[**]

THE SYSTEM DEVELOPING OF FORMING RESEARCH SCHOOLS BASIS OF PUBLICATION ELEMENTS ANALYSIS

Abstract

In this paper the method of research publications elements analysis that is determining common qualities of research publications and their clustering as an instrument of selecting and sorting out the information about research schools has been introduced. 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.

1. INTRODUCTION

Scientific direction is an area of scientific research team to address some significant fundamental problems. School Science is a Research team, which aims at addressing scientific field.

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 [1,2].

Among the segmental items of scientific publications the following ones can be defined: the author, the research institution, the subject, and 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 [1].

[*] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, natalya233@gmail.com

[**] Lviv Polytechnic National University, Ukraine, 79013, Lviv, Bandera str., 28a, rnoha@gmail.com

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 [2,5,6].

2. THE PROBLEM FORMULATION

Among the available methods for the analysis of scientific publications is to develop ontologies. The important role in the analysis of the set of text documents and Internet there is playing World Wide Web. The ontologies usage for searching allows the user to formulate a query at a higher level of abstraction than is possible when searching for keywords.

Let us consider the examples of systems using ontologies for the Internet [4]. Observer (http://siul02.si.ehu.es/~jirgbdat/OBSERVER). This system provides an approach using the many existing ontologies to access heterogeneous, distributed and independently of developed repositories of data. The implement-tation of this approach is the ideology broker ontology domains. It is assumed that there is a lot of pre-existing ontologies domains; user optionally can build new terms for a specific ontology. User formulates a query certain language in terms of one or more ontologies and broker "seeking" relevant information resources, performing broadcast request in appropriate ontologies and, if necessary, and a combination of multiple ontologies for a precise answer to the request [7,9].

OntoSeek is designed to retrieve information from the context of online "yellow pages" directories and products. The system can handle both homo-geneous and heterogeneous catalogs of products. For exact fixing of context this system can be applied to interactive approach where user progressively refines the meaning of key words with linguistic database.OntoSeek a centralized server that hosts the database of lexical conceptual graphs known system resources, but create the graphs client.

The approach used in OntoSeek is different from the approach used in the model W3C Resource Description Framework (W3C RDF, Netlogon! Unacceptable Object hyperlinks. In the RDF description of the data structure (ie, schema data as <subject, predicate, object>) object is added in HTML/XML document and it is not stored separately. There are no additional conditions on the semantic consistency RDF requires.

WordNet is a linguistic database consisting of sensitive groups of words equivalent in meaning. WordNet is both lexical dictionary (created for several European languages), and ontology that reflects relationships between words in the dictionary. Description resource is implemented as a lexical conceptual graph, where peaks correspond to words, and referred to the arc present the semantic relations between words (eg, relations of "part" or "subclass", etc.). Names of vertices and arcs are also taken from WordNet when you create a conceptual graph of a particular resource. Finding resources is relevant to a user query. This process based on a comparison of ontologies (lexical conceptual graphs) of these resources. For resources selection that relevant to a user query WordNet performs comparison of conceptual graph queries from existing resources conceptual graphs or parts of graphs.

Ontological approach is famous to search for papers specific scientific school. However there is small number of Ukrainian language ontologies. Therefore, we can use the ontological search on the next stages constructing a system analysis of the scientific schools.

As you can see, nearly all of these areas there are scientific investigations. But these non-integrated work, do not provide a single processing and rigidly attached to the data model that is completely unacceptable in the context of data space. Therefore, the problem of formalization of data space is important.

3. MAIN PART

3.1. The method of research school 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* [3]:

$$D = \{T, K, A, M, L\} \tag{1}$$

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.

For the task of bringing to the final ranking factor "information novelty" use the following method:

1. 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 $q_i : P(i)_i = q_i, i = 1,2,..., N$.
2. The sentences from set A sort Descending $P(i)_i$.
3. 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 $k > 0$ is a factor clipping similar sentences.
4. Is A empty? If NOT, go to 1.

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

1. Setting up the number of neighbors – k [7,8,10,11].
   Since the features of clustering (author, research institution, subject, keywords) have not been arranged properly the d-isolated points matrix is to be applied:

$$l(X.x, Y.x) = \begin{cases} 1, X.x = Y.x \\ 0, X.x = Y.x \end{cases} \quad (2)$$

$$d(X, X_i) = \sum_i^p l(X.A_i, Y.A_i) + \sum_j^r l(X.D_j, Y.D_j) \sum_t^w l(X.B_t, Y.B_t) + l(X.C, Y.C) \quad (3)$$

   where: p – number of authors of both of the publications,
   r – total number of keywords, w is a total number of scientific institutions,
   $X.A_i$ – the author with number i for scientific publication X etc.

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.
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:

$$d_s(X, X_i) \le \frac{\max d(X, X_i)}{3} \tag{4}$$

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 serach 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 2 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 2 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.

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.

For estimating of classificator k-*NN* quality there is used indicator of the probability of correct classification: *TP* (true positive) – number of documents with true classified, *FP* (false positive) – error of 2 range (count of documents, which are incorrect been in cluster); *FN* (false negative) – error of 1 range (count of documents, which must be in cluster, but they are not here; *TN* (true negative) – count of documents, which not belong to the cluster. Values *TP* and *TN* evaluate by formula:

$$TP = Np - FN$$
$$TN = Nn - FP \tag{5}$$

where: Np – count of documents, which clustered without error and Nn is count of documents which clustered with error.

Next we norm values TN , TN , FN , FP :

$$\text{nFN} = \frac{\text{FN}}{\text{Np}} * 100\%; nFP = \frac{FP}{Nn} * 100\%,$$
$$nTN = \frac{TN}{Nn} * 100\%; nTP = \frac{TP}{Np} * 100\%. \tag{6}$$

Also we estimate the quality of document clustering. We use such metrics:

- Recalc: $\frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}$,
- Precision: $\frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}$,
- Fall-out: $\frac{|D_{nrel} \cap D_{retr}|}{|D_{nrel}|}$,

where: D_{rel} – the set of relevant documents in cluster,

D_{nrel} – the set of not relevant documents in cluster (error in clusterization),

D_{retr} – the set of relevant documents in another clusters.

3.2. APPROBATION

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:

1) Sentence 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) 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 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 (fig. 1).

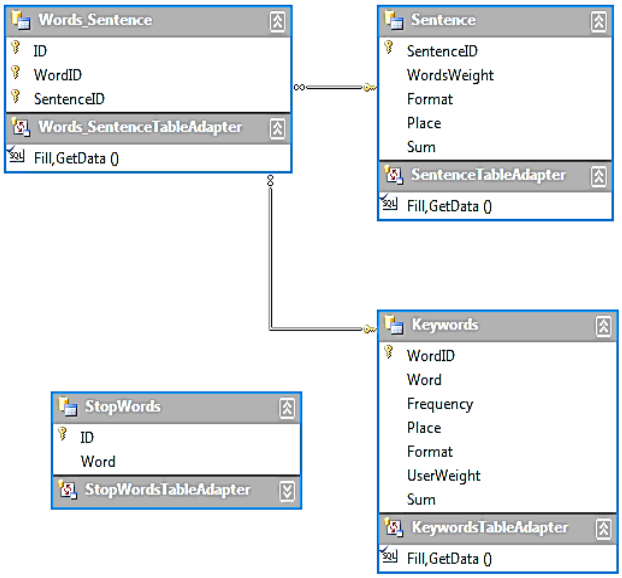

Fig. 1 The database schema [source: own study]

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.

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 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 (table 1).

Tab. 1. Result of publication clustering

Relation database	**Count**
Lviv Politechnic National University	42
Kharkiv national university of radioelectronics	42
Ternopil State University	27
Cloud computing	**Count**
National Technical University of Ukraine :KPI	45
National Aerospace University	32
Lviv Politechnic National University	28

The interface of developed system is given belong.

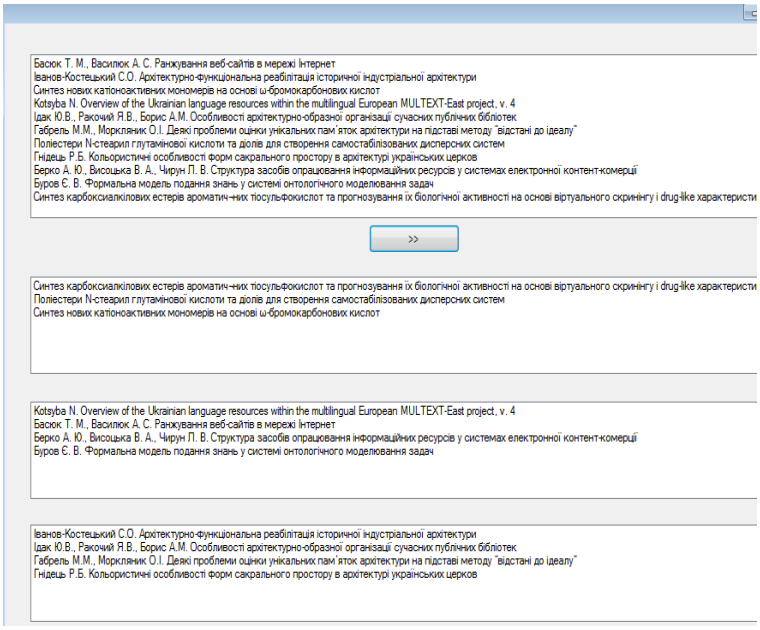

Fig. 2. Result of clustering [source: own study]

There is presented the result of scientific school clustering. In the first window the list of publications is given. After that we can see 3 schools consists of publications with similar keyword.

4. CONCLUSIONS

In this paper the method of research publications elements analysis that is determining common qualities of research publications and their clustering as an instrument of selecting and sorting out the information about research schools has been introduced. Depending on where the file is, it can be a path to a file on the local disk or URL on the Internet. There is described the method, which analyses the document. 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.

The document consists of name, keywords, author, main part, literature. Defining elements of the document is based on the allocation of such features text: 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 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.

The algorithm to create a database publishing features is introduced. The architecture of scientific school forming system is described. In this paper there is described 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.

Our experimental results demonstrate the efficiency of the proposed algorithm.

REFERENCES

[1] BRANDOW R., MITZE K., AND RAU L. F.: *Automatic condensation of electronic publications by sentence selection.* Information Processing and Management, 31 (5), 1995, pp. 675-685.

[2] SOLTON J.: *Dynamic library – information systems.* M: the World, 1979.

[3] SHAKHOVSKA N., NOHA R.: *One method of analysis of research publications' elements*. MEST Journal, 15 01, 2(1), 2014, pp. 94-102.

[4] SALTON G. et al.: *Automatic Text Structuring and Summarization.* Information Processing & Management, vol. 33, no. 2, 1997, pp.193-207.

[5] RADEV D. R., MCKEOWN K. R.: *Generating Natural Language Summaries from Multiple Online Sources.* Computational Linguistics, vol. 24, no. 3, 1998, pp. 469-500.

[6] CARBONELL J.G., GOLDSTEIN J. G.: *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, ACM Press, New York, 1998, pp. 335-336.

[7] SHI ZHONG: *Efficient Online Spherical K-means Clustering*. Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN 2005), Montreal, Canada, 2005, pp. 3180-3185.

[8] HOU, JUN, NAYAK, RICHI: *The heterogeneous cluster ensemble method using hubness for clustering text documents.* Lecture Notes in Computer Science [Web Information Systems Engineering - WISE 2013: 14th International Conference, Nanjing, China, Proceedings, Part I], 2013, pp. 102-110.
[9] STREHL A., GHOSH J.: *Cluster ensembles – a knowledge reuse framework for combining partitions.* Journal of Machine Learning Research, no. 3, 2002, pp. 583-617.
[10] STREHL A., GHOSH J., MOONE R. J.: *Impact of similarity measures on web-page clustering*. AAAI Workshop on AI for Web Search, 2002, pp. 58-64.
[11] KARYPIS G.: *CLUTO - a clustering toolkit*. Dept. of Computer Science, University of Minnesota, 2002. (http://www-users.cs.umn.edu/karypis/cluto)

.