EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Spatial Ambiguities Optimization in GIR

Arun Kumar Yadav[1], Jay Kant Pratap Singh Yadav[2], Divakar Yadav[1,*]

[1]National Institute of Technology, Hamirpur (H.P.), India
[2]Ajay Kumar Garg Engineering College, Ghaziabad (U.P.), India

Abstract

INTRODUCTION: Huge amount of geographically referenced information is available on World Wide Web and has become an excellent source for retrieval of desired information. Extraction of relevant information from such a huge unstructured source is not an easy task.
OBJECTIVES: In this paper, we propose a location based search engine, designing two indexes namely spatial index and inverted index. Using these indexes, we are able to solve toponym ambiguity and overlapping of documents.
METHODS: To handle toponym ambiguity problem, we designed and implemented architecture for directional web document search. The architecture poses spatial followed by textual indexing for toponym resolution and to reduce overlapping of documents in geographical locations.
RESULTS: The proposed architecture was implemented and tested on spatial and textural data sets.
CONCLUSION: The performance is measured in terms of precision and false positive parameters and found that the proposed architecture performs better for geo/geo and geo/non-geo search queries.

Keywords: Geographical information retrieval, geoparsing, geotagging, gazetteer, spatial indexing, textual indexing.

Received on 16 April 2020, accepted on 07 May 2020, published on 14 May 2020

doi: 10.4108/eai.13-7-2018.164556

*Corresponding author. Email: divakaryadav@nith.ac.in

1. Introduction

Now-a-days, searching from Web is dominated by search engines like Google, Yahoo, Bing etc. where, web documents are extracted according to user's need. To search relevant documents from billions of web documents is not a easy task for any search engine. Specially, due to large number geographical as well as non-geographical references, it is not easy to search relevant data with respect to correct geographical reference. In the past, researchers proposed different models of search engine.

Geographic information retrieval [1, 2] is a specific field of information retrieval which focuses on spatial and thematic indexing. Accessing information through geographical references is very useful when users are looking for several resources on a territory or when planning rescue operations during emergencies due to natural disasters. For geographical searching, many search engines have been proposed and developed, some of them are SPIRIT (spatially aware information retrieval on the internet) [3, 4] project which mainly focus on improvement in GIR, and GIPSY (georeferenced information processing system) which rely on gazetteer to identify geographic names.

The addition of spatial knowledge in search engine [5] brings various challenges. There should be facility to recognize the place name in GUI of search engine. For the development of generalized search engine, it would need to maintain knowledge of every geographical location. Some techniques or algorithms have been developed to identify the geographic scope of web documents. Set of geographical coverage of a web document is known as document footprints [6, 7]. As document footprints are determined for fast retrieval of relevant web documents using spatial indexing pertaining to query footprints. Geography is the main criterion to search the resources in a location. It affects our day to day lives, thus we expect a spatially aware search

engine i.e. a framework which may support geographic references and would have great impact on search technology. Existing work in this area generally does not focus on removal of ambiguity problem. Ambiguities are of two types as defined in [8] i.e. geo/geo and geo/nongeo. Geo/geo ambiguity is in which one place have multiple names or one name is used for several places and geo/nongeo ambiguity is one in which organization's name or person's name is used for place name. These two ambiguities come under toponym resolution. In past, researchers proposed many spatial as well as textual indexing mechanism to retrieve relevant web documents [12, 22, 28]. In the paper [13], authors proposed geographical information retrieval system using geographical markup language. It is extended version of extensible-markup language (XML) which is more supportive for retrieval of spatial as well as non-spatial documents with improved search attributes. It supports good trade-off on spatial as well as textual indexing mechanism to reduce the search time. In the paper [12], authors proposed hybrid indexing technique using wavelet tree to reduce search time and in the paper [28], authors proposed dual indexing mechanism using wavelet tree. These indexing methods (dual and hybrid) outperform in terms of space and time complexity. As per best of my knowledge, no one has discussed about search result optimization by resolving toponym resolution.

In this paper, we propose an architecture of location based search engine for toponym resolution to optimize search results.

The organization of this paper is as follows: Section 2 briefly introduces previous and related research. Section 3 illustrates architecture of spatial search engine, database and design of user interface. Section 4 discusses evaluation and experimental results. Finally, we conclude paper in section 5 followed by references..

2. Backgrougnd

Many contributions have been made in geographical information retrieval. Nevertheless, due to web documents regular growth, research in this area is still infancy. In this section, we discuss geographical search engines that have been developed and some studies on proposed models [10].

Geo-referenced information processing system (GIPSY) is a model for geographically based access to text. With the help of this model, geographic words or phrases are extracted from web documents. Georeferenced documents are extracted according to the location of its text. SPIRIT(spatially-aware information retrieval on internet) [3, 4,14,15] is a research based project that has been employed in the design and implementation of search engine to find out the web documents related to places mentioned in the query. Milestones for this project are: a) Geographic ontology that make it capable for expansion of query and include the web documents that are nearby. b) User interface which allows description of place using text, and c) Ranking of results according to geographic and thematic concepts [16, 17].

Geographic search engine based on vector space model (GeoVSM) [28] is an abbreviation for geographic vector space model. The project combines coordinate based geographic indexing with a keyword based vector space model to represent information. Space relevancy depends on both spatial and textual measures which can be integrated into single measure system. In this work focus scoring algorithm is developed which is used for calculation of score of web page. In this algorithm a threshold is defined and values are assigned to places. Places which have values above threshold are more important than other places.

Spatio-Textual Extraction on the Web Aiding Retrieval of Documents (STEWARD) [18] search engine is for extraction of geographic references and determine Geographic focus. It uses document tagger, which identify nouns and tag as geographic references if these resemble to the names of location present in gazetteer. Geo-referenced determination process is divided into two parts (a) Determine the geographical references (b) Ranking of these references to determine geographical focus of web document. Applications of this system are used for collection of web documents in hidden web, for reading news articles, also for diseased monitoring system.

Paper [8, 21], has given system architecture for geographic information retrieval. This architecture contains 3 layers: a) index construction b) processing services c) user interface. Bottom layer contains document abstraction module and indexing module. Indexing can be hybrid structure or double index structure. In this work author proposes some improvements in SPIRIT project. Two algorithms, text first and geo first are described. Ontology is presented here which shows relationship between objects and this relationship is not presented in any other architecture. Index structure shows the combination of textual and spatial indexing. Though in dereferencing of web documents, a geo footprint is assigned to each web document but the problem remains for ambiguity of geographic references. New research topics have emerged in this young area. Firstly dereferencing technique must be improved to resolve ambiguity problem and new index structure should be developed.

In Geographical constrained information retrieval by Andogah, Geoffrey [23] 18 % of information aspirants search for geographically intelligent information retrieval systems. Generally, information retrieval system lack geographical intelligence, needed to effectively answer geography dependent questions. Two research objectives are: a) how to hoard and analyze the geographic information present in the text and b) How geographical knowledge is utilized to build models for answering geography dependent questions. It is assumed that every document and query has its own geographic scope (i.e. where the events described are situated). For utilization of notion geographical scope, techniques are developed to detect the location. A common GIR processing system is as shown in Figure 1.

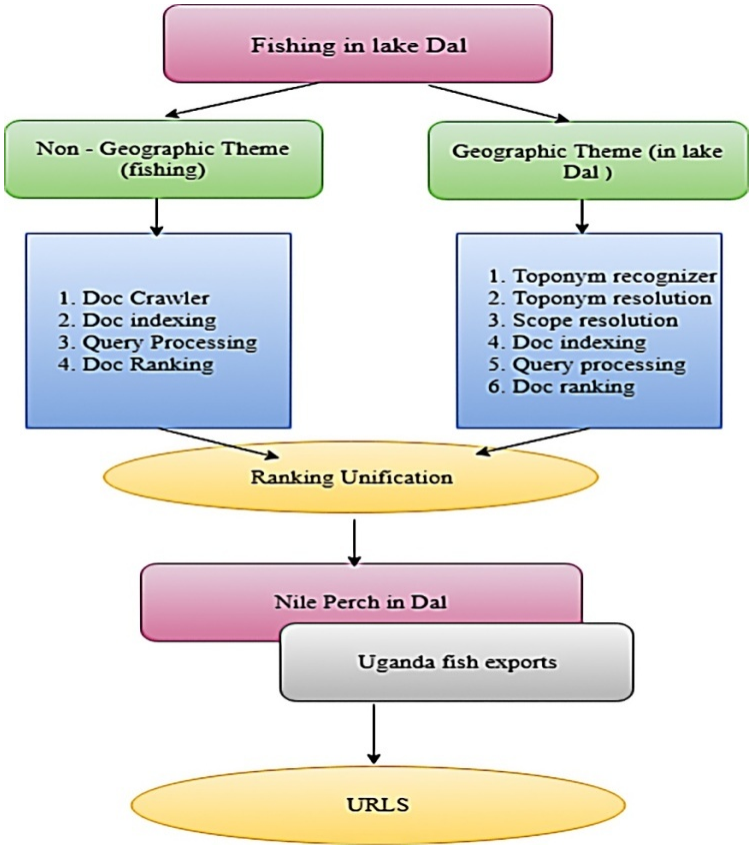

Figure 1. Common GIR processing Process (Andogah Geofrrey, 2010)

3. Methodology to Design Location Based Search Engine

Figure 2 shows the system architecture for proposed location based search engine. The architecture is divided into following components: user interface, repository of web pages, relevance ranking, web crawler, geoparser, textual parser, and ranking. Functionalities associated with all these components are described as below.

Web Crawler: It is a program which searches the web documents according to user's need. These are specially used to create a copy of all visited pages for later processing by search engine that index the crawled web pages.

Repository: Results in the form of URLs are stored in the indexing after removal of stop words etc.. This is connected with spatial and textual indexing with the help of word id (Table 1).

Gazetteer: It is implemented in the form of database which contains location names as shown in Table 2. In the gazetteer, one column is used to store synonym of location name. When user provides query with different names of same location, which is geo/geo ambiguity, it is removed with the help of this field i.e due to synonyms field in the table, results are same for different names of same location. Other columns are to store latitude and longitude coordinates of location.

Table 1. Repository

Id	Word-id	Word	url
365	20	colleges	http://www.employeement.org
366	20	colleges	http://www.upcolleges.org
367	20	colleges	http://www.indiancolleges.org
368	21	hospitals	http://justdial.com
369	21	hospitals	http://www.Worldhotels.com

Indexing Structure: It refers to the capability to process a textual document and identify keywords and phrases that have spatial context [24, 25]. It involves reading of the text,

finding place names and hyperlinks in them, shown in Figure 3.

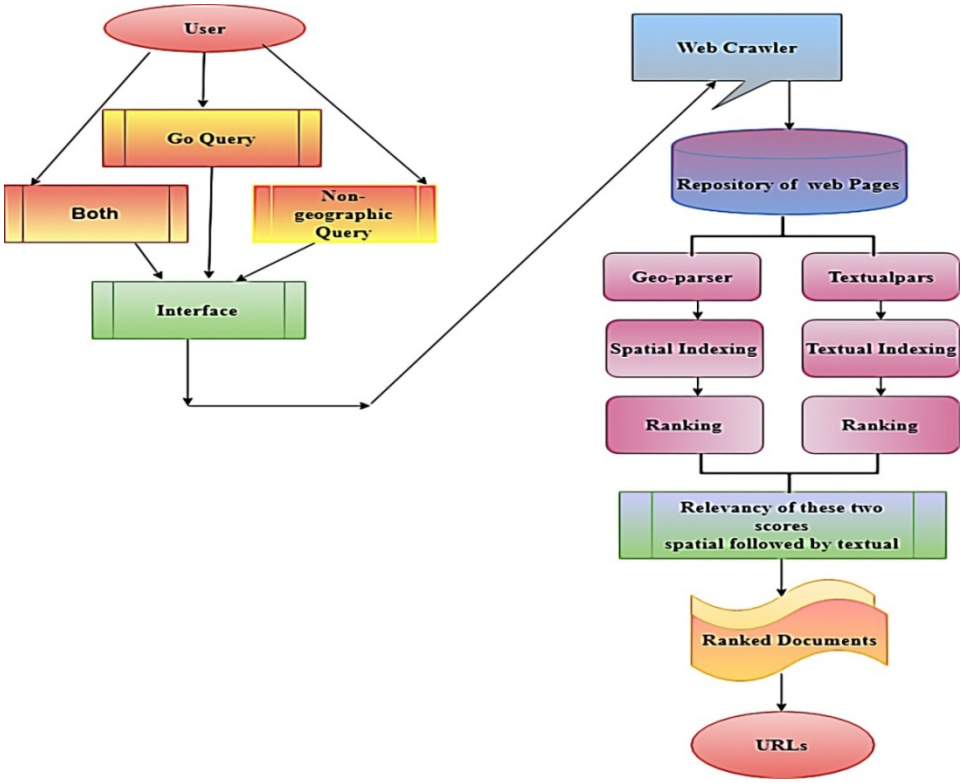

Figure 2. Architecture of Proposed Model

Spatial index: This index contains same columns as in textual indexing and score is calculated [26]. This index is formed on the basis of spatial words which are location names as shown in Table 3.

Textual parser: It refers to the capability to process a textual document and identify keywords, phrases that have a textual context [26]. It involves reading of text and hyperlinks in them. If a word is not a location name then it will be treated as a textual word.

Textual Index: Columns in textual indexing are *word* id, Doc_id, Keywords and other relevant information used for calculating the weights and ranking thereafter. Using different attributes like page title, meta_keywords etc., total weight is calculated which gives a score for the document as shown in Table 4.

Table 2. Gazeteer

Id	Location	Synonym	Latitude	Longitude
1	Abdul Nagar	(NULL)	276900	797300
2	Shastri Nagar	NULL	264200	764250
3	Amrawati	NULL	301700	795500

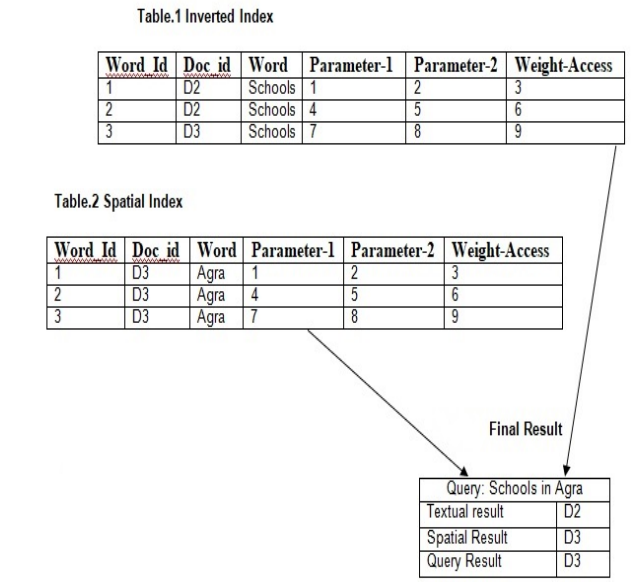

Figure 3. Index Structure

Table 3. Spatial Index

Id	Search_result_id	Location	Page_Title	Meta_keywords	Meta_description	Link_a
1	1	Agra	0	0	0	1
2	1	Agra	0	0	0	6
3	1	Kanpur	3	4	9	3
4	1	Kanpur	3	4	9	3

Table 4. Textual Index

Id	Search_result_id	Location	Page_Title	Meta_keywords	Meta_description
1	1	Hotel	0	0	0
2	2	Hotel	0	0	0
3	3	Hotel	3	4	9
4	4	Hotel	3	4	9

Spatial queries: The term spatial queries imply querying a spatially indexed database based on relationships between particular items in that database within a particular coordinate system. Spatial querying is a general term, and can be defined as queries about spatial relationships (containment, intersection, and boundary) of entities geometrically defined and located in space without regard to the nature of coordinate system. Types of spatial queries are Containment query, Region query, Enclosure query, Clipping, Line intersection query, Adjacency query, Proximity queries and Range query. In this work we covered containment, proximity, direction based enclosure and range queries.

Containment query: If a query is containment type such as "hotels in Agra" or "Schools in Ghaziabad" then resulting links will be related to all directions of Agra and Ghaziabad within a given distance.

Directional queries: For the queries such as 'hotels in north of Agra' a concept based on bearing is used to divide the space into cardinal directions i.e. N, S, E, W. Bearings are a measure of direction. It is measured from north to south in clockwise direction. If user is travelling in north direction then bearing is 000 degree and if the user is travelling in any other direction then bearing is measured clockwise from north. This bearing is used to differentiate cardinal directions i.e. North, South, East and west directions.

Spatial Indexing: It is the most important stage of our work. First, the system analyses the document's fields that are marked as spatially indexable and identify candidate location names from texts. In second step, these candidate locations are processed in order to determine whether candidate locations are real location names or not. To compute the geolocations, there are some problems that can happen at this stage and those are geo/geo ambiguity problem and geo/nongeo ambiguity problem. It extracts frequency of spatial words in different sections of web documents such as Title, Keyword, Description, and Body.

Ranking: The ranking component considers results retrieved from spatial search engine database and ranks the documents with respect to spatial and non spatial elements of the query separately [27]. Separate scores are calculated for both type of indexing spatial as well as textual. Spatial score is calculated using the formula given in equation 1.

$$Spatial\ Score = (frequency\ of\ spatial\ keyword * wt\ of\ attribute_1 + frequency\ of\ spatial\ keyword * wt\ of\ attribute_2) * 1000 \quad (1)$$

Where, attributes are page title, meta_keywords, meta_descrption, link etc. as shown in Table 5.

For example, if weight of all parameters is considered as 1 then score is calculated as follows for "Kanpur" location according to query in Table 3. Spatial Score = ((0*1) + (0*1) + (0*1) + (0*1)+(0*1)+(8*1))*1000 = 8000 (highest score for "kanpur" location in query)

Similarly, textual score are calculated using the formula given in equation (2) and example is shown in Table 6.

$$Textual\ Score = frequency\ of\ textual\ keyword*wt\ of\ attribute_1 + frequency\ of\ textual\ keyword* wt\ of\ attribute_2 \quad (2)$$

For example textual score for row two in table 4 = (2*1) + (1*1) + (2*1) + (12*1) + (0*1) + (108*1) = 125 (highest score for textual word "school" in query).

Table 5. Highest Score in spatial index

Definition	Paper Title	Meta. Key	Meta_des.	Link_a	Meta_alt	Html_body	Totall_Weigh (T+S)
NULL	1	0	3	3	8	5	4125
NULL	0	0	0	0	0	8	8125
NULL	3	8	5	4	11	4	4125
NULL	4	3	2	7	3	11	4172

After calculating both the score i.e. spatial and textual final score of a document is obtained by summing it. One example of such calculation is shown in Table 5 in which the maximum score obtained is 8125, the most relevant url score, obtained after addition of highest spatial and highest textual score.

Table 6. Highest Score in textual index

Definition	Paper_Title	Meta.Key	Meta_des	Link_a	Meta_alt	Html_body	Weigh
NULL	1	0	3	3	7	1	2
NULL	2	1	2	12	0	108	125
NULL	3	9	2	3	9	7	8
NULL	4	3	2	4	9	10	11

User interface: It allows the user to specify subject of interest and geolocations. The terms that form subject of the query are combination of non- spatial terms such as 'hotels', 'schools', 'colleges' and spatial terms such as north, south, east, west, near, around, in etc. Thus it has the capability to recognize textual and spatial keywords (Fig. 6).

Figure 6. User Interface

4. Implementation Results and Discussion

The experimental work was done on windows 7, 64 –bit operating system with Intel (R) core i5-2430M CPU @ 2.40 GHz processor and java, jsp programming language has been used. Apache Tomcat server is used to run the web pages as front end. MySQL database is used to create database of spatial search engine. Queries are divided into two categories: query for geo/geo and query for geo/nongeo. We crawled approximate 5000 locations with latitude/longitude and web references from India and design spatial index, Same way we crawled approximate 100000 web pages with web references and design textual index for textual search. The proposed model was tested on varieties of queries of both types geo/geo and geo/nongeo. Some examples of geo/geo queries on which model were tested are "Schools in Ghaziabad", "schools in Gzb" etc. Similarly, the example of geo/nongeo queries is "schools in Ram" where Ram is person name, a nongeo query word.

Table 7 and Figure 7 show the comparative performance of proposed model with other tools. The performance of location based search engine is better than other search engines. Further, Table 8 and Figure 8 shows that the Geo/Nongeo ambiguity is minimized in location based search engine in comparison to other search tools such as Google, Yahoo and Bing which always gives higher false positive results for these Geo/Nongeo ambiguity queries as shown in Table 8.

Due to non accessibility of data sets/repository of popular search engines, all the generated queries explained in Table 7 and Table 8 were executed on different search engines such as Yahoo, Bing, and Google and the results obtained were analysed manually and latter these results were compared with the results, obtained from proposed location based search engine. This process continues for both textual and spatial search results. Precision in Table.7 and false-positive results explain in Table.8 shows that proposed system out perform in terms resolving toponym ambiguity and accuracy.

Table 7. Comparison of Geo/Geo search results

Precision				
	Location based search engine	Google	Yahoo	Bing
Colleges in Ghaziabad	85%	71%	74%	68%
Colleges in GZB	73%	62%	65%	59%
Hotel in Kanpur	87%	71%	64%	63%
Hotels in CNB	78%	79%	73%	61%

Table 8. Comparison of Geo/Non-Geo results

False Positive Results				
	Location based search engine	Google	Yahoo	Bing
Schools in Vidushi	4	52%	48%	57%
Colleges in Vidushi	5	57%	67%	56%
Schools near Amit	7	47%	52%	51%

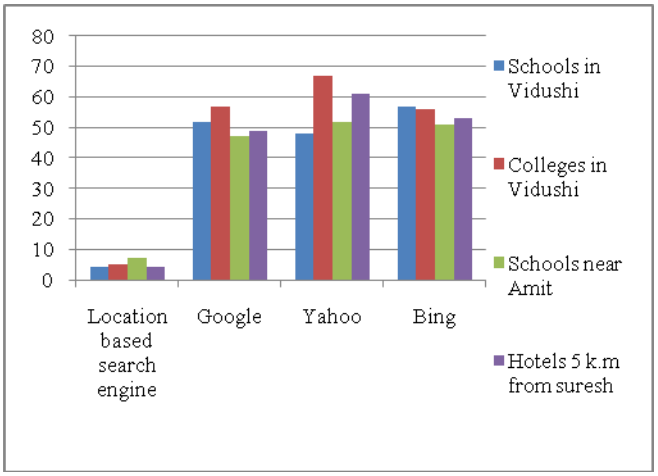

Figure 7. Geo/geo ambiguity

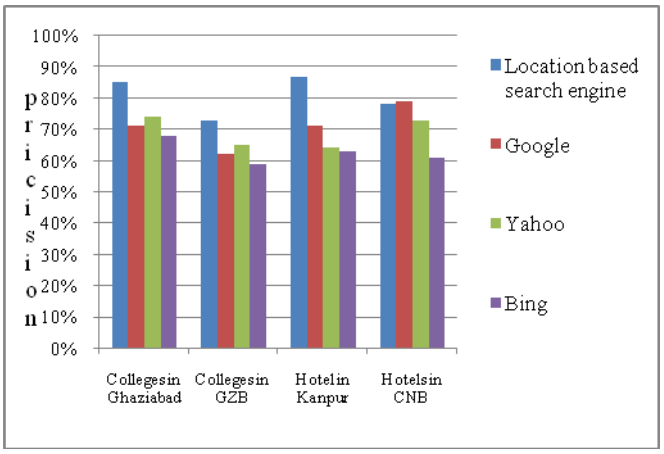

Figure 8. Geo/Non Geo ambiguity

5. Conclusion and Future Work

In fact, we worked and developed a prototype to resolve the Spatial Ambiguities problem in GIR. In this work we proposed architecture of spatial search engine and focused on unification of textual and spatial score on the basis of a number of attributes of web documents. The designed architecture demonstrated the viability of overall design. Bearing and coordinates plays a key role in providing support for location name disambiguation. Frequency and weight of terms plays important role in score calculation for ranking and to display top-k results according to user's requirement. Experimental results show that the improvement in proposed architecture with respect to standard search engines such as Google, Yahoo, and Bing etc.

It would be interesting to consider the best way to make geographical expansion of the query and improvement in ranking process. In addition, exploring the use of different ontologies and determine how each ontology affects different resulting index. We also plan to include other types of spatial relationships in the index structure in addition to inclusion (e.g. adjacency). This type of relationship can be easily represented by ontology-based structure and indexing structure can be extended to support them. It would be better to design both type of repository i.e. file based and server based repository.

Due to resource constraints, we tested the proposed model on a limited number of data which it can be scaled up in future for larger set of data. Further, machine learning approaches can be explored to rank the documents.

References

[1] G. Bordogna. Geographic information retrieval: Modeling uncertainity of user's context. Proceedings of advances in soft computing applied to database and information system. June 2012, pp. 105-124.

[2] Christopher B. Jones and Ross S. Purves. Geographic Information Retrieval. Proc. of International Journal of geographic information science. 2008, pp. 219-228.

[3] Christopher B. Jones et al. Spatial information retrieval and geographic and geographic ontologies: An overview of SPIRIT project. 2002, pp. 1-2.

[4] Christopher B. Jones et al. The SPIRIT spatial Search Engine: Architecture, Ontologies and spatial indexing. 2004, pp. 1-16.

[5] Alexander Markowetz, Yen-Yu Chen, Torsten suel, Xiaohui Long, and Bernhard Seege. Design and implementation of a Geographic search Engine. 8th international work shop on web and databases. 2004, pp. 1-6.

[6] G. Andogah, G.Bouma, and J.Nerbonne. Every document has geographical scope. Data and Knowledge Engineering 2012, Vol. 81-82, pp. 1-20.

[7] Nieves R. Brisaboa, Miguel R.Luaces, Angeles S. Places, and Diego Seco. Exploiting Geographic references for Web Documents in a Geographical Information Retrieval System Using an Ontology-based Index. Database laboratory, University of A Coruna. 2010, pp. 1-23.

[8] Nieves R. Brisaboa, Miguel R.Luaces, Diego Seco. New Methodologies in GIS: Improving the information retrieval process. 2010, pp 1-16.

[9] Vidushi Vidyarthi, Arun Yadav, Divakar Yadav. New Methodology in GIR systems: Improving web document searching. Proceedings of Sixth international conference on contemporary computing. 2013, pp. 208-212.

[10] Alexander Markowetz, Yen-Yu Chen, Torsten suel, Xiaohui Long, and Bernhard Seege. Design and implementation of a Geographic search Engine. 8th international work shop on web and databases. 2004, pp. 1-6.

[11] Allison Gyle woodruff et al. GIPSY: Automated Geographic Indexing of Text Documents. Proceedings of journal of America for information science. 1994, pp. 1-21.

[12] A. Yadav and D. Yadav. Wavelet tree based hybrid geo-textual indexing technique for geographical search. *Indian J. Sci. Technol.*. 2015, vol. 8, no. 33, pp. 1–7.

[13] Fang, Caili & Zhang, Shuliang. Geographic Information Retrieval Method for Geography Mark-Up Language Data. ISPRS International Journal of Geo-Information. 2018, 89(7), pp. 2-23.

[14] Ross S. Purves, Paul Clough, Christopher B. Jones. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. International Journal of GIS. 2007, 21(7), pp. 717-745.

[15] Subodh vaid and Christopher B. Jones. SPIRIT report on Spatial indexing methods. 2004, pp. 1-37.

[16] F. Fu, C.B. Jones et al. Ontology Based spatial query expansion in information retrieval. 2005, pp.1-16.

[17] Christopher B. Jones et al.. Maintaining ontologies for geographical information retrieval on the web. 2003, pp.1-18.

[18] Michael D. Lieberman. STEWARD: Architecture of a spatio textual search engine. 2007, pp. 186-193.

[19] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web–a- Where: Geotagging web Content. SIGIR. 2004, pp. 273-280.

[20] Compelo et al.. Spatial Search Engine. Encyclopedia of Information Science and Technology. Second Edition, IGI Global. 2009, pp. 3554-3558.

[21] Miguel A. Garcia Cumbreras et al. Information retrieval with geographic references. Relevant documents filtering vs query expansion. 2009, pp. 605-614.

[22] Selvaganesan, S.; Haw, S.C.; Soon, L.K. Effective XML keyword search using dual indexing technique. Inf. Technol. J. 2014, 13, pp. 643–651.

[23] G. Andogah. Geographically constrained information retrieval. PhD thesis. 2010, pp. 1-205.

[24] Orkut Buyukkokten, Junghoo cho, and Hector Gracia-Molina. Exploiting Geographic Location Information of Web pages. Deptt. Of Computer Science, Stanford university. 1999, pp. 1-6.

[25] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing Scopes of web resources. 26thinternational conference on very large databases. 2000, pp. 545-556.

[26] Subodh vaid. Spatio textual indexing for geographic search engine on web. 2005, pp. 1-18.

[27] Martins, B. Silva, M.J., & Andrade, L. Indexing and ranking in Geo-IR systems. Workshop on geographical information retrieval, GIR'2005. 2005, pp. 31-34.

[28] A. Yadav and D. Yadav. Wavelet Tree based Dual Indexing Technique for Geographical Search. *Int. Arab J. Inf. Technol.* 2019, vol. 16, no. 4, pp. 624–632.

[29] CAI, G. Geo. VSM: An integrated retrieval model for geographic information, in Proceedings of the Second International Conference on Geographic Information Science, Boulder, CO, United States. 25-28 Sept 2002, pp. 65-79.