Volume : 4 | Issue : 6 | June 2015 • ISSN No 2277 - 8179

Nearest Keyword Search by Measuring Semantic Similarity

Engineering

KEYWORDS : Nearest neighbor search, keyword search, Semantic similarity,Swoogle,Thesaures.

Harsha Aravind M M.Tech CSE AWH Engineering College Calicut, India

ABSTRACT

Spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects geometric properties. A spatial database manages multidimensional objects(such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. Now-a-days many applications call a new form of queries to find the objects that satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of considering all the restaurants, a nearest neighbor query would instead ask for the restaurant that is the closest among those whose menus contain the specified keywords all at the same time.Concept of IR2-tree and spatial inverted index is used in the existing system for providing best solution for finding nearest neighbor and also it is a simple web application that stores a collection of documents in Database. This method has few deficiencies. So proposed a new technique which includes extracting synonyms and measuring semantic similarity to improve the query results.

I. Introduction

Spatial data mining is a special kind of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks use not only non-spatial attributes (as it is usual in data mining in non-spatial data), but also spatial attributes. Spatial data mining is the application of data mining methods to spatial data. The objective of spatial data mining is to find patterns in data with respect to geography.

A spatial database manages multidimensional objects (such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modeling entities of reality in a geometric manner. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in geography information system, range search can be deployed to find all restaurants in a certain area, while nearest neighbour retrieval can discover the restaurant closest to a given address.

Today, the widespread use of search engines has made it realistic to write spatial queries in a brand new way.Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other. have seen some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts. For example, it would be fairly useful if a search engine can be used to find the nearest restaurant that offers "steak, spaghetti, and brandy" all at the same time. Note that this is not the globally"nearest restaurant (which would have been returned by a traditional nearest neighbour query), but the nearest restaurant among only those providing all the demanded foods and drinks. There are easy ways to support queries that combine spatial and text features.

For example, for the above query,could first fetch all the restaurants whose menus contain the set of keywords {steak,spaghetti, brandy}, and then from the retrieved restaurants, find the nearest one. Similarly, one could also do it reversely by targeting first the spatial conditions – browse all the restaurants in ascending order of their distances to the query point until encountering one whose menu has all the keywords. The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs. A typical example is that the real nearest neighbour lies quite far away from the query point, while all the closer neighbours are missing at least one of the query keywords. This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme.

In this paper, propose a method for measuring semantic similarity between entities by using Swoogle API and getting synonyms by using Thesaurus. The method utilizes the existing semantic knowledge embedded in the Swoogle web .It does not require domain specific knowledge engineering.

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their SYNTACTICAL representation (e.g. their string format). Computing semantic similarity between words/phrases has important applications in natural language processing, information retrieval, and artificial intelligence. There are two prevailing approaches to computing word similarity, based on either using of a thesaurus (e.g., WordNet) The value of the proposed method in capturing semantic similarity of entities in two knowledge acquisition tasks. First, we present a method for constructing a thesaurus of real-world entities using the semantic similarity computed from the web directory. For each entity, the thesaurus lists the entities most similar to the target entity in different aspects. The thesaurus can be used in many applications such as entity disambiguation and query expansion. Second, we built a categorization of entities by clustering them based on their semantic similarities. The categorization is at much finer level than common named entity recognition, which is limited to a small set of broad categories.

II. RELATED WORKS

DBXplorer:A System for keyword-Based search over Relational Databases[2] is a system that enables keyword based search in relational databases.It has been implemented using a commercial relational database and web server and allows to interact Via a browser front end.The main goal is to perform keyword search on multiple databases without requiring the users to know the schema of respective databases.The problem is that DBXplorer supports conjunctive keyword queries, i.e., retrieval of only documents that contain all query keywords

Signature Files:An Access Method for Documents and its Analytical Performance Evaluation[6] In this method the documents are stored sequentially in the "text file." Their abstractions are stored sequentially in the "signature file." When a query arrives, the signature file is scanned sequentially, and a large number of non- qualifying documents are discarded. The rest are either checked. A document is called a "false drop" if it does not actually qualify in a query, although signature indicates it does. The method is faster than full text scanning but is expected to be slower than inversion It requires much smaller space overhead than inversion and it can handle insertions easily.The problem is that signature files are slower for large databases.

Keyword Searching and Browsing in Databases using BANKS[4] a system which enables keyword based search on relational databases, together with data and schema browsing. BANK models tuples as nodes in a graph, connected by links induced by foreign key and other relationship. Answers to a query are modeled as rooted trees connecting tuples that match individual keywords in the query. Answers are ranked using a notion of prestige of nodes based on in links, similar to techniques developed for web search.The problem is that when we ignore the directionality would cause problems because of hubs which are connected to large numbers of nodes.

Spatial Keyword Query Processing[5] In this method some spatial indexing scheme such as R tree,Grid,Space filling curve are used. R-trees are a N-dimensional extension of B^{+-} trees, useful for indexing sets of rectangles and other polygons.Basic idea is that generalize the notion of a one-dimensional interval associated with each B+-tree node to an N-dimensional interval, that is, an N-dimensional rectangle.In Grid based method uses a grid to partition the space each cell is associated with one page.In space filling curve is a curve whose range contains the entire two dimensional unit square and it uses Z ordering and Hilbert curve. Z ordering Maps multidimensional data to one dimension while preserving locality of the data points.Hilbert curve need points that are close in 2 dimensional to be close in the 1 dimensional.The problem with R tree is that Maximum Coverage and overlap may occur and in Grid based memory space cost is high.

The R*-tree: An Efficient and robust Access Method for points and rectangles[3] This method uses a variant of R-tree Supports point and spatial data efficiently at the same time Implementation cost only slightly higher than that of other R-trees. Supports map-overlay operation – Spatial Join E.g. of Spatial Join queries: Two spatial relations S1 and S2, find all pairs: x in S1, y in S2 s.t. x rel y = true where rel = intersect, inside etc. This method is completely Dynamic.

Fast Nearest Neighbor Search with Keywords using spatial inverted index[1] only retrieves the results having exact match although it reduces space and improve the query speed.Here the concept of semantic similarity is not considered

III. existing system

A spatial database manages multidimensional objects(such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modeling entities of reality in a geometric manner. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in a geography information system, range search can be deployed to find all restaurants in a certain area while nearest neighbor retrieval can discover the restaurant closest to a given address. Today, the widespread use of search engines has made it realistic to write spatial queries in a brand new way. Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other. some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts. For example, it would be fairly useful if a search engine can be used to find the nearest restaurant that offers "steak, spaghetti, and brandy" all at the same time. Note that this is not the globally" nearest restaurant (which would have been returned by a traditional nearest neighbour query), but the nearest restaurant among only those providing all the demanded foods and drinks. There are easy ways to support queries that combine spatial and text features.

For example, for the above query, we could first fetch all the restaurants whose menus contain the set of keywords {steak, spaghetti, brandy} and then from the retrieved restaurants, find the nearest one. Similarly, one could also do it reversely by targeting first the spatial conditions browse all the restaurants in ascending order of their distances to the query point until encountering one whose menu has all the keywords. The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs. A typical example is that the real nearest neighbour lies quite far away from the query point, while all the closer neighbours are missing at least one of the query keywords. This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme. The technique used in the Existing system is SI(Spatial inverted Index) An SI index preserves the spatial locality of data points, and comes with an Rtree built on every inverted list at little space overhead. As a result, it offers two competing ways for query processing. It can (sequentially) merge multiple lists very much like merging traditional inverted lists by ids. Alternatively, we can also leverage the R-trees to browse the points of all relevant lists in ascending order of their distances to the query point. It contains the set of points and the points are related to the set of keywords and the keywords are related to derive the set of documents. Here using the new concepts of the merge list and the distance alignment to retrieve the documents from the user requirement.

The problem is that Existing system doesn't give importance to semantics so when a keyword exactly matches with the entry in the Dataset the result is retrieved otherwise not.Near points are selected based on distance calculation from the user query point.

A. System Architecture

In this section we describe the system architecture of existing system

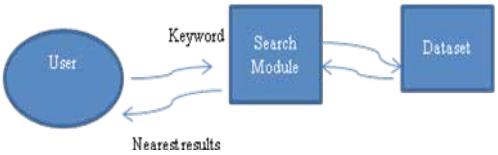

Fig 1. Architecture of Keyword search module

The system architecture consists of four entities they are

- User: The user can be end user who searches for the nearest locations offering the keywords
- Search module: In this module nearest locations offering the keywords are retrieved from the Dataset maintained using spatial inverted index and compression scheme and the nearest locations offering the keywords are retrieved to the user

IV. proposed system

The importance of named entities in information retrieval and knowledge management has recently brought interest in characterizing semantic relationships between entities. Here propose a method for measuring semantic similarity, an important type of semantic relationship, between entities.The method is based on Swoogle, crawler-based indexing and retrieval system for the Semantic Web the semantic similarity between entities can be measured in different dimensions.

Named entities play a vital role in information and language processing. Recently, there has been increasing interest in characterizing semantic relationships between entities. Analogous to the synonym relationship holding between common words, the relationship of semantic similarity holds between entities. Measurement of semantic similarity between entities can provide particular value for tasks concerning the semantics of entities, such as ontology generation, automatic annotation of web pages and question answering.

For common words, general thesauri such as WordNet are important knowledge resources for measuring semantic similarity. However, these general lexical databases provide very limited coverage of named entities. Therefore, similarity measures based on these thesauri are not applicable for named entities.

Thesaurus, a collection of words and phrases arranged according to the ideas they express, presents a solid framework for a lexical knowledge base. Its explicit ontology offers a classification system for all concepts that can be expressed by English words.It is a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. The main purpose of such reference works is to help the user to find the word, or words, by which idea may be most fitly and aptly expressed.

B. Algorithm

1. Input the keyword
2 .Convert the query into tokens
3. Perform synset retrieval by using the concept of web mining
4. HTTP request along with the keyword is refered to the site http://www.thesaurus.com/browse/keyword for getting synonyms
5. Output will be a table consisting of required synonym set Example: suppose keyword is Best Biriyani synonym set will give super Biriyani,perfect Biriyani etc
6. Synonym set is retrieved as HTML doc from this HTML doc ,the required synonyms are fetched by using Tags
7. Synonyms for the keywords are matched for their similarity measure by passing the url http://swoogle.umbc.edu/SimService/GetSimilarity?operation=api&phrase1=best biriyani&phrase2=perfect biriyani
8. Get the GPS location of the user
9. Nearest results are shown to the user

C. System Architecture

In this section we describe the system architecture of proposed system

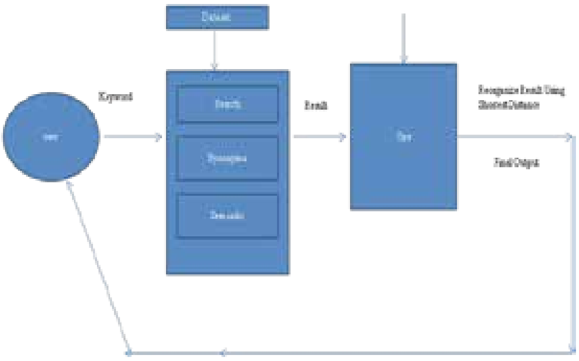

Fig 2. Architecture of keyword searchl in proposed scheme.

The system architecture include entities, they are

1) User: End user who performs the search
2) Search: In search module there are two sections Semantic and synonyms.synonyms for the keyword is extracted and checked for semantic similarity measure
3) Gps:Reorganize the query results and nearest locations are identified by using Gps system

V. Conclusion

In this paper, developed a method for measuring semantic similarity between named entities. The method exploits the semantic knowledge embedded in the Swoogle semantic web,.crawler-based indexing and retrieval system for the Semantic Web By passing the url with the name of an entity,we can obtain the semantic measure relevant to the entities, which capture various features of the entity and once we find the semantic measure the nearest data corresponding to the particular keyword is shown to the user

REFERENCE

[1] Yufei Tao and Cheng Sheng "Fastest Nearest Neighbor search with keywords "IEEE Transcations on Knowledge and Data Engineering" | [2] S. Agrawal, S. Chaudhuri, and G. Das, "Dbxplorer: A System for Keyword-Based Search over Relational Databases,"Proc. Int'lConf. Data Eng. (ICDE) | [3] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, "The Rtree: An Efficient and Robust Access Method for Points and Rectangles,"Proc. ACM SIGMOD Int'l Conf. Management of Data | [4] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S.Sudarshan, "Keyword Searching and Browsing in DatabasesUsing Banks,"Proc. Int'l Conf. Data Eng. (ICDE) | [5] X. Cao, L. Chen, G. Cong, C.S. Jensen, Q. Qu, A. Skovsgaard, D.Wu, and M.L. Yiu, "Spatial Keyword Querying,"Proc. 31st Int'lConf. Conceptual Modeling (ER) | [6] C. Faloutsos and S. Christodoulakis, "Signature Files: An AccessMethod for Documents and Its Analytical Performance Eval-uation,"ACM Trans. Information Systems