Query Mining for Image Retrieval System Using Markov Chain Model

Archana J. Waghchawre **Prof. J.V. Shinde**

Abstract — **In the routines of many users, they need to interact with applications that deal with information retrieval process. These applications always deal with the documents in the database. There is a software which interacts and responds with the top 'n' documents from the database. The expectations of the end users are that the application should retrieve the most relevant document in a very less time and provide accuracy. Even though the application is capable of retrieving the documents, there are some constraints that the application needs to deal with. Implementing the retrieving strategy is a challenging task. In this paper we have proposed a method for automatic annotation, indexing and annotation-based retrieval of images. The proposed method initially analyses the user's query and tries to recognize the relevance between the index structures that were produced for collecting the images. After analysing, the system returns the covet images to the end user with higher clarity and potency. This paper also provides a good survey on the various aspects of information retrieval model.**

Key Words -Information retrieval, Markov chain , *Content-based image retrieval, LSI, PLSI.*

I. INTRODUCTION

An **image retrieval** application is a computer system for browsing, searching and retrieving images from a large database of images.

Most traditional and common methods of image retrieval utilize some method of focusing on metadata such as captioning, keywords and other aspects of that image which will describe those images so that retrieval can be performed over the annotation words. There are types of images, **analog** and **digital**, used in variety of applications.

Capturing the multimedia data by digital devices and storing the data has been the rapidly growing task nowadays. Many researchers are working and resolving the key challenges of the image retrieval techniques. In the past decade, the most attracting topic of interest for the researchers has been the CBIR (Content-Based Image Retrieval). Many different computer communities have been working on CBIR. Although, the multimedia database requires careful studies, image finding techniques, etc; thus, making it a very challenging and open issue.

The fast growing and active area according to the user's interest is the Content-Based Image Retrieval. In this technique, visual contents are used for searching images from a larger pool of image database; also the focus should be on the metadata of the given objects. There are some of the existing CBIR systems like QBIC [2], Photobook [3],Virage[4], Visual Seek [5], WebSeek [5], Netra[6], Cypress[7], that are gaining the attention but still are uncommon. As the CBIR system depends on low level features, the user needs to provide a query similar image to the retrieval system, which is becoming a challenging task. There are several surveys on the CBIR systems that can be found in [8-13]. We refer one of the CDA protocol, the trading proceeds intwo stages. In the first stage, the Market Agent matches the bids and asks based on their individual QoS values and shout prices. After matching, a provisional transaction is created. This provisional transaction enters the second stage. In the second stage, the Buyer Agent compares all the Asks returned as provisional transactions.

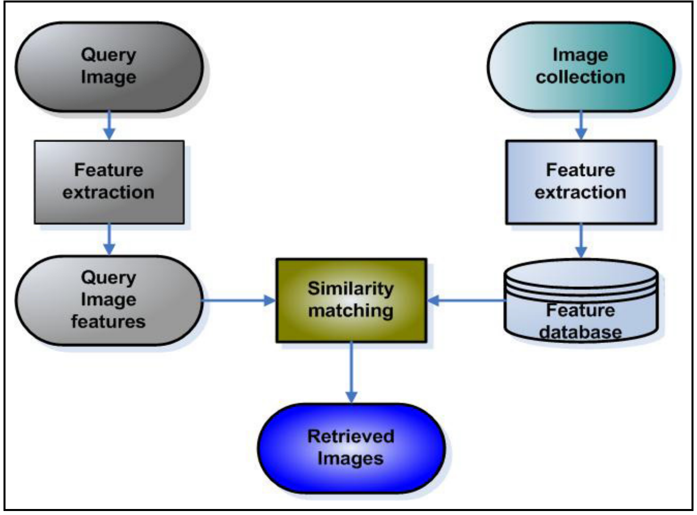

Fig.1.Content Base Image Retrieval

There is a technique that is related to the online image retrieval technique known as the Markovian Semantic Indexing (MSI).

In this technique, the per image annotation data is limited and thus make this technique more suitable for ABIR tasks. Also the characteristics of MSI, make it particularly applicable to the context of online image retrieval systems.
Annotation-Based Image Retrieval (ABIR) is a system where an attempt is made for incorporating efficient semantic contents into text-based queries and into image caption. Also the properties of MSI make the technique suitable of ABIR tasks when the per image annotation data is limited.

II. Literature Survey

The theory of text retrieval systems is used by the annotation based image retrieval techniques. There are many document retrieval techniques, which are assimilated into ABIR systems. Here we discuss some of the document retrieval techniques.

S. Deerwester introduced Latent Semantic Indexing (LSI) [14], where a document retrieval technique is been proposed to deal with some of the shortcomings of traditional lexical matching techniques. LSI deals with the problems of synonymy where many words refer to same object and polysemy where many words have multiple meanings. In the process of Latent Semantic Indexing (LSI), it searches for something that is closer to semantics of a document rather than only matching specific keywords. It starts with the creation of terms by document matrix. The matrix is an high dimensional matrix and is then deteriorated into diminished matrix called as Singular Value Decomposition (SVD). The SVD filters the noise that is found in a document, so that if there are two documents that have same semantics will be located close to one another in a multi-dimensional space [14]. But there are some drawbacks with LSI, it is computationally expensive, performance and speed level degrades when applied to large scale collection, difficult to interpret. An alternative to LSI has been proposed by Hofmann [16] with the probabilistic LSI (PLSI) model because LSI has many shortfalls due to unsatisfactory and incomplete theoretical foundation. The roots of PLSI are attached to the basic LSI technique. Similarly, it deals with synonymous as well as polysemous words.

The process of automated document indexing is indulged in the technique of PLSI, where every document is presented by the word frequency. Sometimes PLSI is familiar with the aspect model which is a latent variable model that includes latent variables related to the observed variables. Therefore, the PLSI method has a more resilient statistical foundation, and is capable for providing a more proper generative data model. The algorithm used by the PLSI systems is known as the Expectation Maximization (EM) algorithm. Although the method of PLSI is good in text analysis technique there are some drawbacks associated with it. Some of which are: it is incomplete since it provides no probabilistic model at the level of documents, leads to over fitting problems if there are too many parameters in the model and it's not clear how to assign how to assign probability to a document outside of the training data. To address the limitations of PLSI, Blei et al. [17]

proposed a unsupervised, generative model called Latent Dirichlet Allocation (LDA). It is closely related to PLSI. It is a powerful generative probabilistic model developed for

modelling words in a document. In LDA each document is a mixture of a small number of latent topics, here each topic is characterized by a distribution over words.

II. III. Proposed System

The proposed system consists of a properly indexed database containing collection of images. The images are stored into the database and are correctly indexed. If the user desires to search an image then the very first task of the user is to give a query 'q'. With accordance to that query 'q', the indexed will be checked and will output some links and few images. The user will then select one of the images from the output images that are more relevant to the users query. Fig 3. describes the flow of the proposed system.

The images can be represented in many different ways and feature vector representation is one of the types. The images are pre-computed and are stored as index in one file into an database. The process of indexing and retrieving is carried on all the images that are stored in the database. While retrieving the image, the images that are near to the query images are given to the end user. A method where a computer automatically assigns the metadata is known as image tagging. The process of automatic image annotation is also known as image tagging, where an image or object is described in more detail in the form of keywords.

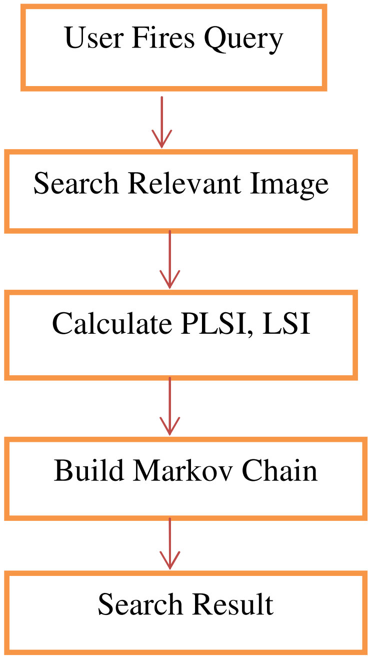

Fig.3.Flow of Proposed System

There are many image classification and distribution classes. The proposed process is considered as one of the type of 'n' class. As soon as a new image is added, it needs to be analysed. Initially the image is analysed in the form of extracted feature vectors and describes the particulars of the image. Then the annotation words will be assigned so that it can be used by the machine learning technique, for automatically applying the annotations.

The proposed method will attempt to find out the relation between the documents and vocabulary.

There is a matix known as TF.IDF matrix which notes the frequency of the terms in a collection of documents. This TF.IDF matix is a mathematical matrix such that let "d" be the collection of documents in the database, "t" be the collection of terms in the collection D.

The term frequency (tf) for a given term ti within a particular document dj is defined as the number of occurrences of that term in the dj th document, which is equal to ni j the number of occurrences of the term ti in the document dj .Therefore term frequency which will specify the occurrence of character / word in a given sentence or file known as term frequency. And inverse document frequency will specify the occurrence of word in multiple documents in a collection. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. I Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

We can easily find out which document consists of a particular word and its frequency, as rows represent the documents in collection and columns represent the terms or phrases. The representation of the TF.IDF matrix is as follows:
We can easily find out which document consists of a particular word and its frequency, as rows represent the documents in collection and columns represent the terms or phrases. The representation of the TF.IDF matrix is as follows:

Sr. No	I	Like	Hate	Mangos
D1	1	1	0	1
D2	1	0	2	1

Mathematical representation will be as follows:

$$w_{ij} = tf_{ij} \times idf_j$$

$$idf_j = log\left(\frac{N}{df_i}\right)$$

In the formula, *tf*$_{ij}$ is the number of keyword *j k* in document *di* which is also called Term Frequency (TF, *idf*$_i$ is the Inverse Document Frequency (IDF Of the keyword *j k* , N is the number of all of training documents. *j df* is the number of documents which contains keyword *j k* . Order the word by their weight and extract a certain number of the words with high weight as the topic feature vector.

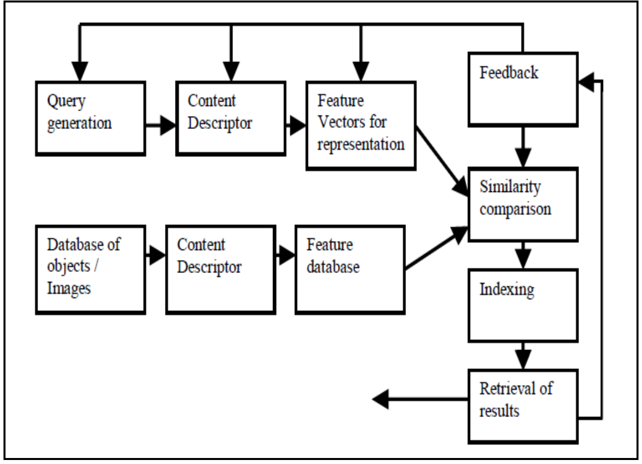

Fig.2.Proposed System Design

IV. RESULTS

We compare the proposed method to the LSI and pLSI approaches in two scenarios. The first experiment is acomparison to LSI, since the limited number of images used in this experiment does not permit reliable comparison to pLSI. The full features of the proposed distance (MSI) are demonstrated in this experiment since the generative process of the aggregate Markov chain during the automatic annotation of images was available to us as is explained later on. Sixty four images that form two intuitive classes were used for this experiment, 32 images related to the term Greek and considered to belong to the first class, and 32 images related to the term Hawaiian are considered to belong to the second class. First, the distance of the 64 images from the query Greek Islands is calculated and ranked for both methods and the results are examined. Then, we calculate the precision and recall for those images and the MSI shows the better result as compare to both methods.

We calculate the precision and recall for all database images. Also compare the LSI, PLSI, MSI of that images.

V. CONCLUSION

The Markovian Semantic Indexing, a new method for mining user queries by defining keyword relevance as a connectivity measure between Markovian states modeled after the user queries. The proposed system is dynamically trained by the queries of the same users that will be served by the system. Consequently, the targeting is more accurate, compared to other systems that use external means of nondynamic or non-adaptive nature to define keyword relevance.

The main intention of this application is to develop an image retrieval application which can perform identity check of an image. The objective is to progress towards the user fulfillment by returning images that have a higher probability to be accepted, meaning is that those images will be highly relevant to the user’s query. Each feature defines a multidimensional space where images are points, and the similarity between images is computed as the distance between points. Therefore we need to focus on various aspects related with the image.

Retrieving images online has been the latest tasks of content based image retrieval. A new technique known as Morovian Semantic Indexing is been used for mining user queries by defining keywords that are used as a connectivity measures between Monrovian states modelled after user’s queries. This method gives better performance as compared to many existing systems.

ACKNOWLEDGEMENT

Prof.J.V. Shinde, for her guidance and support. I will forever remain grateful for the constant support and guidance extended by guide, in making this report. Through our many discussions, she helped me to form and solidify ideas. The invaluable discussions I had with him, the penetrating questions she has put to me and the constant motivation, has all led to the development of this project.

I wish to express my sincere thanks to the Head of department, Prof. N. R. Wankhade also grateful thanks to the departmental staff members for their support.

REFERENCES

[1] Konstantinos A. Raftopoulos, Klimis S. Ntalianis, Dionyssios D. Sourlas, and Stefanos D. Kollias, “Mining User Queries with Markov Chains: Application to Online Image Retrieval”, IEEE Transaction On Knowledge and Data Engineering, Vol. 25, nO. 2, February 2013.

[2] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q.,Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D.,And Yanker, P. 1995. Query by image and video content: The QBIC system. IEEE Comput. 28, 9, 23–32.

[3] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *IJCV*, 18(3):233–254, 1996.

[4] A. Gupta, R. Jain, Visual information retrieval, Commun. ACM 40(5) (1997) 70–79.

[5] J. R. Smith and S.-F. Chang. Visualseek: a fully automated contentbased image query system. In *Proc.ACM Multimedia '96*, 1996.

[6] Ma, W. And Manjunath, B. 1997. Netra: A toolbox for navigating large image databases. In Proceedings of the IEEE International Conference on Image Processing (ICIP).

[7] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan ,T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images.In *Proc. ECCV 96 Workshop on Object Rep.*, 1996.

[8] A.W.M. Smeulders, M. Worring, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.

[9] M.S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: state of the art and challenges, ACM Transactions on Multimedia Computing, Communications and Applications 2 (1) (2006) 1–19.

[10] N. Vasconcelos, From pixels to semantic spaces: advances in content-based image retrieval, Computer 40 (7) (2007) 20–26.

[11] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences and trends of the new age, ACM Computing Surveys 40 (2) (April 2008).

[12] F. Long, H.J. Zhang, D.D. Feng, Fundamentals of content-based image retrieval, in: D.D. Feng, W.C. Siuandg, H.J. Zhan (Eds.), Multimedia Information Retrieval and Management, Springer, 2003.

[13] Y. Rui, T.S. Huang, S.F. Chang, Image retrieval: current techniques, promising directions and open issues 10 (1999) 39–62 Journal of Visual Communication and Image Representation 10 (1999) 39–62.

[14] S. Deerwester, S. T. Tumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman Indexing by latent semantic analysis, J. Soc. Inform. Sci. 41, 6 (1990), 391_407.

[15] M.W. Berry, S.T. Dumais, and G.W. O’Brien, “Using Linear Algebra for telligent Information Retrieval,” SIAM Rev., vol. 37, no. 4, pp. 573-595, 1995.

[16] T. Hofmann, “Probabilistic Latent Semantic Indexing,” Proc. 22ndInt’l Conf. Research and Development in Information Retrieval (SIGIR’99), 1999.

[17] D.M. Blei and A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,”J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

AUTHOR’S PROFILE

Archana. J. Waghchawre

P.G. student: Department of Computer Engineering, Late G.N. Sapkal College of Engineering, Anjaneri,City: Nasik, Country: India.

University: Savitribai Phule Pune University

Email id:archanagaikwad17@gmail.com.

Prof. J.V. Shinde

Associate Professor: Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Anjaneri, City: Nasik, Country: India.

University: Savitribai Phule Pune University

Email id: jv.shinde@rediffmail.com