ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 402-406
doi: 10.18178/wcse.2016.06.065

Framework Deduplication Image Detection Assisted Multimedia System Using Multi Technique

NadiahYusof, NazatulAiniAbd Majid and Amirah Ismail [+]

Faculty of Information Science and Technology, National University of Malaysia (UKM), 43600 Selangor, Malaysia

Abstract. Research in image deduplication detection on internet has increasingly since 2008. Image storage in online database are too many and innumerable, including the same image uploaded by the user repeatedly and it is called as deduplication images. The big amount of deduplication images in database can lead waste of memory space for cloud database usage provided by the provider. This could make a user pays more for memory usage in cloud storage for similar image. Deduplication image will be reduced by using deduplication image detector either by plugin, middleware or software. However there still lack of research on deduplication image detector software or plugin for cloud storage. This is because, many researchers emphasize more on deduplication image detector in a standalone database. This paper compares standalone image deduplication detector, to identify a detail about technique using in deduplication detection and a relevant detection element of image deduplication. A new framework for deduplication detection in cloud has been proposed in this paper to begin early into the research image deduplication detection in cloud storage.

Keywords: image deduplication, cloud, database, Multimedia system, multimedia information retrieval

1. Introduction

Multimedia is increasingly becoming the "biggest big data" as the most important and valuable source for insights and information. It covers from everyone's experiences to everything happening in the world. There will be a lots of multimedia big data surveillance video, entertainment and social media, medical images, consumer images, general image, voice and video (Special issue 2015).

To name a few, only if their volumes increase to the extent that the traditional multimedia processing and analysis systems cannot handle effectively (Special issue 2015). Among them is that there is dumping a duplicate image in online either normal database or database in cloud. Database in cloud storage can be classified as an image or data storage in a database managed by the service provider for cloud storage such as Amazon Elastic Compute Cloud (EC2) has 6521 public virtual machine image (Amazon 2015), images or data operate independently of the cloud (L. Zhou et. al. 2014).

Based on observation, emphasis in the use of software image or plugin detector for cloud storage deduplication still less do in image deduplication detector and the studies in image deduplication detector has done by many research in standalone database or cloud database. Hence this study discussed the pilot test conducted on the image deduplication detector for a standalone database and the result for pilot test will be discuss in the preliminary studies section.

This paper has been divided into six part, introduction, literature review, preliminary studies, discussion, conclusion and future work

2. Literature Review

[+] Corresponding author. Tel.: +6 017-8720007, +6 013-3497108; fax: +6 03 - 8925 6732.
E-mail address: amirahismail@ukm.edu.my.

Automated robust methods for duplicate detection of images is getting more attention recently due to the exponential growth of multimedia content on the web. The large quantity of Multimedia data makes it infeasible to monitor them manually (G. Pratim et. al. 2007). Many duplicate detection (Y. Maret et. al. 2005, S. Roy 2005, G. Pratim et. al. 2007) and sub-image retrieval schemes have been proposed in the previous work. G. Pratimet. al. 2007 has been proposed a system that can detect duplicate image in a large scale database and focus in scalability issue. Y. Maretet. al. 2005 proposed duplicate detection based on support vector classifier. G. Pratimet. al. 2007 has been proved that scalable duplicate detection method has been demonstrated for the web and it applicable to use.L. Zhou et. al. 2014 used data deduplication method to detect image deduplication.

Detail research in deduplication image in cloud area are done by L. Zhou et. al. 2014. Their technique can significantly reduce the transmission time of image files to reduce the transmission time of image files that have already existed in storage. Also the deletion rate for image groups which have the same version of operating systems but different versions of software application is up about 58% (L. Zhou et. al. 2014). Fig 1 show new data deduplication solution (L. Zhou et. al. 2014). To store an image file that has been stored in the image storage server, the traditional data deduplication solution first divides the image file into image block, and then compares the fingerprint of each block with the fingerprints in database. This approach spends a lot of time on chunking and fingerprint matching. Therefore, a mechanism being able to quickly detect whether the image file is stored will effectively avoid segmentation of image storage server (L. Zhou et. al. 2014).

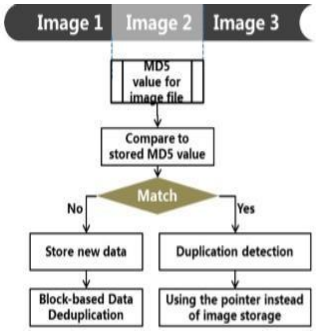

Fig. 1: Flow chart of image file deduplication (L. Zhou et. al. 2014).

Based on the literature review, research and discussion about deduplication image it's still increasingly (K. Saehoon et. al. 2015) and continue, in this paper has try to find current technique, method, category of image scope of data and user using in image deduplication detection because need to know the effectiveness detection deduplication image based on pilot test is only on existing software. In this paper, discussion based on deduplication image software in standalone database. Detail about deduplication image in standalone database in preliminary section.

3. Preliminary Studies

A pilot test was performed to an existing system using a standalone database where the images contained in the database that is the songket motives images. Total of songket motives images in database is 326 species. The total number of songket motives images can be divide into a number of categories of images. Songket motives image is chosen as a domain for the study because there are more than 1000 images (Arbaiyah 2009) and a number of the specific image is relatively large. Fig 2 show some of songket motives images in standalone database.

In this paper the aim of the pilot test is to evaluate existing software are related in data or image deduplication detection to find about advantage and disadvantage this software. Further evaluation was conducted to determine the image deduplication technique used to find similar image in the database, domain, the system founder, development goals and the effectiveness of the software in detecting duplicate images contained in the database and the result shown in Table 1.

Fig. 2: Some of songket image in standalone database.

TABLE I: COMPARISON IMAGE DEDUPLICATION DETECTION STANDALONE SOFTWARE USING MULTI TECHNIQUE

Bill	Founder Cloud/OS	Year	Software	Content (Image duplication detection based on?)				Technique							Duration
				Text	Sketch	Color	Image	Hash	Map reduce	SIFT/ GIFT	CRC	Pixel Based	Spatial Layout	Visual Similarity	
1	James (DigitalVolcano Software) http://www.duplicatecleaner.com/ (Standalone)	2015	Duplicate Cleaner (Support 17 different language)	√				√			√				0.1 S/ 16.6 MB 4/326
2	Jorg Rosenthal (http://www.anti-twin.com/) (Standalone)	2010	Anti-Twin (Support 15 different language)	√								√			2 S/ 16.6 MB 4/326 Image deduplication
3	Nirsof (http://www.nirsoft.net/utils/search_my_files.html) (Standalone)	2008	SearchMyFiles (Support 30 different language)	√				√							5 S/ 16.6 MB 0/326
4	Tago Software http://www.similarimagefinder.com/ (Standalone)	2012	Similar Image Finder				√						√		2 S / 16.6 MB 44 /549 Deduplication image
5	Bolide Software (http://www.bolidesoft.com/imagecomparer.html?dvs/) (Standalone)	2011	Image Search Pony				√							√	8 S / 16.6 MB 4/326
6	Alexander Nikolaev http://www.duplicate-finder.com/photo.html (Standalone)	2010	Awesome Duplicate Photo Finder				√							√	18 S/ 16.6 MB 14/326 deduplication image detect
7	UngSoft Developer Group http://www.ungsoft.com/ (Standalone)	2011	Similar Picture Find				√						√		3 S/ 16.6 MB 0/326 Deduplication image finder
8	Nils Maier https://tn123.org/about/ (Standalone)	2006	Similar Image				√							√	2 S/ 6.16 MB 14/326 deduplication image found in this software
9	Indeep Software http://indeepsoft.blogspot.my/p/exact-duplicate-finder.html (Standalone)	2015	Exact Duplicate Finder	√											2 S/ 6.16 MB 2/326 Deduplication image found
10	MindGems (http://www.mindgems.com/products/VS-Duplicate-Image-Finder/VSDIF-About.htm) (Standalone)	2009	Visual Similarity Duplicate Image Finder	√			√	√							2 S/ 16.6 MB 24 Duplicate image found
11	Phash http://phash.org/ (Cloud)	2008	Phash				√	√							Comparison similarity based one by one image

4. Discussion

Result of preliminary studies show, all the ten software that has doing a pilot test is using standalone database, as a place to detect deduplication image. Then software or plugin to detect image deduplication at cloud database is still new and increasingly in further. But for research in mobile cloud deduplication image detector it's has existing since 2008 (G. H. Jacob et. al. 2010).

After that, result of this pilot test shows, all the ten software use four different technique in image deduplication detection. The four of that technique is; hash (3 software), visual similarity (3 software), pixel based (1 software) and spatial layout (1 software), another left one software (exact duplicate finder) still didn't know which technique the developer use in image deduplication detection.

Based on the comparison recall of deduplication image detection using songket motives image as a domain, shows three from ten of that software get four deduplication image from 326 image. The software is duplicate cleaner, anti-twin and Image Search Pony. After that, result for similar image finder software, similar image and SearchMyFiles is 0 deduplication image detection. Awesome duplicate photo finder and exact duplicate finder software is 14 deduplication image detection. Lastly is visual similarity duplicate image finder get recall result is 24 deduplication image.

This pilot test is to see the results of the technical precision of duplicate images that are applied in software. Then 2 of 10 techniques that yield high precision combined with the technique used and the assumptions map reduce early techniques help reduce map calculation in terms of image indexing calculation. Two techniques that yield high-precision image deduplicationdetector is; hash and visual similarity. In this paper framework for this research has been proposed in 4.1 section.

4.1. A framework image deduplication detector

As Fig 13 shows the initial impression would duplicate image detection. Image uploaded by users must be filtered first through a duplicate image detection techniques aimed at ensuring the uploaded image is not available at the database in cloud. It is as sure a memory for the cloud is not filled with duplicate images. If the memory is filled with data clouds or overlapping image, users who rent storage space for cloud will face a loss, because the principle of memory space rental for the cloud is pay per use, the more huge memory space user use, become expensive user must pay. So that, this research to help reduce deduplication image in a cloud.

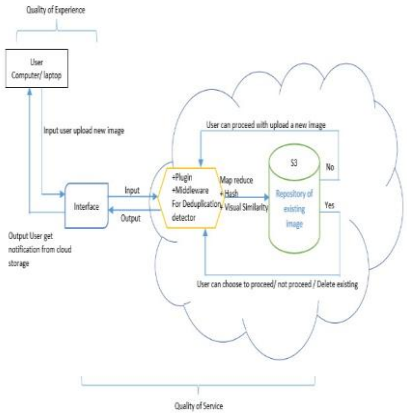

Fig. 3: Deduplication image detector framework.

5. Conclusion

Lastly, this research to support multimedia cloud computing concept to provide a better quality of services for user choose to using software and hardware as cloud and to achieve a better quality of experience using cloud storage (G. H. Jacob et. al. 2010).Image deduplication detection middleware in cloud is very important because, the middleware will be help user realize the existing deduplication image or data in their cloud database storage. Another that, the middleware will be help to reduce space for deduplication image in the cloud database.

6. Acknowledgement

This research was supported by research university grant (GUP-2015-008).Alhamdulillah and Thank to Ministry of Higher Education under MyPhd program for supporting my financial education at PHD level education. National University of Malaysia, especially Faculty of Information Science and Technology for support and allowing us to study at here. Thanks to HjYusof bin Ismail to support from the beginning and thanks to my supervisor to guide me write this article from the scratch and all the related person.

7. References

[1] Special Issue of IEEE Transaction on Multimedia “Multimedia: The Biggest Big Data”. www.signalprocessingsociety.org. 2015.

[2] Amazon Cloud Drive. Unlimited Cloud Storage from Amazon. www.amazon.com. 2015.

[3] L. Zhou, X. L. Zhou, L. Yu, B. YanLing, H. Luokai, S. Wenfeng. An Improved Image File Storage Method Using Data Deduplication. 13th International Conference on Trust, Security and Privacy in Computing and Communications. IEEE. 2014. Pp 638643.

[4] Microsoft Azure. https://azure.microsoft.com/enus/documentation/services/sql-database/. 2016.

[5] Y. Maret, F. Dufaux, and T. Ebrahimi. Image replica detection based on support vector classifier. In optical information system III, SPIE, Vol. 5909. 2005. pp. 173-181.

[6] G. Pratim, E.D. Gelasca, K.R. Ramakrishnan and B.S. Manjunath. Duplicate Image detection in Large Scale Databases. World Scientific. Volume-9.7x6.5. 2007. pp. 1-17.

[7] D. Wei, W. Zhe, C. Moses and L. Kai. HighConfidence Near-Duplicate Image Detection. ACM. 2012.

[8] S. Roy and E. C. Chang. A unified framework for resolving ambiguity in copy detection. In ACM Multimedia. 2005. pp. 648-655.

[9] K. Saehoon, W. Xin Jing, Z. Lei, C. Seungjin. Near Duplicate Image Discovery on One Billion Images. Winter conference on Application of Computer Vision. IEEE. 2015. Pp- 943-950.

[10] Arba'iahAbd Aziz. Othman Yatim. 2009. Falsafah di sebalik motif-motif songketMelayu Terengganu. Seminar AntarabangsaTenunan Nusantara: Kesinambungantradisidanbudaya. http://elib.uum.edu.my/kip/Record/um782898 [12/1/2013].

[11] G. H. Jacob, J. Eric. Lithium: Virtual Machine Storage for the Cloud. SoCC. 2010.