Efficient Web Navigator for Multi-Constrained Spatial Keyword Queries

K.B. Priya Iyer, and Shilpa T

Original scientific paper

Abstract **- The mobile technology revolutionizes the world of communications opening up new possibilities for applications such as location based web search. It involves retrieving the user point of interest (POI) from the web documents based on the query relative to a particular place or region. Existing Location based Applications on mobiles finds the nearest neighbors from the POI database not from the web documents. The existing query searches are limited to POI and do not include data objects brands like Model name, color and price etc. This paper introduces a Spatial Web Crawler (SWC) for multi-constrained keyword queries. New algorithms were developed which provides desired data objects from the web pages basing on the working hours of data objects, keywords and priority such as cost, quality and popularity of data objects etc. The SWC also provides the shortest path to reach point of interest based on travel time.**

Index Terms **- Location based Web Search, multi-constrained keyword queries, Content based Mining, GIS, Location based services, GPS.**

I. INTRODUCTION

Communication is the best invention of science which primarily focuses on exchanging information among parties at locations geographically apart. After its discovery, telephone has replaced the telegram and letter. Similarly, the term mobile has completely revolutionized the communication by opening up innovative applications that are ahead of one's imagination. Today, mobile communication has become the backbone of the society. The mobile technologies have advanced beyond the boundaries and have impact on the way of living.

A Location Based Service (LBS) is an information service based on the user location provided by operators that is accessible by mobile devices. LBS services utilize the geographic position of the mobile device to provide location information to the consumer. Uses of LBS include mapping, navigation and social networking services based on location with technologies embedded either in the handset or placed in the network. LBS also include services to identify a location of a person or object, such as finding the nearest ATM machine, business or the location of a friend or employee.

Manuscript received February 21, 2015; revised June 16, 2015.

Authors are with the M.O.P Vaishnav College for Women, India. (E-mail: priya_balu_2002@yahoo.co.in).

Web mining is the Data Mining technique that routinely determines or billets out the in order from web credentials. It is the extraction of appealing and latently useful patterns and implicit information from artifacts or movement associated to the World Wide Web. Search engines are information retrieval systems designed to help find information stored on a computer system such as on the Web, inside corporate or proprietary networks, social networks or blogs. The search engine allows the user to locate content meeting specific criteria (usually based on a given word of phrase) and retrieves a list of items matching the given criteria. This list is often sorted with respect to some measure of relevance of the results.

Search engines use regularly updated indexes to operate quickly and efficiently. Despite the popularity of search engines, there have not been many comprehensive location-based search engines that enable the user to search for locations around their area. Examples of commercial local search engines include Google (Google Earth, 2009, http://earth.google.com; Google Local, 2009, http://local.google.com), Yahoo (Yahoo! Local Maps, http:// maps.yahoo.com), and Microsoft (Microsoft, http://maps.live.com). These systems are based on a map-and-hyperlink architecture. They only search business addresses in Yellow Pages or other kinds of paid lists. In this paper, a more general form of local search, that is, to search local content
on the Web. The approach is that each web page will be first assigned to a few geographical locations according to its content and then spatially indexed in the search engine. Therefore, it can be later retrieved by its locations.
To sum up the following are the contributions:

1. Finds nearest data objects for user constraints like keywords, priority from the available web documents.

2. The objects from the web documents are retrieved based on the working hours of the POI.

3. Designing efficient algorithm for finding data objects thru four phases namely Query Initiator and Tokenizer, Category Indexer, Web Crawler, Path Finder.

4. The algorithm also gives nearest POI and the shortest path to reach the POI.

The reminder of this paper is organized as follows. In section 2, the related work on location based web search is reviewed. In section 3, we formally define System model for setting up

location based web search. In section 4, the spatial web crawler and its functioning are explained. Section 5 presents the results of the experimental evaluation of the proposed approaches with a variety of spatial network with large number of data, query objects and web pages. Finally section 6 concludes the paper with future research.

II. RELATED WORK

Generally, location information can be represented as either textual keywords (set space) or two-dimensional spatial objects (Euclidean space). Textual keywords include postal codes, telephone numbers, place names, etc. Among them, place name is more convenient to express the location hierarchy and can be easily transformed to other representations. Place name is very useful for extracting and detecting the location information in web content. However, it cannot easily describe the detail shape of a place and spatial relationships among different places. Two-dimensional spatial objects can be represented using vector model or raster model. Compared with textual keywords, they are more powerful in describing the region shapes. As to the raster model, the precision of the representation heavily depends on the size of grid cells. In [1], the authors superimposed a grid of 1024x1024 tiles on the total area of Germany. MBR is a simple approximation to a region's shape. Only two diagonal points are needed to represent the location information. Therefore, computation based on MBR is much simpler. In [2] this research work, a novel approach using weighted technique is introduced to mine the web contents catering to the user needs. Every result of the keywords and content words are compared against dictionary by full word matching. If a match is found then a point is awarded to each words based on their position (keyword / content) using weighted technique. Finally all matched keywords and content words are summarized and normalized so that the cumulative total must be less than or equal to 1. In [4] the web sites are classified based on the content of their home pages using the Naïve Bayesian machine learning algorithm. Pierre [6] discusses various practical issues in automated categorization of web sites. Machine and statistical learning algorithms have also been applied for classification of web pages [6]. An effort has been made to classify web content based on hierarchical structure [9]. In [9], paper utilizes Ant-Miner – the first Ant Colony algorithm for discovering classification rules in the field of web content mining. In [3], a geographic information retrieval system is described that is able to search for location-specific information in Singapore-based Web sites. The user is able to view their search locations on a satellite map instead of the two-dimensional maps currently used in street directories. In [7], the spatial semantic search problem to find top k relevant sets of documents with spatial constraints and semantic constraints is addressed. For devising an effective solution of the spatial semantic search, a hybrid index strategy, a ranking model and an efficient search algorithm is proposed.
Thus the current work focuses on developing a spatial navigator for retrieving POI based on web documents and visiting hours of POI.

III. SYSTEM MODEL AND ROAD NETWORK

The road network is modeled as a weighted undirected graph *G* (N, *E*), in which N consists of all vertices (nodes) of the network, and *E* is the set of all edges. It was assumed that all facility instances (data objects) lie on the road. If a data object is not located at a road intersection, the data object are treated as a node and further divide the edge it lies on into two edges. Hence N is a node set comprised of all intersections and data objects and *E* contains all the edges between them. Each edge is associated with a nonnegative weight representing the time cost of traveling or simply the road distance between the two neighboring nodes. A data objects are set of objects lying on the graph. The POI refers to Point of Interest of the User which the user wishes to visit.

Definition 1: A spatial network G (N, E) is represented by a set N of nodes (intersections) and a set E of edges (road segments). For any given two

points p1 and p2 on G(N, E), the distance DIST(p1, p2) is the distance via the shortest path from p1 to p2 and TT(p1, p2) is the travel time via the shortest path from p1 to p2.

Definition 2: A network distance Dn(n1, n2) between two nodes n1 and n2 as the length of the shortest path SP(n1, n2) connecting n1 and n2.

Definition 3: A path P from node s to destination t is represented by a series of nodes P = {n1, n2, ・・・ , nr}, in which n1 = s, nr = t and the length |P| is the sum of the weight of all edges on P.

Definition 4: Given a set of data points P, constraints c, a query point q, web documents w, and an integer k > 0, the multi-constrained keyword query is to find a result set kNN (SwebcrawlQ) that consists of k data points from the web pages such that for any p $\in$ (P − kNN) and any $p^1 \in$ kNN, | traveltime(p^1,q,c) $\leq$ traveltime(p,q,c).

The framework involves two phases: Off Line Phase and On Line Phase. During Off Line Phase, the spatial road network is partitioned into clusters and each edge in the network is encoded to be part of a cluster. The clusters are divided into C_1, C_2, C_3….C_n cluster(s) by taking Latitude, Longitude as cluster size as shown in Figure 1. The function Node_Cluster_Map() is executed which maps each node in the road network to a cluster C_i.
Instead of searching entire database to find out minimum distance node from user location, the nodes in cluster C_j are matched for minimum distance node. The nodes within cluster C_j are sorted in descending order of their distances from user location (u) where the top most nodegives the nearest neighbour of the road vertex. This improves the efficiency of search and computation speed is increased. Harvesine Formula is used in finding the distance between two locations.

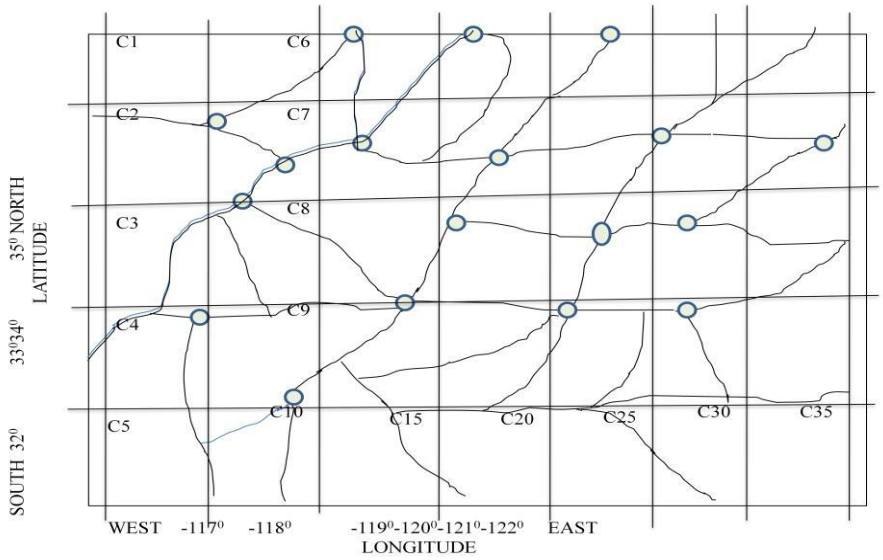

Figure 1 Road network using cluster technique

Harvesine Formula:

$$a = \sin^2(\Delta\varphi/2) + \cos(\varphi 1).\cos(\varphi 2).\sin^2(\Delta\lambda/2)$$
$$c = 2.\text{atan2}(\sqrt{a}, \sqrt{(1-a)}) \text{ distance} = R.c$$

where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km).

During the Online phase, the network is expanded which explores the graph based on distance and travel time to identify next state nearest neighbor nodes. The algorithm first finds the nearest vertex of the query origin using NearestVertex() function. Offline processing also includes extracting geographic scopes and indexing web pages according to their scopes, while online processing includes retrieving location aware information, ranking and presenting the retrieved results.

IV. SWC SYSTEM ARCHITECTURE

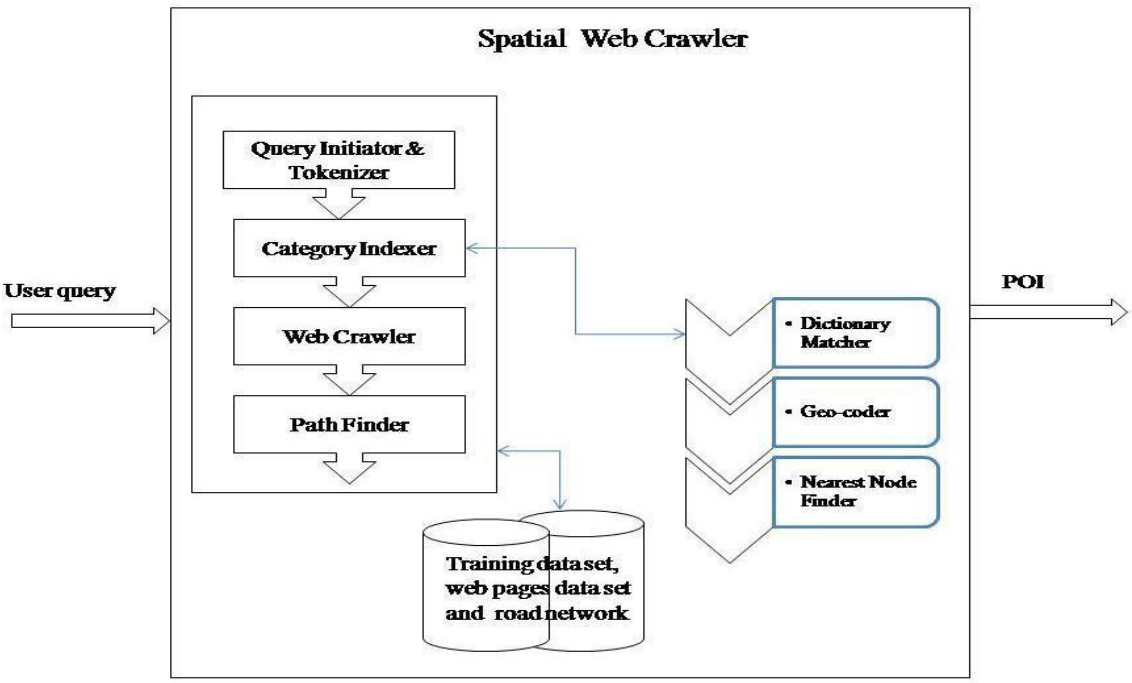

Figure 2 SWC Architecture

The SWC consists of four phases namely Query Initiator and Tokenizer, Category Indexer, Web Crawler, Path Finder.

Phase I: Query Initiator and Tokenizer

In this phase, the user query is taken for processing and it is tokenized. The possible segments of this phase are query input and user priority. The query input is the point of interest of user. The POI may or may not be mentioned directly. For example direct POI is banks, hospitals, ATMs etc. Indirect POI is baby suits, Zee kurti, XRN printer etc. The user priority is to get POI based on cost, quality or popularity etc.

Phase II: Category Indexer

This phase takes the user input and searches the structure to find correct category of the POI. For
Example, if user asks for baby suit, the category is Textile. If user asks for printer, the category is computers etc. Index structures are maintained for categories. The matching is done by using dictionary of words found in web pages by training data set**.**

Phase III: Web Crawler

This phase does the actual process of relating the web documents with the user constraints. The geo-coder matches the user location with the document spatial location. Geo-web indexer maintains the links country, state, province-wise which helps in less search process. The web pages that matches within user spatial region are processed which matches the user priority or keyword. This optimizes the search process further. Finally the web crawler retrieves the most suitable web pages that contain the user query preferences and their spatial location is given.

Phase IV: Path Finder This phase extracts the exact nearest neighbor POI from the list of web pages and finds the shortest path to reach POI based on travel time**.**

V. SWC ALGORITHM

TABLE I
SYMBOLS AND DESCRIPTIONS – SWC

Notation	Description
upos	User current location
Q	Multi-constrained User query
u	Nearest road vertex of User
Areacode	Index structure that retrieves area code of user
qcategory	Index structure that retrieves user preferred POI category
qbrand	Attribute that describes user brand requirement and priority
urllist	list of web pages that matches user query
N	Nearest POI of user preference
P	Path to reach POI

Algorithm 1 : SwebcrawlQ (upos,Q)

1. Find the nearest road vertex for the query origin upos
   u = NearestVertex(upos)

2. Areacode= Geocoder(upos)

3. qcategory=Tokenizer(Q) /* finds the user POI category */

4. qbrand=Prioritizer(Q) /* retrieves the user priority and attributes*/

5. urllist= webcrawl(areacode,qcategory,qbrand) /* retrieves all web documents that matches user query based on working hours of POI */

6. N=nearestobject(urllist,upos) /* finds nearest POI from user location */

7. P=shortestpath(N,upos) /* finds shortest path to reach POI based on travel time */

8. Display (N,P)

/* display POI and shortest path by travel time to reach POI */

The algorithm first finds the nearest road vertex of the user using the NearestVertex() function. The Geocoder() function finds the exact area code of the user current location. This is done by mapping the index structures with user latitude, longitude. The functions Tokenizer() and Prioritizer() returns the user category of POI and priority respectively.

The webcrawl() function retrieves all the web pages that satisfies user preferences. The web pages are retrieved by considering the working hours of the POI. Finally out of all web pages the Poi that is nearest to user location is displayed. The system also display the shortest path to reach POI based on the travel time.

VI. QUERY OPTIMIZATIONS

a) *Search Space reduction*: In finding the nearest vertex of the query origin, the search space is reduced by the road network clustering technique. Here instead of searching the entire database, the nodes in the user cluster are examined to find nearest neighbor node.

b) Keyword Matcher: Instead of searching all the web pages, only the pages that matches the given spatial region and keyword are taken for processing.

VII. EXPERIMENTAL EVALUATION

The experiments were conducted on California road network which contains 21,050 nodes and 21693 edges. The algorithm is implemented in Java and tested on Windows Platform with Intel Core 2 CPU and 80GB memory. The .gov data is taken as

benchmark dataset for web documents. It is a collection of real web resources of major USA government sites whose top domain is .gov. These data are mainly crawled in the year 2002 and used by TREC2003. The dataset covers a wide geographical range of USA. The edges weights are considered for both their network distance as well as travel time.

TABLE II
STATISTICS OF THE WEB DOCUMENT DATASET

Statictics	Value
The number of all pages	1053,111
The number of local pages	197,775
The number of all keywords	2,684,633
The number of geo-keywords	3,535,505

TABLE III
CLASSIFICATION EVALUATION FORMULA DEFINITION

Measure	Formula	Intuitive Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct.

TABLE IV
RESULTS OF SPATIAL WEB CRAWLER

User Preference List	Category	TP	FP	FN	TN	P	R	A
Boy suit	Textile	75	5		1	93%	98%	92%
Media course	Education	110	15		25	88%	95%	87%
DX Samsung TV	Home Electronic	102	3	19	5	97%	84%	82%
Mini Pizza pack	Food	99	25	15	3	79%	86%	71%

To analyze the detection capability of the technique, evaluation measures that include True Positive, True negative, False Positive, False Negative are employed. The results are shown in table 4.

1) Impact of Cluster in search space

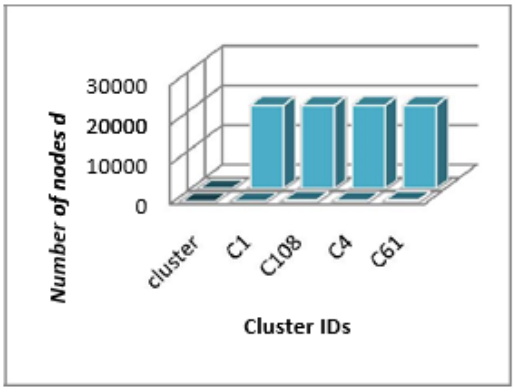

Figure 3 Nearest vertex - search space impact

Figure 3 shows the impact of clustering the road network to find nearest neighbor. The search for nodes are reduced by clustering the given road network.

2) Impact of k in execution time

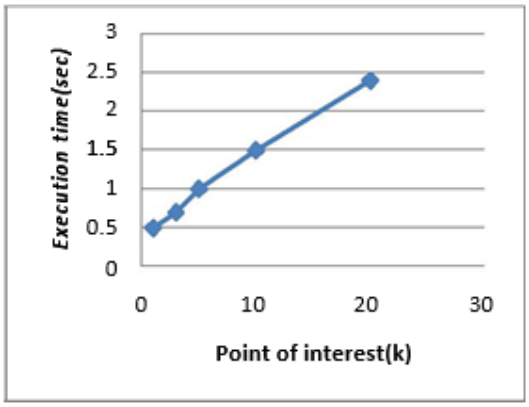

Figure 4 POI vs execution time

Figure 4 shows the impact of k on execution time. The point of interest are retrieved in short time as number of k increases.

3) Impact of k on visited nodes

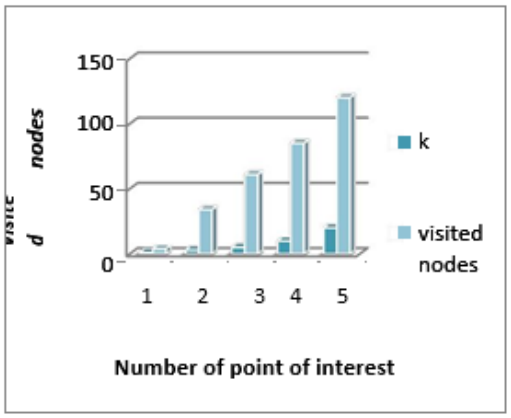

Figure 5 POI vs visited nodes

Figure 5 shows the impact of k on number of nodes accessed.

VIII. CONCLUSIONS

This paper introduces a spatial web crawler that retrieves location based web search. The user preferences are multi-constrained such as keyword, priority and web pages are extracted based on working hours of POI. The algorithm efficiently searches and computes the data object to query origin with user preferences. Different index structures are maintained for fast retrieval of geo-web extraction. The algorithm also provides the shortest path to reach the POI based on travel time. The optimization helps in fast execution of queries. The work can be extended to multiple POIs across web pages.

REFERENCES

[1] Beckmann, N., Kriegel, H., Schneider, R. and Seeger B. The R*-tree: an efficient and robust access method for points and rectangles. In Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data (SIGMOD 1990), Atlantic City, NJ, USA, 1990, 322-331.

[2] Buyukkokten, O., Cho, J., Garcia-Molina, H., and Shivakumar, N. Exploiting geographical location information of web pages. In ACM SIGMOD Workshop on The Web and Databases (WebDB 1999), Philadelphia, Pennsylvania, USA, 1999, 91-96

[3] Ding, J., Gravano L., and Shivakumar N. Computing geographical scopes of web resources, In Proceedings of 26th International Conference on Very Large Data Bases (VLDB 2000), Cairo, Egypt., 2000, 545-556.

[4] Flora S. Tsai . Web-based geographic search engine for location-aware search in Singapore. Expert Systems with Applications 38 (2011) 1011–1016.

[5] Google Local http://local.google.com.

[6] Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. Categorizing web queries according to geographical locality. In Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2003), ACM Press, New Orleans, Louisiana, USA, 2003, 325-333.

[7] Jeong-Hoon Park, Chin-Wan Chung. Spatial Semantic Search in Location-Based Web Services. In proceedings of International World Wide Web Conference Committee (IW3C2). WWW'14 Companion, April 7–11, 2014, Seoul, Korea. ACM 978-1-4503-2745-9/14/04.

[8] Lee, R.,et al. Optimization of geographic area to a web page for two-dimensional range query processing. In Proceedings of Fourth International Conference on Web Information Systems Engineering Workshops(WISEW 2003), IEEE Computer Society 2003, Roma, Italy, 2003,9-17.

[9] Leutenegger, S.T., Edgington, J.M., and Lopez, M.A. STR: a simple and efficient algorithm for R-tree packing. In Proceedings of the Thirteenth International Conference on Data Engineering (ICDE 1997), IEEE Computer Society 1997, Birmingham, U.K., 1997, 497-506.

Priya Iyer K B received her M.C.A from S.V University, Tirupati, Andhra Pradesh, India in 2003 and Ph.D in Privacy Aware Location based Services and recommendation systems for Spatial Networks from Sathyabama University, Chennai, India in 2015. She is an Associate Professor of Computer Science department at the M.O.P. Vaishnav College for Women (Autonomous), Chennai, India. Her research interests include Spatial Networks, Location based Services, Mobile Computing and privacy techniques. She is author of about 18 scientific papers published on international journals and conferences with international diffusion.

Shilpa T is an Under-graduate Student of M.O.P Vaishnav College (Autonomous). She is passion towards research and her areas include Spatial Mining and Mobile Computing.