Content-Based Multimedia Information Retrieval for Unstructured Data

Kyar Nyo Aye
University of Computer Studies, Yangon, Myanmar
kyarnyoaye@gmail.com

Abstract

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, multimedia search engines that search and rank them based on their relevance to user queries become essential for information seeking. In other words, high search efficiency is one of the key design and implementation objectives of multimedia search engines. In this paper, we propose efficient indexing and search system for unstructured data by combining text-based information retrieval and content-based information retrieval methods.

Keywords: unstructured data, text-based information retrieval, content-based information retrieval, multimedia search engine

1. Introduction

A large fraction of the data that will be stored and accessed in future systems is expected to be unstructured, in the form of images, audio files, video files, etc. All organizations are aware that a considerable amount of technical and business information and knowledge resides in both the structured databases and in unstructured repositories (e.g. documents, emails, etc). Unstructured information accounts for more than 70%-80% of all data in organizations and is growing 10-50x more than structured data. So, unstructured data processing is a very important emerging class of applications. There are a number of unstructured data processing applications that are already in use today. These applications include text searches, content-based searches of image, video, and audio files.

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. 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 of 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.
2) Manual annotations fail to deal with the discrepancy of subjective perception.
3) Some media contents are difficult to describe concretely in words.

Content-based methods are necessary when text annotations are nonexistent or incomplete.

Furthermore, content-based methods can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media collections. 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.

There was evidence of a growing synergy between traditional text-based and content-based retrieval techniques and the development of system that combined the two may yield better results. So, text- based information retrieval can be used for textual unstructured data and for non-textual unstructured data content-based multimedia information retrieval can be used. In this paper, we develop the general framework of content-based multimedia information retrieval for searching potentially large collections of unstructured data. The goal (and contribution) of this paper is to provide powerful search capabilities by using efficient indexing mechanisms depending on the type of unstructured data with the help of proposed content-based multimedia information retrieval framework. The rest of the paper is organized as follows: In section 2, we present related work and explain content-based multimedia information retrieval 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 text-based information retrieval and content-based multimedia information retrieval in various applications and systems that use Lucene for full text retrieval. Chun Liu [1] 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 [7] developed paper duplication detection system to decrease duplicate excerpt rate based on Lucene. Guan et al. [2] outlines several multimedia systems that utilize a multimodal approach. These systems include audiovisual based emotion recognition, image and video retrieval, and face and head tracking.

Lew et al. [3] presented challenges of content-based multimedia information retrieval such as semantic search, interactive search or relevance feedback system, multi-modal analysis and retrieval algorithms, evaluation with emphasis on representative test sets and usage patterns, etc. Lili et al. [4] surveyed the art of video query and retrieval and proposed a basic framework for video retrieval based on an iterated sequence of navigating, searching, browsing, and viewing. Singhai et al. [5] 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.

3. Background Theory

3.1. Content-Based Multimedia Information Retrieval (CBMIR)

Content-based multimedia information retrieval uses the contents of multimedia to represent and index the data. In typical content-based retrieval systems, the contents of the media in the database are extracted and described by multi-dimensional feature vectors, also called descriptors. The feature vectors of the media constitute a feature dataset. To retrieve desired data, users submit query examples to the retrieval system. The system then represents these examples with feature vectors. The distances (i.e., similarities) between the feature vectors of the query example and those of the media in the feature dataset are then computed and ranked. Retrieval is conducted by applying an indexing scheme to provide an efficient way to search the media database. Finally, the system ranks the search results and then returns the top search results which are the most similar to the query examples. A conceptual architecture for content-

based multimedia information retrieval is shown in figure 1.

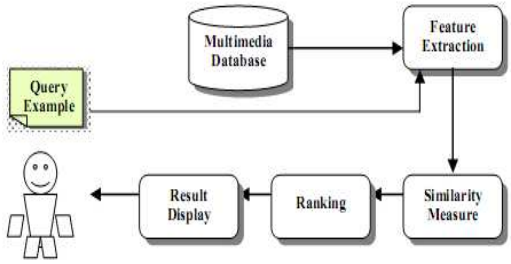

Figure 1. A conceptual architecture for content-based multimedia information retrieval

Content-based multimedia information retrieval consists of: 1) content-based image retrieval, 2) content-based audio retrieval, 3) content-based video retrieval and 4) content-based text retrieval (full-text retrieval).

3.1.1. Content-Based Image Retrieval (CBIR)

Content-based image retrieval, also known as query by image content and content-based visual information retrieval is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based means that the search makes use of the contents of the images themselves, rather than relying on human-input metadata such as captions or keywords. Figure 2 shows the block diagram of content-based image retrieval.

In CBIR each image that is stored in the database has its features extracted and compared to the features of the query image. It involves two steps:

1) Feature Extraction: The first step in this process is to extract the image features to a distinguishable extent.

2) Matching: The second step involves matching these features to yield a result that is visually similar.

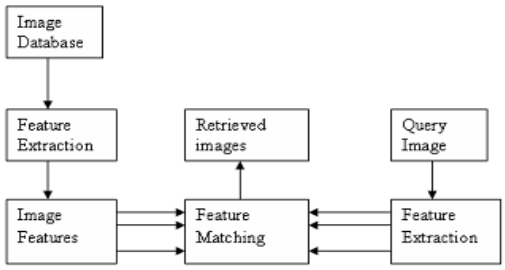

Figure 2. Block diagram of content-based image retrieval

3.1.2. Content-Based Audio Retrieval (CBAR)

In content-based audio retrieval, the goal is to find sound recordings (audio documents) based on their acoustic features. This content-based approach differs from retrieval approaches that index media files using metadata such as file names and user tags. Online audio content is available both isolated (e.g., sound effects recordings), and combined with other data (e.g., movie sound tracks). Earlier works on content-based retrieval of sounds focused on two main thrusts: classification of sounds to (usually a few high-level) categories, or retrieval of sounds by content-based similarity. For instance, people could use short snippets out of a music recording, to locate similar music. This “more-like-this” or “query-by-example” setting is based on defining a measure of similarity between two acoustic segments.

3.1.3. Content-Based Video Retrieval (CBVR)

Content-Based Video Retrieval (CBVR) systems appear like a natural extension (or merge) of Content-Based Image Retrieval (CBIR) and Content-Based Audio Retrieval (CBAR) systems. However, there are a number of factors that are ignored when dealing with images which should be dealt with when using videos. These factors are primarily related to the temporal information available from a video document. While these factors may complicate the querying system, they also may help in characterizing useful information for the querying.

A system that supports video content-based indexing and retrieval has, in general, two stages [6]. The first one, the Database population stage, performs the following tasks:
1) Video segmentation: Segment the video into its constituent shots,
2) Key frames selection: Select one frame or more to represent each shot, and
3) Feature extraction: Extract low-level and other features from key frames or their inter-relationships in order to represent these frames, hence shots.

The second stage, the retrieval subsystem processes the presented query (usually in the form of QBE), performs similarity matching operations, and finally displays results to the user.

3.1.4. Content-Based Text Retrieval (Full-Text Retrieval)

Full-text retrieval is a very popular technology in recent information search area. After information search technology had been developed for several years, it has been raised requirements on precision and speed. Database search technology based on SQL is ineffective on fuzzy searching because of its own structure features. Defect of database search technology is low efficiency. Full-text search technology resolves the efficiency problem. Full-text search scan every word of documents and create index, user search index with highly-speed and highly-efficiency [7].

4. Proposed System Architecture

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. Providing efficient indexing and search on unstructured data is not a simple task. A large number of search engines (e.g. Google, Altavista and Yahoo) support indexing and content-based retrieval of documents but only the textual information is taken into account. There is an emerging need for a new generation of search engines that try to exploit the full multimedia information. In this paper, we propose text-based information retrieval and content-based multimedia information retrieval for unstructured data to achieve efficient index and search. Figure 3 describes the proposed architecture of the unstructured data indexing and retrieval system.

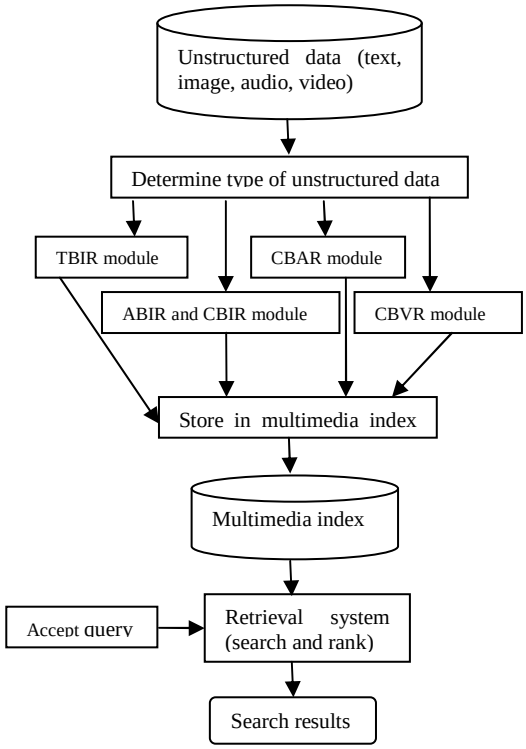

Figure 3. Architecture of unstructured data indexing and retrieval system

First, the system determines the type of unstructured data based on file extensions. Depending on the type of unstructured data, corresponding retrieval module can be used to retrieve indices as shown in figure 3.

4.1. Text Based Information Retrieval (TBIR) Module

In this system, text based information retrieval (TBIR) module is implemented by Lucene. It is an excellent technology framework of full-text retrieval engine, which is a basic

technology widely used in information retrieval field. Lucene is an tool package of open source code, which can be easily embedded into retrieval engine of application through expanding its function.

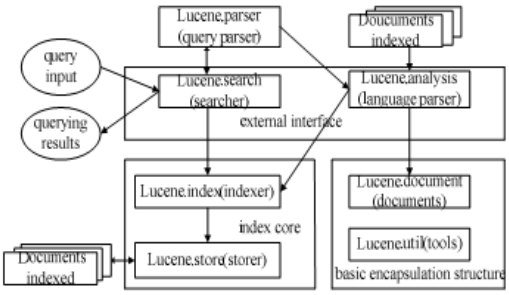

Figure 4. Lucene system architecture

4.2. Content- and Annotation-Based Image Retrieval Module

There are many queries for which visual similarity does not correlate strongly with human similarity judgments. This can lead to a semantic gap between user and machine. To minimize this gap, there are two solutions: 1) automatic annotation and 2) relevance feedback. Figure 5 describes the architecture of content- and annotation-based image retrieval module. First of all, all images in the database must be annotated with text caption to retrieve easily by text query. For each concept, the corresponding training images are converted to a general form by using low level features (i.e. color, edge and texture) and normalization. For "car" concept, the system uses the set of training images labeled with "car" to learn the model for the associated visual concept. When a query image is entered to retrieve similar images, the system extracts visual features from it, matches these features with the general forms for all concepts and then output similar images. This framework can be used to automatically annotate a new image with corresponding text caption and then this annotated image can be added into the database.

In this proposed system, we use color, texture and edge features to retrieve exact and user satisfied result. Feature extraction and representation consists of four stages. They are:

1) Color histogram: A color histogram for each image is defined, using a number of color bins for ranges of colors, each bin having a bin count and then stored as a feature vector for that image in multidimensional feature database.
2) Gray-scale histogram: A gray-scale, binned histogram of the image luminance is defined. For each pixel in the image, the intensity is determined and a bin count for the corresponding intensity bin is incremented. The gray-scale histogram is reduced to include and is stored in database.
3) Edge detection and edge histogram: Edges in the image are sharpened using standard edge detection algorithms. Then, for each edge fragment, the shape of the fragment along with neighboring fragments is histogrammed. The entire edge histogram is stored in database.
4) Texture detection and texture histogram: Textures in the image can be determined by scanning across the image for high frequency variations in the image, and histogramming the variations based on the size and frequency of the spikes, and then stored in database.

For feature matching or similarity matching, kd-tree based indexing and nearest neighbor searching technique is used.

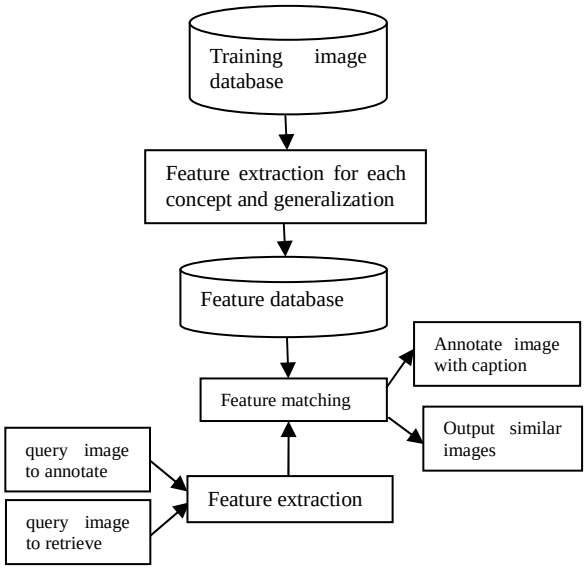

Figure 5. Architecture of content- and annotation-based image retrieval module

4.3. Content-Based Audio Retrieval (CBAR) Module

In content-based audio retrieval (CBAR) module, user can use text query for desired music and audio retrieval. To understand user entered text query, system needs to convert audio data to equivalent text data. This conversion consists of two steps:

1) Speech recognition: Third-party speech recognition software is used to recognize words in the audio data. These words are stored in database as an index.

2) Extract word alternatives: some speech recognition systems optionally supply alternates for recognized words, in case the word is improperly recognized. These alternates are also stored in multimedia index.

4.4. Content-Based Video Retrieval (CBVR) Module

Generally, video data is comprised of a running stream of three data types: image frames, audio data, and (optionally) closed caption text. Content-based video retrieval module consists of: 1) determine video data type, 2) convert video to standard type, 3) extract and process audio data by content-based audio retrieval method, 4) extract and process image by content-based image retrieval method.

In this system, user's query can be metadata query, keyword query, an exemplar image query and text query. Metadata and keyword query that can be entered to retrieve any type of unstructured data are text only, so Lucene can be used as full text search engine. User can enter an example image to search similar images or to annotate with text caption and simple text query to retrieve audio data.

5. Conclusion

We have proposed a system that combines content-based multimedia information retrieval and text based information retrieval for unstructured data indexing and searching. For textual unstructured data, Lucene is used as a full text search engine and content-based approaches are used for non textual unstructured data. This combined approach 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 user satisfied 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," in Proceedings of the first international workshop on Education Technology and Computer Science, IEEE Computer Society, 2009.

[2] L. Guan, "Multimedia multimodal methodologies" in Proceedings of International Conference on Mulimedia and Expo. IEEE Computer Society, 2009.

[3] M.S. Lew, N. Sebe and C. Djeraba, "Content-based multimedia information retrieval: state of the art and challenges," in ACM Transactions on Multimedia Computing Communications, and Applications, Feb. 2006.

[4] N.A. Lili, "Hidden markov model for content based video retrieval," in Proceedings of the 3rd Asia International Conference on Modelling & Simulation. IEEE Computer Society, 2009, pp. 353-358.

[5] N. Singhai and S.K. Shandilya, "A survery on: content based image retrieval systems," in International Journal of Computer Applications, vol. 4, no. 2, July 2010.

[6] W. Farag and H. Abdel-Wahab, Video Content-based Retrieval Techniques in Multimedia systems and Content-based Image Retrieval (Idea Group Inc., 2004).

[7] Y.H. Ding, K. Yi and R.H. Xiang, "Design of paper duplicate detection system based on Lucene," in Proceedings of Asia-Pacific Conference on Wearable Computing Systems. IEEE Computer Society, 2010, pp, 36-39.