Semantic Search for Self-describing Scientific Data Formats

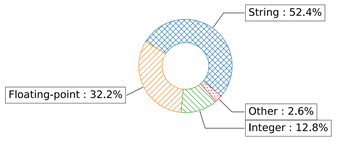
Chenxu Niu†, Wei Zhang†, Suren Byna‡, Yong Chen†
† Texas Tech University, ‡ Lawrence Berkeley National Laboratory

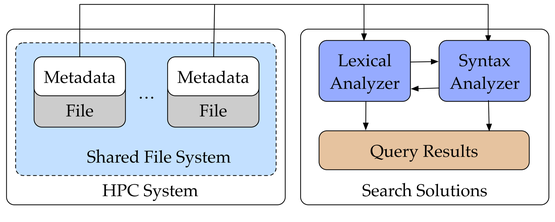
Introduction

- Finding relevant self-describing scientific datasets is critical to data discovery, data reuse, and data management
- Existing solutions[1] serve queries based on *(lexical) exact or partial match* of search keywords in metadata
- We propose to integrate semantics into the process so that search considers the meanings of keywords and metadata
- We build metadata semantic indexes too and provide a semantic-level search solution for scientific datasets
- Our experimental evaluation demonstrates the effectiveness and semantic search accuracy of our approach

Background

- Large-scale scientific applications typically store their experimental, observational and simulation datasets in self-describing data formats, such as HDF 5[2]
- In self-describing data files, the metadata can be seen as a collection of attributes
  - Each attribute represented as a key-value pair, where *k* represents the attribute name and *v* represents the attribute value
  - Finding required data can be accomplished by issuing metadata queries that utilize these key-value pairs
- Major data types of attributes are strings and numbers
  - In particular, the attribute names are of strings and attribute values are either strings or numbers
- **Existing dataset search solutions are based on (lexical) pattern matching, *without considering semantic meanings***
  - Match the querying keywords with attributes in the metadata
- Existing semantic solutions in information retrieval and semantic web are not directly applicable to scientific datasets
  - Metadata attributes in the self-describing data formats are in the form of key-value pairs, not sentences/paragraphs/articles

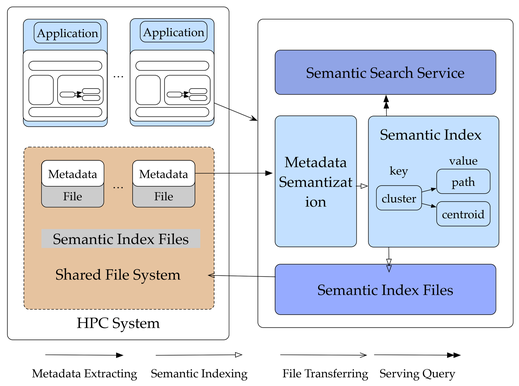


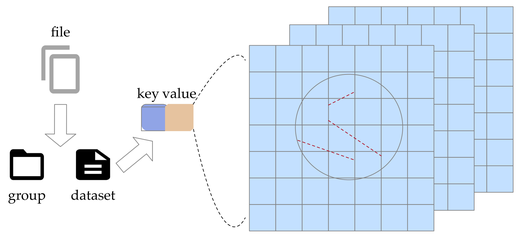


High-level structure of existing metadata search solutions

Methods

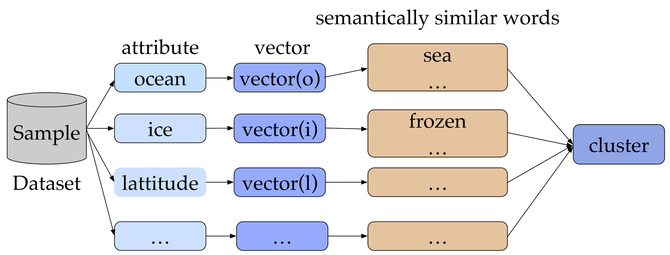
- **Overview of our solution:**





File structure and metadata semantization

- Metadata semantization represents each attribute by a multi-dimensional semantic vector
- **Metadata semantization**
  - Extracts metadata from objects in the files, including groups and datasets, and conducts metadata semantization to represent each attribute by a multi-dimensional semantic vector[3]
- **Semantic index construction**
  - Index builder retrieves the semantic vectors and develops clusters along with the corresponding object and file path
  - Similarity measures are used to determine the similarity between two semantic vectors
- **Semantic search**
  - Records the paths of the files and corresponding similarity scores in the values and ranks the files according to the similarity scores between the centroid and keywords

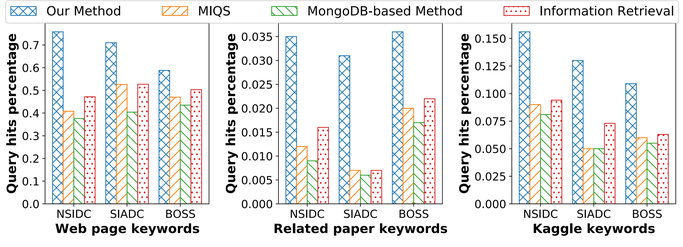


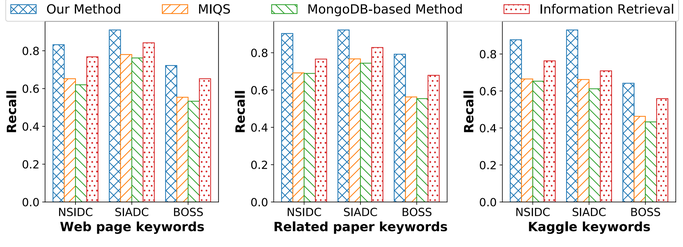
Sample of cluster construction over a real-world dataset

- For example, extracts each attribute and finds semantically similar words and corresponding vectors
- These vectors create a cluster and our approach builds indexes to link each dataset with its cluster
- When serving queries, the keyword is converted into a semantic vector with word-vector mapping to determine the similarity between the semantic vector and the clusters

Evaluation

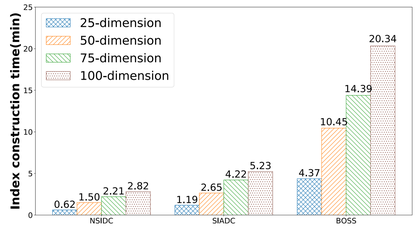
- We have performed experiments with three sets of real-world datasets: NSIDC, SIADC and BOSS
- We randomly selected keywords from science data website, related papers and Kaggle to perform queries
- Compared with existing metadata search solutions, we measure the search accuracy by calculating query hits percentage and recall accuracy
  - Recall accuracy: the rate of # of retrieved relevant datasets and # of the relevant datasets
- Our results demonstrate that our approach achieves better precision and recall accuracy than existing search solutions



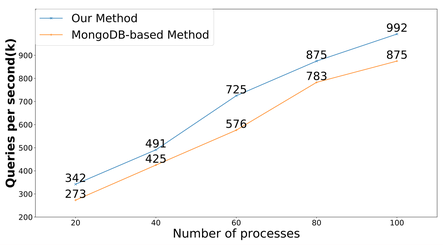


Performance Results

- We also tested the index construction performance
  - The results show that the index construction is proportional to the dimensions of the vectors



- We further tested the performance of queries
  - The results indicate that our approach is scalable, and the query throughput outperforms the MongoDB-based solution



Conclusion & Future Exploration

- In this research, we have explored a methodology to achieve semantic search for self-describing datasets
- We demonstrate the potential of this approach through experiments of real-world scientific datasets
- Our approach outperforms existing solutions in terms of semantic precision and recall accuracy
- We will continue improving the efficacy of the semantic search methodology by optimizing the metadata semantization and cluster construction

Acknowledgement

- We are thankful to the anonymous reviewers for their valuable feedback. This research is supported in part by the National Science Foundation under grant CCF-1409946, CCF-1718336, and CNS-1817094.
- This work is supported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References:

[1] W. Zhang, S. Byna, H. Tang, B. Williams, and Y. Chen, "MIQS: Metadata Indexing and Querying Service for Self-describing File Formats," SC 2019.

[2] M. Folk, A. Cheng, and K. Yates, "HDF5: A File Format and I/O Library for High Performance Computing Applications," SC 1999.

[3] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," EMNLP, 2014.