© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

An Efficient Approach for Search Engine Using KNN Search

Vignesh V, Suganth R, Ramesh Kannan C

Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, Tamilnadu, India

ABSTRACT

Data mining techniques are the result of a long process of research and product development. This evolution began when business data were first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. We focus on a typical query called the KNN query. Formally, given a point-of-interest dataset P and a road network G, the KNN query reports k points such that their shortest path distances distN(q, p) from q are minimized.
Keywords: KNN, Geo, Bit Array Construction

I. INTRODUCTION

We aim to develop a compact main-memory index for road network such that: (i) it can fit into main memory, and (ii) it can process KNN queries efficiently. This is important in reducing the number of servers and the operational cost at the service provider. We propose two memory efficient KNN search solutions, Object-Last and Guide-Forest, that exploit the highway property of road networks. It offers very low latency (0.1 ms) per KNN query on 10-million-node networks on a commodity machine. This translates to a query throughput of 10,000 queries per second per commodity machine.

Indexing data for K-NN search is a closely related open problem that has been extensively studied. A K-NNG can be constructed simply by repetitively invoking K-NN search for each object in the dataset. Various tree-based data structures are designed for both general metric space and Euclidean space. Locality Sensitive Hashing is a promising method for approximate KNN search. Such hash functions have been designed for a range of different similarity measures, including hamming distance, lp with cosine similarity, etc. However, the computational cost remains high for achieving an accurate approximation, and designing an effective hash function for a new similarity measure is non-trivial.

II. METHODS AND MATERIAL

A. Scope of The Work

The scope of this project is to find the exact location with the shortest path and the point of interest. And the solutions offer very low query latency (0.1 ms) and require only small index sizes, even for 10-million-node networks.

By designing better optimizations to reduce the processing time per query, each server can provide a higher query throughput, thus enabling the service provider to cut the number of servers (and operational cost).

The key in boosting performance is to replace disk-Based solutions by main-memory solutions; since main memory has a page access latency (50 NS) significantly lower than that of hard disks (5 ms).

B. Problems In Existing System

- In the existing system, there is no option of tracking particular location with the shortest path.

- And it is the design of searching nearest neighbors with detailed information based on user's current location.
- And by tracking, the exact place of the particular user can not be known.
- The point of interest cannot be calculated so, it examines the same nodes multiple times.
- The social graphs used in the application are un weighed.
- The techniques like contraction early termination are not reducing the index size, but boost the query performance.
- The query process does not incur any disk access.
- This approach needs to search multiple indices if the knee query is interested in more than one single type of objects.
- The incremental network expansion may visit some irrelevant nodes that cannot lead to the KNN results.

C. Proposed System

This paper proposed memory-efficient algorithms for processing KNN queries on road networks. We propose two memory efficient KNN search solutions, Object-Last (OL) and Guide-Forest (GF), that exploit the highway property of road networks. We propose to extend our solutions for queries on multiple object types and for range queries. We can boost the performance of IER by replacing the A* algorithm with modern shortest path algorithms.

D. Advantages

- Object-Last, reduces the search space by half through optimizing the overlay graph structure based on the objects' distribution.
- Guide-Forest adds guidance information into the overlay graph structure and searches on a heterogeneous group composed of both the original graph and the overlay graph.
- Robust to noisy training data (especially we use the inverse square of weighted distance).
- Effective if the training data is large.
- We can minimize our costs when we build a network. It is because the shortest path algorithm will find the shortest path weight from a given source node subject to another node.
- Therefore, we need not build much of router to build path from a node to another.
- Shortest path algorithm also can maximize the performance of your system. The algorithm will find the minimum path weight. The path to weight is propagation delays for a system.

E. Technical Glossary

Spatial Database

A spatial database or Geo database is a database that is optimized to store and query data that represents objects defined in a geometric space. Most spatial databases allow representing simple geometric objects such as points, lines and polygons.

Some spatial databases handle more complex structures such as 3D objects, topological coverage, linear networks, and TINs. While typical databases are designed to manage various numeric and character types of data, additional functionality needs to be added to databases to process spatial data types efficiently.

Overlay Networks

An overlay network is a computer network that is built on top of another network. Nodes in the overlay network can be thought of as being connected by virtual or logical links, each of which corresponds to a path, perhaps through many physical links, in the underlying network. For example, distributed systems such as peer-to-peer networks and client-server applications are overlay networks because their nodes run on top of the Internet

Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that have traditionally been too time consuming to resolve. They scour databases for hidden patterns, finding predictive

information that experts may miss because it lies outside their expectations. Data mining models are,

Verification Model:

The verification model takes a hypothesis from the user and tests the validity of it against the data. The emphasis is with the user who is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis. In a marketing division, for example, with a limited budget for a mailing campaign to launch a new product it is important to identify the section of the population most likely to buy the new product. The user formulates a hypothesis to identify potential customers and the characteristics they share. Historical data about customer purchase and demographic information can then be queried to reveal comparable purchases and the characteristics shared by those purchasers, which in turn can be used to target a mailing campaign. The whole operation can be refined by drilling down' so that the hypothesis reduces the set' returned each time until the required limit is reached.

Discovery Model

The discovery model differs in its emphasis in that it is, the system automatically discovering important information hidden in the data. The data are sifted in search of frequently occurring patterns, trends and generalizations about the data without intervention or guidance from the user. The discovery or data mining tools aim to reveal a large number of facts about the data in as short a time as possible. An example of such a model is a bank database, which is mined to discover the many groups of customers to target for a mailing campaign. The data are searched with no hypothesis in mind other than for the system to group the customers according to the common characteristics found.

III. RESULTS AND DISCUSSION

A. Algorithm Details

Object Last Hierarchies Algorithm

- We develop a solution called Object-Last hierarchies, which exploits the distribution of objects such that every node containing objects is rearranged to the last (top) level of the upward graph.
- This enables the KNN result of any query to be found by traversing edges in GOL" only. OL reduces the query response time significantly since it reduces the search space

Bit Array Construction

- We use a concise data structure called bit-array B as the guidance information.
- We indicate the object reach ability of a node ni by the bit B(ni).
- This bit-array is a compact structure that only requires N/8byte space overhead.
- Its size is negligible as compared to other existing KNN solutions.
- In addition, this bit-array also provides good construction time and low maintenance cost

Developing Kit			
	Processor	**RAM**	**Disk Space**
Net Beans 8.0	**Computer with a 2.6GHz Dual Core processor or higher**	**1GB Minimum**	**Minimum 80 GB**
Database			
MySQL 5.0	**Dual Core processor at 2.6GHz or faster**	**Minimum 1 GB Physical Memory**	**Minimum 80 GB**
HeidiSQL 8.3	**Dual Core processor at 2.6GHz or faster**	**Minimum 1GB Physical Memory**	**Minimum 80 GB**

IV. CONCLUSION

The promise and potential of data mining needs to be matched by a considered approach to collection, storage, licensing and use. Without a well thought through data strategy, remedies for misuse may be hard to find. Traditional copyright protection is unlikely to assist and contract and confidential information remedies are likely to be far more significant. If competitors benefit from observing publically available data, remedies may be particularly very difficult to find. As a result, the battle against scrapers of data will often be more a battle of technologies (such as IP address blocking) than successful assertion of legal remedies.

V. REFERENCES

[1] L. A. Barroso, J. Dean, and U. Heolzle, "Web search for a planet: The google cluster architecture," IEEE Micro, vol. 23, no. 2, pp. 22–28, Mar.-Apr. 2003.

[2] R. A. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri, "Challenges on distributed web retrieval," in Proc. IEEE 23rd Int. Conf. Data Eng, 2007, pp. 6–20.

[3] R. A. Baeza-Yates, "Towards a distributed search engine," in Proc. 7th Int. Conf. Algorithms Complexity, 2010, pp. 1–5.

[4] S. Nutanong and H. Samet, "Memory-efficient algorithms for spatial network queries," in Proc. IEEE 29th Int. Conf. Data Eng., 2013, pp. 649–660.

[5] H. Hu, D. L. Lee, and J. Xu, "Fast nearest neighbor search on road networks," in Proc. 10th Int. Conf. Extending Database Technol., 2006, pp. 186–203.

[6] K. C. K. Lee, W.-C. Lee, B. Zheng, and Y. Tian, "Road: A new spatial object search framework for road networks," IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 547–560, Mar. 2012.

[7] D. Papadias, J. Zhang, N. Mamoulis, and Y. Tao, "Query processing in spatial network databases," in Proc. 29th Int. Conf. Very Large Data Bases, 2003, pp. 802–813.

[8] K. Deng, X. Zhou, H. T. Shen, S. W. Sadiq, and X. Li, "Instance optimal query processing in spatial networks," VLDB J., vol. 18, no. 3, pp. 675–693, 2009.

[9] R. Zhong, G. Li, K.-L. Tan, and L. Zhou, "G-tree: An efficient index for knn search on road networks," in Proc. 22nd ACM Int. Conf. Inform. Knowl. Manage., 2013, pp. 39–48.

[10] H. Samet, J. Sankaranarayanan, and H. Alborzi, "Scalable network distance browsing in spatial databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008, pp. 43–54.

[11] H. Hu, D. L. Lee, and V. C. S. Lee, "Distance indexing on road networks," in Proc. 32nd Int. Conf. Very Large Data, 2006, pp. 894–905.

[12] M. R. Kolahdouzan and C. Shahabi, "Voronoi-based k nearest neighbor search for spatial network databases," in Proc. 13th Int. Conf. Very Large, 2004, pp. 840–851.

[13] A. V. Goldberg and C. Harrelson, "Computing the shortest path: A search meets graph theory," in Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms, 2005, pp. 156–165.

[14] J. M. Kleinberg, A. Slivkins, and T. Wexler, "Triangulation and embedding using small sets of beacons," J. ACM, vol. 56, no. 6, pp. 32:1–32:37, 2009.

[15] I. Abraham, A. Fiat, A. V. Goldberg, and R. F. F. Werneck, "Highway dimension, shortest paths, and provably efficient algorithms," in Proc. 21st Annu. ACM-SIAM Symp. Discrete Algorithms, 2010, pp. 782–793.

[16] R. Mertens, T. Steffens, and J. Stachowiak, "Path searching with transit nodes in fast changing telecommunications networks," in Proc. GI Jahrestagung, 2008, pp. 158–163.

[17] R. Geisberger, P. Sanders, D. Schultes, and D. Delling, "Contraction hierarchies: Faster and simpler hierarchical routing in road networks," in Proc. 7th Int. Conf. Exp. Algorithms, 2008, pp. 319–333.

[18] R. J. Gutman, "Reach-based routing: A new approach to shortest path algorithms optimized for road networks," in Proc. 6th Workshop Algorithm Eng. Exp., 2004, pp. 100–111.

[19] R. Bauer and D. Delling, "Sharc: Fast and robust unidirectional routing," in Proc. Workshop Algorithm Eng. Exp., 2008, pp. 13–26.

[20] A. D. Zhu, H. Ma, X. Xiao, S. Luo, Y. Tang, and S. Zhou, "Shortest path and distance queries on road networks: towards bridging theory and practice," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 857–868.