Acta Technica *62* No. 2B/2017, 73–84

Text mining technology based on cloud computing

Chong Xing[1], Kunhao Wang[2,3]

Abstract. The purpose is to study the text mining technology based on cloud computing. Text mining is the process of getting user care and valuable information from unstructured text data. With the developing of informatization, the massive data processing has become an urgent problem. Based on this, the Hadoop cloud computing platform and the MapReduce programming model are described in detail. The Chinese word segmentation and the new word recognition algorithm are mainly studied. In addition, based on the Hadoop platform, the MapReduce solution of the two algorithms were presented. By setting up Hadoop experiment platform, the two improved algorithms are programmed. Finally, the performance and advantages of the new algorithm are analyzed by experiments. The results show that the MapReduce programming model can improve the efficiency of processing large-scale data. Therefore, it can be concluded that it is very meaningful to combine cloud computing with text mining to deal with massive text data.

Key words. Chinese word segmentation, new word recognition, cloud computing, mining.

1. Introduction

At present, most of the information on the Internet is in the form of text. Therefore, it is important to extract resources efficiently from a large number of unstructured text messages [1]. The study of text mining effectively solves this problem. As an important research direction of data mining, mining has been applied to many fields such as search, classification, recommendation system, public opinion and viewpoint mining. It has social and economic significance for scientific research and enterprise applications [2]. Today, the blowout of the Internet's massive information also poses a serious challenge to traditional information processing methods. The traditional information processing methods cannot deal with large amounts of data [3]. The vast majority are in the stand-alone processing, which is easily restricted by computer hardware devices such as processors and storage media. When dealing with large amounts of data, it seems powerless. Nowadays, the industry has

[1]Department of Information and Technology, Changchun Finance College, Jilin, 130028, China
[2]School of information engineering , Changchun Sci-Tech university, Jilin, 130600, China
[3]Corresponding author

paid more and more attention to this problem, and has accumulated some experience. Among them, the parallel processing of data is an important and effective means [4]. Cloud computing is the development and continuation of parallel computing, distributed computing, and grid computing. It is the result of a combination of virtualization, utility computing, Iaas (infrastructure as a service), Paas (platform as a service), Saas (software as a service), and so on [5]. The basic principles of cloud computing are as follows. By distributing computing on a large number of parallel computers, the operation of enterprise data centers is more similar to the use of the internet. As a result, the enterprise can switch the resource to the required application at any time, and access the computer and storage system according to the requirements [6]. It is foreseeable that cloud computing technology has a broad development prospects. Therefore, the application of cloud computing to other areas has become the focus of the current research. Based on this, text mining and cloud computing technology have been studied. Combining the classical algorithm of text mining and cloud computing technology, the improved text mining algorithm overcomes the shortcomings of traditional algorithms. It meets the requirement of dealing with mass data [7].

2. State of the art

2.1. The word segmentation algorithm

An example is given to illustrate the main flow of the segmentation algorithm. Entering a sentence "Liu Shuang welcomes you", the participle of the body flow is as follows. First, Participle Liu / double / welcome / you"; second, posTagging (part of speech tagging) Liu /q double j/ welcome /v you /r"; third, NE identification (name, transliteration name, place name), identification "Liu /q, double j/, welcome /v, you /r", Liu Shuang /nr"; fourth, re-word: "Liu Shuang / welcome / you"; fifth, re-posTagging (verbose) "Liu Shuang / nr / welcome / v you / r". finally, the participle is finished. The main idea of Chinese segmentation algorithm is to segment the word by CHMM (cascaded Markov model) first. Through layering, it not only increases the accuracy of word segmentation, but also ensures the efficiency of word segmentation. It is divided into five layers. The Chinese lexical analysis framework based on CHMM is shown in Fig. 1.

First, the dictionary is loaded. Then, the atoms are segmented. On this basis, the N- shortest path segmentation is carried out, and the segmentation results of the previous N are found out. The binary wordbook is generated. Then, the word formation result is generated. Finally, the part of the annotation is carried out and the main word segmentation step is completed.

2.2. The new word recognition algorithm

New words are one of the unregistered words. It is a word that does not appear in the dictionary [8]. Language develops with the development of society. In vocabulary, its big performance is the new words and the emergence of new phrases.

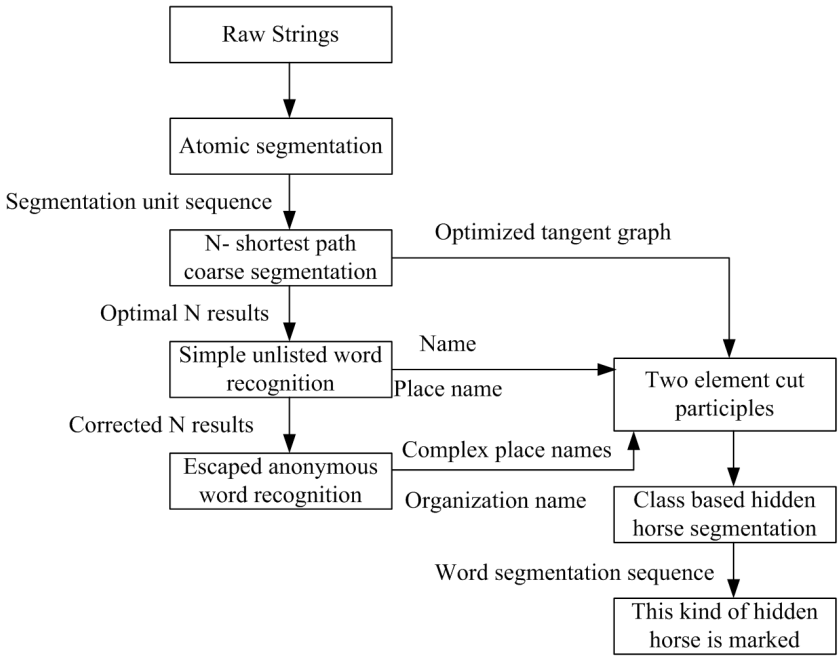

Fig. 1. The Chinese lexical analysis framework based on CHMM

The new words are first proposed in a particular field. After that, the frequency of its repetition increases. Finally, the new words stabilize [9]. It not only appears many times in a document, but also occurs repeatedly in many documents, which is a prerequisite for the recognition of new words. The emergence of new words reflects the emerging of new things, but it brings challenges to the processing of new Chinese words. Therefore, the emergence of new words has aroused special attention of linguists in recent years [10].

The new words mainly fall into two categories: the first category is named entities. It includes geographical names, names and institutional names. The second category is the emergence of new things with the words, such as "Super Girl", "Ray", "scientific concept of development" and so on [11]. The two main indicators to judge the pros and cons of the new word recognition algorithm are accuracy rate and recall rate. The following are the definitions of these two indicators:

$$\text{Accuracy rate P} = \frac{\text{Identification of new words correctly}}{\text{Total number of candidate words}} \times 100\%\,, \tag{1}$$

$$\text{Recall R} = \frac{\text{Identification of new words correctly}}{\text{New number of candidate words in total}} \times 100\%\,. \tag{2}$$

Among them, the total number of new words in the candidate words is generally calculated by manual.

3. Methodology

MapReduce is a major feature and core part of the Hadoop platform [12]. Therefore, the focus of the algorithm design is how to make the original single machine algorithm reasonably MapReduce, and transplant it to Hadoop. The proposed new word recognition algorithm based on cloud computing uses the combination of rules and statistics to identify [13]. It cooperates with the dictionary and noise dictionary. In order to achieve better accuracy and recall rate, a pruning strategy is used to filter out noise words. The new word recognition algorithm MapReduce process is shown in Fig. 2.

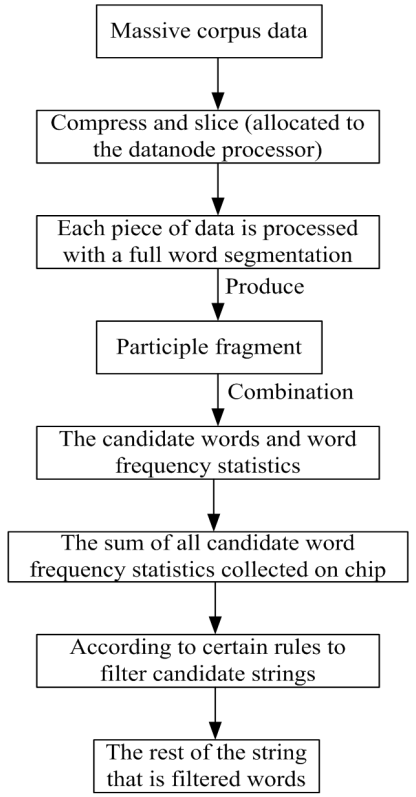

Fig. 2. Process of new word recognition algorithm

The new words are extracted from the word segmentation after the word processing. The recognition process of the new words is essentially the process of fusing these successive string fragments. That is, to find a string that is adjacent to each other and should not be cut, and combine them to form new words. According to statistics, most new words are combinations of single words or combinations of single words and multiple words. The combination of multi-word and multi-word is less frequent. Therefore, the candidate string extracted from the corpus data is mainly concerned with the string that contains the single word. Due to the relationship

between massive data, the corpus data that used in this paper is handled by pre-compression. The compression format is Sequence File Input Format. This format is the compression method that used on the Hadoop platform. It can reduce the size of the data and can effectively improve the speed of reading and writing.

3.1. MapReduce design of Chinese word segmentation

Since the Chinese word segmentation process is a global serial structure, it cannot be split into parallel processing. So, this approach is as following. First, the system configures the parameters globally and initializes the dictionary, so that each Data node does not have to separate the dictionary separately. It saves a lot of overhead and time to initialize the dictionary. Second, the input data is fragmented, and then assigned it to each Datanode for a complete word segmentation process. After the Map operation, the effective word segmentation has been obtained. Therefore, this algorithm does not require Reduce operation, as long as the segmentation results of each segmentation is compressed and output to the file directly. The flow chart is shown in Fig. 3. MapReduce parameters and logic instructions is as shown in Table 1.

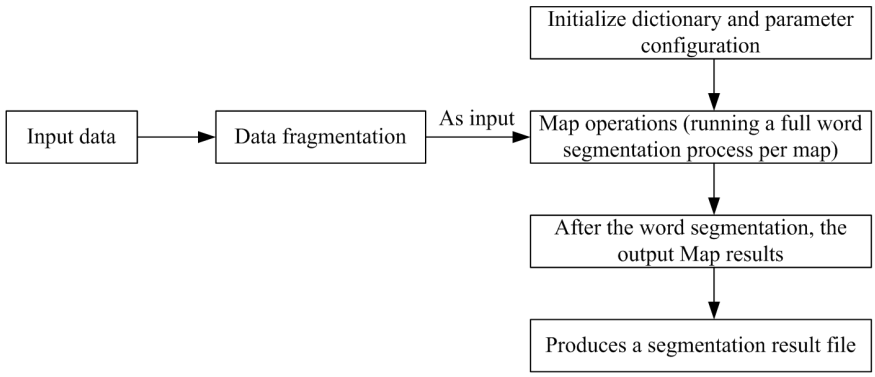

Fig. 3. Chinese word segment MapReduce process

3.2. The MapReduce design of new word recognition algorithm

The MapReduce flow chart is shown in Fig. 4. The main idea of the new word recognition algorithm MapReduce is as follows. On Namenode, configuration information such as dictionaries and parameters is initialized. This information is shared by all nodes in the cluster (including Namenode and Datanode). According to the size of the input data and the number of machines in the cluster, the input data is sliced and sent to the corresponding Datanode to wait for processing. A Chinese word segmentation is performed on the data fragmentation, and each Datanode performs a complete and independent Chinese word segmentation process for the assigned slice data. The advantage of our Hadoop platform design is that the designer cannot care about this data allocation and communication problem at

all, because the platform itself provides a stable and intelligent solution for distribution. The extraction of candidate words integrates the successive words remaining in the segmentation fragments after the end of the participle, and forms the candidate words.

Table 1. MapReduce logic instructions

	Map function	Reduce function
Input data segmentation logic	Sequence file A key-value pair is a cut	
Enter Key, value	Key: it is the key of the Sequence file, that is, the path of the source text file Value: the contents of the text	
Output Key, value	Key: the path of the source text file; Value: word segmentation result	
Function calculation logic	The value corresponding to each input key, that is, the text, is processed by word segmentation	
Global data	Word segmentation dictionary and part of speech display parameters	

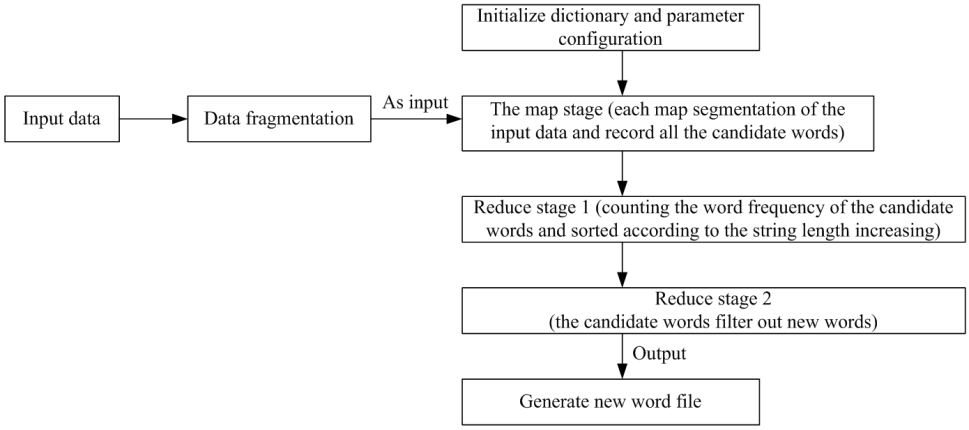

Fig. 4. The new word recognition algorithm MapReduce process

The so-called continuous word, refers to a combination of two or more consecutive words. For example, the "melamine incident is a hot spot of recent concern", the result of word segmentation (without part of speech) is "three / poly / cyanide / amine / event / yes / near / attention / a /". The single words are "melamine", "one", which can generate candidate words for "trimeric", "melamine", "cyanide", "melamine", "cyanamide", " one "and so on. The candidate word frequency statistics collects the candidate words and word frequency statistics processed on all Datanodes (the same word frequency is added). The candidate words are filtered first, and the frequency

of candidate candidates is counted. By comparing it with the pre-set threshold, the candidate words whose frequency is smaller than the threshold are filtered out. Then, in conjunction with a stop dictionary, a noisy dictionary, a pruning strategy is used to filter the remaining candidates. The candidates left by the above steps are considered new words.

3.3. New word filtering method based on pruning strategy

Because the garbage strings and the redundant information exists in a large number of candidate words from the word fragments, such as melamine". If "melamine" is selected, then "trimer" and "cyanuric chloride" must also be selected. However, these are redundant information, and they should be deleted. Based on the frequency filtering of candidate words, and combined with the stop word dictionary and noise dictionary, a pruning strategy to effectively filter the noise words and redundant information is proposed. The pruning strategy is defined as follows.

First, if $N(c_ic_{i+1}...c_{i+j}) = N(c_{i+1}c_{i+2}...c_{i+j+1}) = N(c_ic_{i+1}...c_{i+j}c_{i+j+1})$, then, $N(c_ic_{i+1}...c_{i+j})$ and $N(c_{i+1}c_{i+2}...c_{i+j+1})$ are deleted.

Second, if $N(c_ic_{i+1}...c_{i+j}) > N(c_ic_{i+1}...c_{i+j}c_{i+j+1})$ or $N(c_{i+1}c_{i+2}...c_{i+j+1}) < N(c_ic_{i+1}...c_{i+j}c_{i+j+1})$, then, $N(c_ic_{i+1}...c_{i+j}c_{i+j+1})$, where N is the frequency at which strings appear, and c is a single word.

If the frequency of a parent string is equal to the frequency of its two largest substrings, it is also assumed that the parent string is a word, the substring is deleted, and the pruning ends. On the contrary, if any of the largest substring frequencies are greater than the parent string, the parent string is not considered a word and the parent string is deleted. The cycle was repeated until no string substring, and the pruning is ended. The overall filtration process is shown in Fig. 5.

The new word recognition algorithm only takes into account the recognition of four words (including four words). The reason is that after statistics, the word more than four words is little, and the frequency is not high. Through the experimental summary, the frequency of the occurrence of four words and more than four words is obviously different. Therefore, this paper proposes a double threshold strategy to deal with four words and four words respectively. It achieves better results.

4. Result analysis and discussion

4.1. Chinese word segmentation

The test data uses lab internal data for plain text files, and the total size of the source file is 1.6 G, totaling more than 9000 files. After transformation, the size of sequence file was 635 MB. The test result is related to the machine condition and network at that time. The tests were conducted in a cluster with Datanode numbers of 2 units, 4 units, 6 units, and 8 units. After several tests, the acceleration ratio curve is obtained. The test acceleration is shown in Fig. 6.

It can be seen from the figure that as the number of clusters increases, the acceleration ratio is getting higher and higher. Due to the use of a specific compression

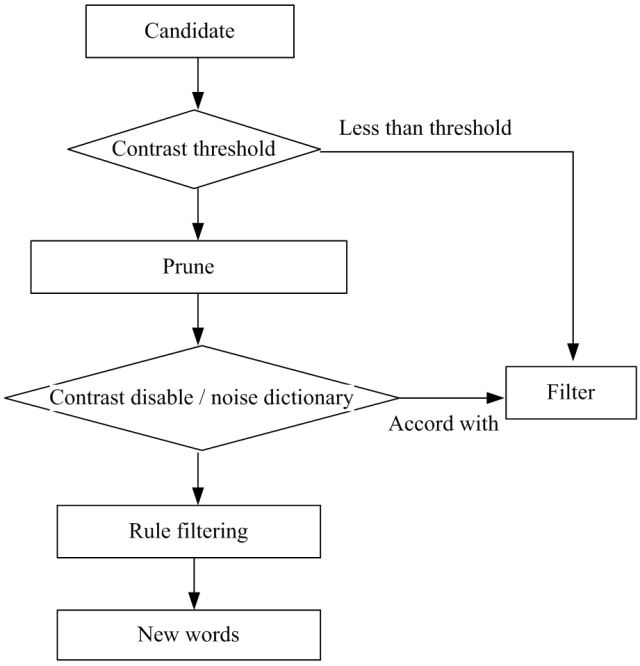

Fig. 5. Filtering process of candidate words

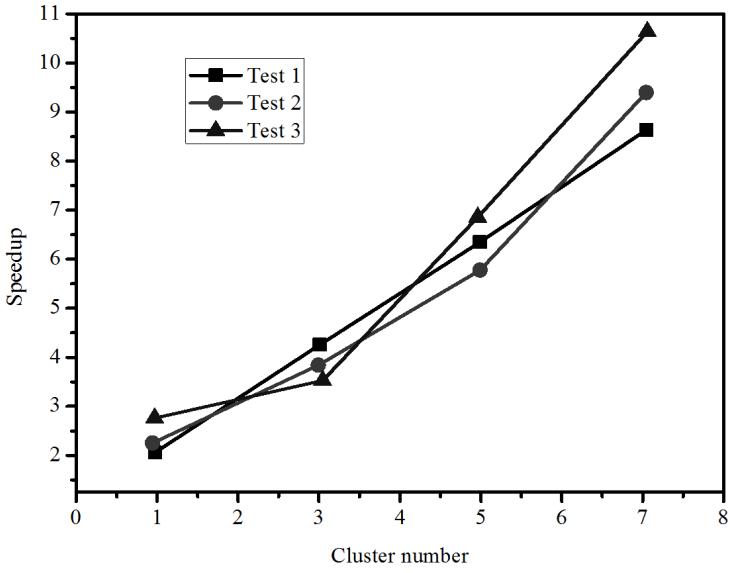

Fig. 6. Test of acceleration ratio

format, it saves a lot of time compared to processing the raw data directly. However, the experimental results show that the number of cluster machines is relatively small when the overall performance increase is not obvious. There are two main reasons for this situation. The communication between clusters takes a certain amount of time. The advantage of Hadoop platform is to deal with massive data through large-scale cluster mode. The number of experimental clusters is small, and the test data is small. It is difficult to play Hadoop platform of this advantage.

From this, it can be guessed that if the data scale is increased, the speedup ratio will be improved obviously. To verify the conjecture, another experiment is carried out in this paper. The scale of the experimental data is increased by two times to 4.8 G. After compression, it is 1.8 G, and the experiment is repeated many times with the number of two Datanode added. The experimental results are shown in Fig. 7.

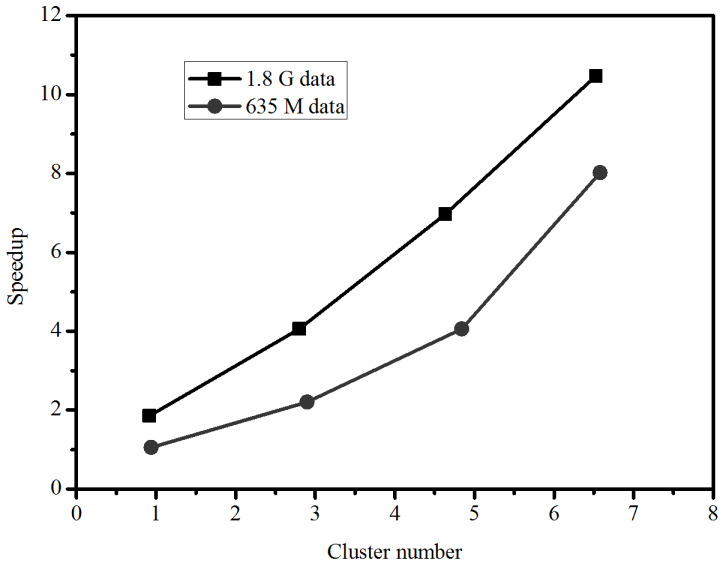

Fig. 7. Acceleration of ratio comparison

As can be seen from the Fig. 7, as the amount of data increases, the acceleration ratio is significantly improved. Moreover, as the number of clusters increases, the speedup is faster. The experimental results are in line with the expected conjecture, which verifies the advantages of the Hadoop platform for processing large data. The test data used in this article, even if it has reached nearly 5 G. However, in the mass storage and massive data of today, it is not too large. And the size of the 9 cluster is also small. Because of the constraints of these experimental hardware, this article has not done more testing. However, it can be inferred that the advantages and efficiency of the Hadoop platform, especially the MapReduce programming model in dealing with large-scale data, are quite objective.

5. Conclusion

On the Hadoop platform, the MapReduce improvements of the two text mining algorithms are implemented. It solves the problem that traditional mining algorithms are relatively difficult to deal with large-scale mass data sets. By briefly introducing the principle of Chinese word segmentation and new word recognition algorithm, a method of MapReduce of these two algorithms is proposed. At the same time, through the cluster experiment, the feasibility of the method is verified. It gets the ideal speedup effect.

References

[1] A. M. Espín, F. Exadaktylos, B. Herrmann, P. Brañas-Garza: *Short- and long-run goals in ultimatum bargaining: Impatience predicts spite-based behavior.* Frontiers in Behavioral Neuroscience (2015), No. 9, 214.

[2] S. Bennett, T. Wernberg, S. D. Connell, A. J. Hobday, C. R. Johnson, E. S. Poloczanska: *The 'Great Southern Reef': Social, ecological and economic value of Australia's neglected kelp forests.* Marine and Freshwater Research *67* (2016), No. 1, 47–56.

[3] L. A. Woodham, R. H. Ellaway, J. Round, S. Vaughan, T. Poulton, N. Zary: *Medical student and tutor perceptions of video versus text in an interactive online virtual patient for problem-based learning: A pilot study.* Journal of Medical Internet Research *17* (2015), No. 6, e151.

[4] K. Kirubhakaran, K. M. Parammasivam: *Understanding blowout phenomena to the induced angle of V-gutter-stabilized flames.* International Journal of Turbo & Jet-Engines *33* (2016), No. 1, 81–85.

[5] B. J. Ellis, A. A. Volk, J. M. Gonzalez, D. D. Embry: *The meaningful roles intervention: An evolutionary approach to reducing bullying and increasing prosocial behavior.* Journal of Research on Adolescence *26* (2016), No. 4, 622–637.

[6] L. Qian, R. Agarwal, G. Hoetker: *Configuration of value chain activities: The effect of pre-entry capabilities, transaction hazards, and industry evolution on decisions to internalize.* Journal Organization Science *23* (2012), No. 15, 1330–1349.

[7] D. Dai, H. Wu, W. Zhang: *Utilization of field enhancement in plasmonic waveguides for subwavelength light-guiding, polarization handling, heating, and optical sensing.* Materials (Basel) *8* (2015), No. 10, 6772–6791.

[8] H. Santos, A. Latgé, J. E. Alvarellos, L. Chico: *All-electrical production of spin-polarized currents in carbon nanotubes: Rashba spin-orbit interaction.* Physical Review B *93* (2016), No. 16, paper 165424.

[9] J. Liu, W. Zhu, T. Ebrahimi, J. Apostolopoulos, X. S. Hua, C. Wu: *Introduction to the special section on visual computing in the cloud: Fundamentals and application.* IEEE Transactions on Circuits and Systems for Video Technology *25* (2015), No. 12, 1885–1887.

[10] T. Hirai, H. Masuyama, S. Kasahara, Y. Takahashi: *Performance analysis of large-scale parallel-distributed processing with backup tasks for cloud computing.* Journal of Industrial and Management Optimization *10* (2014), No. 1, 113–129.

[11] S. L. Smith, M. K. Pichora-Fuller, G. Alexander: *Development of the word auditory recognition and recall measure: a working memory test for use in rehabilitative audiology.* Ear Hear *37* (2016), No. 6, e360–376.

[12] T. R. McRackan, J. B. Ahlstrom, W. B. Clinkscales, T. A. Meyer, J. R. Dubno: *Clinical implications of word recognition differences in earphone and aided conditions.* Otology & Neurotology *37* (2016), No. 10, 1475–1481.

[13] T. Aitamurto: *Crowdsourced democratic deliberation in open policymaking: Definition, promises, challenges.* International Reports on Socio-Informatics (IRSI), Proc. CSCW 2016 – Workshop: Toward a Typology of Participation in Crowdwork *13* (2016), No. 1, 79–90.

Received August 7, 2017