Contemporary Engineering Sciences, Vol. 7, 2014, no. 22, 1095 - 1103
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49137

Big Data Summarization Using Semantic Feture for IoT on Cloud

Yoo-Kang Ji

School of Information and Communication GIST
123, Cheomdangwagi-ro, Buk-gu, Gwangju 500-712 Korea

Yong-Il Kim

Dept. of Internet Content, Honam University
417, Eodeung-ro, Gwangsan-gu, Gwangju, Korea

Sun Park

School of Information and Communication GIST
123, Cheomdangwagi-ro, Buk-gu, Gwangju 500-712 Korea
(Corresponding Author)

Abstract

Data management is a crucial aspect in the Internet of Things (IoT) on Cloud. Big data is about the processing and analysis of large data repositories on Cloud computing. Big document summarization method is an important technique for data management of IoT. Traditional document summarization methods are restricted to summarize suitable information from the exploding IoT big data on Cloud. This paper proposes a big data (i.e., documents, texts) summarization method using the extracted semantic feature which it is extracted by distributed parallel processing of NMF based cloud technique of Hadoop. The proposed method can well represent the inherent structure of big documents set using the semantic feature by the non-negative matrix factorization (NMF). In addition, it can summarize the big data size of document for IoT using the distributed parallel processing based on Hadoop. The experimental results demonstrate that the proposed method can summarize the big data document comparing with the single node of summarization methods.

Keywords: Internet of Things (IoT), big document data summarization, semantic features, non-negative matrix factorization (NMF), distributed parallel processing, Cloud, Hadoop

1. Introduction

Various application areas of Internet of Things (e.g., Smart Cities, Smart Car and mobility, Smart Home and assisted living, Smart Industries, Public safety, Energy & environmental protection, Agriculture and Tourism, etc.) have received high attention. The goal of the IoT is to enable things to be connected anytime, anyplace, with anything and anyone ideally using any path/network and any service. A potential of IoT is estimated by a combination with related technology approaches and concepts such as Cloud computing, Future Internet, big data, robotics and Semantic technologies. Integrated environments of IoT have been taking up smartphone platforms and capable of running a multiplicity of user-driven applications and connecting various sensors and objects [1].

Data management is a crucial aspect in the Internet of Things on Cloud computing. When considering a world of objects interconnected and constantly exchanging all types of information, the volume of the generated data and the processes involved in the handling of those data become important. There are many technologies and factors involved in the data management within the IoT context. Some of the most relevant concepts of data management are data collection and analysis, big data, semantic sensor networking, virtual sensors, and complex event processing [1]. This paper focuses on the big document data (i.e., big text data) for data management of IoT on Cloud computing.

The expansion of Internet data (e.g., web pages, image and video application, social networks, mobile devices, apps, sensors, and so on), according to IBM, more than 2.5 quintillion byte per day, to the extent that 90% of the world's data have been created over the past two years [1]. In addition, with the fast growth of the Internet access by user (i.e., smartphone, mobile devices, data of IoT, etc.), has increased the necessity of the information seeking from big data of IoT. However, it is difficult to find suitable information for user from Internet of Cloud environment. Summary information of Internet data can help to users, which the user can save time not only in deciding whether it is interesting or not but also in finding the information without having to read the full information with respect to big document data of IoT on Cloud.

Document summarization is the process of reducing the sizes of documents while maintaining their basic outlines. That is, it should distill the most important information (i.e., topics of document) from the document. The summarization method can involve either generic summaries or query-based summaries. A generic summary distills an overall sense of a document's contents, whereas a query-based summary distills only the contents of a document that is relevant to a user's query. It can also divide into single-document summarization or multi-document summarization according to the scope of the summary target. The purpose of multi-document summarization is to produce a single summary from a

set of related documents, whereas single-document summarization is intended to summarize only one document [2]. Traditional document summarization methods are restricted to summarize suitable information from the exploding Internet big data (i.e., SNS, email, message, blog, collection data of smartphone, etc.), since it have been studying for enhancing the summarization precision which it uses various statistical or natural language processing methods on single computer or server. In order to resolve the limitations of the traditional document summarizations for document size, this paper study a big document data summarization which the information is summarized by sentences extraction from a big document data of IoT on Cloud computing. The proposed method uses the extracted semantic feature of document by distributed parallel processing of NMF (i.e., distributed NMF) based cloud technique of Hadoop [3]. It can well represent the inherent structure of big data document set. In addition, it can summarize the big data document using the distributed parallel processing in connection with Hadoop of Cloud technique.

2. Related Works

2.1 Document Summarizations

Gong and Liu proposed summarization method using LSA (Latent Semantic Analysis). This method extracts the important sentence which has the largest index value with respect to the important singular vector by LSA [4]. Zha employed the mutual reinforcement principle (MRP) and sentence clustering for the generic summarization. Their method clusters sentences of documents into several topical groups by sentence clustering method. And then, sentences are extracted from each topical group by saliency scores using the MRP (i.e., modified LSA method) [5]. Yeh et al. proposed the summarization method using LSA and the text relationship map (TRM). Their method finds semantic sentences using LSA. TRM is constructed by the semantic sentences, and the important sentences are extracted by the number of links in TRM [6]. Li et al. extended the generic multi-document summarization using LSA for the query based document summarization [7]. Han et al. proposed a text summarization method using relevance feedback with query splitting (i.e., a query expansion process by splitting the initial query into several pieces) [8]. Diaz and Gervas proposed an item summarization method for the personalization of news delivery systems. The method uses three phrase-selection heuristics that build summaries using two generic summarizations and one personalized summarization depending on RF from news items [9]. Also, they proposed an automatic personalized summarization using a combination of generic and personalized methods. Their generic summarization methods combine the position method with the thematic word method. Their personalized method selects those sentences of a document that are most relevant to a given user model [10]. Kumar et al. generated personalized summaries using generic and user-specific methods based on proba-

bility. This method extracts the top ranking sentences by means of the generic sentence scoring and the user-specific sentence scoring [11]. Ko et al. proposed a web snippet generation method from web pages using PRF and a query-biased summarization based on the probability model [12]. Li and Chen extracted personalized text snippets using the probability sequence analysis and the hidden Markov model [13]. In our previous works [14, 15, 16], we proposed the document summarization methods using sentence ranking depending on the semantic features of the NMF. Our methods are can be divided into the generic document summarization using NMF [14], the multi-document summarization using clustering and NMF [15], and the personalized document summarization using pseudo relevance feedback and semantic feature [16]. However, these methods [4-16] are restricted by the big data document size and the summarization methods based on the single computer environment.

3. Non-negative Matrix Factorization Section

This section reviews NMF theory. In this paper, we define the matrix notation as follows: Let X_{*j} be j'th column vector of matrix X, X_{i*} be i'th row vector, and X_{ij} be the element of i'th row and j'th column. A matrix A (i.e., term-by-sentence matrix) is the preprocessed document set which it is comprised of sentences set of big data document of IoT. NMF is to decompose a given $m \times n$ matrix A into a non-negative matrix W and H as shown in Equation (1). Lee defines the matrix W and H which are a non-negative semantic feature matrix W and a non-negative semantic variable matrix H *respectively* [17].

$$A \approx WH \tag{1}$$

Where W is a $m \times r$ non-negative matrix, H is a $r \times n$ non-negative matrix, and r is a number of semantic feature vector. Usually r is chosen to be smaller than m or n, so that the total sizes of W and H are smaller than that of the original matrix A.

The objective function is used minimizing the Euclidean distance between each column of A and its' approximation $\tilde{A} = WH$, which was proposed by Lee and Seung [17]. As an objective function, the Frobenius norm is used:

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij} - \sum_{l=1}^{r} W_{il}H_{lj}\right)^2 \tag{2}$$

Updating W and H is kept until $\Theta_E(W,H)$ converges under the predefined threshold or exceeds the number of repetition. The update rules are as follows:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^T A)_{a\mu}}{(W^T WH)_{a\mu}}, \qquad W_{ia} \leftarrow W_{ia}\frac{(AH^T)_{ia}}{(WHH^T)_{ia}} \tag{3}$$

The advantage of the two non-negative matrices W and H are described as follows: All semantic variables (H_{lj}) are used to represent each sentence. W and H are represented sparsely as shown in Figure 1. Intuitively, it make more sense for

each sentence to be associated with some small subset of a large array of topics (W_{*l}), rather than just one topic or all the topics with respect to big data document set. In each semantic feature (W_{*l}), the NMF has grouped together semantically related terms from big document data set [17].

4. Proposed Big Document Summarization Method

This paper proposes a big document summarization method using semantic feature by distributed NMF based Hadoop for IoT on Could. The proposed method consists of two phases: summarization module, and distributed parallel processing module, as shown in Figure 1. In the subsections below, each phase is explained in full.

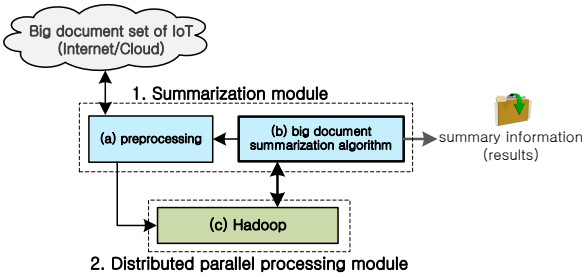

Fig. 1. Big document summarization method using Hadoop and semantic features

4.1. Summarization module

This section describes how to extract the sentences from the big document set. In the summarization module consists of preprocessing and summary algorithm. In the preprocessing phase of Figure 1(a), the big document set is decomposed into individual sentences, Rijsbergen's stop words list [18, 19] is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [18, 19]. Then, the term sentence frequency matrix A is constructed.

In this paper, we use Mahout's tools [20] to construct frequency matrix for saving the matrix to the Hadoop Distributed File System (HDFS). The Sequence File Directory tool is used to convert sentences set into sequence file form. The seq2sparse tool is used to convert the sequence file into vector matrix. The constructed term sentence frequency matrix is saved into HDFS by distributed parallel processing of Hadoop framework in Figure 1(c). The distributed parallel processing method for extracting the semantic features (i.e., distributed NMF) is explained in next subsection. In the big document summarization algorithm phase of Figure 1(b), Semantic features of big document for summarizing are extracted by the modified our previous method [14, 15, 16] and Liu's distributed NMF (dNMF) method [21] based on distributed parallel processing on Hadoop MapReduce programming. The summarization algorithm phase is as follows: In the first step, the dNMF (i.e., chapter 3.2) are performed after the preprocessing phase. Second step, the important sentences are selected by the semantic weight based on the semantic variable matrix of the dNMF. We define the semantic weight *sweight*() as Equation (4).

$$\text{sweight}(H_{*j}) = \sum_{i=1}^{r} \left(H_{ij} \times \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{r} \sum_{q=1}^{n} H_{pq}} \right) \quad (4)$$

The semantic weight denotes how much the sentence reflects significant topics of the big document data, which are represented as semantic variables with respect to sentences.

Distributed parallel processing module calculates semantic features using the distributed NMF and the saved term by sentence matrix on the HDFS. In this paper, we modify Liu's distributed NMF method for our big document summarization method. Liu's NMF method is designed by MapReduce to the distributed parallel processing on Hadoop framework. Table 3 shows the summary of Liu's NMF method using MapReduce of Hadoop [21].

5. Experimental Results

For our experiment data, we used real data of Google Korea (i.e., www.google.co.kr). We gave a 1000 topic to retrieve document set from the searched results and the related sites. Table 1 shows the experiment environment for the single node and multi node summarization methods. In this paper, we used the recall (R), precision (P), and F-measure to evaluate the performance of the proposed method using real data of Google Korea. Let S_{man}, S_{sum} be the set of topic selected by the human evaluators, and the summarizer, respectively. The standard definitions of recall (R), precision (P), and *F*-measure are defined as follows [18, 19]:

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|}, P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, F = \frac{2RP}{R+P} \quad (5)$$

Table 1. Experiment Environment

	Single node (1 computer)	Multi node (4 computers)
Hardware specification (computer & network)	Intel i3 CPU * 1, RAM 8G * 1 HDD 1TB * 1	Intel i3 CPU * 4, RAM 8G * 4 HDD 1TB * 4, LAN 1GB * 4
Operating system	CentOS 6.3	CentOS 6.3
Framework	-	Hadoop

We implemented five different summarization method based semantic feature, such as LSA, MRP, TRM, NMF, and BDS. LSA denotes Gong and Liu's summarization method using latent semantic analysis [4]. MRP denotes Zha's method using the mutual reinforcement principle [5]. TRM denotes Yeh's method using the text relationship map [6]. NMF denotes our previous method using non-negative matrix factorization [14]. BDS (big data summarization) denotes the proposed method in this paper. Our proposed method is only big document data summarization method based on multi node computer environment.

In the evaluation results, the average recall of the BDS is approximately 32.17% higher than that of the LSA, 28.42% higher than that of the MRP, 11.45% higher than that of the TRM, and 5.23% higher than that of the NMF. The average precision of the BDS is approximately 23.58% higher than that of the LSA, 9.25%

higher than that of the MRP, 20.60% higher than that of the TRM, and 5.39% higher than that of the NMF. The average *F*-measure of the BDS is approximately 27.30% higher than that of the LSA, 18.11% higher than that of the MRP, 17.14% higher than that of the TRM, and 5.33% higher than that of the NMF.

Table 2 shows the experiment results with relation to calculation time of NMF and BDS according to change data size. Original document data size (i.e., before preprocessing phase) is increased by double size from 300M to 4.8G. In this experiment, we choose the NMF and BDS because these methods have a top score in Figure 4. In our experiment, single node method (i.e., LSA, MRP, TRM, NMF) is not worked over 600 M sizes of original document data (i.e., before preprocessing phase) since the method do not support to calculate big document data size. However, our proposed method can calculate the big document data size since our method is designed to apply the distributed parallel processing based Hadoop framework.

Table 2. Experiment results of calculation time regarding NMF and BDS.

Environment (method)	300 MB	600 M	1.2 G	2.4 G	4.8 G
Single node (NMF)	194 minute	-	-	-	-
Multi node (BDS)	25 minute	34 minute	167 minute	427 minute	531 minute

6. Conclusion

Traditional document summarization methods are restricted for summarizing suitable information from the big document data in the Internet of Things on Cloud computing, since it have been proposing for enhancing the summarization precision which it uses various statistical or natural language processing methods based on single node computer environment. In order to resolve the limitations of the summarizations for big document data of IoT, this paper proposed big document summarization method which the information is summarized from a big document data of IoT on Cloud computing. The proposed method can well represent the inherent structure of big documents set using the semantic feature by the distributed NMF based on Hadoop MapReduce. In addition, it can summarize the big data document using the distributed parallel processing in connection with Hadoop framework comparing with the summarization methods based on single node computing.

References

[1] O. Vermesan, P. Friess, “Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems”, River Publishers, 2013.
[2] I. Mani, “Automatic Summarization”, John Benjamins Publishing Company, 2001.
[3] The Apache Hadoop project, “http://hadoop.apache.org/”, 2013.
[4] Y. Gong, X. Liu, “Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis”, in proceedings of the 24th annual international

ACM SIGIR conference on research and development in information retrieval (SIGIR'01), pp.19-25, New Orleans, USA, 2001.
[5] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering", In proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'02), pp.113-120, Tampere, Finland, 2002.
[6] J. Y. Yeh, H. R. Ke, W. P. Yang, I. H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis", Information Processing and Management 41, pp.75-95, 2005.
[7] W. Li, B. Li, M. Wu, "Query Focus Guided Selection Strategy for DUC 2006", In proceedings of the Document Understanding Conference (DUC'06), 2006.
[8] K S Han, D H Bea, and H C Rim, "Automatic Text Summarization Based on Relevance Feedback with Query Splitting," In proceedings of the 5th International Workshop on Information Retrieval with Asian Language, Hong Kong, pp.201-2, Sep. 2000.
[9] A Diaz, and P Gervas, "Item Summarization in Personalisation of News Delivery Systems," In proceeding of the 7th International Conference on Text, Speech and Dialogue (TSD), LNAI 3206, Brno, Czech Republic, pp. 49-56, Sep. 2004.
[10] A Diaz, and P Gervas, "User-model based personalized summarization," Information Processing and Management, vol. 43, pp.1715-34, Mar. 2007.
[11] C Kumar, P Pingali, and V Varma, "Generating Personalized Summaries Using Public Available Web Documets," In proceeding of the International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, pp.103-6, Dec. 2008.
[12] Y J Ko, H K An, and J Y Seo, "Pseudo-relevance feedback and statistical query expansion for web snippet generation," Information Processing Letters, vol. 109, pp.18-22, 2008.
[13] Q Li, and Y P Chen, "Personalized text snippet extraction using statistical language models," Pattern Recognition, vol. 43, pp.378-86, 2010.
[14]. J H Lee, S Park, C M Ahn, and D H Kim, "Automatic Generic Document Summarization Based on Non-negative Matrix Factorization," Information Processing and Management, vol. 45, pp.20-34, Jan. 2009.
[15] S Park, B R Cha, and D U An, "Automatic Multi-document Summarization Based on Clustering and Nonnegative Matrix Factorization," IETE TECHNICAL REVIEW, vol. 27, no. 2, pp.167-78, Mar. 2010.
[16] S Park, B R Char, C. U. Kwon "Personalized Document Summarization using Pseudo Relevance Feedback and Semantic Feature" IETE JOURNAL, vol. 58, no. 2, pp.155-165, 2012.
[17] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, 401, pp. 788-791, Oct., 1999.
[18] B. Y. Ricardo, R. N. Berthier, "Moden Information Retrieval: the concepts and technology behind search Second edition, ACM Press, 2011.

[19] W. B. Frankes, B. Y. Ricardo, “Information Retrieval: Data Structure & Algorithms”, Prentice-Hall, 1992.
[20] S. Owen, R. Anil, T. Dunning, E. Friedman, “Mahout in Action”, Manning Publications, 2011.
[21] C. Liu, H. C. Yang, J. Fan, L. W. He, Y. M. Wang, "Distributed Nonnegative Matrix Factorization for Web-Scale Dyadic Data Analysis on MapReduce," in Proceeding of the International World Wide Web Conference Committee, USA, pp.1-10, 2010.

Received: August 4, 2014