Computers & Geosciences 106 (2017) 164–170

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

A relevancy algorithm for curating earth science data around phenomenon

Manil Maskey[a,*], Rahul Ramachandran[a], Xiang Li[b], Amanda Weigel[b], Kaylin Bugbee[b], Patrick Gatlin[a], J.J. Miller[b]

[a] NASA, Marshall Space Flight Center, Huntsville, AL, USA
[b] University of Alabama in Huntsville, Huntsville, AL, USA

ARTICLE INFO

Keywords:
Data curation
Search paradigms
Information retrieval
Earth science phenomena
Relevancy algorithm

ABSTRACT

Earth science data are being collected for various science needs and applications, processed using different algorithms at multiple resolutions and coverages, and then archived at different archiving centers for distribution and stewardship causing difficulty in data discovery. Curation, which typically occurs in museums, art galleries, and libraries, is traditionally defined as the process of collecting and organizing information around a common subject matter or a topic of interest. Curating data sets around topics or areas of interest addresses some of the data discovery needs in the field of Earth science, especially for unanticipated users of data. This paper describes a methodology to automate search and selection of data around specific phenomena. Different components of the methodology including the assumptions, the process, and the relevancy ranking algorithm are described. The paper makes two unique contributions to improving data search and discovery capabilities. First, the paper describes a novel methodology developed for automatically curating data around a topic using Earth science metadata records. Second, the methodology has been implemented as a stand-alone web service that is utilized to augment search and usability of data in a variety of tools.

1. Introduction

Earth science domain is no stranger to explosion of data volume and variety. For example, a quick search on data.gov for the term "earth science" returns over 46,000 data collections. Data discovery has become an inherent issue for sites like data.gov, which harvests metadata on all open data from a wide range of federal agencies, state governments, and other organizations within the United States. Earth science data can be, and typically are, used for novel applications by unanticipated users, who must know what and where to search in order to discover relevant data for a specific research investigation or application. This requirement of knowledge on these unanticipated users becomes both difficult and time consuming, and has generated the need for data curation.

Curation, which typically occurs in museums, art galleries, and libraries, is traditionally defined as the process of collecting and organizing information around a common subject matter or a topic of interest. More specifically, the act of searching, selecting, and synthesizing Earth science data/metadata around information from across disciplines and repositories into a single, cohesive, and useful collection has been defined by Ramachandran et al. (2016) as geocuration. For consistency throughout the paper, the term curation will be used to refer to geocuration since the focus of this paper is on Earth science data and information. Curating data sets around topics or areas of interest is a potential solution to improve the data discovery problem, especially for unanticipated users. Curation can be a manual process where the domain experts search, identify, and package the relevant data sets. The Climate Data Initiative (CDI) project, described by Ramachandran et al. (2016), utilized Subject Matter Experts (SMEs) from different federal agencies to manually curate and share data around key climate resiliency themes and openly available climate data from various federal agencies.

However, curation can also be achieved in an automated fashion. In this paper, we present a methodology to automate curation around well-defined topics. The topics of our focus are a specific set of Earth science phenomena. According to the American Meteorological Society (2016), a phenomenon is an observable occurrence of particular physical significance. Instances of specific phenomena (also referred to as events), such as Hurricane Katrina and the volcanic eruption of Chaitén, are of a great interest in Earth science because these events form the basis of case studies. Case studies are scientific investigations that examine the underlying governing dynamical and physical processes that drive the occurrence of a specific event and are a popular scientific research approach within the Earth sciences, Atmospheric science in particular (Schultz, 2009). Curating data around specific phenomenon or events improves Earth scientist's ability to discover data for scientific investigation.

* Corresponding author.
E-mail address: manil.maskey@nasa.gov (M. Maskey).

http://dx.doi.org/10.1016/j.cageo.2017.06.007
Received 9 August 2016; Received in revised form 12 April 2017; Accepted 6 June 2017
Available online 10 June 2017
0098-3004/ Published by Elsevier Ltd.

This paper presents a novel curation methodology that automates search and selection of data around a specific Earth science phenomenon and returns data sets ranked according to their relevancy to the specific phenomenon. This particular methodology contains several components (i.e., assumptions, reference query definition, and relevancy ranking algorithm) and has been implemented as a stand-alone operational web service that can be utilized to augment searches in other tools. Furthermore, the described methodology uses Earth science metadata records to compute relevancy ranking to enhance data search and selection. To our knowledge, such an approach has not been investigated within the field of Earth science.

2. Information retrieval

Information retrieval is defined as the task of finding resources of unstructured nature from a large collection of resources to satisfy an information need (Manning et al., 2008). A typical information retrieval consists of several steps. First, the user identifies a task (e.g., "assess the impact of Hurricane Katrina on coastal shorelines"), which generates an information need (e.g., "find all relevant data sets needed to study Hurricane Katrina") encoded as a query that can be executed by a search engine. Search engine utilizes underlying information retrieval model to analyze the encoded query and returns results for that search. *Note: Encodings of the query will depend on search engines.* In the final step, the user refines the query and reviews the results in an iterative manner until the results satisfy his or her needs.

Two challenges must be addressed while designing an information retrieval system: misformulation, where the user is unable to encode their information need to an effective query, and customization of an information retrieval model for the user's particular application. Forming the right query requires the use of not only correct combinations of keywords, but also domain knowledge, which unanticipated users of data might not have, to obtain the best results. The customization of an information retrieval model depends upon the following: knowing the types of documents in your collection, understanding the documents in your collection, and leveraging domain knowledge to improve relevancy ranking scores.

Providing a mechanism for query expansion is a widely used technique employed in information retrieval to avoid misformulation. Query expansion involves expanding the original query with synonyms in order to improve retrieval performance. Qiu and Frei (1993) proposed a probabilistic query expansion model based on a similarity thesaurus that reflects domain knowledge about the particular collection. In Qiu and Frei's model, queries are expanded by adding terms that are similar to the concept of the query rather than by selecting terms that are similar to the query terms. Ontology-based query expansion is another widely used method (Shamsfard et al., 2006). Bhogal et al. (2007) and Carpineto and Romano (2012) provide the latest review of ontology-based query expansion techniques. More recently, ways to compute similarity between related entities using ontologies have been presented by Zheng et al. (2015). However, knowledge engineering to construct robust ontologies tends to be labor and time intensive.

A number of information retrieval models have been developed in the past, including Boolean retrieval model, vector space model (Turney and Pantel, 2010), and probability retrieval model (Manning et al., 2008; Singhal, 2001). Most search tools available for finding Earth science data use a Boolean retrieval model, wherein a user query is constructed as a Boolean expression of search terms that can be combined with different operators such as AND, OR, and NOT. The returned results are an unranked list of documents where the search terms match and meet the operator criteria. Search tools based on the Boolean retrieval model are useful for expert users with a precise understanding of their needs and of the collection. Users of these search tools must be familiar with not only the data sets, but also how the data sets are represented in the metadata catalog.

Boolean retrieval models are plagued with feast problems—a return of too many results without any ranking—and famine problems—a return of zero results. The feast and famine problems associated with Boolean retrieval models force users either to wade through a very large list of unranked results or to expend time and energy contriving a correct query that will produce sufficient results. Therefore, Boolean retrieval models are not useful for unanticipated users of data where the burden is on the user to formulate the right query attuned to the search tool.

Unlike the Boolean retrieval model where a document is either matched or not matched to the query, the vector space model, introduced by Salton et al. (1975), ranks the returned documents based on document scores, with the most relevant documents appearing at the top of the list. The vector space model approach models a set of documents as vectors in a common vector space, with each dimension defined by the terms (also known as bag-of-words) in the whole document collection. The "document vector" can be in binary form, where its components are prescribed 1 if the term is in the document or 0 if the case is otherwise. A user query comprising of terms of the user's interest is represented as another vector in the vector space. This "query vector" can be constructed with terms of equal weights or of different weights assigned using some quantifiable scheme. The closeness of a document to a query is determined by the similarity measure between query vector and document vector, with scores assigned accordingly.

Cosine similarity is a widely used similarity measure that calculates the angle between query vector and document vector (Manning et al., 2008; Salton and McGill, 1986). The smaller the angle or larger the cosine value, the more similar the document is to the query. The Jaccard coefficient (Kim and Choi, 1999) is another similarity measure, but it accounts for the term overlap between the query vector and document vector normalized by the union of terms in both of them (Manning et al., 2008; Salton and McGill, 1986).

A better approach for presenting the document is to assign weights to vector components—also known as Term Frequency-Inverse Document Frequency (TF-IDF) (Manning et al., 2008; Salton and McGill, 1986). In the TF-IDF weighting scheme, the weight is directly proportional to the frequency in which the term occurs within the document, and indirectly proportional to the popularity of the term, which is determined by the number of documents where the term occurs (Manning et al., 2008).

The effectiveness of an information retrieval system is assessed using two key statistics—precision and recall. Precision indicates the percentage of the returned results that are relevant to the user's information need, while recall indicates the percentage of the relevant documents in the total collection retrieved by the system (Manning et al., 2008). Although a high precision and high recall is the goal of a retrieval system, the gain of one metric often leads to the loss of another.

Information retrieval methods can also be applied to other resources besides metadata text. Specifically, for Earth science, browse images are possible resources, whose image features can characterize underlying data sets. However, Earth science images are published for only limited data sets and without any standardization, making the image features difficult to generalize for retrieval.

We frame the data curation need as a specialized information retrieval problem with a well-defined scope. Since we are targeting a limited set of phenomena, we can address misformulation by using a predetermined set of science keywords identified from a controlled vocabulary using domain knowledge as terms for the query. We designed a customized information retrieval model using our domain knowledge of the document collections, which are the individual records in a metadata catalog. Each metadata record in the catalog contains science keyword annotations from the same controlled vocabulary.

3. Understanding the metadata records

Metadata, which is data about data, plays an integral role to ensure that data can be discovered, navigated, and analyzed. The NASA Earth Science's Common Metadata Repository (CMR) (EOSDIS, 2016a, 2016b, 2016c) is designed as a high performance, high quality, continually evolving metadata system that merges all existing metadata into one source. CMR provides a unified, authoritative repository for NASA's Earth Science metadata. The CMR catalog currently contains metadata for 32,195 data sets and over 300 million files (EOSDIS, 2016a, 2016b, 2016c), and so is a rich resource of information that can be mined for useful information to discern the relevance of a data set for a particular phenomenon.

NASA's CMR is built on a Unified Metadata Model (UMM) (EOSDIS, 2016a, 2016b, 2016c), which is an extensible model that can provide a cross-walk for mapping between CMR-supported metadata standards, such as ISO 19115. The use of the UMM model allows each standard to be mapped centrally to the UMM model rather than mapping CMR-supported metadata standards to each other. This process drastically reduces the need for the number of required translations from n × (n-1) to 2n where n is the number of metadata standard. The UMM describes the metadata related to key concepts (*collection*, *granule*, etc.) for NASA's Earth science data using UMM metadata "Profiles." Each UMM profile is a document that provides a schema-agnostic representation of the elements necessary to provide high quality metadata for its related Earth Observing System Data and Information System (EOSDIS) concept and maps those elements to each CMR-supported metadata standard (EOSDIS, 2016a, 2016b, 2016c).

Our approach exploits the UMM-C profile, or the metadata elements that describe a data collection or data set. The collection-level metadata schema describes the metadata for the whole data set and either requires or recommends certain fields. The required fields include the "data set short name" and "long name" and a "description" for the data set, while the recommended fields include spatial and temporal resolutions and extents and science keywords to describe the data set. The long name is the reference name used to describe the scientific contents of the data collection. The description field allows data providers to describe, in detail, the content of the data collection. Both long name and description fields are usually in free-text format. The science keywords describe the contents of the data set as defined by the Global Change Master Directory (GCMD) vocabulary (GCMD, 2016). The GCMD controlled science keyword vocabularies allow metadata to be described in a consistent manner and enable precise searching of metadata records and subsequent retrieval of data and services. The GCMD vocabulary is constructed into seven facets, Earth Science, Data Services, Data Centers, Locations, Instrument/Sensors, Platforms/Sources, and Projects, with each facet represented as a taxonomy working from a general concept at the root toward specialized concepts at the leaf. GCMD keywords are consequently organized into five hierarchical levels from Topic to Term, along with three variable levels. These GCMD Earth science keywords describe physical variables, such as temperature, wind, water, radiation, and aerosols, that are considered relevant to the phenomena. An example of a GCMD Earth Science keyword is "Atmosphere > Aerosols > Aerosol Optical Depth/Thickness > Angstrom Exponent."

Our methodology uses the GCMD Earth Science keywords, long name, and description fields to determine the relevancy-rankings. We further utilize the GCMD science keywords to define a phenomenon. The method presented in this paper is built on a certain set of assumptions, which are as follows:

- The GCMD vocabulary is complete enough for such use and has the proper granularity to comprehensively characterize an Earth science phenomenon.
- The metadata records stored in the CMR catalog are consistent, correct, and complete. Specifically, the metadata description long name and keywords fields have consistent, correct, and complete metadata values (i.e., the GCMD vocabulary is used properly with each record having the correct annotation and the correct granularity; the GCMD vocabulary is used consistently across all records from different data providers).

4. Methodology

As stated earlier, data curation for a given set of phenomena can be framed as a specialized information retrieval problem with a well-defined scope. We address the misformulation issue by using a predetermined set of terms for the query. Domain knowledge is used to identify science keywords from the GCMD controlled vocabulary. We also utilize our knowledge and expertise of the metadata records in the CMR catalog to design a custom information retrieval model.

4.1. Defining reference queries for different phenomena

We tasked three Earth science experts to select a relevant subset of GCMD science keywords from version 6.0 to describe a specific phenomenon. Hurricane, volcanic eruption, flood, and fire were selected as the initial set of phenomena based on the Earth science expertise deeming these phenomena most monitored by NASA Earth Observing Systems. Earth science keywords selected by the different experts for each phenomenon were aggregated to construct the bag-of-words set to serve as the reference query. These keywords are considered equally important with regard to ranking collection-level science keyword metadata. Another set of keywords and/or phrases, each corresponding to the word or phrase in the five hierarchical levels of the keywords, was generated from these Earth science keywords. The generated keyword set is referred to as a "free-text keyword set" in order to distinguish it from the GCMD science keyword set. The free-text keyword set was used to rank the long name and description metadata (described further in this paper). Weights were assigned to each selected keyword based on its depth level within the taxonomy. The weight of 0.2 was assigned to the topic (root) level of the GCMD Earth science keyword, 0.4 to the term level keyword, and weights of 0.6, 0.8, and 1.0 were assigned to keywords at variable levels 1, 2, and 3, respectively. Higher weights imply higher specificity; therefore, the keywords with higher weights serve as a better discriminator. *Note: Even though our initial approach only considered four phenomena, the same approach can be extended to other Earth science phenomena, such as earthquake and landslide. Domain experts for such phenomena will need to select appropriate Earth science keywords from GCMD as the bag-of-words.*

4.2. Vector space model for science keyword fields

A vector space model was used to rank each Earth science keyword field (denoted as s) in a metadata collection. Assuming k is the number of GCMD Earth science keywords identified by domain experts, the vector space is a k-dimensional space with each dimension being one of the k Earth science keywords.

We denote via $\boldsymbol{V}(c_s)$ the vector derived from an Earth science keyword field of a collection metadata c, represented as follows:

$$\boldsymbol{V}(C_s) = [c_1, c_2, \ldots, c_k], \tag{1}$$

where $c_i = 1$ if the keyword field of a collection metadata c contains the i^{th} keyword, or $c_i = 0$ if no keywords are present in the record.

Similarly, the reference query vector for a phenomenon is represented as follows:

$$\boldsymbol{V}(q_s) = [q_1, q_2, \ldots, q_k] or [1, 1, \ldots, 1], \tag{2}$$

Since phenomenon-relevant keywords are used only once, $q_i = 1$ (where $i = 1 to\ k$).

In vector space model document retrieval, all keywords in a document are not treated equally important for relevancy. A weight based on the $TF-IDF$ scheme is often assigned to a keyword t in scoring the document relevancy. The $TF-IDF$ weight of a t is defined as follows:

$$TF\text{–}IDF(t) = TF(t)\bullet IDF(t), \tag{3}$$

where $TF(t)$ is the number of occurrences where term t appears in a document. The more frequently t appears in a document, the more weight t is assigned.

Additionally,

$$IDF(t) = \log(N/DF(t)),$$

where N is the total number of documents in the document set and $DF(t)$ is the number of documents in the set that contain t. If a rare t appears in documents, the more unique t is to the document and thus more weight is assigned to t.

In our metadata records, unique keywords can occur only once per record; therefore, $TF(t) = 1$ for all t.

So,

$$TF\text{–}IDF(t) = IDF(t).$$

The IDF values are calculated for all Earth science keywords in all of the metadata records. As a result, the modified document vector is as follows:

$$\boldsymbol{V}_m(c_s) = [c_1\bullet IDF_1,\ c_2\bullet IDF_2, \ldots,\ c_N\bullet IDF_N],\ , \tag{4}$$

where $c_i = 1$ if the keyword field of a collection metadata c contains the i^{th} keyword, or $c_i = 0$ if no keywords are present in the record.

Similarly, the query vector is represented as

$$V_m(q_s) = [IDF_1,\ IDF_2, \ldots,\ IDF_k], \tag{5},$$

since $q_i = 1$ for i = for 1, k in (2).

4.3. Vector space model for long name (title) and description

The long name field in the metadata record provides a descriptive title for the data set. The document vector for a long name is defined as follows:

$$\boldsymbol{V}(c_l) = [c_1\bullet IDF_1\bullet w_1,\ c_2\bullet IDF_2\bullet w_2, \ldots,\ c_N\bullet IDF_N\bullet w_N],$$

where c_i is the number of occurrences of term i in the long name field, IDF_i is the inverse document frequency of term i in the long name field of all collection metadata, and w_i is the weight assigned to the term in the free-text keyword set.

Correspondingly, the query vector $\boldsymbol{V}(q_l)$ for the long name field and $\boldsymbol{V}(q_d)$ for the description field are defined as follows:

$$\boldsymbol{V}(q_l) = \boldsymbol{V}(q_d) = [IDF_1\bullet w_1,\ IDF_2\bullet w_2, \ldots,\ IDF_N\bullet w_N].$$

The description field, which is a free text field, in the metadata record provides additional information about the data set. The document vector for the description field is defined as follows:

$$\boldsymbol{V}(c_d) = [c_1\bullet IDF_1\bullet w_1,\ c_2\bullet IDF_2\bullet w_2, \ldots,\ c_N\bullet IDF_N\bullet w_N],$$

where c_i is the number of occurrences of term i in the description field, IDF_i is the inverse document frequency of term i in the description field of all collection metadata, and w_i is the weight assigned to the term in the free-text keyword set.

4.4. Similarity measures

Ranking of the metadata record is computed using two commonly used similarity metrics: Jaccard coefficient and Cosine similarity. Jaccard coefficient, a similarity measure between two data sets, is defined as the size of the intersection divided by the size of the union of the two data sets.

For document vector $\boldsymbol{V}(c_s)$ and query vector $\boldsymbol{V}(q_s)$ in (1) and (2), the Jaccard coefficient for metadata record c is defined as follows:

$$Jaccard(c) = |\boldsymbol{V}(c_s)\cap\boldsymbol{V}(q_s)|/|\boldsymbol{V}(c_s)\ \cup\ \boldsymbol{V}(q_s)|, \tag{6}$$

where $\bigcap$ indicates set intersection and $\bigcup$ indicates set union.

Cosine similarity of data collection c using Earth science keyword metadata is defined as follows:

$$CosSim(c_s) = \boldsymbol{V_m}(c_s)\bullet\boldsymbol{V_m}(q_s)/|\boldsymbol{V_m}(c_s)\bullet\boldsymbol{V_m}(q_s)|, \tag{7}$$

where $\boldsymbol{V_m}(c_s)\bullet\boldsymbol{V_m}(q_s)$ is the inner product of the document vector and the query vector and the denominator, $|\boldsymbol{V_m}(c_s)\bullet\boldsymbol{V_m}(q_s)|$, is the product of their Euclidean lengths.

$CosSim(c_s)$ is also referred to as $S_c(s)$, the similarity score from the science keyword field. Similarly $S_c(l)$, the similarity score from the long name field, and $S_c(d)$, the similarity score from the description field, are calculated using Eq. (7), where $\boldsymbol{V_m}(c_s)$ and $\boldsymbol{V_m}(q_s)$ are replaced with $\boldsymbol{V}(c_l)$, $\boldsymbol{V}(q_l)$ and $\boldsymbol{V}(c_d)$, $\boldsymbol{V}(q_d)$, respectively.

We calculated the Jaccard coefficient and Cosine similarity for each collection metadata and then ranked these collections in decreasing order. The metadata records with larger values appeared first on the list and are considered the most relevant to the phenomenon of interest.

4.5. Weighted zone ranking (ensemble approach)

The algorithm generated scores for each collection metadata record using three fields: Earth science keyword, long name, and description. We combined all three scores and generated an overall score, known as the ensemble score, for each collection record. Using the zone ranking approach, we defined the ensemble score, $S_c(e)$, for a collection c as a linear combination of the three individual scores, as defined in the following:

$$S_c(e) = w_s\bullet S_c(s)+w_l\bullet S_c(l)+w_d\bullet S_c(d),$$

where $S_c(s)$, $S_c(l)$, and $S_c(d)$ are the similarity measure values from the Earth science keyword field, long name field, and description field of collection c, respectively. w_s, w_l, and w_d are corresponding weights for the three metrics with the sum of w_s, w_l, and w_d equaling 1.

Fig. 1 illustrates graphical overview of our methodology for a phenomenon.

5. Results

Next, we describe our experiments and results.

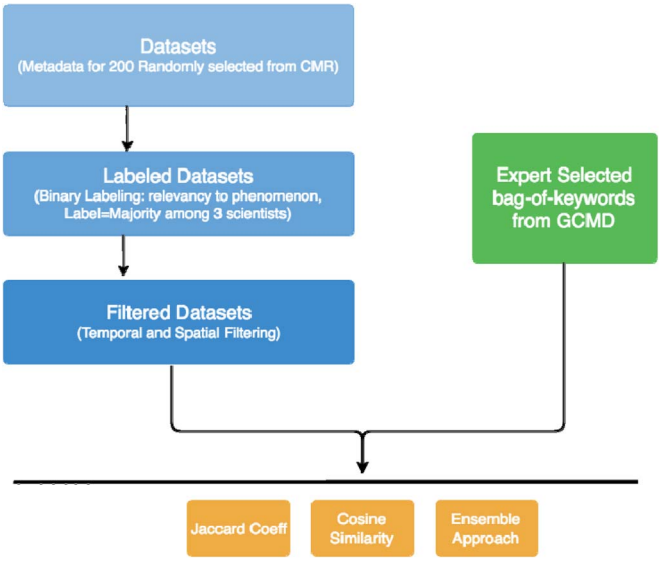

Fig. 1. Diagrammatic overview of relevancy ranking approach for a phenomenon.

5.1. Experiment setup

The methodology explained in Section 4 was tested for ranking data sets for the four phenomena. For each phenomenon, 200 metadata records describing data sets were randomly selected from the catalog to create a truth set. Since various phenomenon occur during different seasonal time frames, transpire over a specific duration of time, and exist within certain geographical areas, data sets can be filtered using heuristics based on phenomenon characteristics. For instance, hurricanes originate in tropical ocean regions and have been found to have a life cycle of up to two to three weeks (University of Illinois Urbana-Champaign, 2010). Filtering data sets using domain knowledge about the phenomenon helps remove irrelevant data sets from the ranking process. For hurricanes, data sets with a temporal resolution larger than 'daily' were removed. The remaining metadata records for the data sets in the truth set were then labeled by three domain experts to be either "relevant" or "not relevant" to a specific phenomenon. The final relevancy label was determined by a majority of votes from the three experts.

5.2. Comparison of similarity measures

First, the ranking performances of the Jaccard coefficient and Cosine similarity metrics on Earth science keyword metadata were compared for top 10, 20, and 30 collection returns. The amount of relevant collections returned for hurricane and volcanic eruption are shown in the table below.

The results presented in Table 1 suggest that both of the measures—the Jaccard coefficient and the Cosine similarity—performed similarly. However, we selected Cosine similarity measure as the similarity metric for relevancy ranking because it is commonly used in space vector model information retrieval.

5.3. Data curation results

Since we are using three fields from within the metadata records—Earth science keywords, long name (title), and description—we needed to assign weights to the similarity measure computed for each of these fields per Section 4.5. We calculated these weights by optimizing precision and changing each weight (w_s, w_l, w_d) from 0.0 to 1.0 in increments of 0.1.

Precision and recall are two of the most often used performance metrics for document retrieval. Precision is the fraction of the retrieved documents that are relevant, whereas recall is the fraction of relevant documents that are retrieved. For a document set that contains a total of N documents in which M documents are relevant, when the query returns n documents out of m relevant documents, precision equals n/m. When relevant documents m are returned from M relevant documents, recall is computed as m/M. For an optimal retrieval system, both precision and recall are high. Since more than one combination of weight sets may produce the same precision value, we utilized tie-breaking measure T defined in the equation below:

$$T = \sum_{i=1}^{n} \frac{S_i}{i},$$

where S_i is the status of returned document i. S_i equals 1 if document i is relevant and 0 if otherwise. Therefore, for a set of weights that has equal precision value, the optimal weight is the one that maximizes the T value.

Ranking results for the top 20 returns using the ensemble method with an optimal weight and with equal weight were compared against a random selection of data sets for all four phenomena. It is assumed that the top 20 data sets should satisfy most users' data search needs.

Based on the experimental setup, two factors should be considered while analyzing the results. First, there are different amounts of "relevant" data sets within each truth set for each phenomenon. The precision values from random selection reflect this variation. Over 60% of volcanic eruption and fire data sets in the truth set were relevant, while only 35% of flood data sets were relevant. For hurricane, the amount of relevant and not relevant data sets in the truth set was roughly equal. Second, the recall values from the random selection depend upon the collection size. While there are 40 data sets for volcanic eruption, there are over 70 data sets for hurricane, fire, and flood. As a result, the recall value for volcanic eruption was 50% (20/40) for 20 returns while the recall value for hurricane was 28.6% (20/70).

It is therefore better to compare the curation results against a random selection rather than compare the performance of the methods for each phenomenon against each other.

Based on the results presented in Table 2 below, precision when using optimal weights is 5% better than when using equal weights, and the recall values are about 3.5% better on average. More importantly, when comparing the results of the ensemble method using optimal weights to the results of random returns, precision values improved by 35%, 22%, 11%, and 30% for hurricane, volcanic eruption, fire, and flood, respectively, and recall values improved by 19%, 18%, 4%, and 22%, respectively. On average, precision improves about 25% when using our method and recall improves about 16%.

Analyzing the retrieval performance for specific phenomena, the results for fire are lower than those for hurricane, volcanic eruption, and flood. The quality of the metadata records may partly contribute to these differences.

The top 20 return results for each phenomenon are shown in Figs. 2–5. Each figure displays the precision improvement between our method and the random selection's results (denoted by a dotted line). For hurricane events, precision is 100% when recall reaches 0.45, suggesting that the top 17 returns are relevant (0.45 × 38, where 38 is the total number of relevant data sets). For flood events, precision is low when the recall value is small, and improves with increasing recall values. This correlation is caused by the first data set returned being "not relevant" in addition to 3 out of the 5 top returns being "not relevant".

We evaluated the contribution of the three fields of the metadata records to the ranking algorithm based on the weight distributions. These results are presented in Table 3 below. On average, the weight for Earth science keyword is largest when the weight for description is smallest. This relationship is expected since the Earth science keywords

Table 1
Similarity measures results for hurricane and volcanic eruption.

	Hurricane		Volcanic eruption	
	Jaccard Coefficient	Cosine Similarity	Jaccard Coefficient	Cosine Similarity
Top 10 retrieval	10	9	6	7
Top 20 retrieval	17	16	15	15
Top 30 retrieval	23	24	22	21

Table 2
Ranking results from top 20 returns using the ensemble method.

	Optimal weight		Equal weight		Random	
	Precision	Recall	Precision	Recall	Precision	Recall
Hurricane	90.0%	47.4%	85.0%	44.7%	54.3%	28.6%
Volcanic eruption	85.0%	68.0%	80.0%	64.0%	62.5%	50.0%
Fire	75.0%	30.0%	75.0%	30.0%	64.1%	25.6%
Flood	65.0%	48.1%	55.0%	40.7%	35.5%	26.3%

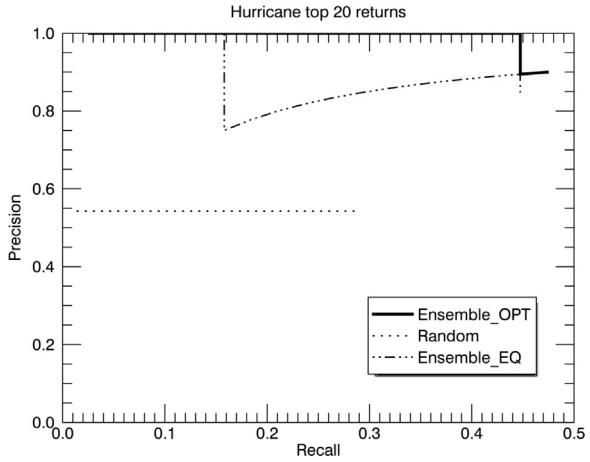

Fig. 2. Precision and Recall plot for Hurricane; top 20 results. The chart shows the precision improvement between our method and the random selection results (denoted by a dotted line).

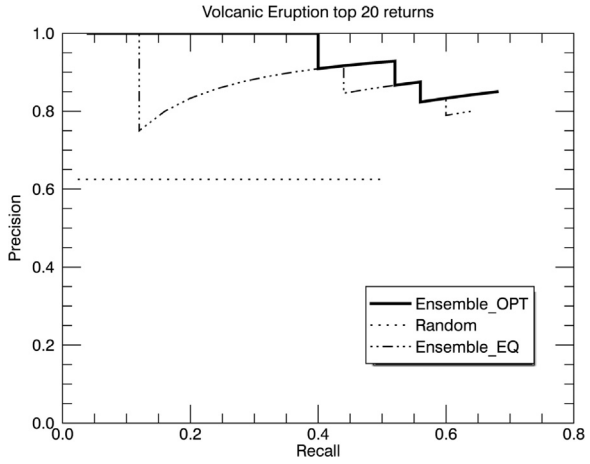

Fig. 3. Precision and Recall plot for Volcanic Eruptions; top 20 results. The chart shows the precision improvement between our method and the random selection results (denoted by a dotted line).

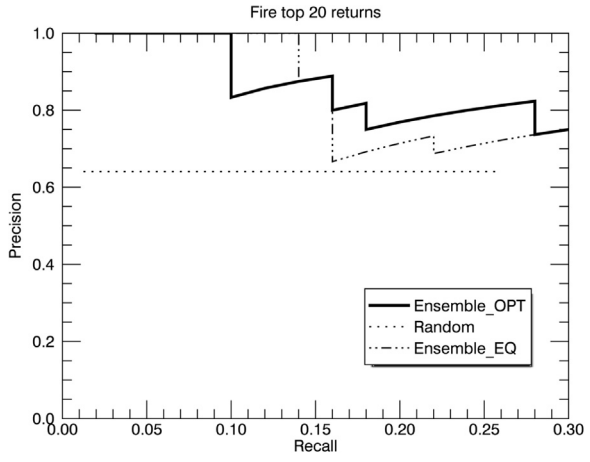

Fig. 4. Precision and Recall plot for Fire; top 20 results. The chart shows the precision improvement between our method and the random selection results (denoted by a dotted line).

metadata fields, which use a controlled vocabulary, are accurate and consistent in describing data product while the description field is free-text and has the most variability in quality.

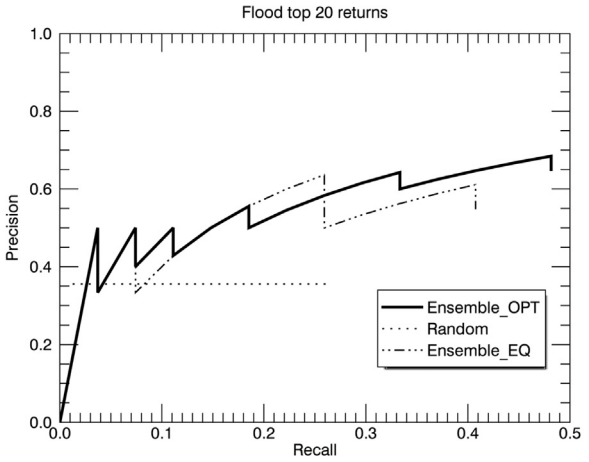

Fig. 5. Precision and Recall plot for Flood; top 20 results. The chart shows the precision improvement between our method and the random selection results (denoted by a dotted line).

Table 3
Optimal ensemble weights for each phenomenon.

	Optimal Weight Set ($W_{sciencekeyword}$, $W_{longname}$, $W_{description}$)
Hurricane	(0.6, 0.1, 0.3)
Volcanic eruption	(0.2, 0.6, 0.2)
Fire	(0.6, 0.2, 0.2)
Flood	(0.5, 0.4, 0.1)

5.4. Web service implementation

The relevancy ranking algorithm has been implemented as a web service. The web service follows the REST (Representational State Transfer) architectural style and is implemented using Java Spring framework. Results from relevancy ranking algorithm are read and converted into JSON encoding. The web service accepts phenomena type as a request parameter and returns data set name, relevancy score, version, data set shortname, and processing level in descending order of relevance. Any other metadata elements can be added to the response using a configuration in server side. The service can be utilized by other search services or analysis applications and can be accessed here for hurricane: http://34.192.255.219:8080/ecstest/relevancy?type=hurricane&bbox=-180,-90,180,90&starttime=2010-10-20&stoptime=2010-10-30.

6. Discussion

Our proposed methodology has both strengths and limitations—each discussed below.

6.1. Strengths

6.1.1. Approach is data driven

Unlike other domain ontology-driven approaches that are top-down, our methodology uses a data (metadata) driven approach. The curation methodology was developed after analyzing the metadata schema and the existing records. Both the reference query and the relevancy ranking algorithm are dependent on the controlled vocabulary used in the metadata records and the specific fields used in our method.

6.1.2. Construction of reference query is simple

The methodology defines a reference query using a controlled vocabulary. This approach is effective for a search tasks that are well scoped such as data discovery for a specific phenomenon. Furthermore, defining a

reference query using a controlled vocabulary is simpler and less labor intensive than trying to knowledge engineer a formal ontology.

6.1.3. Methodology is scalable

Our approach is scalable to the addition of new records in the metadata catalog and does not require any modifications since new metadata records that are added to the catalog all utilize a controlled vocabulary.

6.2. Limitations

6.2.1. Modeling the search intent is difficult

It is difficult to predetermine the exact information need of a user. For instance, one user may be interested in only a specific aspect of a phenomenon (e.g., flooding caused by a hurricane) whereas another user may only be interested in studying a unique characteristic of a phenomenon (e.g., hurricane intensification). These two users' data needs are going to be different. We mitigate this issue by ensuring that the reference query for a specific phenomenon is broad and covers all possible relevant keywords. While this may not provide the exact results for a specific user, it does substantially reduce the list of data for the search results. Furthermore, additional facets such as application areas can be added to the reference query to improve search results.

6.2.2. Quality of metadata records is variable

One of our key assumptions is that the metadata records stored in the CMR catalog are consistent, correct, and complete. More specifically, we assumed the following:

- The metadata description and science keywords fields are complete
- The GCMD vocabulary is used correctly to annotate each metadata record
- The correct granularity is used in each metadata record
- The GCMD vocabulary is used consistently across all records by different data providers

Our initial analysis of the methodology results showed that part of our assumptions to be incorrect. Mainly, we observed incomplete metadata for problematic data sets. Also, the quality of metadata records in the CMR is variable in terms of consistency. To address this limitation, we have launched a new project to improve NASA's Earth science metadata record quality in the CMR catalog. The new project seeks to address all of the critical metadata quality issues uncovered so far.

6.2.3. Dependency on the controlled vocabulary

Ranking results from our methodology depend on two aspects of the controlled vocabulary: its richness in detail and future changes. A rich, detailed, and controlled vocabulary provides a better level of annotation granularity to represent different phenomena and helps disambiguate data sets. Whereas the use of a poor controlled vocabulary will limit it usefulness. Also, any major changes to the controlled vocabulary will carry a substantial impact on our methodology and will require reformulation of the reference queries.

6.2.4. Truth set labels may be biased

There may be labeling bias in the truth sets created by the domain experts. The Earth science domain experts on our team have stronger expertise in certain areas, such as hurricanes, and weaker expertise in others, such as floods and fire. This bias is possibly reflected in the overall results of the methodology. We plan to expand the pool of domain experts to assist in both defining reference queries and labeling truth data to improve the relevancy ranking results.

7. Summary

Curating data sets around topics or areas of interest solves the data discovery problem in the field of Earth science, especially for unanticipated users. Towards that end, this paper provides methodology in building a relevancy ranking-based Earth science data curation service around phenomena. Applications of the service for various Earth science phenomena are also presented.

As part of our future work, we plan to expand the algorithm to encompass the variable levels stored within data files (granules) instead of remaining at just the data set level. We designed an initial algorithm for this problem and it is currently being tested. We also plan to expand our approach beyond using the metadata records and plan to incorporate information from journal publications. One approach being considered is to construct graphs linking information extracted from publications along with the information stored in the metadata catalog. These graphs can be used to develop relevancy ranking algorithms to improve curation results. To address the misformulation problem, we also plan to explore auto-generating reference queries for topics by mining selected papers.

Acknowledgements

The authors would like to acknowledge Mike Little, Nikunj Oza, and the NASA ESTO team for providing technical guidance. This project was funded by a NASA ESTO AIST grant. The authors would like to acknowledge other members on this project for their contributions: Steve Kempler, Chung-Lin Shie, Suhung ShenSu Hang, Maksym Petrenko (NASA/GSFC) and Peter Fox, Stephan Zednik, Anirudh Prabu (RPI). The authors are also grateful to Chris Lynnes (NASA/GSFC) for his insights and suggestions to improve the methodology. Finally, the authors are grateful to Kala Golden for editing and improving the manuscript.

References

American Meteorological Society (AMS), 2016. 'Phenomenon'. Glossary of Meteorology. August (Online). ⟨http://glossary.ametsoc.org/wiki/Main_Page⟩ (Accessed 03 August 2016).

Bhogal, J., Macfarlane, A., Smith, P., 2007. A review of ontology based query expansion. Inf. Process. Manag. 43 (4), 866–886.

Carpineto, C., Romano, G., 2012. A survey of automatic query expansion in information retrieval. ACM Comput. Surv. 44 (1), 1–50.

Earth Observing System Data and Information System (EOSDIS), 2016a. NASA's Common Metadata Repository, August, (Online). ⟨https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository⟩ (Accessed 03 August 2016).

Earth Observing System Data and Information System (EOSDIS), 2016b. Unified Metadata Model (UMM), August (Online). ⟨https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm⟩ (Accessed 03 August 2016).

Earth Observing System Data and Information System, (EOSDIS), 2016c. CMR Search, August (Online). ⟨https://cmr.earthdata.nasa.gov/search/⟩ (Accessed 03 August 2016).

Global Change Master Directory (GCMD), 2016. Discover Earth Science Data and Services, August (Online). ⟨http://gcmd.nasa.gov/index.html⟩ (Accessed 03 August 2016).

Kim, M., Choi, K., 1999. A Comparison of collocation-based similarity measures in query expansion. Inf. Process. Manag. 35 (1), 19–30.

Manning, C., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, 2008.

Qiu, Y., Frei, H., 1993. Concept based query expansion. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160–169.

Ramachandran, R., Bugbee, K., Tilmes, C., Privette, A., 2016. Climate data initiative. Comput. Geosci. 88 (C), 22–29.

Salton, G., McGill, M., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc, New York.

Salton, G., Wong, A., Yang, C., 1975. A vector space model for automatic indexing. Commun. ACM 18 (11), 613–620.

Schultz, D., 2009. Eloquent Science: A Practical Guide to Becoming a Better Writer, Speaker, and Atmospheric Scientist. American Meteorological Society, Boston.

Shamsfard, M., Nematzadeh, A., Motiee, S., 2006. Orank: an ontology based system for ranking documents. Int. J. Comput. Sci. 1 (3), 225–231.

Singhal, A., 2001. Modern information retrieval: a brief overview. IEEE Data Eng. Bull. 24, 35–43.

Turney, P., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37 (1), 141–188.

Zheng, J., Fu, L., Ma, X., Fox, P., 2015. SEM+: tool for discovering concept mapping in Earth science related domain. Earth Sci. Inform. 8 (1), 95–102. http://dx.doi.org/10.1007/s12145-014-0203-1.