Journal of Computer Science, 9 (1): 83-93, 2013
ISSN 1549-3636

doi:10.3844/jcssp.2013.83.93 Published Online 9 (1) 2013 (http://www.thescipub.com/jcs.toc)

Efficacious Geospatial Information Retrieval Using Density Probabilistic Document Correlation Approach

[1]Uma, R. and [2]Muneeswaran, K.

[1]Department of Computer Science and Engineering,
Jayamatha Engineering College, Aralvaimozhi-629301, Tamil Nadu, India
[2]Department of Computer Science and Engineering,
Mepco Schlenk Engineering College, Sivakasi-626005, Tamil Nadu, India

Received 2012-10-09, Revised 2012-12-18; Accepted 2013-02-16

ABSTRACT

Information Retrieval (IR) is a profound technique to find information that addresses the need of query. Processing of normal text is easier and information can be retrieved efficiently. There are plenty of algorithms in hand to carry out the normal text retrieval. Whereas retrieving geospatial information is very complex and requires additional operations to be performed. Since geospatial data contain complex details than general data such as location, direction. To handle geographical queries, we proposed a Density Probabilistic Document Correlation (DPDC) approach. This approach, initially categorize the geographical features from text that satisfies the given queries. Existing text classification techniques are unsuitable for geospatial text classification due to the exclusivity of the geographical features. Depending on the DPDC approach result we predict overlap of the feature set for a document. Based on overlap and document correlation, the documents are ranked. Highly relevant documents are extracted depending on the score obtained through ranking. Documents with high score are considered the most relevant. The experimental results show that our proposed method efficiently retrieves the list of relevant documents.

Keywords: Geospatial Documents, Information Retrieval, Ranking, Feature-Selection

1. INTRODUCTION

For the past several years, geographical data has been collected enormously by various organizations and archived at globally distributed location. Research scholars, educational institutions, governments, students, engineers, scientists and other interested person are able to access this precious spatial data through the internet. However, the source of data and diversity in spatial data poses challenges for the user. Since the user tries to assemble multidisciplinary data set for a specific learning. This brings the need for mining spatial data in order to help the user to retrieve the useful information.

Knowledge discovery from a large geospatial database began approximately a decade ago. Mining spatial data is the progression of discovering interesting patterns, which were formerly unknown, but potentially useful for large spatial data sets. Mining of geospatial information requires more knowledge about the economic, environmental and social phenomena. Documents of geospatial consist of various thematic maps with multiple objects in each layer and also mention the relationships and auto-correlation properties. Geospatial data mining is more difficult than the traditional categorical and numeric data. Since complicated data type and intrinsic relationship between non-spatial and spatial components as well as the association between spatial data makes geospatial data mining difficult. Spatial data type may be either numerical and categorical or spatial. Graph, polygons,

Corresponding Author: Uma, R., Department of Computer Science and Engineering, Jayamatha Engineering College, Aralvaimozhi-629301, Tamil Nadu, India

line string and points are some of the spatial data type. Such data types are correlated through the relationship like topographical, metric and directional.

Spatial database incorporates more irrelevant and redundant attribute, which should be removed efficiently. Feature selection is highly complex in spatial data mining for the reason that spatial data are frequently related to each other. The correlation and relationship exist among spatial data are frequently handled by the algorithms of data mining. It includes the spatial relationship through transforming spatial to non-spatial attributes. The intricate of spatial data mining is that of transforming spatial to non-spatial features. Therefore, it is essential to select the features efficiently.

Selected features, plays a key role in further steps of data mining. Based on the features chosen the relevant documents are retrieved. User expresses their interest in the form of queries to a component, which performs the search operation. As a result of search operation, lists of documents are listed in an increasing order through a ranking function. The function of ranking is to compute the score of each document. Ranking algorithms depends greatly on query that is executed and the information that is required.

This study presents a framework for retrieving relevant information. **Fig. 1** shows the architecture of the proposed method.

We initially preprocess the documents that are retrieved from the database for a user query. During preprocessing, stop words are removed in the documents since they carry very less important meaning while comparing with the keywords. Stop words has the following impact on the geospatial information retrieval. Stop words have the impact on the retrieval process since they have high frequency of appearing in document with less meaning and affect the weighting process, which is carried out in our Density Probabilistic Document Correlation (DPDC) approach. This is because that stop words affects the document length. Consequently, the length of the document affects the DPDC weighting process. Therefore, stop word list is necessary to remove the words that do not add any significant importance to the text. There are two different type of stop word list (1) Domain Dependent (2) Domain Independent. There exist many tools for removal of stop words in the market, which are commonly used for English language that can be used for domain independent removal of the stop words. Using a tool that is available online can be used, or we can also create a tool that is specific to the domain for the removing stop word.

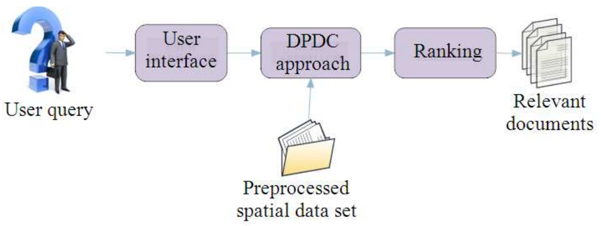

Fig. 1. Overall architecture

Following the stop word removal, we perform POS tagging, which is the process of assigning a part-of-speech like verb, adverb, preposition, noun, pronoun and other lexical class maker to each word in a sentence. POS-tagging algorithm is based on either rule based or stochastic based. It plays a vital role in various areas of natural language processing. On removing stop words, most of the words within a document will be nouns and adjectives. Retrieve those nouns and adjectives for processing. For weighting the document, geographical features that are relevant for the query are chosen, which is already trained.

The preprocessed documents or data sets and the user query are given to the DPDC approach component. In the DPDC approach, we determine the document correlation through probability and density of the given documents. Depending on the result of the DPDC and the degree of features overlapped on a document is used for ranking of the documents.

Ranking is normally employed by the search engines in order to retrieve most related documents for a given query. The effectiveness of a search engine is based upon the ranking method it is used for retrieving the documents. Ranking are taken in two different fashions. (1) Considers the importance, content score and relevance of a document. (2) Depends on the popularity score of the document, i.e., the link structure. Some ranking method includes both the aforementioned technique. If the user required information are not displayed according to his/her interest, then the retrieval system looses it popularity. So, in order to retrieve relevant document, we use the ranking algorithm, which determines the occurrence trained features in a document. Depending on the degree the documents are ranked, they are listed to the users in increasing order (descending order). Most relevant documents for the user queries are in the top of the list, whereas the irrelevant documents are not retrieved, which are eliminated by the DPDC approach.

2. RELATED WORK

Searching of relevant information across a geospatial database is an increasing and also interesting research area. Many of the research scholars are trying for a method to retrieving information that is related to a given query. Here, we present some of the related works carried out earlier by various scholars.

A lot of information stored on the web pages contains the geographical context, but older search engine treat those information as a normal way as all other context. It is required to design and implement spatially aware search engine. Scholars of an article in (Purves *et al*., 2007) designed such a spatially aware search engines, which is capable of queries that were in the triplet format (i.e., <theme><spatial relationship><location>). To design such search engine they identified geographical references in documents and assigned corresponding footprints to the documents and stored along with the term in the document terms in an appropriate indexing structure. It explored the related results for the query that was ranked based on both the query that was ranked based on both thematic and spatial relevance. Usability study was undertaken by the authors of (Purves *et al*., 2007) and the analysis showed that users were satisfied with the range of spatial relationships available and intuitively understand how to use search engine.

Huge collection of web pages was essential for research in geospatial information retrieval system. A study was presented by the authors of (Joho and Sanderson, 2004) in which they provided an overview of huge web-page document collections that were used for the SPIRIT project meant for the testing and design of the spatially-aware information retrieval system. With the use of SPIRIT, a method was proposed as an article (Clough, 2005) to retrieve geographic information from the web. They annotated 900,000 web pages containing geospatial information that was focused on the regions of UK, Germany, Switzerland and France and taken from a 1TB web crawl. Authors also discussed a tool for extracting the spatial metadata, which was based upon the GATE Information Extraction (IE) system and besides a simple geo-coding program to dispense spatial coordinates to extracted locations. Along with that analysis was made for geo-parsing and geo-coding is provided together with an initial statistical and geographical analysis of the SPIRIT collection presented.

Early methods of Geospatial Information System (GIS) allowed only the experienced user to access the information presented in it. However, it is essential to access the information even to final end users by content. Obviously, the final end user may not be expert in the GIS system, but he may be expert in some specific application area. Therefore, it is necessary to GIS should be designed to manage both the geographic and structured data. Furthermore, GIS must be able to provide access points for geographical information collection to the final users (non-expert user of the application). In order to establish this author of (Agosti *et al*., 1993) introduced architecture and design approach for Geographical Information Retrieval System, which is capable of supporting the retrieval by content and browsing on textual data. It provided the framework for managing the geographical systems.

Larson (1996) author's pinpoints the problems that were relevant to the retrieval and browsing of spatial information. Larson (1996) examine the prospects of retrieval and indexing methods that were highly suited for the digitized materials with geographic details or their associations. Authors discussed the difficulties of information retrieval that are based on the geographic features. As well as methods and requirements for automatic information retrieval were studied. In addition to that authors also discussed the general issues and distinctiveness of the geo-referenced multimedia, information retrieval system was discussed.

Cai (2002) a model was developed depending on the coordinate-based geographic indexing and keyword-based vector model for denoting spatial information were proposed. As a consequence, (Martins *et al*., 2005) discussed the possible indexing structures. GeoIRIS was developed in (Shvu *et al*., 2007) for automatic feature extraction and high dimensional database indexing for retrieval of relevant geospatial information for a complex query. For fast retrieval of relevant information (Purves *et al*., 2007) proposed SPIRIT method, which used an indexed structure, which identified and assigned a footprint for the geographic references in a document and stored along with document terms.

Shvu *et al*. (2007) and Barb and Shyu (2010) and used semantic models and concept-based methods respectively to link low-level image features along with the high-level visual descriptors. Barb and Shyu (2010) a set of association rules were generated to correlate semantic terms with visual patterns and a mathematical model was used for relevant feature measurement. Graph-based approach was employed in (Purves *et al*., 2007) for mining geospatial data. Error-tolerant graph matching (Gautama *et al*., 2007) was applied to discover a relationship among detected image feature and geospatial vector data.

Galileo was designed in (Malensek *et al*., 2012), which took the data stream based on geospatial and chronological distinctiveness of time-series for efficacious storage and relevant retrieval of geospatial information. Evaluation of Galileo on a benchmark showed that it supported high-throughput storage and effective retrieval from a large data set to a given complicated queries.

For better retrieval of spatial data, the geographic features that satisfy the queries were classified. The classification of the geographic features was carried out by different authors addressed in distinct papers. In (Kavouras and Kokla, 2002) concept lattice, a mathematical approach was implemented for classification of different geographic features and the relationship was managed. It considered the semantic heterogeneity to achieve semantic interoperability. Conceptual integration of cognitive science was applied by the authors of (Kuhn, 2002) for geographic categorization. A new way for classification of geographic features was carried out in (Huang, 2011) through latent semantic analysis and domain knowledge regarding the specified information desires of the user.

The extracted features were taken and processed with documents to retrieve the relevant documents. Relevance was the central theme in information science, geographical information retrieval extends the classical information retrieval technique where relevance is the more challenging one (Cai, 2011). The more related documents were arranged through ranking. Several ranking methods were projected by different scholars. Profound technique was found in (Beard and Sharma, 1997) employed multidimensional ranking method for finding the most significant document that was available for a given query. Time, space and theme were the dimensions considered in (Beard and Sharma, 1997; Cai, 2002). Followed the work in (Beard and Sharma, 1997), used the relevance degree in both spatial and thematic domains for ranking. A logistic regression from the sample of text collection for ranking was envisioned in (Larson and Frontiera, 2004). A survey about the ranking methodologies was given in (Martins *et al*., 2005). Yu and Cai (2007) suggested a new way for ranking dynamically by combining the thematic and geographic relevance measures on per-query basis. This method determined weights of different documents of ranking substantiation for each query. A metric was proposed in (Meeks and Dasgupta, 2004), which was applied to find the degree of utility of accessed data of geospatial. Multi-attribute utility theory was used to find the information that was discovered in distributed scores.

GIR systems used one of the following methods in order to improve the selection process (1) query expansion (2) Filtering of relevant documents. An article in (Garcia-Cumbreras *et al*., 2009) evaluates the effectiveness of filtering the relevant documents depending on the user query. To measure the effectiveness of this technique Cross Language Evaluation Forum (CLEF) framework was used. The experimental results presented by the authors in the study (Garcia-Cumbreras *et al*., 2009) represents that filtering the relevant document worked significantly in GIR environment since, it the relevant documents were not recorded on the final list. Authors of (Perea-Ortega *et al*., 2007) described a GIR system named GEOUJA. Main objective of the article in (Perea-Ortega *et al*., 2007) is to filter the documents retrieved from an Information Retrieval (IR) system. Analysis from the result showed, increasing the number of documents retrieved by the IR subsystem also improves the final result.

A new model for the geographic information retrieval was proposed in (Bordogna *et al*., 2012) and implemented the system to represent uncertainty in indexing the geographic contents and the user's perspective and preferences during the manipulating the spatial quires. To denote the geographic content authors used fuzzy footprints. Fuzzy footprints were the distinct locations on the earth along with text. They also evaluated the system for two different kinds of user quires through combining the content-based condition with spatial condition, which is interpreted as the closeness between the users's perceived distance among the query and document footprints. Relevance scores were computed for the documents that were retrieved depending on the query conditions which were combined with to create on the whole ranked list document. Authors, allow the users to choose either the asymmetric or compensative aggregation to define the linear combination of the two conditions in order to specify the relative preference between the two conditions, which is used to achieve personalization and effectiveness. Geo-finder a geospatial information retrieval system was described, which was dependent on this model. Moreover, their performance was evaluated and analyzed.

3. PROPOSED METHODOLOGY

The proposed method for geospatial information retrieval from a large data set is detailed. As shown in Fig. 2, the overall process of our proposed approach comprises of four main phases: feature selection, document preprocessing, DPDC approach and ranking.

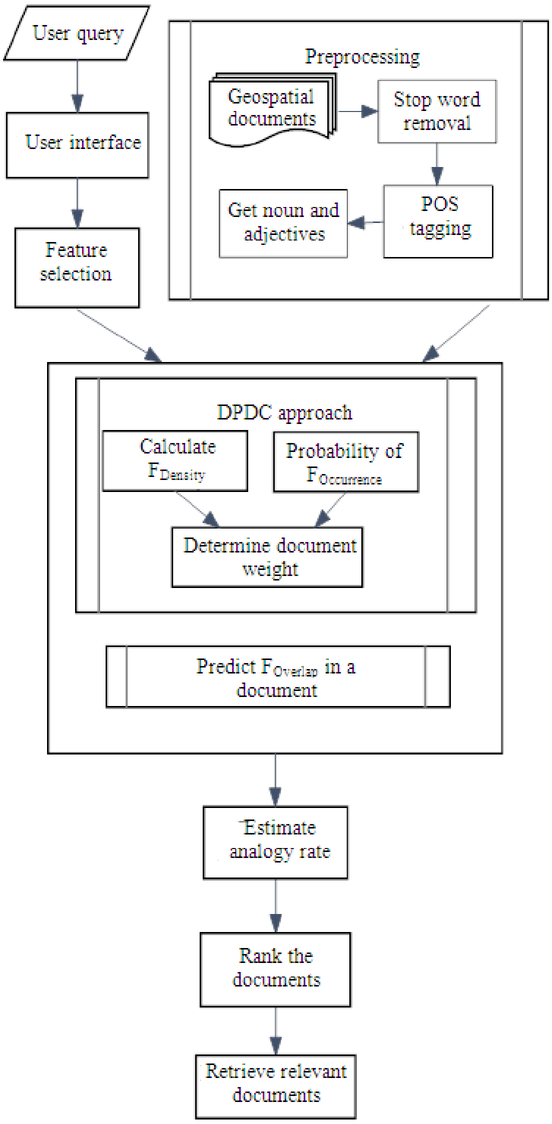

Fig. 2. Overall flow of proposed method

Each phase is described as below Algorithm 1 represents the procedure to determine the most relevant documents.

3.1. Feature Selection

Users express their need in the form of query. The query contains the keyword, which represents the required geospatial information. They interact with the user interface of the geospatial information retrieval system to issue the query. On receiving the query the system chooses a set of features that are exactly related to the keywords contained in the user query. These features are trained in prior and they are the default for particular keywords. Feature selection is an important step in the information retrieval of spatial information since depending on this the related and more accurate documents are given as output to the user. Feature selection in prior helps improve the comprehensibility of the results of the system.

For example, for a keyword, Kanchenjunga in the user query, we train the feature set as third highest mountain, five peaks, repositories of God, gold Ganga, offer greetings, Eastern Himalayas, Romantic Mountain. Similarly, for the keyword Himalaya the selected features are snow abode, world's highest mountain chain, awe-inspiring power, magnificent mountain, massive mountain, worlds highest, world's highest peaks, permanent ice, extreme cold, fold mountain, Indo-Australian plate, fresh water, large perennial rivers, Indus Basin, Ganges-Brahmaputra Basin.

Likewise, we collect features for all the possible keywords that are particular about the information retrieval system. Feature selection method speed up the retrieval approach. These selected features are used in the following phases of the geospatial information retrieval method.

3.2. Document Preprocessing

The task of this phase is to collect the documents and to make them flexible for retrieval process. The goal of document preprocessing is to represent the documents in terms of both space and time. It is a complex process. Preprocessing of document takes raw document as input and makes the document more apt for the information retrieval system.

This phase begins with the deletion of stop words from the documents that are collected for the given query from the repository. Removal of stop words will decrease the length of the document size effectively which minimizes the time required for retrieval process. We use a domain dependent stop word removal because the spatial information requires to be processed with plenty of domain knowledge. Therefore, applying domain-independent stop word removals are not applicable in spatial documents. For better results, it advised to use both domains dependent and independent removals. POS tagging is applied after removing the stop words. This process assigns lexical class makers such as verb, noun, to each word in the sentence. As a result of the stop word removal and POS tagging the documents contain the maximum of nouns and adjectives.

3.3. DPDC Approach

Preprocessed document along with the user query and the selected features are given as input to this phase. Here, we calculate density of features that are selected for a given keyword in user query and probability of a feature that can appear in a given document. Depending on the calculated values, our DPDC approach determines the document correlation. Following subsection deals with the computation of all above-mentioned values.

3.3.1.Calculation of Feature Density

Keyword density provides the percentage of the number of times a keyword appears in a document compared to the total number of words in the document. Keyword density is one among the factors in determining whether a document is relevant to a specified geospatial keyword in a user query. Mathematically, it can be represented as in equation 1, denotes that the frequency of appearance of a particular keyword in a document (dissertation):

$$F_{\text{Density}} = \left(\frac{N_{\text{Fr}}}{\text{Total}_{\text{wordDOC}}} + \frac{T_{\text{DOC}}}{N_{\text{DOCFr}}} \right) \times 100 \qquad (1)$$

In equation 1, F_{Density} denotes the keyword density, N_{Fr} is the frequency of a feature and $\text{Total}_{\text{wordDOC}}$ represents the total number of words in the documents. T_{DOC} describes the total number of documents in the corpus and N_{DOCFr} personifies the number of documents having the given feature.

3.5. Estimate Probability of Feature Occurrence

Another attribute that enhances the retrieval of information is the probability of feature occurrence. It is calculated using the equation 2:

$$P_{\text{Focc}} = \frac{N_{\text{Fr}}}{\text{Total}_{\text{wordDOC}}} \qquad (2)$$

Number of times a feature among the list of features selected as in section, 3.1 appeared in the document is symbolized as N_{Fr} in equation 2. Likewise, $\text{Total}_{\text{worddoc}}$ delineate the total number of words in the document.

3.6. Estimate Document Weight

Document weight is determined from the equation 3. Document weight value depends on the values of equation 1 and 2. It is computed for each document in the corpus individually. This value is highly required for computing the score of a corresponding document:

$$\text{DOC}_{\text{weight}} = \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{f} (F_{\text{Density}} + P_{\text{Focc}}) \qquad (3)$$

where, $\text{DOC}_{\text{weight}}$ in equation 3 denotes the document weight, n and f represents the total number of documents in the corpus and total number of features selected for a given query. Further process of relevant information retrieval depends on document weight so, it should be calculated exactly. Otherwise, irrelevant documents may be retrieved.

3.4. Ranking

This ranking phase retrieves the most related documents that satisfy the user requirement. Ranking specifies the importance of the document. The document retrieval can be carried out as below.

3.4.1. Predict Feature Overlap

The Feature overlap as in (Joho and Sanderson, 2004) reveals the number of features among the total number of features selected appears on a document. Spatial data set of features may appear in any one of the ways as shown in **Fig. 3**.

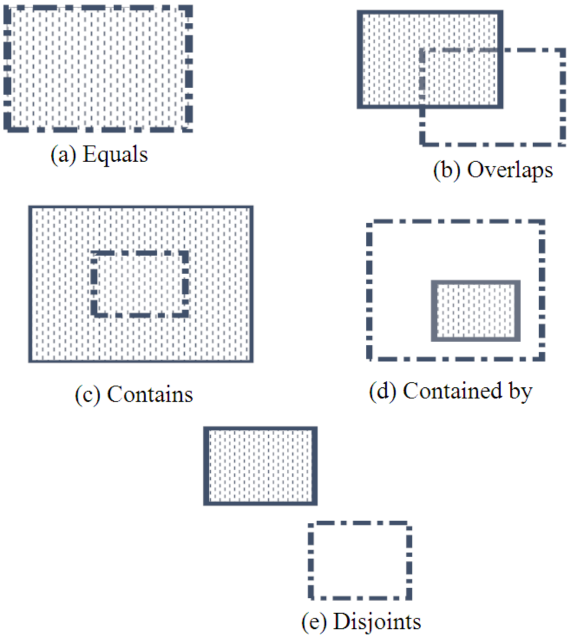
(a) Equals
(b) Overlaps
(c) Contains
(d) Contained by
(e) Disjoints

Fig. 3. Relationship between features and documents

The overlapping values can be determined through as follows:

If all the features entirely present in a document, then it represents the figure in (a) of 3, which carries the feature overlap ($F_{overlap}$) value as 1. In case if only certain features are present in the given documents as in (b) of 3, then they carry the value of 0.25 for $F_{overlap}$. The figure (c) of 3 denotes the features that are contained in a document, (d) of 3 represents the features that are contained by the documents. The (c) and (d) takes $F_{overlap}$ values as 0.75 and 0.5 respectively. Disjoint in (e) of 3, skhes that none of the features are in the given documents. Therefore, its $F_{overlap}$ is 0. Disjoint nature of a document shows that it is irrelevant to the query and also illustrates the corresponding document will not satisfy the user need.

3.4.2. Determine Document Score

The values predicted in the above step and the document weights are used to predict the score of the document. Document score shows the degree of relevance of a document for a given user query. Document score can be calculated as shown in equation 4:

$$DOC_{score} = DOC_{weight} + F_{overlap} \tag{4}$$

Each document in the corpus is evaluated and the score is calculated. In the above equation DOC_{score}, specifies the score for a document, DOC_{weight} and $F_{overlap}$ are the document weights calculated through the equation 2 and overlap of features in a given document respectively.

3.4.3. Rank and Retrieve the Documents

Documents are ranked in this phase using the document score obtained through equation 4. The documents in the corpus are ordered in the increasing order of the score it has. Then, they are ranked in ascending order. Therefore, the documents that are having the highest rank are considered the irrelevant and the documents with lowest rank are treated as the most relevant one. For example:

```
Algorithm: Relevant Document Retrieval
Input: User Query,
Output: Relevant documents
begin
        Get user query as input
        Choose feature set as related to user query (trained earlier)
Document Preprocessing
        Domain dependent removal of stop-word
        Assign lexical class makers to each word in the sentence (POS tagging)
DPDC Approach
        Estimate the percentage of number of times a keyword appears in a document
          $F_{Density} = \left(\frac{N_{Fr}}{Total_{wordDOC}} + \frac{T_{DOC}}{N_{DOCFr}}\right) \times 100$
        Calculate probability of feature occurrence
          $P_{POCC} = \frac{N_{Fr}}{Total_{wordDOC}}$
        Compute Document Weight
          $DOC_{weight} = \sum_{i=1}^{n} \sum_{j=1}^{n} (F_{Density} + P_{Pocc})$
        Ranking Predict Feature overlap
        if Equals Then
                $F_{overlap} = 1$
        end if
        if overlaps then
                $F_{overlap} = 0.25$
        end if
        if contains Then
                $F_{overlap} = 0.75$
        end if
        if contained by Then
                $F_{overlap} = 0.5$
        end if
        if disjoints Then
                $F_{overlap} = 0$
        end if
        Determine Document Score
                $DOC_{score} = DOC_{weight} + F_{overlap}$
        Retrieve the most pertinent documents
end
```

If a set of documents such as d1, d2, d3 and d4 has the score as 10,8,15 and 3 respectively. They are ranked as d1 = 2, d2 = 3, d3 = 1 and d4 = 4 and shows that d3 is the most pertinent document and d4 having the highest rank is not as much as relevant to the user query.

4. EXPERIMENTAL EVALUATION

Retrieval of pertinent document through our proposed approach is evaluated by an experiment. The experiment was conducted to retrieve the documents that are relevant to the mountains and river. In this study, we focus particularly on the rivers and mountains of India.

Initial work of spatial data information retrieval of our proposed method carried out through the feature set collection. For selecting features about the Indian mountains and rivers we referred the website http://www.ecoindia.com, a Hephaestus book named "Articles on Mountains of India" and various other articles. These resources are used to gather enormous domain knowledge for feature selection. The features of mountains and rivers are selected from the topics location, distinct characteristics, ecosystem, elevation and dimension. Instead of choosing the entire topics mentioned above we use only the two topics namely location and distinct characteristic of the respective mountains and rivers. For effective analysis and selection of features we examine only the location and characteristics.

Then in order to minimize the time and memory requirement we preprocess the documents present in the corpus. Preprocessing of the documents reduces the length of the documents, which in turn the time and memory requirement of the proposed approach for processing and finding retrieval of information. Result of preprocessing generates only the nouns and adjectives of a document. The nouns and adjective of the document expresses the location and characteristics of the rivers and mountains. Therefore, the best related documents are retrieved effectively from analyzing the preprocessed documents using the features.

Table 1. Document score

Documents	DOC_{score}	Documents	DOC_{score}
Doc1	0.000000	Doc11	0.000000
Doc2	0.000000	Doc12	0.000000
Doc3	0.000000	Doc13	0.000000
Doc4	1.167562	Doc14	0.000000
Doc5	0.380277	Doc15	0.497623
Doc6	0.854950	Doc16	0.000000
Doc7	0.937777	Doc17	0.000000
Doc8	0.867270	Doc18	0.000000
Doc9	0.653831	Doc19	0.000000
Doc10	0.000000	Doc20	1.283553

Table 2. Document ranking

Documents	DOC_{score}	Rank	Documents	DOC_{score}	Rank
Doc20	1.283553	1	Doc11	0	11
Doc4	1.167562	2	Doc12	0	12
Doc7	0.937777	3	Doc13	0	13
Doc8	0.867270	4	Doc14	0	14
Doc6	0.854950	5	Doc16	0	15
Doc9	0.653831	6	Doc17	0	16
Doc15	0.497623	7	Doc18	0	17
Doc5	0.380277	8	Doc19	0	18
Doc1	0.000000	9	Doc2	0	19
Doc10	0.000000	10	Doc3	0	20

The preprocessed documents are processed to find the document weight using the DPDC approach and they are scored. The scoring values depend on the relevancy. During our experimental analysis, for a query "Mount Everest" the corpus containing 20 documents is processed as below. Our approach selects "Mount Everest debacle, Nepalese side, Earth's highest mountain, Chomolungma and British began" these set of features as the feature list and searches the corpus for the documents that carries the features selected.

The DPDC approach calculates document weight and the overlap values are determined. Depending on the values calculated by DPDC approach we determine the document score. **Table 1** represents the scores of the documents in a corpus.

Depending on the score the documents are arranged in descending order and ranked as shown in **Table 2**.

Documents that have the value higher scoring are ranked first and that are considered the most related documents for the query "Mount Everest". Documents whose values are zero are discarded and not displayed to the user. Therefore, for this query we retrieve eight documents as the pertinent with m20.txt as the most related documents among 20 documents in the corpus.

The effectiveness of the geospatial retrieval results are evaluated based on precision, recall, prediction accuracy, time required for the proposed approach for processing and retrieving documents. Following figures show comparison between our proposed and existing method. The existing method uses only keyword, whereas our proposed method use features which are getting substituted for the keywords given as a query for retrieving the relevant documents.

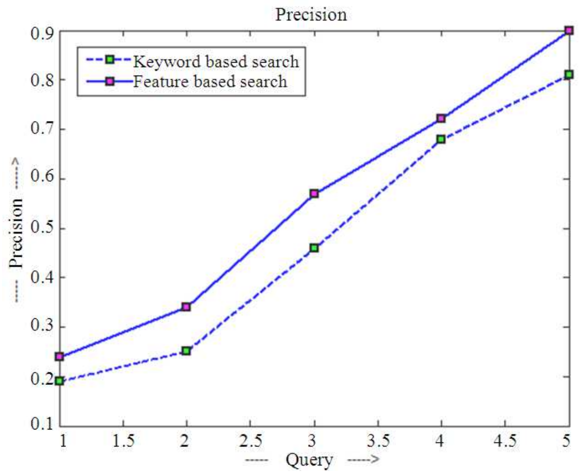

Fig. 4. Precision study for existing and proposed

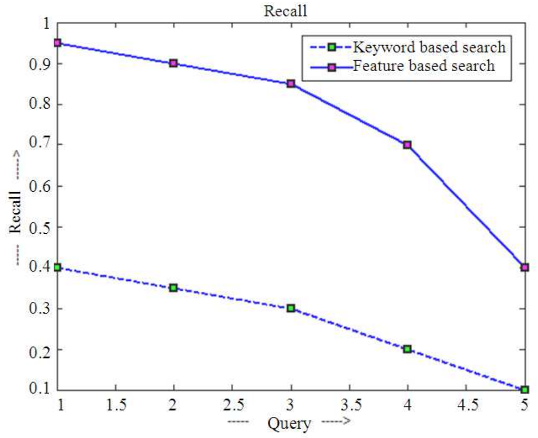

Fig. 5. Recall study for existing and proposed

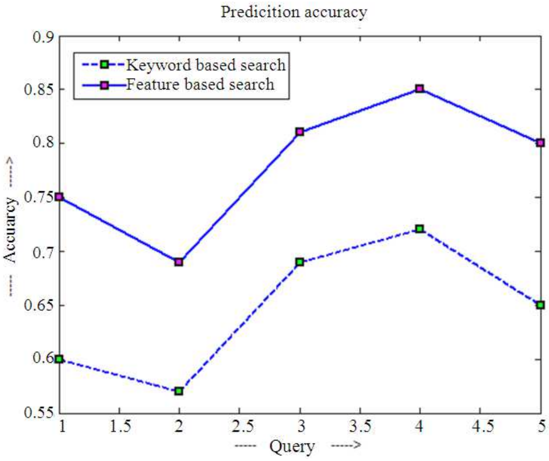

Fig. 6. Comparison of prediction accuracy existing Vs proposed

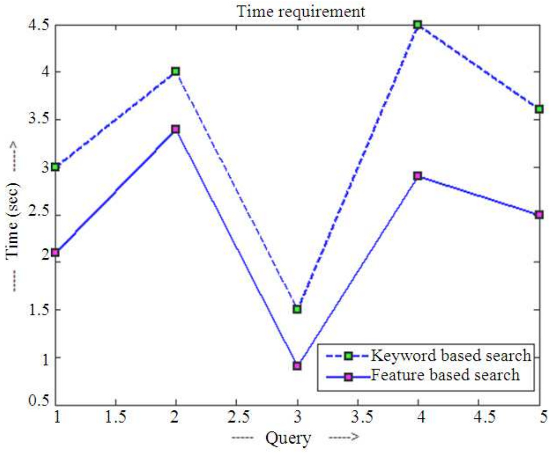

Fig. 7. Time required for retrieval

Precision is a measure that expresses the fraction of returned (retrieved) documents that are relevant, which is purely based on the measure and understanding of relevance. Precision can be calculated using the Equation 5:

$$\text{Precision} = \frac{P_{Doc} \cap R_{Doc}}{R_{Doc}} \quad (5)$$

In the above equation P_{Doc}, R_{Doc} interpreted as relevant document and retrieved documents respectively. **Fig. 4** illustrates that our proposed method has higher precision value than the existing one for a geospatial query.

For analysis, five different queries are given as input to both the systems and estimate the precision values for all five quires.

It represents that retrieved result of the proposed was more pertinent than the existing work.

Similarly the existing and proposed approaches are compared using Recall another measuring factor. The recall value can be estimated through the equation 6:

$$\text{Recall} = \frac{P_{Doc} \cap R_{Doc}}{P_{Doc}} \quad (6)$$

Fig. 5 reveals the recall values for proposed and existing techniques.

An efficiency retrieval technique's accuracy depends on how accurately it retrieves the documents for a given query. In order to predict the proposed methods efficiency we calculate its overall prediction accuracy. Five quires are taken and for each query the accuracy is carried out using equation 7. Proposed and existing methods are experimented with same queries and the results are expressed in **Fig. 6**:

$$\text{Prediction}_{ACC} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FF} + N_{FN} + N_{TN}} \quad (7)$$

The values N_{TP}, N_{PP}, N_{PN}, N_{TN} are the number of true positive, false positive, false negative and true negative.

Fig. 6 illustrates that the Prediction accuracy of our approach, which is evident that the proposed is superior to the keyword based searching algorithm. Experimental result shows that our approach retrieves 78% of the pertinent documents accurately, whereas the existing method retrieves 64.6% of related documents.

Running time of the approach plays a vital role in the evaluation of the effectiveness of the retrieval system. We measured running time for both the proposed and existing methods for the same set of queries executed at different time. **Fig. 7** portrays the time required for both proposed and existing techniques for retrieving spatial information.

Fig. 7 reveals that the proposed method consumes less time than the existing method. This explicitly denotes that proposed method retrieves the related documents faster than the existing method. Computation time is calculated in seconds.

The experimental results emphasize that our proposed method retrieves the relevant spatial documents with an accuracy of 78%, which is 13.4% higher than the existing method. Also, our proposed method retrieves the document 9.6 sec faster than the existing technique as an average. Therefore, the result pictured in **Fig. 4-7** implicitly express that our proposed method retrieves the related documents efficiently and also effectively than the existing method.

5. CONCLUSION

Due to the complex nature of spatial data type and the correlation relationship exists among the spatial data; information retrieval of spatial data becomes laborious. We proposed a method based on DPDC approach for effective retrieval. For efficient retrieval we collect the feature set for a given query based on which the documents are computed to find the relevancy. Documents are preprocessed in order to reduce the length of the documents presented in the document. DPDC approach in our method is used to calculate the weight of all documents, which is computed, based on the features collected for keywords and the documents that contain the features. Quantity of features overlaps on the document is found along with its value, document weight is used to estimate the document score. Based upon the score the documents are ranked in increasing (descending) order. Documents with the high score are considered the most pertinent document that satisfies the user requirement.

The experimental results show that our proposed method predicts the relevant document with an accuracy of 78% and also reveals that it consumes less time than the existing method. Our proposed algorithm outperforms the existing method. To enhance the proposed approach, we will carry out the retrieval of document for all types of spatial data along with the retrieval of images related to the given query that satisfies the user need.

REFERENCES

Agosti, M., F. Crivellari, G. Deambrosis and G. Gradenigo, 1993. An architecture and design approach for a geographical information retrieval system to support retrieval by content and browsing. Comput. Environ. Urban Syst., 17: 321-335. DOI: 10.1016/0198-9715(93)90029-5

Barb, A.S. and C.R. Shyu, 2010. Visual-semantic modeling in content-based geospatial information retrieval using associative mining techniques. IEEE Geosci. Remote Sens. Lett., 7: 38-42. DOI: 10.1109/LGRS.2009.2017214

Beard, K. and V. Sharma, 1997. Multidimensional ranking for data in digital spatial libraries. Int. J. Digital Libraries, 1: 153-160. DOI: 10.1007/s007990050011

Bordogna, G., G. Ghisalberti and G. Psaila, 2012. Geographic information retrieval: Modeling uncertainty of user's context. Fuzzy Sets Syst., 196: 105-124. DOI: 10.1016/j.fss.2011.04.005

Cai, G., 2011. Relevance ranking in geographical information retrieval. SIGSPATIAL Special, 3: 33-36. DOI: 10.1145/2047296.2047304

Cai, G., 2002. GeoVSM: An integrated retrieval model for geographic information. Proceedings of the Second International Conference on Geographic Information Science, (GIS' 02), ACM Press, Springer-Verlag London, UK., pp: 65-79.

Clough, P., 2005. Extracting metadata for spatially-aware information retrieval on the internet. Proceedings of the 2005 Workshop on Geographic Information Retrieval, Oct. 31-Nov. 05, ACM Press, Bremen, Germany, pp: 25-30. DOI: 10.1145/1096985.1096992

Garcia-Cumbreras, M.A., J.M. Perea-Ortega, M. Garcia-Vega and L.A. Urena-Lopez, 2009. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. Inform. Process. Manage., 45: 605-614. DOI: 10.1016/j.ipm.2009.04.006

Gautama, S., R. Bellens, G.D. Tre and W. Philips, 2007. Relevance criteria for spatial information retrieval using error-tolerant graph matching. IEEE Trans. Geosci. Remote Sens., 45: 810-817. DOI: 10.1109/TGRS.2007.892006

Huang, Y., 2011. A latent semantic analysis-based approach to geographic feature categorization from text. Proceedings of the 5th IEEE International Conference on Semantic Computing, Sept. 18-21, IEEE Xplore Press, Palo Alto, CA., pp: 87-94. DOI: 10.1109/ICSC.2011.15

Joho, H. and M. Sanderson, 2004. The SPIRIT collection: An overview of a large web collection. ACM SIGIR Forum, 38: 57-61. DOI: 10.1145/1041394.1041395

Kavouras, M. and M. Kokla, 2002. A method for the formalization and integration of geographical categorizations. Int. J. Geographical Inform. Sci., 6: 439-453. DOI: 10.1080/13658810210129120

Kuhn, W., 2002. Modeling the semantics of geographic categories through conceptual integration. Geographic Inform. Sci., 2478: 108-118. DOI: 10.1007/3-540-45799-2_8

Larson, R.R. and P. Frontiera, 2004. Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. Res. Adv. Technol. Digital Libraries, 3232: 45-56. DOI: 10.1007/978-3-540-30230-8_5

Larson, R.R., 1996. Geographic Information Retrieval and Spatial Browsing. In: Geographic Information Systems and Libraries: Patrons, Maps and Spatial Information, Smith, L.C. and M. Gluck (Eds.), Clinic on Library Applications of Data Processing, pp: 81-124.

Malensek, M., S.L. Pallickara and S. Pallickara, 2012. Exploiting geospatial and chronological characteristics in data streams to enable efficient storage and retrievals. Future Generat. Comput. Syst., 29: 1049-1061. DOI: 10.1016/j.future.2012.05.024

Martins, B., M.J. Silva and L. Andrade, 2005. Indexing and ranking in Geo-IR systems. Proceedings of the workshop on Geographic Information Retrieval, Oct. 31-Nov. 05, ACM Press, Bremen, Germany, pp: 31-34. DOI: 10.1145/1096985.1096993

Meeks, W.L. and S. Dasgupta, 2004. Geospatial information utility: An estimation of the relevance of geospatial information to users. Decision Support Syst., 38: 47-63. DOI: 10.1016/S0167-9236(03)00076-9

Perea-Ortega, J.M., M.A.G. Cumbreras, M.G. Vega and L.A.U. Lpez, 2007. Filtering for improving the geographic information search. Adv. Multilingual Multimodal Inform. Retrieval, 5152: 823-829. DOI: 10.1007/978-3-540-85760-0_104

Purves, R.S., P. Clough, C.B. Jones, A. Arampatzis and B. Bucher *et al*., 2007. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the internet. Int. J. Geographical Inform. Sci., 21: 717-745. DOI: 10.1080/13658810601169840

Shvu, C.R., M. Klaric, G.J. Scott, A.S. Barb and C.H. Davis *et al*., 2007. GeoIRIS: Geospatial information retrieval and indexing system-content mining, semantics modeling and complex queries. IEEE Trans. Geosci. Remote Sens., 45: 839-852. DOI: 10.1109/TGRS.2006.890579

Yu, B. and G. Cai, 2007. A query-aware document ranking method for geographic information retrieval. Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, Nov. 6-10, ACM Press, Lisbon, Portugal, pp: 49-54. DOI: 10.1145/1316948.1316962