Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.578 – 584

RESEARCH ARTICLE

Fastest Nearest Neighbour Search Using Keywords

Ameya Patil, Aafra Qazi, Akshay Patil, Amar Thombare, (Professor)Nilesh Rathod
Department of Information Technology,Rajiv Gandhi Institute of Technology,Mumbai India
ameya1092@yahoo.co.in, aafra_q@yahoo.com, kshpatil432@gmail.com,
thombareamar8@gmail.com, nilesh_rahod2004@yahoo.co.in

***Abstract*: We propose to develop a new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data, and comes with algorithms that can answer nearest neighbour queries with keywords in real time. As verified by experiments, the proposed techniques outperform the IR2-tree in query response time significantly, often by a factor of orders of magnitude.**

A spatial database manages multidimensional objects (such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modelling entities of reality in a geometric manner.

Keyword search in document performed with various approaches ranked retrieval results, clustering search results & identifying the nearest neighbour Keyword search on xml document categorized as two different approaches one is Keyword search on xml document which can be performed by ranking the searched results based on match or the answer to keyword & finding the nearest neighbour of keyword by using GST or by Xpath Query.

Keywords: real time data, multidimensional data, Data Mining, spatial, clustering, keyword

I. INTRODUCTION

A. Introduction to area:

Data mining is a new powerful technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as the fast analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized.

B. Fast analysis:

Data mining automates the process of shifting through historical data in order to discover new information. So this is one of the main differences between data mining and statistics. Here a model is usually devised by a statistician to deal with a specific analysis problem. It also differentiates data mining from expert systems and the model is built by a knowledge engineer from rules extracted from the experience of an expert.

C. Large or complex data sets:

One of the attractions of data mining is that it makes it possible to analyst very large data sets in a lesser time scale. Also Data mining is suitable for complex problems involving relatively small amounts of data with relatively many fields or variables to analyze. However there may be simpler, cheaper and more effective solutions for small and relatively simple data analysis problems.

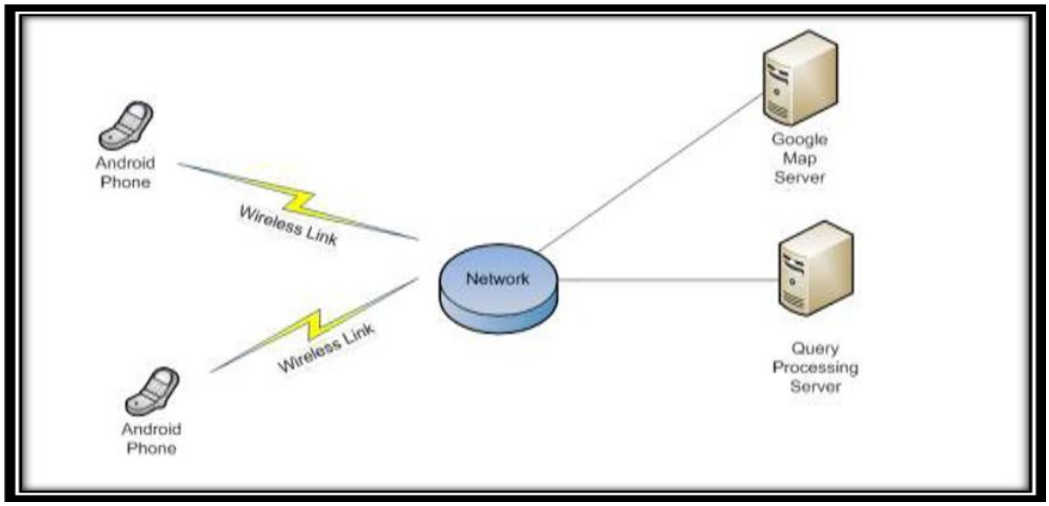

Fig1. Block Diagram.

II. ENTERPRISE OBJECTIVES

To design a variant of inverted index that is optimized for multidimensional points. This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme. To provide efficiently a support for novel forms of spatial queries that are integrated with keyword search.

We have 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 neighbour search in time that is at the order of dozens of milliseconds.

III. SCOPE OF THE PROJECT

1. The concept or scope of the project is very vast. It could be implemented to various fields just as the need be.

2. The areas to which fastest nearest Neighbor search can be implemented is Hotels, Medical services and Health Centre, Entertainment and etc.

3. With any field that we select, a more broader scope can be obtained. If for the field of Entertainment, the sub categories could be Cinemas, Gaming Centers, Live concerts.

4. The application can be used by anybody, it has no limitations restricting the usage to any age bar or group of people.

IV. METHODOLOGY

Android is an operating system for mobile devices such as smart phones and tablet computers. It is developed by the Open Handset Alliance led by Google. Google released most of the Android code under the Apache License, a free software license. The Android Open Source Project (AOSP) is tasked with the maintenance and further development of Android. Android consists of a kernel based on the Linux kernel, with middleware, libraries and APIs written in C and application software running on an application framework which includes Java-compatible libraries based on Apache Harmony. Android uses the Dalvik virtual machine with just-in-time compilation to run compiled Java code. Android has a large community of developers writing applications ("apps") that extend the functionality of the devices. Developers write primarily in a customized version of Java. Apps can be downloaded from third-party sites or through online stores such as Android Market, the app store run by Google.

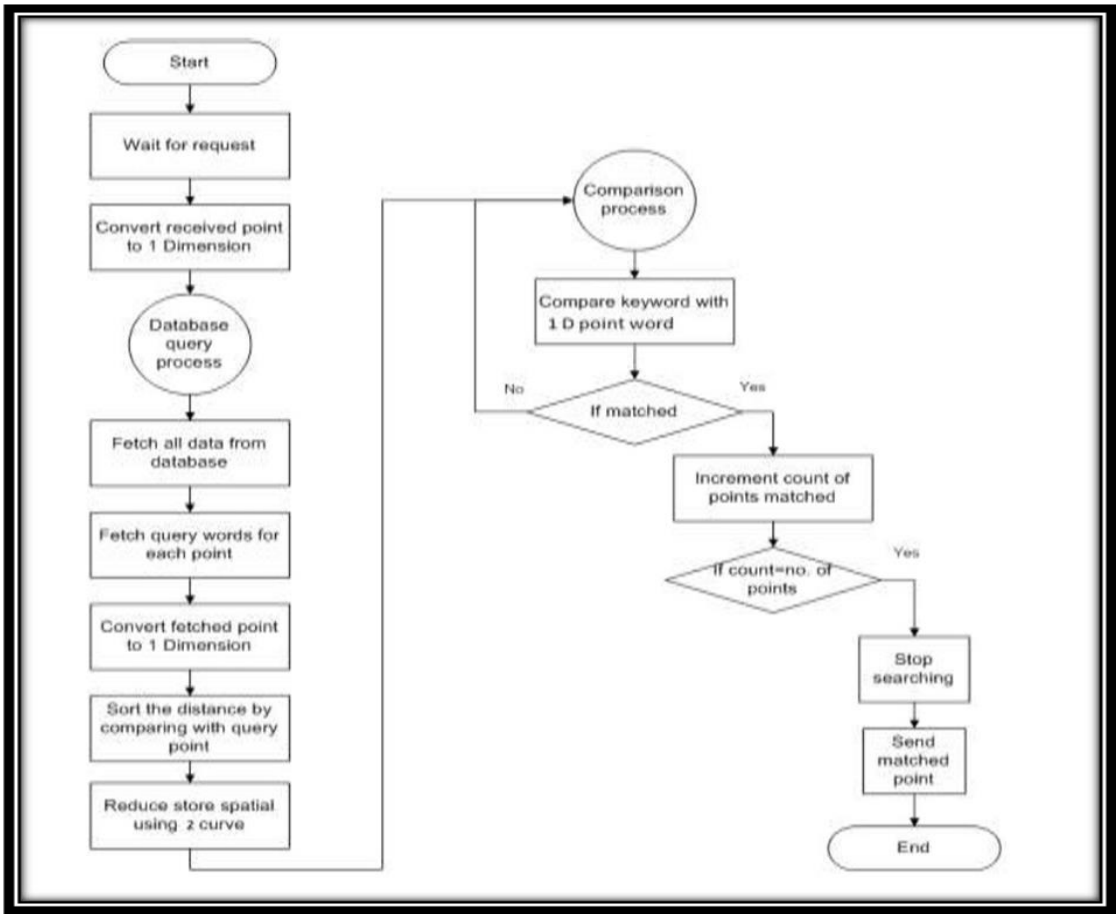

Fig2. System Flow

Operations:

- **Classification and prediction:**

Classification is most ordinarily supported operation by industrial data processing tools. This operation allows organizations to get patterns in massive or complicated information sets so as to resolve specific business issues. Classification is that the method of sub dividing knowledge set with relation to variety of specific outcomes.

- **Clustering:**

Clustering is associate unsupervised operation. It is used after you would like to seek out groupings of comparable records in your information with none preconditions on what that similarity could involve. Agglomeration is employed to spot attention-grabbing teams during a client base that will not are recognized before.

- **Association analysis and sequential analysis:**

Association Analysis is an unattended sort of data processing that appears for links between records during an information set. Association analysis is typically named as market

basket analysis its commonest application. The aim is to get that things are usually purchased at a similar time to assist retailers organize client incentive schemes and store layouts additional expeditiously.

- **Forecasting:**

Classification identifies a particular cluster or category to that an item belongs. A prediction supported a classification model thus it will be a separate outcome, distinctive a client as a communicator or non-responder, or a patient as having a high or low risk of heart condition. In distinction prognostication considerations the prediction of continuous values like share values, the amount of the stock exchange, or the longer term value of an artefact like oil. Data processing tools also can offer prognostication functions.

Techniques and Algorithms:

A data mining operation is achieved exploitation one amongst variety of techniques or ways. Every technique will itself be enforced in numerous ways in which, employing a kind of algorithms.

- **Clustering Algorithms:**

Cluster analysis is that the method of distinctive the relation- ships that exist between things on the idea of their similarity and difference. In contrast to classification, cluster doesn't need a target variable to be known beforehand. A cluster algorithmic rule takes Associate in Nursing unbiased investigate the potential groupings at intervals an information set Associate in Nursing makes an attempt to derive an optimum delineation of things on the idea of these teams. To spot things that belong to a cluster, some live should be used that gauges the similarity between things at intervals a cluster and their difference to things in alternative clusters. The similarity and difference between things is often measured as their distance from one another and from the cluster canters at intervals a mulch- dimensional house, wherever every dimension represents one in all the variables being compared.

- **Nearest Neighbour:**

Nearest Neighbour could be a prophetical technique appropriate for classification models. Not like alternative prophetical algorithms, the coaching information is not scanned or processed to form the model. Instead, the coaching information is that the model. Once a brand new case or instance is conferred to the model, the algorithmic rule appearance in the

slightest points the information to search out a set of cases that are most almost like it and uses them to predict the result. There are two principal drivers within the k-NN algorithm: the quantity of nearest cases to be used (k) and a metric to live what‟s meant by nearest. Every use of the k-NN algorithmic rule needs that we tend to specify a positive number worth for k. This determines what number existing cases are checked out once predicting a brand new case. K-NN refers to a family of algorithms that we tend to may denote as 1-NN, 2-NN, and 3- NN, so forth. For instance, 4-NN indicates that the algorithmic rule can use the four nearest cases to predict the result of a brand new case. K-NN is predicated on a thought of distance, and this needs a metric to work out distances.

- **Neural Networks:**

A Neural Network could be a set of connected input/output units wherever every association includes a weight associated with it. Throughout the learning section the network learns by adjusting the weights thus on are able to predict the proper category label of the input samples. Neural network learning is additionally observed as connectionless learning because of the connections between units. A key distinction between neural networks and lots of different techniques is that neural nets solely operate directly on numbers. As a result, any non numeric information in either the freelance or dependent (output) columns should be reborn to numbers before we are able to use the info with a Neural Network.

V. Conclusion

In this research paper we have focused on getting the best access to hospital and various medical services. This methodology will ease the search time for the user in an event of emergency when the location of the client would be tracked GPS and the nearest medical help can be sought easily.

REFERENCES

[1] S. Agrawal, S. Chaudhuri, and G. Das.Dbxplorer: A system for keyword-based search over relationaldatabases. In Proc. of International Conference on Data Engineering (ICDE), pages 5–16, 2002.
[2] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*tree: An efficient and robust access method for pointsand rectangles. In Proc. of ACM Management of Data (SIGMOD), pages 322–331, 1990.
[3] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan.Keyword searching and browsing in databasesusing banks. In Proc. of International Conference on Data Engineering (ICDE), pages 431–440, 2002.

[4] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu. Spatial keyword querying. InER, pages 16–29, 2012.
[5] X. Cao, G. Cong, and C. S. Jensen.Retrieving top-k prestige-based relevant spatial web objects. PVLDB, 3(1):373–384, 2010.

