International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Incremental Query Processing by Relevance Feedback Using Big-Data Streams

D. Ravi, V. Viknesh, A. Lakshmakarthi, S. Sugumaran

Department of Computer Science and Engineering, Kathir College of Engineering, Coimbatore, Tamil Nadu, India

ABSTRACT

This paper presents deals social network in large scale distributed server data storing and retrival process is more complex.There has been an explosive increase in media data,such as images,videos and social media in the internet,mobile devices, and desktops.Engineers and researches are dealing with data sets of petabyte scale in the cloud computing paradigm.Thus, the demand for building a service stack to distribute,manage and process massive data sets has risen drastically.Data collection has become easy due to the rapid development of both mobile devices and wireless networks.During the processing of image queries. Many factor are affecting quality of the retrievel system. Image searching and ranking , indexing are the insufficient factors to affect the quality of image search results. There are many factors which affect the quality of image search results.The learning of the model is from the image output extracts the designed with the evolutionary feedback system to perform the image retrievel by processing the image search query.

Keywords : Query Optimiztion using Feedback Processing, Backtracking Process.

I. INTRODUCTION

Data mining is a developing science and it can be defined and categorized in a number of ways depending on the specific knowledge domain. For example, this has manifest in the domain of biological science where the technology of data mining has been applied successfully and categorized as bioinformatics. Various techniques have been employed within bioinformatics to filter out the useful data to gain valuable information. High dimensional big Media data like audios, images and videos are growing rapidly nowadays. Emerging with this increasingly growing volume of data is the need to retrieve relevant contents from such large databases. The fundamental scientific problem behind this need is the nearest neighbor search problem. Typical graph applications include predicting biological activity of molecules, identifying errors in computer programs, and categorizing scientific publications. Unlike traditional vector data, graphs are only characterized by node-edge representation and no features are readily available for training prediction models. Taxonomies are the key to developing successful applications in a domain, such as information retrieval, knowledge searching and classification. In particular, considering the ever-growing amount of text digital data per year, taxonomy learning from text is a primary research area for developing such applications nowadays. Kernel methods have emerged as a versatile mechanism to handle generic data. The growing interest in kernels is mainly motivated by the positive impact they have in important applications such as data clustering and classification.

II. METHODS AND MATERIAL

A. Proposed System

The idea of the project is to implement the query optimization using the Feedback collection such as positive & negative Feedback.The Learning process is made by Supervised Learning in this we know the result which is going to produce.For this Ranking and Learning Process is used.

IMAGE COLLECTION:For query q and an image+ collection, multiple search result lists can be derived using different search algorithms. Each search result list is a permutation/ranking of the N images sorted in descending order of their ranking scores, which are generated by the search algorithm. The ranking list variable is to denote a search result list. Assuming there are ranking lists generated for a query, they constitute a set of search result lists.

Ranking:For two ranking lists, the one with more relevant images ranked at the top gives a better performance than the one with fewer relevant images ranked at the top. If we have the ground truth label of each image (its relevance to query q), then it can be derived by using AP or NDCG and the best search result selection in problem is straightforward. However, in real applications, the ground truth, relevance labels for images are unavailable.

Due to the well-known semantic gap problem, any queries (especially the queries with large intra-class appearance variance) are hard to represent with descriptive visual features. The intuition behind this feature is that, for a good ranking result list, more relevant images have higher ranks. In other words, images belonging to the top part should share higher similarity than those inother parts. The re-ranking ability measures to what degree ranking can improve text-based search results. For a query, if its ranking ability is positive (suitable to be re-ranked), the re-ranking result list will be presented to users; otherwise the text-based search result list will be presented. In other words, the search engine can achieve a guaranteed performance enhancement by only referencing queries which are suitable for re-ranking while leaving the remaining unsuitable ones unchanged.

S.No	Meta Ranking	Visual Ranking
1	Histogram	Color
2	Corrlogram	Color Corrlogram
3	Color Moment	Shape and Edge

Table 1: Ranking information

Machine learning:In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

Backtracking : Backtracking is a general algorithm for finding all (or some) solutions to some computational problems, notably constraint satisfaction problems, that incrementally builds candidates to the solutions, and abandons each partial candidate c ("backtracks") as soon as it determines that c cannot possibly be completed to a valid solution.. In the common backtracking approach, the partial candidates are arrangements of k queens in the first k rows of the board, all in different rows and columns. Any partial solution that contains two mutually attacking queens can be abandoned.

Backtracking depends on user-given "**black box procedures**" that define the problem to be solved, the nature of the partial candidates, and how they are extended into complete candidates. It is therefore a meta heuristic rather than a specific algorithm – although, unlike many other meta-heuristics, it is guaranteed to find all solutions to a finite problem in a bounded amount of time. The rankings themselves are totally ordered. For example, materials are totally preordered by hardness, while degrees of hardness are totally ordered.

B. System Implementation

Create distributed storage server: Set of independent storages is designed to hold the large amount of files. These files are either uploaded from user or stored from any other external storage options. The independent storages are connected by using internet or by using the wired links. The meta data information is stored with the corresponding file in the data base

Image search query processing system

From the user end, query is collected from the user and the same is processed by the distributed server. The server searches the images based on the image file

name and its corresponding meta information. Query results are significantly large amount of results and hence results are ordered based on file name relevancy and displayed to the user.

It is crucial to understand the scope and nature of image data in order to determine the complexity of image search system design. The design is also largely influenced by factors such as the diversity of user-base and expected user traffic for a search system. Along this dimension, search data can be classified into the following categories:

Archives - usually contain large volumes of structured or semi-structured homogeneous data pertaining to specific topics.

Domain- Specific Collection - this is a homogeneous collection providing access to controlled users with very specific objectives. Examples of such a collection are biomedical and satellite image databases.

Enterprise Collection- A heterogeneous collection of images that is accessible to users within an organization's intranet. Pictures may be stored in many different locations.

Personal Collection- usually consists of a largely homogeneous collection and is generally small in size, accessible primarily to its owner, and usually stored on a local storage media.

Web- World Wide Web images are accessible to everyone with an Internet connection. These image collections are semi-structured, non-homogeneous and massive in volume, and are usually stored in large disk arrays.

C. Preference Learning Feature Extraction System

Learning model is designed by capturing the large set of training data that comprising of informative samples will lead to a high-performing model. Images are returned for difference types of queries have different visual distributions. As a consequence, it is designed to handle all queries using a universal preference learning model Ranking, Indexing and re-ranking with optimal search engine

Predefined query categorization is applied to categorize the image queries. From the visual distributions, visual similarities are identified and extracted using features of the images. The search result images are indexed using the indexing operation, and ranking is applied by estimating the relevancy of the search results to search query. A re-ranking process accomplished to improve the accuracy level of the search results.

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning', keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation. Additionally, the increase in social web applications and the semantic web have inspired the development of several web-based image annotation tools.

D. Image Segmentation

Kernel Fuzzy C Means is the improved version of fuzzy c means clustering algorithm to perform the segmentation process. KFCM is an iterative optimization algorithm which produces the optimized solution based on heuristic modeling. It produces the mathematical optimization by performing the local search operation to outcomes the best combination of clusters to obtain the best results. This iterative process begins with the arbitrary set of solutions to the clustering problem. Inevery iteration, an incremental modification is performed to the new solution. In this fuzzy operation, each and every data pixel point is belongs to the cluster centroid point. The association with each pixel point with centroid is defined by the membership function. This member function defines the strength of connectivity between pixel points and the cluster centroid point.

Similar to the FCM clustering,KFCM takes the degree of membership as the primary input to the objective function. And this evaluates the significance of the pixel point with the cluster groups. If the pixel point retains the higher significance in terms of membership

value,then the objective value of the membership function will be greater value.

KFCM make use of kernel distance function as objective function along with the membership function to find the best combination of centroids in the images.Kernel distance is extracted from the Euclidean distance between cluster centroid and pixel data points. This distance value is taken as input for the negative exponential with the dividing ratio of the sigmoid function. This Sigmoid function is the logistic function which divides he Euclidean distance. The following steps are executed to perform the segmentation in KFCM clustering approach. KFCM takes the de-noised images as input for the segmentation process.

III. RESULTS AND DISCUSSION

Support Vector Machine (SVM) Classifier: Support vector machines are best case supervised learning approach which is based on associated learning algorithms that is used to analyze data and recognize patterns, used for classification and regression analysis. There are two phases are executed in the classification process. That is, training phase and testing phase. Given an input set is trained with the SVM classifier to build the association between the input pixel points based on the extracted features. This classification process is executed in the form of non-probabilistic binary linear classification. It observes the replication, relativity connectivity and missed points. It formulates the group which consists of the linear representation for each row represents the class. This corresponds to the logical class for each row generalized.

Quad Programming (SVM): Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A First normal formal form of SVM with optimization is modeled as Quadratic Programming (QP) SVM. It is purely based on quadratic programming optimization. QP-SVM is a kernel based classifiers which accumulate the quadratic operations. The objective of this Quadratic programming has to solve the given problem by means of quadratic function which takes the decision based on the variable and constraints are alinear function of the variables It estimates the portfolio optimization variance based on the sum of the variances and covariance of individual values and the linear constraints which indicates the lower and upper boundary points.

Backtracking Using Feedback System: Once images are displayed as results for the search query, the relevancy of query results are estimated by collecting the feedback from the user. The feedback value is either positive or negative. From the collected feedback the query results are filtered and reordered to the next level query processing. Here the backtracking technique is applied to re-modify the results of the search queries that is produced based on ranking and collected feedbacks

S.NO	MSE	PSNR
1	1.32	4.96
2	3.54	42.67

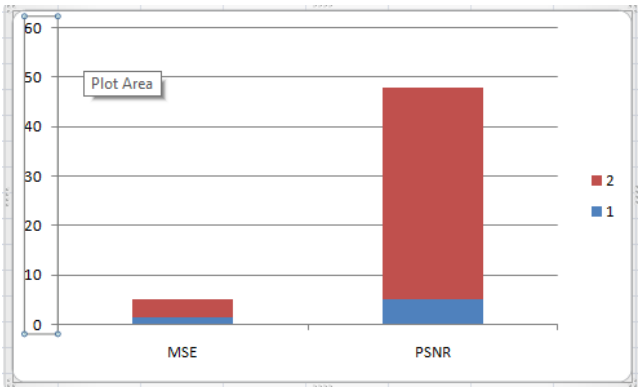

Figure 1. Graph

IV.CONCLUSION

Searching the image and producing the best results in the bigdata storage is theinnovative problem which is needed to solve by arobust solution. The proposed system is accomplished with the responsive feedback system for user suggestions. In order to produce the finest outcomes, the proposed model is executed based on User feedback system such as backtracking which improves accuracy and speed. Our proposed scheme has significant features like easier image storing and retrievel, precise query processing, ranking, indexing.

V. FUTURE ENHANCEMENT

An interesting extension of this work is to exploit the re-ranking ability measures to what degree ranking can improve both text-based and visual-based search results.the system may also allow extending the query

processing support to all file types. A second promising direction is the incorporation of relevance feedback from user to increase the retrieval performance of the system.

VI. REFERENCES

[1]. D. J. Abadi, D. Carney, U. Cetintemel, et al. Aurora: A New Model and Architecture for Data Stream Management. In VLDB Journal, 12(2):120–139, 2003.

[2]. B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and Issues in Data Stream Systems. In Symposium on Principles ofDatabase Systems (PODS), pages 1–16, 2002.

[3]. O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom. ULDBs: Databases with Uncertainty and Lineage. In International Confer-ence on Very Large Data Bases (VLDB), pages 953–964, 2006.

[4]. D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya. An Annotation Management System for Relational Databases. In International Conference on Very Large Data Bases (VLDB), pages 900–911, 2004.

[5]. P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquin. Incoop: Mapreduce for Incremental Computations. In ACM Sym-posium on Cloud Computing (SoCC), 2011.

[6]. O. Boykin, S. Ritchie, I. O'Connell, and J. Lin Summingbird: A Framework for Integrating Batch and Online MapReduce Compu-tations. In International Conference on Very Large Data Bases (VLDB), pages 1441–1451, 2014.

[7]. D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A Recursive Model for Graph Mining. In Fourth SIAM International Conferenceon Data Mining (SDM), pages 442–446, 2004.

[8]. B. Chandramouli, J. Goldstein, M. Barnett, R. DeLine, D. Fisher,

[9]. J. C. Platt, J. F. Terwilliger, J. Wernsing. Trill: A High-Performance Incremental Query Processor for Diverse Analytics. In InternationalConference on Very Large Data Bases (VLDB), pages 401–412, 2014.

[10]. S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin,

[11]. J. M. Hellerstein, W. Hong, S. Krishnamurthy, S. Madden, V. Ra-man, F. Reiss, and M. Shah. TelegraphCQ: Continuous Data flow Processing for an Uncertain World. In Conference on Innovative DataSystem Research (CIDR), 2003.

[12]. T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmeleegy, and R. Sears. Mapreduce Online. In USENIX Symposium onNetworked Systems Design and Implementation (NSDI), 10(4), 2010.