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.