EXPLOITING DARK INFORMATION RESOURCES TO CREATE NEW VALUE ADDED SERVICES TO STUDY EARTH SCIENCE PHENOMENA

Rahul Ramachandran[+], *Manil Maskey*[*], *Xiang Li*[*], *Kaylin Bugbee*[*]

+ NASA Marshall Space Flight Center, * University of Alabama in Huntsville

1. INTRODUCTION

Organizations routinely collect, process, and store information resources for regular business or operational activities. However, most of these organizations fail to recognize, identify and effectively utilize these "dark information resources" for other purposes. Metadata catalogs at data centers are an example of a dark information resource. These catalogs contain structured information, free form descriptions of data, and browse images that are seldom used beyond enabling traditional query based searches on fields. For example, the NASA Earth science metadata catalog holds more than 6000 data collections, over 127 million metadata records for individual data files and browse images surpassing 67 million in number. The information stored in these metadata catalogs 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 will present two research applications exploiting these metadata resources to provide improved data discovery and exploration capabilities. The first application is a data curation service that exploits the metadata catalog records and recommends to users the relevant data sets and data variables useful in studying a particular phenomenon. The second application is an image retrieval service that mines browse imageries. This image retrieval service allows users to to select a type of phenomena and then retrieves images from the catalog containing the desired phenomenon. The image retrieval service can be used by researchers to discover new, previously unknown occurrences of a phenomenon type for case study analysis.

2. DATA CURATION SERVICE

The goal of this research application is to design and develop a stand alone data curation service. A user or program can invoke this service, designate a specific type of Earth science phenomena such as a hurricane, and receive a list of relevant data sets and pertinent variables within each data file.

A traditional search process consists of a user task that requires specific information that is then mapped to a search query. The user's interaction with the search engine is iterative, where the user refines the query based on each search result until his or her needs are met. In our approach, data curation is framed as a specialized, well scoped search problem where the set of phenomena (i.e. search query) can be predefined using domain expertise. A phenomenon is represented via a "bag of words" using a controlled vocabulary. A specialized relevancy ranking algorithm is designed to find data set matches based on the science keywords using different similarity measures and approaches such as Jaccard Coefficient and Cosine Similarity [1], matches based on the data set name and collection as well as a weighted combination of all called the Ensemble approach. A set of experiments were

conducted to test this approach using two different phenomena – hurricanes and volcanic eruptions. These phenomena were defined by experts as a "bag of words" using NASA Global Climate Change Directory (GCMD) vocabulary [2]. GCMD keywords are used in the metadata record to annotate several fields. Two hundred metadata records were randomly selected and were labeled as either relevant or not-relevant for a particular phenomenon by three domain experts. The final label was determined by a majority vote. The data sets were filtered spatially and by time, based on the known climatological information of a phenomenon. The data sets ranking results using different similarity measures and schemes were compared against the expert label. The best results are returned by using the Ensemble approach. Figure 1 shows the precision and recall plot from the top 20 hurricane data set search returns using one weight set. Of the top 20 data set returns, 17 of them are labeled as relevant. Hurricane data sets are retrieved with a precision of 1 at recall of 0.88. The results for volcanic eruptions are lower with a precision of 0.95 and a recall of 0.7.

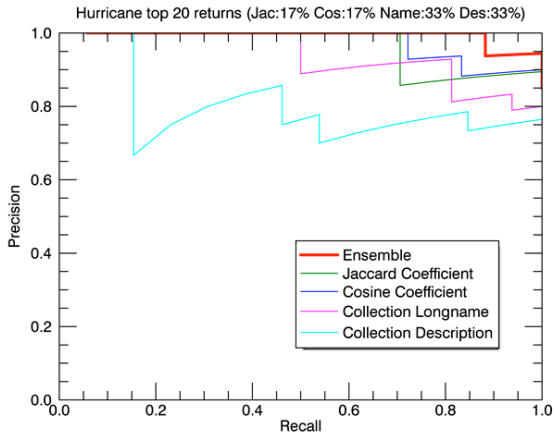

Figure 1: Precision vs. recall for ranking algorithms and schemes. A heuristically weighted Ensemble approach gives the best results

After the relevant data sets have been identified, the next step is to find relevant data variables within the actual files as most data files contain many variables and not all of them are germane for studying a specific phenomenon. The relevancy ranking approach is extended to the data set level keywords which are used as the reference search query. Science key words are extracted from the data set level metadata records, and processed where special characters are removed and words are tokenized. The processed keywords are then normalized by using text processing methods including stemming, lemmatization, and the removal of stop words. Similarly, data variables and their descriptions are extracted from the data files as keywords. These keywords are then processed and normalized. A look up table is created for handling acronyms/abbreviations for variable names. The intersection of keywords from data sets and data files are used for the final mapping and to identify the relevant variables.

3. IMAGE RETRIEVAL SERVICE

The image retrieval service allows researchers to find unknown occurrences of a specific phenomenon by simply using the browse imagery. The identified occurrences of a phenomenon can potentially be useful for research focused on 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.

Traditional machine learning approaches for image retrieval utilize user-defined image features. Our empirical experimental results show that the user-defined features perform fairly well on characterizing images with a specific phenomenon; however, no particular feature is able to perform satisfactorily for all phenomena. In addition, domain experts are required to characterize the image features manually. Thus, we investigated evolving deep learning techniques for image retrieval where the algorithms auto construct the image features most suitable for the

retrieval. Towards that end, we provide initial experimental results in this paper.

Advancements in deep learning shed light on narrowing the aforementioned semantic gap by learning high-level image representation from raw pixels. Deep learning mimics the human brain which is organized in a deep architecture and processes information through multiple stages of transformation and representation. One of the major advantages of deep learning algorithms is that they learn complex functions that directly map pixels to outputs without relying on human-crafted features. Deep learning techniques, specifically, Convolution Neural Network (CNN), have achieved great success on image classification and retrieval tasks [4]. In general, CNN, consists of two parts: (i) convolution layers and max--pooling layers, and (ii) fully connected layers and the output layers [3]. CNNs are configured for desired accuracy and computation speed by varying the depth and breadth of the network, learning rate, and convolution filter sizes. CNNs have fewer connections and parameters when compared to standard feed-forward neural networks.

For our experiments, MODIS browse imagery from NASA's Rapid Response website [http://rapidfire.sci.gsfc.nasa.gov/realtime] were used. In total, about 900 images were collected and labeled into four categories by experts. Three of the categories represent a specific Earth science type phenomenon namely hurricanes, dust and smoke/haze. The fourth is the "other" category where images that appeared visually similar to three phenomena were selected. We further increased the number of labeled sets of images to close to 5000 by using various transformations as mentioned in [3]. 70% of the images for each category were used for training the CNN and the remaining 30% were used for testing accuracy. To date, the best results are obtained from using a 7 layer CNN with a learning rate of 0.003.

True\Predicted	Others	Dust	Smoke	Hurricane	Row Total
Others	**106**	122	37	11	276
Smoke	38	**291**	21	10	360
Dust	38	32	**371**	35	476
Hurricane	14	0	11	**365**	390
Column Total	196	445	440	421	**1502**

Overall Accuracy = 1133/1502 = **75%**

Producer's Accuracy
Others = 106/276 = 38%
Smoke/Haze = 291/360 = 81%
Dust = 371/476 = 78%
Hurricane = 365/390 = 94%

User's Accuracy
Others = 106/196 = 54%
Smoke/Haze = 291/445 = **65%**
Dust = 371/440 = **84%**
Hurricane = 365/421 = **87%**

Figure 2: Confusion Matrix showing the CNN results

The overall accuracy of CNN for this application is around 75%. The user's accuracy is important since the algorithm will be used as an image retrieval service. The algorithm correctly retrieves images for hurricanes 87%, dust 84% and smoke/haze 65% of the time.

4. SUMMARY

This paper presents two research applications exploiting unused metadata resources in novel ways to aid data discovery and exploration capabilities. The results based on the experiments are encouraging and each application has the potential to serve as a useful standalone component or service in a data system.

There were also some interesting lessons learned while designing the two applications and these are presented next.

Data Curation Service

There are two assumptions in the approach used. The first assumption is that the catalog is rich, complete, and that most metadata records have proper tags with the appropriate vocabulary terms. However, one of the first obstacles uncovered while designing and testing the data curation service is that this is assumption is optimistically flawed. Even the best metadata records have errors and gaps. There

are large inconsistencies in tagging especially when one compares the data set tags to the data file variables. However, the data curation algorithm can be modified and used to check the quality of the metadata records. The algorithm can identify inconsistencies and suggest keywords that a data curator at a data center may not consider when creating the metadata entry. The second assumption in the approach is that a phenomenon can be broadly defined by a bag of keywords using vocabulary terms and that this broad definition would cover the search intent of most users. However, this approach cannot distinguish nuances in the search intent of users interested in a phenomenon. For example, one researcher may be interested in hurricanes with a focus on finding the cause for hurricane intensification while another researcher may also be interested in hurricanes but with a focus on environmental impacts. These researchers will have different data needs.

Image Retrieval Service

The retrieval results of images with Earth science phenomena using CNN outperformed our empirical results using traditional machine learning algorithms with several handcrafted features. However, there are several issues to using deep learning for retrieving images with Earth science phenomena. First, retrieval accuracy for certain phenomena (e.g., Smoke/Haze) only improved slightly compared to accuracy using traditional algorithms. We believe we need to improve the representative sample images of those phenomena in our training set. Second, the speed of training is painstakingly long while executing the algorithm in commodity machine. We observed that the accuracy improved after the number of images in the training set was considerably increased. Furthermore, a larger number of layers and slower learning rate were required to obtain better accuracy. Both of these factors slow down the learning process. After we switched to parallel implementation of CNN on NVIDIA Tesla K20C GPU with 5GB memory for training, the process completed in about 7 hours with 7 layers and learning rate of 0.003.

5. REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008. ISBN: 0521865719

[2] Global Change Master Directory (GCMD). 2015. GCMD Keywords, Version 8.1. Greenbelt, MD: Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). URL: http://gcmd.nasa.gov/learn/keywords.html

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," In NIPS, pp: 1106–1114, 2012.

[4] J. Wan, D. Wang, S. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study," In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). ACM, NY, pp: 157-166, 2014.