ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

Fast Searching With Keywords Using Data Mining

Chandrashekhar

M.Tech (CS) IVth SEM KLESCET Belgaum

Under the Guidance of Prof. Gambhir Halse

Abstract: **Conventional spatial queries, such as vary search and nearest neighbour retrieval, involve solely conditions on objects' geometric properties. Today, several trendy applications call for novel varieties of queries that aim to seek out objects satisfying both a spatial predicate, and a predicate on their associated texts. As an example, rather than considering all the restaurants, a nearest neighbour query would instead invite the restaurant that is the nearest among those whose menus contain "steak, spaghetti, brandy" all at identical time. Presently the simplest solution to such queries relies on the IR2-tree, which, as shown in this paper, includes a few deficiencies that seriously impact its efficiency. impelled by this, we have a tendency to develop a brand new access methodology called the spatial inverted index that extends the standard inverted index to address multidimensional information, and comes with algorithms which will answer nearest neighbour queries with keywords in real time. As verified by experiments, the projected techniques outperform the IR2- tree in query reaction time significantly, typically by an element of orders of magnitude.**

Keywords: **Information Retrieval Tree, Keyword Search, Spatial Inverted Index.**

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. Some of the key elements that make data mining tools a distinct form of software are:

• 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.

• 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 analyse. However there may be simpler, cheaper and more effective solutions for small and relatively simple data analysis problems.

• Discovering significant patterns or trends that would otherwise go unrecognized:

The main goal of data mining is to unearth relationships in data so that it provides useful insights. Data mining tools can also automate the process of finding predictive information in large databases. Data mining techniques can yield the benefits of automation on existing software and hardware platforms to enhance the value of existing information resources, and can be implemented on new products and systems.

The core components of data mining technology have been under development for decades in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

a) To understand the different facets of data mining is to distinguish between data mining applications, operations, techniques and algorithms.

1) Applications: A data mining application is associate implementation of knowledge mining technology that solves a particular business or analysis downside. Example application areas include:

• A drug company will analyse its recent sales department activity and their results to boost targeting of high-value physicians and verify that selling activities can have the best impact within the next few months. The information has to embrace challenger market activity still as information regarding the native health care systems. The results will be distributed to the sales department via a wide-area network that permits the representatives to review the recommendations from the angle of the key attributes within the call method. The continuing, dynamic analysis of the information warehouse permits best practices from throughout the organization to be applied in specific sales things.

• A MasterCard company will leverage its large warehouse of client dealing knowledge to spot customers possibly to have an interest in a very new credit product. Employing a little check mailing, the attributes of shoppers with associate affinity for the merchandise will be known. Recent comes have indicated over a twenty fold decrease in prices for targeted mailing campaigns over standard approaches.

• A heterogeneous company with an oversized direct sales department will apply data processing to spot the most effective prospects for its services. Mistreatment data processing to investigate its own client expertise, this company will build a singular segmentation characteristic the attributes of high- value prospects.

• A giant shopper package merchandise company will apply (data mining/data methodology) to boost its sales process to retailers. Knowledge from shopper panels, shipments, and challenger activity will be applied to grasp the explanations for complete and store shift. Through this analysis, the manufacturer will choose promotional ways that best reach their target client segments.

2) 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. For an example, We would need to classify our customers into 'high' and 'low' classes with relation to credit risk. The class or 'class' into that every client is placed is that the 'outcome' of our classification. The foremost common techniques for classification square measure call trees and neural networks. If a call tree is employed, it will offer a group of branching conditions that in turn split the purchasers into teams outlined by the values within the freelance variables. The aim is to supply a group of rules or a model of some kind that may determine a high share of responders.

• 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. For a an example, it will be accustomed determine similarities in customers' phone usage, so as to plot and market new decision services. Clustering is sometimes achieved mistreatment applied math strategies, like a k-means algorithmic rule, or a special kind of neural network

referred to as a Kohonen feature map network. The fundamental operation is that the same. Every record is compared with a collection of existing clusters that are outlined by their center. A record is assigned to the cluster it is nearest to, and this successively changes the worth that defines that cluster. Multiple passes are created through a knowledge set to re assign records associated modify the cluster centers till an optimum answer is found.

• 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.

3) 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.

• Naive-Bayes:

Naive-Bayes could be a classification technique that is each prophetic and descriptive. It analysis the connection between

every experimental variable and also the variable to derive a chance for every relationship. Nave-Bayes needs just one experience the coaching set to come up with a classification model. This makes it the foremost economic data processing technique. However, Naive-Bayes doesn't handle continuous information, therefore any freelance or dependent variables that contain continuous values should be binned or bracketed.

- **Decision Trees:**

Decision trees are one among the foremost common data processing technique and therefore the best liked in tools geared toward the business user. They are simple to line up, their results are graspable by AN end-user, they will address a large vary of classification issues, they are strong within the face of various knowledge distributions and formats, and that they are effective in analysing giant numbers of fields. a call tree rule works by cacophony a knowledge set so as to make a model that with success classifies every record in terms of a target field or variable. The foremost common forms of call tree rule are CHAID, CART and C4.5. CHAID (Chi- square automatic interaction detection) and CART (Classification and Regression Trees) were developed by statisticians. CHAID will turn out tree with multiple sub-nodes for every split. CART needs less knowledge preparation than CHAID, however produces solely two-way splits. C4.5 comes from the globe of machine learning, and relies on scientific theory.

B. Concepts/Basic of topics:

A spatial database is used to store large amount of space related data such as maps, medical imaging data etc. and manages multidimensional objects (such as points, rectangles, etc.), and provides quick access to those objects based on different choice criteria. The importance of spatial databases is it provides a convenient way to model the entities of reality in a geometric manner[key-1]. As an example, locations of restaurants, hotels, hospitals and then on are typically shown as points in an exceedingly map, whereas larger extents like parks, lakes, and landscapes typically as a mixture of rectangles. Several functionalities of a spatial information are helpful in numerous ways that in specific contexts. For example, in an exceedingly geographic data system, vary search may be deployed to search out all restaurants in a certain space, whereas nearest neighbour retrieval will discover the restaurant nearest to a given address. Today, the widespread use of search engines has created it realistic to write down spatial queries in an exceedingly novel approach. Conventionally, queries focus on objects' geometric properties solely, like whether or not a point is in a rectangle, or however close two points are from one another. We have seen some trendy applications that have an ability to pick objects supported each of their geometric coordinates and their associated texts. As an example, Search engine will be fairly useful if it finds nearest restaurant that will offers the demanded food. Note that this is often not the "globally" nearest building (which would are came back by a standard nearest neighbour query), however the closest restaurant among solely those providing all the demanded foods and drinks. During this paper, we tend to design a variant of inverted index that's optimized for multidimensional points, and is therefore named the spatial inverted index (SI-index). This access technique with success incorporates point coordinates into a standard inverted index with little additional space, attributable to a delicate compact storage theme. Meanwhile, an SI-index preserves the spatial locality of information points, and comes with an R-tree designed on each inverted list at space overhead.

C. Issues and Challenges:

- Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other.
- Some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts.
- The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs.

D. Problem Statement:

Today, the widespread use of search engines has made it realistic to write spatial queries in a brand new way.

Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other. We have seen some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts. For example, It would be fairly useful if a search engine can be used to find the nearest restaurant that offers "steak, spaghetti, and brandy" all at the same time. Note that this is not the "globally" nearest restaurant (which would have been returned by a traditional nearest neighbour query), but the nearest restaurant among only those providing all the demanded foods and drinks. There are easy ways to support queries that combine spatial and text features. For example, for the above query, we could first fetch all the restaurants whose menus contain the set of keywords {steak, spaghetti, brandy}, and then from the retrieved restaurants, find the nearest one. Similarly, one could also do it reversely by targeting first the spatial conditions – browse all the restaurants in ascending order of their distances to the query point until encountering one whose menu has all the keywords. The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs. A typical example is that the real nearest neighbour lies quite far-away from the query point, while all the closer neighbours are missing at least one of the query keywords.

E. Aim and 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.

II. LITERATURE SURVEY

A. Previous Research work:

Many applications need finding objects that are nearest to a given location that contains a group of keywords. An increasing variety of applications need the economical execution of nearest neighbour (NN) queries affected by the properties of the spatial objects. Owing to the recognition of keyword search, notably on the net, several of those applications enable the user to produce a listing of keywords that the spatial objects ought to contain, in their description or alternative attribute. For example, real estate websites enable users to go looking for properties with specific keywords in their description and rank them in line with their distance from a given location. We tend to decision such queries spatial keyword queries [12]. A spatial keyword query consists of a query space and a group of keywords. The solution could be a list of objects hierarchical in line with a mix of their distance to the query space and also the connection of their text description to the query keywords. A simple nevertheless widespread variant that is employed is the distance-first spatial keyword query, wherever objects square measure hierarchical by distance and keywords square measure applied as a conjunctive filter to eliminate objects that don't contain them. Unfortunately there's no economical support for top-k abstraction keyword queries, wherever a prefix of the results list is needed. Instead, current systems use ad-hoc combos of nearest neighbour (NN) and keyword search techniques to tackle the matter. For an example, associate points R-Tree is employed to seek out the closest neighbours associate points for every neighbour an inverted index is employed to envision if the query keywords area unit contained. The economical methodology to answer top-k spatial keyword queries is predicated on the combination of knowledge structures and algorithms utilized in spatial in- formation search and Information Retrieval (IR). Particularly, the strategy consists of building an Information Retrieval R- Tree (IR2- Tree) that could be a structure supported the R- Tree. At query time an incremental algorithm is employed that uses the IR2-Tree to efficiently produce the top results of the query. The IR2-Tree is a R-Tree wherever a signature is supplementary to every node v of the IR2- Tree to denote the matter content of all spatial objects within the sub tree non-moving at 'v'. The top-k spatial keyword search formula that is impressed by the work of Hjaltason and Samet [14] exploits this data to find the highest query results by accessing a bottom portion of the IR2-Tree.

This work has the subsequent contributions:

• The matter of top-k spatial keyword search is outlined.

• The IR2-Tree is as an economical categorization structure to store spatial and matter data for a group of objects.

Economical algorithms also are bestowed to take care of the IR2-Tree, that is, insert and delete objects.

• An economical progressive formula is bestowed to answer top-k spatial keyword queries mistreatment the IR2-Tree.

The IR2-Tree could be a combination of a R-Tree and signature files. Specially, every node of an IR2-Tree contains each abstraction and keyword information, the previous within the kind of a minimum bounding space and therefore the latter within the kind of a signature. An IR2-Tree facilitates each top-k abstraction queries and top-k abstraction keyword queries. R-tree, a preferred special index, and signature file, a good technique for keyword-based document retrieval. IR2-tree (Information Retrieval R-Tree) structure developed, that has the strengths of each R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity that is the key to determination spatial queries expeditiously. Like signature files, the IR2-tree is in a position to filter a substantial portion of the objects that does not contain all the query keywords, so considerably reducing the amount of objects to be examined.

R-TREE:

R-Tree makes use of solely Associate in Nursing R-Tree organization [16]. Given a distance-first top-k spatial keyword query, the algorithmic rule initial finds the top-1 nearest neighbour object to the query purpose Q.p. Then it retrieves that object (since the R-tree solely contains object point- ers) and compares that object's matter description with the keywords of the query. If the comparison fails then that object is discarded, and therefore the next nearest object is retrieved. The progressive NN algorithmic rule is employed. This method continues till Associate in nursing object is found whose matter description contains the query keywords. Once a satisfying object is found it's came back and therefore the method repeats till k objects are came back. The drawback of this algorithmic rule is that it's to retrieve each object came back by the NN algorithmic rule till the top-k result objects are found. This doubtless will result in the retrieval of the many "useless" objects. Within the worst case (when none of the objects satisfies the query's keywords) the whole tree must be traversed and each object must be inspected.

SIGNATURE FILE:

Signature files were introduced by Faloutsos and Christodoulakis [11] as a technique to with efficiency search a group of text documents. Signature files appear to be a promising access technique for text and attributes. Consistent with this technique, the documents (or records) are keep consecutive in one file ("text file"), whereas abstractions of the documents ("signatures") are keep consecutive in another file ("signature file"). So as to resolve a query, the signature file is scanned 1st, and plenty of no qualifying documents are at once rejected. In general signature file refers to a hashing-based framework, whose internal representation in keyword search on spatial information is thought as superimposed committal to writing (SC). It's designed to perform membership tests that confirm whether or not a query word w exists in a very set W of words. SC is conservative, within the sense that if it says "no", then w is certainly not in W. on the opposite hand, if SC returns "yes", truth answer will be either manner, during which case the total W should be scanned to avoid a false hit.

In the context of keyword search on spatial information [12], SC works within the same means because the classic technique of bloom filters. In pre-processing, it builds somewhat signature of length 'l' from W by hashing every word in 'W' to a string of l bits, and so taking the disjunction of all bit strings. For instance, allow us to denote by h(w) the bit string of a word w. First, all the l bits of h(w) are initialized to zero. Then, SC repeats the subsequent m times: randomly choose a bit and set it to one. Terribly significantly, organization should use w as its seed to make sure that constant w forever finishes up with a identical h(w). Moreover, the m selections are mutually independent and will even happen to be constant bit. The concrete values of l and m have an effect on the space price and false hit chance.

Table I

Word	Hashed Bit String
A	101
B	1001
C	11
D	110
E	10010

Example of bit string computation with 1 = five and m=two

Table 1 gives an example to illustrate above process, assuming l = five and m = two. For an example, within the bit string h(a) of a, the third and fifth (counting from left) bits area unit set to one. As mentioned earlier, the bit signature of a collection W of words merely OR's the bit strings of all the members of W. As an example, the signature of a collection {a,b} equals 01101, whereas that of {b,d} equals 01111. Given a query keyword w, SC performs the membership ,take a look at in W by checking whether or not all the 1's of h(w) seem at constant positions within the signature of W. If not, it's secured that w cannot belong to W. Otherwise, the take a look at can't be resolved victimization solely the signature, and a scan of W follows. A false hit happens if the scan reveals that W truly doesn't contain w. As an example, assume that we wish to check whether or not word c could be a member of set {a,b} using the set's signature 01101. Since the fourth little bit of h(c) = 00011 is one however that of 01101 is 0, SC in real time reports "no". As another example, take into account the membership take a look at of c in {b,d} whose signature is 01111. This time, SC returns "yes" as a result of 01111 has 1's in the least the bits wherever h(c) is about to 1; as a result, a full scan of the set is needed to verify that this can be a false hit.

The inverted index system may be a central element of a typical computer programme categorization formula. A goal of a research engine performance is to optimize the speed of the query: realize the documents wherever word happens. Once associate index is developed, that provisions lists of words per document; it's next inverted to develop associate inverted index. Querying the index would need serial iteration through every document and to every word to verify an identical document. The time, memory and process property to execute such a query don't seem to be forever in theory realistic. Rather than listing the words per article within the index, the inverted index system is developed that lists the documents per word. The inverted index made, the query will currently be determined by jumping to the word id within the inverted index. These were effectively inverted indexes with a little quantity of supplementary clarification that needed an implausible quantity of decide to turn out.

In NN process with IR2-tree, a point retrieved from the index should be verified (i.e., having its text description loaded and checked). Verification is additionally necessary with I- index, except for precisely the opposite reason. For IR2-tree, verification is as a result of we have a tendency to don't have the careful texts of a points, whereas for I-index, it's as a result of we have a tendency to don't have the coordinates. Specifically, given associate points NN query Q with keyword set Wq, the query rule of I-index initial retrieves (by merging) the set Pq of all points that have all the keywords of Wq, and then, performs |Pq| random I/Os to get the coordinates of every purpose in Pq so as to judge its distance to Q. When Wq has solely one word, the performance of I-index is extremely dangerous, that is anticipated as a result of everything within the inverted list of that word should be verified. apparently, because the size of Wq will increase, the performance gap between I index and IR2-tree keeps narrowing specified I- index even starts to trounce IR2-tree at |Wq| = four. This can be not as shocking because it could appear. As |Wq| grows, not several objects have to be compelled to be verified as a result of the quantity of objects carrying all the query keywords drops rapidly. On the opposite hand, at now a advantage of I index starts to pay off. That is, scanning associate points inverted list is comparatively low cost as a result of it involves solely successive I/Os , as against the random nature of accessing the nodes of associate points IR2-tree. The spatial inverted list (SI-index) is essentially a compressed version of an I-index with embedded coordinates. Query processing with an SI-index can be done either by merging or together with R-trees in a distance browsing manner. Furthermore, the compression eliminates the defect of a conventional I index such that an SI-index consumes much less space.

B. Existing System and Its Effects:

The best method to date for nearest neighbour search with keywords is due to Felipe ET AL [12]. They nicely integrate two well-known concepts: R-tree [2], a popular spatial index, and signature file [11], an effective method for keyword-based document retrieval. By doing so they develop a structure called the IR2-tree [12], which has the strengths of both R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files, the IR2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined. The IR2-tree, however, also inherits a drawback of signature files: false hits. That is, a signature file, due to its conservative nature, may still direct the search to some objects, even though they do not have all the keywords. The penalty thus

caused is the need to verify an object whose satisfying a query or not cannot be resolved using only its signature, but requires loading its full text description, which is expensive due to the resulting random accesses. It is noteworthy that the false hit problem is not specific only to signature files, but also exists in other methods for approximate set membership tests with compact storage. Therefore, the problem cannot be remedied by simply replacing signature file with any of those methods.

C. Proposed System and Its Advantages:

In this project, we design a variant of inverted index that is optimized for multidimensional points, and is thus named the spatial inverted index (SI-index). This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme. Meanwhile, an SI-index preserves the spatial locality of data points, and comes with an R-tree built on every inverted list at little space overhead.

As a result, it offers two competing ways for query processing.

- We can (sequentially) merge multiple lists very much like merging traditional inverted lists by ids.
- Alternatively, we can also leverage the R-trees to browse the points of all relevant lists in ascending order of their distances to the query point.

III. SOFTWARE REQUIREMENT SPECIFICATION

A Software Requirements Specification (SRS) is a complete description of the behaviour of the system to be developed. It includes the functional and non-functional requirement for the software to be developed. The functional requirement includes what the software should do and non-functional requirement include the constraint on the design or implementation of the system. Requirements must be measurable, testable, related to identified needs or opportunities, and defined to a level of detail sufficient for system design. What the software has to do is directly perceived by its users either human users or other software systems. The common understanding between the user and developer is captured in requirements document. The writing of software requirement specification reduces development effort, as careful review of the document can reveal omissions, misunderstandings, and inconsistencies early in the development cycle when these problems are easier to correct. The SRS discusses the product but not the project that developed it; hence the SRS serves as a basis for later enhancement of the finished product. The SRS may need to be altered, but it does provide a foundation for continued production evaluation.

A. Overall Description

Hardware Requirements

- Processor: Pentium 3 or More.
- Ram: 512MB or More.
- Disk Space: 200MB.
- Android Mobile Phone

Software Requirements

- JAVA (JDK 1.6 or More).
- Net Beans IDE 6.1 or More.
- Eclipse Juno.
- Android SDK.
- Programming language JAVA.

B. Specific Requirements

1) Functional Requirements: The functional requirements for a system describe what the system should do. These requirements depend on the type of software being developed, the expected users of the software and the general approach taken by the organization when writing requirements. When expressed as user requirements, the requirements are usually described in a fairly abstract way. However, functional system requirements describe the system function in detail, its inputs and outputs, exceptions, and so on.

Functional requirements are as follows:

- The developed system should be able to perform keyword-augmented nearest neighbour search in time.
- It should be incorporable in a commercial search engine that applies massive parallelism, implying its immediate industrial merits.

2) Non-Functional Requirement: Non-functional requirements, as the name suggests, are requirements that are not directly concerned with the specific functions delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy.

Alternatively, they may define constraints on the system such as the capabilities of I/O devices and the data representations used in system interfaces.

- SI-index can be done either by merging, or together with R-trees in a distance browsing manner.
- Compression eliminates the defect of a conventional I- index such that an SI-index consumes much less space.

IV. SYSTEM DESIGN

Design is one of the most important phases of software development. The design is a creative process in which a system organization is established that will satisfy the functional and non-functional system requirements. Large Systems are always decomposed into sub-systems that provide some related set of services.

A. Design Consideration

The purpose of the design is to plan the solution of the problem specified by the requirements document. This phase is the first step in moving from problem to the solution domain. The design of the system is perhaps the most critical factor affecting the quality of the software and has a major impact on the later phases, particularly testing and maintenance. System design describes all the major data structure, file format, output as well as major modules in the system and their Specification is decided.

1) Development Methods: The development method used in this software design is the modular/functional development method. In this, the system is broken into different modules, with a certain amount of dependency among them.

The system has the following modules:

- Location Manager
- HTTP Communicator
- Spatial Inverted Index
- Spatial Data Display

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

B. System Design

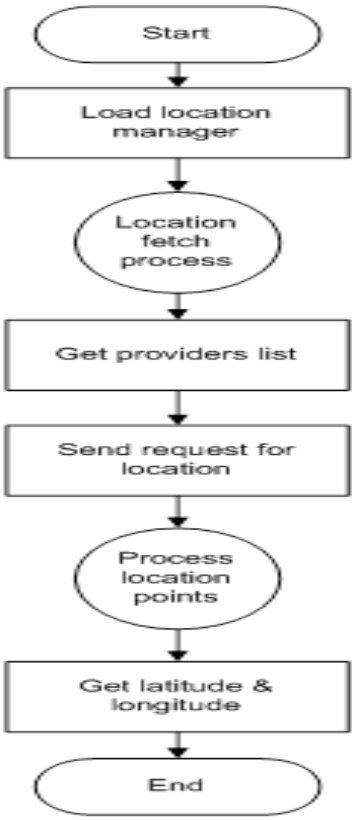

Figure 1. FC of Location Manager

1) Location Manager Module: The Figure shows the Flow chart of Location Manager Module. This is first module of the system and is used to get the current location of the user that will serve as input to the next module. As show in the figure initially it will load the location manager and location fetch process will start. Then it will send the location request and in response it will get latitude and longitude of the location.

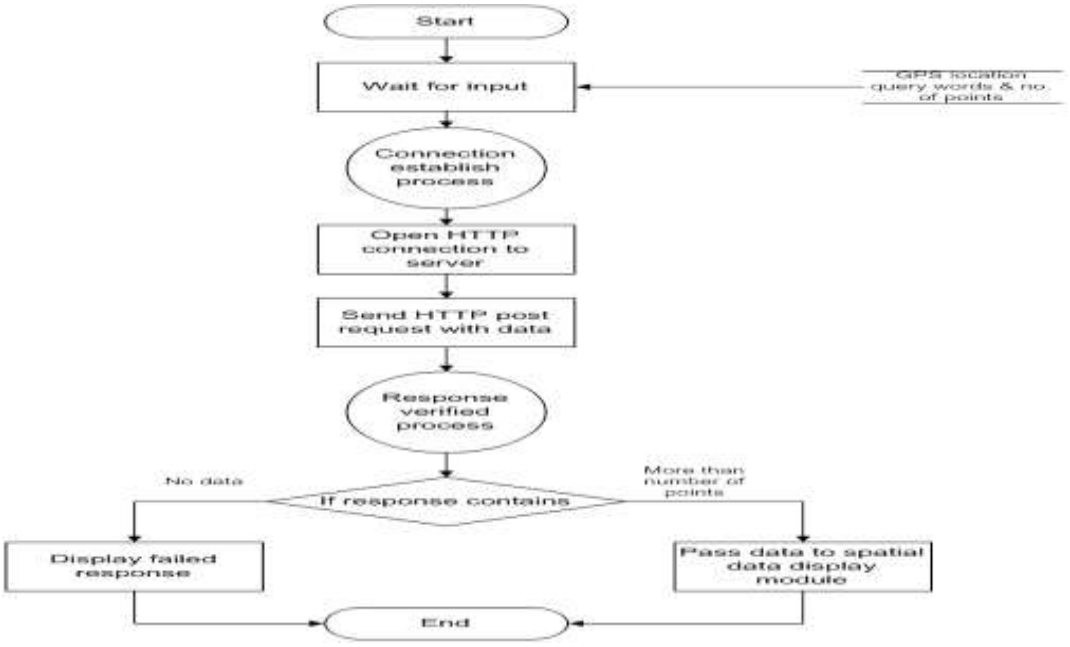

Figure 2. FC of HTTP Communicator

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

2) HTTP Communicator Module: The Figure shows the Flow Chart of HTTP Communicator Module. This module is used to send the request and to receive the response. It will take GPS Location, Query Keywords and Number of Minimum points as input and then establishes the connection to the server and sends the request to severs. It will then wait for the response if the response contains the data then it will pass the data to the Spatial Data Display module.

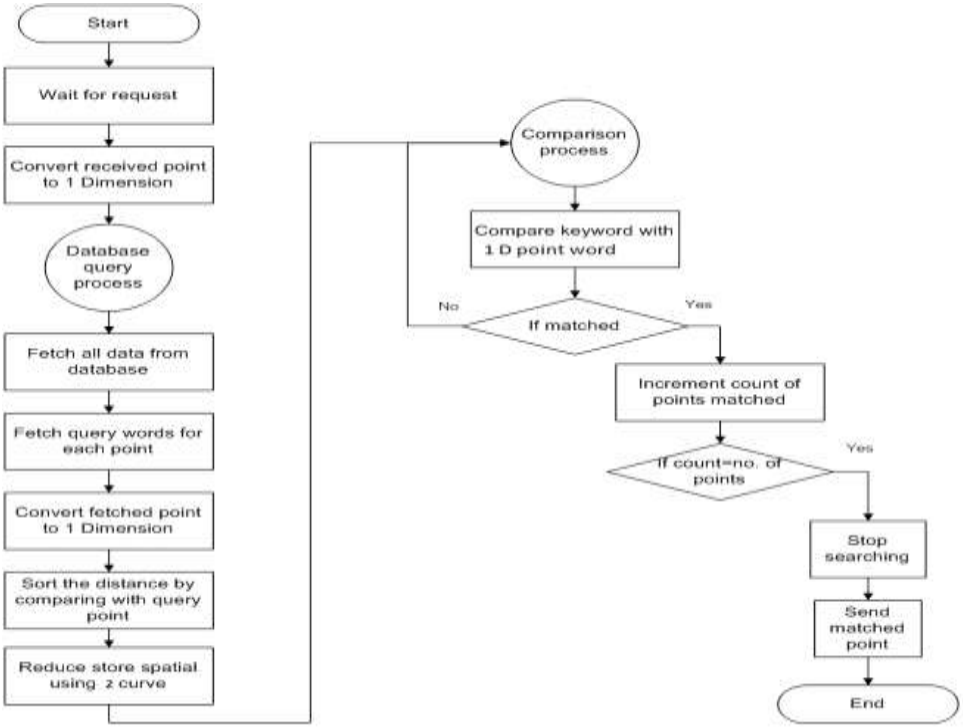

Figure 3. FC of Spatial Inverted Index Module

3) Spatial Inverted Index Module: The Figure shows the Flow Chart of Spatial Inverted Index Module. This is the most important module of the system. Initially it waits for the request from HTTP Communicator module and converts the points into one dimension. Database query process will begins and will fetch the all data from the database including query words for each point and converts the points into one dimension. Then it will sort the distance by comparing with the query points. Next comparison process will start, it will compare the keywords with one dimensional point word, if it matches then increment the count of points matched and if count becomes to number of point it will stop searching and send the matched points.

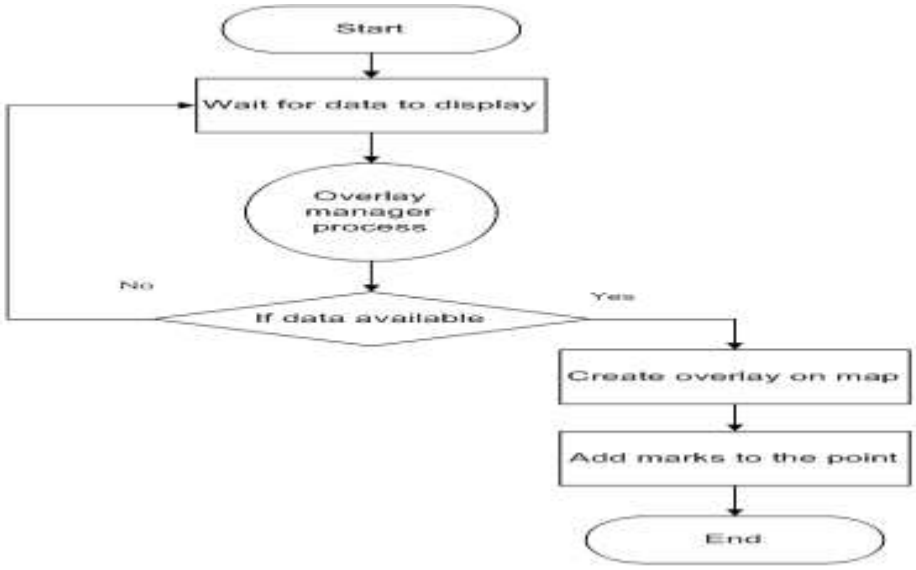

Figure 4. FC of Spatial Data Display Module

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

4) Spatial Data Display Module: The Figure shows the Flow Chart of Spatial Data Display Module. This is the final module of the system. It waits for the data to be displayed from HTTP Communicator module and then initiates Overlay manager process if the data is found and then it creates overlay on the map.

C. System Architecture

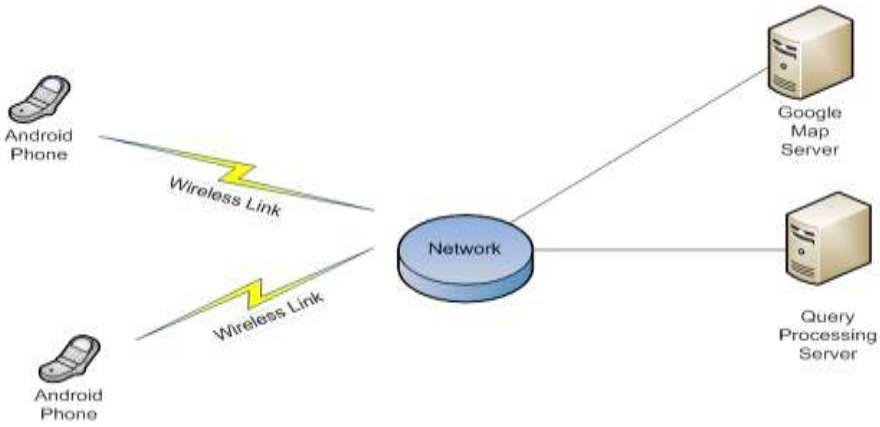

Figure 5. System Architecture

The Figure shows the System Architecture. The above diagram shows the working of our approach which involves a Query processing sever, Google map server and an android mobile phone. The query processing server stores the spatial data (Latitude, Longitude) of a place and information about a place such items available at that place. The mobile phone performs the search on the data stored on the server by using the phone's current position and keywords and retrieves the resulted nearest Neighbour's and display's the places on the map by interacting with the Google map server.

D. Data flow diagram

A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system. Data Flow models are used to show how data flows through a sequence of processing steps. The data is transformed at each step before moving on to the next stage. These processing steps or transformations are program functions when Data Flow diagrams are used to document a software design. DFD diagram is composed of four elements, which are process, data flow, external entity and data store.

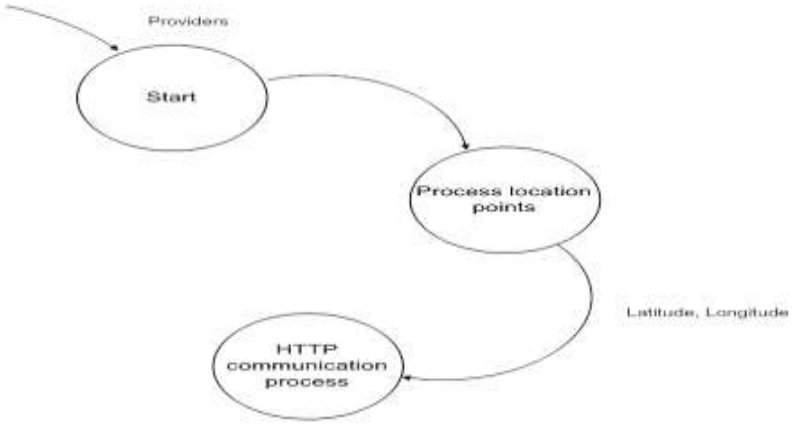

Figure 6. DFD of Location Manager Module

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

1) DFD for Location Manager Module: The Figure shows Data Flow Diagram of Location Manager Module. This module will gives the current location of the user (Latitude, Longitude) and the output of this module is given as input to the next module.

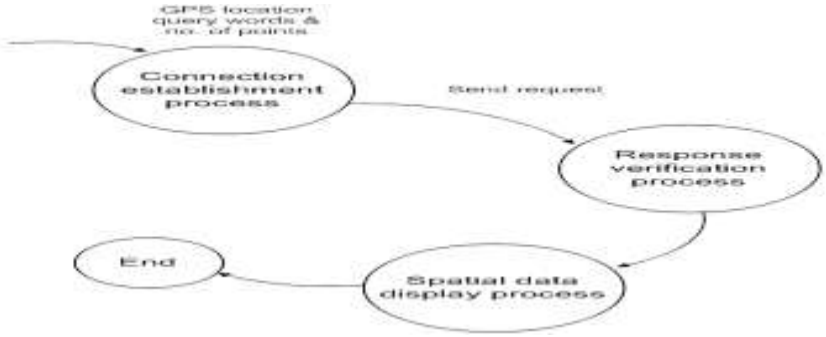

Figure 7. DFD of HTTP Communicator Module

2) DFD for HTTP Communicator Module: The Figure shows the Data Flow Diagram of HTTP Communicator Module. This module will take GPS Location, Query Keywords and Number of Minimum points as input. Then Connection Establishment Process will takes place to the Server and output of this module is fed as input to Spatial Data Display Module.

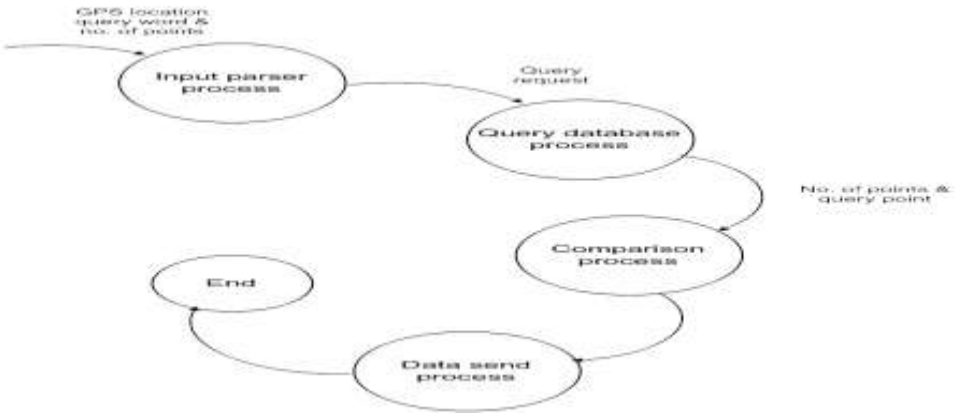

Figure 8. DFD of Spatial Inverted Index

3) DFD for Spatial Inverted Index: The Figure shows Data Flow Diagram of Spatial Inverted Index. This is One of the main module among the all. This will take GPS Location, Query Keywords and Number of Minimum points as input. Then it will query the database and comparison process will start. It will compare inputs with data stored in database and then sends the result to the HTTP communicator module.

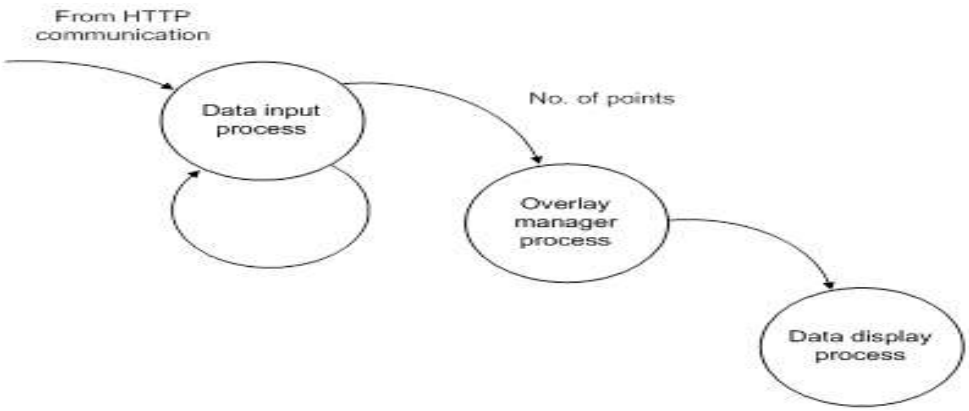

Figure 9. DFD of Spatial Data Display Module

4) DFD for Spatial Data Display Module: The Figure shows the Data Flow Diagram of Spatial Data Display Module. This module take input (Result) from the HTTP Communicator module, Overlay manager will start and displays the result on the map.

E. Sequence diagram

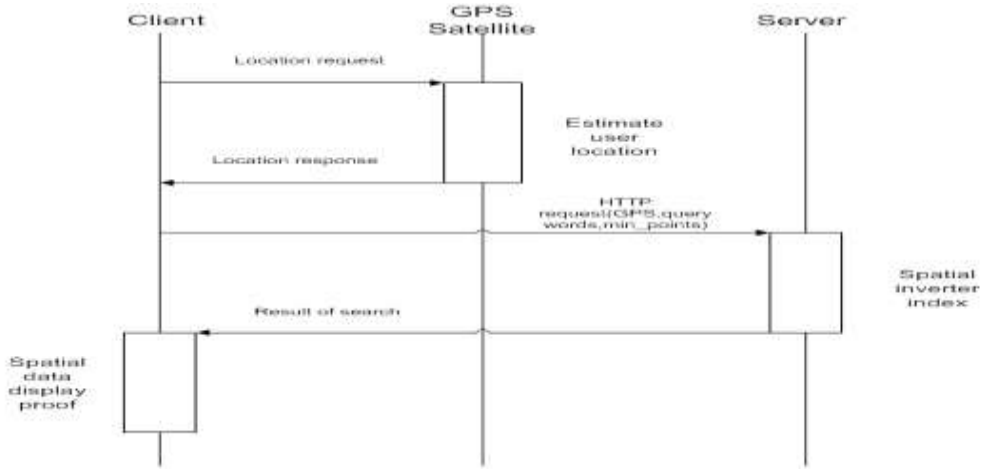

Figure 10. Sequence Diagram

The Figure shows the Sequence diagram of System. As shown in the figure, initially client will send the location request to the GPS Satellite and gets the current location. Then Client will send HTTP request containing Location, Keywords, and number of minimum points as input to the server. Server will take the input and compare with data stored in database and gives final result and is displayed on map.

V. IMPLEMENTATION/SIMULATION

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 that the first task within the system implementation. Designing involves picking the method and also the duration to be adopted. At the time of implementation of any system people from completely different departments and system analysis involve. to verify the sensible drawback of dominant numerous activities of individuals outside their own

processing departments. The road managers controlled through an implementation coordinative committee.

The committee considers ideas, issues and complaints of user department, it should conjointly consider:

The implication of system atmosphere

(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 close work flow analysis to the formal go-live of the new system. Throughout system implementation organizations could refine the initial work flow analysis that had been completed as a part of the requirements analysis section. With the help of the seller they will additionally begin mapping out the planned new work-flow. The system implementation section needs the seller to play a really outstanding role. Additionally to the work-flow analysis it's throughout this section that full system testing is completed. Alternative key activities that will occur throughout this section include piloting of the new system, formal go-live and also the immediate post implementation amount throughout that any application problems are resolved. Systems design can naturally cause another stage wherever it becomes nearer to the particular preparation of the planned software package. Since the planning is already there, developers have an inspiration on however the software package really seems like. the requirement is to place all of them along to understand the supposed software package.

The Modules specified in the design are implemented by using fallowing tools.

- **HTML:**

HTML means Hypertext mark-up language .HTML is method of describing the format of document ,which allow them to be viewed on the computer screen .HTML documents are dis- played by web browser ,programs which can navigate across networks and display a wide variety of types of information. HTML page can be developed to be a simple text or to be complex multimedia containing sound, moving ,images ,virtual reality ,and java applets. The global publishing format of internet is HTML. It allows authors to use not only text but also format that text with heading ,lists, and tables .Readers can access the pages of information from anywhere in the world at a click of mouse button.HTML pages can also be used for entering the data as a front end for commercial transactions.

- **Java Script:**

Java Script is fairly simple language ,which is only suitable for fairly simple tasks. The language is best suited for task ,which runs for short time ,and is most commonly used to manipulate the piece of document object model. The idea behind finding the java script is to find the language which could be used to provide client side browser application but which was not complicated as Java. Java Script is Netscape cross platform object oriented scripting language. Core Java Script contains a core set of objects such as array, date and Math and core set of language elements such as operators, control structure and statements .It is mainly used here for validation.

- **Java:**

It is a product of Sun Micro system and is purely object oriented programing language mainly used for development of internet, web based applications. It uses C++ style syntax but has been designed to be much easier to use. Java language was designed to be small, simple, and portable across platforms and operating systems, both at the source and at the binary level . Applets appear in a Web page much in the same way as images do, but unlike images, applets are dynamic and interactive. Applets can be used to create animations, figures, or areas that can respond to input from the reader, games, or other interactive effects on the same Web pages among the text and graphics. Platform-independence is a program's capability of moving easily from one computer system to another. Platform independence is one of the most significant advantages that Java has over other programming languages, particularly for systems that need to work on many different platforms. x Java is platform-independent at both the source and the binary level.

Advantages:

1) Java is simple: No language is simple, but Java considered a much simpler and easy to use object-oriented programming language when compared to the popular programming language, C++. Partially modelled after C++, Java has replaced the complexity of multiple inheritance in C++ with a simple structure called interface, and also has eliminated the use of pointers.

2) Java is Object-oriented: Object oriented programming models the real world. Everything in the world can be modelled as an object.

3) Java is Distributed: Distributed computing involves several computers on a network working together. Java is designed to make distributed computing easy with the networking capability that is inherently integrated into it. Writing network programs in Java is like sending and receiving data to and from a file.

4) Portability: Program once, Run anywhere (Platform In- dependence)

5) Java is Interpreted: An interpreter is needed in order to run Java programs. The programs are compiled into Java Virtual Machine code called byte code. The byte code is machine independent and is able to run on any machine that has a Java interpreter.

6) Java is one of the first programming languages to consider security as part of its design.

7) Java is Robust, Multithreaded, Portable, Reliable etc.

- **Database:**

A database is collection of related information. Defining a database involves specifying the data type, attributes and constraints for data to be stored. Constructing a database is process of storing data itself on some storage medium like disk or tape that can be handled by DBMS. Manipulating database includes such function like querying a database to retrieve specific data, updating the database to reflect the change in the mini world and generating report from the data. A DBMS represents complex relationship with data.

Database constitutes the primary data resource in enterprise application. The JDBC API facilitates access to relational data from Java. This API provides cross vendor connectivity and data access across relational database from different vendor. A database vendor such as Oracle, Sybase typically provides set of proprietary API for accessing the data managed by the database server. Client applications written in C/C++ can make use of these API calls for database access directly. The JDBC API provides a Java language alternative to these vendor specific API. JDBC API does not eliminate the access to the native API's for database access ,the implementation of JDBC API still make these native calls for data access. The JDBC is a middle layer that translates the JDBC calls to the vendor specific API. The JVM uses the JDBC driver to translate the generalized JDBC calls into vendor specific database call.

- **JSP:**

A Java Server pages is a template for a web pages that uses Java code to generate HTML document dynamically. JSP's are run in server side component known as JSP container, which translate them into equivalent Java Servlets. For this reason JSP pages and servlets are intimately related. The Source code for the system is organized in various file and is compiled using the JAVA C utility provided in Java.

VI. RESULT ANALYSIS

When look on to the comparisons between existing and proposed system, the primary set of experiments is to check the performance of various mixtures of fast neighbour search and existing search methods. All methods are tested below two request patterns: information analysis and results. In additional specific the chapter particularly curious about the overall number of results and search delay during a spatial data search and also the average interval of an information extraction since they are the dominant factors affecting service quality experienced by the users.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014, Available at: www.researchpublish.com

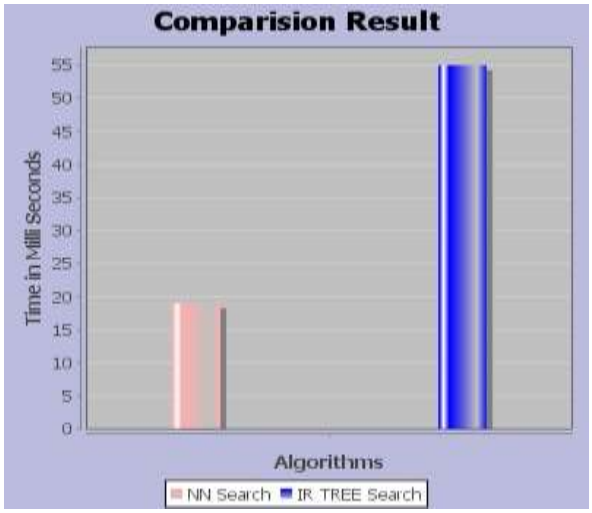

Figure 11. Comparison of time Consumption

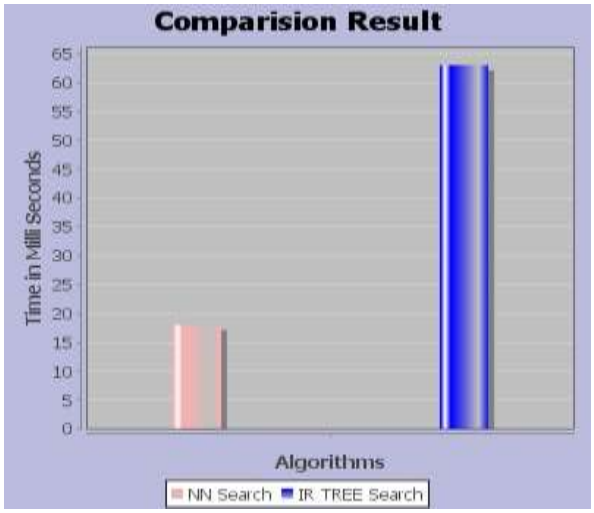

Figure 12. Comparison of time Consumption

The above figures shows that in all strategies the proposed system perform significantly better than the existing system.

VII. CONCLUSION

There are many applications seen for calling a search engine that's ready to with efficiency support novel varieties of abstraction queries that are integrated with keyword search. The present solutions to such queries either incur preventative space consumption or are unable to provide real time answers. The planned system has remedied the situation by developing an access methodology referred to as the abstraction Inverted index (SI-index). Not solely that the SI-index is fairly space economical, however additionally it's the flexibility to perform keyword-augmented nearest neighbour search in time that's at the order of dozens of milliseconds. Moreover, because the SI- index relies on the standard technology of inverted index, it's readily incorporable in a business search engine that applies huge similarity, implying its immediate industrial merits.

REFERENCES

[1] S. Agrawal, S. Chaudhuri, and G. Das. Dbxplorer: A system for keyword-based search over relational databases. 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 points and 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 databases using 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. In ER, 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.

[6] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In Proc. of ACM Management of Data (SIG- MOD), pages 373–384, 2011.

[7] B. Chazelle, J. Kilian, R. Rubinfeld, and A. Tal. The bloomier filter: an efficient data structure for static support lookup tables. In Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 30–39, 2004.

[8] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In Proc. of ACM Management of Data (SIGMOD), pages 277–288, 2006.

[9] E. Chu, A. Baid, X. Chai, A. Doan, and J. Naughton. Combining keyword search and forms for ad hoc querying of databases. In Proc. of ACM Management of Data (SIGMOD), 2009.

[10] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. PVLDB, 2(1):337–348, 2009.

[11] C. Faloutsos and S. Christodoulakis. Signature files: An access method for documents and its analytical performance evaluation. ACM Trans- actions on Information Systems (TOIS), 2(4):267–288, 1984.

[12] I. D. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In Proc. of International Conference on Data Engineering (ICDE), pages 656–665, 2008.

[13] R. Hariharan, B. Hore, C. Li, and S. Mehrotra. Processing spatial keyword (SK) queries in geographic information retrieval (GIR) systems. In Proc. of Scientific and Statistical Database Management (SSDBM), 2007.

[14] G. R. Hjaltason and H. Samet. Distance browsing in spatial databases. ACM Transactions on Database Systems (TODS), 24(2):265–318, 1999.

[15] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. In Proc. of Very Large Data Bases (VLDB), pages 670–681, 2002.

[16] I. Kamel and C. Faloutsos. Hilbert R-tree: An improved r-tree using fractals. In Proc. of Very Large Data Bases (VLDB), pages 500–509, 1994.