Efficient Nearest and Score Based Ranking for Keyword Search

[1]Rajkumar.R, [2]Manimekalai.P, [2]Mohanapriya.M, [3]Vimalarani.C

Abstract - Conventional spatial queries, such as range search and nearest neighbour retrieval, involve only conditions on objects' geometric properties. The proposed system uses an efficient algorithm to find the exact nearest neighbor based on the Euclidean distance for large-scale computer vision problems. We embed data points nonlinearly onto a low-dimensional space by simple computations and prove that the distance between two points in the embedded space is bounded by the distance in the original space. Instead of computing the distances in the high-dimensional original space to find the nearest neighbor, a lot of candidates are to be rejected based on the distances in the low-dimensional embedded space; due to this property, our algorithm is well-suited for high-dimensional and large-scale problems. We also show that our algorithm is improved further by partitioning input vectors recursively. Contrary to most of existing fast nearest neighbor search algorithms, our technique reports the exact nearest neighbor not an approximate one and requires a very simple preprocessing with no sophisticated data structures. We provide the theoretical analysis of our algorithm and evaluate its performance in synthetic and real data.

Index Terms - Keyword Search, Nearest Neighbor Search, Spatial Index.

I. INTRODUCTION

In modern century technology place an important role, it shows an immense presence of every person, and it has reduced human works. Under certain circumstances they act to be tedious, preferably to the middle educated people. Here in our nearest neighboring search, it finds the nearest locations available to the user. We may guess that we can find solution through internet. Yes it is possible, but it takes some smothering steps to locate our destination. It gives a lot of inappropriate details and it takes time to solve the process. This can be done when there is no emergency, but in the most of situation time place a predominant role.

People search instantaneously, so time act as a crucial factor. Here by using our method we can easily track down the exact place. This process involves the use of KNN algorithm, according to these algorithms the user enters the search area with keywords which are processed and it provides the list of areas which are closely related to the user. These two algorithms are synchronized with offline map. Hence through these maps the exact position of the user is identified and in accordance to that the search procedure takes place.

We propose an efficient algorithm to find the exact nearest neighbor based on the Euclidean distance for large-scale computer vision problems. We embed data points nonlinearly onto a low-dimensional space by simple computations and prove that the distance between two points in the embedded space is bounded by the distance in the original space. Instead of computing the distances in the high-dimensional original space to find the nearest neighbor, a lot of candidates are to be rejected based on the distances in the low-dimensional embedded space due to this property, our algorithm is well-suited for high-dimensional and large-scale problems.

We also show that our algorithm is improved further by partitioning input vectors recursively. Contrary to most of existing fast nearest neighbor search algorithms, our technique reports the exact nearest neighbor – not an approximate one – and requires a very simple preprocessing with no sophisticated data structures. We provide the theoretical analysis of our algorithm and evaluate its performance in synthetic and real data.

An increasing number of applications require the efficient execution of Nearest Neighbor (NN) queries constrained by the properties of the spatial objects. Due to the popularity of keyword search, particularly on the Internet, many of these applications allow the user to provide a list of keywords that the spatial objects (henceforth referred to simply as objects) should contain, in their description or other attribute. A spatial keyword query consists of a query area and a set of keywords. The answer is a list of objects ranked according to a combination of their distance to the query area and the relevance of their text description to the query keywords. The proposed system deals the spatial approximation string search based on the Euclidean space and road space.

Learning to rank is a kind of learning-based information retrieval techniques specialized in learning a ranking model with some documents labeled with their relevancies to some queries where the model is hopefully capable of ranking the documents returned to an arbitrary new query automatically. Various machine learning methods are Ranking SVM, Rank Boost, Rank Net, List Net, and Lambda Rank. The learning to rank algorithms has already shown their promising performances in information

Manuscript received Feb, 2014.

***Rajkumar.R**, Department of Computer Science and Engg., Anna University/ SNS College of Technology/ SNS Institution,Coimbatore, India, Mobile No:9597824661.*

***Manimekalai.P Mohanapriya.M,** SNS Department of Computer science and Engg., Anna University/ SNS College of Technology/ Coimbatore, India,*

***Vimalarani.CAP/CSE**, Department of Computer science and Engg.,,Anna University/ SNS College of Technology/ Coimbatore,India,*

retrieval, especially web search. However, as the emergence of domain-specific search engines, more attentions have moved from the broad-based search to specific verticals for hunting information constraint to a certain domain.

Different vertical search engines deal with different topicalities, document types or domain specific features. For example a medical search engine should clearly be specialized in terms of its topical focus, whereas a music, image or video search engine would concern only the documents in particular formats. Since currently the broad-based and vertical search engines are mostly based on text search techniques, the ranking model learned for broad based can be utilized directly to rank the documents for the verticals.

Most of the current image search engines only utilize the text information accompanying images as the ranking features, such as the Term Frequency (TF) of query word in image title, anchor text, alternative text, surrounding text, Uniform Resource Locator (URL). Therefore, web images are actually treated as text-based documents that share similar ranking features as the document or webpage ranking, and text-based ranking model can be applied here directly.

II. PROBLEM DEFINITIONS

Spatial search engine is used to retrieve the results for the requested query from the data base. Nowadays most of the spatial search engine retrieves the most rated results instead of the most desired results. Ultimately, users have to spend more time on searching for the desired information from the search results. In spatial database, it does not give real time answers. A spatial keyword query consists of a query area and a set of keywords. The answer is a list of objects ranked according to a combination of their distance to the query area and the relevance of their text description to the query keywords. The proposed system deals the spatial approximation string search based on the Euclidean space and road space. To enhance it, adapting reranking method is used as a proposed technique.

III. RELATED WORK

Section 3.1 reviews the Information retrieval R-Tree (IR^2-tree), which is the state of the art of answering the nearest neighbor queries has been defined.

A. The IR^2-tree

Many applications require finding objects closest to a specified location that contains a set of keywords. For example, online yellow pages allow users to specify an address and a set of keywords. In return, the user obtains a list of businesses whose description contains these keywords, ordered by their distance from the specified address. The problems of nearest neighbour search on spatial data and keyword search on text data have been extensively studied separately. However, to the best of our knowledge there is no efficient method to answer spatial keyword queries, that is, queries that specify both a location and a set of keywords. In this work, we present an efficient method to answer top-k spatial keyword queries. To do so, we introduce an indexing structure called IR^2-Tree [6] (Information Retrieval R-Tree) which combines an R-Tree with superimposed text signatures. We present algorithms that construct and maintain an IR^2-Tree, and use it to answer top-k spatial keyword queries. Our algorithms are experimentally compared to current methods and are shown to have superior performance and excellent scalability.

A spatial keyword query consists of a query area and a set of keywords. The answer is a list of objects ranked according to a combination of their distance to the query area and the relevance of their text description to the query keywords. A simple yet popular variant, which is used in our running example, is the distance-first spatial keyword query, where objects are ranked by distance and keywords are applied as a conjunctive filter to eliminate objects that do not contain them.

B. Ranking model adaptation

With the explosive emergence of vertical search domains, applying the broad-based ranking model directly to different domains is no longer desirable due to domain differences, while building a unique ranking model for each domain is both laborious for labelling data and time consuming for training models. In this paper, we address these difficulties by proposing a regularization-based algorithm called [2] ranking adaptation SVM (RA-SVM), through which we can adapt an existing ranking model to a new domain, so that the amount of labelled data and the training cost is reduced while the performance is still guaranteed. Our algorithm only requires the prediction from the existing ranking models, rather than their internal representations or the data from auxiliary domains.

Since currently the broad-based and vertical search engines are mostly based on text search techniques, the ranking model learned for broad based can be utilized directly to rank the documents for the verticals. For example, most of current image search engines only utilize the text information accompanying images as the ranking features, such as the term frequency (TF) of query word in image title, anchor text, alternative text, surrounding text, URL, and so on. Therefore, web images are actually treated as text-based documents that share similar [2] ranking features as the document or webpage ranking, and text-based ranking model can be applied here directly. However, the broad based ranking model is built upon the data from multiple domains, and therefore cannot generalize well for a particular domain with special search intentions.

C. Signature files

The signature-file access [5] method for text retrieval is studied. According to this method, documents are stored sequentially in the “text file”. Abstraction of the documents is stored in the “signature file”. The latter serves as a filter on retrieval: It helps in discarding a large number of nonqualifying documents. In this paper two methods for creating signatures are studied analytically [5], one based on word signatures and the other on superimposed coding. Closed form- formulas are derived for the false-drop probability of the two methods, factors that affect it are studied, and performance comparisons of the two methods based on these formulas are provided.

Traditional database management systems are designed for formatted records. Recently there seem to be many attempts to extend these systems so that they will be able to handle unformatted free text. The major application of such extended system is office automation.

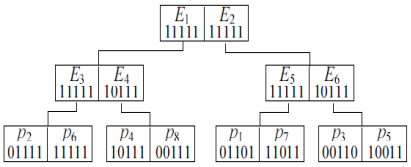

Fig 3.1 gives the signature of the entries

Much type of messages circulates in an office: correspondence, memos, reports, etc. Another important application of text retrieval method is the computerized library. We focus our attention on text retrieval methods only. It follows some methods to retrieve text such has, full text scanning, inversion, signature files, clustering, multi attribute hashing.

D. Ranking web pages

One of the greatest things about the Internet is that everyone can use it and owns it. It is a collection of networks, both big and small which can be shared worldwide. These networks connect in many different ways to form the single entity that we know as the [4] Internet. The Internet carries an extensive range of information resources and services, like as the linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support email. The terms Internet and World Wide Web are often used in everyday speech without much distinction, means, and these terms can be used vice versa. However, the Internet and the World Wide Web are not the same. The Internet can be a global system of interconnected computer networks. In contrast, the Web is one of the applications that run on the [4] Internet through web browser. It is a collection of text documents and other resources, which are linked through hyperlinks and URLs, usually accessed by web browsers from web servers. In short, the Web can be thought of as an application or services "running" on the Internet providing various information to the end user.

IV. K NEAREST NEIGHBOR ALGORITHM

We first present a K Nearest Neighbor algorithm, K Nearest Neighbour is a Lazy Learning Algorithm Defer the decision to generalize beyond the training examples till a new query is encountered we have a new point to classify, we find its K nearest neighbours from the training data.

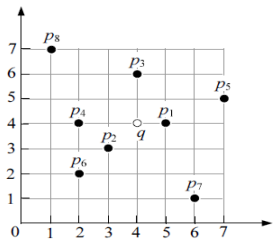

Fig 4.1 shows the locations of the points

If K = 5, then in this case query instance p will be classified as negative since three of its nearest neighbours are classified as negative. The distance is calculated by using the Euclidean Distance.

A. Euclidean Distance

Euclidean distance is the distance between two points in Euclidean space. Euclidean space was originally devised by the Greek mathematician Euclid around 300 B.C.E. to study the relationships between angles and distances. This system of geometry is still in use today and is the one that high school students study most often. Euclidean geometry specifically applies to spaces of two and three dimensions. However, it can easily be generalized to higher order dimensions.

The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them PQ. In Cartesian Coordinates if $\mathbf{p} = (p_1, p_2,..., p_n)$ and $\mathbf{q} = (q_1, q_2,..., q_n)$ are two points in Euclidean n-Space, then the distance from **p** to **q**, or from **q** to **p** is given by,

$$d(p,q)=d(q,p)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+.....+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

V. EXPERIMENTS

System implementation is the process of making the newly designed system fully operational and consistent in performance. That is, implementation is the process of having the personnel check out and put new equipment into use, train the users to use the new system and construct any file that are needed to use it. At this stage the main workload, the major impact on the existing practices shifts to the user department.

If the implementation is not carefully planned and controlled, it can cause chaws. Thus it can be considered to be the most crucial stage in achieving a successful new system and in giving the users confidence that the new system will work and be effective. Before the development of the system, the user specification, the forms are prepared. The user can specify the change if any, then the design department examines the changes and if accepted then the requirement of the user are taken care of. This is the stage where the system design begins the theoretical design is converted into a working system.

All the technical errors are fixed and the test data is entered. Then the reports are prepared and compared with that of the existing system. If the new system is not working properly, then once again we can go back to the existing system and after rectification; the new system can be installed.

System implementation is the important stage of project when the theoretical design is tuned into practical system. The main stages in the implementation are as follows:

- Planning
- Training
- System testing and
- Changeover Planning

Planning is the first task in the system implementation. Planning involves deciding on the method

and the time scale to be adopted. At the time of implementation of any system people from different departments and system analysis involve. To confirm the practical problem of controlling various activities of people outside their own data processing departments. The line managers controlled through an implementation coordinating committee. The committee considers ideas, problems and complaints of user department, it must also consider:

The implication of system environment

(i) Self-selection and allocation for implementation tasks
(ii) Consultation with unions and resources available
(iii) Standby facilities and channels of communication

System implementation covers a broad spectrum of activities from a detailed workflow analysis to the formal go-live of the new system. During system implementation organizations may refine the initial workflow analysis that had been completed as part of the requirements analysis phase. With the aid of the vendor they may also start mapping out the proposed new workflow. The system implementation phase requires the vendor to play a very prominent role. In addition to the workflow analysis it is during this phase that full system testing is completed. Other key activities that would occur during this phase include piloting of the new system, formal go-live and the immediate post implementation period during which any application issues are resolved. Systems Design will naturally lead to another stage where it becomes closer to the actual deployment of the planned software. Since the design is already there, developers have an idea on how the software actually looks like. The need is to put them all together to realize the intended software.

When look on to the comparisons between existing and proposed system, the first set of experiments is to compare the performance of different combinations of fast neighbor search and existing search strategies. All strategies are tested under two request patterns: data analysis and results.

In more specific the chapter particularly interested in the total number of results and search delay during a spatial data search and the average processing time of a data extraction since they are the dominant factors affecting service quality experienced by the users.

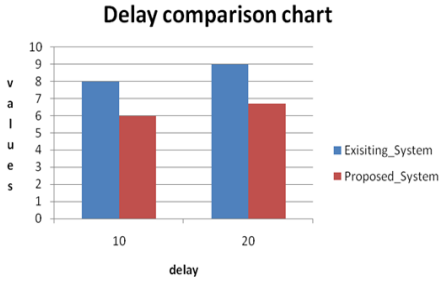

Fig 5.1 Comparison of time consumption

The above figure 5.1 shows that all strategies perform significantly better than traditional search strategy with simple look ahead k nearest neighbor.

VI. CONCLUSION

There are plenty of applications seen for calling a search engine that is able to efficiently support novel forms of spatial queries that are integrated with keyword search. The existing solutions to such queries either incur prohibitive space consumption or are unable to give real time answers. The proposed system has remedied the situation by developing an access method called the Spatial Inverted index (SI-index). Not only that the SI-index is fairly space economical, but also it has the ability to perform keyword-augmented nearest neighbor search in time that is at the order of dozens of milliseconds. Furthermore, as the SI-index is based on the conventional technology of inverted index, it is readily incorporable in a commercial search engine that applies massive parallelism, implying its immediate industrial merits.

By adapting KNN algorithm and ranking adaptation model, it finds the nearest neighbor information based on the user intension and it decreases the information retrieval time. Hence we are planning to propose the future enhancement as to reduce the information retrieval time gradually.

REFERENCES

[1] Agrawal .S, S. Chaudhuri, and G. Das. Dbxplorer (2002), In *Proc. Of International Conference on Data Engineering (ICDE)*, A system for keyword-based search over relational databases. pages 5–16.
[2] Bo Geng (April 2012), IEEE Transaction on knowledge and Data Engineering," Ranking model adaptation for specific search", Vol.24, No.4.
[3] Beckmann .N, H. Kriegel, R. Schneider, and B. Seeger (1990), In *Proc. of ACM Management of Data (SIGMOD)*, The R*-tree: An efficient and robust access method for points and rectangles. pages 322–331.
[4] Das Shivani Gupta .N. N, (July – 2013), International Journal of Engineering Research & Technology (IJERT), "An Algorithm For Ranking The Web Pages Of Search Engine**",** ISSN: 2278-0181, Vol. 2 Issue 7.
[5] Faloutsos .C and S. Christodoulakis (1984), *ACM Transactions on Information Systems (TOIS)*, Signature files: An access method for documents and its analytical performance evaluation. 2(4):267–288.
[6] Felipe I.D, V.Hristidis (2008), In Proc. Of International Conference on Data Engineering (ICDE), "Keyword search on spatial databases". Pages 656-665.
[7] Herbrich.R, T. Graepel, and K. Obermayer (2000), "Large Margin Rank Boundaries for Ordinal Regression," Advances in Large Margin Classifiers, pp. 115-132, MIT Press.
[8] Hjaltason G.R. and H. Samet (1999), *ACM Transactions on Database Systems (TODS)*, Distance browsing in spatial databases. 24(2):265–318.
[9] Hristidis. V and Y. Papakonstantinou (2002), In *Proc. of Very Large Data Bases (VLDB)*, Discover: Keyword search in relational databases. pages 670–681.
[10] Lu.J, Y. Lu, and G. Cong (2011). Reverse spatial and textual k nearest neighbor search. In *Proc. of ACM Management of Data (SIGMOD)*, pages 349–360.
[11] Yue.Y, T. Finley, F. Radlinski, and T. Joachims (2007), "A Support Vector Method for Optimizing Average Precision," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 271-278.
[12] Zhou.Y, X. Xie, C. Wang, Y. Gong, and W.-Y (2005). Ma. Hybrid index structures for location-based web search. In *Proc. of Conference on Information and Knowledge Management (CIKM)*, pages 155– 162.

Rajkumar.R received his diploma degree in computer science and engineering in the year of 2010 and his currently pursuing B.E Computer science and Engineering in SNS College of Technology under Anna University, Chennai.

Manimekalai.P, she is currently pursuing B.E. Computer science and Engineering in SNS College of Technology Under Anna University, Chennai

