IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Crawling Hidden Objects with KNN Queries

P. Krithika[1*], G. Sowmiya[2]

[1,2]M.Sc Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India

Corresponding Author: kirthi90@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—**Many websites offering Location Based Services (LBS) provide a k NN search interface that returns the top-k nearest-neighbor objects (e.g., nearest restaurants) for a given query location. This paper addresses the problem of crawling all objects efficiently from an LBS website, through the public k NN web search interface it provides. Specifically, we develop crawling algorithm for 2D and higher-dimensional spaces, respectively, and demonstrate through theoretical analysis that the overhead of our algorithms can be bounded by a function of the number of dimensions and the number of crawled objects, regardless of the underlying distributions of the objects. We also extend the algorithms to leverage scenarios where certain auxiliary information about the underlying data distribution, e.g., the population density of an area which is often positively correlated with the density of LBS objects, is available. Extensive experiments on real-world datasets demonstrate the superiority of our algorithms over the state-of-the-art competitors in the literature.

***Keywords*—**Knn, Hidden Objects, Location Based Services, Crawling, Density.

I. INTRODUCTION

Database Systems and Knowledgebase Systems share many common principles. *Data & Knowledge Engineering (DKE)* stimulates the exchange of ideas and interaction between these two related fields of interest. *DKE* reaches a world-wide audience of researchers, designers, managers and users. The major aim of the journal is to identify, investigate and analyze the underlying principles in the design and effective use of thesesystems.*DKE* achieves this aim by publishing original research results, technical advances and news items concerning data engineering, knowledge engineering, and the interface of these two fields. Data & Knowledge Engineering (DKE) is a journal in database systems and knowledgebase systems. It is published by Elsevier. It was founded in 1985, and is held in over 250 academic libraries. The editor-in-chief is P.P. Chen (Dept. of Computer Science, Louisiana State University, USA) This particular journal publishes 12 issues a year. All articles from the Data & Knowledge Engineering journal can be viewed on indexing services like Scopus and Science Citation Index.
The DKE delivers in-depth knowledge and competences on Data & Knowledge Engineering, one of the most promising career areas for ambitious computer scientists. Its subject area is "Engineering" for Data and for Knowledge, aiming to turn passive data into exploitable knowledge:
It focuses on the representation, management and understanding of data and knowledge assets. It encompasses technologies for the design and development of advanced databases, knowledge bases and expert systems, methods for the extraction of models and patterns from conventional data, texts and multimedia, modelling instruments for the representation and updating of extracted knowledge. The Master DKE can be studied on German or English and is thus open to students mastering either of the two languages.

II. METHODOLOGY

We start with addressing the *k*NN crawling problem in 1-D spaces, and propose a 1-D crawling algorithm with upper bound of the query cost being $O(n/k)$, where n is the number of output objects, and k is the top-k restriction. We then use the 1D algorithm as a building block for *k*NN crawling over 2-D spaces, and present theoretical analysis which shows that the query cost of the algorithm depends only on the number of output objects n but not the data distribution in the spatial space. We extend the *k*NN crawling problem to the general case of m-D spaces - which is the first such solution in the literature. Our contributions also include a comprehensive set of experiments on both synthetic and real-world data sets. The results demonstrate the superiority of our algorithms over the existing solutions.

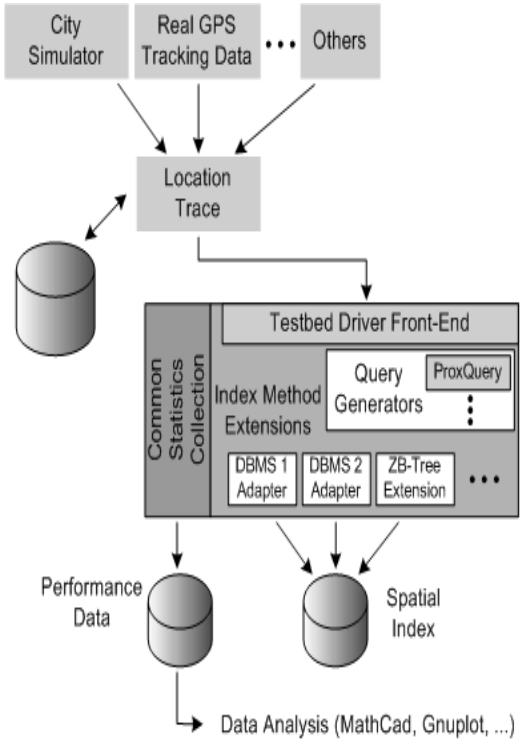

FIGURE 1: PROPOSED ARCHITECTURE

ADVANTAGE OF PROPOSED SYSTEM:

- For 2-D space, we take external knowledge into consideration to improve the crawling performance.
- The experimental results show the effectiveness of our proposed algorithms.

III. MODULE DESCRIPTION

Our project modules are given below:

- One dimensional analysis
- Two dimensional analysis
- Multiple dimensional analysis

ONE DIMENSIONAL ANALYSIS

In this module, we develop our crawling algorithm for databases with kNN interfaces in 1-D spaces. Specifically, we start with introducing an example of heavily overlapped queries issued with poor crawling strategies. Then we develop our OPTIMAL-1D-CRAWL algorithm for databases in 1-D spaces which can avoid the above mentioned problem. Finally, we give the theoretical analysis of the proposed algorithm.

TWO DIMENSIONAL ANALYSES

2-D spatial databases are popularly used in the real world, for which users are often allowed to perform kNN queries with a digital map. Take Yahoo Local as an example, we can search for restaurants by simply clicking a point through a map-like interface. Then, k nearest objects (restaurants) will be delivered by the spatial database for us to browse. We now investigate how to design an effective algorithm for crawling all points (objects) from a 2-D database without knowing their data distribution.

MULTIPLE DIMENSIONAL ANALYSIS

Though 2-D spatial databases are the most popular ones in the real world, there still exist some applications of kNN spatial databases in higher dimensional spaces (three or more dimensions). For example, the coastal.com website allows users to perform kNN queries for looking for glasses in 4-D space, with dimensions including temple arm length, lens height, lens width and DBL (distance between lenses). In order to give a solution for higher dimensional spaces and make our approach more complete.

IV. CONCLUSION

In this paper, we study the problem of crawling the LBS through the restricted *k*NN search interface. Although hidden points usually exist in 2-D space, there are some applications with points in higher dimensional spaces. We extend the 2-D crawling algorithm to the general m-D space, and give the m-D crawling algorithm with theoretical upper bound analysis. For 2-D space, we take external knowledge into consideration to improve the crawling performance. The experimental results show the effectiveness ofour proposed algorithms. In this study, the proposed algorithms crawl data objects by given a rectangle (cube) in the spatial space. In the general situation when the bounded region of the objects is irregular, it can be pre-partitioned into a set of rectangles (cubes) before using the techniques proposed in this paper.

REFERENCES

[1] Mcdonalds, "Mcdonalds page, http://www.mcdonalds.com/," [Accessed: Aug. 6, 2014]. [Online]. Available: \url{http://www.mcdonalds.com/us/en/restaurant locator.html}

[2] S. Byers, J. Freire, and C. T. Silva, "Efficient acquisition of web data through restricted query interfaces," in *Poster Proceedingsof the Tenth International World Wide Web Conference, WWW 10,Hong Kong, China, May 1-5, 2001*, 2001. [Online].Available:http://www10.org/cdrom/posters/1051.pdf

[3] W. D. Bae, S. Alkobaisi, S. H. Kim, S. Narayanappa, and C. Shahabi, "Web data retrieval: solving spatial range queries using k-nearest neighbor searches," *Geoinformatica*, vol. 13, no. 4, pp. 483–514, 2009.

[4] G. E. Glasses, "Great eye glasses page, http://www.greateyeglasses.com/shop/search.php," [Ac cessed: Jan. 20, 2014]. [Online]. Available: \url{http://www.greateyeglasses.com/shop/search.php}

[5] Yahoo, "Yahoo local page, https://local.yahoo.com/,"[Accessed: Dec. 2012]. [Online]. Available: \url{https://local.yahoo.com/}

[6] U. Census, "Us census, http://www.census.gov/cgibin/geo/shapefiles2013/layers.cgi," [Accessed: Dec. 2013].[Online]. Available: \url{http://www.census.gov/cgi-bin/geo/shapefiles2013/layers.cgi}

[7] L. Devroye, "Sample-based non-uniform random variate generation," in *Proceedings of the 18th conference on Winter simulation*. ACM, 1986, pp. 260–265.
[8] L. Barbosa and J. Freire, "Siphoning hidden-web data throughkeyword-based interfaces," in *SBBD*, 2004, pp. 309–321.
[9] A. Ntoulas, P. Pzerfos, and J. Cho, "Downloading textualhidden web content through keyword queries," in *DigitalLibraries, 2005.JCDL'05.Proceedings of the 5th ACM/IEEE-CSJoint Conference on*. IEEE, 2005, pp. 100–109.
[10] K. Vieira, L. Barbosa, J. Freire, and A. Silva, "Siphon++: ahidden-webcrawler for keyword-based interfaces," in *Proceedings of the 17th ACM conference on Information and knowledgemanagement*. ACM, 2008, pp. 1361–1362.
[11] L. Jiang, Z. Wu, Q. Feng, J. Liu, and Q. Zheng, "Efficient deep web crawling using reinforcement learning," in *Advances inKnowledge Discovery and Data Mining*. Springer, 2010, pp. 428–439.
[12] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," in *VLDB 2001, Proceedings of 27th InternationalConference on Very Large Data Bases, September 11-14, 2001,Roma, Italy*, 2001, pp. 129–138. [Online]. Available: http://www.vldb.org/conf/2001/P129.pdf
[13] S. W. Liddle, D. W. Embley, D. T. Scott, and S. H. Yau, "Extracting data behind web forms," in *ConceptualModeling - ER 2002, 21st International Conference on ConceptualModeling, Tampere, Finland, October 7-11, 2002, Proceedings*, 2002, pp. 402–413. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-45275-1 35
[14] P. Wu, J. Wen, H. Liu, and W. Ma, "Query selection techniques for efficient crawling of structured web sources," in *Proceedingsof the 22nd International Conference on Data Engineering, ICDE2006, 3-8 April 2006, Atlanta, GA, USA*, 2006, p. 47. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.124.
[15] M. Alvarez, J. Raposo, A. Pan, F. Cacheda, F. Bellas, and ´ V. Carneiro, "Crawling the content hidden behind web forms," in *Computational Science and Its Applications–ICCSA 2007*. Springer, 2007, pp. 322–333.

