IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

User Query Analyses Using Improvised Markov Chain Model

S.Mariammal[1], Mr.K.Dhakshana Moorthy[2]

[1]PG student, Department of Computer Science and Engineering, S.Veerasamy Chettiar College of Engineering and Technology Puliangudi-627 855

[2]Assistant Professor, Department of Computer Science and Engineering, S.Veerasamy Chettiar College of Engineering and Technology Puliangudi-627 855

Abstract

A novel method for automatic annotation, indexing and annotation based retrieval of images is proposed. The Markovian Semantic Indexing (MSI) is presented in the context of an online image retrieval system. The user's queries are used to construct an Aggregate Markov Chain (AMC) through which the relevance between the keywords seen by the system is defined. The user's queries are also used to automatically annotate the images. A stochastic distance between the images, based on their annotation and the keyword relevance captured in the AMC is also introduced. geometric interpretations of the proposed distance are provided and its relation to a clustering in the keyword space is investigated. The proposed method shows theoretical advantages and also achieves better precision and recall rate when compared to the state of the art approaches.

Index Terms— *Similarity Measure, Image Utility, Markovian Chain, Transition Probability*

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection and data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data degrading, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing strategies for their product. They can use data to compare and contrast among competitors. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company.

The great interest and a wealth of promise in content-based image retrieval is an emerging technology. While the last decade laid foundation to such promise, it also paved the way for a large number of new techniques and systems, got many new people involved, and triggered stronger association of weakly related fields. The survey almost 300 key theoretical and empirical contributions in the current decade related to image retrieval and automatic image annotation, and in the process discuss the spawning of related subfields. The challenges involved in the adaptation of existing image retrieval techniques to build systems that can be useful in the real world. As part of an effort to better understand the field of image retrieval, compiled research trends in image retrieval using Google Scholar's search tool and its computed citation scores. Graphs for publication counts and citation scores have

been generated for: (1) subfields of image retrieval, and (2) venues/journals relevant to image retrieval research.

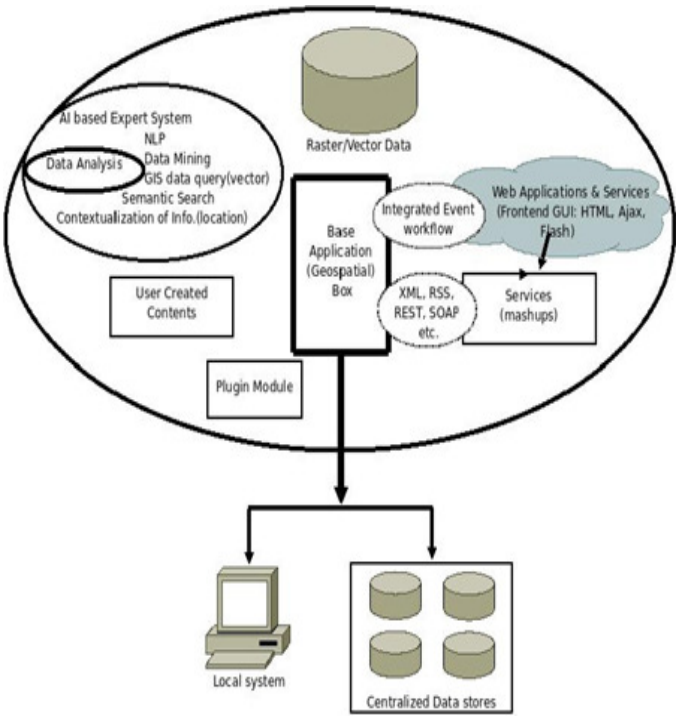

Fig 1 Data Minig Architecture

Latent Semantic Analysis (LSA) [1] is an approach to automatic indexing and information retrieval that attempts to overcome these problems by mapping documents as well as terms to a representation in the so called latent semantic space. Probabilistic Latent Semantic Indexing is used to automated document indexing which is based on a statistical latent class model for factor analysis of count data. Fitted from a training corpus of text documents by a generalization of the Expectation Maximization algorithm, the utilized model is able to deal with domain - specific synonymy as well as with polysemous words. In contrast to standard Latent Semantic Indexing (LSI) by Singular Value Decomposition, the probabilistic variant has a solid statistical foundation and defines a proper generative data model. Retrieval experiments on a number of test collections indicate substantial Performance gains over direct term matching methods as well as over LSI. It can also take advantage of statistical standard methods for model fitting, over fitting control, and model combination. The empirical evaluation has clearly confirmed the benefits of Probabilistic Latent Semantic Indexing which achieves significant gains in precision over both, standard term matching and LSI.

The transaction logs containing 51,473 queries posed by 18,113 users of Excite, a major Internet search service. The data provided by three ways: (i)sessions - Changes in queries during a session, number of pages viewed, and use of relevance feedback; (ii)queries - The number of search terms, and the use of logic and modifiers; and (iii)terms - Their rank/frequency distribution and the most highly used search terms. Then, the analysis from the query to the user gains insight to the characteristics of the Web user. A unique query was the first query by a user (this represents the number of users). A modified query is a subsequent query in succession (second, third . . .) by the same user with terms added to, removed from, or both added to and removed from the unique query. Unique and modified queries together represent those queries where the user did something with terms. Identical queries are queries by the same user that are identical to the query preceding it. They can come about in two ways. The first possibility is that the user retyped the query. Studies have shown that users often do this (Peters, 1993). The second possibility is that the query was generated by Excite. When a user views the second and further pages (i.e., a page is a group of 10 results) of results from the same query, Excite provides another query, but a query that is identical to the preceding one. The rest of the paper is as follows. In the next section, we present the proposed approach. Section III presents experimental results. Section IV presents conclusion and future works.

II. PROPOSED METHOD

A novel (alternative)probabilistic approach for Annotation Based Image Retrieval that, compared to LSI and PLSI, is better suited to sparsely annotated domains, like in image databases where, the per image sparse keyword annotation is also limited. It addresses in a more natural way the zero frequency problems, defined as the fact that the probability to find common keywords even in closely related images is typically small because the images are not annotated with exactly the same keywords. This problem is addressed here by means of an explicit relevance link between keywords that carries a probabilistic weight. We show that assigning logical connections between keywords by means of a Markovian model, permits better generalization over a sparsely annotated domain hence the proposed approach raises the reasoning aspect next to the numerical aspect of probabilities. The key idea behind the approach is to compensate for the sparse data by incorporating an annotation procedure of probabilistic qualitative reasoning that will propagate partial beliefs regarding connections between keywords. A mechanism that gains performance from mining the structure of the existing data rather that incorporating new data, as it happens with traditional models is hence introduced.

The proposed system can be divided into the following six modules.

i) RGB Projection
ii) Image Utility
iii) Comparable Image

iv) Similarity Images
v) Semantic Search
vi) Graphical Representation

RGB Projections

The RGB color model is an additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors. The name of the model comes from the initials of the three additive primary colors, red, green, and blue. The main purpose of the RGB color model is for the sensing, representation, and display of images in electronic systems, such as conventional photography. In this module having two dialog controls like folder browser dialog and open file dialog. In this folder browser dialog can be used to open and browse for the particular file folder and view the overall content for that file folder. And also, we can use the open file dialog, to open the particular files of the content for view the content.

In this module the RGB Projections is used to find the size of the image vertically and horizontally. And also to get the output result for the images like vertically and horizontally. This module is used to view the color of the images and adjustment of those colors for the particular images.

Image Utility

Whenever minimizing the error of classification is interesting for CBIR, this criterion does not completely reflect the user satisfaction. If the user mostly gives relevant labels, the system should propose new images for labeling around a higher rank to get more irrelevant labels. To get the output result of the RGB projection image to use in this module. The source and destination images are been used in this module. To adjust the color of the images like jpg image, and to adjust or reduce the size of the image like horizontally and vertically. Source image to be modified for the image utility clarification to the destination images to be used.

Other utility criteria is closer to this, such as precision, should provide more efficient selections. Our main strategy leads to a fast and efficient active learning scheme to retrieve sets of online images. The main purpose of the model is for the sensing, representation, and display of images in electronic systems, such as conventional photography. The image utility classification to the destination images to be used.

Comparable Image

In this module a reselection technique to speed up the selection process, which leads to a computational complexity negligible compared to the size of the database for the whole active learning process. All these components are integrated in our retrieval system, called RETIN and the user gives new labels for images, and they are compared to the current classification. If the user mostly gives relevant labels, the system should propose new images for labeling around a higher rank to get more irrelevant labels. The experiments of the image to be compare several powerful classification techniques in this information retrieval context. Experiments on large databases show that the RETIN method performs well in comparison to several other active strategies.

The main purpose of this model is for the sensing, representation, and display of images in electronic systems, such as conventional photography. The method which aims at minimizing the error of generalization is the less efficient active learning method.

Similarity Measure

The results in terms of mean average precision according to the training set size (we omit the KFD which gives results very close to inductive SVMs) for both ANN and Corel databases. One can see that the classification-based methods give the best results, showing the power of statistical methods over geometrical approaches, like the one reported here (similarity refinement method). The main concepts that are the most difficult to retrieve are very small and/or have a much diversified visual content. Our strategy leads to a fast and efficient active learning scheme to retrieve sets of online images. The average precision is then computed using the ranking. The similarity measures for the source and destination images and to get the overall result for those images. The method which aims at minimizing the error of generalization is the less efficient active learning method. Focusing on interactive methods, active learning strategy is then described.

Semantic Search

Finally, the image will take the relevant image what the user search. One can see that we have selected concepts of different levels of complexities. The performances go from few percentages of Mean average precision to 89%. The concepts that are the most difficult to retrieve are very small and/or have a much diversified visual content. The method which aims at minimizing the error of generalization is the less efficient active learning method. The most efficient method is the precision- oriented method.

These labeled images are used to train a classifier, which returns a ranking of the database. The average precision is then computed using the ranking. One can see that we have selected concepts of different levels of complexities. The criterion of generalization error to optimize the active learning selection is modified to better represent the CBIR objective of database ranking. And also to get the output image like a fast and efficient strategy to retrieve the query Concept in content-based image retrieval.

Graphical Representation

This module is used to determine relationships between the two Images. The precision and recall values are measured by simulating retrieval scenario. For each simulation, an image category is randomly chosen. Next, 100 images are selected using active learning and labeled according to the chosen category. These labeled images are used to train a classifier, which returns a ranking of the database. The average precision is then computed using the ranking. These simulations are repeated 1000 times, and all values are averaged to get the Mean average precision. Next, we repeat ten times these simulations to get the mean and the standard deviation of the MAP.

The three steps of the proposed method can be explained using the following diagrams.

Step 1:

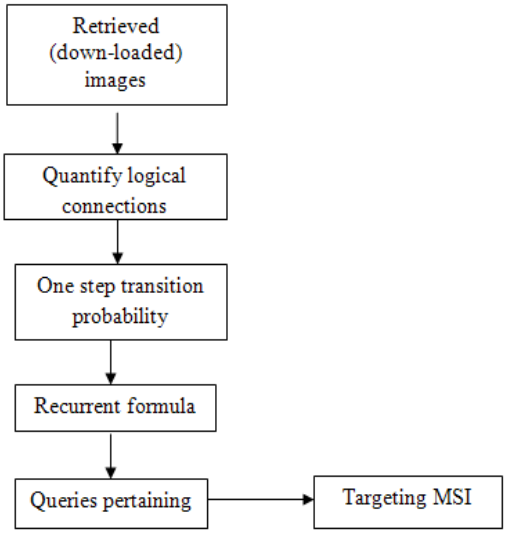

Fig 2 Step 1 of the proposed system

Step 2:

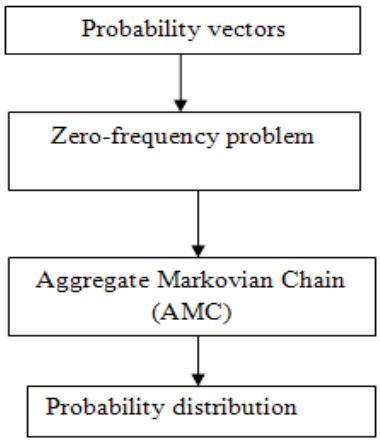

Fig 3 Step 2 of the Proposed System

Step 3:

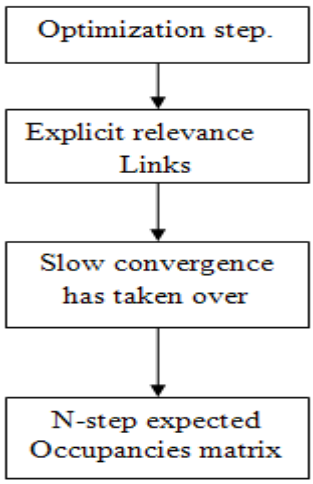

Fig 4 Step 3 of the proposed system

III. Experimental Results

We compare the proposed method to the LSI and pLSI approaches in two scenarios. The first experiment is a comparison 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, a complete distance table is built for all the in-between distances of these 64 images using both methods.

In the second experiment, the full features of our method cannot be demonstrated since the scenario involves a publicly available ground-truth database, during the annotation of which we had no control [23]. The Aggregate Markov Chain, necessary for our method cannot be reliably constructed. Nevertheless this experiment serves as a comparison to pLSI in the ability to extract latent features in the case of already annotated databases, the annotation having been performed with unknown methods. For such cases, we propose a modification to the standard MSI approach that involves an explicit step for dimensionality reduction (since the implicit dimensionality reduction through the clustering of the keyword space cannot by applied). This comparison to the pLSI, again with precision versus recall diagrams, is performed using the ground-truth database.

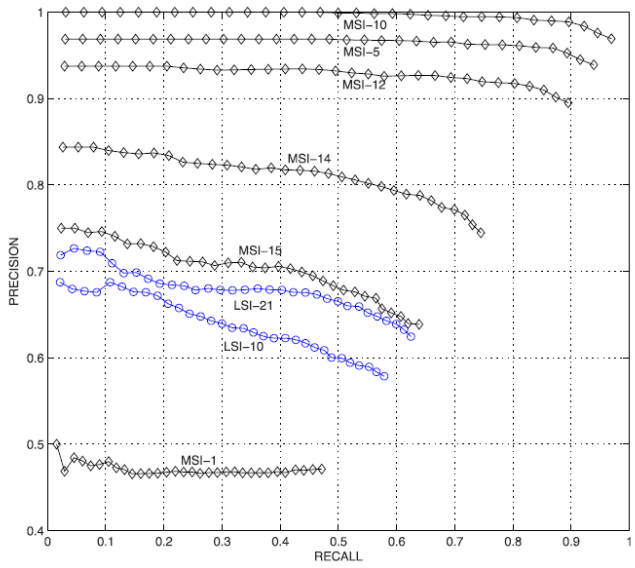

Fig 5 Precision and Recall Rates

IV. CONCLUSION AND FUTURE WORK

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 non-dynamic or non-adaptive nature to define keyword relevance. A stochastic distance, in the form of a generalized Euclidean distance, was constructed by means of an Aggregate Markovian Chain and proved to be optimal with respect to certain Markovian connectivity measures that were defined for this purpose.

A comparison to Latent Semantic Indexing and probabilistic Latent semantic Indexing revealed certain theoretical advantages of the proposed method (MSI). Experiments have shown that MSI achieves better retrieval results in sparsely annotated image data sets. A comparison to LSI on 64 images gathered from the Google Image Search and annotated in a transparent way by the proposed system, revealed certain advantages for the MSI method, mainly in retrieving images with deeper dependencies than simple keyword co-occurrence. Another comparison to PLSI was performed using the ground-truth annotated database of 1,109 images after modifying the proposed method to incorporate AMC construction and dimensionality reduction in external annotations. The results of Precision versus Recall for this experiment revealed that MSI at 200 dimensions achieve better scores than PLSI does at any dimensionality.

In future work, using ranking based image retrieval (RBIR) method can be done to provide ranking for each user query. Many times to give same query for image retrieval, the query will arrange priority based and the query easily retrieve the images.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Z. Guo, S. Zhu, Y. Chi, Z. Zhang, and Y. Gong, “A Latent Topic Model for Linked Documents,” Proc. 32nd Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2009.

[3] W.J. Stewart, Numerical Solution of Markov Chains. Princeton Univ. Press, 1994.

[4] G. Zhen, Z. Shenghuo, C. Yun, Z. Zhongfei, and G. Yihong, “A Latent Topic Model for Linked Documents,” Proc. 32nd Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’09), 2009.

[5] O. Tuzel, F. Porikli, and P. Meer, “Pedestrian Detection via Classification on Riemannian Manifolds,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 10, pp. 1713-1727, Oct. 2008.

[6] R.O. Duda, P.E. Hart, and N.J. Nilsson, "Subjective Bayesian Methods for Rule-Based Inference Systems," Proc. Nat'l Computer Conf. and Exposition (AFIPS), vol. 45, pp. 1075-1082, 1976.

[7] U. Montanari, "Networks of Constraints, Fundamental Properties and Applications to Picture Processing," Information Science, vol. 7, pp. 95-132, 1974.

[8] L.G. Shapiro, "GroundTruth Database," http://www.cs.washington.edu/research/imagedatabase/groundtruth/, Univ. of Washington, 2012.

[9] J. Fan and Y. Gao, and H. Luo, "Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classier Training for Multilevel Image Annotation," IEEE Trans. Image Processing, vol. 17, no. 3, pp. 407-426, Mar. 2008.

[10] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," Proc. Int'l Conf. Computer Vision, vol. 2, pp. 408-415, 2001

[11] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, vol. 40, no. 2, article 5, pp. 1-60, 2008.

[12] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic Author-Topic Models for Information Discovery," Proc. 10th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, 2004

[13] T.L. Griffiths and M. Steyvers, "Finding Scientific Topics," Proc.Nat'l Academy of Sciences USA, vol. 101, no. suppl. 1, pp. 5228-5235, 2004.

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

[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. 22nd Int'l Conf. Research and Development in Information Retrieval (SIGIR 99), 1999

[16] D. Joshi, J.Z. Wang, and J. Li, "The Story Picturing Engine – A System for Automatic Text Illustration," ACM Trans. Multimedia Computing, Comm. and Applications, vol. 2, no. 1, pp. 68-89, 2006.

[17] J. Li and J. Wang, "Real-Time Computerized Annotation of Pictures," Proc. ACM 14th Ann. Int'l Conf. Multimedia, 2006.

[18] A. Bhattacharya, V. Ljosa, J.-Y. Pan, M.R. Verardo, H. Yang, C. Faloutsos, and A.K. Singh, "Vivo: Visual Vocabulary Construction for Mining Biomedical Images," Proc. IEEE Fifth Int'l Conf. Data Mining, Nov. 2005.

[19] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, 2008.

[20] B.J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," Information Processing and Management, vol. 36, no. 2, pp. 207-227, 2000.

[21] K. Stevenson and C. Leung, "Comparative Evaluation of Web Image Search Engines for Multimedia Applications," Proc. IEEE Int'l Conf. Multimedia and Expo, July 2005.

[22] S. Santini and R. Jain, "Similarity Measures," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 9, pp. 871-883, Sept. 1999.