ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Keyword Based Efficient Retrieval for Documents and Images

Twinkle Pardeshi, Prof. P. V. Kulkarni

Department of Computer Engineering, Maharashtra Institute of Technology, Pune, India

ABSTRACT: Due to continuous improvement in data storage and image acquisition technologies, large number of image datasets are been developed. Due to which needs are increasing day by day to improve the retrieval quality of the systems. There are mostly two approach to deal with image dataset. One is Text based image retrieval and other is content based image retrieval (CBIR). The goal is to support image retrieval based on content properties (e.g., shape, color, texture), usually encoded into feature vectors. One of the main advantages of the CBIR approach is the possibility of an automatic retrieval process, instead of the traditional keyword-based approach, which usually requires very laborious and time-consuming previous annotation of database images. The CBIR technology has been used in several applications such as fingerprint identification, biodiversity information systems, digital libraries, crime prevention, medicine, historical research, among others.

KEYWORDS: Image retrieval, image database, image descriptors, indexing, content-based image retrieval

I. INTRODUCTION

Data mining is to extract or mine knowledge from a lot of data called Knowledge Discovery in Databases (KDD), which is the result of information technology natural which is the result of information technology natural evolution. In past years, the data mining technology produced great attention among the information industry. It is an inter-discipline subject, influenced by multiple disciplines. At present, according to the various types of mining knowledge and mining the different objects, many data mining methods and special tools are available. There are three main components in a search engine known as Crawler, Indexer and Ranking mechanism. The Crawler is also called as a robot or spider that navigates the web and downloads the web pages. The pages that are downloaded are being transferred to an indexing module that parses the web pages and erect the index based on the keywords in individual pages. An index is normally sustaining using the keywords. When a query is being drifted by a user, it means the query transferred in terms of keywords on the interface of a search engine, the query mainframe section examine the query keywords with the index and precedes the URLs of the pages to the user. But before presenting the pages to the client, a ranking mechanism is completed by the search engines to present the most relevant pages at the top and less significant ones at the substructure. It makes the search outcomes routing easier for the user.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Algorithms	Main Technique	Input Parameters	Methodology	Limitations
Page Rank	Web Structure Mining	Back Links	Computesthe page score at thetime of indexing of the Pages.	Results are at the time of indexing and Not at the query time.
Weighted Page Rank	Web Structure Mining	Back links and Forward links.	Weight of web page is calculated on the basis of input and output linksand on the basis of weight the importance of page is decided.	Relevancy is ignored.
Distance Rank	Web Structure Mining	Forward links	Based on reinforcement Learning.	If new page inserted between two pages then the crawler should perform a large calculation to calculate the distance vector.
Time Rank	Web Usages Mining	Original Page Rank and Sever Log	Here the visiting time is added tothe score of the original page rank of that page.	Important pages are ignored because it increases the rank of those web pages which are opened for long duration
Relational Based Page Rank	Web Structure Mining	Keywords	A semantic search engine would take into account keywords and would return page only if both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.	Here every page is to be annotated with respect to some Ontology, which is the very tough task.

Table 1. Comparison of various algorithms

II. IMAGE MINING

Image mining is a technique which handles the mining of information, image data association, or additional patterns not unambiguously stored in the images. It utilizes methods from computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence. The main intention of image mining is to produce all considerable patterns without any information of the image content, the patterns types are different. They could be classification patterns, description patterns, correlation patterns, temporal patterns and spatial patterns. Image mining

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

handles with all features of huge image databases which comprises of indexing methods, image storages, and image retrieval, all regarding in an image mining system.
The shortcomings of keyword search have been addressed by the so-called Content-Based Image Retrieval (CBIR) systems.In such systems, image processing algorithms (mostly automatic)are used for extraction of feature vectors that deals with representation ofimage properties such as color, texture, and shape. In this technique, it is possible to retrieve images similar to one chosen by the user's query image. One of the main advantages of this approach is the possibility of an automatic retrieval process, contrasting to the effort needed to annotate images. It requires the translation of high-level user perceptionsinto low-level image features. To index visual features, it is common to use numerical values for the n features and then to represent the image or object as a point in an-dimensional space. Multi-dimensional indexing techniques and common similarity metrics are factorsto be taken into account.Set of visual marks: e.g., points, lines, areas,volumes, and glyphsset of visual properties: e.g., position, size, length, angle, slope, color, gray scale, texture, shape, animation, blink, and motion.

III. IMAGE DESCRIPTOR:

Image Descriptors = feature vector extraction function + distance function

- Color Descriptors

Due to its simplicity the RGB color space is most commonly used. It is represented by red (R), green (G), and blue (B) chromaticity. The final color is defined by the additive combination of those three primary colors. The Hue, Saturation, and Value (HSV) color space separates the intensity from the chromaticity and represents them independently. Hue describes the position of the color in a 360◦ spectrum. Saturation describes the pureness of the color: it measures the difference between the color and a grayscale value of equal intensity. Value, as the third channel, is the measurement of brightness. Measurement of brightness. The CIE L*a*b* and CIE L*u*v* spaces are selected to represent a uniform color space. These two color spaces are derived from the CIE XYZ color space and attempt to produce a coordinate system in which perceptual distances correspond to Euclidean distances. In CIE L*a*b* color space, L* represents the lightness of color going from 0 (dark) to 100 (white), while a* and b* channels are the two chromatic components. The first of these two (a*) represents the colors position between red/magenta (+a) and green (-a). Similarly, b* indicates its position between yellow (+b) and blue (-b). In practice, their range goes from −128 to 127 with 256 levels. Similar to L*a*b*, the CIE L*u*v* color space has one lightness channel and two chrominance components referring to the same chrominance.The opponent color space has been claimed to give better performance in several image processing tasks. In this space, two channels, O1 and O2 , are used to store the red-green and blue-yellow opponent pairs, while the O3 channel is equal to the intensity channel in the HSV color space.

- **Texture Descriptors**

  There is no widely accepted definition of texture. However, this image property can be characterized by the existence of basic primitives, whose spatial distribution creates some visual patterns defined in terms of granularity, directionality, and repetitiveness. There exists different approaches to extract and represent textures. They can be classified into space-based, frequency-based models, and texture signatures

- **Shape Descriptors**

  In pattern recognition and related areas, shape is an important characteristic to identify and distinguish objects Shape descriptors are classified into boundary-based (or contour-based) and region based methods. This classification takes into account whether shape features are extractedfrom the contour only or from the whole shape region. These two classes, in turn, canbe divided into structural (local) and global descriptors. This subdivision is based on whetherthe shape is represented as a whole or represented by segments/sections.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

III. CONTENT BASED IMAGE SEARCH

- Usually, two kinds of queries are supported by CBIR systems:

1. K-nearest neighbor query (KNNQ):
   The user specifies the number k of images to be retrieved that are closest to the query pattern.

2. Range query (RQ)
   The user defines a search radius r and wants to retrieve all database images whose distances to the query pattern are less than r. In this case, both the specification of k in the KNNQ and the specification of r needs to be incorporated into Q.

- Several CBIR systems have been proposed recently

1 Chabot integrates image content retrieving based on color information with textbasedqueries. Its interface allows user to search and update the image database. This systemdoes not include texture and shape descriptors.

2 The QBIC (Query by Image Content) system was developed by IBM. QBIC uses color, shape, and texture to retrieve image databases. Query specification follows the queryby-example paradigm. A user can sketch a shape, select colors, indicate color distributions,or pre-defined textures.

IV. CONCLUSION

This paper has presented a brief overview of efficient retrieval of documents and images and various techniques for the same. Also Content based image retrieval and text based image retrieval are considered as a main components for image retrieval. Various existing algorithms and systems are been discussed in the paper.

REFERENCES

1. Vishwakarma Singh, Bo Zong, and Ambuj K. Singh, "Nearest Keyword Set Search in Multi-Dimensional Dataset" ,in IEEE Transactions on knowledge and Data Engineering, vol. 28, no. 3, march 2016.
2. W. Li and C. X. Chen,"Efficient data modeling and querying system for multi-dimensional spatial data", in Proc. 16thACMSIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., 2008, pp. 58:1-58:4.
3. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", in Proc. 20th Int. Conf. Very Large Databases, 1994, pp. 487-499.
4. N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger,"The R*- tree: An efficient and robust access method for pointsand rectangles", in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1990, pp. 322-331.
5. D. Zhang, B. C. Ooi, and A. K. H. Tung,"Locating mapped resources in web 2.0", in Proc. IEEE 26th Int. Conf. Data Eng., 2010, pp. 521-5.
6. V. Singh, S. Venkatesha, and A. K. Singh,"Geo-clustering of images with missing geotags", in Proc. IEEE Int. Conf. Granular Comput., 2010, pp. 420-425.
7. G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects", Proc. VLDB Endowment, vol. 2, pp. 337-348, 2009.
8. D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa, "Keyword search in spatial databases: Towards searching by document", in Proc. IEEE 25th Int. Conf. Data Eng., 2009, pp. 688-699.
9. X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, Collective spatial keyword querying, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 373384.