IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-12 E-ISSN: 2347-2693

Implementation of Nearest Neighbor Retrieval

Reddy S.P.[1*] and Govindarajulu P[2]

[1*]Department of Computer Science, Sri Venkateswara University, Tirupati, India
[2]Department of Computer Science, Sri Venkateswara University, Tirupati, India

[*]*Corresponding Author: pg.sunilkumar@gmail.comm*

Available online at: www.ijcseonline.org

Received: 02/Feb/2017 Revised: 07/Feb/2017 Accepted: 20/Feb/2017 Published: 28/Feb/2017

***Abstract*—Conventional pensiveness queries, like contrast search and nearby neighbor retrieval involve completely on conditions imposed on objects of geometric properties. Nowadays, various applications absorb new types of queries that aspire to hunt out objects satisfying every generalization predicate and a predicate on connected texts. as Associate in Nursing example, instead of considering all the restaurants, a nearest neighbor question would instead elicit the edifice that is the utmost among those whose menu contain "steak, spaghetti, sprite" all at a similar time. Presently the foremost effective resolution to such queries is based on the IR2-tree, which, as shown throughout this paper, aims at a couple of deficiencies that seriously impact its efficiency. motivated by this, we have a tendency to tend to develop a replacement access methodology called the abstraction inverted index with the intention of extends the quality inverted index to deal with flat data, and comes with algorithms that will answer nearby neighbor queries through keywords in real time. As verified by experiments, the projected techniques outgo the IR2-tree and are subjected to significantly, generally by a component of, orders of magnitude.**

***Keywords*- SI Index,IR Tree, Fast Nearest, Neighbor**

I. INTRODUCTION

An abstraction info manages two-dimensional things (such as points, rectangles, etc.), and provides swift access to those items supported entirely diverse preference criteria. The consequence of abstraction databases is mirrored by the expediency of modeling entities of authenticity in a very geometric manner. For instance, locations of restaurants, hotels, hospitals then on are typically depicted as points in a very map, while larger extents like parks, lakes, and landscapes classically as a mix of rectangles. Several functionalities of generalization info are supportive in various ways in which the explicit contexts. As an instance, in an earth science system, the search will be deployed to search out all restaurants in a incredibly certain space, whereas nearby neighbor retrieval will determine the edifice highest to a given address. Today, the prevalent use of search engines has formed it realistic to put in script pensiveness queries in a exceedingly greenhorn method. Predictably, queries contemplate on objects' geometric properties exclusively, like whether or not various extents is in a very parallelogram, or though shut two points are from one another. We've got seen some fashionable applications that decision for the flexibility to pick out objects supported each of their geometric coordinates and their associated texts.

For instance, it might be fairly helpful if a pursuit engine will be accustomed realize the closest edifice that gives "steak, spaghetti, and sprite" all at a similar time. Note with the intention of this is frequently not the "globally" nearby edifice (which would came reverse by a conventional nearest neighbor query), nevertheless the nearby edifice among solely given that all the demanded foods and drinks.

In this work, we be inclined to style a alternative of inverted index that is optimized for two-dimensional points, and is so named the pensiveness inverted index (SI-index). This access method with triumph incorporates purpose coordinates into a standard inverted index with petite further house, attributable to a fragile compact storage theme. Meanwhile, Associate in SI-index preserves the pensiveness neighborhood of in sequence points; Associate in R-tree is engineered on each reversed list at very little abode overhead. As a result, it offers 2 aggressive ways in which for question progression. We are able to (sequentially) merge multiple lists noticeably like merging ancient inverted lists by ids. Instead, we are able to conjointly leverage the R-trees to browse the purposes of all appropriate lists in rising order of their distances to the subject point. As per the experiments, the SI-index considerably outperforms the IR a pair of -tree in question potency, typically by an element of orders of magnitude.

II. SURVEY

The IR^2 - Tree

The IR2-tree [1] includes the R-tree with signature files. Next, we'll review what a signature file is before explaining the most points of IR2-trees. Our discussion assumes the information of R-trees and conjointly the best-first algorithm [2] for NN search, every of these unit of measurement well-known techniques in special databases. Signature undergo refers to a hashing-based framework, whose illustration in [3] is understood as superimposed writing (SC), that's shown to be a lot of sensible than various instantiations [4]. It's designed to perform membership tests: ensure whether or not or not a look word w exists during a set W of words. SC is conservative, means if it says "no", then w is definitely not in

W. If, on the contradictory hand, SC returns "yes", verity answer could also be either methodology, throughout that case the whole W ought to be scanned to avoid a false hit. inside the context of [5], SC works inside constant methodology as a result of the classic technique of bloom filters.

In pre-processing, it builds to alittle degree signature of length l from W by hashing each word in W to a string of l bits, and then taking the disjunction of [6] all bit strings. for instance, denote by h(w) the bit string of a word w. First, all the l bits of h(w) unit of measurement initialized to zero. Then, SC repeats the following m times: indiscriminately opts to alittle degree and set it to 1 and also the organization ought to use w as its seed to verify that an analogous w invariably finally [7] finally ends up with a fair h(w). Moreover, the m selections unit of measurement reciprocally freelance, and will even happen to be an analogous bit. The concrete values of l and m have a sway on the house value and false hit likelihood, as square measure mentioned later.

word	*hashed bit string*
a	00101
b	01001
c	00011
d	00110
e	10010

Figure 1 . bit string computation

Fig offers academic degree example as Associate in Nursing instance the on prime of methodology, assume l = five and m = two. as Associate in Nursing example, at intervals the bit string h(a) of a, the third and fifth (counting from left) bits unit set to 1. As mentioned earlier, the bit signature of a set W of words just ORs the bit strings of all the members of W. as an example, the signature of a set (a; b) equals 01101, whereas that of (b;a) metric weight unit equals 01111. Given an issue keyword w, SC performs the membership, and appears in W by checking whether or not or not all the 1s of h(w) appear at constant positions at intervals the signature of W. If not, it's guaranteed [8] that w cannot belong to W. Otherwise, it can not be resolved exploitation alone the signature, and a scan of W follows.

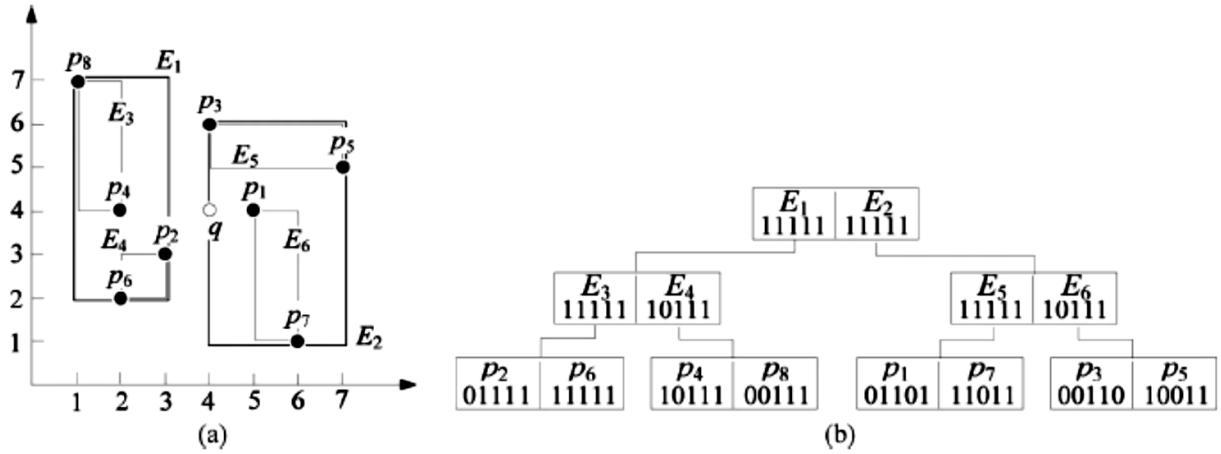


Figure 2. examples of IR^2 Tree (a) MBRs of R-Tree (b) Signatures of entries

Solutions Based on Inverted Indexes

Inverted indexes (I-index) have tested to be AN economical access technique for keyword-based document retrieval. at intervals the spatial context, nothing prevents North yankee country from treating the text description Wp of a degree p as a document, and then, building associate I-index. Fig.3 illustrates the index for the knowledge set of Fig. 1. each word at intervals the vocabulary has associate inverted list, enumerating the ids of the points that have the word in their documents.

word	*inverted list*
a	p_1 p_4
b	p_1 p_2 p_7
c	p_5 p_6 p_8
d	p_2 p_3 p_6 p_8
e	p_4 p_5 p_6 p_7

Figure 3. Examples of inverted index

Note that the list of each word maintains a sorted order of operate ids, that has extended convenience in question method by allowing associate economical merge step. as an example, assume that we might wish to rummage around for the points c and d. typically|this can be} often primarily to reason the intersection of the two words' inverted lists. As every lists square measure sorted inside a similar order, we tend to area unit ready to handle merging them, whose I/O and element times square measure every linear to the total length of the lists. Recall that, in NN method with IR2-tree, some extent retrieved from the index ought to be verified (i.e., having its text description loaded and checked). Verification is additionally necessary with [9] I-index, except for precisely the opposite reason. For IR2-tree, Specifically, given associate NN question alphabetic character with keyword set Wq, the question formula of I-index first retrieves (by merging) the set Pq of all points that have all the

keywords of Wq, and then, performs jPqj random I/Os to induce the coordinates of each purpose in Pq thus on choose its distance to q.

III. SYSTEM MODEL

Our proposed system model consists of 5 modules.

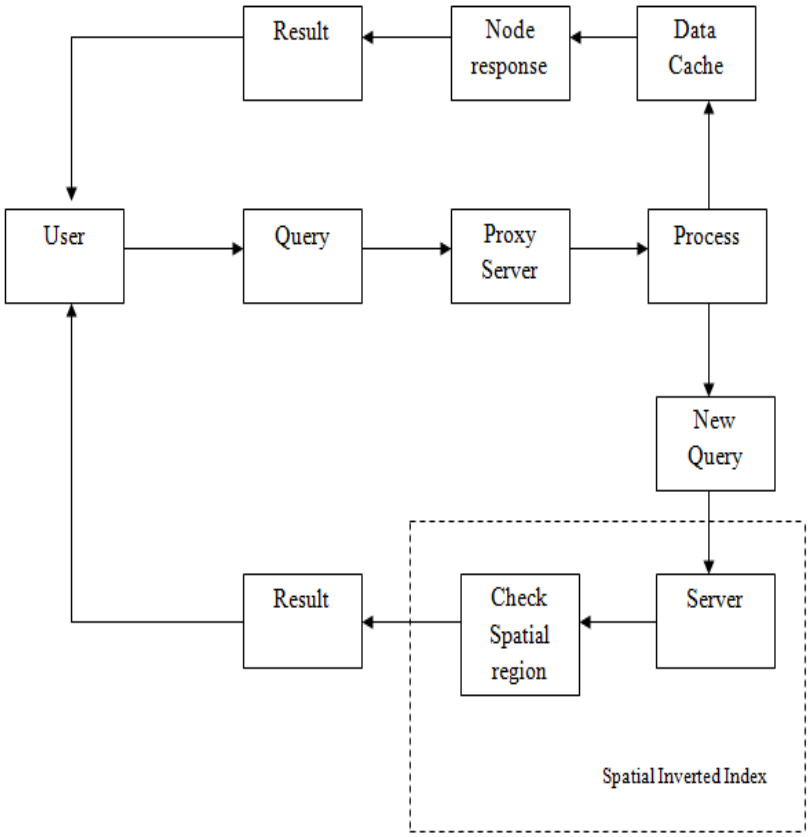


Figure 4. System Architecture

User

User provides question keyword he/she has to search. Let P be a gaggle of flat points. As our goal is to combine keyword search with the prevailing location-finding services on facilities like hospitals, restaurants, hotels, etc. assume that the points in P have variety coordinates, such each coordinate ranges in [0,t], where t might be a full variety. typically this can be} often not as restrictive as a result of it may appear, as a results of though one would like to place operative real valued coordinates, the set of varied coordinates represents below a region limit remains finite and enumerable; thus, it would additional convert everything to integers with correct scaling. it's assumed associate variety value sort of a hundred,200,300 etc. to represent the varied special locations.

Proxy Server

Proxy server gets user query and processes it. The proxy either checks for native question process to cut back the work load on server, if the query not found, then it'll forward to server. Server processes the question and sends result to proxy server. The query cached by node is holding on within the cache memory, if the other requests same query, then the node, that cached the question can reply to the request node.

LBS server

A continuous question typically consists of variety of queries to the LBS server, thereby increasing the load on the LBS server. Associate in LBS server is ready to answer special queries quickly exploitation R tree- like index structures. The LBS server is responsible for managing static information objects and respondent the queries submitted by the proxies. Note that the LBS server will use any index structure (e.g., R-tree or grid index) to method spacial queries.

Nearest Neighbor Queries

Nearest neighbor queries are, If a user submits a question, it's sent to proxy server, wherever proxy checks for cache objects within the info. If the cached objects area offered, then the question is forwarded to the cached node, then cache node replies for the requesting node. Else the question is forwarded to server, then server executes the question and reply to user. NN queries retrieves services within the queried region

Spatial Inverted Index

Spatial inverted index could be a list of question result, that maintains a sorted order of purpose ids, that provides substantial convenience in question process by permitting associate degree economical merge step. for instance, assume

that we wish to seek out the points that rebuke c and d. this is often basically to calculate the intersection of the 2 words inverted lists. As each lists are sorted within the same order, we are able to do thus by merging them, whose I/O and mainframe times are each linear to the overall length of the lists. Compression is already wide wont to scale back the dimensions of associate degree inverted index within the standard context wherever every inverted list contains solely ids. Size of the index is reduced by preferring the placement / abstraction primarily based results

IV. DESIGN ANALYSIS

UML Diagrams: UML might be a technique for describing the system style all right oppression the blueprint. UML verified among the modeling of huge and complex systems. UML might be a necessary a vicinity of developing objects homeward software package and additionally the software package development methodology. UML uses mainly graphical notations to specific the design of software package comes. The UML helps project teams communicate, explore potential designs, and validate the topic kind of the software package package. represents a collection of best engineering practices that have Definition: UML might be a general visual modeling language that is accustomed specify, visualize, construct, and document the artifacts of the pc code.

UML might be a language: it will offer vocabulary and rules for communications and performance on abstract and physical illustration. UML Specifying: Specifying implies that building models that unit precise, unambiguous and complete.

Use-Case diagram

A use case is a set of scenarios that describing an interaction between a user and a system. A use case diagram displays the relationship among actors and use cases. The two main components of a use case diagram are use cases and actors

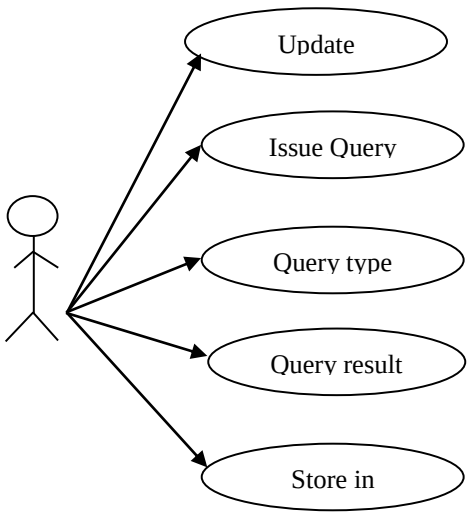


Figure 5. Use case Diagram

The use-case diagram contains two actors client and server, it also contains update current location use case which updates the current location of the user, issue query use case which gets query from use case and issued to the server, the query type use case which classifies the query type, after classifying the query the query is executed at server side and it responds to the client request with query result use case, the store in cache memory use case did store the result in cache memory

Class Diagram

Class diagrams wont to describe the categories of objects in an exceedingly system and their relationships. class diagrams model social organization and contents victimization style parts like categories, packages and objects. Class diagrams describe 3 totally different views once planning a system, conceptual, specification, and implementation. In the majority modeling tools a category has 3 components, name at the highest, attributes within the middle and operations or ways at all-time low.

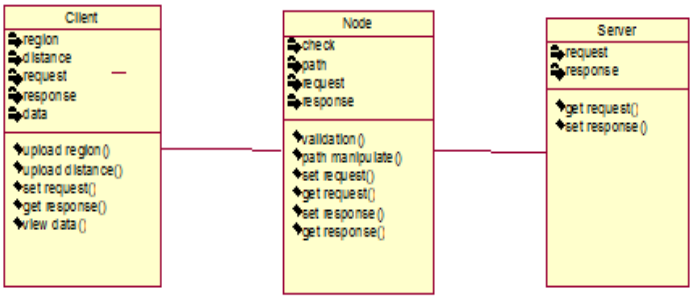


Figure 6. Class Diagram

Class diagram consists of three classes named Client, Node and Server, the Client class contains five attributes called region, distance, request, resposnse and date, the operations of client class is upload region, upload distance, set request, get request and view data operations. The Node class contains four attributes named check, path, request and response attributes, Node class consists of validation, path manipulation, set request, get request, set response and get response operations. The server class contains two attributes request and response, it also contains two operations get request and set response operations.

Sequence diagram

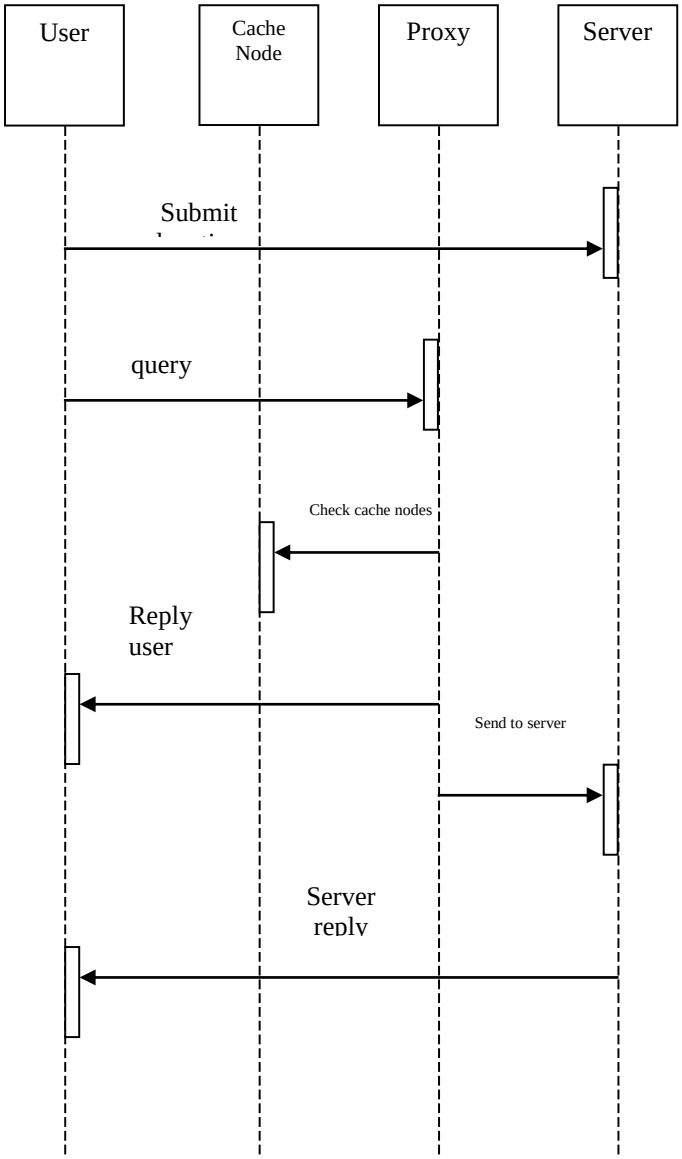


Figure. 7. Sequence Diagram

The server gets the current location of the user when the user tries to query the required location, the user submits his query, the proxy server which acts as mediator between the user and server will acts as cache memory and executes the query over server, the serer will executes the query and send results to proxy server, the proxy server then stores the results in cache memory and send the same to the client. Proxy server gets user query keyword and processes it. The proxy either checks for local query processing to reduce the work load on server, if the query keyword not found, then it will forward to server. Server processes the query and sends result to proxy server. The query keyword cached by node is stored in the cache memory, if any other request same query keyword, then the node, which is cached the query will reply to the requested node

Collaboration diagram

Communication diagram was called collaboration diagram in UML. It is same as sequence diagrams it concentrates on information transfers between objects. .

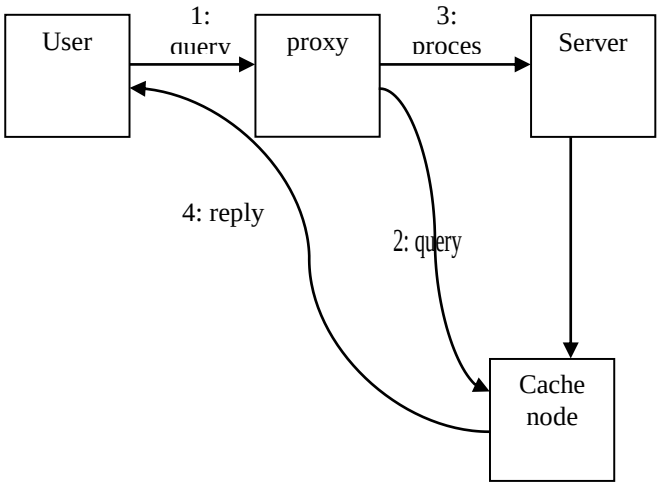


Figure 8. Collaboration Diagram

Activity diagram

Activity diagrams depict the workflow activities of a system. Activity diagrams are same as state diagrams since activities are the status of liability something. The diagrams depict the state of activities by showing the sequence of activities performed. Activity diagrams can show activities that are provisional or equivalent. Activity diagrams show the flow of activities through the system. Diagrams are read from top to bottom and have branches and forks to describe conditions and parallel activities. A junction is used when multiple activities occur at the same time

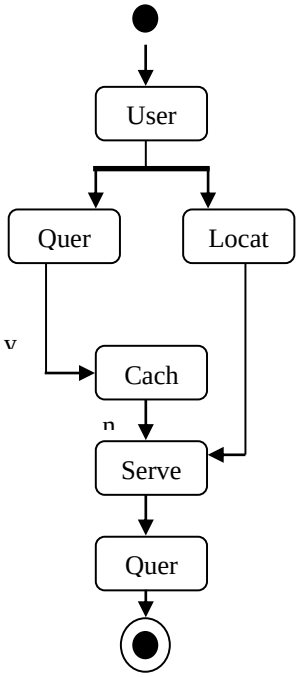


Figure 9 Activity Diagram

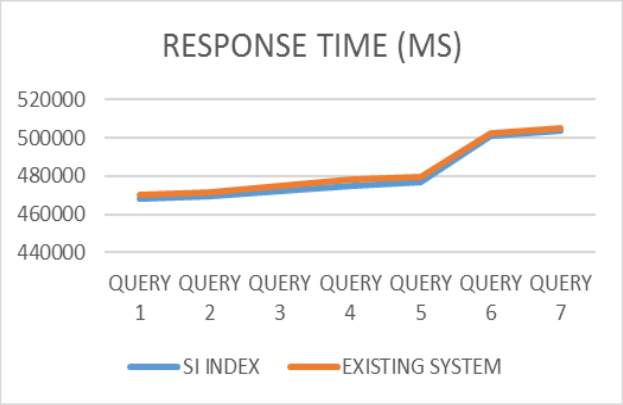
The activity diagram consists of six major activities user, query, location, cache node, server and query result, all the actives are executed sequentially.

V. RESULT ANALYSIS

The comparative study of the proposed SI Index over existing system is shown in the below table with line graph by considering the response times of each approach the orange line indicates the SI Index and blue line indicates the existing approach.

Table 1. Comparison of Response time in ms

RESPONSE TIME (MS)		
	SI INDEX	EXISTING SYSTEM
QUERY 1	467895	469895
QUERY 2	469595	471384
QUERY 3	472095	474773
QUERY 4	475095	477884
QUERY 5	476974	479319
QUERY 6	500871	502105
QUERY 7	503871	505114



Graph 1. Comparision of Response time

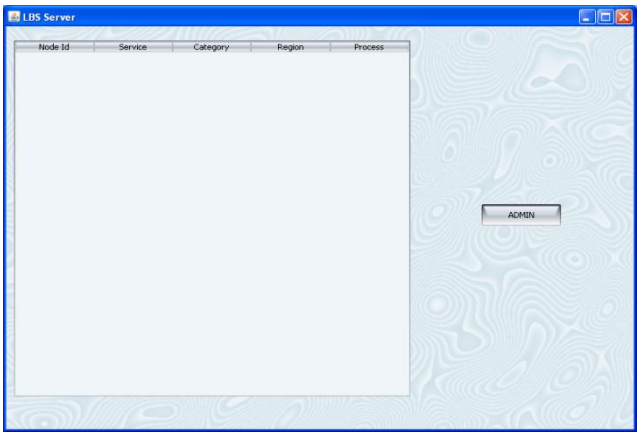


Figure 10. Lbs server

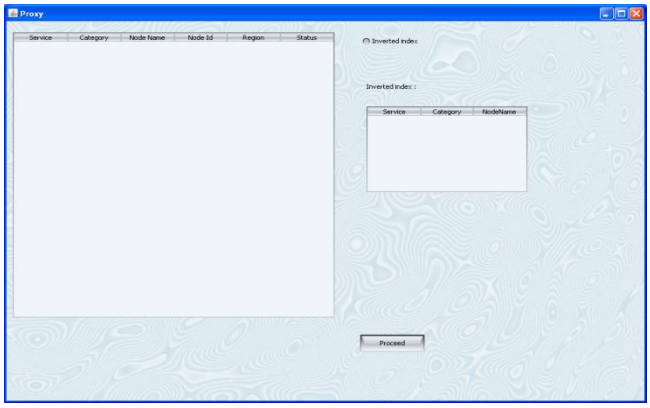


Figure 11. Lbs server SI Index

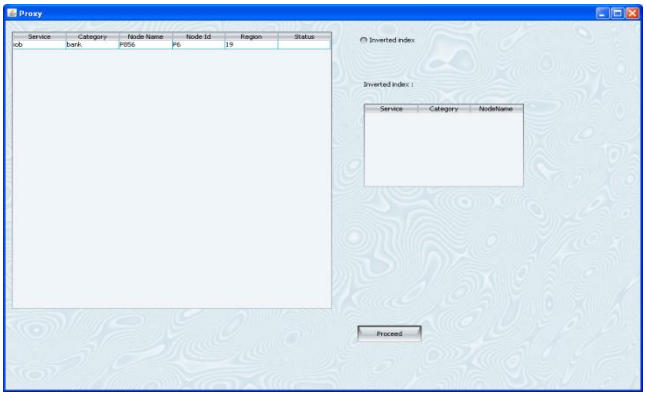


Figure 12. Proxy getting nodes data

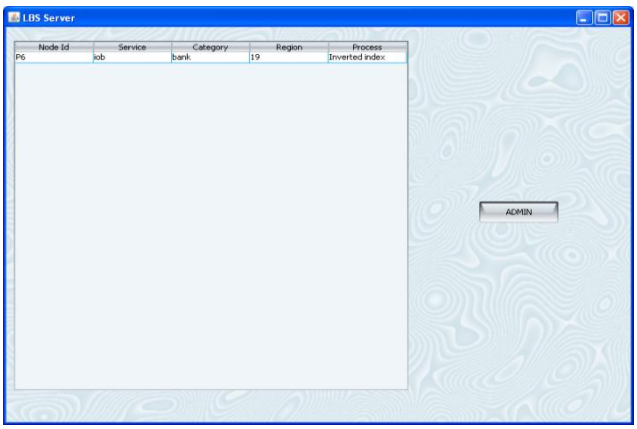


Figure 13. Lbs server Inverted Index

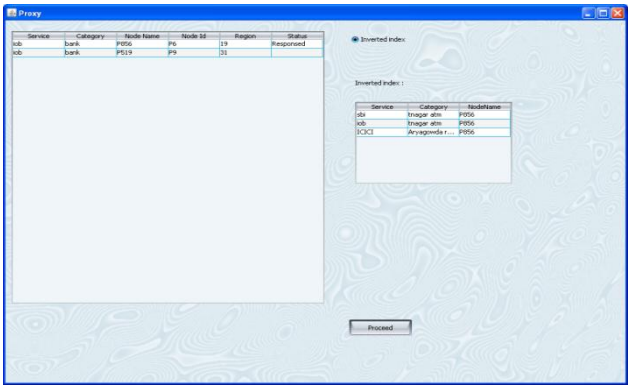


Figure 14. Proxy nearby neighbor information

VI. VI CONCLUSION

We have seen many applications line for a groundwork engine that's ready to expeditiously support novel kinds of abstraction queries that area unit integrated with keyword search. the prevailing solutions to such queries either incur preventive house consumption or area unit unable to convey real time answers. during this paper, we've got remedied matters by developing AN access methodology referred to as the abstraction inverted index (SI-index). Not solely that the SI-index is fairly house economical, however conjointly it's the flexibility to perform keyword-augmented nearest neighbor search in time that's at the order of dozens of milliseconds. what is more, because the SI-index relies on the standard technology of inverted index, it's without delay incorporable during a industrial computer programme that applies huge similarity, implying its immediate industrial deserves.

REFERENCES

[1] J. Lu, Y. Lu, and G. Cong. Reverse spatial and textual k nearest neighbor search. In Proc. of ACM Management of Data SIGMOD) pages 349–360, 2011

[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. Sudar- shan. Keyword searching and browsing in databases using banks. In Proc. of International Conference on Data ngineering (ICDE), pages 431–440, 2002.

[4] V. Maniraj, R .Mary, "Productive K-Nearest Neighbor (PKNN) and Index Based Positioning for Keyword Search", International Journal of Computer Sciences and Engineering, Vol.4(4), pp.379-383, Apr -2016

[5] X. Cao, G. Cong, and C. S. Jensen. Retrieving top-k prestige-based relevant spatial web objects. PVLDB, Vol. 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.

Authors Profile

P.Sunil Kumar Reddy received MCA from Bharathiar University in2004 and M.Phil Computer Science from Madurai Kamaraj University. He is pursuing Ph.D in SV University Tirupati. His research areas are Databases and Data Mining.

P.Govindarajulu, Professor,Department of Computer Science, Sri Venkateswara University,Tirupati, India. He received his M.Tech. from IIT Madras(Chennai), Ph.D. from IIT Bombay(Mumbai). His area of research:Databases, Data Mining, Image Processing, Intelligent Systemsand Software Engineering,Parallel Computing.

