International Journal of Geo-Information

MDPI

Article

A Smart Web-Based Geospatial Data Discovery System with Oceanographic Data as an Example

Yongyao Jiang [1], Yun Li [1], Chaowei Yang [1,*], Fei Hu [1], Edward M. Armstrong [2], Thomas Huang [2], David Moroni [2], Lewis J. McGibbney [2], Frank Greguska [2] and Christopher J. Finch [2]

1 NSF Spatiotemporal Innovation Center and Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA 22030, USA; yjiang8@gmu.edu (Y.J.); yli38@gmu.edu (Y.L.); fhu@gmu.edu (F.H.)

2 NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA; Edward.M.Armstrong@jpl.nasa.gov (E.M.A.); Thomas.Huang@jpl.nasa.gov (T.H.); David.F.Moroni@jpl.nasa.gov (D.M.); Lewis.J.Mcgibbney@jpl.nasa.gov (L.J.M.); Francis.Greguska@jpl.nasa.gov (F.G.); Christopher.J.Finch@jpl.nasa.gov (C.J.F.)

* Correspondence: cyang3@gmu.edu; Tel.: +1-703-993-4742

Received: 15 January 2018; Accepted: 1 February 2018; Published: 11 February 2018

Abstract: Discovering and accessing geospatial data presents a significant challenge for the Earth sciences community as massive amounts of data are being produced on a daily basis. In this article, we report a smart web-based geospatial data discovery system that mines and utilizes data relevancy from metadata user behavior. Specifically, (1) the system enables semantic query expansion and suggestion to assist users in finding more relevant data; (2) machine-learned ranking is utilized to provide the optimal search ranking based on a number of identified ranking features that can reflect users' search preferences; (3) a hybrid recommendation module is designed to allow users to discover related data considering metadata attributes and user behavior; (4) an integrated graphic user interface design is developed to quickly and intuitively guide data consumers to the appropriate data resources. As a proof of concept, we focus on a well-defined domain-oceanography and use oceanographic data discovery as an example. Experiments and a search example show that the proposed system can improve the scientific community's data search experience by providing query expansion, suggestion, better search ranking, and data recommendation via a user-friendly interface.

Keywords: knowledge base; semantic search; user behavior; metadata; search ranking; recommendation; big data

1. Introduction

The global ocean plays several critical roles in the physical climate system of the Earth. The oceans receive more than half of the solar radiation entering the climate system, and evaporative cooling balances much of the solar energy absorbed by the oceans, making them the primary source of water vapor and heat for the atmosphere [1]. Currents in the oceans can move water over great distances and carry heat and other ocean properties from one geographic area to another. The poleward energy transport by the ocean is important in reducing the pole-to-equator temperature gradient. Horizontal and vertical transport of energy by the ocean can also alter the nature of regional climates by controlling the local sea surface temperature [2]. The recent extreme ocean-related weather events (e.g., Hurricanes Harvey, Irma, and Maria) have led to multiple natural disasters in the United States and around the world, resulting in catastrophic levels of damage to our society and environment. To accurately track, predict and assess the consequences of these disasters and to enhance disaster preparedness

and emergency response, near real-time and high spatiotemporal resolution satellite and in-situ oceanographic data has become more important than ever.

However, discovering and accessing oceanographic data in a manner that precisely and efficiently satisfies user demands presents a significant challenge for the ocean science community [3]. For example, the current difficulties that researchers face in discovering and accessing the most applicable observational data at NASA has detrimental consequences for meeting the challenges of climate and environmental change, identified in the 2011 NASA Strategic Plan [4]. At present, the satellite observations needed by the scientific community to evaluate and improve model simulations are under-utilized because the appropriate data are extremely difficult to find among the petabytes of available data [5]. Since the volume of data is only increasing as a function of time, a new paradigm of more open, user-friendly data access is needed [6].

In this context, many online portals have been built to improve the accessibility of oceanographic data. For example, the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) serves physical oceanographic satellite data to the Earth science community. In reality, scientists are still limited to the use of datasets that are familiar to them and they often have little knowledge of the existence of datasets that could be a better fit for their model or application due to the inefficiency of current geospatial search engines [7]. Specifically, finding appropriate geospatial data efficiently and accurately is challenging in three aspects.

(1) Lack of semantic context. Keyword-based search is widely adopted in operational geospatial data portals. Since keyword search uses string matching without considering the semantic context, precision, and recall, the two important measurements for search relevance are hard to be guaranteed [8]. For example, when querying "sea surface temperature" using a keyword search, the query is interpreted as a Boolean query "sea AND surface AND temperature." The search results likely contain the terms "sea", "surface" and "temperature" within their textual content but may not result in documents containing its common abbreviation "sst".
(2) Only single attribute based ranking. There are typically hundreds or even thousands of datasets related to the given query. Current search engines in most geospatial data portals tend to induce end users to focus on one single data attribute (e.g., spatial resolution) [9]. PO.DAAC provides several features to rank the search results, including all-time popularity, monthly popularity, grid spatial resolution, etc. This approach largely fails to take account of users' multidimensional preferences for geospatial data, which often results in less than optimal user experience [10].
(3) Lack of data relevancy. There exist hidden relationships among data hosted by a search engine. For example, after a user clicks on a data, he or she should be informed of the latest version of the clicked data which often has a better accuracy. In addition, Earth system scientists often need to interconnect their research using multiple physical parameters because important discoveries and the overall progress of science often transcend the domain of a single discipline [11].

To address the above challenges, we propose a smart web-based geospatial data discovery system that mines and utilizes data relevancy from metadata, user behavior, and ontology. The contributions of the proposed system are as follows: (1) the system enables semantic query expansion and suggestion to assist users in finding more relevant data; (2) machine learned ranking is utilized to provide the optimal search ranking based on a number of identified ranking features that can reflect users' search preferences; (3) a hybrid recommendation module is designed to allow users to discover related data considering metadata attributes and user behavior; (4) an integrated graphic user interface design is developed to quickly and intuitively data consumers to the appropriate data resources. As a proof of concept, we focus on a well-defined domain-oceanography and use oceanographic data discovery as an example.

2. Related Work

Previous work has attempted to solve the semantic problem through manual creation of ontologies [12]. The associations and concepts (e.g., polysemy and synonym) is often used to provide semantic context for a given query. Geospatial ontologies such as the Semantic Web for Earth and Environmental Terminology (SWEET) [13] capture concepts and relations in the geospatial domain. European INSPIRE (Infrastructure for Spatial Information in the European Community) implemented a semantic-based search approach based on ontology [14]. The problem with the manual creation of ontologies is that it is very labor intensive and hard to maintain up-to-date. Another approach to this challenge has been applied through document-clustering and dimension-reduction techniques such as Latent Semantic Analysis (LSA) [15] and Latent Dirichlet allocation (LDA) [16]. Li, Goodchild [7] developed a geospatial semantic search algorithm integrating LSA in the broad domain of Earth science, and Hu, Janowicz [17] performed topic modeling using LDA in geospatial portals. The advantage of these solutions lies in their automaticity and human and language independence. However, this approach is prone to noise and hard for a human to understand or to interact directly. We, therefore, propose an approach to discovering latent semantic relationships by mining user search logs.

Although various ranking algorithms are adopted by the existing geospatial data portals, such as term frequency-inverse document frequency (TF-IDF) and Okapi BM25 [18], all of them only focus on measuring the overlap between user query and metadata content. Attempts have been made to improve the keyword based ranking by performing semantic analysis, but other aspects of the data that can be related to users' search interest are overlooked such as when the data was released [7,17]. Martins and Calado [19] apply machine learning to rank newspaper documents of geographic query. Shaw, Shea [20] from Foursquare proposed a spatial search algorithm using machine learning to infer users' location. Considering the unique needs of geospatial data discovery, we therefore propose a few ranking related features and apply a machine learning approach to automatically learn a function to weight the ranking features.

With the advancement of semantic technologies, an emerging approach to connecting data is to publish data as "Linked Data" [21]. A good example in geospatial domains is the GeoLink EarthCube project [22]. GeoLink allows users to browse the data by clicking on a metadata attribute (e.g., instrument) to view the related data that share the same attribute value. One issue is that it requires the data to be published using the semantic standards such as Resource Description Framework (RDF). Moreover, there could be many data related to the clicked data which can be overwhelming. Recommender system has achieved remarkable success in many commercial products (e.g., Netflix). It typically produces a list of recommendations in one of two ways—through collaborative and content-based filtering [23]. Vockner et al. [24] proposed a recommendation system based on LSA. As an early attempt in the geospatial domains, we therefore propose a hybrid recommendation method of measuring the relatedness of data with a few identified metadata attribute combined with users' browsing behavior.

This paper also discusses how to integrate the above three functionalities into a data discovery system and how each of them is supported by different system components. The overarching objective is to increase the efficiency of data exploration and enable emerging user communities to readily discover and access data appropriate to their endeavors.

3. System Framework

3.1. Architecture

The system consists of three major components: Web graphic user interface (GUI)/services, knowledge base, and smart engine (Figure 1). Users interact with the system through the Web GUI while generating web logs. The knowledge base stores metadata, user behavior data, and machine-learned models. The smart engine includes four subcomponents: profile analyzer, ranker, semantic similarity calculator, and recommender. The profile analyzer extracts user behavior

data from raw Web logs on a regular basis and stores it into the knowledge base. The semantic calculator calculates the semantic similarity between different user search queries based on the user access pattern, which supports both the ranker and recommender. The ranker searches the metadata index and produces an optimal list of ranked results based on a few predefined ranking features, a pre-trained machine-learned ranking model, and the semantic similarity results. The recommender generates a list of related datasets based on the data that is currently being viewed according to a few pre-defined recommendation features, a pre-trained collaborative filtering recommendation model, and the semantic similarity results as well.

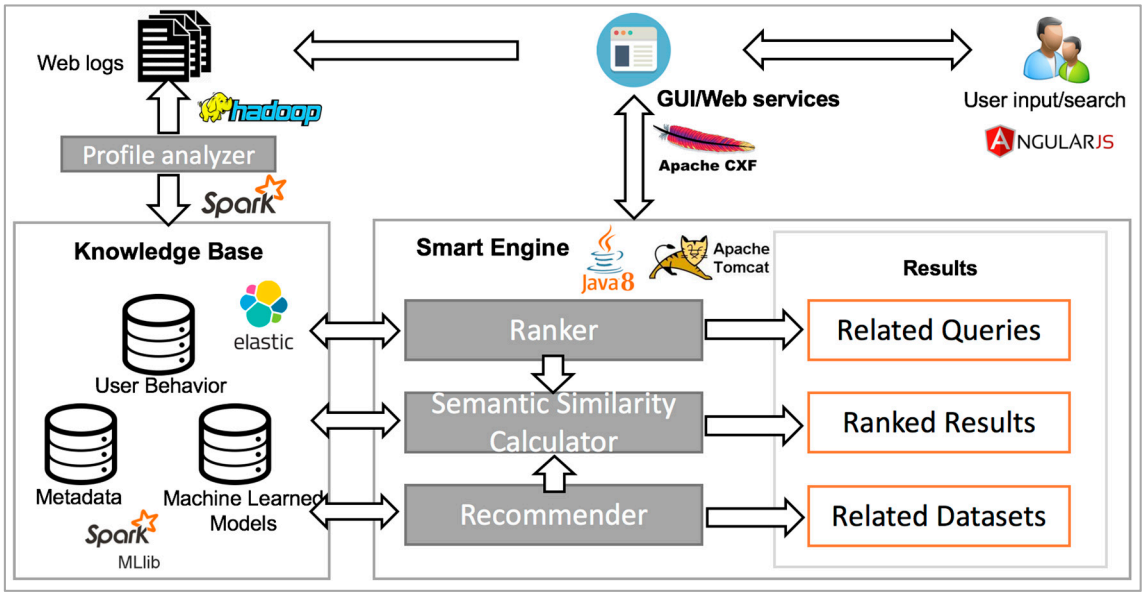


Figure 1. System architecture.

3.2. System Components

3.2.1. System Web GUI

The system Web GUI adopts the user-centered design and provide the interface for user interactions with: (a) search constraints input; (b) ranked results; (c) data exploration based on recommendations; and (d) navigation through query suggestion to find relevant datasets.

Three panels are added to assist users with better data discovery and access: (a) "related queries" panel is provided to display the semantically similar user queries; (b) "machine learning based ranking" panel is developed to provide more relevant results for end users; (c) "related dataset" panel is added once user selected a specific dataset. Domain scientists would be able to use the three functionalities to quickly nail down available datasets and be directed to the data downloading service. Web services of these three components are also developed to support communication with other applications.

3.2.2. Knowledge Base

The knowledge base includes three parts: metadata, user behavior, and machine-learned models. The metadata is the description of data and is indexed in a full-text search engine. The user behavior is the log mining results of the profile analyzer, which lays the groundwork for the ranker, semantic similarity calculator, and recommender. The machine-learned models include the pre-trained ranking model, the co-occurrence matrix of user search history and clickstream, and the pre-trained collaborative filtering recommendation model.

3.2.3. Smart Engine

As the most crucial component of the system, the smart engine consists of four subcomponents: profile analyzer, ranker, semantic similarity calculator, and recommender. The profile analyzer performs log mining and updates user access pattern in the knowledge base periodically. At query time, the smart engine takes the search input and coordinate the search against the metadata index. The search return of the metadata index is then re-ranked by the ranker. Given a user query, the similarity calculator can produce a list of highly related user queries. Once users select a data in the ranked results, the recommender would provide a list related data.

3.2.4. Profile Analyzer

Profile analyzer extracts user access pattern from raw web logs. The log processing workflow has four steps: user identification, crawler detection, session identification, and structure reconstruction (Figure 2). The user identification step identifies each individual user through IP address and web browser. The crawler detection step detects and removes web logs generated by the robotic activities. The session identification splits a sequence of web logs of each user into sessions representing single visits of that user. The session reconstruction step connects user actions according to the previous page information of the web log. More details can be found at Jiang, Li [25].

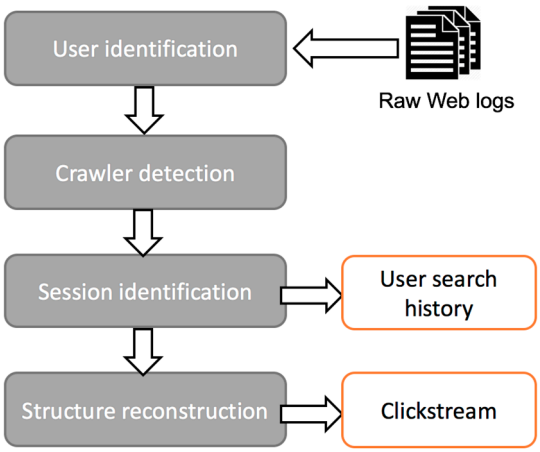


Figure 2. Workflow of the profile analyzer.

There are two types of output from the profile analyzer: user search history, and clickstream. User search history refers to the query searched by a given user in a certain pre-defined time period. Clickstream stands for a series of mouse clicks made while visiting a website. This information is kept in the knowledge base to support other components of the smart engine.

3.2.5. Semantic Similarity Calculator

The similarity calculator computes the semantic similarity between user queries. The assumption is that if two queries are similar, (1) the more frequent they would co-occur in distinct users' search histories; (2) the clicked data would be also similar in the context of large-scale user behaviors. Based on this assumption, the LSA is applied to the query co-occurrence matrix of user search history and clickstream to uncover the latent links between semantically-related terms (Figure 3). The results from both sides are independently scored and intersected to remove noise unique to each side. The resulting similarity values range from 0 (i.e., no relation) to 1 (i.e., identical). The similarity results are stored in the knowledge base and updated periodically. More details can be found at Jiang, Li [26]. The highly-related terms along with their associated similarity values can be used for query expansion and suggestion.

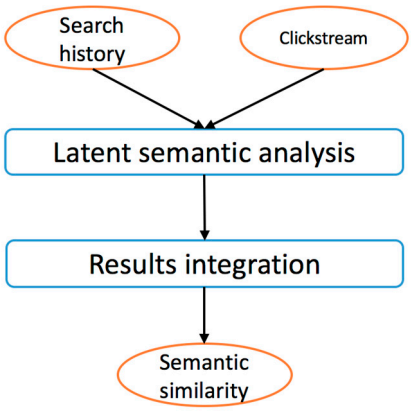


Figure 3. Workflow of semantic similarity calculator.

3.2.6. Ranker

The ranking module is designed to improve the ranking of the search results (Figure 4). When a user submits a query, it is then converted into a semantic query based on the returned results of semantic similarity calculator. For example, query "sea surface temperature" would be converted to "sea surface temperature OR sst". The search index would then return the top K results for the semantic query. After that, feature extractor would extract the ranking features for each of the search results. The ranking features include text-based relevance score, spatial similarity, version number, processing level, release date, spatial resolution, temporal resolution, all-time popularity, monthly-popularity, and user popularity. Once all the features are prepared, the top K results would then be put into a pre-trained RankSVM ranking model, which would finally re-rank the top K retrieval.

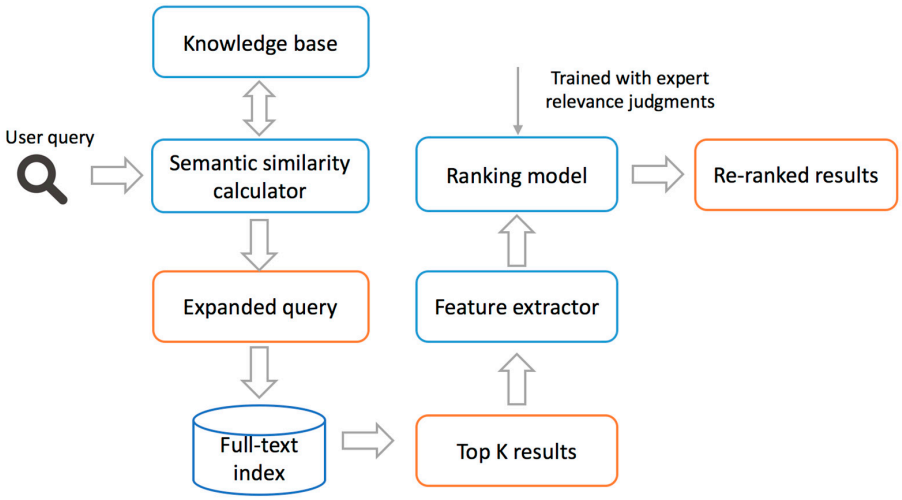


Figure 4. Workflow of the ranker.

3.2.7. Recommender

The recommendation module is developed to predict data that users might be interested in. Recommendations are made based on two types criteria: metadata content and user behavior. The goal is to identify the most similar data based on the data that is being viewed. Figure 5 describes the workflow of the recommendation algorithm. In the metadata content based calculation, after being weighted, metadata attributes (e.g., topic, processing level, spatial resolution) are divided into three categories: spatiotemporal, ordinal and categorical. Corresponding similarity algorithms are designed for each category. In the user behavior based method, data co-occurrence matrix with respect to

user sessions is constructed from user behavior data and then the LSA is applied to calculating the similarity. The intuition is that two data are more likely to be similar if they co-occur in distinct users' web sessions more frequently. Finally, the weighted average of these two methods is used to rank the recommendation results.

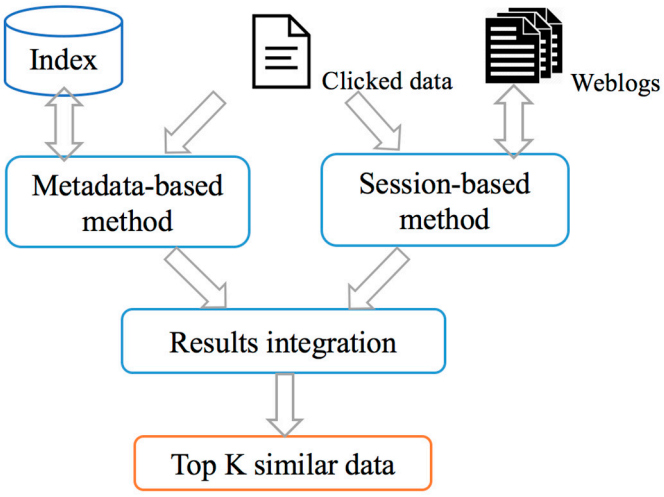


Figure 5. Workflow of the recommender.

3.3. Implementation

3.3.1. Data

Our experimental metadata come from all publicly available collection-level metadata from PO.DAAC. PO.DAAC distributes hundreds of unique datasets relevant to the world's oceans including those in the areas of ocean wind, topography, temperature, circulation, salinity and sea ice. The breadth and number of datasets in each domain is a key challenge to search relevance from the perspective of a user. For example, as shown in Table 1, the sea surface temperature (SST) catalog contains 205 datasets covering multiple disciplines.

Table 1. List of number of domain specific datasets available via the PO.DAAC and their community applications (Parameter abbreviations are defined as follows: Chl-A = Chlorophyll-A Concentration, G = Gravity, OSC = Ocean Surface Currents, OST= Ocean Surface Topography, OSW = Ocean Surface Wind Speed, OSWV = Ocean Surface Wind Vectors, SIAC = Sea Ice Age Classification, SSS = Sea Surface Salinity, SST = Sea Surface Temperature. Discipline abbreviations are defined as follows: ASI = Air Sea Interaction, Met = Meteorology, OB = Ocean Biology, PO = Physical Oceanography.).

Dataset Family	Example Source(s)	Number of Datasets	Parameter(s)	Discipline(s)
Ocean Wind	QuikSCAT, ASCAT, OSCAT	94	OSWV	PO, Met, ASI, Climate, OB
Ocean Radar	QuikSCAT, ASCAT, OSCAT	50	OSWV	PO, Met, ASI, Climate, OB
Ocean Temperature	AVHRR, MODIS, AMSR-E, TMI	205	SST	PO, Met, ASI, Climate, OB
Ocean Circulation	Multi-Sensor	5	OSC	PO, ASI, Climate, OB
Ocean Salinity	Aquarius	147	SSS, OSW, OST	PO, ASI, Climate
Ocean Topography	T/P, Jason -1, -2, Envisat	29	OST	PO, Met, ASI, Climate, OB
Gravity	Grace	70	G	PO, Climate

SST datasets are an essential resource for monitoring and understanding climate variability and climate change. Historically, SST measurements have been made from ships. Ship data have been compiled into databases like International Comprehensive Ocean-Atmosphere Data Set (ICOADS), which in turn form the main input into long-term climate datasets. Moored and drifting buoys are another primary source of in-situ SST data, especially in remote regions like the Southern Ocean, where ARGO floats offer much-improved coverage. Over the tropical Pacific, the dense Tropical Atmosphere Ocean (TAO) project-Triangle Trans-Ocean Buoy Network (TAO-TRITON) array provides key measurements for monitoring the emergence and evolution of El Niño events [27]. In-situ data are also the primary reference for calibrating satellite-based SST estimates [28]. Satellite-based estimates utilize measurements from infrared (IR) and microwave wavelengths. Microwave observations are less sensitive to clouds than IR measurements, but are more sensitive to scattering by rain, and have lower spatial resolution. For climate research, the longest satellite-based dataset is NOAA's OI SSTv2, extending from 1981 to present, with the Advanced Very-High Resolution Radiometer (AVHRR) IR measurements as the primary source data. The Group for High-Resolution SST (GHRSST) is an umbrella mission coordinating the development of multi-spectral SST data products for both the operational and climate communities. Currently, one of the longest global GHRSST products is the Multi-Scale Ultra-High-Resolution (MUR) SST analysis, a 0.01 degree gridded dataset developed by JPL, NASA, covering 2002-present [29].

Our experiments were run using one year of search records from PO.DAAC data search engine, which is nearly 120 million records in 30 gigabytes. These Web logs are in the Apache Common Log Format, the most widely used log format maintained by W3C. Each Web log has several fields including client IP address, request date/time, page requested, HTTP code, and bytes served.

3.3.2. System Implementation

The system is developed using Java 8, JavaScript, HTML 5, and CSS. The Angular JS JavaScript framework is used in the frontend development, which has a data-binding function that updates the view whenever the model changes, as well as updates the model whenever the view changes. The communication between the backend and frontend uses standard RESTful web-service interfaces enabled by Apache CXF and Tomcat. Elasticsearch is used as the full-text search index. The LSA algorithm in the semantic similarity calculator and the RankSVM algorithm in the ranker are implemented with Spark MLlib.

We used a Hadoop cluster with 5 data nodes each having a 2.4 GHZ AMD Opteron Processor with 4 to 8 cores and 8 to 16 GB RAM. It took about 1.5 h to index, query the one-year of Web logs, and build the required models (i.e., the co-occurrence matrices of user search history and clickstream). The database and models are updated monthly. The technologies used to implement the proposed system for PO.DAAC's dataset are: HDFS, Map/Reduce jobs, Spark, Elasticsearch, and DC2 [30,31]. The experiment was conducted on the NASA AIST cloud platform, a hybrid cloud computing environment provided for scientific research. The source code of the system has been published along with this paper as an open source software (https://github.com/mudrod/mudrod) under the MUDROD project [32].

3.4. User Scenario

After users log into the system, they type a query (e.g., ocean temperature) into the search box (Figure 6). The auto-completion function helps during typing by predicting the rest of words users intend to enter. When users hit the search button, a list of results is retrieved which has the default ranking of machine learning based ranking. Users can also choose to sort the list by other metrics such as popularity and spatial resolution. On the right-hand side, a list of related searches is displayed (Figure 7). In this particular case, the similar queries of "ocean temperature" are "sst", "sea surface temperature", "ghrsst", etc. Next to each related search is a number in parenthesis representing the semantic similarity value. Users can choose to click on any of these related searches to explore other

datasets. If users would like to know more details of a particular dataset in the search list, they can click on the "name" attribute (e.g., VIIRS_NPP-NAVO-L2P-v2.0) and more information such as version, processing level, coverage will be displayed. According to the recommendation algorithm behind the scene, the top related datasets are listed on the right (Figure 8). In this case, the most related dataset is the version 1.0 of collection "VIIRS_NPP-NAVO-L2P" as the dataset that is being viewed is the version 2.0 of it. An online demo system has been made available at https://mudrod.jpl.nasa.gov/#/.

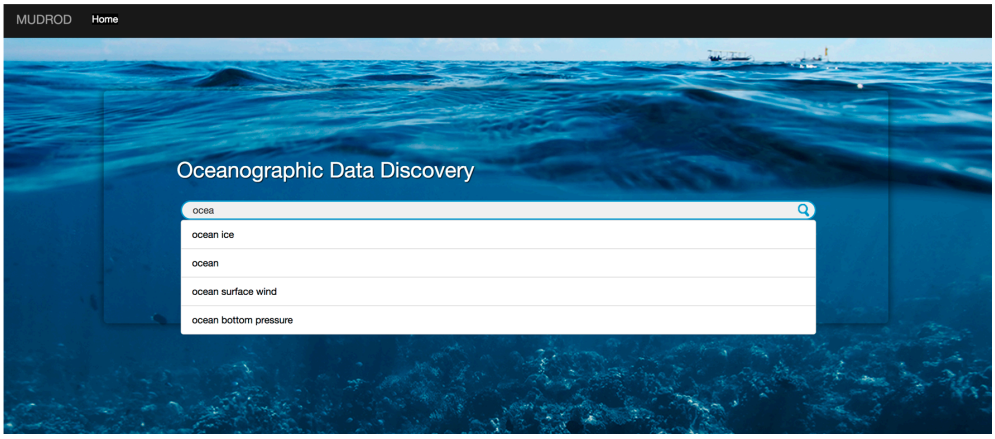


Figure 6. System main page.

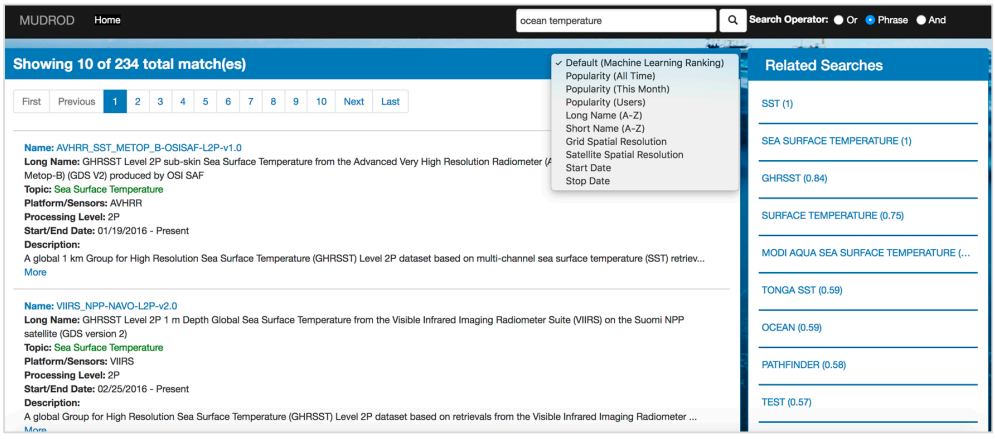


Figure 7. Search results page.

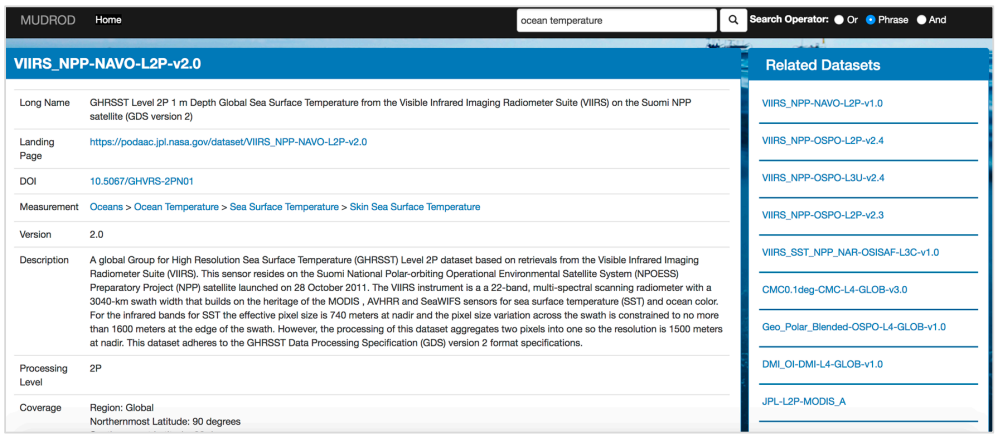


Figure 8. Dataset detail page.

3.5. Use Cases

Sea surface temperature data discovery is used as an example to demonstrate how the proposed system can improve the scientific community's data search experience by providing query expansion, suggestion, better search ranking, and recommendation via a user-friendly interface. Common considerations for utilizing SST datasets in climate research and model evaluation include (1) spatial and temporal resolutions—are features like the Gulf Stream and its fronts are eddies resolved? (2) quantity being measured—is it a "skin" temperature of a very thin surface layer or a bulk temperature of the upper meter or more? (3) processing level—do we need level 2 ungridded data that contains ancillary data fields as well as complete error characteristics for each pixel? (4) latency—is near real-time data needed? (5) spatial and temporal coverage—is the study area and period covered? (6) spatial interpolation—have the data been statistically interpolated in some manner, and what effect does this have on the spatial and temporal variance of climate signals?

3.5.1. Query Suggestion

Figure 9 shows the query suggestion results of the query "sea surface temperature". "sst" and "ocean temperature" are the first two queries in the "related searches" list, which have the similarity values of one. "sst" is a common abbreviation in the ocean science community. In the context of oceanographic satellite data which the experiment is designing around, "ocean temperature" and "sea surface temperature" are nearly synonymous because there are few sub-surface/deep datasets at PO.DAAC. This fact has been verified by data engineers of PO.DAAC. Given that the goal is to improve data discovery, this result is therefore reasonable. In fact, if more sub-surface/deep datasets are made available on PO.DAAC, the proposed method can automatically update the similarity according to the user access pattern. The search recall and precision can be improved by query expansion based on these synonymous queries, which has been systematically evaluated at Jiang, Li [26]. The third query is "ghrsst" with the similarity value of 0.83. "ghrsst" is the shorthand for The Group for High-Resolution Sea Surface Temperature (GHRSST) which is aimed to develop a new generation of global, multi-sensor, high-resolution near real-time SST products. Due to the quality ghrsst provides, it has become one of the most popular sea surface temperature data collections. Other SST oriented missions include AQUA, AVHRR-Pathfinder, Moderate Resolution Imaging Spectroradiometer (MODIS), Suomi National Polar-orbiting Partnership (S-NPP), TERRA, which can be found in the remaining related searches.

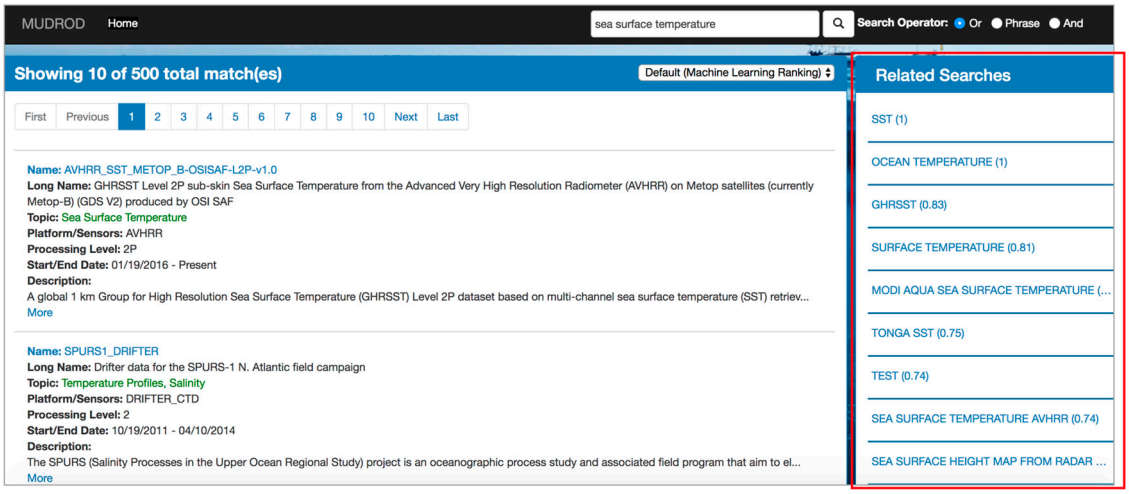
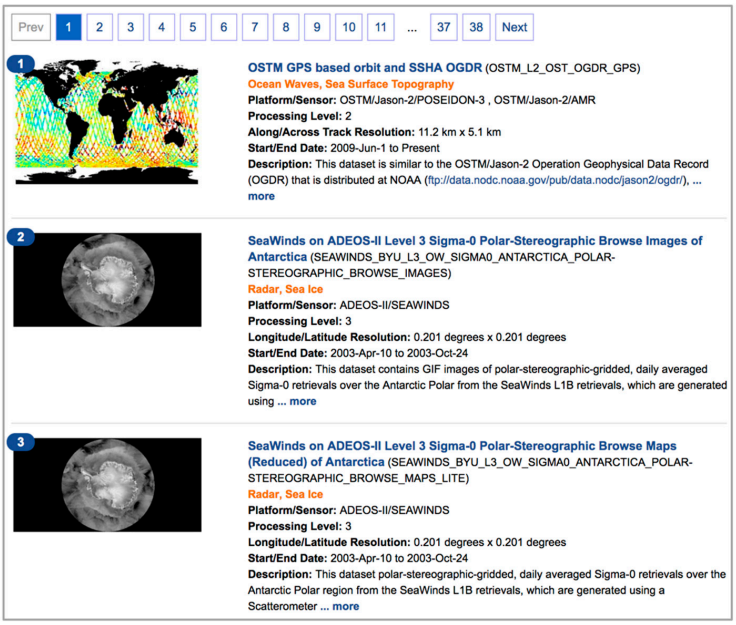


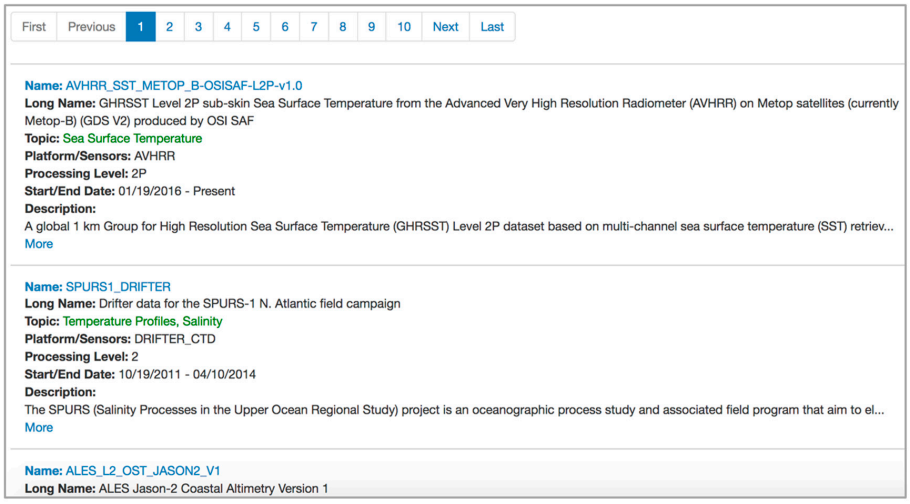
Figure 9. Query suggestion results.

3.5.2. Search Ranking

Figure 10 is the comparison of the top search results of "sea surface temperature" between PO.DAAC search engine and the proposed system. According to the data topic in orange on PO.DAAC website and green on the system's user interface, the topics of the first two data on PO.DAAC are "ocean waves, sea surface topography", "radar, sea ice", while that of the proposed system are "sea surface temperature" and "temperature profiles". This is because of the different rankings used by these two systems. PO.DAAC uses all-time popularity by default to rank the search results. Just because the "ocean waves" data has more downloads than "sea surface temperature" data, those data of little relevance is ranked to the top. The weakness of only considering one data characteristics has been overcome by the machine learning based ranking of the proposed system. This was a substantial precision improvement in ranking problems since the ultimate goal was to put the most desired data to the top of the search results.



(a)



(b)

Figure 10. Comparison between the proposed system and PO.DAAC's search results. (**a**) Top search results of "sea surface temperature" of PO.DAAC; (**b**) Top search results of "sea surface temperature" of the proposed system.

Another example is the order of dataset "AVHRR Pathfinder Level 3 Daily Nighttime SST Version 5" and "AVHRR Pathfinder Level 3 Daily Nighttime SST Version 5.1". These two datasets are the same AVHRR Pathfinder Level 3 Nighttime SST data of different versions. The second one is the newer version with better quality. Just because the former has been downloaded more historically, it outranks its replacement. A systematic evaluation based on precision at K and normalized discounted cumulative gain suggests that the machine learning approach outperforms other methods such as monthly popularity [9].

3.5.3. Recommendation

Figure 11 shows the recommendation results of a selected dataset—"AVHRR_SST_METOP_B-OSISAF-L2P-v1.0", which is the GHRSST Level 2P sub-skin Sea Surface Temperature from the Advanced Very High-Resolution Radiometer (AVHRR) on Metop-B satellites produced by OSI SAF. The first three datasets are AVHRR SST datasets of different satellite platforms, processing levels, and versions. The fourth and fifth ones are the AVHRR sensor data produced by the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) and the US Naval Oceanographic Office (NAVO), respectively. The recommendation function allows users to explore relevant data more easily, which in turn helps find the most desired data in a more timely manner.

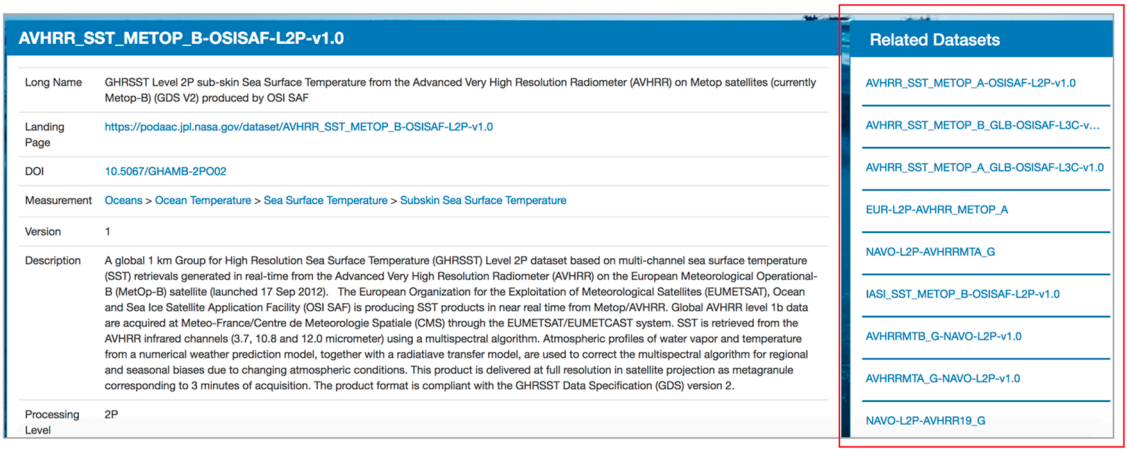


Figure 11. Recommendation results of a selected dataset.

4. Conclusions and Discussion

This article introduces the architecture and methodologies of MUDROD, a smart web-based geospatial data search engine aiming to improve data discovery by mining and utilizing data relevancy from metadata and user behavior. To assist users in finding and exploring more relevant data, a semantic similarity calculator is designed to support query expansion and suggestion. To help users find the most relevant data, a machine learning-based ranker is developed to provide the optimal search ranking based on a few identified ranking features. Additionally, a hybrid recommender is utilized to allow users to discover related data considering metadata attributes and user behavior. To improve users' search experience, an integrated graphic user interface design is developed to quickly and intuitively guide data consumers to the appropriate data resources.

There are several limitations with the current system. One is that the system can only process web logs in a batch mode, which means the users' search interest cannot be learned by the system in real time. We plan to integrate the real-time log ingesting function as it is crucial in many cases [33]. For example, during the course of a hurricane, the most relevant data should be changing as a hurricane region proceeds. Another limitation is that the ranking model is pre-trained using expert relevance judgments, which is both time- and labor-intensive. We are exploring methods of using user behavior to automatically create the training data for the machine learning ranking algorithm [34]. The last

concern is about the ranking feature identification. While the attributes reflect our intuition and discussion with domain experts, these are very likely not optimal. We plan to add more features (e.g., temporal similarity) in the future work. Additionally, a query understanding algorithm which can parse multi-phrase query to enable better semantic search is being actively developed.

Acknowledgments: This project is funded by NASA AIST (NNX15AM85G) and NSF (IIP-1338925 and ICER-1540998). The research was partially carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Author Contributions: Chaowei Yang came up with the original research idea and advised Yongyao Jiang and Yun Li on algorithms. Yongyao Jiang developed the semantic similarity and ranking algorithm. Yun Li developed the recommendation algorithm; Yongyao Jiang, Yun Li, Chaowei Yang and Fei Hu developed the workflow and implemented the system; Edward M. Armstrong, David Moroni, Thomas Huang and Christopher J. Finch provided substantial feedback to the system development; Lewis J. McGibbney and Frank Greguska packaged and documented the work as an open source project. Yongyao Jiang, Yun Li, Chaowei Yang and Fei Hu wrote the paper; Edward M. Armstrong, Thomas Huang and David Moroni revised the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hartmann, D.L. *Global Physical Climatology*; Elsevier: Amsterdam, The Netherlands, 2015.
2. Fan, Y.; Ginis, I.; Hara, T. The effect of wind–wave–current interaction on air–sea momentum fluxes and ocean response in tropical cyclones. *J. Phys. Oceanogr.* **2009**, *39*, 1019–1034. [CrossRef]
3. Devarakonda, R.; Palanisamy, G.; Wilson, B.E.; Green, J.M. Mercury: Reusable metadata management, data discovery and access system. *Earth Sci. Inf.* **2010**, *3*, 87–94. [CrossRef]
4. NASA. NASA Strategic Plan. 2011. Available online: https://www.nasa.gov/pdf/516579main_NASA2011StrategicPlan.pdf (accessed on 7 April 2017).
5. Yang, C.; Yu, M.; Hu, F.; Jiang, Y.; Li, Y. Utilizing Cloud Computing to address big geospatial data challenges. *Comput. Environ. Urban Syst.* **2017**, *61*, 120–128. [CrossRef]
6. Overpeck, J.T.; Meehl, G.A.; Bony, S.; Easterling, D.R. Climate data challenges in the 21st century. *Science* **2011**, *331*, 700–702. [CrossRef] [PubMed]
7. Li, W.; Goodchild, M.F.; Raskin, R. Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *Int. J. Digit. Earth* **2014**, *7*, 17–37. [CrossRef]
8. Jiang, Y.; Xia, J.; Liu, K. Polar CI Portal: A Cloud based polar resource discovery engine. In *Cloud Computing in Ocean and Atmospheric Sciences*; Vance, T.C., Merati, N., Yang, C., Yuan, M., Eds.; Academic Press: Amsterdam, The Netherlands, 2016; pp. 163–185.
9. Jiang, Y.; Li, Y.; Yang, C.; Hu, F.; Armstrong, E.M.; Huang, T.; Moroni, D.; McGibbney, L.J.; Finch, C.J. Towards intelligent geospatial data discovery: A machine learning framework for search ranking. *Int. J. Digit. Earth* **2017**, 1–16. [CrossRef]
10. Ghose, A.; Ipeirotis, P.G.; Li, B. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Mark. Sci.* **2012**, *31*, 493–520. [CrossRef]
11. NRC. *New Research Opportunities in The earth Sciences*; National Academies Press: Washington, DC, USA, 2012.
12. AlJadda, K.; Korayem, M.; Grainger, T.; Russell, C. Crowdsourced query augmentation through semantic discovery of domain-specific jargon. In Proceedings of the 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 27–30 October 2014; pp. 808–815.
13. Semantic Web for Earth and Environmental Terminology (SWEET). Available online: https://www.researchgate.net/publication/250346856_Semantic_Web_for_Earth_and_Environmental_Terminology_SWEET (accessed on 10 May 2017).
14. Gunay, A.; Akcay, O.; Altan, M.O. Building a semantic based public transportation geoportal compliant with the INSPIRE transport network data theme. *Earth Sci. Inf.* **2014**, *7*, 25–37. [CrossRef]
15. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2004**, *38*, 188–230. [CrossRef]
16. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **2002**, *1*, 601–608.
17. Hu, Y.; Janowicz, K.; Prasad, S.; Gao, S. Metadata topic harmonization and semantic search for linked-data driven geoportals: A case study using ArcGIS Online. *Trans. GIS* **2015**, *19*, 398–416. [CrossRef]

18. Gormley, C.; Tong, Z. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2015.
19. Martins, B.; Calado, P. Learning to rank for geographic information retrieval. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18–19 February 2010; p. 21.
20. Shaw, B.; Shea, J.; Sinha, S.; Hogue, A. Learning to rank for spatiotemporal search. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 717–726.
21. Linked Data-The Story So Far. Available online: https://eprints.soton.ac.uk/271285/ (accessed on 10 May 2017).
22. Krisnadhi, A.; Hu, Y.; Janowicz, K.; Hitzler, P.; Arko, R.; Carbotte, S.; Chandler, C.; Cheatham, M.; Fils, D.; Finin, T. The GeoLink modular oceanography ontology. In Proceedings of the 14th International Semantic Web Conference, Bethlehem, PA, USA, 11–15 October 2015; pp. 301–309.
23. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl. Based Syst.* **2013**, *46*, 109–132. [CrossRef]
24. Vockner, B.; Richter, A.; Mittlböck, M. From geoportals to geographic knowledge portals. *ISPRS Int. J. Geo-Inf.* **2014**, *2*, 256–275. [CrossRef]
25. Jiang, Y.; Li, Y.; Yang, C.; Armstrong, E.M.; Huang, T.; Moroni, D. Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. *ISPRS Int. J. Geo-Inf.* **2017**, *5*, 54. [CrossRef]
26. Jiang, Y.; Li, Y.; Yang, C.; Liu, K.; Armstrong, E.; Huang, T.; Moroni, D.; Finch, C. A comprehensive methodology for discovering semantic relationships among geospatial vocabularies using oceanographic data discovery as an example. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2310–2328. [CrossRef]
27. McPhaden, M.J. Genesis and evolution of the 1997-98 El Niño. *Science* **1999**, *283*, 950–954. [CrossRef] [PubMed]
28. UCAR. SST Data Sets: Overview & Comparison Table. 2014. Available online: https://climatedataguide.ucar.edu/climate-data/sst-data-sets-overview-comparison-table (accessed on 10 May 2017).
29. Martin, M.; Dash, P.; Ignatov, A.; Banzon, V.; Beggs, H.; Brasnett, B.; Cayula, J.-F.; Cummings, J.; Donlon, C.; Gentemann, C. Group for High Resolution Sea Surface temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE). *Deep Sea Res. Part II* **2012**, *77*, 21–30. [CrossRef]
30. Li, Y.; Jiang, Y.; Hu, F.; Yang, C.; Huang, T.; Moroni, D.; Fench, C. Leveraging cloud computing to speedup user access log mining. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 Septembe 2016; pp. 1–6.
31. Jin, B.; Song, W.; Zhao, K.; Wei, X.; Hu, F.; Jiang, Y. A high performance, spatiotemporal statistical analysis system based on a Spatiotemporal Cloud Platform. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 165. [CrossRef]
32. Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access. Available online: http://adsabs.harvard.edu/abs/2015AGUFMIN51B1809J (accessed on 10 May 2017).
33. Ranjan, R. Streaming big data processing in datacenter clouds. *IEEE Cloud Comput.* **2014**, *1*, 78–83. [CrossRef]
34. Agichtein, E.; Brill, E.; Dumais, S. Improving web search ranking by incorporating user behavior information. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 19–26.