RA Journal of Applied Research
||Volume||1||Issue||1||Pages-13-17|| Feb. -2015|| ISSN (e): Applied
www.rajournals.in

Content Based Image Searching Using Focused Crawler

Authors
Ayush Agrawal[1], Kanchan Hans[2]
[1] Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India
[2] Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India
Email- ayusha98@gmail.com, khans@amity.edu

ABSTRACT

In today's world, the metadata of the image is looked up by image search engine, when a search query is performed. However some search engines can identify a limited range of visual content also e.g. faces, tree, sky, buildings etc. In this paper, we describe how to enhance the capabilities of an image search engine by applying content based image searching.

The search query is matched with the text contained in the image itself along with the metadata of the image. Focused crawler is used in the process. This results in providing more accurate results, matching with search query. Later, the text contained in the image can be added to the index of the image along with the metadata which reduces the searching time occupied by text extraction algorithms.

INTRODUCTION

Content based image searching involves text extraction from images and use crawling approach through extracted text. In this paper, we are using a two-step process that involves focused crawling and text extraction from crawled images. Text extraction is done in only specific domain to improve the performance of crawling process and save lot of time as well. This is achieved by using focused crawler. A focused crawler searches in specific domains only and text extraction process takes place for crawled pages one by one. User input's the text document or textual query and these documents are in form of character object. These all documents are read for finding the occurrence of n number of consecutive characters. Textual documents or queries are based on document vector dot production of similar text. We have document image features for extraction and these are two features Vertical Traverse Density (VTD) & Horizontal Traverse Density (HTD). Now test the image textual documents for text similarities and retrieve text document. Extracted text is stored along with the metadata of image to improve the further crawling process for same query search.

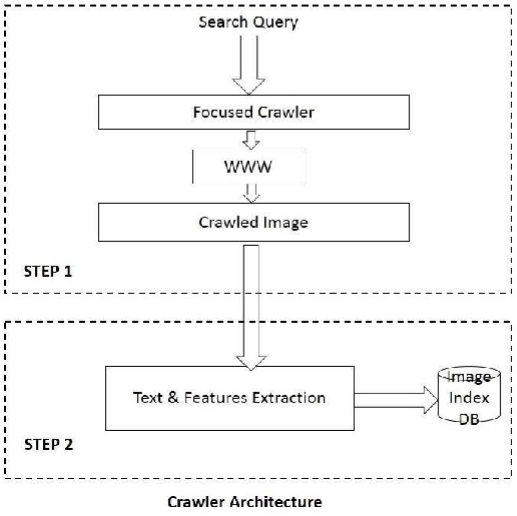
IMPLEMENTATION

The widespread adoption of the World-Wide Web has created challenges both for society as a whole and for the technology used to build and maintain the Web. Looking at societal level, the Web is expanding faster than we can comprehend its implications or develop rules for its use. The

RA Journal of Applied Research
||Volume||1||Issue||1||Pages-13-17|| Feb. -2015|| ISSN (e): Applied
www.rajournals.in

ubiquitous use of World Wide Web has raised important social concerns in the areas of privacy, censorship, and access to information. This dissertation describes WebCrawler, the Web's first comprehensive full-text search engine. WebCrawler has played a vital role in making the Web easier to use for a lot of people. Its invention and subsequent evolution, helped fuel Web's huge growth by creating a new way of navigating hypertext: searching. [8]

Content based Image searching process can be more efficient if image features along with the text contained in the image are part of the searching process and the use of focused crawler will make the search process faster. Focused crawler crawl relevant pages and reduces crawling time [2].



Crawler Architecture

Focused Crawler is designed in such a way that it seeks, acquires, indexes, and maintains pages only on a specific set of topics that represent a relatively narrow segment of the Web. It involves a very little investment in hardware and network resources and achieves a very respectable coverage at a rapid rate, as because there is relatively little to do as compared to without using of focused crawler. Thus, Web content can be managed by a focused crawler, specializing in one topic. A focused crawler will be far more nimble in detecting changes to pages within its focus than a crawler that is crawling the entire Web. The focused crawler uses a classifier which learns to recognize relevance from examples embedded in topic taxonomy, and also a distiller which identifies topical vantage points on the Web. [6] The crawler is designed to crawl entire web repeatedly, keeping a local copy up to 1 MB of the text of each page, plus metadata, in a repository, which can later be used for indexing, mining, etc. Also, this crawler is incremental, i.e., the repository copy of each page is updated as soon as the actual web page is crawled. Also, using this technique, the repository keeping local copy must always be out of date to some (we hope minimal) extent.[9]

Preprocessing of the imaged documents is achieved involving de-skew, noise removal and layout analysis to remove headlines and pictures or photographs leaving finally the main text body of an article to be processed which is typically of one predominant font type and size. Connected component analysis is then performed to identify

RA Journal of Applied Research
||Volume||1||Issue||1||Pages-13-17|| Feb. -2015|| ISSN (e): Applied
www.rajournals.in

character objects. There are three kinds of character objects. First kind characters have only one connected component. The second kind refers to characters having more than one connected component, such as characters "i" and "j". The third kind consists of characters that are connected to each other, such as "ft" and "ff".

Based on horizontal projections, we divide the document image into many rectangular zones each denoting one text line. Here, connected components in different zones belong to different character objects. Hence, these components can be expressed as Ci,j, where i is the text line number and j is the sequence number of connected components in each zone from left to right. Within the same text line, if the horizontal overlapping extent in Ci,j and Ci,j+1 is larger than a threshold, then they belong to the same character object. Otherwise, they belong to different objects. Punctuation marks such as commas and periods are not considered significant in similarity measure and are thus removed during the preprocessing. [1]

For each character object, we use two vectors to store the object features: Horizontal Traverse Density (HTD) and Vertical Traverse Density (VTD). HTD is a vector whose elements denote the numbers of line segments as we scan the character horizontally line by line from top to bottom. And, VTD is another vector obtained from vertical scanning from left to right. One class set of character objects can be obtained from each imaged document.

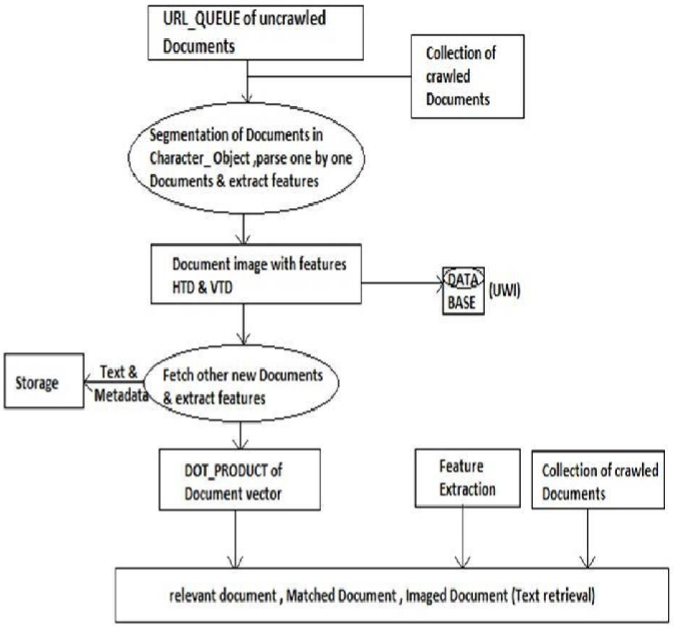


Image features (along with text and metadata) are stored in the storage as image metadata information which can be used in further searches to reduce searching time for same query and produce faster results.

RELATED WORK

The goal of a focused crawler is to find many pages of interest without using a lot of bandwidth. Hence saving a lot of bandwidth usage [10]

Method proposed in [1] for text retrieval is distinct from OCR technology, in the sense that we create ngrams using simple image features to achieve

RA Journal of Applied Research
||Volume||1||Issue||1||Pages-13-17||Feb. -2015|| ISSN (e): Applied
www.rajournals.in

effective retrieval of multilingual documents. OCR technology is based on language dependency, so it becomes time consuming and also multiple OCR's are required for various documents. This method is robust in case of degraded documents and also provides other features like changing fonts etc. while OCR technique will take more cost in case of large or different documents because degraded text and simulate OCR output for required results.

Following diagram explains the approach to achieve text retrieval from images without the use of OCR.

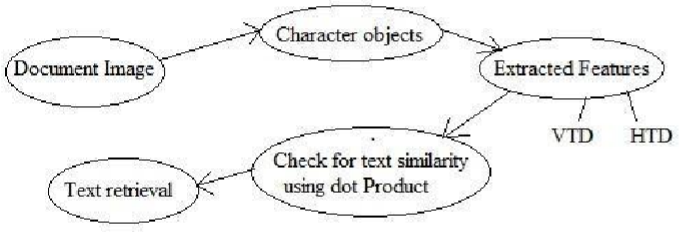


IMAGE FEATURES

Extracted text is stored along with the metadata of the image. The text is stored in database that keeps all image features, visual content of the image, including texture, color, and shape features, as the image index. Hence extracted text is treated as a feature of the image. So that, one image is processed only once for text extraction process and further searches can pick up the text from index database.

CONCLUSION

A concept has been proposed in the research paper which can be used to build a new web crawler. Such crawler will search for images much faster and provide more accurate results. Focused crawler searches in specific domain and provides only relevant data, thus resulting in fast searching process. Content based image searching using text will help in finding more accurate search results. Image searching results will be more accurate. People will be able to search using image features along with text contained in the image itself.

The text-based approach is also significantly more efficient than the CBIR system on the Internet, in terms of computational cost as well as image transmission and storage cost as CBIR systems only uses the visual content of the images, such as texture, color, and shape features, as the image index. [16]

REREFENCES

[1] Chew Lim Tan, Weihua Huang, Zhaohui Yu, Yi Xu : Imaged Document Text Retrieval without OCR

[2] Dongming Jiang, Arvind Krishnamurthy, Jaswinder Pal Singh, Randolph Wang : Method For Apparatus For Focused Crawling (US 7,080,073 B1)

[3] Junghoo Cho and Sougata Mukherjea : Crawling for Images

[4] Bo Luo, Xiaogang Wang, and Xiaoou Tang : A World Wide Web Based Image Search Engine Using Text and Image Content Features

[5] Soumen Chakrabarti, Martin van den Berg, Byron Dom : Focused crawling: a new approach to topic-specific Web resource discovery

[6] Ka-Ping Yee, Kirsten Swearingen, Kevin Li,Marti Hearst : Faceted Metadata for Image Search and Browsing

[7] Marc Najork : Web Crawler Architecture

[8] Brian Pinkerton : Finding What People Want: Experiences with the WebCrawler

[9] Jenny Edwards, Kevin McCurley, John Tomlin : An Adaptive Model for Optimizing Performance of an Incremental Web Crawler

[10] Vladislav Shkapenyuk, Torsten Suel : Design and Implementation of a High-Performance Distributed Web Crawler

[11] Prasant Singh Yadav, Mrs Mala Kalra, Dr. K.P Yadav : Enhancing the performance of web Focused CRAWLer using ontology

[12] Yanni Li, Yuping Wang, Jintao Du : E-FFC: an enhanced form-focused crawler for domain-specific deep web databases

[13] MARC NAJORK : Web Crawler Architecture

[14] Serge Belongie, Chad Carson, Hayit Greenspan and Jitendra Malik : Color- and Texture-Based Image Segmentation Using EM and Its Application to Content-Based Image Retrieval

[15] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra : Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval

[16] Stan Sclaroff, Leonid Taycher, and Marco La Cascia : ImageRover: A Content-Based Image Browser for the World Wide Web

[17] Yihong Gong, Hongjiang Zhang, Chuan, H.C., Sakauchi, M. : An image database system with content capturing and fast image indexing abilities.