This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



# Are Datasets Like Documents?: Evaluating Similarity-Based Ranked Search Over Scientific Data

V.M. Megler and David Maier, *Senior Member, IEEE*

**Abstract**— The past decade has seen a dramatic increase in the amount of data captured and made available to scientists for research. This increase amplifies the difficulty scientists face in finding the data most relevant to their information needs. In prior work, we hypothesized that Information Retrieval-style ranked search can be applied to datasets to help a scientist discover the most relevant data amongst the thousands of datasets in many formats, much like text-based ranked search helps users make sense of the vast number of Internet documents. To test this hypothesis, we explored the use of ranked search for scientific data using an existing multi-terabyte observational archive as our test-bed. In this paper, we investigate whether the concept of varying relevance, and therefore ranked search, applies to numeric data – that is, are data sets are enough like documents for Information Retrieval techniques and evaluation measures to apply? We present a user study that demonstrates that dataset similarity resonates with users as a basis for relevance and, therefore, for ranked search. We evaluate a prototype implementation of ranked search over datasets with a second user study and demonstrate that ranked search improves a scientist's ability to find needed data.

**Index Terms**—Scientific databases, information retrieval and relevance, similarity search

## 1 Introduction

Imagine you are an ocean microbiologist studying the effects of temperature on a population of some organism. You've collected 10 biological samples. You know that for each sample, a vertical temperature profile was collected at the same time and place, and each profile stored in a separate dataset. You are now trying to locate the corresponding temperature data for each sample in the collection of temperature profile datasets. You might be able to figure out which is which from the file names; at worst you can open each file to check the Latitude, Longitude and Date columns.

Time passes. You've now collected 100 samples, some over three years ago; not all samples have nearby temperature profiles available, and the instruments, data formats and naming conventions for profiles have changed. For the sample at hand, you have location L and time T, but you can't recall where to find the relevant temperature data. It is still possible to go through each dataset individually looking for the right combination (but what you are looking for may not exist).

Now other scientists have started contributing datasets, and there are over 1000 temperature profiles. If you can check whether a temperature profile dataset is near a given sample in 20 seconds, a direct search takes over five hours. If accurate metadata on time and location for each profile was collected and stored in a catalog with query capability, a query on time or on space might reduce the number of datasets to check.

Once there are 100,000 temperature profiles, if you query on a range around the T and L of a sample, you might still get 1,000 datasets to consider; alternatively, you might get zero. You can iterate your query, making it more or less strict. Perhaps you are willing to look through 10 or 20 profiles for applicability, but how do you form a query that gives you the "best" or "most likely" 10 or 20? It would help immensely if they were arranged roughly in order of similarity to your information need. These data sizes are not unrealistic; archives are now routinely terabytes in size and may contain thousands of datasets, and the rate of increase continues to accelerate [1], [2].

As data archive sizes grow, methods scientists have used to find data begin to fail. Some systems rely on manual navigation of catalogs; the scientist is expected to be able to choose the correct option at each step that will eventually lead to the desired dataset. Some systems rely on purely geographic metadata comparisons such as *contains* or *intersects*; others, on Boolean queries for specific words in metadata. Metadata collection, curation and maintenance is an acknowledged and ongoing problem, and reliance on manual collection of metadata is considered a prescription for failure [1], [3]. Both manual navigation and metadata-query approaches often result in time-consuming, repeated actions. This problem was highlighted at a National Research Council workshop [2], and in working with one scientific archive, the Center for Coastal Margin Observation and

- *V.M. Megler is with the Department of Computer Science, Portland State University, Portland, OR 97201. E-mail: vmegler@ cs.pdx.edu.*
- *David Maier is with the Department of Computer Science, Portland State University, Portland, OR 97201. E-mail: maier@cs.pdx.edu.*

*Manuscript received January 24, 2013.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE
Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



Prediction (CMOP), the scientists brought this issue of finding relevant data to our attention as one of their highest priority problems with CMOP's archive.

The Internet has seen similar explosive growth, and web search techniques now allow users to easily find relevant documents despite that growth. The task faced by a scientist searching for relevant data strongly resembles that faced by a user searching a large document collection: the scientist hopes for an item that exactly matches her information need, but should it not exist, would still be interested in the closest matches. Research shows that 95% of users would rather have an approximate answer or a near match than none at all [4]. Could ranked IR techniques be applied to datasets? Are datasets like documents? Does representing a dataset by a summary of its content – for example, summarizing contained geographic data by its spatial footprint, as shown in Fig. 1 – make sense to scientists? At first, the comparison of datasets to documents may seem strange. On the other hand, if a feature-space model can be used to calculate an overall similarity score between a search consisting of several words and a document containing hundreds or thousands of words, adapting the model to comparing similarities between numeric search conditions and numeric data with hundreds or thousands of attribute values seems viable.

To adapt IR techniques to scientific-dataset search, we need three things: a way to express a scientific information need as a set of search conditions; a method for extracting features from datasets; and a similarity measure to compare search conditions to the extracted features. Further, we must validate that any proposed set of features and similarity measure resonates with potential searchers; that validation is the focus of this paper. That is, we show that the search system has utility, and that the similarity measure embodies a notion of relevance that mimics the judgment of potential users. As noted by Saracevic [5], the notion of relevance differentiates IR from database retrieval (although databases may be used to implement IR). The concept of different levels of relevance for different items, and approximation of those levels via a similarity measure, supports ranked retrieval based on relative similarity scores for different items. We could thus present a research scientist with a ranked list of all available datasets that is ordered by decreasing estimated relevance to a posed search. If these concepts can be confirmed, then the application of IR measures, such as mean average precision, to the resulting approaches should also be valid.

Traditional text IR treats a document as a bag of words, with each distinct word a feature; further, a frequently used word is seen as having less value than a less frequently used word, leading to the tf-idf similarity measure. A text IR query also consists of a bag of words, and thus each search term can be matched to a document feature. Our scientists, however, do not search for specific values found in a dataset ("air temperature = 14.93615C"), but rather express their information needs in terms of an observational variable with values in some range ("water temperature between 5 and 10C"). Thus, we rejected the bag-of-words model and tf-idf measure in favor of using variable names and value ranges as our features, and developing a similarity measure that allows us to compare them. (Appendix A has further detail on our similarity measure.) We used an existing observational data archive as our "document database" and created a feature-extraction tool, a metadata catalog and a search engine to further test the practicality of our concepts. Our search tool embodies the candidate similarity measure based on the proximity of a search to a set of features that summarize a dataset's contents. This article reports on results of experiments with our user community and their data archive. While our user studies necessarily reflect our current user base, we believe the concepts are generalizable to other scientific fields.

Based on our experience, we assert the IR concept of relevance, IR similarity measures and IR evaluations are all applicable to ranked retrieval of scientific datasets. The importance of this claim should not be underestimated. Without such a capability, the usefulness of a scientific archive decreases as the archive grows beyond the ability of an individual scientist to navigate it. In Section 2 we describe a user study to test the feasibility of ranked search of scientific datasets via a relatively straightforward similarity measure. This first study focuses on geospatial and temporal characteristics of observational datasets, two features that are critical in many areas of scientific research. Section 3 summarizes the prototype feature-extraction and search tools we developed based on the encouraging results of the study. These tools are in current use on a large environmental repository. Section 4 describes a second user study in the form of an operational test of the search tool against 30,400 datasets totaling more than 0.5 TB, a section of the repository. This data is in active use, and represents over a decade of environmental observations. Thus, the task we are studying in this work is both topical and real. The searchers are scientists using the repository, who formulated searches representing their own information needs. This study structure gave us a tight linkage between real users with their own information needs, and the assigned relevance ratings or "ground truth" for their search results. We then reflect on what we learned in Section 5, discuss related work in Section 6, and describe future work in our conclusion, Section 7.

Our contributions in this paper are:

1. We demonstrate via our first user study that the concepts of "dataset relevance" and "dataset similarity" are meaningful, implying that Information-Retrieval-style ranked search over scientific data is reasonable.
2. We show that we can directly map these principles into a ranked retrieval system for datasets; and, we implemented these principles in a prototype [6], [7].
3. We present a second user study that demonstrates the prototype improves scientists' ability to find relevant data, thus removing a significant impediment to research productivity.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



4. We demonstrate that IR measures (such as RBP and DCG) are applicable to dataset search, and they indicate our candidate similarity measure performs well compared to several alternatives.

While our experiments are undertaken within one particular scientific research discipline and archive of observational data, the same issues and problems we consider are seen in many other research disciplines [2]. The related issue of attribute name normalization ("temp", "air_temperature", etc.) in scientific datasets is an important one, and is addressed elsewhere [8].

In this paper we focus on assessing dataset relevance and similarity, not techniques for reducing search latency, such as indexes. This work is focused on search effectivesness as assessed by our users, the scientists.

## 2 User Study 1: Feasibility

The Center for Coastal Margin Observation and Prediction (CMOP) is an NSF Science and Technology Center based in Hillsboro, Oregon, focused on coastal-margin and near-ocean issues. It is a multi-institution research partnership with a collection of observational data from the Columbia River and off the Washington and Oregon coasts spanning more than a decade. CMOP's observational archive contains thousands of datasets in a variety of file formats and in an RDBMS—over 0.5TB in aggregate. CMOP scientists analyze historical observations and run complex simulation models, producing additional terabytes of (human-generated) observations [7].

Almost all data is accessible for public download via CMOP's portal and THREDDS server.[1] The observation platforms are both fixed and mobile; the observations may be as frequent as every few milliseconds, or as rare as a single sample. Each observation consists of a time, a geospatial location, and a set of environmental or biological variables. Observations that are related in some way (often, collected from the same source or at the same time, or supporting a single research project) are stored together in datasets. The set of environmental variables (hereafter called simply variables) that is observed changes over time; there are frequent changes in the instruments deployed as new instruments are developed and new research topics studied, leading to changes in the data structures storing the data. The datasets are heterogeneous in content, format and storage type; multiple tools are needed to access and read the different dataset types, and no single interactive query or search capability spans all the data [6]. Converting all existing data into a common format is not feasible. However, we posited that creating a common abstraction or summary of datasets for use in a dataset catalog might be feasible.

The scientists at CMOP often search for data using geospatial areas, temporal ranges, or both; therefore, those were the initial focus for our application of IR techniques. (In Study 2, we expanded to include search terms for observed variables.) We proposed summarizing each dataset by the minimum and maximum times of its observations, the minimum and maximum values of observed environmental variables (in their source units), and by the "footprint" of the geospatial location. These items together constitute features of the dataset's contents that are brief, can be displayed easily in a form similar to a text snippet, and are suitable for search terms and search comparisons.

We can represent a desired time or variable range by a one-dimensional interval and, similarly, a geospatial search by a two- or three-dimensional spatial footprint. We approximate similarity between a search and a dataset by a notion of distance; anecdotally, we can describe a dataset as "close" to or "far" from the search, whether we are talking temporally, geospatially, or referring to a variable range. While it is well known that people are inaccurate in their estimates of absolute distance, research shows that they are relatively consistent in ordinal rankings [9]. Thus, if a distance measure provides ordinal rankings similar to those a user would give, it should suffice. We chose a basic measure for simplicity and speed, while recognizing that it is an overly primitive approximation. For each search dimension – such as, space and time – we identify the center of the condition's range. As described in Appendix A, we calculate a similarity score by calculating the distance (in units of search radius) from that center to the closest distance and the furthest distance of the footprint, averaging them, and scaling them by the size of the range. If the feature overlaps the search term, we adjust the score by the percentage of overlap. We average the scores for the search terms into a final score for the dataset.

We wished to test a number of hypotheses about our approach (in each case, our null hypothesis is the negation of the statement):

1. Searchers can relate to a brief summary of a dataset, similar to a webpage snippet used in web search.
2. There is general agreement about what is considered "closer" between collections of data, at least with respect to time and space, and we can approximate this distance (or similarity) in a simple way.
3. Users accept joint comparisons of space and time.
4. Relative distance is more difficult for users to judge consistently for items at similar distances to the target.
5. Geospatially, "closer" makes sense across geometry types, such as points, lines and polygons.
6. Scientists and non-domain experts, in general, have similar views on what constitutes "closer".
7. Our candidate distance measure sufficiently captures our users' intuitive notion of distance.

### 2.1 Methods

Two populations, each of size 20, of scientists and non-domain experts, were asked to respond to a paper questionnaire. The scientists consisted of CMOP professors, post-docs and graduate-level students; these scientists study spatial and temporal distributions of phenomena

[1] Data can be accessed via CMOP's website, http://www.stccmop.org. Some data is not available to the public until quality assurance has completed.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



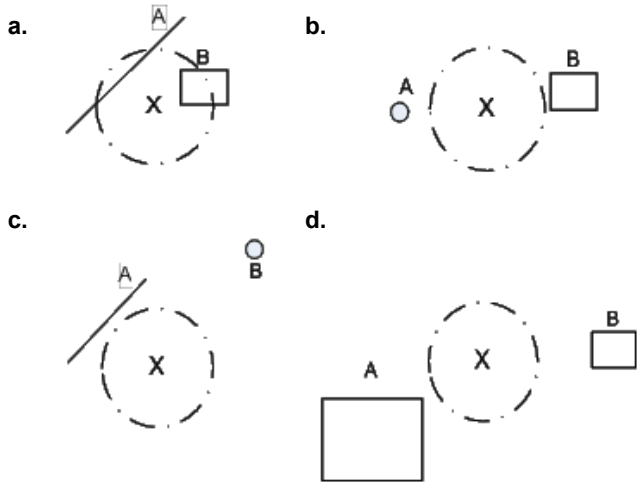



Fig. 1. Four examples of the spatial dataset summary comparisons. The circle marked X marks the area of interest (search). A and B represent the two-dimensional spatial extent of two datasets to be compared to the search circle. In (a) and (b), most study respondents selected B as being closer to the search than A; in (c) and (d), most respondents selected A as being closer.

or populations. The non-domain experts included professors, graduate students and college-educated professionals, primarily in the field of Information Technology. While accustomed to analytical and problem-solving activities, they do not generally search for large scientific, spatial or temporal datasets.

Drawing on psychophysical ordinal-scaling techniques used in cognitive-distance research [9], the questionnaire contained 60 pair-wise comparisons, each between a graphical representation of a search and two datasets represented graphically. Respondents were instructed that, given no other information, they should presume the dataset's contents were spread equally across the entire spatial and temporal "footprint". (Such a uniformity assumption is common in dealing with data summaries, such as in database indexing [10].) Respondents were asked if one dataset (marked A or B) was closer to the search, or whether they considered the two datasets to be equidistant. Each questionnaire included comparisons of just the time feature, just space, and combined space and time features. Some datasets overlapped the search area. The spatial representations included points, lines and polygons; like and unlike shapes were compared. Fig. 1 shows four examples of the spatial comparisons.

### 2.2 Results

Fig. 2 plots the change in respondent agreement against increasing distance between the geometries compared. Two levels of agreement are plotted: the percent of respondents who agreed with the candidate distance measure's assessment of which alternative is closer, as well as that agreement plus the percentage who judged the two options equidistant ("non-disagreement"). While respondents had the option of judging the alternatives equidistant, the distance measure almost always calculates that one is closer, although the difference may be very small. The graph demonstrates that as the difference in distance from the search center to the two shapes becomes small (less than around one-third of the search "radius"), the respondents' level of agreement become inconsistent. In fact, in this range, the respondents often disagree with each other (data not shown), not just with the distance measure. Certain complex configurations or shapes (for example, complex multi-segment lines) increase respondent variability. Plots of time and of time-and-space comparisons (not shown) are almost identical to Fig. 2, despite the difference in search type (time versus space versus space plus time), shapes, and dimensionality (one dimension for time only, two for space only, or three for space plus time).

Visual inspection of respondent agreement across graphs (such as that shown in Fig. 2) of difference in scaled distance between each choice and the search (by the proposed distance measure) revealed a consistent pattern, with strong shifts at approximately .35 and 1 radii difference. In order to statistically test this pattern, we separated the questions into three groups: difference < 0.35, difference > 1, and those between. An ANOVA was used to test for the variation in agreement with the distance measure amongst the three groups. The ANOVA showed the proportion of "equidistant" responses differed significantly among the three groups, $F(2, 56) = 20.45$, $p<<0.001$, with the variability within the "> 1"

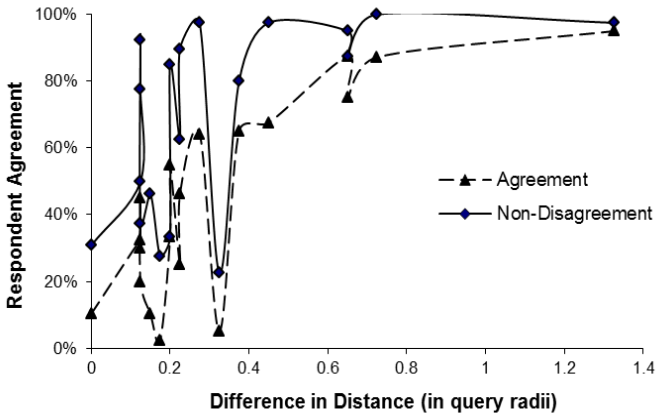



Fig. 2. Level of respondent agreement as a function of the difference in distance of two spatial-only choices, scaled by search radius.

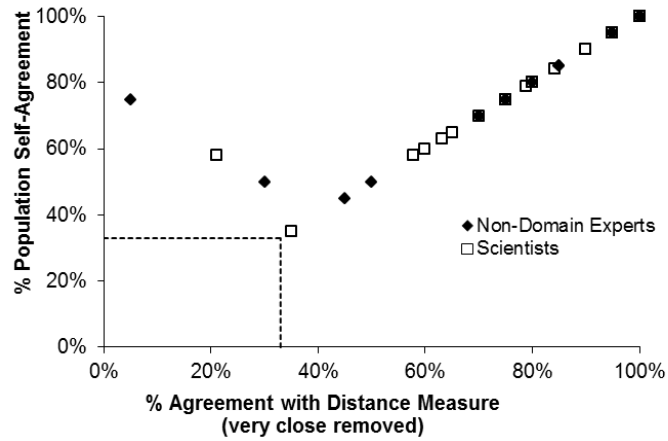



Fig. 3. Percent agreement with distance measure

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



group being smaller than in the "< 0.35" group, as expected. In addition, the proportions of inter-respondent agreement with the proposed distance measure differed significantly across the three groups, $F(2, 56) = 30.93$, $p<<0.001$, with the level of agreement increasing as the difference in distances increases, as expected.

This data is represented in a different way in Fig. 3, which plots for each question the percent of the population that chose the same option. The horizontal axis represents the percent of respondents who agreed with the distance measure, while the vertical axis represents the highest percentage of inter-respondent agreement out of the three options. For example, the left-most point represents a question for which 75% of respondents agreed with each other, but only 5% agreed with the candidate distance measure. Fig. 1d illustrates this case, where most respondents chose A as being closer to the search while the proposed distance measure selected B. Points within 0.30 radii of the same distance from the search center (representing the inconsistency seen in Fig. 2) are removed from Fig. 3. The remaining scattering of points in the top left represent differences of less than 0.35 (according to our distance measure); with greater differences, we see a high level of agreement amongst most respondents and the distance measure. The lower left quadrant, below 33%, is empty since the maximum possible level of disagreement amongst the respondents is when 1/3 choose each option.

The study found only one statistically significant difference between the two populations: scientists had a larger standard deviation in their responses to time comparisons. This difference can be explained by scientists' comments that they included additional factors, such as seasonality, in their assessments of temporal relationships; for example, some regarded September 2002 as "closer" to September 2003 than to July 2002, for certain research questions.

### 2.3 Discussion

There were no questions, comments or objections from respondents in either population with respect to representing dataset contents graphically or as a dataset "footprint", or with the concepts of dataset closeness to a search or ranking datasets by distance from a search.

From this preliminary study, it appears that the candidate distance measure approximates user expectations of which dataset is judged "nearer" when the difference between them is greater than approximately one-third of the search radius. The consistency in relative ordering agrees with findings in spatial cognition literature [9]. We do not consider the inconsistency seen for nearly equidistant datasets a major issue for our measure; such datasets are likely to appear close to each other in a results list. Note also that we cannot be more consistent with our user population than our user population is with itself. Many respondents commented on the difficulty in providing what they felt would be consistent judgments across the different questions; despite this concern, the results we received were remarkably consistent outside of the expected ambiguous cases. While the study focused on expected confounding cases and asked few questions comparing choices with widely different distances, the results are statistically significant, supporting the utility of the candidate distance function as a similarity measure.

Opportunities for improvement in the candidate measure exist where the level of overall respondent agreement with the measure is low. Users appeared to weight the dataset edge closest to the search more heavily than the centroid; it appears that adjusting the distance measure to match that weighting could further improve formula-respondent agreement. The optimal weighting could not be determined from this user study, and remains an opportunity for further research. Other methods of estimating similarity, such as weighting by the data contents (such as using histograms), may also be applicable. In all cases, however, the accuracy of any formula in replicating respondent judgments is limited by the amount of agreement amongst the respondents themselves; where the respondents' responses are highly diverse the formula can at best replicate the most popular response.

Of our study hypotheses, we conclude that there is general agreement about what is considered "closer" to a search with respect to time and space, and that we can approximate this distance in a simple way. We substantiated that "closer" applies across geometry types, such as points, lines and polygons. We confirmed that users understand joint comparisons of space and time and that relative distance is viewed fairly consistently by respondents when the items to be judged are placed at distinct distances to the target search. It appears that respondents can relate to a footprint of a dataset. We conclude scientists and non-domain experts in general have similar views on what constitutes "closer", which provides potential to extend this approach to users beyond the core scientific community.

The results of the user study support the hypothesis that a ranking approach based on the concept of "dataset distance" is feasible. We judged that these results were sufficiently consistent and the candidate distance measure was a sufficiently good approximation to justify implementing these concepts for further testing.

## 3 Catalog and Search Tool

In this section, we describe the test collection and search engine we developed to further test our ideas. We use these components in the second user study, described in Section 4.

### 3.1 Test Collection

We constructed a catalog of the datasets in CMOP's observational archive, and a search tool to operate over it [6], [7]. The catalog acts as our test collection; all searches are performed against the catalog entries. Each dataset within the archive has a corresponding catalog entry. In addition, we may create a catalog entry to represent either a subset or superset of data, in the same way that a book might be listed in a text retrieval system

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



both as a single entry and as individual chapters. Such entries may be linked into a containment hierarchy.

Fig. 4 shows an example of a catalog entry. The entry represents the features gathered from four sources: information such as file name, file type and date last updated is collected from the file system; variable names for the data stored within the file, their data types and units can often be automatically extracted from metadata stored within the file or database; occurrence counts and representative values (currently we are using minimum and maximum values, except for spatial extent) for the variable are produced by reading the file; and lastly, additional metadata may be provided manually by the data owner or curator.

Our overall model of metadata capture is semi-curated. In general, the data curator must configure or code certain options once for each type of data indexed, and these options are used in generating metadata entries for additional datasets of the same type.

A dataset may be a file in one of many data formats, or may exist as a subset of table rows within a relational database. Many scientific disciplines store their data in structured datasets [2], although individual datasets may vary widely in their structure. A dataset might represent a single, detailed observation taken at an instant in time, millions of observations at a single location spanning decades, or thousands of readings spanning hundreds of miles over days or weeks. An archive can have datasets ranging over all these scales. With a few exceptions, the datasets in our test collection are available for public download.

Almost all observations have a three-dimensional location and a time. Scientists frequently search on these characteristics, thus they are important features. We capture minimum and maximum time and elevation. We describe a dataset's geographic footprint by a simplified shape that summarizes the locations of its individual observations as a geographic shape. That shape may be a point (for stationary observation stations), polyline (for mobile platforms), or polygon (for grids or mobile platforms with complex tracks). Since we store the catalog in a geographically enabled relational database, we use built-in tools to summarize the geometry of each dataset's data points.

Table 1 shows the characteristics and counts of datasets by observation class. Using the described combination of techniques, an archive of nearly 800 million observations spanning over a decade is summarized by around 30,400 entries. Of these, around 2,000 entries are "parents", with more detailed subsets defined as "children". Three-dimensional geometries created from environmental simulation outputs were excluded from the test collection as they were early proofs-of-concept and did not yet represent complete, searchable data.

### 3.2 Data Search Tool

The data search tool consists of a user interface and an underlying search engine. Our current user interface, shown in Fig. 5, allows the searcher to specify one or more search conditions. The user can indicate the geographic area of interest either by entering geographic coordinates or by manipulating the box corners on the map (provided by Google Maps). He can choose dates to designate the appropriate time range, and can add one or more variables of interest by choosing from the drop-down list. Available units are then shown for the selected variable; the user can optionally identify a range within which the variable's values are most relevant. If no range is provided, we interpret the search condition as a search for datasets for which that variable is present; we call this condition "variable existence". The user interface design is naïve; alternate designs are possible, but are not the focus of this research.

| Summary Field | Dataset Example |
|---|---|
| Dataset id: | saturn01.ctd.201005 |
| Description: | Saturn-01 Profiler, May 2010 |
| Quality: | "Verified" |
| # Observations: | 247,377 |
| Data Location: | http://... |
| Data Format: | NetCDF |
| Times [start:end]: | 2010-05-14 :2010-05-31 |
| Geolocation, datum: | Point(-123.87,46.23), WGS84 |
| Elevations, datum: | -13 .. 2.5 (m), NGVD27 |
| Variables (units) [values]: | salinity (psu) [0 :29.6]<br>temperature (C) [8.2 :14.6]<br>time (secs since epoch) [1,273,869,578 :1,275,378,800] |

Fig. 4. Representation of sample dataset metadata

The scoring formula in our search engine adapts a feature-space model. The fit of each search term to the relevant features in a catalog entry is estimated by our proposed similarity measure. We use a distance-based measure that compares the search range to the data range; we further use the search term interval size or range (and units, if necessary) as a normalizing "yardstick"; this approach creates a unitless distance that is implicitly weighted by the user. Each resulting normalized distance estimate is converted to a score for that search condition, and the scores combined. We detail our similarity measure in Appendix A.

Conceptually, each search generates a rating of every catalog entry (although some implementations may avoid doing so). Each search condition is evaluated and a score calculated for the entry. The resulting scores are

TABLE 1. COUNT OF TEST COLLECTION CATALOG ENTRIES, REPRESENTED OBSERVATIONS AND GEOMETRY

| Observation Class | Geometry | Count of Entries | Total Observations |
|---|---|---|---|
| Stationary platform | Point | 14,648 | 744,174,016 |
| Stationary, variable depth | Point | 6,677 | 42,850,403 |
| Mobile, fixed depth | Point, line, polygon | 7,938 | 3,922,736 |
| Mobile, variable depth | Point, line, polygon | 1,161 | 6,982,008 |
| Totals | | 30,424 | 797,929,163 |

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



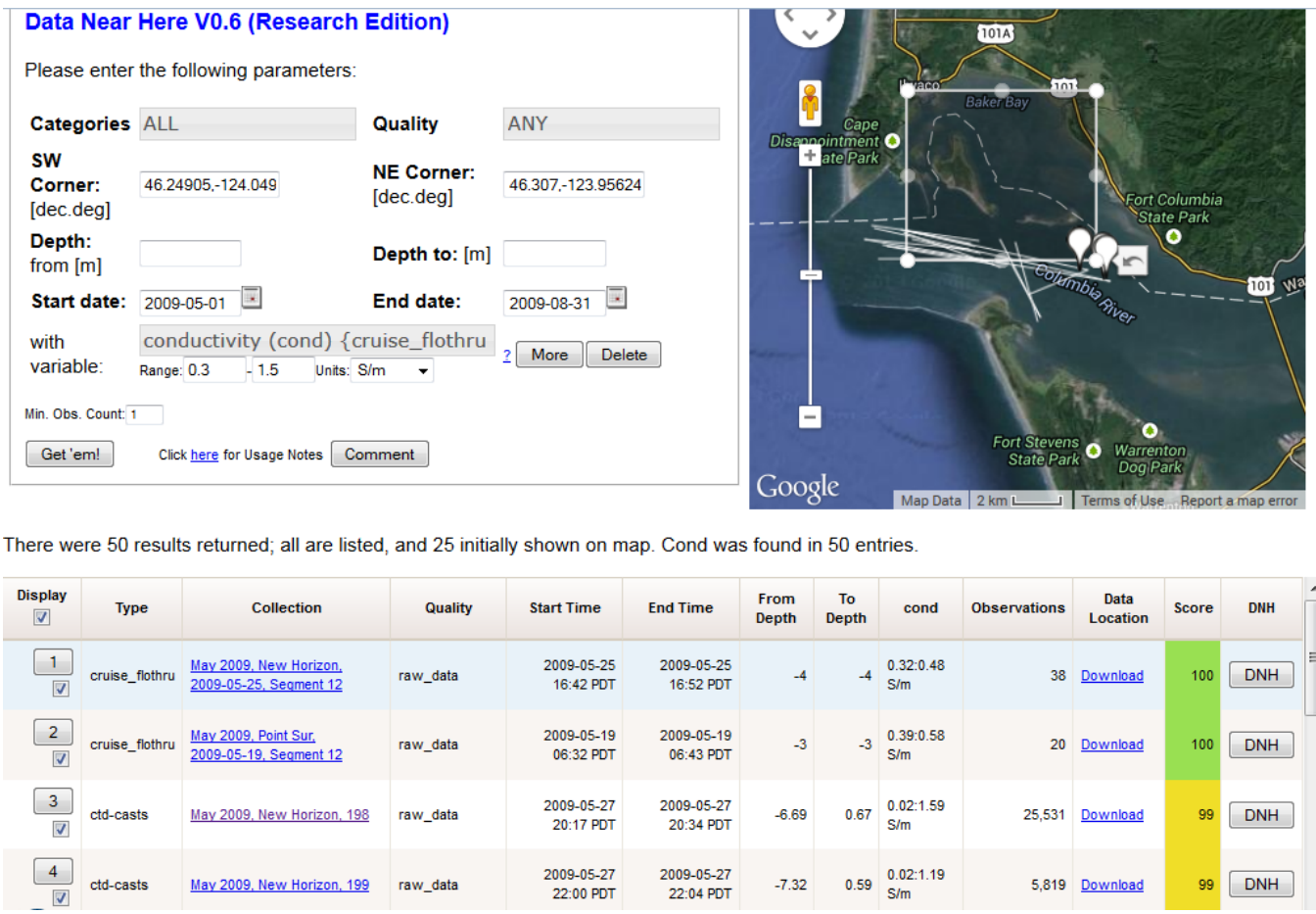
sorted and the catalog entries with the highest scores returned. Appendix B describes the algorithm. Results are returned to the user in the form of a ranked list of dataset snippets, as shown in Fig. 5. A user can directly download any of the datasets. He may also navigate to a page showing details about the dataset's contents, including the kind of details in Fig. 4 and any containment relationships to other datasets. The details page also supports plotting variables from the dataset in a data-analysis tool.

## 4 User Study 2: Fidelity

With a catalogued test collection and prototype search engine in hand from the work described in Section 3, we wished to test the system behavior against the desired qualities. We explore two questions: First, is the search tool a useful one? Second, does the proposed scoring and ranking method provide a good – or "good enough" – approximation of user views of comparative relevance?

To address the first question, we asked respondents (described below) questions regarding the overall performance of the system in responding to their information need, separate from rating dataset relevance. Sanderson reviews studies that show that little agreement exists between IR measures and user satisfaction [11]. Su found that value of search results was more highly correlated with search success than precision [12]. Su's test population was similar to ours (Ph.D. students and faculty members in scientific disciplines); if the precision reported by our study was high but user satisfaction low, or vice versa, those results would influence our future research.

Researchers have shown that relevance judgments are inconsistent across judges [5], [13]. A less frequently discussed concern is the gap between the person with the actual information need and the relevance ratings made by assessors for specific documents. The need as interpreted by the assessor may be different from that intended by the person who originally framed the description of the need for use in the study; this gap would obviously influence the relevance judgments made. We address this concern by asking each study respondent to use one of his own information needs as the source for his searches, and asked him to judge the relevance of the returned items for his own searches. While we might lose some theoretical repeatability (although it does not appear that repeatability has been proven in text retrieval [5], [14]), we gain insight into the applicability of the approach and implementation with this rating scheme.




There were 50 results returned; all are listed, and 25 initially shown on map. Cond was found in 50 entries.

| Display | Type | Collection | Quality | Start Time | End Time | From Depth | To Depth | cond | Observations | Data Location | Score | DNH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | cruise_flothru | May 2009, New Horizon, 2009-05-25, Segment 12 | raw_data | 2009-05-25 16:42 PDT | 2009-05-25 16:52 PDT | -4 | -4 | 0.32:0.48 S/m | 38 | Download | 100 | DNH |
| 2 | cruise_flothru | May 2009, Point Sur, 2009-05-19, Segment 12 | raw_data | 2009-05-19 06:32 PDT | 2009-05-19 06:43 PDT | -3 | -3 | 0.39:0.58 S/m | 20 | Download | 100 | DNH |
| 3 | ctd-casts | May 2009, New Horizon, 198 | raw_data | 2009-05-27 20:17 PDT | 2009-05-27 20:34 PDT | -6.69 | 0.67 | 0.02:1.59 S/m | 25,531 | Download | 99 | DNH |
| 4 | ctd-casts | May 2009, New Horizon, 199 | raw_data | 2009-05-27 22:00 PDT | 2009-05-27 22:04 PDT | -7.32 | 0.59 | 0.02:1.19 S/m | 5,819 | Download | 99 | DNH |

Fig. 5. User interface for "Data Near Here", showing a sample search for a geographic region (shown as a rectangle on the map) and date range, with temperature data in the range 5-10C. Result datasets (or subsets) are shown as points and lines in the output pane, together with their relationship to the search region. In the section of ranked list of results visible here, two full matches for the search conditions were found; two partial matches to a search with time, space and a variable with limits are listed, and more are shown on the map.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



### 4.1 Methods

Our second study followed a common IR approach, adapted appropriately for datasets. The study used a convenience sample of 12 scientists. These scientists were professors, post-docs and graduate-level students, all intended future users of the tool at CMOP and existing users of CMOP's data archive. None of the scientists had previously used the tool.

The search tool's user interface was modified for the user study, adding features to administer survey and rating questions and capture the responses. In addition, the results page was modified to return exactly 100 results, if necessary including results judged by the system to have low relevance. The searches and survey responses were captured using Google Analytics. No data was captured that linked a respondent to his or her responses or searches.

The study procedure is summarized in Fig. 7. Each respondent was given a ten-minute tour of tool operations; the same information was provided as an appendix to the survey instrument. The instructions then asked her to think of a recent information need for data supporting her research, and to perform three or more searches. In order to collect a range of searches we asked for (at least) three different combinations of conditions: one search using only combinations of location, time and (if desired) elevation constraints; one search adding a variable existence condition to the search; and the third search adding constraints on the values of the variable (minimum and maximum values, in some units). In order to capture searches representative of real operations, no restrictions were specified on the kind of information need, locations, times, or variables to be used. The respondent was asked to review the results returned for the search.

1. Respondent is given a brief overview of tool usage, and is given the opportunity to familiarize themselves with the tool, if desired.
2. Respondent considers a recent information need. Respondent formulates the need as a set of 3 or more searches. The set should contain at least one of each of the following types of search: a spatial and/or temporal search; a variable existence search; a search containing variable limits.
3. For each search:
   a. Respondent enters search conditions.
   b. System retrieves and presents a ranked list of 100 items. Respondent briefly reviews results, then proceeds to "survey" step.
   c. System presents the 5 qualitative questions. Respondent rates questions using Likert scale.
   d. Systems presents 25 datasets selected from the results, in random order. Respondent rates each dataset on a 4-point Likert scale from "not relevant" to "relevant".

Fig. 6. Second user-study process

To measure tool utility, we asked five questions for each search, shown in Table 2. Our questions were adapted from Su [12] and represent the major categories of search success (Question 1), utility (Questions 2, 4 and 5), efficiency (Questions 4 and 5), and user satisfaction (Question 3). The answers were rated on a 7-point Likert scale, with 7 = excellent and 1 = not at all.

After answering the five questions, the scientist was presented with a subset of 25 of the 100 results. Included in the list of 25 were the top 10 results returned, the lowest three in the list, and 12 randomly chosen items. These items were chosen to ensure that we could report traditional "at ten" IR measures; they also ensured variety in the items presented for rating, in the absence of a large collection of pre-existing ratings. The dataset score and position on the original list were removed, the 25 items were ordered randomly, and the items were renumbered. The scientist was asked to rate the relevance of each result to her search. We used a four-point scale (3 = high relevance, 0 = no relevance ) adapted from Sormunen [14]. Our focus was on the search behavior, as the archive is already known to fulfill only a subset of scientist information needs. Our analog to topic relevance is the applicability of the dataset's contents to the searcher's search. Our chosen analog of Sormunen's "degree of topical relevance (the extent to which the text discusses the topic)" is the proportion of a dataset's contents that the user believes is directly relevant to the search. These choices allow independence of relevance from dataset size while providing the same intent as topic coverage; a small dataset of highly relevant observations may be more useful than a large dataset with few relevant observations.

### 4.2 Results

The 12 scientists returned 35 responses during the study period. Of these, five were tests submitted with no ratings, leaving 30 usable responses.

#### *4.2.1 Results for Overview Questions*

In order to better understand search-type differences we present the overall results, then break out the searches by type: geospatial-temporal only, searches with variable existence, and lastly searches with limits on variables. Fig. 7 shows the results graphically; Table 2 presents the median and interquartile range for each question, for all searches and by search type. The median for each question is 6 (very good) or better. With the exception of question 3, "confidence in completeness", answers were clustered fairly closely about the mean.

To assess the overall utility of the tool, we compared the proportion of high scores (> 4) to low scores (< 4) for each question, using a two-sample test for the differences in proportions. Results are shown in Table 3. The high-satisfaction responses to the overall-success question were statistically significant in all cases. Separating out by search type, the geospatial-temporal and variable-existence searches had statistically meaningful high responses. We could not calculate the z-statistic for a

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



TABLE 2. RESPONSES TO USER SATISFACTION QUESTIONS AND COMPARISON OF HIGH VS. LOW SCORES: ALL SEARCHES AND BY TYPE

| Median [Interquartile Range] High versus Low Scores: z-score (probability) | All (n=30) | Space + Time (n=8) | Variable Existence (n=13) | Variable with Limits (n=9) |
|---|---|---|---|---|
| 1. How successful was this search in helping with your information need? [success] | 6.5 [0.5] | 6.5 [0.5] | 7 [1] | 6 [2] |
| | 2.79 (<0.01) | NM | 2.05 ( 0.03) | 1.04 (0.16) |
| 2. How well does this style of query allow you to express your information need? [qryexpr] | 6 [1.0] | 6 [0.25] | 6 [0] | 6 [1] |
| | 3.74 (<0.01) | NM | NM | 1.84 (0.05) |
| 3. How confident are you in the completeness of search results? [confcomp] | 6.5 [2.5] | 6 [1.25] | 7 [3] | 6 [4] |
| | 2.02 (0.03) | NM | 1.40 (0.09) | 0.61 (0.28) |
| 4. Was using this tool quicker than finding the most relevant results by other means? [quicker] | 6.5 [0.5] | 6.5 [0.75] | 7 [1] | 6 [0] |
| | NM | NM | NM | NM |
| 5. How valuable are the search results versus time expended? [time/effort] | 7 [1.0] | 7 [1.25] | 7 [1] | 7 [1] |
| | 2.98 (<0.01) | NM | 1.78 (0.05) | NM |

number of the search subsets or for the "quicker" question as there were no low scores in the responses for these sets (shown as "NM", not meaningful). In all combinations but three, the high satisfaction responses were statistically meaningful. The exceptions are all in the variable-with-limits searches. For these questions, the median response was highly positive, but the variance high. Responses for overall search success were highly correlated with the other responses (Pearson's r for correlation of overall search success with search expression, 0.72; with confidence in completeness, 0.85; with quicker, 0.98; with time versus effort 0.95; n = 30, p<.0001 in all cases).

### 4.2.2 *Ratings Results*

Of the 30 usable responses, two judged relevance for three or fewer datasets. These two responses were excluded from the search-level analysis, leaving 28 usable searches with associated dataset judgments. The mean number of judgments for these remaining searches was 24.5, as a few datasets were not rated.

For eight searches, all four values (0-3) were assigned to datasets; an additional seven searches assigned three values, and six searches assigned only two values. In seven searches all datasets were given the same rating. In six of these cases, all datasets were rated as highly relevant; five of these six were variable existence searches. In one case, all were rated as not relevant. Not surprisingly, the searches with "highly relevant" (value of 3) assigned to all datasets were associated with high satisfaction measures, whereas the sole search with "not relevant" (value of 0) assigned to all datasets was associated with the lowest satisfaction measures in the study. Even for this search, however, the "quicker" and "query expression" scores were high, signifying that even when no relevant data is found, the fact that this situation can be ascertained quickly is likely to be of value. This experience is consistent with Su's findings [12]. As our relevance

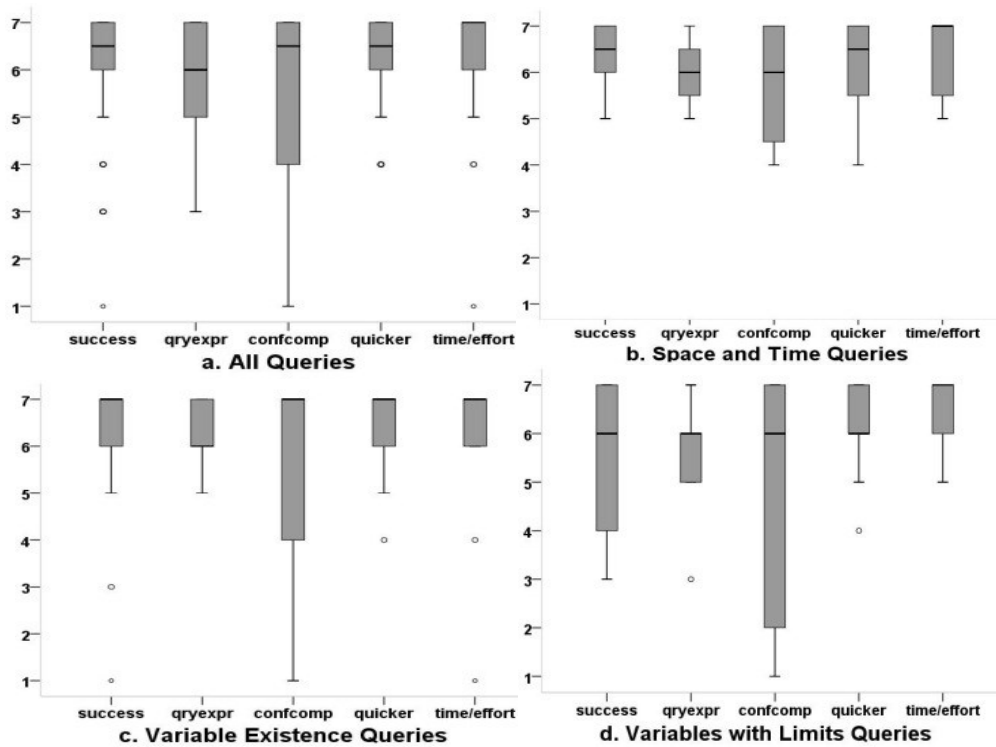



Fig. 7. Summary results for user-satisfaction survey questions. The questions are shown in Table 2.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



data is similar to that collected for IR studies such as TREC and INEX, we believe using IR metrics is justified.

In Table 3 we report precision measured at rank 10 (P@10) and mean reciprocal rank (MRR). Precision at rank 10 gives a measure of the number of relevant documents found in the top ten returned; MRR measures the average position of the first relevant document found [15]. We include in these measures all datasets judged to have any relevance; this choice is in line with the threshold of relevance used in binary evaluations [15]. We report measures for all relevance levels together (P@10, MRR). Overall mean precision at rank 10 was 0.96. Overall MRR was 0.95, and was 1.0 for two search types. In one variable-existence search all datasets were found to be not relevant; no dataset from any other top 10 was rated not relevant. We also report separately precision and MRR for the combination of the "medium" (2) and "high" (3) relevance ratings (denoted as 2+3@10 and MRR2+3); likewise, we report separately precision at 10 and MRR for the "highly relevant" (3) ratings only (denoted as 3@10 and MRR3). As before, we report these measures for the full set of searches, then for each search type separately. Even excluding low-relevance datasets, precision at rank ten and MRR remain respectable, but highlight areas for possible exploration and improvement. Analysis of ratings below position 10 is presented below.

Measures of recall target the completeness of search results, assessing the likelihood of missed, relevant items (false negatives). Calculating recall requires knowing the set of all relevant items for a search. Rating every item in a large corpus can be prohibitive in terms of human effort. Collections such as TREC and INEX approximate the relevant set by only judging items returned by one of the different test systems, and combining those results (thus underestimating the relevant set and overstating recall [15], [16]). In our case, such a pooling strategy is unavailable, as different subjects judged relevance on different queries. The alternative of having each subject rate all 30,000+ datasets on each of three queries was impractical.

We focused instead on assessing "effective recall". Typical behavior for a user examining a ranked result list is to inspect items working down from the top, but stopping if a sequence of low-relevance items is encountered, as lower items are assumed to also be of low relevance. Hence we have concentrated our evaluation of search effectiveness on how well our ranking order corresponds with subjects' relevance judgments, as a relevant item can be "hidden" by higher-ranked non-relevant items, becoming a false negative for practical purposes; this is measured by RBP in Section 4.2.3.

We also conducted a test specifically to detect whether relevant items were being "buried" at the bottom of our ranked result list. We approximate this approach by including items originally ranked 98-100 in the returned list, below the presumed attention span of the user, in the relevance-judgment subset. In Fig. 8 we compare the ratings given to top 10 ranked datasets versus "bottom 3" ranked datasets (positions 98-100). The percentage of top 10 datasets rated as "highly relevant" is significantly higher than the percentage of "highly relevant" in the bottom 3 ($z = 4.63$, $p<<0.001$), despite several searches where all items received the same rating; thus, we conclude that few false positives exist and recall, while not directly quantifiable, is likely to be high. Taking all judged-relevant top 10 datasets as the true positives and all judged-relevant bottom 3 datasets as false negatives, we calculate an estimated recall figure of 0.821.

Question 3 on our survey targets completeness, and the median score there was high. There were low scores on a few individual queries, but as discussed in Section 4.3, some of these are likely due to subjects misremembering properties of datasets they had seen before.

### 4.2.3 *Dataset-Level Results*

A total of 685 datasets was rated. Of these, 351 (51%) were rated as highly relevant; 147 (21%) were rated medium; 118 (17%) were rated low, and 69 (10%) were rated as not relevant. Of the 685 datasets rated, 90 were rated more than once, with 26 of these rated three times. Of the datasets rated more than once, 53 received the same rating each time, while 37 received different ratings. In each of the 53 same-rating cases, the ratings came from the same respondent in the same search set; for example, a respondent rated a dataset as highly relevant for a location and time-based search, then added a variable to the search conditions and found the same dataset highly relevant when it was returned for the modified search. Eight datasets of the 37 were rated

TABLE 4. PRECISION AND MEAN RECIPROCAL RANK (MRR) BY QUERY TYPE AND RELEVANCE JUDGMENT

| | All (n=28) | Space/ Time (n=8) | Variable Existence (n=12) | Variable with Limits (n=8) |
|---|---|---|---|---|
| P@10 | 0.96 | 1.00 | 0.91 | 1.00 |
| 2+3@10 | 0.82 | 0.96 | 0.74 | 0.81 |
| 3@10 | 0.55 | 0.69 | 0.58 | 0.38 |
| MRR | 0.95 | 1.00 | 0.88 | 1.00 |
| MRR2+3 | 0.86 | 0.92 | 0.78 | 0.92 |
| MRR3 | 0.72 | 0.76 | 0.73 | 0.67 |

TABLE 3. COMPARISON OF AVERAGE RBP RANGES AT RANK 25 FOR IDEAL, CURRENT AND ALTERNATIVE SCORING FORMULAE

| **Scoring Alternative** | **Average RBP Range at Rank 25, $p$=0.7** | **Average RBP Range at Rank 25, $p$=0.83** |
|---|---|---|
| Ideal | 0.90 – 0.93 | 0.79 – 0.92 |
| Current | 0.74 – 0.76 | 0.66 – 0.79 |
| SN | 0.72 – 0.74 | 0.64 – 0.77 |
| S2 | 0.80 – 0.83 | 0.70 – 0.83 |
| S3 | 0.74 – 0.76 | 0.67 – 0.79 |
| S4 | 0.73 – 0.76 | 0.66 – 0.79 |
| SX | 0.72 – 0.75 | 0.65 – 0.78 |
| Euclidean | 0.70 – 0.72 | 0.63 – 0.76 |
| Pessimal | 0.36 – 0.38 | 0.36 – 0.49 |

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



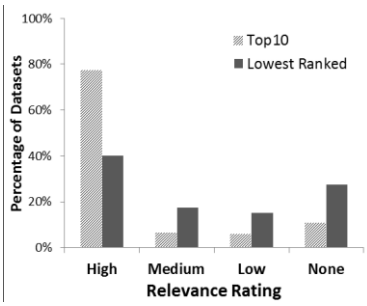



Fig. 8. Proportion of datasets by rating

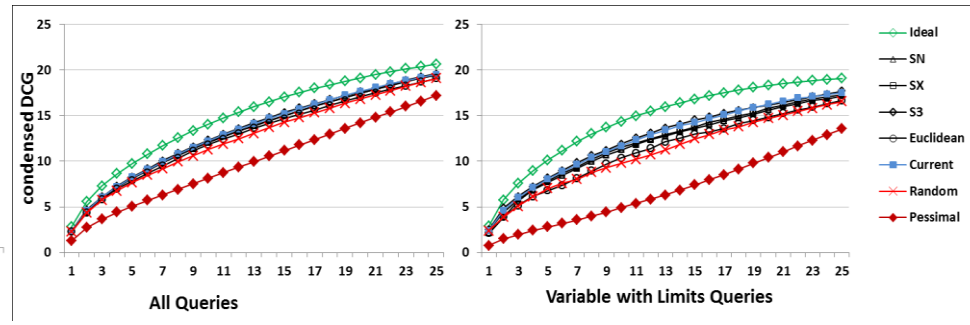



Fig. 9. Condensed discounted cumulative gain. (a) for all searches; (b) for variable-with-limits searches only.

from "highly relevant" to "not relevant"; in each case, the different ratings came from a different search set (hence a different respondent). The original position in the returned list for any single dataset varied from 3 to 97 due to the differences in the search for which it was returned.

We saw no significant difference in the proportion of different ratings of datasets representing different geometries types (as listed in Table 1), that is, datasets represented by points versus lines or polygons. Nor did we see significant differences in user responses between datasets from the lowest level of the hierarchy versus datasets from higher in a hierarchy. We conclude that the subsetting of data and representing subsets as datasets is well-accepted by our users.

We further explored rankings within each search and potential variations of our scoring formula.

We used two methods to explore rankings within each search. First, we applied a compressed version of Discounted Cumulative Gain (DCG) [17]. We have relevance judgments for rank positions one through 10, but we only have relevance judgments for 15 of the datasets in rank positions 11 through 100. Therefore, we condensed the results and treat the judged datasets as though they had been returned in positions 1 through 25, omitting the non-judged datasets. We compare the order of datasets returned with an ideal order for the rated datasets, with all highest-rated datasets returned first, followed by all medium, and so forth. In absolute terms this approach gives arguable results, though without rating all intervening datasets, it is not clear in which direction the results will be slanted. However, since our primary interest is in exploring modifications to the current scoring formula in order to improve ranking results, including and discounting the unjudged items would reduce the differentiation between the reported curves without adding any counterbalancing diagnostic capability. Fig. 9 shows results for all searches and for variables with limits; the plots for space plus time and for variable existence are visually identical to Fig. 9(a).

We tested the current distance measure against five variations (described in Appendix A). These variations modify the distance measure by moving the locus from the center of the dataset's range successively closer to the closest edge, with SN using the closest edge exclusively, while S2, S3, and S4 move successively closer to the center, and SX being further than the center from the closest edge. Based on the insight from the first user study that the closer edge should be a little more heavily weighted, we expected S2 or S3 to perform the best. In Fig. 9, we plot the results from our current scoring formula and three variations: SN, S2 and S3. To contrast our measures with other orderings, we included four controls: the ideal (optimal) and "pessimal" (reverse of the optimal) curves, plus a randomized and a Euclidean ordering of the given ratings (described in Appendix A). Any possible performance curve will be bounded by the ideal and pessimal curves. As can be seen in Fig. 9, relative performance of the current and alternative scorings overlay each other. We applied a one-way ANOVA to the condensed DCG at rank 25 against these alternatives. The results implied that alternatives S2 and S3 might perform 10-20% better than our current approach, but while suggestive, the results were not statistically significant under a Tukey's HSD. The randomized order returned a mean score of around 0, as expected.

Secondly, we applied Rank Biased Precision (RBP) [18] with the extensions for non-binary relevance judgments and for missing judgments. RBP discounts each succeeding position in the ranking by a probability of examination, $p$. Chapelle et al. [19] found in their analysis of Internet search engine click logs that RBP with $p$ = 0.7 closely models user behavior, while DCG overestimates the likelihood of examination of lower-ranked documents. Moffat and Zobel [18] provide a calculation of the RBP accuracy for different result set sizes and values of $p$. Using 25 datasets gives an RBP accuracy to 4 decimal places assuming a "user persistence" factor of 0.7, and to 2 decimal places with a factor of 0.83, which might be expected from a scientific audience. We also felt that asking our scientific users to rate relevance for 25 datasets was testing the limit of their persistence. In order to accentuate possible differences under different scoring formulae, we removed the searches in which all ratings were the same, leaving 22 searches. Using $p$ = 0.7 and $p$ = 0.83, we calculated the mean RBP range for the alternative scoring formulae against all searches. We assumed all unjudged documents were not relevant for the lower bound and assumed all were highly relevant for the upper bound. Results are shown in Table 4, and the average ideal and pessimal RBP ranges are also given. The upper bound is less than the theoretically achievable 1.0, reflecting that ratings below "highly rel-

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



evant" were given to a substantial proportion of returned datasets. With the ideal RBP as our target, we see that scoring alternative S2 more closely approximates user rankings than the current formula. Formula S2 weights datasets even more heavily towards the closest edge than do either the current or S3 formulae. The results varied little across different search types. Again, the results are suggestive but not statistically significant.

### 4.3 Discussion

We were encouraged by the strong, positive response to the search style for expressing the respondents' information needs, especially given that none of the users had used the tool prior to the study. We did not hear of any difficulties or concerns with conflating geographic, temporal, variable existence and variable ranges into a single set of search conditions. Although we did not ask for comments in the study, several respondents approached us with unsolicited comments about their experiences. While we cannot tie respondents to specific searches, we presume that these same experiences flavored their responses to the survey questions.

The biggest frustration respondents expressed was with variable searches. The current prototype treats each column name as a variable name. In cases where different parts of the archive use different names for the same environmental variable (e.g., temperature, airtemp, air_temperature) these are treated as separate variables. At present, only the variable name specified in the search is counted as a match; similar names are not returned. Multiple similar names in a search are treated as separate search conditions. We believe that multiple names for the same variable is one of the key causes of the lower "completeness confidence" scores for searches involving variables. Future enhancements may allow multiple variables to be identified as "the same" for searching purposes. In addition, variable units are not currently standardized; we have experimented with unit translations and believe that this problem is tractable. These concerns are reflected in the wide range of responses for the question concerning confidence in complete results. Despite this spread, in six of the searches for variable existence, all but two datasets were judged highly relevant.

Several respondents commented that the tool did not return datasets that they knew existed and matched their search, leading to reduced confidence in completeness. In several cases the respondent demonstrated a search to us and identified supposedly missing datasets. In each case we investigated, the dataset turned out not to be similar to the search. This effect was most prevalent for searches with variables, where in several cases a long-running observation platform did not have the relevant sensor for that variable during the search time period, or the variable had a different name from that used in the search. The individual searches all focused on locations and time periods in which there were many potentially highly relevant datasets; thus, there are few low-scoring datasets in our judged sample. This effect is the result of two interacting processes: the center collects data in its area of interest, and their scientists are focused on that area of interest.

We found RBP and the condensed DCG useful in exploring the performance of variations of the scoring formula using the existing dataset ratings. We are encouraged by the consistent performance of the scoring approach across the different search types; the distance-centric weighting of data ranges across space, time and variable values seems to produce results relatively consistent with user expectations. Although applying RBP and condensed DCG did not result in statistically significant support for any one of the scoring alternatives over the others, they are suggestive that the weighting in the current formula could be improved, perhaps moving to formula S2. However, given the small differences reported, careful assessment of the effort invested versus the potential improvement is warranted.

Based on the similarities between the results for scientists and IT professionals in our first user study, we do not expect to see significant differences from these results when we expand tool usage to a non-research-scientist population.

## 5 Reflections

So, are datasets like documents? We first discuss similarities, then some differences.

Our scientists easily translated their experience with ranked document search into this new setting with nominal training. Despite their extensive previous experience with database-style Boolean retrieval of data, no concerns were raised about ranked retrieval, representing datasets by summaries, or the contents of the dataset summaries. The users accepted our similarity score and accepted without comment the combination of seemingly different distance units of space, time and variable values. Our overall success ratings were high. We attribute the positive response to the ease with which they can now perform a task they had been struggling with; this functionality, after all, is the goal of our research.

We found it relatively easy to adapt IR metrics to assessing ranked datasets and user ratings. User-study approaches from IR were also easily applied. The areas where we did encounter ambiguity tended to be ones that are also ambiguous in text document retrieval. For example, how should we account for the large number of unrated datasets in our test database? What should the relative weight be for a "highly relevant" rating versus a "medium" one? These issues are familiar to IR researchers.

In contrast to text or XML document-retrieval systems, the searches our system must handle a potential for greater dynamic range in granularity: a scientist may be searching for a single day or week, or for data spanning a year. Search engines usually index material at a single granularity, such as the web page. We have situations where a single sample with many attributes is considered a valuable scientific dataset unto itself, even though many such samples are stored together in a single file. In a similar vein, multiple datasets are stored in

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



a single large database table, but represented as separate datasets in our index. While we began with a view of disjoint datasets as our indexable units, we quickly moved to using additional grouping concepts. In essence, our catalog approximates a concept of "the most useful meaning-bearing unit" [20], recognizing that this unit varies across scientists or even across a single scientist's tasks. We use a hierarchy to intermediate between meaning-bearing units at multiple scales, thus gracefully adapting to differences in research foci. Furthermore, some web-search engines only index a prefix of long documents [21]. In our context, a dataset with a million observations may have subsets with widely varying relevance to a search; a prefix of a dataset might be very unrepresentative of the whole (for example, a cruise that transits from fresh to salt water). Also, treating the whole dataset as a "bag of numbers" (as documents are treated as a "bag of words") does not assist the scientist; the numbers listed in the search may not appear in the dataset at all.

Text retrieval systems currently use a wide variety of ranking criteria. Some of these, such as click frequency, have ready analogs in the dataset world: for example, download frequency as a surrogate for utility. Other criteria, such as reference frequency, will require adaptation of existing scholarly practices; for example, methods to consistently cite a dataset and discipline around retaining accessibility to the cited datasets [2]. Research into how to apply these concepts datasets is warranted.

## 6 Related Work

There is a considerable body of research [16], [22], [23] into ranked relevance of unstructured text documents or XML against text searches; our work focuses on numeric data ranges. Numbers in HTML tables can be extracted and searched [24], but that work focuses on extracting additional semantics. Numbers are also matched by Agrawal and Srikant [25]. Both these approaches assume each "document" is small, by our standards. To our knowledge, ours is the first application of IR techniques to collections of diverse, potentially large, heterogeneous datasets.

Scientific archives support searching for text in metadata associated with datasets [26], [27]; however, these searches are primarily Boolean in nature. In geospatial search, Grossner et al. [28] note that the contents of cataloged digital objects are neither exposed nor searchable. Goodchild [29] notes that most geographic search systems score items based on word matches against metadata without considering the temporal span or geographic content of the items returned. State-of-the-art portals such as Geospatial One-Stop (GOS) [30] and Global Change Master Directory's Map/Date Search [31] allow searches using both geographic and temporal criteria. Three spatial tests are supported (the map view *intersects*, *mostly contains*, or *completely contains* the dataset), and temporal search appears Boolean. In contrast [6], [7], we explicitly rank returned items based on temporal, geographic and variable "distance" of the dataset contents from the search.

The spatial component of our work draws on research in the field of spatialization of data [20], [32]. Spatial cognition researchers have shown that judging relative distance between individual spatialized data points is practical, and that similarity represented as relative distance is naturally understood. We apply their notions of distance more broadly to large sets of combined temporal, spatial and environmental-variable values. Although their work has identified anomalies and inaccuracies in user perceptions at the detail level, we believe a fast approximation of similarity between a search and a dataset has significant value.

Su, [12], [33] and others have discussed the relationships between user satisfaction and IR measures. Chapelle et al. [19] and others have applied DCG and RBP to evaluate systems results for text retrieval. Our user studies were informed by their work, and we adapt their methods to dataset relevance evaluation and for validating the utility of our prototype.

## 7 Conclusion

The prototype system developed during this project is now in use by scientists within CMOP; after a validation period, the system will be made publicly available. We are beginning to incorporate datasets from other sources into the catalog, allowing users to search for data across multiple organizations' archives. Such datasets will be served from their original location, with only an entry added to our catalog. Planned research includes adding mechanisms for similarity of variable names, and for searching over categorical data. We have also discussed applying the techniques described in Section 3 to a different field of science.

We are encouraged by our experiences in applying IR techniques to dataset ranked search, and by the enthusiasm of the scientists for our work. We believe these techniques have broad applicability, and address a need by scientists that will only become greater as data volumes and heterogeneity continue to grow. Large archives of data only have value commensurate with the use and reuse that can be made of their contents; and data cannot be used if it cannot be found [2]. With the constant need to achieve more with fewer resources, tools such as ours are required to reduce the overhead experienced in current research.

Returning to our title question: Are datasets like documents? Our answer is: like enough to profitably apply Information Retrieval search and evaluation techniques.

### Acknowledgment

The authors thank Alistair Moffat for his detailed review and input. They also thank the staff of CMOP for their support. This work is supported by NSF award OCE-0424602.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TKDE.2014.2320737



## REFERENCES

[1] P. Lord and A. Macdonald, "e-Science curation report," 2003.
[2] S. Weidman and T. Arrison, "Steps toward large-scale data integration in the sciences: Summary of a workshop." National Research Council of the National Academies, Aug-2009.
[3] J. K. Batcheller, "Automating geospatial metadata generation", *Computers & Geosciences*, vol. 34, no. 4, 2008.
[4] A. D'Ulizia, F. Ferri, A. Formica, and P. Grifoni, "Approximating geographical queries," *Journal of Computer Science and Technology*, vol. 24, no. 6, pp. 1109–1124, 2009.
[5] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Parts II, III," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, 2007.
[6] V. M. Megler and D. Maier, "Finding haystacks with needles: Ranked search for data using geospatial and temporal characteristics," vol. 6809, Springer Berlin / Heidelberg, 2011.
[7] D. Maier, V. M. Megler, A. Baptista, A. Jaramillo, C. Seaton, and P. Turner, "Navigating oceans of data," in *Scientific and Statistical Database Management*, 2012, vol. 7338, pp. 1–19.
[8] V. M. Megler, "Taming the Metadata Mess," presented at the Phd Workshop for ICDE 2013, Brisbane, 2013.
[9] D. R. Montello, "The measurement of cognitive distance: Methods and construct validity," *Journal of Environmental Psychology*, vol. 11, no. 2, pp. 101–122, 1991.
[10] V. Markl, M. Kutsch, T. Tran, P. Haas, and N. Megiddo, "MAXENT: consistent cardinality estimation in action," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 2006, pp. 775–777.
[11] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas, "Do user preferences and evaluation measures line up?," in *Proceedings of SIGIR*, 2010, pp. 555–562.
[12] L. T. Su, "The relevance of recall and precision in user evaluation," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 207–217, 1994.
[13] S. P. Harter, "Variations in relevance assessments and the measurement of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 47, no. 1, 1996.
[14] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?," in *Proceedings of SIGIR*, 2002
[15] E. Voorhees and D. M. Tice, "The TREC-8 question answering track evaluation," in *TREC*, 1999, vol. 8.
[16] G. Demartini, T. Iofciu, and A. de Vries, "Overview of the INEX 2009 entity ranking track," *Focused Retrieval and Evaluation*, pp. 254–264, 2010.
[17] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.
[18] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 1, p. 2, 2008.
[19] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proc. 18th CIKM*, 2009, pp. 621–630.
[20] A. Skupin and B. P. Buttenfield, "Spatial metaphors for visualizing very large data archives," in *Proceedings of GIS/LIS '96*, 1996, vol. 1, pp. 607–617.
[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1–7, pp. 107–117, 1998.
[22] C. D. Manning, P. Raghavan, and H. Schütze, *An introduction to information retrieval*. Cambridge University Press, 2008.
[23] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM (JACM)*, vol. 7, no. 3, pp. 216–244, 1960.
[24] P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu, "Recovering semantics of tables on the web," *Proc. of VLDB 37*, vol. 4, no. 9, pp. 528–538, 2011.
[25] R. Agrawal and R. Srikant, "Searching with numbers," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 855 – 870, Aug. 2003.
[26] S. L. Pallickara, S. Pallickara, M. Zupanski, and S. Sullivan, "Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections," in *2nd IEEE Intnl Conf. on Cloud Computing Technology and Science*, 2010, pp. 573–580.
[27] A. Rajasekar and R. Moore, "Data and metadata collections for scientific applications," in *High-Performance Computing and Networking*, 2010, pp. 72–80.
[28] K. E. Grossner, M. F. Goodchild, K. C. Clarke, "Defining a digital earth system," *Transactions in GIS*, vol. 12, no. 12008.
[29] M. F. Goodchild and J. Zhou, "Finding geographic information: Collection-level metadata," *GeoInformatica*, vol. 7, no. 2, pp. 95–112, 2003.
[30] "Geospatial One Stop (GOS)." [Online]. Available: http://gos2.geodata.gov/wps/portal/gos. [Accessed: Jan-2011].
[31] "Global Change Master Directory Web Site." [Online]. Available: http://gcmd.nasa.gov/. [Accessed: 19-Jan-2011].
[32] S. I. Fabrikant, D. R. Montello, M. Ruocco, and R. S. Middleton, "The distance-similarity metaphor in network-display spatializations," *Cartography and Geographic Information Science*, vol. 31, no. 4, pp. 237–252, 2004.
[33] A. Al-Maskari, M. Sanderson, and P. Clough, "The relationship between IR effectiveness measures and user satisfaction," in *Proc. of SIGIR*, 2007, pp. 773–774.

**V.M. Megler** is currently a PhD candidate in Computer Science at Portland State University, having received an M.Sc. there in 2012 and a B.Sc. from Melbourne University, Australia. Megler's most recent industry position was as Executive IT Architect at IBM, publishing more than 20 industry technical papers on applications of emerging technologies to industry problems. Current PhD research centers on applying Information Retrieval techniques to scientific data; general research interests include applications of emerging technologies, scientific information management and spatio-temporal databases. V.M. Megler can be reached at vmegler@cs.pdx.edu.

**David Maier** is Maseeh Professor of Emerging Technologies in the Department of Computer Science at Portland State University. His research interests include scientific information management, data stream systems, superimposed information, and declarative cloud programming. Maier has a PhD in Electrical Engineering and Computer Science from Princeton University. He is an ACM Fellow, a Senior Member of IEEE and a member of SIAM. Contact him at maier@cs.pdx.edu.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.