International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13471-13475
© Research India Publications. http://www.ripublication.com

A Novel Approach for Accurate Content Based Search Using Hash Techniques

Srisailapu D Vara Prasad [1] & Dr. K. Rajasekhara Rao [2]

[1]Research Scholar, Department of Computer Science & Engineering, University College of Engineering & Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh 522510, India & Assistant Professor, Department of Computer Science & Engineering, School of Technology, GITAM (Deemed to be University), Hyderabad, Telangana-502329, India.

[2] Professor, Department of Computer Science & Engineering & Director URCE, Vijayawada, Andhra Pradesh-521109, India.

[1]Orcid: 0000-0002- 5112-5909

Abstract

Considering the indigenous digital data designs to new, highly refined and massive amount of binary data, the digital data always attempts to take parts of physical world and reproduce them for technological use. Most of this data isn't in text form, e.g. images, sounds and video. The utility of this data would be immensely rised by good tools for searching and organizing it. However, presently available search tools are inherently text based and depend upon text annotations for such feature-rich datasets. Conducting content-based similarity search on such high dimensional data could be a difficult drawback. This paper involves economical general-purpose ways to go looking and index such datasets, building on recent theoretical work on constructing sketches - compact representations for data. These procedures are enforced during a toolkit modeled for assisting system builders quickly construct content-based similarity search systems for a huge datasets. This toolkit, some of the search and indexing technology it incorporates, and the main goal involves in it to quickly construct content-based similarity search systems for various types of feature-rich data.

INTRODUCTION

Digital data volume increased at an exceptional rate from the past few decades. The "Moore's law curve" (doubling each eighteen months) currently doesn't refers only to the exponential improvement rate of processor performance, storage density and network information measure, but to boot to the data growth permission to create digital or hard copies of all or a region of this work for personal or classroom use is granted without any fee on condition that copies do not appear to be created or distributed for profit or industrial advantage which copies bear this notice and additionally the complete citation on the first page. On the other hand to copy, to republish, to post on servers or to redistribute to lists, desires prior specific permission and/or a fee rates of the various disciplines. The main data types are peculiar like audio, digital photos, videos, and other sensor data. Stepping into a digital society where all information is digitized and where the world is interconnected by digital means, it's extremely fascinating for following-generation systems to produce users with the potential to access, search, explore and manage feature-rich information. A crucial challenge for implementation of a content-based similarity search system for feature-rich data is that such the data is noisy and complicated to handle. For example, consider two different pictures that look alike, or two separate recordings of someone speaking a similar sentence. In spite of that, the high degree of similarity among the two pictures or between the audio recordings, the digital representations completely have variance at the bit level. Comparison of noisy, feature-rich data needs similarity search rather than precise match but the point is similarity search in high-dimensional areas is notoriously tough (the thus referred to as curse of dimensionality). Therefore, practical advanced search solutions like database tools and search engines (e.g. Google), are restricted to find out precise matches and have a tendency to build just for text documents and text annotations.

This paper presents a toolkit designed to help system builders quickly construct content-based similarity search systems for varied varieties of feature-rich information. Building a toolkit as the common software system infrastructure that resolves the core content-based similarity search issues, system implementers will simply add new content-based search capabilities by developing and plugging in data-type specific segmentation and also have feature extraction modules. This permits implementers to separate their design considerations for specific data types from the core capability of content-based similarity search. By using this toolkit, a computer engineer can plugin her image segmentation, feature extraction, and distance perform modules to build a content-based similarity search system without concerning the way to deal with metadata management, filtering, indexing and ranking. A genomic scientist might use the toolkit as her research platform to experiment in conjunction with her new

distance functions for locating the similar genes. The toolkit for content-based similarity search of feature-rich data will perform the same role as that of the text-based search engines in next-generation operating systems.

RELATED WORK

Cai, Deng, et al. 2004[1] remember the difficulty of clustering Web image search outcomes. Generally, the image search results go back by using a image search engine that include more topics. Organizing the results into completely distinctive semantic clusters allows customers' browsing. This paper, have a tendency to recommend a hierarchical clustering technique with the usage of visual, textual and link evaluation. By employing a vision-based page segmentation algorithm, a web page is partitioned into blocks, and also the textual and link records of an image will be accurately extracted from the block containing that image. By use of block-degree link analysis techniques, an image graph may be created. Then apply spectral techniques obtain a Euclidean embedding of the image that follows the graph shape. Hence for each image, there may be three types of representations and they are Visual feature based example, textual based illustration and graph based illustration. By spectral clustering strategies, we're able to cluster the search results into completely one-of-a-kind semantic clusters. An image search example depicts the capacity of these strategies.

Zamir, Oren, and Oren Etzioni.,1999 [2] Clients of web search engines are generally compelled to filter through the long ordered list of document "snippets" returned by the engines. The IR group has investigated document clustering as a substitute technique of sorting out recovery results, however cluster must be conveyed on most major web search tools. The Northern Light search engine organizes its output into "custom folders" based on pre-computed report labels, however would not screen how the folders are generated or how well they correspond to users' interests. This involves in introducing Grouper – an interface to the outcomes of the Husky Search meta-search engine that dynamically combines the search results into clusters labeled through terms extracted from the snippets. By analyzing Husky Search logs, it is confirmed that massive variations inside the range of files observed, and in the quantity of effort and time expended with the aid of customers having access to seek outcomes via those interfaces.

Kustanowitz, Jack, Ben Shneiderman, and Bill Kules[3], When search results against digital libraries and internet resources have restricted metadata, augmenting them with purposeful and stable class facts can permit better overviews and support consumer exploration. This paper proposes six "fast-feature" techniques that use only options supplied within the search result list, like title, snippet, and URL, to divide outcomes into purposeful categories. They make use of credible knowledge resources, together with a US government organizational hierarchy, a thematic hierarchy from the Open Directory Project (ODP) web directory, and private browse histories, to feature valuable metadata to search results. In three checks the percentage of consequences labeled for five consultant queries was sufficient to recommend practical benefits: general web search (76-90%), government web search (39-100%), and also the Bureau of Labor Statistics web site (48-94%). One more test submitted 250 TREC queries to search engine and with success classify 66% of the top 100 using the ODP and 61% of the highest 350. Fast-feature techniques are enforced in a prototype search engine. Hence proposed directions to expand categorization rates and build guidelines regarding how web site designers may re-arrange their sites to assist speedy categorization of search results.

Ahmed, F. & Siyal, M. Y[4], Authentication of image using hash function is really a typical task because various core issues like tamper detection, security and robustness needs have to be included. In this paper, it was proposed that a hash-based image authentication scheme addresses these core issues simultaneously. When compared to other schemes it was introduced that a secret key is used to randomly modulate image pixels for creating a transformed feature space. Now the image hash is calculated with this key-dependant transformed feature space. A 4-bit quantization scheme is proposed in order to reduce the size of hash. The experimental results reported reveals that the proposed scheme offers good robustness against JPEG compression, low-pass and high-pass filtering. The proposed scheme can also detect minute tampering with localization of tampered area. The receiver operating curve (ROC) and security analysis presented in this work makes the proposed technique a candidate for practical digital image signature systems where the transmitted or stored image might undergo JPEG compression, low-pass or high-pass filtering.

Furht, B., Socek, D. & Eskicioglu, A. M[5], The recent advances in the technology, especially in the computer industry and communications, allowed a potentially enormous market for distributing digital multimedia content through the Internet. With the advancement of technology in computer industry and communications led to the establishment of huge market for distribution of digital multimedia content through internet. However, the proliferation of digital documents, multimedia processing tools, and the worldwide availability of Internet access have created an ideal medium for copyright fraud and uncontrollable distribution of multimedia content. The major challenge today is protection of intellectual property of multimedia content in multimedia networks. In such a case two multimedia security technologies are being developed and they are multimedia encryption and multimedia water marking technology.

Belkin, Mikhail, and Partha Niyogi [6], Compared to

traditional document retrieval a web page isn't a good information unit as it contains irrelevant data from navigation, decoration, and interaction part of page. Compared with simple DOM based segmentation method, our page segmentation scheme utilizes useful visual cues to obtain a better partition of a page at the semantic level. This method of page segmentation scheme utilizes useful visual cues for obtaining better partition of page at semantic level compared to Vision based Page Segmentation. By using our VIPS algorithm to assist the selection of query expansion terms in pseudo-relevance feedback in web information retrieval, this method achieves 27% performance improvement on Web Track dataset.

Manjunath, Bangalore S., and Wei-Ying Ma [7], Image content based retrieval is playing a leading role in research area with application in digital libraries and multimedia databases. This focuses on image processing aspects and texture information for browsing as well as retrieval of large image of data. The use of Gabor wavelet features for texture analysis and provide a comprehensive experimental evaluation was proposed. Compared to other multi resolution texture features using the Brodatz texture database it was identified that Gabor features provide the best pattern retrieval accuracy An application to browsing large air photos is illustrated.

Xiaofei, Wei-Ying Ma, and Hongjiang Zhang [8], in correspondence between spectral clustering, spectral dimensionality reduction, and the connections to the Markov chain theory, using spectral techniques it was presented a novel unified framework for structural analysis of image database. The framework provides a computationally efficient approach to both clustering and dimensionality reduction, or 2-D visualization. In this framework, it was also obtained that the semantic degrees of the images, i.e. image rank, which are used to characterize the richness of semantics contained in the images.

Cai, Deng, et al [9], Visual representation of a new web content structure analysis was proposed. Using this structure many web applications such as information retrieval, information extraction and automatic page adaptation will be benefited. An automatic top-down, tag-tree independent approach to detect web content structure was proposed. It simulates how a user understands web layout structure based on his visual perception. When compared to other existing techniques this approach is independent because HTML documentation representation works good even though it is different from layout structure. The experiments show well satisfactory results.

There is a very big demand for effective and efficient method of organizing and retrieving the images with rapid growth of number of digital images on web. A method for clustering and embedding WWW images are discussed here. With the use of vision-based page segmentation algorithm the Web page is partitioned into blocks, and then the textual and link information of an image can be accurately extracted from the block containing that image. We can construct an image graph by the process of extracting the page-to-block, block-to-image, block-to-page relationships through a link structure and page layout analysis. We use techniques from spectral graph theory for image clustering and embedding along with image graph model[10].

The network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. we developed a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of context on the World Wide Web. The central issue we tend to address among this framework is that the distillation of broad search topics, through the invention of "authorative" data sources on such topics. Perform to test an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of "hub pages" that join them together in the link structure. The formulation has connections to the eigenvectors of certain matrices associated with the link graph while these connections in turn motivate additional heuristics for link-based analysis[11].

METHODOLOGY

Robust and Discriminative Image Perceptual Hashing

A perceptual image hashing system, as shown in Figure. 1, contains four pipeline stages. They are the Transformation stage, the Feature extraction stage, the Quantization stage and the Compression and Encryption stage. In the Transformation stage, the input image experience special and/or frequency transformation so as to create all extracted options rely either on the values of image pixels or on the image frequency coefficients. Feature Extraction stage allows the perceptual image hashing system to extract the image options from the input image and obtain the continuous hash vector. Once the continuous perceptual hash vector is obtained it is then quantized into the discrete hash vector within the Quantization stage. Now the discrete hash vector is converted to binary perceptual hash string in third stage. Finally, the binary perceptual hash string is compressed and encrypted into a brief and a final perceptual hash in the Compression and Encryption stage.

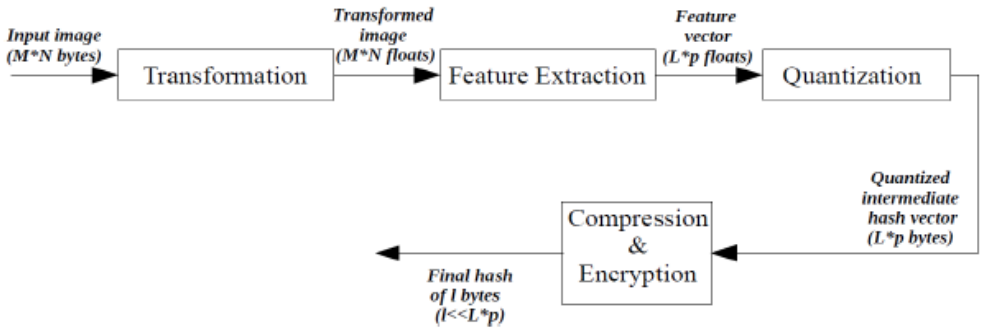

Figure 1: Four pipeline stages of a perceptual image hashing system.

DCT Based Hash

The DCT, similar to Fourier-related transform, expresses a function or signal (a sequence of numerous finite data points) in terms of a sum of sinusoids with different frequencies and amplitudes. The DCT uses only cosine functions whereas the discrete Fourier transform (DFT) uses both cosines and sines. DCT comprises of eight different standard variations, of which the common variant is the type-II DCT. Hence, it is simply referred to as DCT.

Let us consider x[m], m = 0,…N - 1, denote an N-point real signal sequence such that the type-II DCT is defined as follow

$$X[n] = \sqrt{\frac{2}{N}} \cdot \sum_{m=0}^{N-1} x[m] \cdot cos\left(\frac{(2m+1) \cdot n\pi}{2N}\right)$$
$$, (n = 0, \ldots, N-1).$$

This can also be expressed as

$$X[n] = \sum_{m=0}^{N-1} c[n,m] \cdot x[m]$$
$$, (n = 0, \ldots, N-1),$$

where c[n,m] denotes the row number n and column number m of the DCT matrix.

RESULTS

In the system we maintained two types of datasets, one is a trained dataset and the other is a test dataset. The trained dataset maintains the database of legitimate web pages and the test database is created to maintain a database of suspected web pages.

On application of confusion matrix we obtained four potential outcomes of Perceptual Hash and Block mean value hash. From Table 3 on giving the input value 5228 the outcome is True Positive whose predicted and actual value is authentic and on giving value 176 the come is False positive whose actual value isn't authentic. The rejected outcomes are, for the input 84 it is False Negative whose prediction outcome isn't authentic and for value 6258 the outcome is True Negative whose predicted and actual values are not authentic. Similarly The outcomes of Perceptual Hash(Table 4) are obtained by performing confusion matrix test.

To analyze the performance of the proposed system, involved in collection of different web pages by using various keywords i.e., banking, mail, social networking.

Verification System

The integrity verification of media objects is about a two-class prediction drawback (binary classification) in which the results are labeled as either positive or negative. Four potential outcomes are obtained. If the result from a prediction is authentic and the actual value is also authentic, then it's known as a true positive else it is considered to be a false positive. Conversely, a true negative occurs when both prediction outcome and also the actual value don't seem to be authentic. A false negative occurs when the prediction outcome isn't authentic but the actual value is authentic. The confusion matrix in Table 1 illustrates the potential outcomes.

The False Accept Rate (FAR) and the False Reject Rate (FRR) are common metrics to indicate the probability of falsely classified media objects. They rely upon the chosen threshold. The threshold, FAR, FRR and other important metrics are discussed below.

Table 1: Confusion Matrix

Decision / Attempt	Authentic (class 1)	Not authentic (class 2)
Accept	True positive	False positive (Type 2 error)
Reject	False negative (Type 1 error)	True negative

Table 2: Hash for the given Input

	Input	Perceptual Hash
1		008080ffffffffff
		block-based image Hash
2		f2ecfcf4

Comparison Result:

0 degree rotation: 100,

When it is in 0 degree similarity was 100%

90 degree rotation: 53.1,

When it is in 90 degree Similarity was 53.1%

180 degree rotation: 31.3,

When it is in 180 degree similarity was 31.3%

270 degree rotation: 53.1

When it is in 270 degree similarity was 53.1%

Table 3: Block Mean Value Based Hash

Decision / Attempt	Authentic (class 1)	Not authentic (class 2)
Accept	5228	176
Reject	84	6258

Table 4: Perceptual Hash

Decision / Attempt	Authentic (class 1)	Not authentic (class 2)
Accept	5645	131
Reject	68	6047

CONCLUSION

This paper aims in development and implementation of a toolkit modeled to make system builders quickly construct content-based similarity search systems for numerous types of feature-rich knowledge. We proposed perceptual image hashing system using which we can identify authentic and non-authentic web pages using trained and test datasets and extract image options based on four stages as discussed. This algorithm thereby increases the accuracy and performance when compared to existing algorithms.

REFERENCES

[1] Cai, Deng, et al. "Hierarchical clustering of WWW image search results using visual, textual and link information." *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004.

[2] Zamir, Oren, and Oren Etzioni. "Grouper: a dynamic clustering interface to Web search results." *Computer Networks* 31.11 (1999): 1361-1374.

[3] Kustanowitz, Jack, Ben Shneiderman, and Bill Kules. "Categorizing web search results into meaningful and stable categories using fast-feature techniques." *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*. IEEE, 2006.

[4] Ahmed, F. & Siyal, M. Y. (2006). A secure and robust wavelet-based hashing scheme for image authentication, in T.-J.

[5] Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua & L.-T. Chia (eds), Advances in Multimedia Modeling, Vol. 4352 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 51–62.

[6] Furht, B., Socek, D. & Eskicioglu, A. M. (2004). Fundamentals of multimedia encryption techniques, in Multimedia Security Handbook, CRC Press, pp. 93–131.

[7] Belkin, Mikhail, and Partha Niyogi. "Laplacian eigenmaps and spectral techniques for embedding and clustering." *Advances in neural information processing systems*. 2002.

[8] Manjunath, Bangalore S., and Wei-Ying Ma. "Texture features for browsing and retrieval of image data." *IEEE Transactions on pattern analysis and machine intelligence* 18.8 (1996): 837-842.

[9] He, Xiaofei, Wei-Ying Ma, and Hongjiang Zhang. "Imagerank: spectral techniques for structural analysis of image database." *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. Vol. 1. IEEE, 2003.

[10] Cai, Deng, et al. "Vips: a vision-based page segmentation algorithm." (2003).

[11] J. Kleinberg, "Authoritative sources in a hyperlinked environment", Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.