International Journal of Information Sciences and Application.
ISSN 0974-2255 Volume 5, Number 1 (2013), pp. 21-32

http://www.irphouse.com

Extracting Multi Granular Implicit Spatial References in Event Descriptions

V. R. Kanagavalli[1] and Dr. K. Raja[2]

[1]Research ScholarSathyabama UniversityChennai
E-mail:kanagavalli.teacher@gmail.com
[2]Dean (Academics) Alpha College of EngineeringChennai
E-mail:raja_koth@yahoo.co.in

Abstract

The information disseminated in this era comprise of explanation of concepts, description of products and narration of events which obviously has a geographic or spatial reference component present in it. The happening of an event is always tied up to a spatial location. Not only the description of events, but various scientific documents and historical documents are also replete with spatial descriptions in them. This logically leads to the conclusion by the previous studies that more than 80% of the searches are pertaining to geographic locations. Text documents imply the usage of natural language and as such it yields to explicit vague fuzzy descriptions involving linguistic terms such as near to, far from, to the east of, very close and also implicit vague spatial references. Any text document which contains physical location specifications such as place names, geographic coordinates, landmarks, country names etc., are supposed to contain the spatial information. The understanding and extraction of spatial components is a primal area of study not only in the field of information retrieval but also in various other fields such as Robotics, Psychology, Geosciences, Geography, Political Sciences, Geographic Economy, Environmental, Mining and Petroleum Engineering, Natural Resources, Epidemiology, Demography etc., Given a query involving events, the aim of this ongoing research work is to extract both the explicit and implicit spatial references from the text documents, exploiting the granularities. The input to the system would be a text Corpus and a Spatial Query event. The output of the system is the list of implicit and explicit spatial references associated with the event, showing the most possible, disambiguated location of the event queried.

Keywords: Fuzzy Logic, granulation, possibility distribution function, spatial event queries.

1. Introduction

Information is found in various formats such as spread sheets, databases and text documents. Text documents have been a source of descriptive information, instructional materials and narrative documents etc., Queries relating to such documents would more often be a natural language expression most often involving spatial expressions. [5] A spatial expression conveys information about objects in space, their locations, or the frame of reference from which objects are viewed. Understanding a spatial expression includes representing the information it conveys so that it can be used for spatial reasoning. Spatial expressions are found in event descriptions where an event is described as a happening, an occurrence. An event is characterized by: who, what, when, where and some form of attribution [9], the where component of the event may provide implicit or explicit spatial information. Extracting the spatial references from event descriptions would enhance understanding of the text and is a specific branch of information retrieval. A dedicated field of geographic information retrieval actually refers to the entire pipeline of extracting geographic referents from text, indexing them into a spatial index, and allowing spatial search of a corpus using the spatial information [14]. Since it is an active area of research under various domains, the authors have undertaken the work of extracting and inferring implicit spatial references from event descriptions.

This ongoing research work of this author is organized as follows:

- Given an event query determine the level of granulation desired by the user from the query.
- Extract the relevant documents matching the event query from the text corpus.
- Extract the explicit spatial reference from the text.
- Infer the implicit, vague spatial references from the text.
- Generate a list of implicit and explicit spatial references associated with the event from the text.
- Enable the user to select the output based on different levels of granularity.

2. Literature Survey

Extracting spatial data from text is researched under various domains. The most popular application areas include Robotics, Industrial Automation, and Text-to-Scene conversion systems. The task of extracting the spatial component involves the concepts related to information retrieval, Natural language processing and modeling the spatial expressions in text. One of the famous text-to-scene conversion systems, WordsEye is based on VigNet, a unified knowledge base and representational system for expressing lexical and real-world knowledge needed to depict scenes from text [15]. Research in the direction of providing natural language directions to the humanoid by voice instruction suggests extracting a sequence of spatial description clauses from the linguistic input; the humanoid then infers the most probable path through the environment given only information about the environmental geometry and detected visible objects [8].

The four functionally distinct tasks for understanding the spatial expression are extracting spatial expressions and mapping spatial expressions onto spatial primitives. The authors of [5] explore the notion of spatial expression understanding relative to the domain of newspaper captions. Detecting and extracting archaeological events described in natural language text are experimented in the literature [6].

The issues of vagueness that arise in the spatial language that people use when making queries (near to, within walking distance, etc.) and the vagueness in the spatial extent of some geographical regions such as neighborhoods within cities are handled by a body of literature both in GIS and Natural language processing. Qualitative spatial information can be represented by associating qualitative relations with fuzzy sets. Starting with the concept of absolute distance, the authors of [7] have extended the metric notion of proximity to non-metric notions of proximity. The fuzzy methods are employed to model interpretations of spatial language prepositions and of the extent of vague places. In both cases, the parameters for these models are based on data obtained from Web pages that include instances of the various types of vague language and vague neighborhoods. Thus, websites that describe hotels often include textual descriptions that describe their location using phrases such as 'within walking distance' of some prominent location such as a well-known local landmark. The spatial language expressions can then be compared with knowledge of the actual distances involved. The set of actual distances that correspond with the use of a particular expression can be used to create a fuzzy-set membership function.

3. Challenges in handling Indirect Spatial References

Indirect spatial reference is any way to describe a location without using coordinates, generally using a geographic feature to uniquely identify a place. They are important because they are a very common means by which observations of other attribute information are tied to a place. They can serve as a means to link the attribute data to coordinate descriptions of the place to which the attribute data apply. In narrative texts, there is a need for finding spatial relationships that exist between both co referential and non-co referential events to obtain a better understanding.

Spatial referencing is a qualitative process for humans whereas it is a quantitative process for computers. For effective spatial query processing, there is a need of a mechanism to convert the qualitative data to quantitative data. The issues associated with the extraction and representation of vague event descriptions from web documents is discussed in literature [3].

- Vague Spatial Regions and Vague Spatial Relations result in indeterminate boundaries in GIS Modeling.
- There is little support in digital gazetteers for imprecise, vague spatial references.
- Generally digital Gazetteers refer to a point inside the geographic feature (usually the centroid) and do not refer to the actual spatial extent of the geographic features.

- There is currently very less work to incorporate the network features.
- To collect information from multiple documents to resolve the uncertainty in spatial location

Spatial queries are often found in geographic information systems and there is still lot of issues in handling spatial expressions [1].

- Detecting the geographical information within users queries and text documents
- Disambiguating the place names to find the intended **one**
- Interpreting the geometric location of vague place names
- Spatially and thematically indexing the text documents within a GIR System
- Information retrieval model to pick up the relevant documents out of the library and ranking the degree of relevance according to their spatial and non-spatial properties.
- Effective user interface
- Approaches to evaluate the success of a GIR System

4. Proposed Work

The user's query pertaining to specific event is fed into the system. The proposed system retrieves relevant documents from the text corpus. The top ***n*** documents (based on the ranking) are retrieved from the text corpus. The disambiguation may be associated with the spatial query or with the event description in the text document. For now, we deal with the disambiguation of location component of the event description using fuzzy granulation techniques.

Algorithms used for retrieving the relevant documents from the corpus so far have used the footprint of the document, toe print of the document, etc., The algorithms used so far can be technically be divided into two classes either geo-first algorithms or text-first algorithms. The relevance of a document to the given query can be determined by the number of occurrences of the place name in the documents, the place of occurrence, the occurrence of the related places, spatial descriptors etc., This work differs from the existing body of literature in the sense that the spatial relevance and importance of the documents, based on the granulation level in the query are determined using fuzzy distribution functions

The proposed system disambiguates the query and resolves the uncertainty in the text document by considering the fuzzy, ambiguous event description in multiple documents. The advantage of the vector space model is that it allows spatial information to be handled the same way as thematic information with the proper use of proper ontology of places [19, 20].

Taxonomic knowledge of task-relevant geographic layers should be taken into account to obtain descriptions at different granularity levels.

Step 1: Accept the user's spatial keyword query involving an event.

Step 2: Determine the level of granulation required by the user by analyzing the spatial query entered.

Step 3: Identify the most certain areas and least certain areas of the related event
Step 4: Retrieve the documents from the text corpus that most possibly contains the event descriptions and the spatial references of most certain areas.
Step 5: Extract identifiable spatial references, vague and imprecise spatial references using text engineering tool.
Step 6: Use the granulation level to decide the spatial data type for representing the spatial attribute and to determine the possibility of the document matching the user query requirements
Step 7: Use fuzzy logic techniques to model the spatial references.
Step 8: Incorporate the results in a GIS
Step 9: Display the resolved spatial references

4.1 Spatial information in Event Descriptions

The documents generally found in the corpus contain information about products and services also involve spatial component as to where the product can be found or manufactured and the spatial location of the service provided. A query about the event is always tied to a spatial component. Mining spatial information from text reports involves natural language processing and text processing wherein the spatial components expressed in the form of geocodes, place names, addresses etc., would be read in by the system. The spatial relations between events are largely implicit and requires human like intelligence to deduce the spatial bounds of an event. Studying and understanding event descriptions in text is important towards the areas of discourse, semantic processing and general comprehension of textual information. The event descriptions lie not only in verb but also using nouns and adjective modifiers. There are different types of ways in which the events in the text may be related, namely, causation, co reference and temporal ordering [19].

The question that arises in this discussion is, whether it is feasible to use the current techniques of modeling the spatial information embedded in text when maps are so often used in many applications? The authors reiterate and go by the argument that even though the maps are used in many applications still there is lot of scope in modeling the spatial information from the text since there is lot of text documents available in text databases in the form of reports, documents, scientific descriptions, tourist reports etc., The maps as such are not so significant without the associated information with the locations associated with it. Also, the work in this direction would be a stepping stone in achieving better understanding of documents, discourses, add to the efficiency of question and answering systems which would be able to infer the implicit and vague spatial relationships, and also for story telling systems. The spatial information can be acquired through spatial containment relations like same, near, different, overlaps between event descriptions.

Existing models of spatial representation in text, such as SpatialML capture spatial relationships explicitly stated in text or the handling of specific sub-classes of events such as motion events. The concentration is more on geography and culturally-relevant landmarks and not on other domains of spatial language. Also, these annotation schemes don't consider the effect of adjectives or other modifiers associated with spatial locations. None of the models consider implicit spatial

relations between events. All the more, the lack of granularity in event, the actors involved in the event and the spatial locations involved in the event make the understanding of the spatial relations more difficult. Modeling spatial information involves mapping the spatial descriptions on to the logical space using mathematical functions.

So far, the spatial descriptions are mapped using the probability density functions as they possess the properties of formality, practicality, generality and effectiveness. Existing solutions that employ the probability theory are known to be effective and scalable. The authors of this work would like to argue on the same grounds the feasibility of using fuzzy techniques for modeling the spatial information extracted from the text reports.

4.2 Handling Uncertain spatial expressions in Text

Uncertainty can be categorized into two types, vagueness and ambiguity. Vagueness is associated with the lack of clarity of the definition of a class to which a given element belongs whereas ambiguity is associated with the lack of clarity in information about the given element and there exists clear information about the class to it belongs. In the context of spatial information embedded in text reports, there exists a scope of both vagueness and ambiguity. Vagueness exists whenever there is a spatial location specified but it cannot be inferred to which sense it is specified in the document and ambiguity arises whenever there is insufficient information about a spatial location specified in the document. Modeling spatial information is a vital process to visualize the spatial components in the text document. The quality of a model is evaluated based on the complexity of the model and credibility of the model. Uncertainty is a vital factor in increasing the credibility of the model by reducing the inherent complexity present in the real world.

The spatial location names or the spatial expressions associated may be vague which makes it difficult to search the desired location using geographic gazetteers since the gazetteers employ official names of the spatial locations which necessitate the fuzzy extent of the spatial location to be modeled. Being able to interpret such terms will help in analyzing the geographic context of documents and in interpreting user queries that employ vague spatial language [4].

Fuzzy logic is based on the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which membership is a matter of degree. Fuzzy logic allows computing with words closer to human intuition rather than numbers thus allowing the tolerance for imprecision [22]. Fuzzy logic is very apt for systems that would comprehend spatial information from the text repositories since the task calls for more of common-sense and human cognition than any other mathematical methodology. The capability of fuzzy sets to express gradual transitions from membership to non-membership and vice versa provides us with a meaningful representation of vague concepts expressed in natural language. Fuzzy spatial reasoning is a method for handling different types of uncertainty inherent in almost all spatial data. Qualitative spatial information can be best represented using fuzzy logic since it bridges the gap that exists between the real world imprecise or ambiguous system and the perfect objects built in a model. [19]NexTrieve, an information

retrieval technique based on exact search and fuzzy search, the fuzzy logic is applied to position and scoring. The document would get different scores and not all of the words would need to be present in order for the document to get a high score. The drawback is that did not consider word frequency within a document and document length.

4.3 Granularity

Granularity is the concept of breaking down an event into smaller parts or granules such that each individual granule plays a part in the higher level event. Semantic granularity addresses the different levels of specification of an entity in the real world, while spatial granularity deals with the different levels of spatial resolution or representation at different scales [16]. The granularity of the spatial information or spatial query can be used to understand the user's need and provide much relevant information to the user. It can be exploited in multiple domains such as business sectors, disaster response systems, environmental engineering and city planning. The early work of this author throws light on exploiting the granulation of the spatial information for increasing the effectiveness of business activity monitoring [18].

A granularity structure exists only if at least two levels of information are present in text, such that the events in the coarse granularity can be decomposed into the events in the fine granularity and the events in the fine granularity combine together to form at least one segment of the event in the coarse granularity [28, 29]. The authors of the above works have presented the figure [1] in which G_c represents the phrase or sentence with coarse granularity information and G_f represents a phrase or sentence with fine granularity information. Three possible links connect the objects of coarse granularity and the objects of fine granularity - part-whole relations between events, part-whole relations between entities, and a causal relation between the events in the fine granularity and the events in the coarse granularity. The literature study reveals that there are six types of part-whole relationships, Component-Integral, Member-Collection, Portion-Mass, Stuff-Object, Feature-Activity, Place-Area. Their work concentrates on the part-whole relationships including place-area whereas the existing works concentrates on the temporal granulation, and the feature-activity.

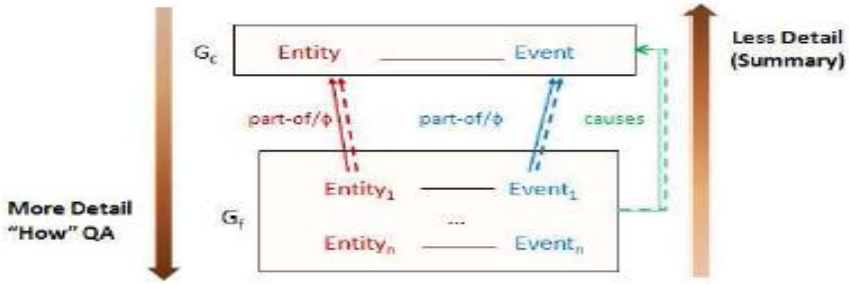

Figure 1: Granularity in Natural Language Descriptions

The difference between afore said work and this work is that the authors of this work are more concerned with the granularity between the spatial entities whereas the

referenced work concentrates on the causal relation between the events.

The issue with implementing granularity in event descriptions in text is that, in case of higher levels of granularity the level of difficulty of annotating the documents is also high and less efficient, whereas lower levels of granularity leads to less inference ability. Granularity can be implemented in all the three components of event descriptions. They are Granularity of an event, granularity of a participant and granularity of a spatial component.

The events may be either static or events-in-action. While considering the events-in-action the starting location, intermediate locations and ending locations are all part of determining the property of event-in-motion. In lower level of granulation, the topological, directional, intermediate points of interest are all to be considered. The actors may exactly define the geographic boundary of the event or may roughly be a part of the event. With respect to the location, the knowledge of spatial relation between the location and the event would lead to better inference. Many of the earlier works have not properly dealt with the granularity of the location. The assumptions made in these works are events are contained in the actor and the location contains the event. The primary goal of this work would be to implement granularity in the location part of the event description. Granularity in query and text documents is to be matched. The higher granularity level would abstract the events or the entities whereas the lower granularity level would throw more light in to the details of the event or the entity. The implicit relationship between the events and the entities are to be explored. The entities in this case are the geographic entities. Granularity in query is previously handled in literature where the granularity level of the information needed is refined by the user himself.

The approaches for granularity identification are of two types as found in the existing literature: namely, Shallow surface level approach and deep semantic approach. The former method uses parse trees to identify entities and events whereas the latter method uses commonsense reasoning for detailed analysis and broader coverage to map part-whole relations from the knowledge base for extracting the granularity structure. The use of fuzzy logic is apt for deep semantic approach since it uses human like reasoning. Also the documents, queries and their characteristics could easily be viewed as fuzzy granular classes of objects with unsharp boundaries and fuzzy membership in many concept areas.

4.4 Fuzzy Vector Space Retrieval Method

Combining fuzzy logic techniques with existing vector based model ensure the simplicity and formalism of the logic based model, and the flexibility and performance of the vector model. Unlike the Boolean model that is based on binary decision criterion {relevant, not relevant}, fuzzy logic expresses relevance as degrees of memberships (e.g., document | query could have a relevance measure with the following degrees of membership: 0.7 highly relevant and 0.5 moderately relevant and 0.1 not relevant) [21]. Fuzzy reasoning is used to determine the importance of a word thus reducing the dimensionality of the vector space and the retrieval result is grouped automatically according to page contents using the vector space model method, the frequency of word appearance and the fuzzy reasoning.

The vector space model is slightly modified where each document is modified as a vector v in the t dimensional space R and the spatial term frequency is defined as the number of occurrences of term spatial term S in the document D, that is, $S_F(D; S)$. The difference between calculating the ranking based on the non-spatial and spatial term is the identification and disambiguation of spatial terms. The (weighted) spatial term-frequency matrix $S_{WF}(D; S)$ measures the association of a term S with respect to the given document D.

Using the traditional vector space method the weight of the spatial term is computed using the following formula,

$$S_{WF}=SF\ X\ IDF=\ SF\ X\ (log_2(N/n)+1) \qquad (1)$$

where *IDF* is the Inverse Document Frequency, N is the total number of documents and n is the number of documents involving the spatial term S. Once the importance of the given spatial term in the document is identified, the next step is to find the spatial similarity between two different documents present in the text corpus. The spatial similarity S_s (D_i, D_j) between two documents D_i and D_j is defined as the measure of possibility that both the documents are referring to the same geographic locations.

By following the traditional information retrieval formula, the spatial similarity of the document is measured as the cosine of the angle between the two document vectors which is calculated as follows:

$$\Sigma_{k=1}^{m}(S_{WF\,i}X\ S_{WF\,j})/[(\sqrt{\Sigma_{k=1}^{m}(S_{WF\,i})^2(\Sigma_{k=1}^{m}(S_{WF\,j})^2]}$$

Where m is the total number of documents, and S_{WFi} and S_{WFj} are the weighted frequency of the spatial term in the document D_i and D_j. The authors propose to apply intuitive fuzzy logic techniques for the retrieval of the documents. It is proposed to be better than the existing methods since, unlike the crisp logic which decides whether the document is relevant or not, fuzzy logic allows computing the relevance as degrees of memberships.

Though there are works using fuzzy logic techniques for the information retrieval, there is no fuzzy logic function applied for the ranking of the documents retrieved. The proposed method would find out whether the terms are spatial references by referring to gazetteer and also applying fuzzy logic memberships and using fuzzy rule base. The word sense disambiguation techniques are used to find out whether a spatial resembling term does actually refer to a geographic location or not. The fuzzy membership functions are assigned to the terms with the use of the fuzzy rule base and the weight of the spatial referring term is calculated based on that. This is much different from the traditional document classification methods.

The fuzzy rules are generated using the fuzzy membership functions for the documents, using the S_{WF} calculated using the formula (1). If SWF is high then the document is highly relevant to the current query involving the spatial term S, if SWF is low, then the document's relevance is estimated to be low. This graded relevance

would help user to resolve the level of uncertainty associated with the documents and the relevance of the document with respect to the query posed by the user to the system.

The parameters for evaluating the efficiency of the information retrieval are precision and recall. The precision helps to evaluate the amount of relevancy of information that is retrieved from the text corpus whereas the recall evaluates the amount of relevant information extracted from the text corpus. For precision, the fuzzy measures can be added to find the degree of relevance of the information whereas for the recall the fuzzy measures can be added to the relevancy of the retrieved documents.

5. Conclusion

The authors are retrieving the documents from the text corpus which are relevant to the spatial queries and are different from the traditional Boolean ranking since the documents are ranked on the basis of relevance using fuzzy logic techniques. The fuzzy membership functions determine the spatial relevance of the documents and the fuzzy rules decide the relevance. The spatial similarity between two documents is also evaluated on basis of the fuzzy rule base. The granularity of the query and the spatial information present in the text are used to resolve the uncertainty of the spatial information. Possibility functions, Fuzzy logic techniques are used to model the uncertainty of the spatial information present in the text instead of the probability logic.

The limitations of this work is that it would answer spatial queries involving spatial attributes only and not spatial queries involving geometric shapes since it is querying the textual data and not the spatial database.

References

[1] C. B. Jones, Ross. S. Purves, 2008. Geographical Information Retrieval, International Journal of Geographical Information Science, 22(3)

[2] S. Kikuchi et al., Place of possibility theory in transportation analysis. Transportation Research Part B 2006. Elsevier

[3] Rock, Nathaniel Robert."Mapping geospatial events based on extracted spatial information from web documents." master's thesis, University of Iowa, 2011.http://ir.uiowa.edu/etd/1068

[4] George J. Klir and Bo Yuan. Fuzzy sets and Fuzzy logic, Theory and applications,

[5] Debra, Rajiv Chopra, Rohini Srihari. Domain Specific Understanding of Spatial Expressions. citeseerx.ist.psu.edu / viewdoc / download ?doi=10

[6] Kate Byrne and Ewan Klein.Automatic Extraction of Archaeological Events from Text. May 2009.

[7] Hans w. guesgen. Reasoning About Distance Based on Fuzzy Sets. Applied Intelligence 17, 265–270, 2002

[8] Thomas Kollar et al., Toward Understanding Natural Language Directions. Naval Research

[9] Geospatial reasoning in a Natural Language Processing (NLP) Environment. Bitters B.

[10] Glander, T. and D¨ollner, J., 2007. Cell-based generalization of 3D building groups with outlier management. In: Hanan Samet and Cyrus Shahabi and Markus Schneider (ed.), GIS, ACM, p.54.

[11] Salton, G. and McGill, M. J., 1983. Introduction to Modern Information Retrieval. McGraw-Hill

[12] Cai, G., 2002. GeoVSM: An Integrated Retrieval Model for Geographic Information. In: Max J. Egenhofer and David M. Mark (ed.), GIScience, Lecture Notes in Computer Science, Vol. 2478, Springer, pp. 65–79

[13] Jones, C. B., Alani, H. and Tudhope, D., 2001. Geographical Information Retrieval with Ontologies of Place. In: D.R. Montello (ed.), Conference on Spatial Information Theory - (COSIT 2001), Vol. 2205 / 2001, Springer-Verlag Heidelberg, Morro Bayand California USA, pp. 322–335.

[14] Kalev H. Leetaru Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia . D-Lib Magazine September/October 2012 Volume 18, Number 9/10 doi:10.1045/september2012-leetaru.

[15] B. Coyne, D. Bauer, and O. Rambow, "VigNet: Grounding language in graphics using frame semantics," in ACL Workshop on Relational Models of Semantics (RELMS), 2011.

[16] Frederico Fonseca et al., Semantic Granularity in ontology-driven geographic information systems. Annals of Mathematics and artificial intelligence. 2011

[17] H.W. Guesgen, J. Albrecht. Imprecise reasoning in geographic information systems Fuzzy Sets and Systems 113 (2000)

[18] V. R. Kanagavalli, K. Raja, Graduated granulation of spatial information for efficient, effective business activity monitoring, Fuzzy Sets and Systems, pp 99-101. 2010

[19] Mulkar-Mehta, R.; Hobbs, J. R.; and Hovy, E. 2011. Granularity in Natural Language Discourse. International Conference on Computational Semantics, Oxford, UK 360—-364

[20] Mulkar-Mehta, R.; Hobbs, J. R.; and Hovy, E. Applications and Discovery of Granularity Structures in Natural Language Discourse. Logical Formalizations of Commonsense Reasoning — Papers from the AAAI 2011 Spring Symposium (SS-11-06)

[21] N. O. Rubens. The application of fuzzy logic to the construction of the ranking function of information retrieval systems. Computer Modeling and New Technologies, 10(1):20–27, 2006.

[22] Zadeh L.A. (2004) Foreword to Fuzzy Logic Toolbox User's Guide. The MathWorks Inc