Clustering based Online Image retrieval using Markovian Semantic Indexing

[1]N.Vijay, [2]R.Vijayalakshmi

[1]PG Student, [2]Assistant Professor
Department of Information Technology, SRM University, Kattankulathur
[1]vijayn88cse@gmail.com, [2]vijisekaran@gmail.com

***Abstract* — Information mining is now emerging term in recent days. This helps the user to gain knowledge and learning new information as for their need image searching is also a type of information searching. The current computer vision techniques extract from images mostly low-level features and the link between low-level features and high-level semantics of image content is lost. Neither a single low-level feature nor a combination of multiple low-level features has explicit semantic meaning in general. Images are searched through keywords, In recent days image searching is processed through content based image retrieval and image based retrieval. Here in this paper we implemented Markovchain Technique is related with clustering to improve efficient image searching. In existing image search in clustering, but there is no identity for our search, but here we are using AMC and clustering we provide identity to each image by giving meta-tag. So this identity and meta tag make the searching more easy and it gives quick result for the user as per the user requirement. Time taken to search the image is improved and it provide relevant as per the user search. Time taken to search the image and quick indexing will provide easy search. Here annotation and keywords are compared efficiently with the system. The MSI distance between minimum vector and maximum vector is compared, the result are obtained efficiently. This makes the search efficiently.**

Keywords- Markovian Semantic Indexing, Ranking, Meta Tag, Identity, Content Based, Semantic Based

1. INTRODUCTION

Retrieving information in a faster way becomes more important in today's information era. Although it is new, accelerating growth of its structure and accommodation of thickness and overwhelming amount of information makes the Internet first resort for people. While the exploding rate of growth in Internet, as the youngest media continues, the search engines that are indispensable tools for retrieving information keep on developing themselves. It is obvious that the more effective the system is the more it will satisfy the user. However, low precision and large amount of results have still been problems for the search engines. Much of the research and development in information retrieval is aimed at improving the effectiveness and efficiency of retrieval and satisfy the user information necessity. The current computer vision techniques extract from images mostly low-level features and the link between low-level features and high-level semantics of image content is lost. Neither a single low-level feature nor a combination of multiple low-level features has explicit semantic meaning in general.

The similarity measures between visual features do not necessarily match human perception. The semantic gap: the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. However, the retrieval process fails also due to the sensory gap: the gap between the object in the world and the information in a (computational) description assigned to a recording of that object. Annotation-Based Image Retrieval (ABIR) systems are an attempt to incorporate more efficient semantic content into both text-based queries and image captions like in Google Image Search, Yahoo! Image Search). The Latent Semantic Indexing (LSI)-based approaches that were initially applied with increased success in document indexing and retrieval, were incorporated into the ABIR systems to discover a more reliable concept association. Mathematical optimisation models are based on classical operations research techniques such as linear programming, integer programming, goal programming or dynamic programming.

The current computer vision techniques extract from images mostly low-level features and the link between low-level features and high-level semantics of image content is lost. Neither a single low-level feature nor a combination of multiple low-level features has explicit semantic meaning in general. The similarity measures between visual features do not necessarily match human perception. The semantic gap: the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. However, the retrieval process fails also due to the sensory gap: the gap between the object in the world and the information in a (computational) description assigned to a recording of that object. Annotation-Based Image Retrieval (ABIR) systems are an attempt to incorporate more efficient semantic content into both text-based queries and image captions e.. Google Image Search, Yahoo! Image Search). The Latent Semantic Indexing (LSI)-based approaches that were initially applied with increased success in document indexing and retrieval were incorporated into the ABIR systems to discover a more reliable concept association. The Markovian Semantic Indexing, a new method for mining user queries by defining keyword relevance as 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, m-pared to other systems that use external means of non-dynamic or non-adaptive nature to define keyword relevance.

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 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 no dynamic or no 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

2. RELATED WORKS

The motivation of our work is to make the search more efficiently using the user query by automated annotation, indexing and annotation-based retrieval of images. In this section we survey the previously proposed approaches for the accurate and efficient methods which helps in image search that makes the user to search in effective way.

K. Stevenson et.al. Proposed a novel ranking model named Eagle Rank for web image search engine. In Eagle Rank, multiple sources of evidence related to the images are considered, including image surrounding text passages, terms in special HTML tags, website types of the images, the hyper-textual structure of the web pages and even the user feedbacks. Meanwhile, the flexibility of Eagle Rank allows it to combine other potential factors as well. Based on inference network model, Eagle Rank also gives sufficient support to Boolean AND and OR operators.

B.J. Jansen at. Al. analyzed transaction logs containing 51,473 queries posed by 18,113 users of *Excite*, a major Internet search service. The method provided input data on: (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 shift the focus of analysis from the query to the user to gain insight to the characteristics of the Web user. With these characteristics as a basis, we then conducted a failure analysis, identifying trends among user mistakes.

J.Z. Wang et.al. took analysis on ACM Computing Surveys which states great interest and a wealth of promise in content-based image retrieval as 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.In this paper, it states that almost 300 key theoretical and empirical contributions in the current decade related to image retrieval and automatic image annotation, and discuss the spawning of related sub-fields in the process. It is also discussed the significant challenges involved in the adaptation of existing image retrieval techniques to build systems that can be useful in the real-world. In retrospect of what has been achieved so far, we also conjecture what the future may hold for image retrieval research.

Real-Time Computerized Annotation of Pictures**,** The capability of annotating pictures by computers can lead to breakthroughs in a wide range of applications, including Web image search, online picture-sharing communities, and scientific experiments. In this work, the authors developed new optimization and estimation techniques to address two fundamental problems in machine learning. These new techniques serve as the basis for the automatic linguistic indexing of pictures - real time (ALIPR) system of fully automatic and high-speed annotation for online pictures. In particular, the D2-clustering method, in the same spirit as K-Means for vectors, is developed to group objects represented by bags of weighted vectors. Moreover, a generalized mixture modeling technique (kernel smoothing as a special case) for no vector data is developed using the novel concept of hypothetical local mapping (HLM). ALIPR has been tested by thousands of pictures from an Internet photo-sharing site, unrelated to the source of those pictures used in the training process. Its performance has also been studied at an online demonstration site, where arbitrary users provide pictures of their choices and indicate the correctness of each annotation word**.**

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. Furthermore, numerical problems have to be addressed when the norm of both vectors is close to zero. The method proposed in this work (MSI/AMC) on the other hand, incorporates automatic indexing and query matching tasks within the whole framework, the distance being defined in an optimal manner, directly interpretable in relation to the clustering.

3. PROBLEM STATEMENT

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. Furthermore, numerical problems have to be addressed when the norm of both vectors is close to zero. The method proposed in this work (MSI/AMC) on the other hand, incorporates automatic indexing and query matching tasks within the whole framework, the distance being defined in an optimal manner, directly interpretable in relation to the clustering.

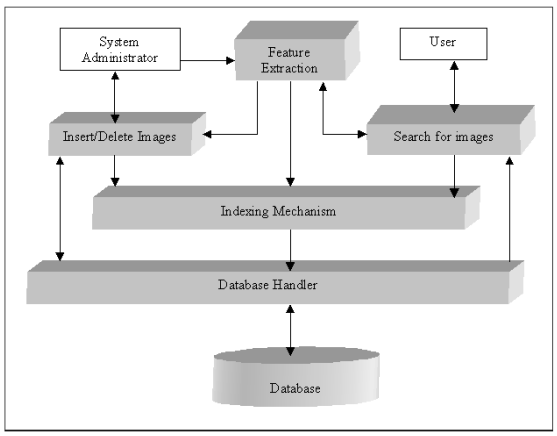
4. METHODOLOGY

The proposed approach (MSI) is presented together with a proximity measure (distance). The geometric interpretation and the optimal properties of the proposed distance are examined. A statistical model with hybrid text/visual characteristics also based on the aspect model. Another model based on Bayesian incremental learning. The reasoning aspect of probabilities, as it defines explicit relevance relationships between keywords. Using network representations for capturing semantics is common in AI systems. 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. The literature the proposed system is novel in the way it unifies these two tasks. Indeed both the automatic annotation and retrieval tasks are assumed in the implicit user interaction context, for dynamically mining semantics towards qualitative probabilistic reasoning. This has implications in the targeting aspect since the training is performed dynamically. During the training phase of the system the images are considered with no annotation.The methodology proposed in this work encompasses 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 problem, 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 interactive procedure has implicit consequences that we exploit one by one in a step by step construction of the proposed system: The user implicitly relates the retrieved (downloaded) images to her/his query. By assuming Markovian chain transitions in the order of the keywords the aim of the proposed approach is to quantify logical connections between keywords. If some user relates image Ii to his query q_i, where keyword k2 follows keyword k1 and this occurs m times, then the one step transition probability $p_i(k_1, k_2)$ is being updated as follows: if $p_i(k_1, k_2)$ is the current probability (before the update) based on M keywords then the new probability (based on M þ m keywords) is calculated by the recurrent formula

$$p_i(k_1, k_2) = \frac{Mp_i(k_1, k_2) + m}{M + m}$$

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 mechanism behind LSI correlates keywords that appear at the same image and images when they are annotated with the same keyword, starting therefore from a keyword/image frequency matrix a singular value decomposition provides a more compact representation of keywords and images in a space of fewer dimensions. The mechanism behind MSI is revealed in the behaviour of the system with respect to the power n. We see that as n increases the results are improved until the best Precision

Recall result is achieved at n ¼ 10. This behavior shows that we managed to improve the retrieval without adding observations in the system but by propagating relevance connections defined on the existing observations. At the time where LSI or pLSI approaches would need more observations to improve results, the proposed system uses the convergence of the Markov chain (by raising the AMC to the appropriate power) to gain on the existing observations by mining deeper qualitative inferences between keywords. When we have no control over the annotation method used for the images neither we have access to the user queries that were used to annotate the images (if such a method has even been used), many of the advantages of the proposed method (MSI) vanish since the AMC cannot be constructed. Nevertheless, we still want to incorporate such cases in the proposed approach and compare to pLSI by modifying our method to permit AMC construction with dimensionality reduction whenever only a keyword image matrix is available. The AMC kernel needs to be raised to the respective power. The EM algorithm on the other hand, is an iterative method which is only guaranteed to find a local maximum of the likelihood function. The inference engine of the MSI approach. The powers of the AMC kernel in the first implementation as well as the PCA treatment in the second implementation that are involved in the respective MSI optimization steps are both computed accurately and efficiently.

This expression will be useful in proving optimality concepts of the proposed distance in the next sections, since it links the mechanics of the convergence to the actual process statistics. Let now i, j be two states of the same chain with j recurrent. The mean first passage time θ_{ij} from state i to state j is the expected number of transitions the process needs to reach j for the first time, starting from i. To calculate cross passage time between two different states in the same chain,

$$\vec{\theta}_{ij} = \frac{t_y(jj) - t_y(ij)}{\pi_j}$$

where $\pi = \pi_1, \pi_2, \pi_3 \ldots . \pi_N$ the equilibrium state of the chain.

The model presented in the frame-work of an online image retrieval system (similar to Google image search) where users search for images by submitting queries that are made of keywords. The queries formed by the users of a search engine are semantically refined, the keywords representing concise semantics when compared to text in documents or other vocabulary related presenta-tions. The aim is to improve user satisfaction by returning images that have a higher probability to be accepted (downloaded) by the user. The assumption is that the users search for images by issuing queries, each query being an ordered set of keywords. The system responds with a list of images. The user can download or ignore the returned images and issue a new query instead. During the training phase of the system the images are considered with no annotation. As the users issue queries and pick images the system annotates the images in an automatic manner and at the same time establishes relevance relations between the keywords as will be explained later on in the manuscript. The user never annotates the images explicitly, this happens by the system transparently from the user. At the testing phase the system uses the annotations available from the training phase but also the keyword relevance probability weights also evaluated during the training phase to return images that better reflect the users preferences and improve user satisfaction.

The new method, Markovian Semantic Indexing (MSI), is presented in the context of an online image retrieval system. The properties of MSI make it particularly suitable for ABIR tasks when the per image annotation data is limited. The characteristics of the method make it also particularly applicable in the context of online image retrieval systems. 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. Annotation-Based Image Retrieval systems incorporate more efficient semantic content into both text-based queries and image captions. A direct consequence is that methods initially developed for document retrieval may be suitable for ABIR systems, as well. Latent Semantic Indexing was initially developed for document retrieval. Furthermore, even though automatic annotation and annotation-based image retrieval systems have been presented in the literature the proposed system is novel in the way it unifies these two tasks. Indeed both the automatic annotation and retrieval tasks are assumed in the implicit user interaction context, for dynamically mining semantics towards qualitative probabilistic reasoning. Let us record their click-through data (query q, returned ranking r, and clicks c) when interacting with Google Image Search using the English language. The students were asked at different periods to simulate different target groups with specific interests (tourists, sports fans, artists, etc.) since the adaptability of the proposed system to the user's preferences had to be tested. A simple proxy was used for keeping a log file with the query-ID, URL, and click-log. From the recorded sessions over a six months interval, for each selected image, its individual Markov chain (based on the query terms as explained above) was built; only the selected images among the images that were returned by Google Image Search were considered.

Aggregate Markov Chain (AMC), calculated at the desired n. Then the MSI distance between images x and y is defined as,

$$d(x,y) = (\pi_x - \pi_y) \sum (F_G^T) (\pi_x - \pi_y)^T$$
$$= \delta_{xy} \sum (F_G T)\, \delta_{xy}^T$$

where the dimensionality of π_x and π_y has been extended to that of $\sum(F_G^T)$ by filling in with zeros the respective coordinates.

The proposed distance is well defined since it is a generalized euclidean distance function, using a covariance matrix, which is always positive definite. The equilibrium state vector of the respective Markovian chain was then used to automatically annotate each image. In a similar manner, from the total number of queries from all users, regardless of the chosen images, a Markov chain, representing the Aggregate Markov Chain, was constructed, serving as a dynamically adaptive taxonomy for the keywords already seen in the user queries. When we have no control over the annotation method used for the images neither we have access to the user queries that were used to annotate the images (if such a method has even been used), many of the advantages of the proposed method (MSI) vanish since the AMC cannot be constructed. Nevertheless, we still want to incorporate such cases in the proposed approach and compare to pLSI by modifying our method to permit AMC construction with dimensionality reduction whenever only a keyword image matrix is available. Our approach assumes that the construction of the AMC has happened progressively from the user queries and in such a case a reduction of dimensionality can be realized by aggregation/disaggregation techniques of the Markovian states, when clusters of states appear in the form of a block diagonal AMC.

Each query is run on the selected image search engine individually. In order to have stable performance measurement of search engines: (1) all the searches performed in a short time period, (2) retrieval outputs of search engines are recorded, and (3) based on the Jansen and Spink study2, the first twenty documents retrieved are evaluated as "relevant" or "non-relevant". Evaluation of each image item is done by the authors separately and the relevancy of the image items is decided by the consensus of authors. During the evaluation of image items retrieved, the following criteria's [7] are used: The relevance judgment was binary, though normalized recall metric allows us to pose more relevancy levels. The same images from different web addresses (i.e., mirror pages) are considered as different ones. Duplicated image items are conceived as good as one for its relevance judgment. If for some reason, the retrieved image is not accessible, or if it is a different image, then it is considered as a non-relevant item.

4. RESULT AND CONCLUSION

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.

In future the model can be further developed using 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. And also in Comparison to Latent Semantic Indexing and probabilistic Latent semantic Indexing revealed certain theoretical advantages of the proposed method (MSI).

REFERENCES

[1] 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.
[2] com Score's Report Article, "Comscore's Qsearch 2.0 Service," comScore's Report Article, www.comscore.com, 2007.
[3] 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.
[4] 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.
[5] 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.
[6] J. Li and J. Wang, "Real-Time Computerized Annotation of Pictures," Proc. ACM 14th Ann. Int'l Conf. Multimedia, 2006.
[7] D. Joshi, J.Z. Wang, and J. Li, "The Story Picturing Engine - ASystem for Automatic Text Illustration," ACM Trans. Multimedia Computing, Comm. and Applications, vol. 2, no. 1, pp. 68-89, 2006.
[8] M.W. Berry, S.T. Dumais, and G.W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," SIAM Rev., vol. 37, no. 4, pp. 573-595, 1995.