Geophysical Research Abstracts
Vol. 18, EGU2016-9664, 2016
EGU General Assembly 2016

Mining dark information resources to develop new informatics capabilities to support science

Rahul Ramachandran (1), Manil Maskey (2), and Kaylin Bugbee (2)
(1) NASA, (2) University of Alabama in Huntsville

Dark information resources are digital resources that organizations collect, process, and store for regular business or operational activities but fail to realize their potential for other purposes. The challenge for any organization is to recognize, identify and effectively exploit these dark information stores. Metadata catalogs at different data centers store dark information resources consisting of structured information, free form descriptions of data and browse images. These information resources are never fully exploited beyond a few fields used for search and discovery. For example, the NASA Earth science catalog holds greater than 6000 data collections, 127 million records for individual files and 67 million browse images. We believe that the information contained in the metadata catalogs and the browse images can be utilized beyond their original design intent to provide new data discovery and exploration pathways to support science and education communities.

In this paper we present two research applications using information stored in the metadata catalog in a completely novel way. The first application is designing a data curation service. The objective of the data curation service is to augment the existing data search capabilities. Given a specific atmospheric phenomenon, the data curation service returns the user a ranked list of relevant data sets. Different fields in the metadata records including textual descriptions are mined. A specialized relevancy ranking algorithm has been developed that uses a “bag of words” to define phenomena along with an ensemble of known approaches such as the Jaccard Coefficient, Cosine Similarity and Zone ranking to rank the data sets. This approach is also extended to map from the data set level to data file variable level. The second application is focused on providing a service where a user can search and discover browse images containing specific phenomena from the vast catalog. This service will aid researchers in uncovering interesting event in the data for case study analysis. The challenge of this second application is to bridge the semantic gap between the low level image pixel values and the semantic concept perceived by a user when he or she sees an image. A deep learning algorithm, specifically the Convolution Neural Network (CNN), has been trained and tested to identify three types of Earth science phenomena – Hurricanes, Dust, and Smoke/Haze in MODIS imagery.

Latest results from both the applications will be presented in this paper.