Efficient Indexing and Searching Framework for Unstructured Data

Kyar Nyo Aye, Ni Lar Thein
University of Computer Studies, Yangon
kyarnyoaye@gmail.com, nilarthein@gmail.com

ABSTRACT

The proliferation of unstructured data continues to grow within organizations of all types. This data growth has introduced the key question of how we effectively find and manage them in the growing sea of information. As a result, there has been an increasing demand for efficient search on them. Providing effective indexing and search on unstructured data is not a simple task. Unstructured data include documents, images, audio, video and so on. In this paper, we propose an efficient indexing and searching framework for unstructured data. In this framework, text-based and content-based approaches are incorporated for unstructured data retrieval. Our retrieval framework can support various types of queries and can accept multimedia examples and metadata-based documents. The aim of this paper is to use various features of multimedia data and to make content-based multimedia retrieval system more efficient.

Keywords: unstructured data, indexing, content-based retrieval.

1. INTRODUCTION

The Internet is a huge collection of data that is highly unstructured which makes it extremely difficult to search and retrieve valuable information. Due to the massive number of unstructured data, search engines that search and rank documents that contain unstructured data based on their relevance to user queries become essential for information seeking. Search engines are required to determine relevant documents within a short latency. In other words, high search efficiency is one of the key design and implementation objectives of search engines. Thus, efficient indexing techniques that organize documents according to their contents are demanded.

Unstructured data is a generic label for describing any corporate information that is not in a database. Unstructured data can be textual or non-textual. Textual unstructured data is generated in media like email messages, PowerPoint presentations, word documents, collaboration software and instant messages. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and flash video files. Unstructured information is typically text-heavy but may contain data such as dates, numbers, facts and multimedia data as well. Unstructured information accounts for more than 70%-80% of all data in organizations and is growing 10-50x more than structured data.

In the past decade, there has been rapid growth in the use of unstructured data especially digital media such as images, video, and audio. As the use of digital media increases, effective retrieval and management techniques become more important. Such techniques are required to facilitate the effective searching and browsing of large multimedia databases. Before the emergence of content-based retrieval, media was annotated with text, allowing the media to be accessed by text-based searching. Through textual description, media can be managed, based on the classification of subject or semantics. This hierarchical structure allows users to easily navigate and browse, and can search using standard boolean queries. However, with the emergence of massive multimedia databases, the traditional text-based search suffers from the following limitations:

1) Manual annotations require too much time and are expensive to implement. As the number of media in a database grows, the difficulty finding desired information increases. It becomes infeasible to manually annotate all attributes of the media content. Annotating a 60-minute video containing more than 100,000 images consumes a vast amount of time and expense.

2) Manual annotations fail to deal with the discrepancy of subjective perception. The phrase “a picture is worth a thousand words” implies that the textual description is not sufficient for depicting subjective perception. Capturing all concepts, thoughts, and feelings for the content of any media is almost impossible.

3) Some media contents are difficult to describe concretely in words. For example, a piece of melody without lyrics or an irregular organic shape cannot be expressed easily in textual form, but people expect to search media with similar

Fourth International Conference on Machine Vision (ICMV 2011): Machine Vision, Image Processing, and Pattern Analysis, edited by Zhu Zeng, Yuting Li, Proc. of SPIE Vol. 8349, 83493F · © 2012 SPIE
CCC code: 0277-786X/12/$18 · doi: 10.1117/12.921130

contents based on examples they provide. In an attempt to overcome these difficulties, content-based retrieval employs content information to automatically index data with minimal human intervention.

Content-based retrieval has been proposed by different communities for various applications. These include medical diagnosis, intellectual property, broadcasting archives, information searching on the Internet, etc [7]. The rest of the paper is organized as follows: Related works are reviewed in Section 2 and background theory such as desktop search engine and web search engine is explained in Section 3. In section 4, we introduce our proposed system architecture and then conclusion is described in section 5.

2. RELATED WORK

We survey some of existing systems dealing with searching multimedia data in various applications and systems that use Lucene for full text retrieval. We also focus on the systems capable of content-based search for multimedia data. Han Guo [1] proposed a framework of multimedia retrieval system basing on the MPEG-7 standard as information schema and discussed the annotation and feature extraction of multimedia data. In addition, a new method of metadata representation is proposed. Yuk Ying Chung [2] proposed and demonstrated a Content Based Multimedia Retrieval System (CBMRS). The proposed CBMRS includes both video and audio retrieval systems. The Content Based Video Retrieval System (CBVRS) based on DCT and clustering algorithms. The audio retrieval system based on Mel-Frequency Cepstral Coefficients (MFCCs), the Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) rule.

Yongbo Ma [3] proposed a framework of multimedia retrieval system which composed of three modules and fully rely on the MPEG-7 standard. This framework provides a rich set of automatic feature extraction components and an independent retrieval interface. Singhai et al. [4] surveyed content based image retrieval systems to provide an overview of the functionality of these systems. The techniques of CBIR are discussed, analyzed, compared and introduced the feature like neuro fuzzy technique, color histogram, texture and edge density for accurate and effective CBIR system. Chun Liu [5] designed a simple web Chinese full text retrieval system based Lucene using Struts2 MVC framework, and expounded the architecture of the web Chinese full text retrieval system and the implementation of the Chinese words segmentation module. YueHua Ding [6] developed paper duplication detection system to decrease duplicate excerpt rate based on Lucene.

3. BACKGROUND THEORY

A typical desktop search engine is displayed in Figure 1. It includes an indexer application that crawls existing and new stored files and extracts information on keywords, metadata, size and location in memory. This information is kept in an index file. Some systems use multiple indexes and indexers, to keep index files from getting too large to work with efficiently. When a user fills out a search form and sends a query, the engine searches the index, identifies the appropriate files, finds their locations on the hard drive, and displays the results.

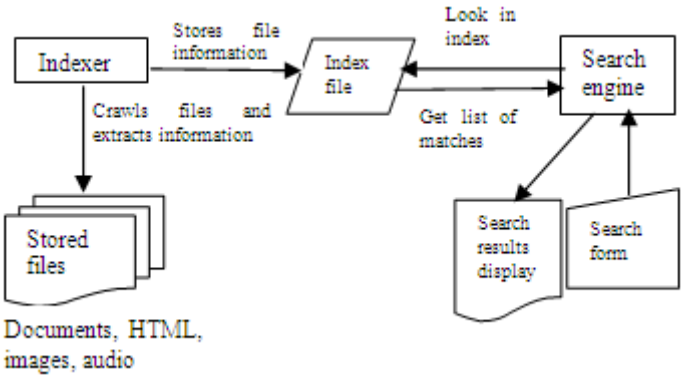

Figure 1. Typical desktop search engine

Figure 2 shows that the architecture of a common web search engine. It contains a front-end process and a back-end process. In the front-end process, the user enters the search words into the search engine interface, which is usually a web page with an input box. The application then parses the search request into a form that the search engine can

understand, and then the search engine executes the search operation on the index files. After ranking, the search engine interface returns the search results to the user. In the back-end process, a spider or robot fetches the web pages from the Internet, and then the indexing subsystem parses the Web pages and stores them into the index files.

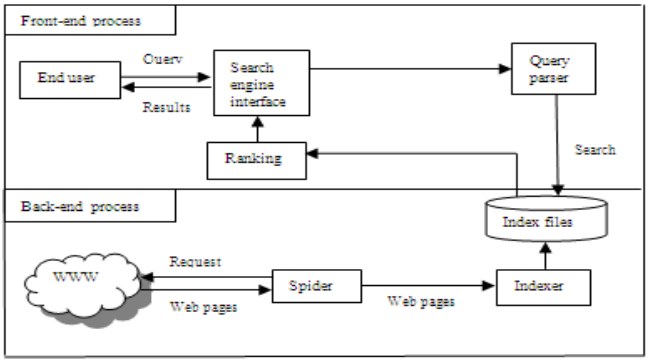

Figure 2. Web search engine architecture.

4. PROPOSED SYSTEM ARCHITECUTRE

Nowadays, efficient indexing and searching system for unstructured data is a challenge. In this paper, we propose efficient content-based unstructured data retrieval system. The proposed system incorporates full-text search approach for text data and content-based approach for audio, video and image data. Multimedia search engines can search not only text data but also image, audio and video data that is unstructured data. They use content-based multimedia retrieval approach that combines content-based text retrieval, content-based image retrieval, content-based audio retrieval and content-based video retrieval approaches for efficient indexing and searching. A functional model of multimedia search engine is shown in Figure 3.

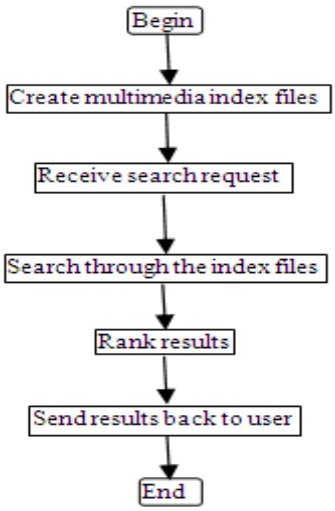

Figure 3. Functional model of multimedia search engine.

Figure 4 describes the proposed unstructured data indexing and retrieval framework.

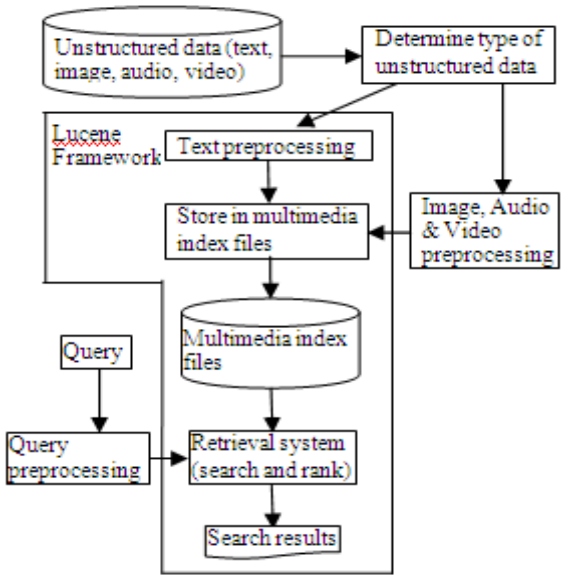

Figure 4. Unstructured data indexing and retrieval framework

First, the system determines the type of unstructured data. If the data is text, the text preprocessing operations are performed by a text preprocessing module.

4.1 Text preprocessing

Text preprocessing module is implemented by Lucene. Lucene is a high-performance, open source, full-featured text search engine library written entirely in Java. It is a technology suitable for adding full-text search function to nearly any application that requires it.

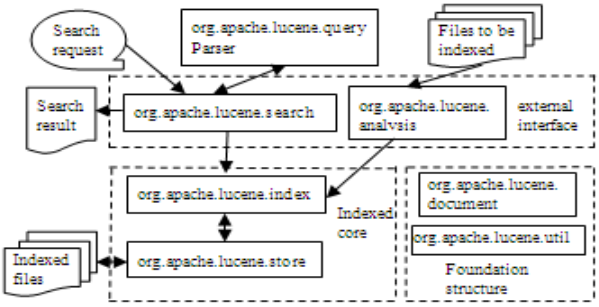

Figure 5. Lucene system architecture

4.2 Image preprocessing

For image data, image preprocessing steps are performed. They are:

1) Determine file type: First, the type of image is determined. File extensions or internal data can be used to determine image type.

2) Convert to standard image type: Once file type is known, it is converted into a standard internal representation. This involves conversion of file type to a standard file type, such as GIF. The size of the image can be standardized and reduced to a common size.

3) Feature extraction: The following image features are extracted to achieve efficient retrieval: 1. Conventional color histogram in RGB and HSV color space, 2. Invariant color histogram in RGB and HSV color space, 3. Fuzzy color histogram in RGB and HSV color space, 4. Auto color correlation feature, 5. MPEG-7 descriptors scalable color, color

layout and edge histogram, 6. Color and edge directivity descriptor, CEDD, 7. Fuzzy color and texture histogram, FCTH, 8. Color, texture and edge density, 9. The tamura texture features coarseness, contrast and directionality and 10. SIFT features.

Some of the above features are included in LIRe (Lucene Image Retrieval) framework. LIRe is a light weight open source java library for content based image retrieval. In this paper, we would like to further extend the functionality of LIRe by the integration of additional image features for effective and efficient retrieval.

4.3 Audio preprocessing

The objective of audio preprocessing module is to transform from audio data to text for easily accessible via lucene indexing mechanisms. Audio preprocessing steps are:

1) Determine audio data type: First, the type of the audio data is determined.

2) Convert to standard format: Next, the data is converted into a standard format. This includes the file format and normalization of the frequency range and sampling frequency.

3) Speech recognition: Third-party speech recognition software is used to recognize words in the audio data.

4.4 Video preprocessing

For video data, video preprocessing operations are executed by a video pre-processing module. Generally, video data is comprised of a running stream of three data types: image frames, audio data, and (optionally) closed caption text. Video preprocessing operations are: 1) determine video data type, 2) convert video to standard type, 3) extract and process audio data by audio preprocessing methods, 4) extract and process image by image preprocessing methods.

4.5 Query preprocessing

Query preprocessing involves text preprocessing and image preprocessing. User query can be metadata query, keyword query, exemplar image query and simple text query. Lucene can support various query types such as Boolean search, field search, wildcard search, fuzzy search, range search and so on. User can search images and videos by text and example image and can also search audio data via text. For example, user can view lecture videos and listen to lecture tracks by only knowing the lecture contents.

5. CONCLUSION

Although content-based multimedia retrieval has been studied for decades, most commercial search engines still rely on text information to index multimedia data. This is because of the many fundamental limitations in current content-based multimedia retrieval technologies when applied to large-scale multimedia data. We have proposed a framework for the content-based multimedia retrieval by combining various global and local features. This framework can be used for any type of unstructured data (text, image, audio and video) with a variety of queries (metadata, keyword, text and exemplar image query) to achieve satisfactory results. To evaluate this system's efficiency and effectiveness, recall and precision metrics can be used and compared with other multimedia search engines.

REFERENCES

[1] L. Chun,"Analysis and Research of Web Chinese Retrieval System Based Lucene," Proc IEEE, (2009)

[2] Y.Y. Chung and K.Y. Ng,"Design of a content-based multimedia retrieval system", Proc WSEAS, (2006)

[3] Y.H. Ding, K. Yi and R.H. Xiang,"Design of paper duplicate detection system based on Lucene," Proc IEEE, 36-39, (2010)

[4] H. Guo, C.Z. Ma, G. Liu, X.G. Dong, "The research about a content-based multimedia retrieval system," Proc IEEE, 148-151, (2010)

[5] Y.B. Ma, Z.Y. Fang, J. Liu and T.Y. Wang,"A content-based multimedia retrieval system base on MPEG-7 Metadata Schema," Proc IEEE, 1200-1201, (2008)

[6] N. Singhai and S.K. Shandilya,"A survey on: content based image retrieval systems," Computer Apps.4(2), (2010)

[7] http://encyclopedia .jrank.org/articles/pages/6567/Content-Based-Multimedia-Retrieval.html