2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

Automatic Document Archiving for Cloud Storage Using Text Mining-Based Topic Identification Technique

Keedong yoo

Dept. of MIS, School of Economics and Commerce, Dankook University, Cheonan, Republic of Korea

Abstract. Cloud storage can deliver users various benefits; however it also has a serious problem in using, which is the difficulty in the process of storing and retrieving documents. Assistance in concluding the directory to store a document can be accomplished by analyzing the contents of the document with respect to the directories defined in the cloud storage. This research proposes a methodology to automatically extract the predefined directory-specific keywords (or topics) of a working document to be stored. Based on the extracted keywords, any documents can be automatically stored under the directories in cloud storage.

Keywords: Cloud storage, Automatic archiving, Text mining, Topic identification

1. Introduction

Companies, nowadays, utilize cloud computing technologies not only to efficiently manage organizational information and knowledge but also to effectively secure them. Cloud storage, one of widely known cloud computing technologies, initiates its function by providing the Internet-based data storage as a service. One of the biggest merits of cloud storage is that users can access data in a cloud anytime and anywhere, using any device [4]. Typical examples of cloud storage services are Amazon S3 [8], Mosso [9], Wuala [11], or uCloud[10]; All of these services offer users clean and simple storage interfaces, hiding the details of the actual location and management of resources [6]. Once a document to be archived is stored in a cloud storage, users can access and download it anytime and anywhere if the right to access has been granted. Because of such advantage in utilizing organizational information resources, more companies and organizations are implementing the online storage under the cloud computing environment.

While cloud storage can deliver users various benefits, it also has not a few weaknesses in network security as well as privacy [7]. Additionally, many users also point out a very serious problem in using cloud storage, that is the difficulty in the process of storing and retrieving documents. To store a working document under any directory (or category) provided by the cloud storage, a user has to determine the directory that exactly coincide with the contents of the document. Since the directory is naturally various and complicated, determining a proper directory is not an easy work. When retrieving a document in which a user is interested, he/she has to spend not little time to locate the file because too many directories exist. Assistance in concluding the directory to store a document can be accomplished by analyzing the contents of the document with respect to the directories defined in the cloud storage. Since any keywords or topics extracted from the document stand for the possible directories under which the document must be stored, users can convenniently perform their job. In retrieving a document from the storage, more accurate and fast searching can be done because each document has been archived according to its keywords and topics.

This research tries to enhance the usability of cloud storage by automatically archiving the working documents. To do so, this research proposes a methodology to automatically extract the predefined directory-specific keywords (or topics) of a working document to be stored. Based on the extracted keywords, any documents can be automatically stored under the directories in cloud storage.

2. Methodology

As Fig.1 illustrates, a process to automatically identify topics (keywords) of the working document is additionally needed to automate the whole process of cloud storage-based archicing. Tasks in the dotted ellipse are required to perform automated topic identification, and they must be processed sequentially. Once the topic of the given working document is identified, the document can be automatically stored in cloud storage with the topic. By writing some programming codes to search corresponding directory with the topic and to save the document with the topic, automatic document archiving can be completed. When the destination directory is concluded, the system may send a message to the user to confirm whether the directory is valid. Although the user may change the directory as he/she intends, the system also automatically store the document in the location where the user designated by simply applying the agent programming. Specific roles of each task are as follows;

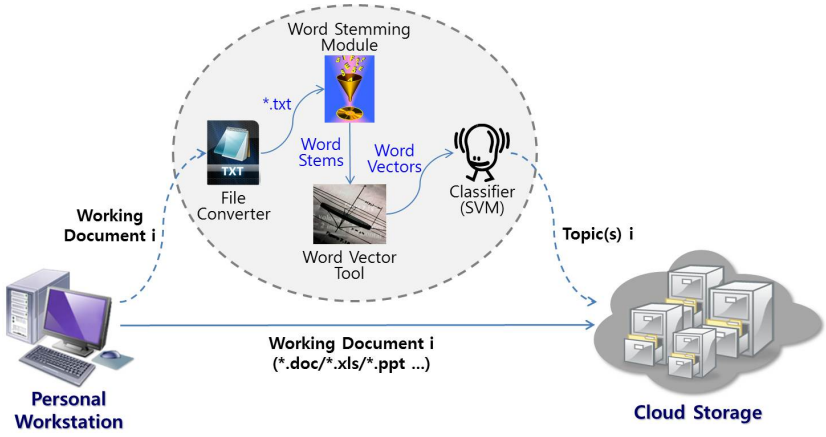

Fig. 1 Conceptual Framework

2.1 File Converter

A file converter changes the format (one of '.doc', '.ppt', '.xls', or '.html') of a working document to an analyzable one ('.txt') so that the following module can read the contents. A file converter plays the role of a file format filter that prepares input documents into a unified format ('.txt' in this research).

2.2 Word Stemming Module

To standardize the words in the document, unnecessary or redundant parts of each word must be eliminated. A stem, in linguistics, is the combination of the basic form of a word (called the root) plus any derivational morphemes, but excluding inflectional elements. This means, alternatively, that the stem is the form of the word to which inflectional morphemes can be added, if applicable. For example, the root of the English verb form destabilized is stabil- (alternate form of stable); the stem is de•stabil•ize, which includes the derivational affixes de- and -ize, but not the inflectional past tense suffix -(e)d.

2.3 Word Vector Tool

Based on the word stems from the word stemming module, the word vector tool transforms each word stem into the vector. To extract the vector value, TF/IDF(Term Frequency/Inverse Term Frequency) is used. TF/IDF is a statistical technique used to evaluate how important a word is to a document. The importance increases proportionally to the number of times a word appears in the document; however is offset by how common the word is in all of the documents in the collection or corpus. A high weight in TF/IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weight hence tends to filter out common terms. The word with the highest TF/IDF can be regarded as a keyword. However, to determine the keyword of a given document, usually every TF/IDF value of meaningful terms (stems) must be respectively calculated.

2.4 Classifier

A classifier extracts resultant keywords by projecting the word vectors of the target document on the vector spaces provided by the training based on a corpus. A corpus is a predefined directories, and each directory possesses a lot of related example documents. To train the classifiers based on the constructed corpus, sample documents can be excavated by browsing conventional Web pages. Because conventional Web pages have been already labeled with corresponding keywords as titles, in a sense, the title of each document can be deemed to be already formalized [3].

This research deploys SVM-based classifier, becasuse it is demonstrated that the SVM outperforms other similar text mining algorithms applicable to topic identification [1, 5]. The SVM determines the keyword of a document by depicting the word vectors on the vector space R^n (n: number of dimensions) and comparing the kernel functions of each document. The accuracy of the SVM was verified to be very high. If the prediction model has been trained sufficiently, then the SVM outputs very accurate and correct results. Comparing to the accuracy of manual classification, that of SVM-based classification was reported to be over 90% [2].

3. Concluding Remarks

Benefits from utilizing cloud storage in companies and organizations can be beyond description because it promotes effective and efficient sharing of organizational information and knowledge regardless to the time and place. If some usability issues around cloud computing, however, are not resolved realistically, then the benefits as well as interests can be scattered away. This research tries to resolve one of such usability issues around cloud storage by suggesting a practical guidance to relieve user's burden in selecting directories of cloud storage. The proposed methodology to identify the topics of working documents and to store documents with repect to the identified topics in an automated manner can contribute higher productivity and convenience of work. Companies can also expect more concentrated management of organizational information and knowledge through the proposed concepts, because more accurate and secured processing of organizational document archive is guarantted.

This research, however, must be further studied so that the proposed methodology can be applied to various mobile devices, such as smartphones and smartpads, which are the essential items of current users. To cope with this requirement, wireless-communication-oriented networking protocols must be additionally considered. Formal corpus, in addition, needs to be also developed to highten the performance of topic identification, because the accuracy of text mining mainly depends on the result of training based on the corpus. Since the corpus may have the same structure with the directory of cloud storage, this adjustment can also reinforce the realistic application of automatic document archiving.

4. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.H00021).

5. References

[1] Basu, A.; Watters, C.; Shepherd, M. Support Vector Machines for Text Categorization, Proceedings of the 36th Hawaii International Conference on System Sciences, 2002

[2] Hsu, C.W.; Chang, C.C.; Lin, C.J. A Practical Guide to Support Vector Classification: LibSVM Tutorial. available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2001

[3] Kim, S.; Suh, E.; Yoo, K. A study of context inference for Web-based information systems. Electronic Commerce Research and Applications, 2007; 6, 146-158

[4] Liu, Q.; Wang, G.; Wu, J. Secure and privacy preserving keyword searching for cloud storage services. Journal of Network and Computer Applications, 2011; article in press

[5] Meyer, D.; Leisch, F.; Hornik, K. The support vector machine under test. Neurocomputing [J], 2003; 55, 169-186

[6] Pamies-Juarez, L.; García-López, P.; Sánchez-Artigas, M.; Herrera, B. Towards the design of optimal data redundancy schemes for heterogeneous cloud storage infrastructures. Computer Networks [J], 2011; 55, 1100-1113

[7] Svantesson, D.; Clarke, R. Privacy and consumer risks in cloud computing. Computer Law & Security Review [J], 2010; 26, 391-397

[8] http://aws.amazon.com/s3

[9] http://www.rackspacecloud.com

[10] http://www.ucloud.com

[11] http://www.wuala.com