Spatial Queries Entity Recognition and Disambiguation using Rule-Based Approach

E. Hamzei [a], F. Hakimpour* , A Forati

[a] School of Surveying and Geospatial Engineering, Faculty of Engineering, University of Tehran
(e.hamzei, fhakimpour, forati)@ut.ac.ir

KEY WORDS: GIS, Spatial Search Engine, Natural Language Processing, Entity Recognition

ABSTRACT:

In the digital world, search engines have been proposed as one of challenging research areas. One of the main issues in search engines studies is query processing, which its aim is to understand user's needs. If unsuitable spatial query processing approach is employed, the results will be associated with high degree of ambiguity. To evade such degree of ambiguity, in this paper we present a new algorithm which depends on rule-based systems to process queries. Our algorithm is implemented in the three basic steps including: deductively iterative splitting the query; finding candidates for the location names, the location types and spatial relationships; and finally checking the relationships logically and conceptually using a rule based system. As we finally present in the paper using our proposed method have two major advantages: the search engines can provide the capability of spatial analysis based on the specific process and secondly because of its disambiguation technique, user reaches the more desirable result.

1. INTRODUCTION

In the modern world, using search engines for specific or general applications is a prevalent issue. Therefore, this subject has risen as a challenging topic for researchers. In general all studies in this field is looking for the more effective desired results and this purpose affects all various parts of a search engine from crawler to indexing mechanism. Some of these parts improve the system's performance and some others directly help the user to access better results. With increasing advances in the Web and its related technologies and infrastructures, the use of different type of data and their various applications increased dramatically, one instance of data types is spatial data type that has not been exempted from this rule. Nevertheless, there is a variety of challenges for researchers in this field which is due to the inherent complexity of spatial data. One of these fundamental challenges is in the way of data access, particularly in the subject of search engines. Currently the major search engines including Google and Bing provide separate section for their spatial services which it makes the importance of spatial data evident.

One of the search engines sections is query processing section which its duty is analysing the user queries. If at this stage all the information that user entered into the search engine, is well analysed, search engines can more effectively guide the user to the desired results. Analysing the queries means to scrutiny of different parts of the queries entered by the user and understanding the connections among them.

It can be very difficult for a general search engine because each type of information can be entered by the user into general search engines. In contrast with this, for special purpose search engines such as spatial search engines, the types of information that the user entered is specified. In this phase, we only want to determine that what type of information each constituent of the user query envelop and what are the relationships between this information. After that, in various search engines, it's much easier to filtering out some of the results and ranks them. The advantage of this approach to special purpose search engines is it make it possible to respond to complex queries. And also, many of the useless results filtered out by the understanding that obtained by the query analysis. In addition the possibility of results weighting in ranking procedure is another added value of this approach. So by the use techniques in language processing NLP (Natural Language Processing), a new algorithm for processing the spatial queries of spatial search engines is presented.

NLP is one of the main research topics in computer science. This field of study is trying to analyse the information that is extracted from the textual data automatically, and figures out the relations between its components. An important part of NLP is the NER (Named Entity Recognition) whose duty is reviewing and classification of the components (words). NER has an important application in query analysis so that the elements of the query classifies to predefined categories and their relationships are determined therefore query result is conceptually closer to the outcome that the user expects.

There has been a lot of research in the area of entity recognition and query processing but as far as we know, however, there was no work on Spatial Entity Recognition in Query using language modelling as defined in this paper. We inspired Named Entity Recognition concept which is usually performed on text documents ERD solution. Early work on NER was based on rules and then machine learning techniques have been applied to NER. D. Nadeau and S. Sekine (2007) presented a survey of named entity recognition and its classification techniques, they reports machine learning as a basis of new Named Entity Recognition and Classification systems and discussed word, dictionary and corpus level representations of words in a document and presented Evaluation techniques, ranging from intuitive exact match to very complex matching techniques with adjustable cost of errors. L. Ratinov and D. Roth (2009) presented a model for NER that uses expressive features to reaches new state of the art performance on the Named Entity recognition task so they explored four fundamental design decisions: text chunks representation, inference algorithm, using non-local features and external knowledge.

Later, Guo et al. (2009) addresses the problem of Named Entity Recognition in Query (NERQ), which involves detection of the named entity in a given query and classification of the named entity into predefined classes.so they have proposed employing a probabilistic approach to perform the task using query log and a topic model, and new weakly supervised learning method for creating the topic model called WS-LDA, in which the topics of a document are assigned. Junwu Du et al (2010) propose a method to utilize the search session information. Their approach

help to improve two classical NER solutions by utilizing the search session context, which are known as Conditional Random Field (CRF) based solution and Topic Model based solution respectively. Guo et al (2011) argue that query similarity should be defined upon search intents, so-called intent-aware query similarity. By introducing search intents into the calculation of query similarity, they claimed to obtain more accurate and also informative similarity measures on queries.
Krishnamurthy and Mitchell (2011) proposed concept resolver, a technique that jointly does both word sense induction and synonym resolution on extracted relations from a given text corpus. Dalvi et al.(2014) developed a four step algorithm named Topic-specific Language Model (TLM method) for doing Entity Recognition and Disambiguation from search queries.in the process that initially candidate entity strings by segmenting the query in different ways then retrieves candidate entities by searching these candidate entity strings in Freebase thereupon ranks the candidate entities using language model based query likelihood scores and finally groups the entity annotations into interpretations.
The proposed algorithm inspired by the Dalvi et al. (2014) approach in order to solve the problem which is associated with spatial search engines. The most important issue emphasized in this paper is changing the viewpoint of spatial search engines from textual perspective to spatial perspective. Textual perspective has led queries with spatial analysis not to be supported or had multiple wrong answers, so in order to fix these flaws primarily query analyzed, and then by using the information which gained from query processing, the process of disambiguation and spatial analysis is applied, thus in this paper we focused on this important spatial search engines principle.
In Figure 1 you can see a simple query that is not answered correctly by the Google search engine, while all of the information needed to respond to this query is provided. This is an example of the impact of query processing in better responsiveness of the search engines, and in fact, this weakness in the responsiveness is because of the purely emphasis on textual information without any query processing.

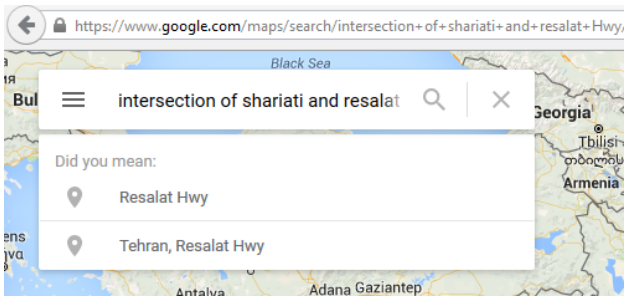

Figure 1 the query which responded quietly wrong by the Google

In Figure 2, two queries with the same format can be seen that in the first case Google search engine is able to respond, but the latter will not gained any results. The cause of search engine ability to response to the first query is the existence of address information in that area. In fact, Google is answered to the first query by using its additional textual data and in the second case due to lack of address information it failed to response. This fact indicates that the wrong approach in query processing not only the cause of failure to respond to a range of queries, but in some cases the search engines behavior turned to an unstable manner.

The proposed approach can solve the aforementioned problems by analyzing and processing the query. Actually, use of this mechanism diminished the highly dependency of search engines on textual data. It will be able to respond stably with a minimum of textual data in the system. One of this method's by-product is its disambiguation ability. In general, ambiguity in spatially searches is the result of two elements: 1) Multiple locations can have the same name, 2) a location can have different names. The former type of ambiguity is only related to the data volume in the system but the latter is effectively surmountable for the spatial relationship included queries in a way that the results domain rarely exceeds a result.

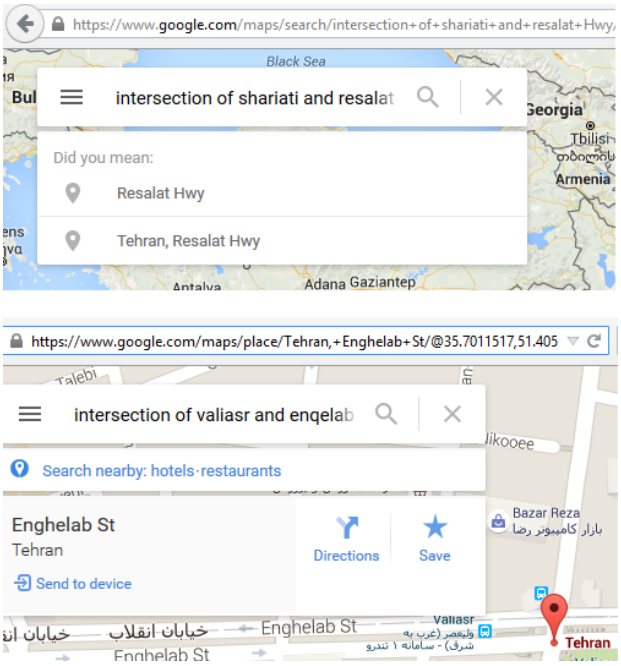

Figure 2 different behaviour of the Google against queries with same structure

2. METHODOLOGY

Spatial queries conceptually is made of three basic elements: locations name, locations type and spatial relations, so we want to provide a mechanism to parse the query to these elements which be followed by improvement in the results of the query. Our proposed algorithm consists of three main steps (depicted in figure 3).

2.1 Splitting

First step is splitting, that means determining all possible states between queries components (words which builds query), the proposed method for splitting is carried out in a top-down process, in a way that queries gradually decompose to small set of words and each time output in the form of a sub-queries list is sent to the next step. This is continued to the lowest level that the query has review word by word. The philosophy of top-down process is that splitting the set of words with more words is more reliable than the set of words with less words, as instance "northern Golpayegani avenue" is Less ambiguous and more important than " Golpayegani".

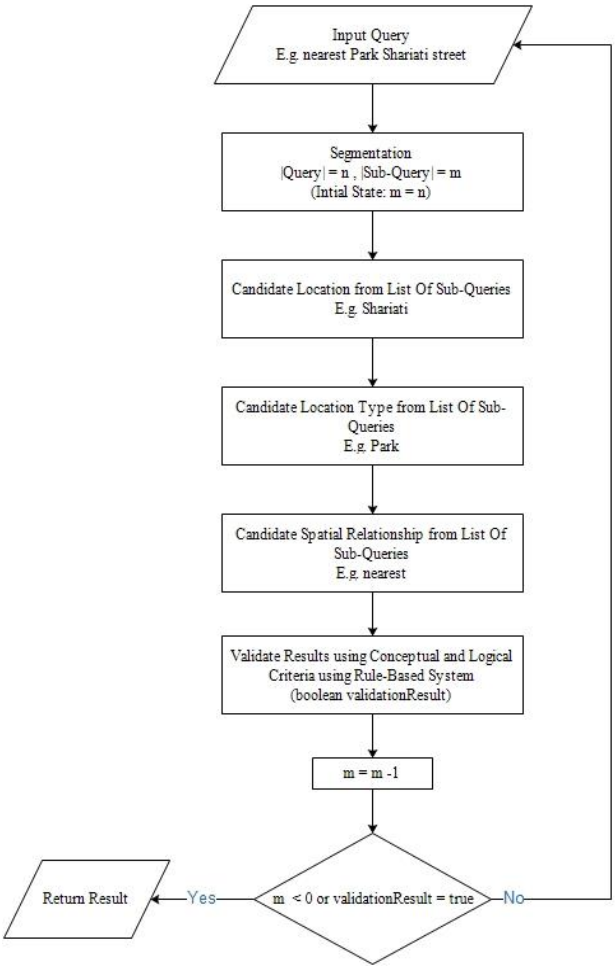

Figure 3 our proposed algorithm

2.2 Candidate sub-queries

The next step is to candidate the location name, so Initially main source of information which means the location name is examined, the outputted sub-queries of the previous section examined whether they are the location name or not. For this purpose a robust and consistent database is needed. Therefore, in this article the Google Geocode API's database is used as our robust and comprehensive data source. One of the important points in this section is that the all ambiguities is provided by Google Geocode web service, So that by entering a name as input, all the locations with that name in the Google database, outputted as the result. If the Web service output contains one or more elements, it can be said that the subquery can be a candidate for the location name. An example of the web service output is depicted in Figure 4.


```
"results" : [
   {
      "address_components" : [
         {
            "long_name" : "Hengam Road",
            "short_name" : "Hengam Rd",
            "types" : [ "route" ]
         },
         {
            "long_name" : "Tehran",
            "short_name" : "Tehran",
            "types" : [ "locality", "political" ]
         },
         {
            "long_name" : "Tehran",
            "short_name" : "Tehran",
            "types" : [ "administrative_area_level_2", "political"
         },
         {
            "long_name" : "Tehran",
            "short_name" : "Tehran",
            "types" : [ "administrative_area_level_1", "political"
         },
```


Figure 4 part of the output of Google Geocode web service

The next step is to candidate the location type, In order to check whether the content of the word is a location type or not. It needs to set up a lexicon of words that has spatial connotation, which it can be done by using existing techniques in NLP. The more explicit and greater lexicon domain, the more queries responsiveness. For this purpose, a rich set of location types which are used by Google web service, along with their synonyms used in this section. In order to compare a word with items in the dictionary there is different NLP solutions which one of those is the use of techniques Word Similarity. We used the Cilibrasi (2007) proposed method for this purpose. If the sub query similarity with that word in the Dictionary was beyond threshold of 80%, the subquery provides the location type and used under the same name. It should be noted that this threshold obtained experimentally by trial and error.

The next step is to candidate spatial relationship, in this section as the pervious section the first step is to prepare the dictionary and the technique used to measure the similarity of the two words is the method described by Cilibrasi. The only thing that distinguishes this part from previous section is the semantic load of words that are in its dictionary.
For example, two words of "in" and "inside", both of them can acts as an expression of the spatial relationship, but they have different semantic loads. In this example, the word "inside" implies more explicitly the spatial relationship of inclusion. To consider this for each element in the dictionary, a number in percentage for words with spatial relationship related semantic load is designated. The multiplication of that percentage number with the rate of similarity between sub queries and the word in the dictionary, used as a measure of being spatial relationship for the sub-queries. Thus for this purpose, the obtained measure compared with the experimental threshold of 70%.

2.3 Results Validation

If the system concludes that the information discovery process is over, the rule based phase started, it happens when the subquery remaining words set did not belong to any of the three categories spatial relationship, location type and location name, the set was empty, or individual elements of it are composed only by one word. Otherwise, the breaking down process resumes for those words that still have not determined efficiently. We used rule based system in order to verifying the candidate outputs and reducing the ambiguity of the results (number of output), for example Intersection is not a unary spatial relationship, so only if query decomposes to a few location's name and intersection relation, it can be done. In fact, implicit constraints on the type of spatial relationship are checked, then the semantic constraints will be considered, as instance intersection between point and polygon is not valid. Study of the implicit constraints can confirm or deny that the set of words have spatial content or not. Then the semantic constraints is applied which be checked out the possibility of output and disambiguates that. Generally In a binary spatial relationships this disambiguation process discard many false cases from the scope of the results, so often in way that output is a unique answer.

3. CONCLUSION

The results of proposed algorithm indicate that changing the search engines perspective from textual to spatial not only expands the domain of supported queries but makes the produced results to be more appropriate with respect to user's

needs. Our studies on 100 spatial queries, show that the system can disambiguates on average, 89.45% of queries result and in fact, by eliminating a large part of results reduced the output ambiguity. This shows the potential of the proposed algorithm in order to disambiguation of locations with the same name; on the other hand, a mechanism for analyzing the spatial search engines is provided by using language processing which makes search engines more intelligent, because while avoiding the complexity and structuring search engine's query, the ability to analyze and understand the elements in the query is added. Integration of this system by major search engines services and other available databases which contains spatial information can improve our system significantly so it can be our suggestion for the future works in order to gaining better results.

REFERENCES

Nadeau, D., & Sekinem S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), pp. 3-26.

Ratinov, L., & Roth, D. 2009. Design challenges and misconceptions in named entity recognition. *In Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 147-155.

Guo, J., Xu, G., Cheng, X., & Li, H. 2009. Named entity recognition in query. *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA. pp. 267-274.

Guo, J., Xu, G., Cheng, X., & Li, H. 2011. Intent-aware query similarity. *In Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM New York, NY, USA. pp. 259-268.

Du, J., Zhang, Z., Yan, J., Cui, Y., & Chen, Z. 2010. Using search session context for named entity recognition in query. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA. pp. 765-766.

Dalvi, B., Xiong, C., & Callan, J. 2014. A language modeling approach to entity recognition and disambiguation for search queries. *In Proceedings of the first international workshop on Entity recognition & disambiguation*, ACM, New York, NY, USA. pp. 45-54.

Krishnamurthy, J., & Mitchell, T. M. Which noun phrases denote which concepts? *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 570-580.

Technologies-Volume 1, 2011 Cilibrasi, R. L., & Vitanyi, P. 2007. The google similarity distance. K*nowledge and Data Engineering, IEEE Transactions*, Stroudsburg, PA, USA. pp. 370-383.