Journal of Intelligent & Fuzzy Systems 39 (2020) 8981–8988
DOI:10.3233/JIFS-189296
IOS Press

Translation of news reports related to COVID-19 of Japanese Linguistics based on page link mining

Liu, Xiaohua*
School of Foreign Studies, Henan University of Urban Construction, Pingdingshan City, Henan Province, China

Abstract. In the face of the current epidemic situation, news reports are facing the problem of higher accuracy. The speed and accuracy of public emergency news depends on the accuracy of web page links and tags clustering. An improved web page clustering method based on the combination of topic clustering and structure clustering is proposed in this paper. The algorithm takes the result of web page structure clustering as the weight factor. Combined with the web content clustering by K-means algorithm, the basic content that meets the conditions is selected. Through the improved translator of clustering algorithm, it is translated into Chinese and compared with the target content to analyze the similarity. It realized the translation aim of new crown virus epidemic related news report of Japanese Linguistics based on page link mining.

Keywords: Web community, link structure, page link mining, minimum description length

1. Introduction

In the 21st century, the spread speed of news information plays a decisive role, especially transnational news, which involves differences in language, culture and values [1, 2]. In the recent investigation of epidemic news, it was found that the lack of subject in translation resulted from different language habits, which seriously affected the spread of epidemic information [3–5]. It leads to many misunderstandings and mistakes in epidemic prevention [6].

With the rapid popularization and development of Internet/Web technology, Web data has become the largest "data warehouse" in the world. How to find knowledge from massive Web data and benefit human beings is the mission of Web data mining technology [7]. However, Web data is heterogeneous, unstructured and dynamic, which requires us to first classify Web pages (clustering), then design different wrappers for different classifications [8, 9], extract information, and finally mine the structured data. The method discussed in this paper is an improved Web page clustering method based on the previous work of Web Clustering. It obtains the site classification model by analyzing some (but not all) representative pages in the Web site, so as to provide training samples for the design of packager and ensure the smooth progress of the whole data mining process.

In the past, page classification was carried out manually, which was very tedious and complicated. The development of machine translation system should fully consider the expression characteristics of source language and target language, only in this way can we enhance the pertinence and improve the quality and efficiency of translation [10–14]. It is based on this principle that we design and develop Japanese Chinese machine translation system. Japanese is a kind of adhesive language, which mainly depends on the

*Corresponding author. Liu, Xiaohua, School of Foreign Studies, Henan University of Urban Construction, No. 1 Longxiang Street, Pingdingshan City, Henan Province, China. E-mail: xiaohua_6901@126.com.

appurtenances attached to the back of the independent words to determine the position and grammatical function of the independent words in the sentence, as well as the structure and meaning of the sentence, so the focus of Japanese analysis should be on the study of the usage of the appurtenances. If you know a Japanese sentence, you can understand the whole sentence if you can determine the meaning of each auxiliary word. Because most of the adjuncts have multiple senses, the selection of the senses has become the primary and core problem in the analysis process. The design given in this paper solves this problem well.

Due to the high accuracy and fast speed of epidemic news, how to translate the text quickly is the focus of research. In addition, the conflict of values and the acceptance of culture should be resolved [15–17]. Therefore, the speed and accuracy of epidemic news depends on the clustering of Web page links and tags.

At present, the research in this field mainly includes Web page topic clustering based on text content (the main clustering algorithm is k-means algorithm, etc.) and clustering based on Web page structure proposed by some scholars. The former only considers the content information of Web pages, and the clustering time efficiency is low. The latter cleverly uses the organization structure of Web pages, but does not use the content information provided by Web pages. If the two can be combined, the clustering quality will be improved.

2. Link analysis and its Web graph representation

2.1. Link analysis and its Web graph representation

In link analysis, page Web is often regarded as a node, and links between pages are regarded as edges, so that the whole world wide Web can be regarded as a huge digraph, i.e. $G = (V, E)$, which can be defined as follows:

(1) V: The node set composed of Web pages, p, q $\in$ v, p $\neq$ q;
(2) E: The directed edge set composed of links between Web pages: $p \to q \in E$;
(3) $p \to q$: p node has a link to q, where p is the link in Web page of q, which is called chain source, q is the link out Web page of p, which is called chain destination;
(4) Link: the link of other nodes to p;
(5) Out of chain: P links to other nodes;
(6) Node out degree: the number of nodes out of chain;
(7) Node access: the number of nodes in the chain

It is also very common for Web graphs to be represented by adjacency matrix. For graph $G = (V, E)$, construct a matrix $A_{ij} = (a_{ij})_{n\times n}$, among which: $A_{ij} = \begin{cases} 1 & i \to j \in E \\ 0 & other \end{cases}$ others, call matrix A adjacency matrix of graph G.

The significance of link p $\to$ q can be considered as follows: page P tells users who have visited page p that they can visit page q along the link created by page p. In this way, the link between the two Web pages can show that p and q have related topics of interest. p $\to$ q indicates that p has a subjective value judgment on q. it can be said that the link is that the Web page p thinks that the Web page q contains valuable information.

2.2. Mining algorithm based on web link structure

Page Rank algorithm: the Web is regarded as a digraph, Ni is the output of page i. B_i is the collection of all pages pointing to page i. then, the PageRank value PR(i) of page i can be calculated in the following two steps:

Step 1: randomly take any page i of the Web with probability (1-d);

Step 2: randomly take the page j pointing to the current page i with probability d, and the specific iteration formula of Page Rank algorithm is as follows [17]:

$$PR(\mathrm{i}) = (1 - \mathrm{d}) + \mathrm{d}\sum_{\mathrm{j}\in \mathrm{B}} \frac{PR(j)}{|N_j|} \quad (1)$$

Where parameter d is the damping coefficient, between 0 and 1, d is usually set to 0.85

Page Rank algorithm is one of the earliest and most successful algorithms in web link structure analysis, which applies link analysis technology to commercial search engines. It is an important tool to evaluate the authority of web pages.

However, the calculation of value in PageRank algorithm is not for query. For a query on a specific topic, there are often some "robust" pages that have nothing to do with the topic in the returned results, but the most important pages are not in the

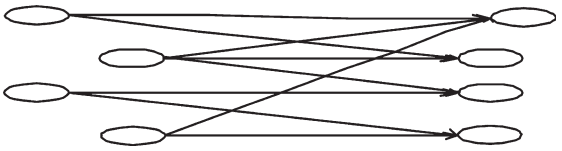

Fig. 1. Relationship between Hubs and Authorities.

result set. Then some scholars of Washington University put forward PageRank algorithm combining link and content information. In addition, we consider the situation that users jump directly from one web page to another page that is not directly adjacent but related to the content. In addition, the Department of computer science of Stanford University has improved the PageRank algorithm, transforming it into a topic related algorithm. The experiment shows that the effect of the PageRank algorithm before the transformation is much better than that before the transformation.

The algorithm based on CO reference and co coupling is based on two web page types (authority page and hub page) and their relationship. Authority page refers to the page recognized as authoritative in a subject, and center page refers to the page with many links to authority page. The center page and authority page form a mutually reinforcing relationship ¡£ A good central page points to many good authoritative pages, and a good authoritative page is pointed to by many good central pages. This relationship describes a web page as a diagram, as shown in Fig. 1:

HITS algorithm mines the hyperlink structure of the web, including authorities and hubs. For these two types of web page extraction, the following two operations can be performed by cycling:

I operation (calculate authority weight):

$$x_p = \sum_{q,q\to p} y_p \tag{2}$$

O operation (calculate hub weight):

$$y_p = \sum_{q,q\to p} x_q \tag{3}$$

The basic operation of weight calculation is shown in Fig. 2:

According to the description of the above web page extraction process, the whole execution steps of HITS algorithm can be summarized as shown in Fig. 3:

Specific steps:

Step 1: submit the query to the traditional search engine, select a certain number of top R pages from

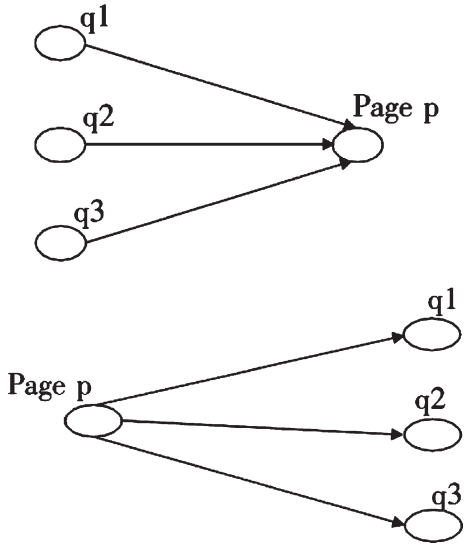

Fig. 2. Setting for document template.

the pages returned by the search engine to form the root set R;

Step 2: expand the size of the root set to n to form the basic set B. the extension rule is: add all the Web pages linked by the Web pages in the root set, and add up to d Web pages linked to the web pages in the root set R;

Step 3: G (B) is used to represent the subgraph derived from the link relationship of web pages in basic set B

Step 4: construct n*n adjacency matrix and its transposition matrix A^T according to G(B), calculate the maximum eigenvalue λ_1 of A^TA, and normalize the main eigenvector e_1 of the maximum eigenvalue corresponding to λ_1;

Step 5: return the element with larger absolute value in the normalized eigenvector e_1 as the authors;

Step 6: calculate the maximum eigenvalue λ_1' of AA^T, and normalize the main eigenvector e_1' of the maximum eigenvalue corresponding to λ_1';

Step 7: return the element with larger absolute value in the normalized eigenvector e_1' as a hub

In HITS algorithm, the authority value of each page converges to the main eigenvector of A^TA, and the hub value converges to the main eigenvector of AA^T. However, when querying a lot of generalized topics, hits algorithm mistakenly assigns many pages that are not related to the topic with high value, resulting in the phenomenon of topic drift. Therefore, after HITS algorithm, many researchers put forward many improved hits algorithms.

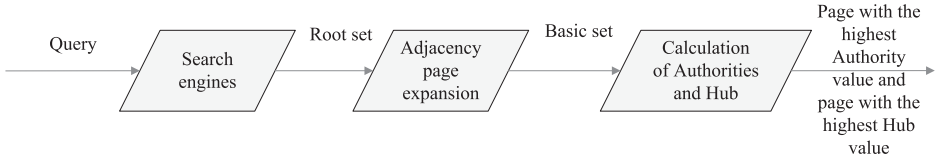

Fig. 3. HITS algorithm.

The engineering group of Bmalmaden research center put forward arc (automatic resource compilation) algorithm, with hits as the core, and tried to solve the problem of topic drift by increasing the use of web content information. ARC algorithm uses the relevance between the text around the link and the retrieval topic to distinguish the importance of the link. If there is a description of the topic in the front and back text of a link, it is reasonable to think that the current page and the target page are related to the topic, then increase the weight of the link accordingly.

It can be seen that arc algorithm uses link text and surrounding text to calculate link weight, and introduces text semantic information of page into HITS algorithm, which has achieved certain effect. Because in some cases, the text around the link is a simple description and evaluation of the content of the link destination page, then using the word frequency information in the text around the link can improve the accuracy of the algorithm. However, the form of the web page is very complex. In many cases, the text around the link cannot represent the content of the chain destination page, even far from the content of the chain destination page.

Some scholars have proposed the trawling algorithm, which is based on the bipartite graph relationship between the central page and the authoritative page on the web page set. A bipartite digraph is such a graph: the node set of graph k_{ij} can be divided into two sets. It is represented by F (fan) and C (Center). There are i nodes in set F, j nodes in set C, and there is a directed edge from each node in set F to each node in set C.

The data source of the trawl algorithm is not based on a certain topic, but on the general crawling results. All potential Fan sets are found by scanning the data set, and the Center set is also determined. Then, all the cores are obtained by repeated include / exclude pruning, and then cluster to a smaller set of cores by association rule mining algorithm. Finally, each core is a group of pages about the theme.

The trawl algorithm is based on the whole web crawling result, so the web result web page is objective and has nothing to do with the topic. However, this kind of complete dichotomous digraph can only identify the very dense core of the graph, and some cores will be lost. To solve this problem, many researchers have proposed the dense dichotomous algorithm.

According to graph theory, some scholars put forward a community discovery method based on traffic algorithm. It defines a community as a collection of pages with such characteristics in a Web diagram. The link density between pages in the community is higher than that between pages in the community. Suppose that digraph $G = (V, E)$, (m, n) is the directed edge from node m to n. Suppose s, t is a fixed point in V, and each edge (m, n) has a allocated capacity $c(m, n) \in Z^+$, the flow from s to t is a non negative integer function: $0 \leqslant f(m, n) \leqslant c(m, n)$, the flow out of s is equal to the flow into t. These scholars have proved that using the maximum flow algorithm can separate a dense network sub-graph and get a network community. That is to say, the problem of community identification is equivalent to the problem of $s \rightarrow t$ maximum traffic / minimum cut.

The definition and execution of the algorithm are clear. However, it assigns constant values to the edge capacity, which results in the extraction of the image structure of some noisy pages.

3. Introduction to relevant technologies and knowledge

3.1. Features of available web pages

Feature 1: the link structure in web page can be used as the feature item of web page. We can preprocess the data of web page to generate XML file, then generate DOM tree, extract the set of web page paths containing link labels, and take “the set of web page link label paths” (hereinafter referred to as “link path") as the feature item of web page, as the main basis to judge the similarity and clustering of web page; Fig. 4 is a DOM tree structure generated by a web page. The link path set of the page is html-table-tr-td, html-table-td-tr-td.

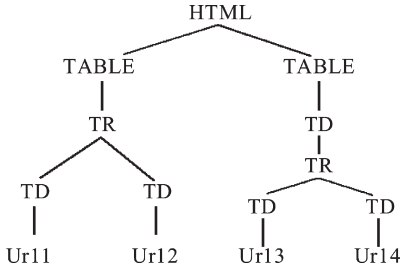

Fig. 4. Structure of net page.

Feature 2: links under the same link path in web pages may point to pages of the same category

As shown in the DOM tree in Fig. 4, there are two links url1, URL2 under the link path html-table-tr-td in the web page. We think that these two links are likely to point to the same category of web pages, because the format and layout of well formed web pages are very regular. The distribution is shown in Fig. 5.

Figure 5 is from http://www.chinadaily.com.cn/. Regions 1, 2, and 3 in the figure represent three different link paths. The link path of hyperlinks in each area is the same, and points to the same page, or even the same page, but only passes different parameters. Therefore, we can compare the web pages pointed by links in the same region before clustering, because they have high similarity, so we can improve the efficiency of clustering.

Feature 3: some labels of web pages can provide page content information. The title tag tells us the title of the page.

The meta tag tells us the description and keywords of the page.

Combined with these content information, we can take both web page structure and page content into account when clustering, so that the clustering results are more accurate.

The theory of minimum description length (MDL) in information theory is described as follows: Model = min(DesLen(Model) + Deslen(data\Model)

Deslen (model) is the description length of the model, and deslen (data|model) is the description length of the data under the given model. The principle of MDL theory is that the sum of the description length of the model and the description length of the data under the given model is required to be the minimum. In this paper, MDL theory is used to constrain the clustering process to ensure the clustering quality.

3.2. *Experimental environment and result analysis*

CWPBLT algorithm implementation environment

Operating system Windows2000, language Java, development platform eclipse + Tom - cat + JDK + sqlserver.

Fig. 5. The distribution of link paths.

The Java packages used are as follows:

Jtidy: during the experiment, we use jtidy package to obtain web page from the specified web address and convert the web page to XML format, which is stored locally;

JDOM: in the experiment, we use JDOM package to obtain the link path set in XML file, and use Java multithreading technology to compare and cluster web pages.

Four large websites or their subsites are selected as the experimental data sources in this paper. They have regular structure, rich links and representativeness.

We use F-measure to evaluate the experimental results:

$$F_{\mathrm{i}} = \max_{j=1,\cdots,m} \left(\frac{2p(C_i, \hat{C}_j)R(C_i, \hat{C}_j)}{P(C_i, \hat{C}_j) + R(C_i, \hat{C}_j)} \right) \quad (2)$$

P (precision) is the accuracy rate, R (recall) is the recall rate, and the formula is:

$$P(C_{\mathrm{i}}, \hat{C}_j) = \frac{\left| C_{\mathrm{i}} \cap \hat{C}_j \right|}{\left| \hat{C}_j \right|} \quad (2)$$

$$R(C_{\mathrm{i}}, \hat{C}_j) = \frac{\left| C_{\mathrm{i}} \cap \hat{C}_j \right|}{|C_{\mathrm{i}}|} \quad (3)$$

P (precision) is the accuracy rate, R (recall) is the recall rate, and the formula is:

$C_1, C_2, C_{\mathrm{i}}, \cdots C_{\mathrm{n}}$ indicates the correct category, $\hat{C}_1, \hat{C}_2, \hat{C}_{\mathrm{j}}, \cdots, \hat{C}_{\mathrm{m}}$ refers to the category of experimental results, $\left| C_{\mathrm{i}} \cap \hat{C}_{\mathrm{j}} \right|$ refers to the number of C_{i} pages appearing in $\hat{C}_{\mathrm{j}}$; we use the following formula to evaluate the system based on all categories.

$$F = \sum_{\mathrm{i}=1}^{\mathrm{n}} F_{\mathrm{i}} \cdot \frac{|C_{\mathrm{i}}|}{\sum_{\mathrm{p}=1}^{\mathrm{n}} |C_{\mathrm{p}}|} \quad (4)$$

After F-measure evaluation, we get: (1) the system; (2) the comparison results based on Web page topic content clustering (often using k-means algorithm) are shown in Table 1.

The corresponding line chart is shown in Fig. 6. From the experimental control data, the combination of Web page link structure and tag content clustering method has a good effect on improving the performance of page clustering, especially in the precision;

Table 1
The evaluation results

NAME	(1) F Value	(2) F Value	(1) Accuracy	(2) Accuracy
SOHU	0.87	0.82	93%	87%
BAIDU	0.92	0.86	97%	89%
163	0.88	0.77	89%	81%
SOWANG	0.69	0.68	71%	71%

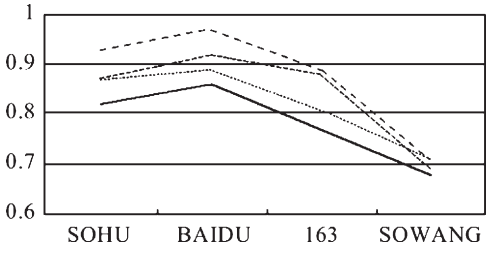

Fig. 6. The corresponding line chart.

at the same time, compared with the structure of the page in the site, the site with relatively simple page structure has higher F value, higher accuracy and better clustering results. For example: the site page structure of BAIDU is much simpler than that of SO-WANG (a Chinese meta search engine, in which links of multiple sites are integrated), and the corresponding F value of the former is much higher. In addition, compared with the previous clustering algorithm based only on Web page structure, although the latter may obtain a higher F value, it only gathers the web pages with similar structure into one group. Because it can't take into account the content of the page, such results often deviate from the actual situation and are difficult to be applied in practice. Therefore, CWPBLT is an improvement on the previous web clustering method, and also a more practical clustering method.

4. Comparison of Japanese and English sentences

The following is a typical Japanese sentence. By comparing the original text with the translation, we can find the similarities and differences in the structure of Japanese sentences and English sentences.

Example sentence:
私は /毎朝 /7時半ごろ /食堂で /朝食を/食べます /
私は /→subject
毎朝 /7時半ごろ /食堂で /→complement
朝食を/→object

食べます /→predicate

/⟶symbol

Translation:

I/have/breakfast /in the canteen /around 7 : 30 /every morning

I/⟶subject

have/⟶predicate

breakfast /⟶object

in the canteen /around 7 : 30 /every morning⟶complement

Similarities:

Japanese sentences and English sentences can be regarded as chunks, each of which is an independent syntactic component. As in example, those separated by '/'. Moreover, there is a one-to-one correspondence between the original and the translated fragments. Therefore, for the translation from Japanese to English, we can adopt the method of "analyzing a single Japanese segment to generate its corresponding English segment, and finally connecting each English segment", that is, based on the level of "segment", we can adopt the transformation generation method to carry out Japanese English machine translation. The third analysis part of our system adopts this analysis strategy.

Difference:

In Japanese sentences, the syntactic components of each segment are determined by the appendages at its tail (for example, the case particle "" indicates that the segment to which it belongs is the object component). In general, English sentences are determined by the sequence of the segments (in the example above, "breakfast" is placed after the predicate verb "eat" to indicate that it is the object), and sometimes it needs the auxiliary function of some function words (for example, "in the canteen" in the example above indicates that the segment is the adverbial component). Therefore, the order of the segments in Japanese sentences is relatively free, except for the principle that the predicate is not in the sentence, which does not change the meaning of the sentence due to the reversal of its order; the order of the components in English sentences cannot be reversed at will. Therefore, in the process of generating English fragments, we need to adjust the order of all fragments according to what they are.

Based on the analysis, it can be seen that the following two problems should be solved in the design of Japanese adjunct Dictionary: (1) the polysemy selection of Japanese adjunct corresponding to English translation; (2) the order of each segment in English translation.

The following briefly describes the whole process of designing the framework of machine dictionary and the final results.

In general, Japanese case auxiliary words correspond to English translation words. For example:

1. 彼は/食堂で/ギョーザを/食べます/

Translation:

He eats dumplings in the canteen

2. 彼は/飛行機で/東京に/行きます/。

Translation:

He flew to Tokyo.

3. 彼は/自転車で/東京に/行きます。

Translation:

He went to Tokyo by bike

4. 彼は/病気の/ため/休学した

Translation:

He dropped out of school because of illness.

5. 彼は5万円でこの着物を買いました。

Translation:

He bought the kimono for 50000 yen.

For the case auxiliary words in the above sentences, the English auxiliary words corresponding to are "in", "ride", "ride", "cause" and "flower", in example 1, "canteen" refers to the action place, in example 2.3, *飛行機で* and *自転車で* refer to the means of transportation. In example 4, C is the reason, and in example 5, "50000 nine" is the coin. Traffic means are divided into different categories. Example 2.3 can be distinguished. Therefore, we can infer its meaning according to the semantic category of the preposition of case auxiliary.

5. Conclusions

Aiming at solving the problems of news information clustering of COVID-19 epidemic situation and Japanese Chinese translation, this paper proposes a novel structure framework of Japanese to Chinese News clustering and Japanese dictionary for machine translation, which solves the problem of ambiguous choice of polysemy words in news translation process, that is, the choice of polysemy words of adjunct words. In addition, in order to solve the problem of COVID-19 epidemic news information redundancy and low accuracy, this paper describes a clustering method based on Web page link structure and tag content information. This method uses DOM and XML technology to assist in the experiment, and improves the original web page clustering method. This method

uses the combination of web structure information and content information to cluster, and the result is more convincing. In this method, MDL theory is introduced into the clustering process to improve the quality of clustering. After testing, the effectiveness of the design of the Japanese Chinese translation system and news clustering system is verified.

Acknowledgments

This paper is supported by the humanities and social sciences research project of The Education Department Of Henan Province of China. ("Corpus based study on the case dominance of Japanese Compound Verbs" 2019-ZZJH-615).

References

[1] S. Chakrabarti, B.E. Dom, S.R. Kumar, et al., Mining the Web's link structure, *Computer* **32**(8) (1999), 60–67.

[2] K.J. Kim and S.B. Cho, Personalized mining of web documents using link structures and fuzzy concept networks, *Applied Soft Computing* **7**(1) (2007), 398–410.

[3] G.J. Jing, Z. Zhang, H.Q. Wang, et al., Mining gene link information for survival pathway hunting, *Systems Biology Iet* **9**(4) (2015), 147–154.

[4] L. Bowers, Evaluation of noise exposures of miners operating highwall mining equipment, *Journal of the Acoustical Society of America* **127**(3) (2010), 1874.

[5] P. Chen, J. Chai, L. Zhang, et al., Development and Application of a Chinese Webpage Suicide Information Mining System (Sims), *Journal of Medical Systems* **38**(11) (2014), 88.

[6] I.É. Ginzburg, V.M. Zhigalkin, S.V. Zhigalkin, et al., Experimental investigation of plastic pure-shear deformation under simple and complex loadings, *Journal of Mining Science* **34**(1) (1998), 27–37.

[7] L. Vaughan, M. Kipp and Y. Gao, Are co-linked business web sites really related? A link classification study, *Online Information Review* **31**(4) (2007), 440–450.

[8] L. Wade, 4-Dimensional computer visualisation as an aid in mining education, *American Journal of Clinical Nutrition* **78**(2) (2008), 197–198.

[9] C.M. Perrott, Tool Materials for Drilling and Mining, *Annual Review of Materials Science* **9**(1) (1979), 23–50.

[10] P. Cheluszka and M. Ciupek, Application of the structured-light scanning for estimation of wear and tear of the link mining chains, *Journal of Neural Transmission* **70**(1) (2015), 42–47.

[11] A.B. Sohrabi, Graph Theory Application And Web Page Ranking For Website Link Structure Improvement, *Behaviour & Information Technology* **28**(1) (2009), 63–72.

[12] K.N. Trubetskoy, A.D. Ruban and V.S. Zaburdyaev, Justification methodology of gas removal methods and their parameters in underground coal mines, *Journal of Mining Science* **47**(1) (2011), 1–9.

[13] V.A. Chanturia, I.Z. Bunin, M.V. Ryazantseva, et al., Surface activation and induced change of physicochemical and process properties of galena by nanosecond electromagnetic pulses, *Journal of Mining Science* **50**(3) (2014), 573–586.

[14] G.R. Bochkarev, V.I. Rostovtsev, Y.P. Veigel't, et al., Effect of accelerated electrons on structural and technological properties of ores and minerals, *Journal of Mining Science* **28**(6) (1993), 571–577.

[15] J. Jonak, Influence of Friction on the Chip Size in Cutting the Brittle Materials, *Journal of Mining Science* **37**(4) (2001), 407–410.

[16] L.A. Nazarova, L.A. Nazarov, G.Y. Polevshchikov..., Inverse problem solution for estimating gas content and gas diffusion coefficient of coal, *Journal of Mining Science* **48**(5) (2012), 781–788.

[17] T. Guerel, L.D. Raedt and S. Rotter, Ranking neurons for mining structure-activity relations in biological neural networks: Neuron Rank, *Neurocomputing* **70**(10–12) (2007), 1897–1901.