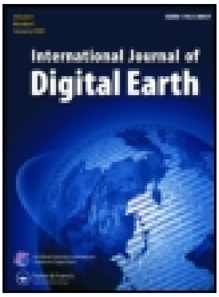


International Journal of Digital Earth
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tjde20

Towards geospatial semantic search: exploiting latent semantic relations in geospatial data

Wenwen Li [a], Michael F. Goodchild [b] & Robert Raskin [c]

[a] GeoDa Center for Geospatial Analysis and Computation, School of Geographical Science and Urban Planning, Arizona State University, Tempe, AZ, USA

[b] Center for Spatial Studies (Spatial@UCSB), University of California, Santa Barbara, CA, USA

[c] NASA Jet Propulsion Laboratory, Pasadena, CA, USA

Available online: 10 Apr 2012

To cite this article: Wenwen Li, Michael F. Goodchild & Robert Raskin (2012): Towards geospatial semantic search: exploiting latent semantic relations in geospatial data, International Journal of Digital Earth, DOI:10.1080/17538947.2012.674561

To link to this article: http://dx.doi.org/10.1080/17538947.2012.674561

International Journal of Digital Earth,
2012, 1–21, iFirst article

Towards geospatial semantic search: exploiting latent semantic relations in geospatial data

Wenwen Li[a]*, Michael F. Goodchild[b] and Robert Raskin[c]

[a]*GeoDa Center for Geospatial Analysis and Computation, School of Geographical Science and Urban Planning, Arizona State University, Tempe, AZ, USA;* [b]*Center for Spatial Studies (Spatial@UCSB), University of California, Santa Barbara, CA, USA;* [c]*NASA Jet Propulsion Laboratory, Pasadena, CA, USA*

(*Received 19 August 2011; final version received 7 March 2012*)

This paper reports our efforts to address the grand challenge of the Digital Earth vision in terms of intelligent data discovery from vast quantities of geo-referenced data. We propose an algorithm combining LSA and a Two-Tier Ranking (LSATTR) algorithm based on revised cosine similarity to build a more efficient search engine – Semantic Indexing and Ranking (SIR) – for a semantic-enabled, more effective data discovery. In addition to its ability to handle subject-based search, we propose a mechanism to combine geospatial taxonomy and Yahoo! GeoPlanet for automatic identification of location information from a spatial query and automatic filtering of datasets that are not spatially related. The metadata set, in the format of ISO19115, from NASA's SEDAC (Socio-Economic Data Application Center) is used as the corpus of SIR. Results show that our semantic search engine SIR built on LSATTR methods outperforms existing keyword-matching techniques, such as Lucene, in terms of both recall and precision. Moreover, the semantic associations among all existing words in the corpus are discovered. These associations provide substantial support for automating the population of spatial ontologies. We expect this work to support the operationalization of the Digital Earth vision by advancing the semantic-based geospatial data discovery.

Keywords: ontology; geospatial semantics; search engine; Digital Earth; similarity; search effectiveness

1. Introduction

Nowadays, geospatial information has been extensively used to support a variety of physical-science and social-science studies, such as natural disaster prediction (Li *et al.* 2009), emergency response (Rauschert *et al.* 2002), and urban economics studies (Anas and Liu 2007). In the past decades, billions of gigabytes of geospatial data have been produced and made available to the public by government agencies and other stakeholders from multiple Earth-orbit missions, ground survey, and in situ measurements. The large volume of data provides science and applied researchers with a valuable resource. To enable the seamless access and visualization of geo-referenced data, the former Vice President of the US Al Gore envisaged a

*Corresponding author. Email: wenwen@spatial.ucsb.edu
This work was conducted at University of California, Santa Barbara.

ISSN 1753-8947 print/ISSN 1753-8955 online

http://dx.doi.org/10.1080/17538947.2012.674561
http://www.tandfonline.com

virtual globe – the Digital Earth – as 'a new wave of technological innovation that allows us to capture, store, process and display an unprecedented amount of information about our planet and a wide variety of environmental and cultural phenomena' (Gore 1998). Ten years later, a number of advanced techniques, such as geobrowsing, distributed geographic information processing (DGIP, Yang *et al.* 2008), and volunteered geographic information (VGI; Goodchild 2007), have been developed to operationalize the Digital Earth concept. However, as a comprehensive goal, the 'Digital Earth' is still facing challenging problems (Xu 1999, Craglia *et al.* 2008). One grand challenge is how to provide an intelligent mechanism to assist users of Digital Earth systems to readily discover, search, and access useful science content from multiple sources. In the position paper from the Vespucci Initiative for the Advancement of Geographic Information Science, Craglia *et al.* (2008) highlighted the importance of establishing 'a dynamic information system to provide reliable, accurate, timely and openly accessible information' for building the next-generation Digital Earth. In 2010, the workshop 'Towards Digital Earth: Search, Discover and Share Geospatial Data 2010' (http://ceur-ws.org/Vol-640/) was held at the Future Internet Symposium and discussed the application of state-of-the-art information technology to enable intelligent discovery of geospatial data. Although several efforts have been made to promote the scientific discovery process, such as establishing data application centers and developing Web catalogs (Li *et al.* 2010) with search capabilities, in reality, scientists are still limited to the use of datasets that are familiar to them (Li *et al.* 2011). These efforts often have little knowledge of the existence of datasets that could be a better fit for their model or application (Gray *et al.* 2005, Singh 2010, Tisthammer 2010) due to the inefficiency of current geospatial search engines. This deficiency brings great challenges to the information-retrieval community to develop more effective mechanisms for intelligent geospatial data discovery and a semantic search platform to support the realization of the Digital Earth vision (Gore 1998, Li *et al.* 2008a, 2008b).

There are two factors that influence the discoverability of a geospatial search engine in the digitized world: accessibility and effectiveness. Accessibility measures whether all existing geospatial data and services can be accessed by as many users as possible; in other words, it involves the process of building the corpus which provides the most up-to-date data. Effectiveness measures whether a search engine is able to find all relevant information by scanning the corpus. One way to improve accessibility is to build a comprehensive corpus containing all available datasets dispersed on the Internet. For example, NASA has built several distributed, discipline-specific active archive centers (DAACs) for scientific modeling and analysis. NASA's Global Change Master Directory (GCMD) and the US Government's Geospatial One Stop (GOS) provide public gateways and catalogs to facilitate the collection and access of geospatial data. Li *et al.* (2010) developed an active crawler to automatically collect the distributed geospatial services that exist on the Web and have not yet been published, and to incorporate them into the aforementioned catalogs to extend the geospatial data corpus. These works have greatly improved the accessibility of geospatial data. However, in terms of improving the effectiveness of a search engine, almost all of the existing geospatial catalogs and Web portals use Lucene, a full-text keyword-matching technique (Hatcher and Gospodnetic 2004). The datasets that are semantically related to a user's query but described differently from the query keyword will be considered irrelevant and excluded from the search results. Hence, improving the effectiveness of a geospatial search engine and making available datasets reachable by scientists is becoming even more significant.

Recently, the emerging semantic technologies are attracting the attention of researchers, who are exploring how to utilize such technology to improve search effectiveness. One direction of the efforts is to incorporate domain ontologies to identify associations and concepts (such as polyseme, synonym) related to a query, recommending a list of related search terms for users to refine their search. These works include Virtual Solar Terrestrial Observatory (VSTO) (Fox *et al.* 2009), Geosciences Network (GEON) (Bowers *et al.* 2004), Linked Environments for Atmospheric Discovery (LEAD) (Droegemeier *et al.* 2005), and Noesis (Movva *et al.* 2008). These solutions rely heavily on the logical representation in the ontology, which is usually developed by humans. The issue is that the words used for indexing a document are often different from those in the pre-defined ontologies. Moreover, different people with different knowledge sets tend to have different perspectives on the categorization of terms and their linkages and relations. This would cause heterogeneous representations and conflicting statements, and eventually influence the effectiveness of a search engine. To overcome this problem, in this paper we propose to use an analytical and human-independent method – latent semantic analysis (Dumais 2004) – which has rarely been applied to the retrieval of geographic data. By applying latent semantic analysis, the semantic structure of documents in the corpus can be discovered and the latent semantics between the occurrences of patterns of words, and clues to the likely occurrence of others, will also be discovered. In this way, even the words with no occurrence in a document will be given weights indicating the correlation between the words and the document.

Latent semantic analysis enables the discovery of more semantically relevant datasets. Meanwhile, these discovered dataset need to be ranked so that the most relevant results will always appear on top. Therefore, we also propose a ranking model based on revised cosine similarity to filter out documents that are not closely related in order to improve the effectiveness of geographic data retrieval. The geospatial metadata sets from the NASA's SEDAC (Socio-Economic Data Application Center) are used as our test corpus in this study.

2. Background and limitation of existing methods

Two criteria are always used to measure the effectiveness of an information-retrieval system: precision and recall. Precision is the ratio between the number of relevant answers retrieved from a search and the total number of answers retrieved. Recall is the ratio of the number of relevant answers retrieved from a search to the total number of relevant answers within the corpus. LSA can improve the recall rate and a good ranking algorithm can improve the precision of an information-retrieval system.

LSA, also known as LSI (Latent Semantic Indexing), was first introduced by Deerwester (Deerwester *et al.* 1990) as a technique to discover the existence of latent structure in the pattern of word usage across documents. It is a variant of the vector space model that uses Singular Vector Decomposition (SVD) and low-rank approximation to enable information retrieval in a reduced-dimensional space of the corpus. The LSA technique has proven to be a valuable analysis tool and has been widely used in information retrieval (Dumais 2004, Gabrilovich and

Markovitch 2009). In addition, several extensions, such as probabilistic LSA (Park and Ramamohanarao 2009), have been proposed to better understand why LSI works. In this paper, we combine standard LSA with stemming and reversed index techniques to enhance semantic association detection, in order to improve the recall aspect of retrieval effectiveness of geographic data.

Cosine similarity is one of the most popular methods used for relevancy ranking based on document similarity theory. It measures the cosine of the angle between the query vector and the document vectors. When the angle is 0, the resulting cosine function equals 1, meaning that the document being measured is the most relevant to the query. Mathematically, the similarity can be represented as:

$$sim(X, Y) = \frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}, \qquad (1)$$

where X is the query vector and Y is the document vector. Cosine similarity supports partial keyword matching and is able to reflect the relevancy between the query and documents on a continuous basis. This clear and simple linear-algebra-based model has a rigorous mathematical foundation. Due to these advantages, the technique has been adopted by a number of well-known search engines, such as Apache Lucene. However, it has also several limitations. For example, the similarity values of long documents can be misleading due to a small scalar product (small value in the numerator) and a large dimensionality (large value in the denominator). Meanwhile, the angle of two vectors is just a relative measure; it ignores the distance between the vectors. This relative measure will result in a false negative match. To overcome the previous problems, we propose a Two-Tier Ranking model based on revised cosine similarity to improve the precision in the retrieval process.

In the next section, the proposed models and system workflow will be introduced.

3. Building an effective platform for retrieving geospatial data

3.1. Pre-processing

To build a search engine for geospatial data retrieval, indexing all the metadata documents in the corpus is an essential step, because indexing optimizes the speed and performance in finding relevant documents for a search query (Gulli and Signorini 2005). The product of indexing is an 'inverted index': the keys are all existing keywords in the corpus and the values of each key are documents containing the occurrence of the specific keyword. The purpose of building an inverted index is to enable the quick location of relevant documents once a query is given. LSI will be performed on top of the inverted index to discover the latent associations between keywords and documents such that even if a keyword does not appear in a document but is detected by semantic analysis as related, the weight of the keyword in the document will be positive instead of 0.

The metadata records in the corpus are encoded in ISO19115 (2003) and need to be preprocessed before indexing for the following reasons. First, there are on average 600 metadata tags in each metadata record and most of the tags are shared across the corpus. If the tags are to be indexed, the similarity among metadata records will be increased due to such high co-occurrences. Although we adopted some techniques to reduce the influence of common terms shared by documents, the indexing time and storage are wasted. Second, traditional indexing techniques conduct full-text index for each document in a corpus. However, this is not necessary when indexing geospatial metadata because a substantial amount of information (such as 'ResponsibilityParty' or 'MetadataStandardName') does not describe the actual content. Therefore, the original metadata records were parsed, and only 'Title', 'Abstract', 'Science Keyword', 'GMCD Keyword', 'Location Keyword', and 'Lineage' were extracted from each metadata record. New text files matching the original metadata documents were generated by streaming out the useful information. We tested in the experiments whether the modified system still maintains high precision and recall rates.

Figure 1 demonstrates the workflow of the search engine, with three phases included: pre-process, indexing, and ranking. The uppermost box shows the pre-processing of the metadata documents discussed previously. Once new text files are generated, the system will scan each text file, extract all the words (we also call them terms), and count the occurrence of these words throughout the whole corpus. This process will generate the 'Word-Frequency List.' During the generation, frequently used words such as 'is', 'the', also known as stop words, were eliminated from the statistics to save disk space. The 1000 most frequent used words reported by Fry and Kress (2006) were used as the vocabulary to filter the stop words.

Another strategy used to improve the retrieval is stemming, which reduces all words with the same morphological root to a common form. For example, 'Antarctica', 'Antarctic' and 'Antarctic's' would all be converted to their root form, 'antarct'. Maximizing the usefulness of a subject word keeps the significance of the words in a corpus. Lovin's stemming algorithm (Lovins 1968) was adopted in this

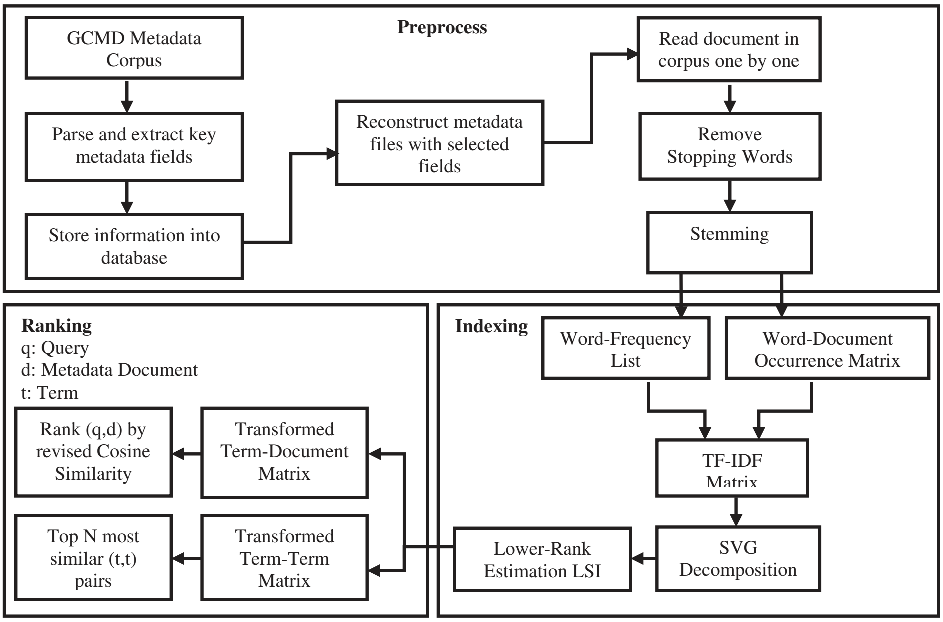

Preprocess
GCMD Metadata Corpus
Parse and extract key metadata fields
Store information into database
Reconstruct metadata files with selected fields
Read document in corpus one by one
Remove Stopping Words
Stemming
Ranking
q: Query
d: Metadata Document
t: Term
Rank (q,d) by revised Cosine Similarity
Transformed Term-Document Matrix
Top N most similar (t,t) pairs
Transformed Term-Term Matrix
Indexing
Word-Frequency List
Word-Document Occurrence Matrix
TF-IDF Matrix
Lower-Rank Estimation LSI
SVG Decomposition

Figure 1. Workflow of the retrieval system for geographical data.

study. The algorithm proposes a list of recoding rules to reduce a word's derivational and inflectional suffixes. But sometimes exceptions occur, for example, 'sediment' was stemmed to 'sedim' and 'sedimental' was stemmed to 'sediment' because the algorithm is not iterative. Thus after the generation of the word-frequency list, the frequencies of words with same root, for example, the substring case, were combined.

3.2. Latent semantic indexing

The term-document occurrence matrix A is the input for LSA. As Figure 2 shows, the rows (unique words) and columns (documents) are all of the metadata records in the corpus. The value in cell (*i, j*) is the number of occurrences of word i in document j, where $i \in [1, m], j \in [1, n]$. The total of each row equals the total frequency of word occurrence in the corpus. The total of each column is the length in words of a metadata document.

In previous studies, the earlier raw matrix was used directly for decomposition in LSI. However, we argue that the cell values indicating the importance of words (currently by the number of occurrences in a document) are biased by the length of the metadata documents and the number of documents in which a keyword occurs. For example, a long document will have a better chance to contain more instances of a given word; and if a word occurs in most or all of the documents in the corpus, it should have less importance. For this reason, we adopted the Term Frequency-Inverted Document Frequency (TF-IDF) to adjust the weight of words (terms) in the Term-Document matrix.

$$tf_{i,j} = \frac{count(word_i, d_j)}{\sum_k count(word_k, d_j)} \tag{2}$$

$$idf_i = \frac{\sum_j d_j}{\sum_j occurrence(word_i, d_j)} \tag{3}$$

$$tf - idf_{i,j} = tf_{i,j} * idf_i, \tag{4}$$

	d1	d1	..	..	..	dn
k1						
k2						
.						
.						
km						

Figure 2. Structure of term-document matrix *A*.

where i and k are words index (row index), j is the document index (column index); *count* (*word*$_k$, *d*$_j$) is the number of times that *word*$_k$ occurs in document *d*$_j$. *occurrence*(*word*$_i$, d$_j$*) is a binary function: if* word$_i$ *occurs in document* d$_j$*, the function returns 1; otherwise, it returns 0.* tf$_{i,j}$ *adjusts the weight of* word$_i$ *in* d$_j$ *by the length of* d$_j$*;* idf$_i$ *adjusts the weight of* word$_i$ *by its co-occurrences across the corpus.* tf–idf$_{i,j}$ *is the product of* tf$_{i,j}$ *and* idf$_i$. *This factor is applied to the whole term-document matrix* A *to obtain* A′.

Once the input matrix is prepared, the LSI can be computed. The foundation of LSI is Singular Value Decomposition (SVD), by which the matrix A is decomposed into three matrices: (1) a term-by-concept matrix W (dimensions: $m \times r$) describing the original column vector as an orthogonal unit vector; (2) a concept-by-document matrix P (dimensions: $r \times n$) describing the original row vector as an orthogonal unit vector; and (3) a diagonal matrix S (dimensions: $r \times r$) containing the scale values, as Equation (5) shows.

$$A' = WSP, \tag{5}$$

in which $WW^T = I$ and for any column vector W_i and W_j, where $i, j \in [1,r]$

$$\|W_i W_j\| = 0, \text{ when } i \neq j \tag{6}$$

$$\|W_i W_j\| = 1, \text{ when } i = j \tag{7}$$

$PP^T = I$ and for any row vector P_i and P_j, where $i, j \in [1,r]$

$$\|P_i P_j\| = 0, \text{ when } i \neq j \tag{8}$$

$$\|P_i P_j\| = 1, \text{ when } i = j \tag{9}$$

S is a diagonal matrix, namely $S_{i,j} = 0$, where $i \neq j$ and $i, j \in [1,r]$. In addition, S satisfies

$$S_{i,i} \geq S_{j,j}, \text{for } \forall i \leq j$$

The Singular Value Decomposition makes the reduced dimension or lower-rank estimation of A' possible. By retaining the largest k-dimensional scale values in the matrix S and setting the remaining scale values to 0, and then combining the three matrices by matrix multiplication, the term-document matrix A' can be represented in a reduced LSI space by $\tilde{A}'$. The value of k is between [1,r]; when $k = l$, the estimated matrix $\tilde{A}'$ is generated using only the largest scale value in the LSI space; and when $k = r$, the estimated matrix $\tilde{A}'$ equals to A'. An important consequence of this lower-rank estimation is that the words are no longer independent in the LSI concept space (Dumais 2004), while they are orthogonal and independent in the original term space. In traditional information retrieval, each document vector is represented in the term space. For example, ‘census’ and ‘population’ will be considered orthogonal and the correlation between them is 0. Therefore, when a query contains only ‘population’, the document containing ‘census’ will not be returned although they are related in meaning. However, by the lower-rank estimation, all terms and documents are represented in the LSI space, so terms are no longer orthogonal and the locations of the term vectors reflect their correlations in terms of their usage in the corpus.

3.3. Indexing of location information aided by the GCMD location taxonomy

There are two primary ways for conventional spatial search engines to handle location information. One method, typified by the Global Earth Observation System of Systems (GEOSS) Clearinghouse metadata search engine (http://clearinghouse.cisc.gmu.edu/geonetwork/srv/en/main.home), considers the location information as a spatial constraint. Although some basic geocoding service is provided, for most spatial queries, users still need to provide the reference of geographic extent by drawing a bounding box on a map or typing in the geographic coordinates. This common search mechanism, requiring users to be familiar with the geography of regions of interest, limits the flexibility and usability of spatial search engines, and restrains the Digital Earth user community, which includes 'all the world's citizens' as stated by Gore (1998). Another type of search engine considers the query place name in the same way as any other keyword and will retrieve datasets that contain the specified name in the metadata description (Jones *et al.* 2004). However, without an effective way of recognizing the presence of place names in a query expression and a mechanism to annotate the geographic extent of a dataset, it is still difficult for a spatial search engine to achieve satisfactory performance.

To overcome the earlier limitations, a geospatial taxonomy – GCMD location keyword – was introduced in the semantic indexing to guide the search related to location. The GCMD location keyword has a six-tier hierarchy: Category > Type > Sub-Region1 > Sub-Region2 > Sub-Region3 > Location. The places are categorized into six classes according to their locations in the space (below, on, and above the surface of the Earth). The six classes, namely, 'Continent', 'Geographic Region', 'Ocean', 'Solid Earth', 'Virtual Location', and 'Space' are further classified by their sub-regions. Figure 3(a) and (b) shows the skeleton of the GCMD location keyword structure. The nodes in gray are examples of locations with full paths. For example, through the hierarchical definition, it can be derived that the island Kiribati is located in the Central Pacific Ocean, which is part of the Pacific Ocean, which is a sub-region of Ocean.

To make use of this spatial taxonomy, each individual metadata record annotated the locations covered by the dataset using the metadata tag <gmd:keyword>. Figure 4 shows a fragment of location keywords being annotated in the ISO metadata entitled 'IPCC IS92 Emissions Scenarios'. There are in total 476 locations encoded in this metadata and more than 2000 keywords are contained on the paths of the location classification tree. These location keywords were extracted during metadata parsing and included in the latent semantic indexing discussed in the previous section. The introduction of indexing locational information has the following advantages. (1) The association of science keywords will be emphasized by the adoption of a spatial taxonomy. An intuitive example for explaining the phenomenon is when two datasets have overlaps in the regions they cover, they would share more location keywords along the paths of hierarchical location annotation. These co-occurred keywords will certainly increase the relatedness of other science keywords being indexed in the metadata documents by the essence of LSI. This finding could also be justified by the First Law of Geography (Tobler 1970), which indicates that geographically closer things are more related than more distant entities. (2) Search engines adopting this spatial taxonomy would better handle spatial queries with locational information as keywords. With the

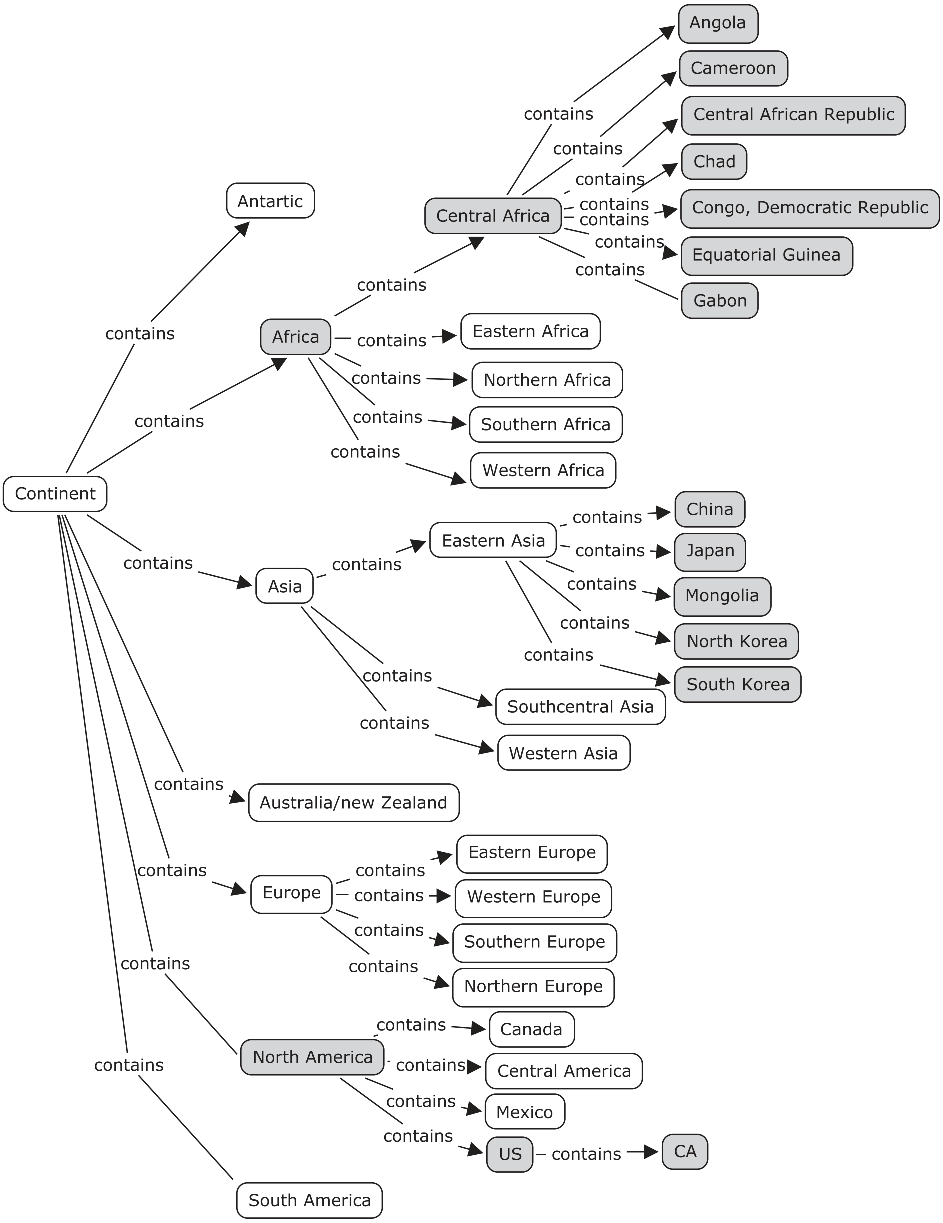


Figure 3. Fragment of GCMD geospatial taxonomy: (a) hierarchical classification of continent; (b) hierarchical classification of geographic region, ocean, atmosphere, and space.

informative annotations of locational information in the metadata documents, the possibility of retrieving spatially matched datasets will be greatly increased.

Besides being used to annotate the geographic regions that a dataset covers in its metadata, the GCMD location taxonomy is also utilized for automatic place name detection from spatial queries by combing with Yahoo! GeoPlanet. As an emerging Internet Location Platform, Yahoo! GeoPlanet provides a series of Application

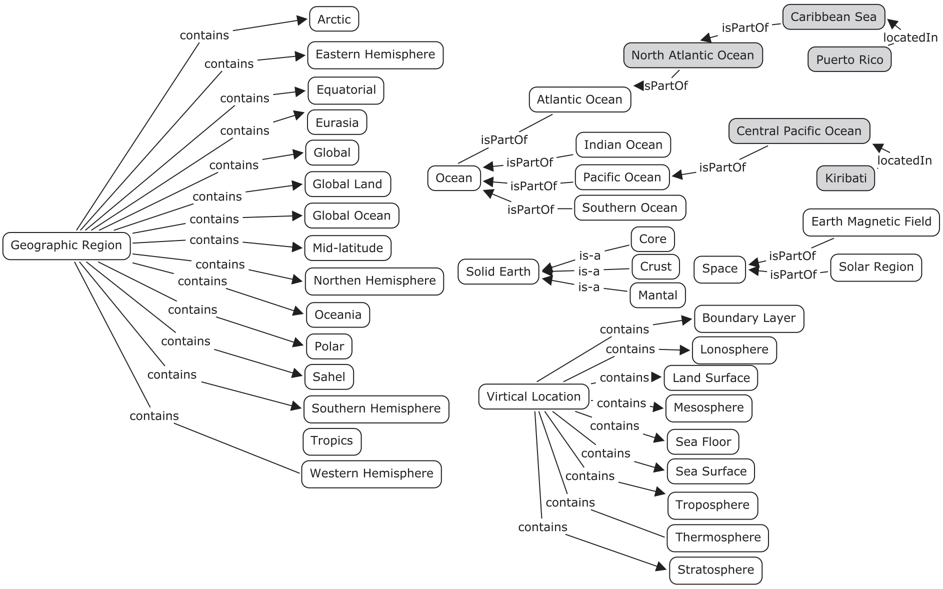


Figure 4. GCMD location keywords annotated in the ISO metadata.

Programming Interface (APIs) to traverse the global spatial hierarchy and geo-code the places (http://developer.yahoo.com/geo/geoplanet/data/). Different from another popular geocoding service, GeoNames (http://www.geonames.org/), GeoPlanet provides a bounding box of a region in addition to the center coordinates provided by GeoNames. This feature enables a spatial filtering function: (1) detecting the place name from spatial queries; (2) geocoding: converting the place name to a bounding box; and (3) comparing the queried region with the geo-extent provided by the dataset. The implementation details are discussed in Section 4.3 and associated experimental results will be discussed as well.

3.4. Two-Tier Ranking by revised cosine similarity

The beauty of LSI lies in its ability to uncover semantically related documents. Using the estimated term-document matrix $\tilde{A}'$ obtained in the previous section, more related documents can be found once a query is given (here the query is represented by keywords in a vector). However, not only do all relevant documents need to be discovered, but also the most relevant documents need to be on top of the returned results because search users will lose interest after checking the first few results. Therefore, a ranking algorithm becomes important to the data retrieval process. As discussed in Section 2, the most commonly used method is cosine similarity, which measures the angle between the query vector (such as Q1 in Figure 5) and document vectors (such as D1 and D2 in Figure 5).

As shown in the figure, the angles between D1 and Q1 and between D2 and Q1 are the same, therefore, they have the same rank in terms of relevance to Q1. However, we observed that: (1) by applying LSI, the weights of terms with original value 0 in a document will be always be reassigned to weights greater than zero and mostly, the values are very small when no strong latent semantic relation is found.

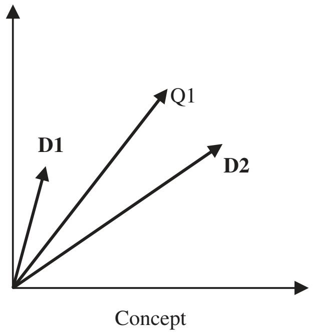


Figure 5. Query and document vectors in LSI concept space.

Therefore, the closer vectors (such as Q1 and D2) should be more similar than those further apart (such Q1 and D1). We call this a 'distance' requirement. (2) The dimension of LSI is usually large (size equals the number of terms in the corpus), while the number of keywords given by users in a query is usually less than eight (Hitwise 2011). So similarity values may be misleading due to a small scalar product (small value in the numerator) and a large dimensionality (large value in the denominator) if weights in every dimension of LSI are used. We call this a 'dimension' requirement.

To satisfy these requirements, we propose a new rank method to measure the similarity between a query vector X and a document vector Y:

$$sim(X, Y) = \frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}}, \tag{10}$$

in which n equals the number of query keywords instead of the total number of words in the corpus. This strategy greatly helps to remove the bias caused by using the vector formed by the complete keyword list. In addition, Equation (10) considers not only the angle but also the separation between vectors, so the requirements on 'distance' and 'dimension' are both satisfied. This will be the tier-1st ranking. Meanwhile, the ranking given by Equation (10) does not guarantee that documents with full matching will be measured as more relevant to the query than those with partial hits. So we propose the tier-2nd ranking, which is to rank on the result obtained from Equation (10) by the number of hits in descending order. In this way, documents containing all j query keywords will still be ranked by Equation (10), while those documents containing $t(t<j)$ query keywords will be ranked lower than the documents containing all j query keywords.

4. Experimental results

In the corpus, there are in total 200,000 geospatial metadata records from GCMD. These metadata are provided by a number of organizations, including SEDAC, NSIDC (National Snow and Ice Data Center), and ACCDC (Atlantic Canada Conservation Data Center). In the experiments in the current phase, we selected the

SEDAC subset (145 documents) to be our selected test corpus for the following reasons: first, through some initial experiments, we found that the clustering pattern of word usages and metadata documents has intra-institutional characteristics. Second, it is easy to examine the actual number of documents related to a given query from a small corpus. So it will be easier to validate the effectiveness of the proposed retrieval methods.

4.1. Precision and recall

Two criteria are used to evaluate the effectiveness of the system: recall and precision. Precision is the ratio between the number of relevant answers retrieved from a search and the total number of answers retrieved by that search. If all the retrieved answers are relevant to the search, the precision reaches its highest peak, which is one. If none of the retrieved answers are relevant, the precision reaches its lowest value, zero. Mathematically, precision can be defined as follows:

$$\text{Precision} = \frac{|\{\text{all relevant records}\} \cap \{\text{all retrieved records}\}|}{\{\text{all retrieved records}\}} \tag{11}$$

Recall is the ratio between the number of the relevant answers retrieved from a search and the total number of relevant answers within the corpus. If all the relevant answers are retrieved by the search algorithm, the recall rate is one. If none of the relevant answers can be retrieved, the recall rate is zero.

$$\text{Recall} = \frac{|\{\text{all relevant records}\} \cap \{\text{all retrieved records}\}|}{|\{\text{all relevant records}\}|} \tag{12}$$

Eight queries, listed in Table 1, were conducted on the corpus. The precision and recall rates of the eight queries from our proposed search engine SIR were compared with those obtained from Geonetwork, one of the most popular metadata search engines relying on Lucene.

For Q 1.1, by typing ‘natural disaster death’, a user expects to retrieve death statistics caused by natural disasters. We determine the relevance of a dataset to the query by meanings instead of keyword occurrences. So even though a document contains all of the aforementioned keywords, or the weights of the aforementioned keywords that come up after conducting semantic analysis, the dataset is still considered to be irrelevant if the subject is not directly related. Results show that Geonetwork returns zero results on this query; in contrast, using our proposed method, there are 29 results returned from SIR. The topics of the returned records include global earthquake/flood/volcano/drought/landslide/cyclone mortality risks and distributions, and the economic losses caused by these disasters. It is clear that all these records are related to Q 1.1. Through examining the test corpus, we found 31 documents are relevant in total. Therefore, by applying Equations (11) and (12), the precision of SIR reaches 100% and that of Geonetwork is 0%; the recall rate of the SIR is 94%, while that of Geonetwork is still 0%. The reason for the significant performance difference of the proposed method and Lucene-based searching is that ’mortality’ is relevant to the query word ‘death’, but it is not present in the metadata document. The proposed method is able to detect this association while the pure full-text indexing cannot.

Table 1. Selected query for evaluating search effectiveness.

Query type	Query	Keyword
1	Q 1.1	Natural disaster death
	Q 1.2	Disaster population impact
	Q 1.3	Natural disaster damage
	Q 1.4	Wildlife distributions by species
	Q 1.5	Global climate change pollution
	Q 1.6	China agriculture food sustainability
	Q 1.7	Census housing condition
	Q 1.8	Africa poverty statistics
2	Q 2.1	Colorado population
	Q 2.2	California population dynamics in the United States
	Q 2.3	Wild life habitat of Costa Rica
	Q 2.4	China County level population data
	Q 2.5	Puerto Rico census data

For Q 1.5 (global climate change pollution), a user expects to see the data indicating the relation between climate change and pollution, such as the dataset showing the climate change caused by air pollution. Using the proposed method, SIR returned 33 documents and Geonetwork only returned 6 documents. By examining the test corpus, we found that 20 of the 33 records are relevant to the query. The extra records returned by SIR are on the subjects of 'Global Multi-hazard Total Economic Loss', etc. In these records, there are occurrences of words 'environmental protection', which is related to 'pollution'. But because the subject is more on the loss estimation from 'hazards' instead of 'pollution', they are not considered relevant. Similarly, in the results returned by SIR for Q 1.2, there is a record named 'China Dimensions Data Collection: Fundamental GIS: Digital Chart of China, 1:1M, Version 1'. The record is about the environmental impact to humans, but the environmental impact is from 'urbanization' rather than 'disaster'. As it is not directly relevant, this record is considered irrelevant when measuring recall and precision.

Figure 6 shows the comparison of overall recall and precision rates for our SIR search engine applying LSA and the Two-Tier Ranking (LSATTR) algorithms and that for Lucene search in Geonetwork. We can tell that the recall rate by the LSATTR method is much higher than using Lucene. Except for Q 1.1, which has the recall rate at 94%, all other queries have 100% recall rate. It means that all relevant records in the corpus could be retrieved by the LSATTR. However, the LSATTR method also returned datasets judged not closely relevant (as analyzed earlier) and this influenced the precision rate. As shown in Figure 6(b), although Geonetwork returns fewer records than SIR, all the records returned are relevant. Therefore, it maintains a higher precision rate than our retrieval system. Note that precision rate is just a relative measure; although it is a little higher when using Luence than using the LSATTR method, our search engine SIR was still able to identify all records found by Lucene.

4.2. Detected word–word associations

In the test corpus, 2541 unique words were extracted (not including the stop words). Besides using LSI to improve the retrieval effectiveness, the semantic associations

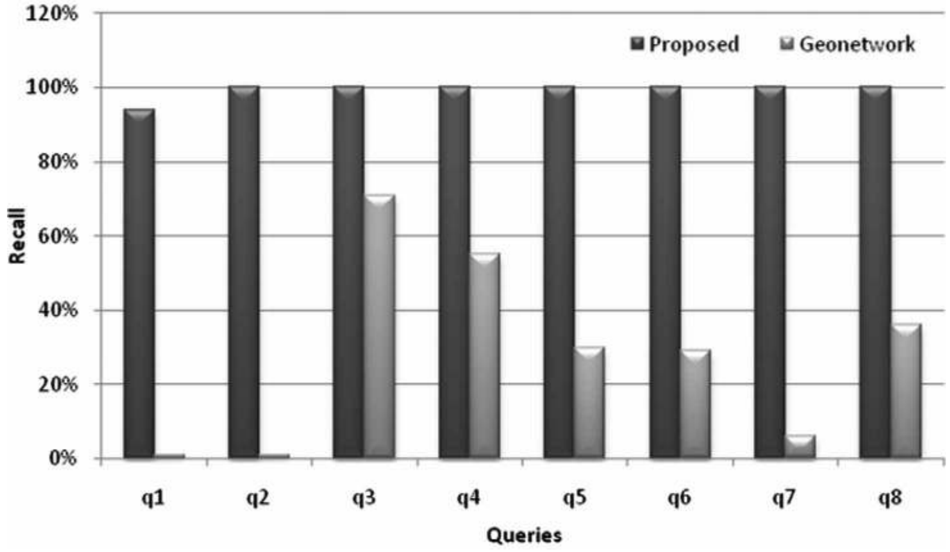


Figure 6a. Recall comparison between proposed method and Geonetwork.

between terms can also be found as a side product. In this case, there are more than 300,000 values indicating the association between term pairs. The square symmetric matrix T_T containing all the word–word dot products can be computed from W and S – the component in Equation (5). According to Deerwester *et al.* (1990),

$$T\_T = WS^2W \qquad (13)$$

where T_T is the 2541*2541 matrix and $T\_T[i, j]$ represents the relatedness between word i and word j.

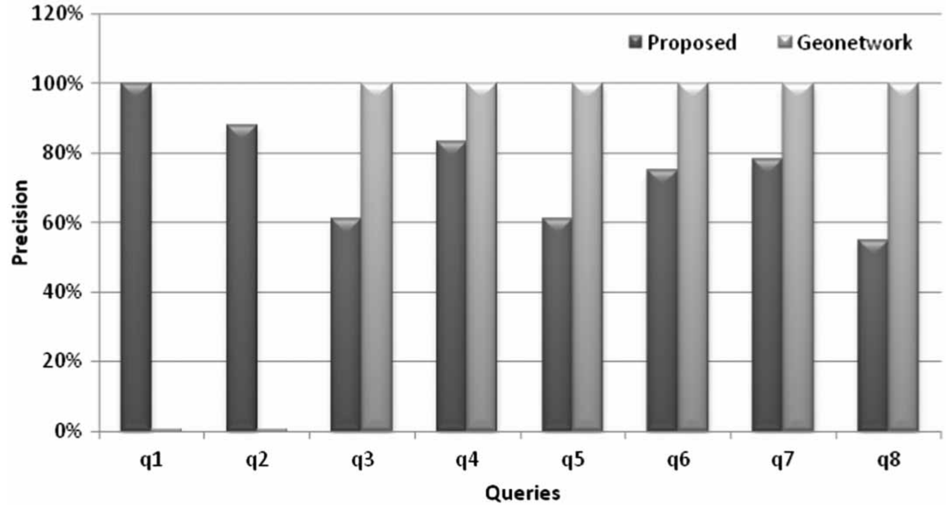


Figure 6b. precision comparison between proposed method and Geonetwork.

The algorithm for detecting the semantic associations is as follows:

(1) Construct a hash table for all entries (terminologies) in the GCMD science keyword taxonomy for quick lookup. The key of the hash table is a science keyword, and the value is the cluster, e.g. Biosphere, that the keyword belongs to. Without the hash table, we need to traverse all the trees in the taxonomy forest to look up a term and it will be a very time-consuming task.
(2) Given a term, e.g. 'Forest', in the GCMD science keyword taxonomy, by searching the *T_T* matrix, all related terms in the corpus could be returned in descending order of relatedness.
(3) Determine whether each related term has an entry in the GCMD science keyword taxonomy by looking it up in the hash table. Once a hit is found, the association 'hasRelatedCluster' can be added automatically by a Jena (http://incubator.apache.org/jena) operation. Note that the 'Cluster' will be replaced by the name of a specific cluster that the related term belongs to. For example, if the association between 'Forest' and 'Flood' was discovered from the *T_T* matrix, then the relation will be defined as 'hasRelatedEcosystem-Term' from 'Forest' to 'Flood'.
(4) After the associated concepts were built from the previous procedure, the new taxonomy needs to be evaluated by the domain experts to assure its accuracy.

Figure 7 shows the 16 words most related to 'Forest' obtained from *T_T* and organized in the clusters of the GCMD science keyword taxonomy (Olsen *et al.* 2007). The different colors (in grayscale) indicate different clusters and the numbers 1–13 listed beside the nodes are the ranks of relatedness. From this figure, we can tell that the most related words to 'Forest' are its ancestors, e.g. 'Biosphere', and its siblings, such as 'Grassland'. The relationships of words from other clusters to 'Forest' are also uncovered, as the dotted arrows show. For example, the relation between 'Forest' and 'Human Dimension' and that between 'Forest' and 'Terrestrial Ecosystem' (through its child node 'Flood') are discovered. The identification of these relations provides a semi-automated way to associate semantically related words in isolated clusters in the taxonomy together. Besides, in the current taxonomy, there are no entries for words 'timber', 'wildlife', and 'woodland', which are also considered to be related through the latent semantic analysis. Adding these semantically related words and associations into the taxonomy would greatly enrich its semantics toward generating a synthesized domain knowledge base.

4.3. Improvement of spatial search by geospatial taxonomy

Besides the advancement of subject-based search by the support of the proposed indexing and ranking algorithm, we also utilized the GCMD location taxonomy to improve the spatial search when a place name appears as part of a query. This procedure includes the automatic detection of a place name in the query and a spatial filter function to exclude datasets not spatially related. The following three-step procedure describes the implementation details:

- Step 1: Link each region/place name encoded in the GCMD location taxonomy to a bounding box. This linkage is automatically built by traversing

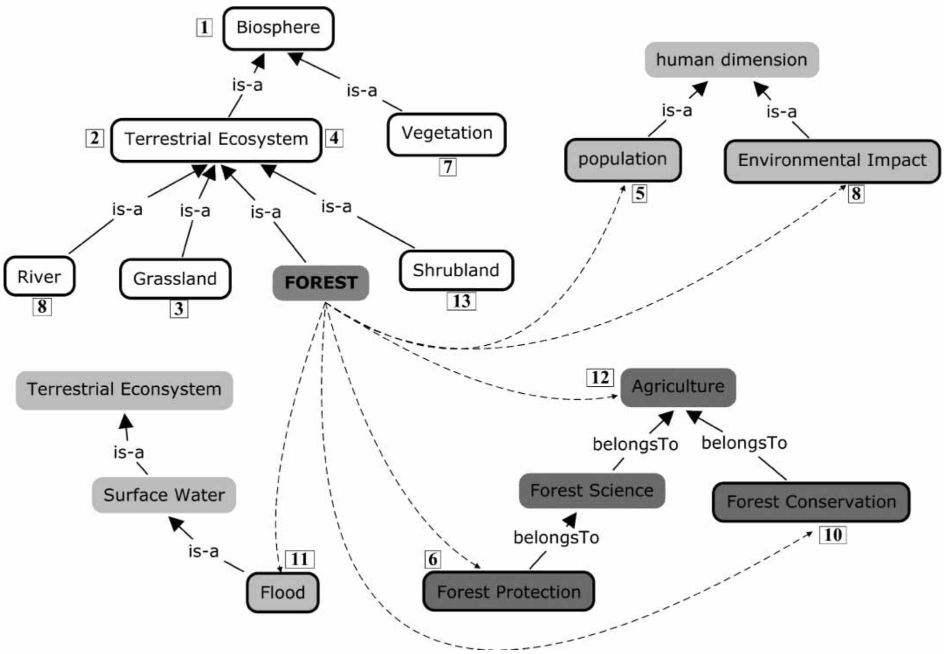
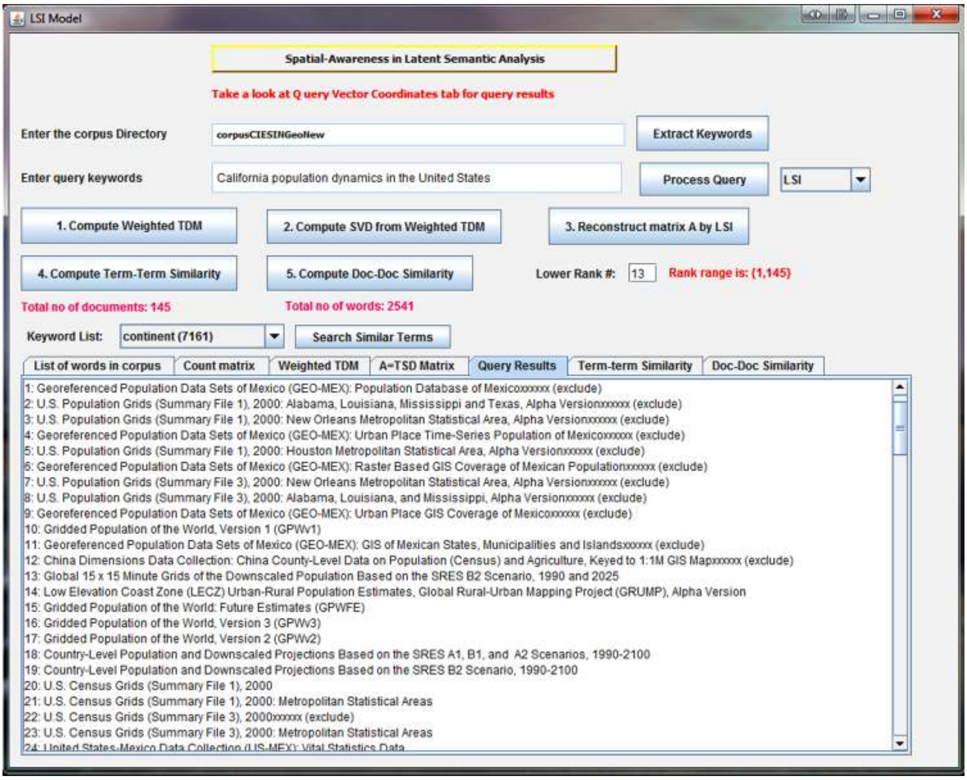


Figure 7. Latent semantic association discovery.

the taxonomy tree and geolocating the place name on the current node using GeoPlanet APIs. For example, to obtain the bounding box and centroid of 'North America', a HTTP request http://where.yahooapis.com/v1/places.q%28%22north%20america%22%29?appid={Ahe1C6HV34HLpJjFSX.svGthr7_1Ddd207T_f37S7tsuL4VUms2VY1P1uiWyAjbsWg} could be constructed. The coordinates of the southwest and northeast corners of the geographical extent of the queried place could be extracted and encoded into the GCMD location taxonomy.

- Step 2: Detect the existence of a place name in a query. Once a query is given, all possible combinations of adjacent keywords can be obtained as the potential place name strings. For a query containing n keywords, there will be $n(n+1)/2$ possible combinations. Each keyword combination will be looked up in the GCMD taxonomy to find a matching node. Once a matching node is found, the geo-extent of the node (a place name) can be acquired. If more than one place name is being detected and if their geo-extents have overlaps, the one with smaller geo-extent will be used.
- Step 3: The 'spatial filter' function helps the SIR semantic search engine to exclude the datasets that do not cover the area of interest given in a user's query. This filter is implemented by comparing the geo-extent covered by a dataset and the requested geo-extent obtained from Step 2. If a dataset does not cover the region of interest at all, it will be excluded from the result set. That is, if the bounding boxes of the queried location and the geospatial dataset have no overlap, the dataset will be filtered out.

The introduction of spatial taxonomy-based location annotation, place name detection, and spatial filter functions further improves the effectiveness of geospatial search, especially the location-based search. To compare the performance of SIR which adopts the proposed mechanism in dealing with spatial queries with Geonetwork search, we used another set of queries (listed as Type II queries in Table 1). Our focus is to examine both search engines in terms of their abilities to automatically identify place names and perform the correct spatial filter function. Therefore, in comparison to Type I queries, Type II queries have less complex semantics but all contain place/region names as part of the queries. From the experiments, very promising results were found. For instance, Query 2.1 and Query 2.2 are to search for population-related datasets in different states: Colorado and California. Query 2.2 includes more than one place name (‘California’ and ‘United States’). Based on the rule we set, ‘California’ has the smaller extent and should be picked. These two queries were designed to test whether SIR can filter out the population datasets covering only New Orleans, Texas, and Louisville instead of the whole US region, given that there exist such datasets in the corpus whose titles indicate that the whole US is being covered. For Query 2.1, only one record (about ‘USDA plant list’) is returned from Geonetwork and it is irrelevant. The reason why this record is returned is that in the abstract of the metadata file, there are occurrences of both ‘population’ and ‘Colorado’, but ‘population’ is in the context of population interaction with species and ‘Colorado’ is the location of a research center that maintains the plant dataset for USDA. For SIR, the top 20 metadata records returned are all relevant and the location condition is satisfied by each of the results, as well. For Query 2.2, Geonetwork returns six records (all about ‘China Dimensions Data Collection’); however, neither the content nor the geospatial extent of the results satisfies the query. Similar to Query 2.1, these six records were returned only because there are matches of the queried keywords in the metadata documents. As for SIR, for the top 20 datasets returned, all are relevant. Figure 8 (the prototype graphic user interface [GUI] of SIR) shows the results for Query 2.2. In the result set, we highlighted the metadata records with ‘(exclude)’ at the end of such records that match the query in terms of topic but not the location condition. The purpose is to demonstrate the effect of the spatial filter function. For example, the first record that SIR returns is a population dataset of Mexico and is automatically filtered out by SIR. For records 2, 3, 7, and 8, although the titles say ‘US population grids’, the datasets only cover a few states in the southeast US, and are excluded as well. Similarly, Record 12 is also excluded because its coverage is in Asia instead of California, US. From the results, we also found that the records (e.g. 20, 21, 23) which do not contain the exact keyword ‘Population’ but contain its synonym ‘Census’ also were ranked highly, although behind those containing ‘Population’. This reflects the benefits of adopting latent semantic analysis and the proposed ranking algorithm to discover relevant datasets even if only words of similar meanings are used in the metadata. SIR also works well for other spatial queries from the experiments. For example, for Query 2.3, Geonetwork returns zero records because of no place name/subject matches in the metadata documents. In contrast, the top 10 records returned from SIR are all closely relevant and the records that were ranked lower are partially relevant. Other tests we conducted, e.g. Q 2.4 and Q 2.5, all returned satisfying results.



Take a look at Q uery Vector Coordinates tab for query results

Enter the corpus Directory: corpusCIESINGeoNew

Enter query keywords: California population dynamics in the United States

LSI

Lower Rank #: 13 Rank range is: (1,145)

Total no of documents: 145

Total no of words: 2541

Keyword List: continent (7161)

```
1: Georeferenced Population Data Sets of Mexico (GEO-MEX): Population Database of Mexicoxxxxxx (exclude)
2: U.S. Population Grids (Summary File 1), 2000: Alabama, Louisiana, Mississippi and Texas, Alpha Versionxxxxxx (exclude)
3: U.S. Population Grids (Summary File 1), 2000: New Orleans Metropolitan Statistical Area, Alpha Versionxxxxxx (exclude)
4: Georeferenced Population Data Sets of Mexico (GEO-MEX): Urban Place Time-Series Population of Mexicoxxxxxx (exclude)
5: U.S. Population Grids (Summary File 1), 2000: Houston Metropolitan Statistical Area, Alpha Versionxxxxxx (exclude)
6: Georeferenced Population Data Sets of Mexico (GEO-MEX): Raster Based GIS Coverage of Mexican Populationxxxxxx (exclude)
7: U.S. Population Grids (Summary File 3), 2000: New Orleans Metropolitan Statistical Area, Alpha Versionxxxxxx (exclude)
8: U.S. Population Grids (Summary File 3), 2000: Alabama, Louisiana, and Mississippi, Alpha Versionxxxxxx (exclude)
9: Georeferenced Population Data Sets of Mexico (GEO-MEX): Urban Place GIS Coverage of Mexicoxxxxxx (exclude)
10: Gridded Population of the World, Version 1 (GPWv1)
11: Georeferenced Population Data Sets of Mexico (GEO-MEX): GIS of Mexican States, Municipalities and Islandsxxxxxx (exclude)
12: China Dimensions Data Collection: China County-Level Data on Population (Census) and Agriculture, Keyed to 1:1M GIS Mapxxxxxx (exclude)
13: Global 15 x 15 Minute Grids of the Downscaled Population Based on the SRES B2 Scenario, 1990 and 2025
14: Low Elevation Coast Zone (LECZ) Urban-Rural Population Estimates, Global Rural-Urban Mapping Project (GRUMP), Alpha Version
15: Gridded Population of the World: Future Estimates (GPWFE)
16: Gridded Population of the World, Version 3 (GPWv3)
17: Gridded Population of the World, Version 2 (GPWv2)
18: Country-Level Population and Downscaled Projections Based on the SRES A1, B1, and  A2 Scenarios, 1990-2100
19: Country-Level Population and Downscaled Projections Based on the SRES B2 Scenario, 1990-2100
20: U.S. Census Grids (Summary File 1), 2000
21: U.S. Census Grids (Summary File 1), 2000: Metropolitan Statistical Areas
22: U.S. Census Grids (Summary File 3), 2000xxxxxx (exclude)
23: U.S. Census Grids (Summary File 3), 2000: Metropolitan Statistical Areas
24: United States-Mexico Data Collection (US-MEX): Vital Statistics Data
```

Figure 8. Prototype GUI of SIR.

5. Graphic user interface

Figure 8 demonstrates the prototype GUI to test the proposed methodology. It contains the following components: (1) a corpus directory is used to designate what datasets are used for Semantic Indexing and Ranking (SIR); (2) the textbox below it is for entering query keywords. There are three modes for conducting the search: 'Brute force', 'TF*IDF', and 'LSI'. 'Brute force' will only match the query by counting the occurrence of keywords in the metadata; 'TF*IDF' adjusts the importance of each keyword in the documents and therefore improves similarity ranking from the 'Brute force' method; 'LSI' uses our proposed method for semantically indexing on the words that exist in the metadata documents, and will adjust the weight of other keywords based on co-occurrences and relatedness across the entire corpus. In the experiments discussed in Section 4, LSI mode was selected. (3) The buttons labeled 1–5 show the procedure to generate the term-document matrix and to compute similarity among terms and documents; (4) each tab, e.g. 'List of words in corpus' or 'Query results', presents statistical information or the search results.

6. Conclusion and discussion

This paper discussed a combined LSATTR' algorithm to improve the effectiveness of geographical data retrieval to address the grand challenge of the Digital Earth vision in terms of intelligent data discovery from vast quantities of geo-referenced data. Experiments show that a retrieval system implementing the proposed method improved the retrieval of relevant documents significantly – for all eight sample subject-based (Type I) queries, the recall rate almost reached 100%. Although the precision is in some cases lower than the Lucene-based retrieval method, the system guarantees that all the records returned by Lucene could be discovered by the proposed retrieval system. Besides the capability of handling subject-based queries, we also introduced the advanced mechanisms of automatic place-name detection and spatial filtering to handle spatial queries with the assistance of the GCMD location taxonomy. Applying the proposed methodology in geographical data retrieval has the advantages of (1) discovering latent semantic associations between terms and enabling fuzzy match (match based on meanings instead of appearances); (2) on-the-fly query answering, because the time-consuming aspects – SVD of the process mostly occurs at the pre-process phase; (3) effective identification of place name as part of a spatial query for spatial filtering; (4) conducting subject-based and location-based query simultaneously; (5) automated discovery of semantic linkage among geospatial data resources to enrich the geospatial taxonomy.

There are several directions that might further improve the research. In the current SIR system, the granularity of the latent semantic analysis is still a single word. That is why the phrase 'Terrestrial Ecosystem' in Figure 7 has two rankings in terms of its relatedness to the given word 'Forest'. Sometimes, a phrase would have stronger semantic meaning than the words in it considered separately. In the future, we will extend the LSI to handle phrase-based latent semantic analysis by measuring the occurrence/distance of words in a document and cross-matching the entries in existing ontologies, such as SWEET (Raskin and Pan 2005). Second, the proposed discovery mode is based on the assumption that the metadata documenting the information of the actual dataset is accurate and complete. To match user queries with data content requires a great amount of time for manual checking by humans and yet is still a very challenging topic for machine processing. Therefore, the quality of metadata from different providers would influence the discoverability of the dataset even with a well-performing retrieval system. In the future, we will release SIR as an open-source geospatial search engine and integrate the advanced techniques into the current popular metadata search engine Geonetwork to benefit the broader Digital Earth user community.

References

Anas, A. and Liu, Y., 2007. A regional economy, land use, and transportation model (relu-tran): formulation, algorithm design, and testing. *Journal of Regional Science*, 47 (3), 415–455.

Bowers, S., Lin, K., and Ludascher, B. 2004. On integrating scientific resources through semantic registration. *In*: *Proceedings of the 16th international conference on scientific and statistical database management*, 21–23 June 2004, Santorini Island, Greece: IEEE Computer Society, 349–352.

Craglia, M., *et al.*, 2008. Next-generation Digital Earth – a position paper from the Vespucci initiative for the advancement of geographic information science. *International Journal of Spatial Data Infrastructures Research*, 3, 146–167.

Deerwester, S., Dumais, S.T., and Harshaman, R., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (6), 391–407.

Droegemeier, K., *et al.*, 2005. Service-oriented environments for dynamically, interacting with mesoscale weather. *Computing in Science & Engineering*, 7 (6), 12–29.

Dumais, S.T., 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38, 189–230.

Fox, P., *et al.*, 2009. Ontology-supported scientific data frameworks: the Virtual Solar-Terrestrial Observatory experience. *Computers and Geosciences*, 35 (4), 724–738.

Fry, E.B. and Kress, J.E., 2006. *The reading teacher's book of lists: grades K-12*. San Francisco, CA: Jossey-Bass.

Gabrilovich, E. and Markovitch, S., 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34, 443–498.

GEO, 2005. *The global earth observation system of systems (GEOSS) 10-year implementation plan* [online]. Available from: http://www.earthobservations.org/docs/10-Year%20 Implementation%20Plan.pdf [Accessed 8 December 2011].

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.

Gore A., 1998. *The Digital Earth: understanding our planet in the 21st century, given at the California Science Center*. Los Angeles, CA: Open Geospatial Consortium.

Gray, J., Liu, D.T., and Dewitt, D.J., 2005. Scientific data-management in the coming decade. *SIGMOD Record*, 34 (4), 34–41.

Gulli, A. and Signorini, A., 2005. The indexable web is more than 11.5 billion pages. *In: Special interest tracks and posters of the 14th international conference on World Wide Web*, 10–14 May 2005, Chiba, Japan: doi:10.1145/1062745.1062789

Hatcher, E. and Gospodnetic, O., 2004. *Lucene in action*. Greenwich, CT: Manning.

Hitwise, 2004. *Hitwise: search queries are getting longer*. Available from: http://www.readwriteweb.com/archives/hitwise_search_queries_are_getting_longer.php] [Accessed 30 March 2012].

ISO19115, 2003. *ISO19115:2003 geographic information-metadata standard*. Available from: http://www.iso.org/iso/catalogue_detail.htm?csnumber = 26020 [Accessed 30 March 2012].

Jones, C.B. *et al.*, 2004. The SPIRIT spatial search engine: architecture, ontologies and spatial indexing. *In*: *Proceedings of 3rd international conferences on geographic information science*, 20–23 October 2004, vol. 3234. MD: Adelphi, 125–139

Li, W., *et al.*, 2011. Semantic-based service chaining for building a virtual Arctic spatial data infrastructure. *Computers & Geosciences*, 37 (11), 1752–1762.

Li, W., Yang, C., and Raskin, R. 2008a. A semantic enhanced model for searching in spatial web portals. *In*: *Proceedings of semantic scientific knowledge integration AAAI/SSKI symposium*, 26–28 March 2008, Palo Alto, CA: Association of American Artificial Intelligence, 47–50

Li, W., Yang, C., and Sun, D., 2009. Mining geophysical parameters through decision-tree analysis to determine correlation with tropical cyclone development. *Computers & Geosciences*, 35 (2), 309–316.

Li, W., Yang, C.W., and Yang, C.J., 2010. An active crawler for discovering geospatial Web services and their distribution pattern-a case study of OGC Web Map Service. *International Journal of Geographical Information Science*, 24 (8), 1127–1147.

Li, W., Yang, C., and Zhou, B., 2008b. Internet-based spatial information retrieval. *Encyclopedia of GIS*, 1, 596–599.

Lovins, J.B., 1968. Development of a stemming algorithm. *Mechanical Translation*, 11 (1–2), 22–31.

Movva, S. *et al.*, 2008. Customizable search engine with semantic and resource aggregation capability. *In*: *Proceedings of the 2008 10th IEEE conference on E-commerce technology and the fifth IEEE conference on enterprise computing, E-commerce and E-services. IEEE Computer Society*, 376–381.

Olsen, L.M. *et al.*, 2007. *NASA/Global Change Master Directory (GCMD) earth science keywords*. Available from: http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html [Accessed 30 March 2012].

Park, L.A.F. and Ramamohanarao, K., 2009. Efficient storage and retrieval of probabilistic latent semantic information for information retrieval. *Vldb Journal*, 18 (1), 141–155.

Raskin, R.G. and Pan, M.J., 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, 31 (9), 1119–1125.

Rauschert, I. *et al.*, 2002. Designing a human-centered, multimodal GIS interface to support emergency management. *In*: *Proceedings of the 10th ACM international symposium on advances in geographic information systems*, 4–9 November 2002, McLean, VA: ACM, 119–124.

Singh, D., 2010. *The biological data scientist. Business, bytes, genes and molecules.* Available from: http://mndoci.com/the-biological-data-scientist/ [Accessed 6 July 2011].

Tan, P.-N., Steinbach, M., and Kumar, V., 2006. *Introduction to data mining*. 1st ed. Boston: Pearson Addison Wesley.

Tisthammer, W.A., 2010. *The nature and philosophy of science.* UFO Evidence, 1386.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46 (2), 234–240.

Xu, G., 1999. Meeting the challenge of "Digital Earth". *Journal of Remote Sensing*, 3 (2), 85–89.

Yang, C., *et al.*, 2008. Distributed geospatial information processing-sharing distributed geospatial resources to support the Digital Earth. *International Journal of Digital Earth*, 1 (3), 259–278.