IJCTA, 9(10), 2016, pp. 4501-4508

Novel Approach to Imrove the Performance of Information Retrieval Using Collocation Rules in Spatial Databases

***Mr. Ramesh Babu Pittala **Dr. M.Nagabhushana Rao ***Mr. P.Srinivas**

Abstract : In this paper, we proposed a novel approach to improve the performance of the spatial data by retrieving the data in the form of the Map (GIS). The user can identify the data quickly instead representing the data in the form to the text; it can retrieve in the geographical data. Proposed Geo-spatial Information Retrieval will improve the relevant retrieved results and reduces the time to collect the required information using Collocation Rules in Spatial Databases.We calculated Precision, Recall and F-Measure to estimate the performance of the spatial and non-spatial data sets and proposed a spatial factor to improve the relevant retrieved value to improve the overall performance of the system.

Keywords : Spatial Data Mining, Information Retrieval, Evaluation, Performance evaluation.

1. INTRODUCTION

Data is the collection of raw facts, text, numbers, digits, etc., which are processed by the computer and are of various formats like operational data such as sales payrolls, sales, inventory, cost and non-operational data like forecast data and meta- data, *i.e*. data about data. It is the process of extracting data from the large data sets. It is the way of analyzation of the data from distinct software systems and with distinct pattern recognition techniques. It also provides how to utilize data and classify data and finally to formulate different scenarios of the data.

Data mining consists of a different mining process like data cleansing, transformation, reduction, and discrimination. Data mining tools anticipate behavior, scope of the data and traditional business trades. In this, data can collect truly enormously and can be stored with high speed [4][6]. The best example for data mining is our mobile phone number's directory. Distinct machine learning algorithms, different pattern recognition algorithm and various statistical operations will provide very high dimensionality of the data. As per as mining is concerned, we have distinctive prediction procedures like by using variables to anticipate the future values of variables. We can collect the unusual type of data like simple measurement text data and graphic data and even spatial data and multimedia data with different type of structures.

A. Spatial Data

Spatial data or Geospatial data is a collection of information from the objects within the space. It is typically in the form of points, lines, Pixels. It is regularly accessed, controlled or examined through Geographic Information Systems. Spatial Data is stored in different data patterns like table, image, graph, etc. This database is a collection of attributes related to the spatial objects represented by a spatial data type and associations among those objects. It optimized to store and queries the data that is about the characteristics of the objects in a geometric space [7].

* Assoc.Professor, HOD CSE Department, Trinity College of Engineering and Technology, Karimnagar Telangana

** Professor, IT Department, SRK Institute of Technology, Vijayawada, Andhra Pradesh, India

*** Assoc.Professor, CSE Department, Trinity College of Engineering and Technology, Karimnagar Telangana prameshbabu526@gmail.com, mnraosir@gmail.com, sri23131@gmail.com

B. Information Retrieval

Text Mining is playing a major role to provide the require information to the user based upon their needs. It is the best methodology to provide the information to the user in simple manner. Only text data type is used to retrieve the relevant data. All the industries are automated their data to provide the information within the organization. To provide such accurate data, Information Retrieval system is used. It is a system used to store, manipulate and retrieve the information to fulfill the requirements from the user. Evaluation of IRS's is very important to provide the relevant information quickly. IRS may retrieve relevant or Non-Relevant, may not retrieve relevant or Non-Relevant information from the database [10].

C. Geographic Information System:

Interoperability inside GIS is troubled with both information and operations. It is very essential that, researchers should be made on, more particularly, on the access to be different, distributed, heterogeneous and self-ruling data sources and on their integration. The extraordinary prerequisites of these applications are dealing with geo-spatial information.There are five components in GIS: Hardware, Software, Data, People, and Methods [9].

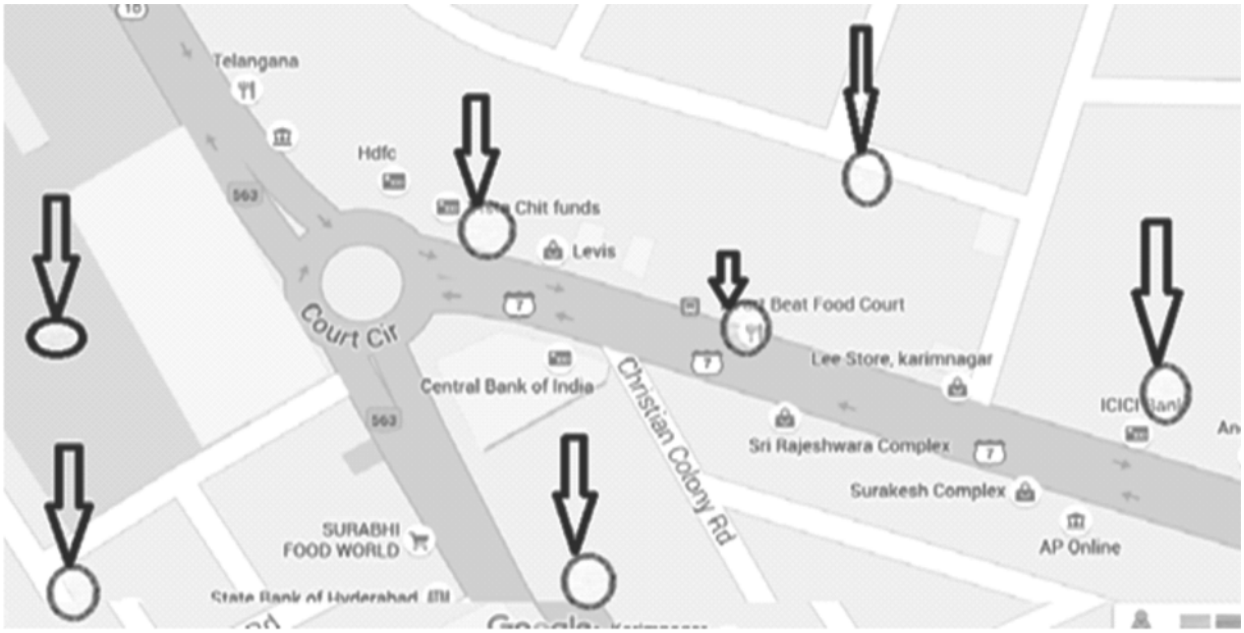

Fig. 1. Representation of Spatial Data in GIS.

D. Evaluation

Assessment is an extremely essential and tedious task in IR framework. There are numerous IR models, algorithms and frameworks in writing so keeping in mind the end goal to declare the best among numerous one, pick it to utilize and improve there are to assess them. The difficult of measuring viability is that it is connected with the relevancy of the retrieved items. This makes relevance the establishment on which IR assessment stands[2].

2. METHODOLOGY

Relevance is the relationship between the items, documents and query of the input. Relevancy is from the System Perspective and User Perspective.

A. System Perspective

There are many algorithms are available to fine the relevance of the document/item to find the required data. It is classified and evaluated into either ranked or unranked retrieval results. Classification of the evaluation techniques are shown below.

B. User Perspective

Relevancy from the human point of view is

1. **Individual :** It depends on the specific user requirements.
2. **Conditional :** It Relates to the current user needs
3. **Cognitive :** It depends on human pinion
4. **Dynamic :** Change over time.

With these issues, it is extremely hard to evaluate the performance of the system because it requires many resources [5].

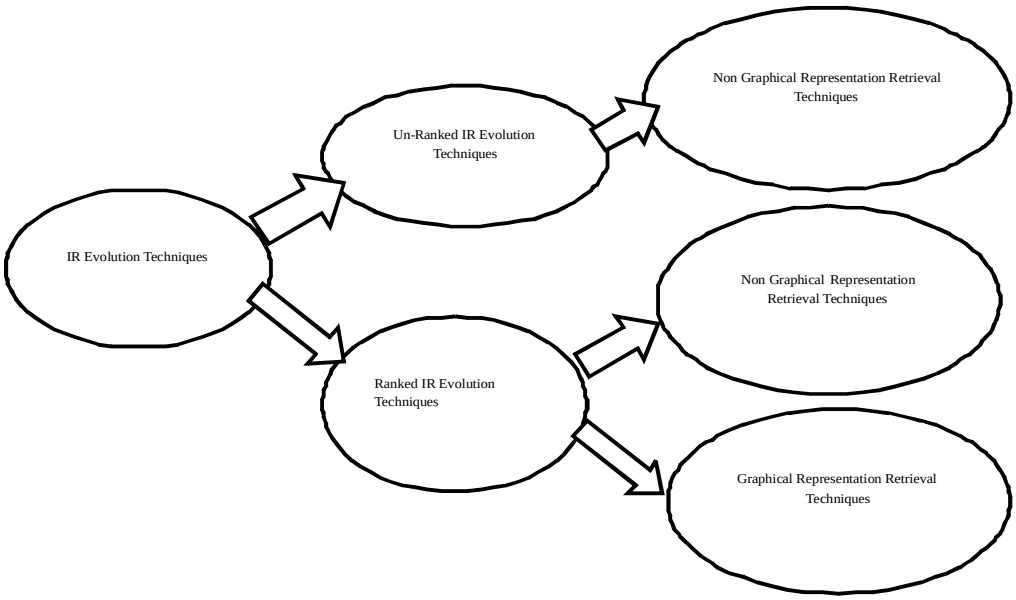

Fig. 2. IR Evolution Techniques.

3. RESEARCH SIGNIFICANCE, LITERATURE REVIEW

In this paper we compared the performance of the Non Graphical Representation with Graphical Representation IR Techniques. Graphical Representation is in the form of the Map's. Two important factor's ***Precision*** and ***Recall [1] [7]*** will be used to estimates the ranking of the relevance items.

A. Recall

Recall is the proportion of the number of retrieved records to the aggregate number of significant records in the database.

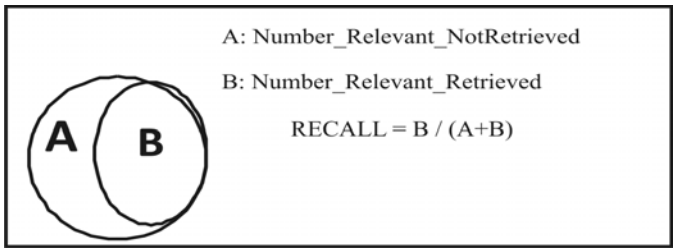

Fig. 3. Set diagram for Recall .

B. Precision

Precision is the fraction of number_relevant_retrieved to the total _ Retrieved (relevant and irrelevant).

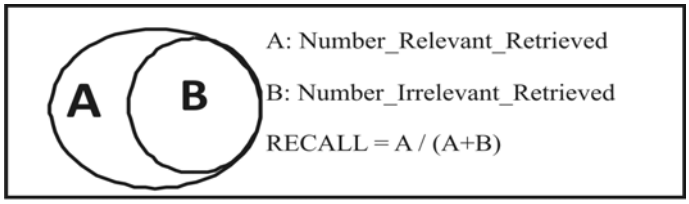

Fig. 4. Set diagram for Precision.

Example : Assume that the database contains 90 records in a specific topic. When the user searched for a topic, he retrieved 70 records. Out of the 70 Records, 45 records are relevant.

The precision and recall values are

$$\text{Precision} = 45 / (45+25) = 45/70 = 0.64\ (64\%)$$
$$\text{Recall} = 45 / (45+45) = 45/90 = 0.50(50\%)$$

C. F-Measure

It computes the average of the information retrieval precision and recall values. It is computed using harmonic mean.

Given "M" points harmonic mean $a_1, a_2, a_3 \ldots a_m$ is

$$H = M^* \sum_{K=0}^{M} X_k \tag{1}$$

So, the harmonic mean of Precision and Recall

$$\text{F-Measure} = 2^* \left(\frac{\text{Precision(P)}^* \text{ Recall (R)}}{\text{Precision(P)} + \text{Recall (R)}} \right) \tag{2}$$

As per the F-measure derived by van Rijsbergen (1979), Fâ "It will find the efficiency of system with respect to the items entered by the user, who assigns â times as much significance to recall as precision".

To fine the effectiveness measure
$$\text{Eff} = 1 - \frac{1}{\frac{\Omega}{P} + \frac{1-\Omega}{R}} \tag{3}$$

Their relationship is
$$F\beta = 1 - E \tag{4}$$

Where
$$\Omega = \frac{1}{1+\beta 2} \tag{5}$$

For each cluster we will calculate the F-measure to estimate the performance.

4. APPLYING SPATIAL DATA MINING

Different Statistical methods are used to retrieve the relevant information. Mostly Baye's theorem is used find the relevant possibilities from the existing data. It defines the probability of an item, based on situations that might be related to the item.

$$P(x1/x2) = P(x2/x1)\ P(x1)/P(x2)$$

Here in the x1 and $x2$ are the terms of the spacial objects. Bayes rule finds the relevant possible results $x1$ of $x2$. For example, we want to know the ***probability*** of getting the dengue fever among the people who are affected with different diseases. The probability may be maximum of 1% can be determined by using this method.

The spatial knowledge comprises of the relevant attributes of the spatial data with their degree of intensity. The degree of intensity of the characteristics is identified as fuzzy values. The relevant information can be retrieved through the Collocation of the objects. The collocation is a set of co-located spatial items with regard to their common features.The collocation identified as pattern can describe the basic nature of the spread over spatial stretch in the domain [8,10].

In a spacial database (SB), let the futures {*a*1, *a*2, *a*3….*a*}. The instances of the characteristics will be stored as {*t*1, *t*2, *t*3*t*}. The colocation rule will optimize the data in every level ad can produce the desirable information.

Implementation of Collocation Rule :

Let us assume that '*y*' is a consequence future of '*x*', forms a first level of collocation is $x \rightarrow y$,(1) similarly, '*z*' is a consequence of '*y*, *y* ' $\rightarrow$ *z* (2) forms a collocation. As '*y*' already have an antecedent (1, 2) '*x*', the consolidated version of collocation, {*y*, *x*} $\rightarrow$ *z* can be formed. Attributes (Items) from the DB are collected and applied the collocation rules among the Items. In every level, ***Geo-spatial Information systems (GSIR)*** is applying a collocation rules, optimize the data and the user can retrieve the different types of information at each stage.

5. SPATIAL INFORMATION RETRIEVAL

In Geographic/ Spatial Information Retrieval, which is concerned with deterministic_retrieval (discovering all data sets that encloses information on a particular coordinates) and probabilistic retrieval (such as finding all towns near a major river)[3][11].

Geographic queries and spatial queries infer challenging a spatially indexed database in view of connections between specific items in that database inside a specific coordinate system. Geographical interactions in the coordinate systems enforced on the physical domain are geometric relationships. Within this system, where space and route can be restrained on a continuous scale, numerous types of relationships between things defined within that space can be determined using geometry. Spatial and geographic queries combine both geometric and topological components.

A. Types of spatial queries

The type of the IR Submitted by the user to retrieve the desire data (Ex. Digital Library) may be difficult in terms of the time, location, period and other queries. If we focus on only the spatial or geographic query, there are different types of input items are provided to the user in query [5].

We can consider five types of spatial queries:

- Point-in-polygon queries.
- Region queries.
- Distance and Buffer Zone queries.
- Path queries.
- Multimedia queries

In this paper, we used Region queries to retrieve the information quickly. In this method, the user can request the distinct types of the queries to retrieve the information from different areas in a Map to the Items. Through this, the user can find the information very easily and can complete the task.

6. ALGORITHM

The following algorithm is used to retrieve the information form the spatial database. User is having flexibility to write a query or to select the items listed for quick retrieval of the desired data.

1. Identify the Spatial Objects.

$$S = \{A1, A2, A3 \ldots\ldots\ldots\ldots An\}$$

2. Apply the spatial data mining and select the items to be Indexed.

$$I = \{I1, I2, I3 \ldots\ldots\ldots\ldots\ldots\ldots In\}$$

3. Apply the Collocation Rule

$$I1 \rightarrow I2$$
$$I1 \rightarrow I3$$
$$\{I1, I2) \rightarrow I3$$

4. Select the Items to know the Information in GSIR
   Level1 : I1, I4
   Level1 : I2, I6, I9 etc.
5. Retrieve the Information in the Form of a GIS.

7. FRAMEWORK

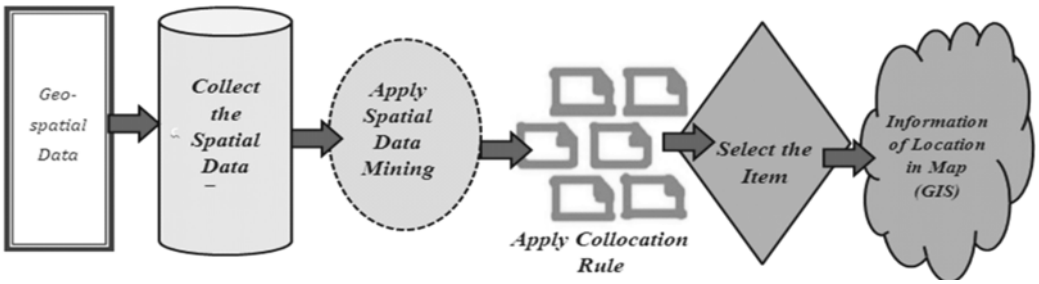

Fig. 5. Framework to retrieve spatial IR.

As shown in the below figure, we will collect the Geo-spatial data and apply the Data mining operations. After, apply the collocation rules to optimize the results on every level. The user can select any combination of the items from the MapServer, and in each stage, they can retrieve the relevant information on GIS.

8. PERFORMANCE EVALUATION

As stated earlier, we have taken the sample database to find the performance of the Spatial IR with the non-spatial DB and results are displayed for the both Databases. In this paper, we concentrated on improving the Relevant Retrieved among the total Retried Items, which will give the better results from the system and also user can retrieve it quickly through the Map [2].

As stated in (1)(2)(3)(4)(5), we have taken the sample database to find the performance of the Spatial IR with the non-spatial DB and results are displayed for the both Databases. In this paper, we concentrated on improving the Relevant Retrieved among the total Retried Items, which will give the better results from the system and also user can retrieve it quickly through the Map [2].

Comparison of the result are shown below by including a spatial factor *f* by including the collocation rule to the items and evaluated to estimate the processing time to retrieve the information from the spatial databases with the non-spatial data bases.

$$S_f = (1/f)*PRE \quad (6)$$

Table 1. Non-Spatial & Spatial Database Performance

	Non-Spatial Database						
Item	1	2	3	4	5	6	7
TR	105	86	70	42	32	20	4
RR	80	69	48	29	18	12	2
REL	53	47	32	16	11	9	1
PRE	0.66	0.68	0.67	0.55	0.61	0.75	0.50
Recall	0.50	0.55	0.46	0.38	0.34	0.45	0.25
F-M	0.57	0.61	0.54	0.45	0.44	0.56	0.33
Ω	0.50	0.20	0.10	0.06	0.04	0.03	0.02
E	0.43	0.43	0.53	0.61	0.65	0.55	0.75
β	0.57	0.57	0.47	0.39	0.35	0.45	0.25

Spatial Database							
Item	1	2	3	4	5	6	7
TR	105	86	70	42	32	20	4
RR	80	69	48	29	18	12	2
REL	61	49	37	19	13	10	1
PRE	0.76	0.71	0.77	0.66	0.72	0.83	0.50
Recall	0.58	0.57	0.53	0.45	0.41	0.50	0.25
F-M	0.66	0.63	0.63	0.54	0.52	0.63	0.33
Ω	0.50	0.20	0.10	0.06	0.04	0.03	0.02
E	0.34	0.41	0.45	0.54	0.59	0.49	0.75
S_f	0.66	0.59	0.55	0.46	0.41	0.51	0.25

The improvement results of the Spatial Database Vs. Non-Spatial Database are shown in below table.

Non-Spatial Database Vs Spatial							
Item	1	2	3	4	5	6	7
β	0.57	0.57	0.47	0.39	0.35	0.45	0.25
S_f	0.66	0.59	0.55	0.46	0.41	0.51	0.25
↑	9%	2%	8%	7%	6%	6%	0

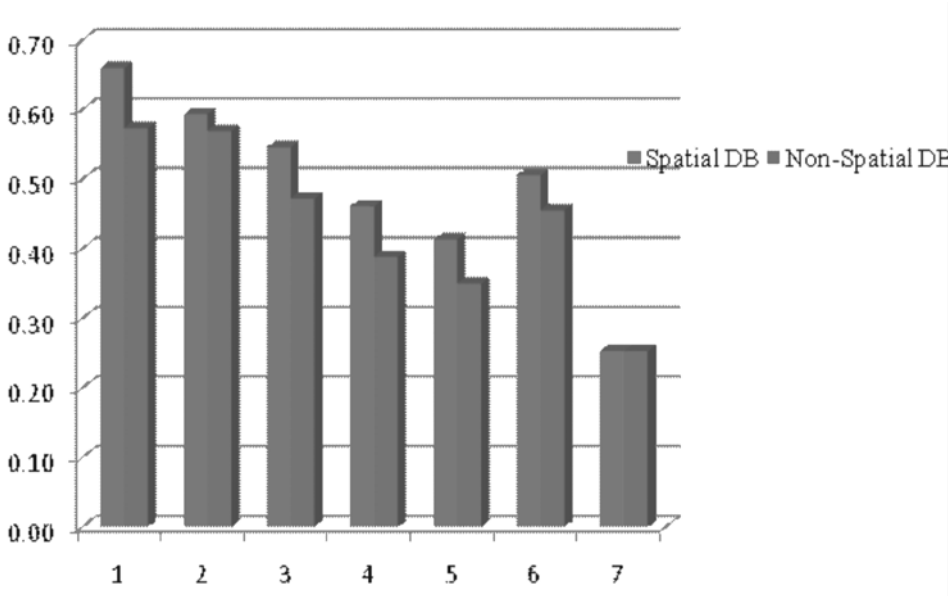

Fig. 6. Comparison graph of spatial and non- spatial DB IR Performance.

The performance graph of the Spatial Database Vs. Non-Spatial Database are shown below.

9. CONCLUSION

Proposed Collocation rule is optimized the search terms in every level and the Spatial IR through maps are improving the browse time to get the required information from the database. We can improve up to 10% of the efficiency when comparing both the DB's. Results are evaluated

10. ACKNOWLEDGEMENT

I would like to take this opportunity to thank Sri. D.Manohar Reddy, MLA,Founder of Trinity Educational Group, Sri.DasariPrashanth Reddy, Chairman, TrinityCollege of Engineering and Technology, Karimnagar, Telangana, India, Dr.P.RajaRao, Principal, Sri. N.Radha Krishna, AO, Trinity College of Engineering and Technology, Karimnagar, Telangana, India, Dr. M.N.Rao, Professor, SRKIT, A.P, India for the continuous encouragement for bringing this article. I would like to extent my special thanks to Sri. K. Bala Showri, Principal of SRKIT Vijayawada, Sri. B.S. Krishna, Secretary SRKIT, Vijayawada, Dr.B.S. Appa Rao Garu, Chairman of SRKIT, Vijayawada to permit me to associate with their organization.

11. REFERENCES

1. CSU, San Jose State University on 2011-08-14
2. Xia Lin, School of Library and Information Science, University of Kentucky, Lexington, KY, Map Displays for Information Retrieval,Journal of The American Society For Information Science, PP 41-53
3. Paul Jen-Hwa Hu , Pai-Chun Ma , Patrick Y.K. Chau, Evaluation of user interface designs for information retrieval systems: a computer-based experiment, Decision Support Systems 27 1999(125–143) www.elsevier.comrlocaterdsw
4. www.wikipedia.com
5. Ray R. Larson, Geographic Information Retrieval and Spatial Browsing University of California, Berkeley
6. Powers, D.M.W.Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness&Correlation, Journal of Machine Learning Technologies ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63
7. Henning Muller , Wolfgang MullerP performance Evaluation in Con ten t-Based Image Retrieval: Overview and ProposalsCentre UniversityAireD'infOrmaTiqueGr Oupe VisionUniversite De Geneve
8. N. K. Kameswara Rao, Dr. G. P. Saradhi Varma, Dr M. Nagabhushana Rao, "Network Architecture to Identify Spatial Knowledge forDengue" International Journal Of Innovative Technology And Research, Volume No. 1, Issue No. 2. pp.155-160 March-2013 (ISSN 2320 –5547).
9. http://mapserver.org/introduction.htmlý.
10. Roddick.JF, Myra Spiliopoulou, "A Survey of Temporal Knowledge Discovery Paradigms and Methods",Vol 16,Issue 4,pp.750-767 IEEE-KDE , July 2002
11. HardyPundt, "Evaluating the relevance of spatial data in time critical situations", Geoinformation for Disaster Management. Springer,pp.779-788,2005.www.springer.com/3-540-249988-5. Item
12. Abdulvahit Torun and ªebnemDuzgun, "Relevance Of Visual Exploration For Strengthening Spatial Thinking & Spatial Knowledge Exploration", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Part B2. Beijing 2008, pp 1049-1056