International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Nearest Neighbor Search Technique for Novel Queries

P. R. Shejawal[1], J. R. Pansare[2], G. S. Pole[3]

[1, 2, 3]Pune University, MESCOE Pune, Late Prin V. K. Joag Path, Wadia College Campus Pune

Abstract: ***Today's applications asking for finding spatial protests nearest to a predefined area in the meantime fulfill limitation of keywords. Best answer for such questions depends on the IR2-tree, which has some inadequacies that truly affect system's efficiency. To defeat those inadequacies another access strategy is produced called the Spatial-inverted Index (SI) that extends the modified file to adapt to multidimensional information, and accompanies calculations that can answer closest neighbor queries with keywords continuously. This new technique SI is produced broadens the capacities of routine modified record makes do with multidimensional information, alongside the arrangement of using so as to move reach queries replied SI results to calculation which tackles the issue continuously***

Keywords: Inverted index, Nearest neighbor search, Spatial queries

1. Introduction

The colossal use of web searchers has made it down to earth to form spatial quires recently. Questions generally focus on things geometric properties, as whether a point is in a rectangle, or how very nearly two centers are from each other. Various bleeding edge applications that require the ability to pick objects considering both of their geometric compose and related compositions with it. For example, instead of considering each one of the lodgings, a nearest neighbor inquiry would rather ask for those hotels that is the closest among those whose administrations contain "focal aerating and cooling, eating territory, wi-fi" all meanwhile. Note this is not the "all around" nearest hotel (which would have been returned by a customary nearest neighbor request), yet the nearest restaurant among it simply giving each one of the civilities.

Today with the appearance of advances in field of Geographic Information Systems (GIS) and rising web applications, a colossal measure of spatial information accessible over the web. Number of new rising GIS applications that slither or wrap different web sources and gives an incorporated answer for the clients, for example, Information Mediators. A Spatial database contains focuses and rectangles speak to multidimensional articles helps in finding spatial items in view of some particular criteria. The significance of spatial databases is that it maps the genuine elements in a geometric way. As a case, area of lodgings, eateries, healing facilities spoke to by focuses while mix of rectangles shows expansive elements like parks, lakes, and scenes. Diverse functionalities of spatial database can be helpful in various setting. For instance, in GIS, sending of reach questions as discover all shopping centers in sure in which the client is moving, while closest neighbor recovery can find the shopping centers nearest to a given location.

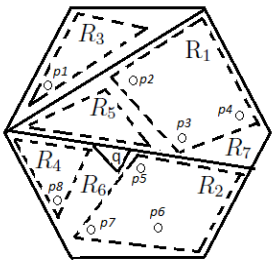


Figure 1: Locations of Objects in particular area

Objects	Words
P1	{a, b, c, d, e, f, g, h, k, l, m}
P2	{a, b, c, d, e, f, g, h, k, l, m}
P3	{c, d, e, f, g, h, k, l, m}
P4	{a, b, c, d, e, f, g, h, k, l, m}
P5	{a, b, e, f, g, h, k, l, m}
P6	{a, b, c, d, e, f, g, h, k, l, m}
P7	{d, e, f, g, h, k, l, m}
P8	{g, h, l, k, l, m}

Table 1: Object Description

As shown in above diagram q is the area of client and p1,p2....pn are the restaurants and words are amenities present at that inn. Every locale is partitioned into littler district so that if client need to go past the range then the threshold value is expansions.

2. Motivation

A spatial inquiry alongside the content takes a customer territory through the GPS and set of customer supplied watchwords as contentions and returns objects that are all the while spatially and literately significant to these watchwords. With the rich semantics of area space and the methodicalness of geological space in human's life, diverse arrangements of applicable spatial catchphrase inquiry value may be envisioned. It improves the capability of inquiry taking care of in geographic web crawlers, e.g., how to enhance the question throughput for a given issue measure and measure of gear. Inquiry taking care of is the genuine execution bottleneck in back and forth movement standard web searchers, and the crucial clarification for the countless used by greater business players and adding geographic necessities to chase request results in additional challenges in the midst of query execution

The rest of this paper is organized as follows. In Section II, it is an overview of nearest neighbor search. Sections III related work and analysis provided for IR-tree. Section IV

Licensed Under Creative Commons Attribution CC BY

based on proposed schema of given system which mainly include use of spatial inverted index and apriori algorithm Section V it concludes the research contribution of this paper.

3. Literature Survey

Web indexes for the most part utilizing catchphrase based query. Client submit question with catchphrases to the web index and a positioned rundown of records is to the client. A different option for catchphrase pursuit is organized inquiry. Both models are vast important accomplishment of both catchphrases seek and the arrangement chain of importance. A lot of the world's endeavor information stays in social databases. It is imperative that clients have the capacity to look and in addition search data put away in these databases. This plan permits clients to coordinate ventures in an organized way.

DBXplorer is associate economical and scalable keyword search utility for relative databases. This method extends to keyword search over multiple databases. The system additionally permits anyone to seek out multiple databases at the same time, however typically it's time intense method [2].

Online objects related to geo-location and a text description, [3], [4] the net is feat a spatial dimension. Mainly, internet users and its content are more and more being geo-positioned. At a similar time, matter descriptions of points of interest like cafes and toured locations are more and more changing into offered on the net. This technique that change the classification knowledge of knowledge| of information} that contains each text data and geographical locations so as to support the economical process of spatial keyword queries that take a geographical location and a group of keywords as arguments and come relevant content that matches the arguments spatial indices helps in analysis of geo-textual indices, R-tree based mostly indices, grid- based mostly indices. However, R-tree strategy desires the additional variety of keywords to look the user specification and signature files are loading the additional no text to match the thing for user specification.

R*-tree can productively be utilized as an entrance information [2] in database frameworks sorting out both, multidimensional focuses and spatial information. The R*-tree performs superior to the 2-level matrix document for point information. The new ideas consolidated in the R*-tree depend on the lessening of the edge, range and cover of the index rectangles since each of the three qualities are diminished, the R*-tree is extremely strong against monstrous information dissemination. [6]

R*-tree can productively bolsters point and spatial information at the same time. Taken a toll for its usage is just marginally higher than that of other R-trees. The mark documents are stacking the all the more no content to coordinating the item for client determination.

4. Related Work

The IR Tree

The IR tree joins the R-tree with mark records, it will audit on both what is a mark document and IR trees. The R-trees and the best-first calculation for closest neighbor seek; both are understood systems in spatial databases. Signature document alludes to a hashing-based system, whose instantiate in is known as Superimposed Coding (SC), [6] which is much powerful than different instantiates. It is intended to perform enrollment tests, figure out if word w in an inquiry exists in a set W of words. Superimposed Coding is moderate, on the off chance that it says "no", then w is unquestionably not show in W. Then again, Superimposed Coding returns "yes", the right answer can be in any case, in which case the entire information must be examined to stay away from any false hit.

SC works in the same as the system [6] called sprout channel. In pre-processing, it makes a touch of mark of length l from set of words W by hashing every word in set W to a string of l bits, and afterward take the disjunction of the greater part of the bits in string. To make it less difficult, it is indicated by h (w) the bit string of every word w. Firstly, all the l bits of h (w) are instated to 0. At that point, Superimposed Coding rehashes it for m times, then haphazardly picks a bit and set that bit to 1. Fundamentally, randomization ought to utilize word was its seed to guarantee that the same word w dependably closes with an indistinguishable h (w). The m decisions are not reliant on one another, and it might happen to be the same piece. The estimations of m influence the space cost and false hit likelihood.

Drawbacks of IR tree

The IR tree is the first system for noting closest neighbor quires with watchwords. The IR tree likewise has an a few downsides which influences proficiency. The most genuine one is false hits can be truly in substantial sum when the object of the last result is far from the query perspective, or the outcomes are essentially vacant. In such cases, the query calculation needs to stack the archives of numerous articles. The IR tree was proposed to perform definite catchphrase look with k closest neighbor inquiries in spatial databases. The m-nearest watchwords query in Euclidean space. This framework focuses on positioning queries that join both the spatial and content significance to that question object. The tree proposed to answer area based estimated catchphrase query. There are numerous comparable capacities have been proposed to measure the closeness between two words in question. Numerous methods have been proposed for distinguishing competitor word inside of a little separation from an inquiry. Framework predominantly chips away at finding top k-closest neighbors, where every hub needs to coordinate the entire questioning catchphrases. It doesn't consider the information objects in the spatial space. Additionally, this system is with low productive for handling any question. The current information structure called the IR tree is utilized for handling the inquiry. Be that as it may, IR tree has a disadvantage of mark records: false hits. Signature record, because of its moderate nature, it seeks a percentage of the articles, despite the fact that they don't have the

Licensed Under Creative Commons Attribution CC BY

watchwords present inside of query. The punishment of it causes the need to check an article who are fulfilling a question catchphrases or can't be determined utilizing just mark.

The general disadvantages of the IR tree [2], it is not supporting spatial estimated word seeks. IR tree is just backings careful catchphrase seeks. Existing word arrangement endures versatility and execution issues. It faces question streamlining issue. R trees experiences high I/O cost and correspondence overhead. It just viewed as Euclidean space or street space.

5. Proposed System

5.1 Problem Statement

Let *P* be a set of multidimensional points and *q* be the set of keywords, each point in *p* is associated with a set of words, which is denoted as *Wp* and termed as the document of *p*. Each keyword in *q* associated with set of keywords denoted as *Wq*.

5.2 System Architecture

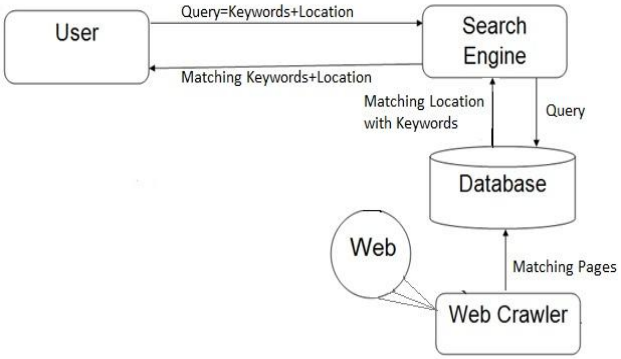


Figure 2: Architecture of the System

Web crawler: It is a system which skims the World Wide Web in robotized way. This procedure is called Web crawling. At whatever point, aftereffect of client's query's result not found in static database, the work of web crawler starts. On the off chance that web crawler discovered such area; it will give the outcome to client and update it to database

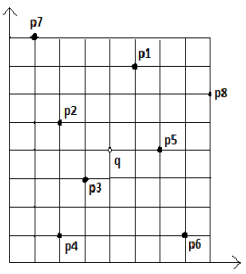
Spatial Database: A geodatabase is a database is put away in type of latitude and longitude estimation of any area i.e. directions of specific area in the meantime it stores the menus present at that place. It is triplet.

Search Engine: It is local search engine which is used to search data within static database which is already stored in system.

5.3 Mathematical Model

$Pq=\{p \in P | Wq \subseteq Wp\}$
where $p= \{p1,p2,p3...pn\}$
p is places found near to point *q*(i.e. given location)
$Wq= \{a,b,c...\}$
It is the set of keywords in given query.

Example:
-If *p* stands for restaurant, *Wp* can be its menu.
-A nearest neighbor search query specifies point *q* and a *Wq* set of keywords in query.



Objects	Words
P1	{a, b, c, d, e, f, g, h, k, l, m}
P2	{a, b, c, d, e, f, g, h, k, l, m}
P3	{c, d, e, f, g, h, k, l, m}
P4	{a, b, c, d, e, f, g, h, k, l, m}
P5	{a, b, e, f, g, h, k, l, m}
P6	{a, b, c, d, e, f, g, h, k, l, m}
P7	{d, e, f, g, h, k, l, m}
P8	{g, h, l, k, l, m}

Fig. 3: Locations of Points In map

Table 2: Text Associated with Points

Let P be an arrangement of multidimensional points and q is the area where client situated. Objective of framework is to join watchword seek with the current area finding administrations in eateries

Each point in *p* is associated with a set of words, which is denoted as *Wp* , which is nothing but amenities provided by particular restaurant and termed the document of *p*. *Wp* can be the menus and facilities. If *p* is a hospital, *Wp* can be the list of its doctors. *Wp* contain only list of words there is no any numerous values.

For example, p consists of 8 points whose locations are as shown in Figure with the black dots, and the text associated with it shown in figure 3b.

Consider a query point q, shown with hollow dot in Figure 1 with the set of keywords *Wq* = {a, b, c, d, h, k, l, m}. Nearest neighbor search finds p3, noticing that all points closer to query point *q* than *p5, p2, p1* and *p4*. But point *p3* missing the value {a, b} in its associated document, if we increase value of threshold i.e. k = 4 nearest neighbors found *p1, p6* and *p4* which is satisfying query of nearest neighbor and text associated with it. So the result of corresponding query will be *p1,* because this point is relatively nearer to query point q.

Algorithm: Nearest Neighbor (NN) Search
Input: q query point (Location of user),
BO (threshold value),
Kw keywords
Output: Location satisfying associated text
1. Initially Heap ← ∅; visited ← ∅; L ← BO.result ( );
2. Let query ***q*** fired by user with location and keywords.
3. nearer point stored in heap which is taken for processing
4. If point is within BO then
5. Add point to visited list
6. If point is satisfying text associated in query
7. Return point.
8. Repeat step 4.

Licensed Under Creative Commons Attribution CC BY

At the point when client fires a query, it comprises of a few keywords and his area, standard web search tool discovers those keywords in significant pages in the meantime it fares thee good latitude and longitude i.e. area of client. Keywords seeking should be possible with spatial transformed indexing, at whatever point new area is added to the database, framework will create reversed record again to make database a la mode. So that at whatever point other client fires same query, framework will give prompt result as it is put away in reserve, it will expand the productivity of framework.

Give D a chance to be the database in which every location is put away with triplet (x_i, y_i, k_w) coordinates and keywords present at that location. Query is likewise fired with same triplet and extent within client want to find result. The extent parameter fired by client will go about as limit in system, it is utilized to avoid unnecessary pursuit it will minimizing time consumption.

Name of the Hotel	Latitude of the Hotel	Longitude of the Hotel	Amenities Provided By the Hotel
Trikaya, Pune	18.52389	73.85010	Air-conditioned, Connecting rooms, Windows soundproof, Fitness center, Spa, Dining
Royal Orchid, Banglore	12.9756	77.5935	Connecting rooms, Windows, soundproof, Fitness center, Dining
J W Marriot, Goa	15.4872	73.81	Air-conditioned, Windows soundproof, Dining , Business services, Wi-fi facility
Le Meridian, Pune	18.5204	73.8567	Connecting rooms, Windows soundproof, Fitness center,
Taj, Mumbai	18.9369	72.8337	Windows soundproof, Air-conditioned, Dining, Wi-fi facility

Figure 4: Format of database

At whatever point any query terminated by customer it will be searched in database which is statically made and comparing area is found and content connected with it as well. It will seek the area until it discovered it. In proposed framework, the framework is powerful, at whatever point a customer fire a query with spatial area and content connected with it, it will seek inside of a static database first within threshed given by admin and afterward it will go further if client needs not fulfilled. As framework is powerful, so google map network given to it, with the goal that framework is capable hunt on the web.

6. Flow of the System

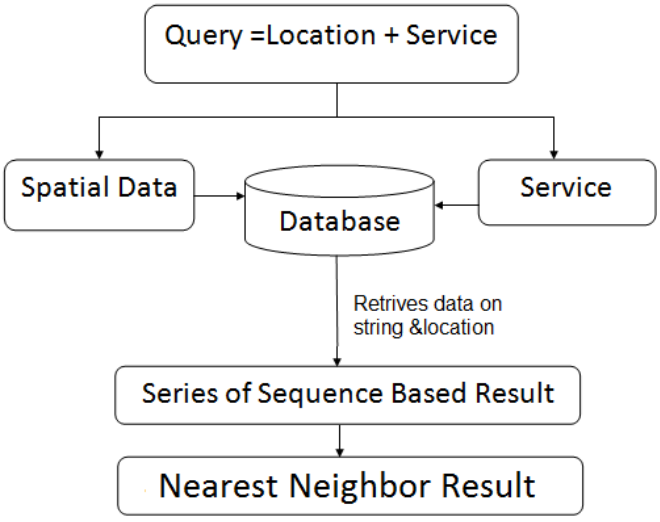


Figure 6: Flow of the System

- Initially question is terminated by client which contains area and administrations required by the client.
- The crawler gets the information of various areas and gathers the upgraded data from the web.
- Then gathered information is put away in spatial database kept up by administrator. The data is put away with its name alongside its co-ordinates values i.e. x-coordinate and y-coordinate. Likewise, the arrangement of watchwords identified with point is additionally put away in database.
- The internet searcher quest for closest neighbor area from the client area which the client is asking for, where it recovers data with the assistance of SI-list (Spatial Inverted) from the database.
- With the assistance of directions data and id values by utilizing spatial upset list it will give the closest neighbors which will fulfill the watchwords which the client is asked.

There is a simple technique known as I-index: It needs only to store the coordinates of each point together with appearances of it in the inverted lists. The coordinates present in the inverted lists motivates the creation of an R* tree, on each list indexing the points present in it. Then perform keyword based nearest neighbor search with combined structure. The R* trees allow [2] to an awkwardness in the way nearest neighbor queries are processed with an I-index. Firstly, get all the points which are carrying all the query words in Wq by merging several lists. This results to appear to be unreasonable if the point, say p, of the final result lies close to the query point q. It would be great if it can discover point p very earring have

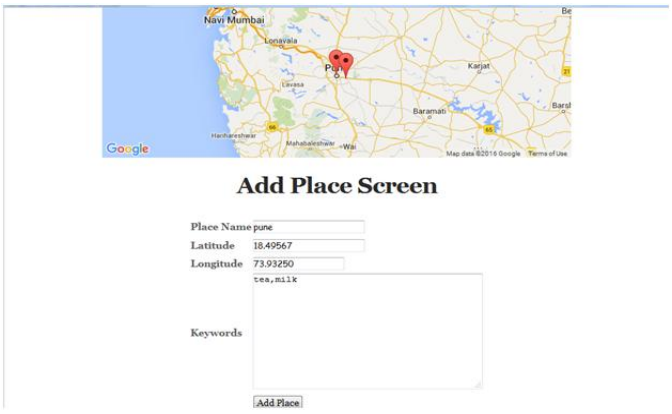
Licensed Under Creative Commons Attribution CC BY

popped up continuously, and terminate by reporting the point once the count reaches |Wq|. At any point, it is sufficient to remember only one count, because whenever a new point occurs, it is safe to forget the previous one.

7. Results

System Administrator

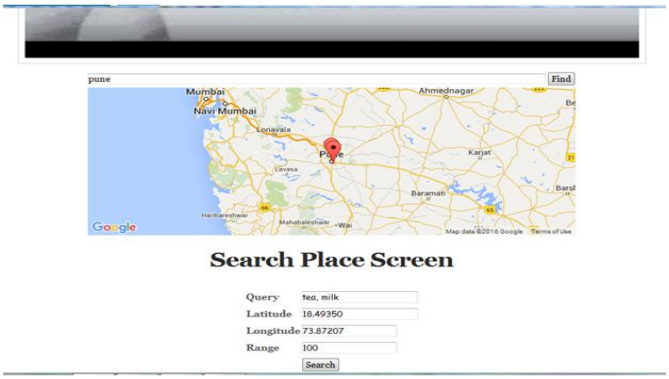
The following snapshot is the image where administrator add the places and amenities provided to that inn. In this way, whenever new hotel is discovered by the admin he is able to update the system.



Screenshot to add places in database (Admin Side)

User:

The following snapshot is the image where user is giving his/her input in keywords as a requirement. The range entered by user is nothing but the threshold value, within which system search for nearest inn.

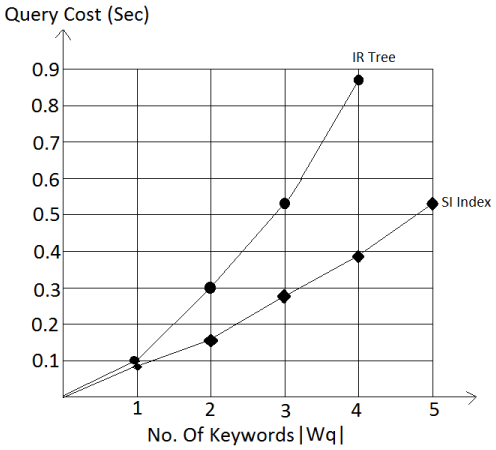


Screenshot to search places (User Side)

Performance Measure:

The main objective of this chapter is to compare the results of Nearest Neighbor Search System using SI indexing, Nearest Neighbor Search System using IR tree.

Let us start with the query performance with respect to the number of keywords |Wq|. For this purpose, parameter k to is increasing, i.e., each query search for k keywords. For each search, report its average query time in processing a workload. The results are shown in following graph, where both competing method use Uniform database,



8. Conclusion

There are a great deal of usages requires a web searcher which can reinforce novel sorts of spatial quires which are facilitated with catchphrase look. The present responses for such sort of quires either stand up to the issue of space multifaceted nature or can't give continuous answers. The proposed procedure cured the condition by building up another access framework called the spatial transformed record (SI list). SI list framework that modestly space mild, and additionally it can perform catchphrase based nearest neighbor look for in time. Besides, as the SI file relies on upon the advancement of changed record, it is quickly incorporable in a business web searcher which applies tremendous parallelism

References

[1] Pooja Shejawal and Jaayshree Pansare "Nearest Neighbor Search Technique Using Keywords and Threshold" ACM *WOMEN'S IN RESEARCH, February* 2016.

[2] Yufei Tao and Cheng Sheng "Fast Nearest Neighbor Search with Keywords" *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* 2014.

[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] 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.

[5] 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.

[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.M. Shell. (2002) IEEEtran homepage on CTAN. [Online].Available:http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[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-*

Licensed Under Creative Commons Attribution CC BY

SIAM Symposium on Discrete Algorithms (SODA), pages 30–39, 2004.

[8] 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.

[9] 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.

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

[11] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In *Proc. of Conference on Information and Knowledge Management (CIKM)*, pages 155–162, 2005.

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

Author Profile



Pooja Shejawal received the B.E. degrees in Computer Engineering from VidyaPratithan College of Engineering in 2014. She now pursing her post-graduation degree in Computer Engineering and her research interest is in Data-Mining.



Jayshree Pansare received the B.E. and M. Tech degrees in Computer Engineering. Now she is working as assistant professor at MESCOE Pune and her research interest is in image processing.



Govind Pole received the B.E. and M. Tech degrees in Computer Engineering. Now he is working as assistant professor at MESCOE Pune and his research interest is in distributed system, information retrieval.

Licensed Under Creative Commons Attribution CC BY