Visual-Interactive Querying for Multivariate Research Data Repositories Using Bag-of-Words

Maximilian Scherer
TU Darmstadt
Interactive Graphics Systems Group
Fraunhoferstr. 5
64283 Darmstadt, Germany
maximilian.scherer
@gris.tu-darmstadt.de

Tatiana von Landesberger
TU Darmstadt
Interactive Graphics Systems Group
Fraunhoferstr. 5
64283 Darmstadt, Germany
tatiana.von_landesberger
@gris.tu-darmstadt.de

Tobias Schreck
University of Konstanz
Data Analysis and Visualization Group
Universitaetsstr. 10
78457 Konstanz, Germany
tobias.schreck
@uni-konstanz.de

ABSTRACT

Large amounts of multivariate data are collected in different areas of scientific research and industrial production. These data are collected, archived and made publicly available by research data repositories. In addition to meta-data based access, content-based approaches are highly desirable to effectively retrieve, discover and analyze data sets of interest. Several such methods, that allow users to search for particular curve progressions, have been proposed. However, a major challenge when providing content-based access – interactive feedback during query formulation – has not received much attention yet. This is important because it can substantially improve the user's search effectiveness.

In this paper, we present a novel interactive feedback approach for content-based access to multivariate research data. Thereby, we enable query modalities that were not available for multivariate data before. We provide instant search results and highlight query patterns in the result set. Real-time search suggestions give an overview of important patterns to look for in the data repository. For this purpose, we develop a bag-of-words index for multivariate data as the back-end of our approach.

We apply our method to a large repository of multivariate data from the climate research domain. We describe a use-case for the discovery of interesting patterns in maritime climate research using our new visual-interactive query tools.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2077-1/13/07 ...$15.00.

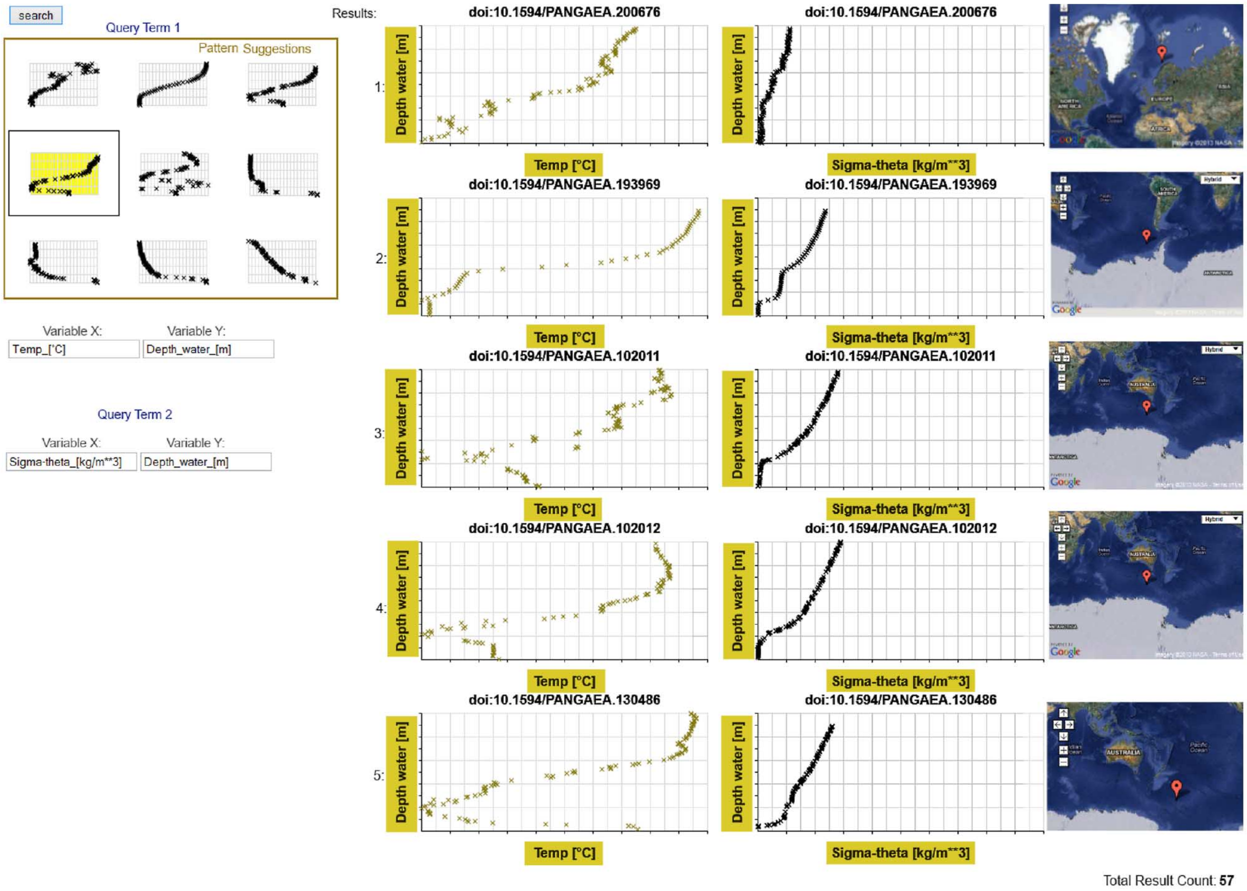
Keywords

Research Data Repositories; Content-Based Retrieval; Bag-of-Words; Query Interfaces; Multivariate Data

1. INTRODUCTION

Multivariate data can be described as tabular data with dimensionality $m \times n$, where n is the number of variables (e.g., water density, water depth or pressure) and m is the number of observations (e.g., time of day or location). Such data arises in many areas of research, industrial production and other commercial applications. Due to increasing efforts in the digital library community over the last decade, such data, particularly that obtained for research purposes, is made publicly available in specialized research data repositories. For example, the PANGAEA repository [6] is a digital library for data-intensive environmental sciences. It hosts very large amounts of earth observation data of various kinds (e.g., time series, multivariate observations, image data, etc.), which are provided for public access. Similar to the search and access paradigms for multimedia databases, content-based access to such repositories has started to receive attention from the Digital Library community. Such access supports users to search and explore data patterns, in addition to annotated textual meta-data. Previous work has considered similarity functions and feature extraction techniques for relevant aspects of research data, including time series, functional, and bivariate data [14, 10, 7, 25], as well as their evaluation [13, 26]. Little research, however, focused on *interactive methods* which use these new similarity functions to help the user with the query formulation process based on data content. Such methods include highlighting of results to show why a document was retrieved, as well as search suggestions to provide the user with an overview of meaningful terms she can search for next. These functions are typically located on the front-end of a visual-interactive retrieval system, but require indexing structures in the back-end to be efficient.

In this work, we present a novel approach for providing the user with interactive search suggestions and result highlighting when querying multivariate data. Such visual-interactive tools are already successfully used in textual search engines and yield similar advantages to users querying non-textual research data documents. Search suggestions provide users with an overview of (often complex) data patterns and vari-



Query Term 1

Variable X: Temp_[°C]

Variable Y: Depth_water_[m]

Query Term 2

Variable X: Sigma-theta_[kg/m**3]

Variable Y: Depth_water_[m]

Total Result Count: 57

search
Pattern Suggestions
Results:
1:
2:
3:
4:
5:
doi:10.1594/PANGAEA.200676
doi:10.1594/PANGAEA.193969
doi:10.1594/PANGAEA.102011
doi:10.1594/PANGAEA.102012
doi:10.1594/PANGAEA.130486
Depth water [m]
Temp [°C]
Sigma-theta [kg/m**3]
Hybrid
AUSTRALIA
Google

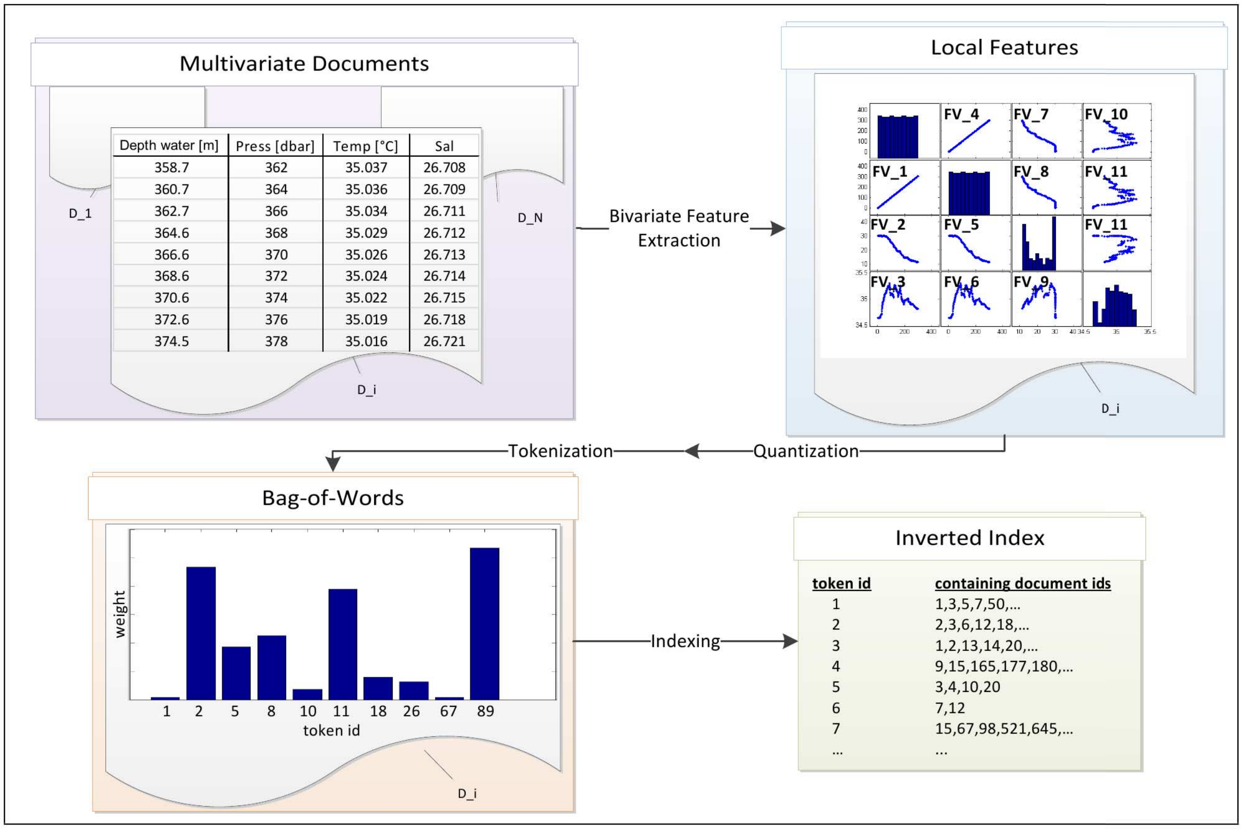
Figure 1: Case Study: We queried for a specific pattern between temperature and water depth to see whether the documents containing this pattern were measured at locations with a similar maritime climate.The first document is situated in the Norwegian sea, while documents 2 to 5 were obtained in the antarctic southern ocean. Both regions have a maritime subarctic climatic zone, explaining why the same pattern was found there. Map Data is attributed to Google Maps.

able names to search for or refine their search; result highlighting shows the user, which part of a document matched her query and thus explains *why* it was retrieved. Figure 1 shows a screenshot of the proposed system in action and illustrates these benefits for the users (the use-case is detailed in Section 4). By searching for a specific pattern of water depth versus temperature, we can find measurements that were obtained in areas with a similar maritime climate. Using only meta-data (the geo-location in this case) such a query would not have been possible.

Akin to search suggestions on popular web- or e-commerce search engines, we present the user with search suggestions and completions, based on her partial query as it is being entered. Figure 4 shows an example of this suggestion-approach. Furthermore we can provide the user with instant search results and also highlight those parts of a retrieved multivariate data-set, that corresponds to the user's query. Similar to paragraph highlighting in text retrieval, we propose to show those scatter-plots of a retrieved data-set, that contain parts of the query (e.g., a textual hit on the axis label or a particular scatter-plot pattern) and to highlight these parts. That way a user can see why a particular document was retrieved in the first place, and quickly skim through the results to find the data-sets she is most interested in. Another example in Figure 5 shows a result list and the highlighted scatter-plots.

To allow for this kind of visual-interactive querying in multivariate data, we develop a novel indexing method based on a bag-of-words approach. The bag-of-words approach has shown to yield state of the art retrieval performance in multimedia databases, e.g., for images, videos or music [12]. We propose to adapt it for retrieval in multivariate research data repositories. The basic idea is shown in Figure 2. By extracting bivariate features from each pair of variables in multivariate data, we obtain a set of local features for each document. We then quantize each feature vector, by assigning the id of the closest cluster centroid (obtained, e.g., via k-means clustering) to each feature vector. Thus we can represent a document with multivariate data by a set of content-based tokens obtained from this quantization. Such a representation allows us to leverage efficient indexing using inverted lists. The details of this indexing approach are described in Section 3.

We provide a case-study of our proposed approach in Sec-



Depth water [m]	Press [dbar]	Temp [°C]	Sal
358.7	362	35.037	26.708
360.7	364	35.036	26.709
362.7	366	35.034	26.711
364.6	368	35.029	26.712
366.6	370	35.026	26.713
368.6	372	35.024	26.714
370.6	374	35.022	26.715
372.6	376	35.019	26.718
374.5	378	35.016	26.721

token id	containing document ids
1	1,3,5,7,50,...
2	2,3,6,12,18,...
3	1,2,13,14,20,...
4	9,15,165,177,180,...
5	3,4,10,20
6	7,12
7	15,67,98,521,645,...
...	...

Figure 2: Overview of our bag-of-words approach for indexing multivariate research data.

tion 4. We index all publicly available research data documents of the data repository *PANGAEA* [6]. We show how our proposed indexing scheme and the newly enabled visual-interactive search tools can be used for this kind of research data documents.

2. RELATED WORK

This work is related to several aspects of digital libraries, multimedia information retrieval and data mining. In the following two subsections we outline recent work related to this paper.

2.1 Content-based and Visual Access Methods

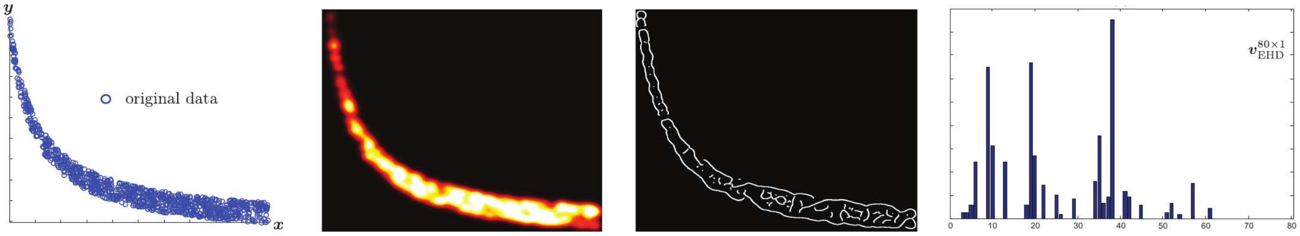
Content-based analysis and indexing is an important research domain within digital libraries to provide additional access paradigms to documents besides access based on annotated meta-data [20].

Examples of recent digital library systems that provide different means of content-based access include systems for 3D models and classical music [3], images [23, 5], time-series data [1], climate data [25] and chemical data [16]. On top of access via annotated meta-data, these digital library systems extract domain-specific *descriptors* from the underlying data as a basis to implement distance functions in support of search and access functionality. Such access includes query-by-example, e.g., supplying an example image and retrieving similar images [23, 5]; query-by-sketch, e.g., drawing a shape and retrieving similar 3D models; or content-based layouts, e.g., clustering time-series by data similarity and presenting the user with an overview [1].

Visual access methods have shown to be highly successful for providing overview and search functionality for users in the Digital Library domain [9]. Effective interfaces can help to more effectively browse, search or analyze large data repositories [31]. The idea behind many approaches is based on Shneiderman's Visual Information Seeking Mantra to provide overview first and details on demand [27]. A recent example of such a system in the digital library context was presented in [2]. There, by analyzing meta-data and time-series based content at the same time, this system generates an interactive layout of research data to enable the discovery of interesting co-occurrences of meta-data based and time-series based patterns. Such approaches can combine traditional meta-data based and content-based methods and can extend the standard search support with elements of explorative search systems useful for hypothesis generation [30].

2.2 Bag-of-Words

The focus of this work is to provide the user with a set of interactive retrieval tools which can respond in real-time to user interaction. Our interactive approaches include instant display of search results, highlighting and search suggestions for querying multivariate research data. All of these functions require an efficient computation of similarities. A suitable, efficient content-based indexing method to this end is



(a) Input data (b) Gaussian kernel density (c) Detected edges (d) Edge histogram

Figure 3: Bivariate feature extraction: Given bivariate input data (a), estimate Gaussian kernel density (b), apply canny edge detector (c) and compute edge histogram descriptor (d). This algorithm by Scherer et al. [26] has shown to yield state of the art performance for bivariate data retrieval.

the bag-of-words (BOW) approach that has become highly popular in multimedia information retrieval. It has been shown to yield state-of-the-art retrieval performance in different domains, including image and music retrieval [17, 22, 12]. In this paper we transfer it for the first time to the domain of multivariate research data. BOW approaches originate from text retrieval and natural language processing, where the inherent tokenization of textual documents was used for proposing efficient indexing and term weighting methods [24].

It was first applied to multimedia documents by Sivic and Zisserman for content-based image retrieval [28]. The basic approach is to extract local features, e.g., SIFT [19] or SURF features for images, quantizing these features via k-means or other suitable clustering methods [32], and finally indexing / weighting these tokens using techniques like tf-idf [24], (probabilistic) latent semantic indexing [11] or latent Dirichlet allocation [4]. This allows for similarity measurements between multimedia objects via their associated bag-of-words (usually the terms are encoded as a histogram), as well as querying or clustering the documents via specific terms (e.g., a predominant color in an image). Most recently, such a bag-of-words approach was also applied successfully to the retrieval of 3D models and 3D scenes [8] as well as to the retrieval of time-series data [18].

3. APPROACH

In our approach, we provide search suggestions and highlighting for querying multivariate data documents. To perform the required computations at interactive rates, we need efficient similarity functions for multivariate data. Therefore, we base our approach on constructing and utilizing a bag-of-words index.

In the following subsections, we first describe the construction of the index itself and then describe the interactive feedback functions for retrieval of multivariate data. We present several examples for retrieval, suggestions and highlighting using our proposed approach using data described in Section 4.

3.1 Bag-of-Words Indexing

As a basis of our approach we provide data indexing, whereby we adapt the bag-of-words approach to multivariate (tabular) data. The flow-chart in Figure 2 gives an overview of the required algorithmic steps. We describe how we adapted each of the these steps for indexing multivariate data.

1. Feature Extraction: extract a set of n local feature vectors $\vec{v}_i$ for each data object (scatter plot in our case)

2. Quantization: quantize each of the feature vectors $\vec{v}_i$ for all documents.(Offline Step: Training a quantizer model $q(\vec{v})$)

3. Tokenization: combine the quantized features with additional categorical information

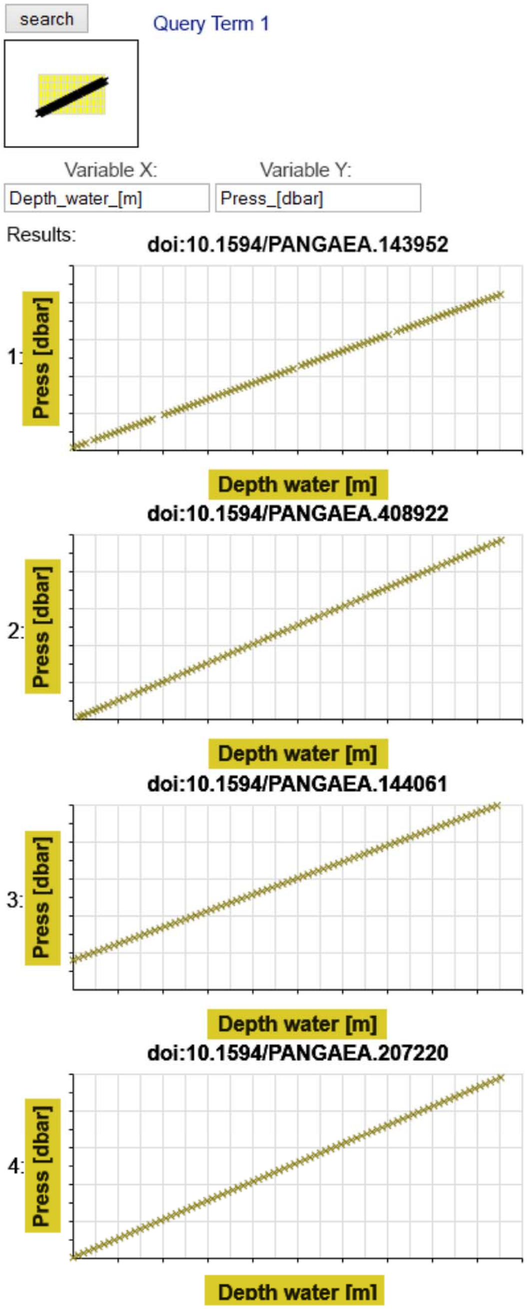
4. Term Weighting: assign a suitable weight to each obtained term

5. Indexing: build inverted lists containing the relevance of a given token (quantized feature vector) for every document

Step 1: Feature Extraction.

We consider multivariate data documents that contain tabular data with dimensionality $m \times n$. In practice that means n different variables (like water density, water depth or pressure) where measured m times. To extract a set of feature vectors from such a document containing multivariate data, we propose to compute all bivariate variable combinations, and compute a feature vector from each of these two-dimensional point-clouds (scatter-plots). Much like the previously mentioned SIFT or SURF features for images, these feature vectors are also *local* in the sense that they represent a local pattern (bivariate) in the whole (multivariate) document.

Based on previous results on feature extraction and benchmarking for bivariate data [26], we will use the algorithm that yielded the best overall results in the benchmark: EDH. It is based on the MPEG-7 descriptor "edge histogram detector" used in shape retrieval. The basic idea of EDH is to render the actual scatter-plot of the bivariate data using Gaussian kernel-density estimation. During this process the scatter-plots is min-/max-normalized, resulting in translation and linear-trend invariance. Then an edge filter is applied to this rendered image and the orientation of the resulting edges are extracted as a histogram. Figure 3 shows an illustration of this extraction process.



Variable X: Depth_water_[m]

Variable Y: Press_[dbar]

Results:

Figure 4: An example query using the *PANGAEA* data repository. We searched for a linear relationship between water depth and water pressure. This tri-gram (variable x, variable y, curve form) we searched for is highlighted in each retrieved dataset.

The result of this feature extraction step is a set of 80-dimensional feature vectors for each multivariate document. The number of feature vectors equals the number of possible scatter-plots, $n \cdot (n-1)$ for multivariate data with n dimensions.

Please note that the benchmark used for evaluation of bivariate descriptors [26] cannot be directly applied to our case, as we consider multivariate data retrieval.

Step 2: Quantization.

The result of the feature extraction is a set of feature vectors for each document. Since we need to obtain a set of tokens for each document, we train a quantization model that is suitable to project each input feature vector to a categorical integer value (the id of the codebook entry). There is a wealth of clustering algorithms suitable for this task [32]. We chose k-means clustering as this has shown good performance for image-retrieval tasks at reasonable computational costs. We choose a random subset of all feature vectors $\vec{v}_i$ and compute a k-means clustering on this subset. The number of clusters k was set to 5000 based on the literature for a compromise between discriminativeness and computational cost [33].

We can then represent an unlabeled feature vector by computing the nearest of the k centroids and assigning the ID of this centroid as the token for this feature vector.

Step 3: Tokenization.

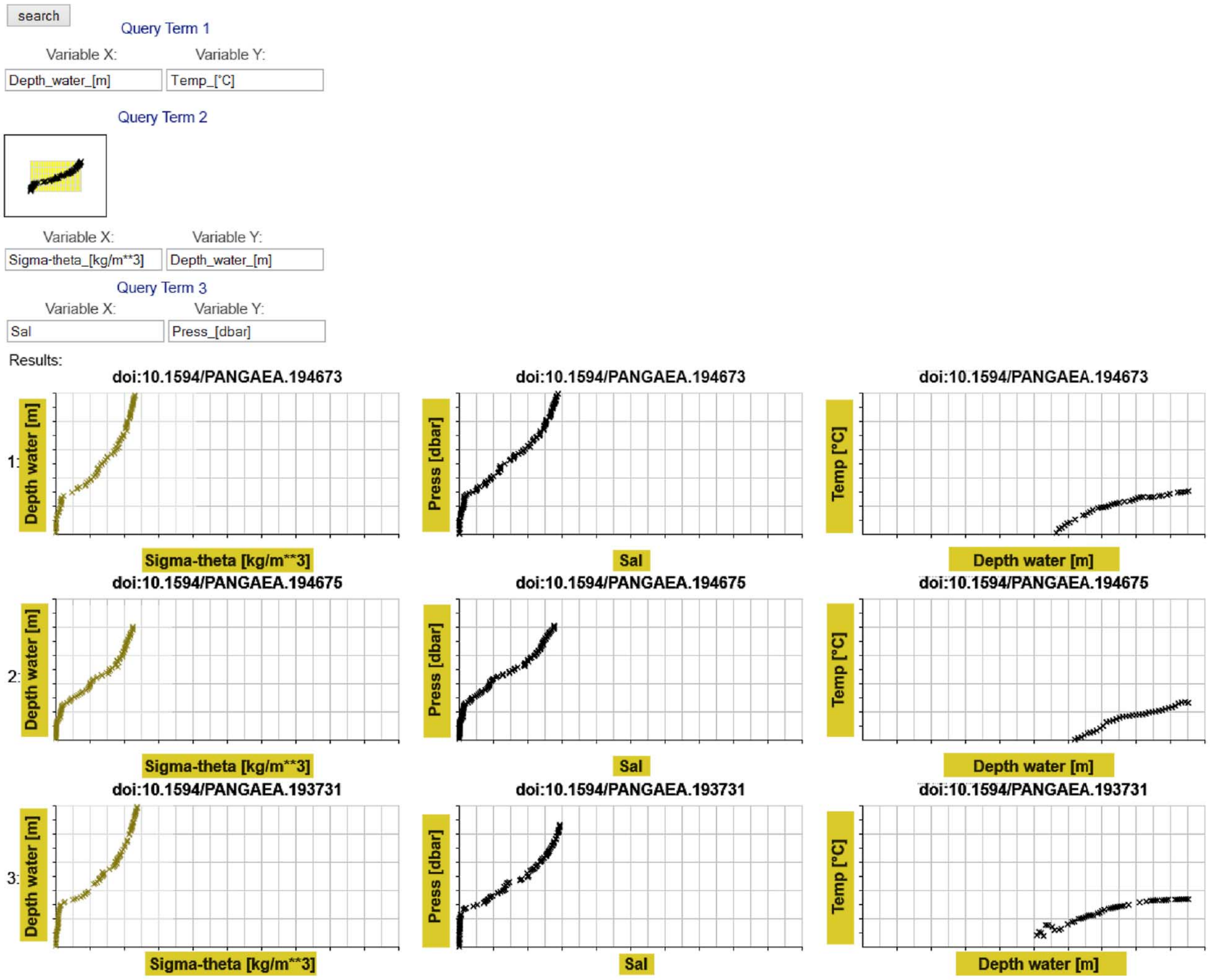
Once we quantized the feature vectors of each document and obtained the categorical cluster ids, we tokenize the document. Since we are not only interested in bivariate data patterns (which are now encoded in the quantized features), but also in the variable combination that exhibits this pattern, we index the data tokens as all possible uni-, bi- and tri-gram terms. For example, if the feature vector of the scatter plot of variable a versus variable b was quantized to cluster id c, we would obtain the terms a, b, c, a_b, b_c, a_c, a_b_c.

Step 4: Term Weighting.

After obtaining a set of terms for each document, we have to choose a weighting scheme for these terms to allow for ranked retrieval (instead of just Boolean retrieval). A straight forward scheme to measure the relevance of a term to a given document is *term frequency* – the number of occurrences of a term in a document. This, however, is not suitable for our terms as the tri-grams – by construction – occur at most once in a given document. Hence, we propose to use the distance to the closest cluster centroid in feature space as the relevance of those terms respectively. As an alternative, we also experimented with introducing an inverse-document frequency (idf) weight, which had little to no effect on retrieval performance and was thus discarded.

Step 5: Indexing.

The final step is to index the set of weighted terms of each document. For each given, distinct term we build an inverted list. This means, that we save a hashed look-up from each term to each document that contains this term along with the associated weight. This allows for ranked retrieval by intersecting the inverted lists of each search term, aggregating the term weights and sorting them in descending order. This approach scales very well. The required main memory for this indexing structures increases linearly with the number of indexed documents. Retrieval time is constant with respect to the index look-up and is dominated by the time required to read the document data from the hard disks.



Query Term 1

Variable X: Depth_water_[m]

Variable Y: Temp_[°C]

Query Term 2

Variable X: Sigma-theta_[kg/m**3]

Variable Y: Depth_water_[m]

Query Term 3

Variable X: Sal

Variable Y: Press_[dbar]

Results:

doi:10.1594/PANGAEA.194673
doi:10.1594/PANGAEA.194673
doi:10.1594/PANGAEA.194673
1:
Depth water [m]
Sigma-theta [kg/m**3]
Press [dbar]
Sal
Temp [°C]
Depth water [m]
doi:10.1594/PANGAEA.194675
doi:10.1594/PANGAEA.194675
doi:10.1594/PANGAEA.194675
2:
Depth water [m]
Sigma-theta [kg/m**3]
Press [dbar]
Sal
Temp [°C]
Depth water [m]
doi:10.1594/PANGAEA.193731
doi:10.1594/PANGAEA.193731
doi:10.1594/PANGAEA.193731
3:
Depth water [m]
Sigma-theta [kg/m**3]
Press [dbar]
Sal
Temp [°C]
Depth water [m]

Figure 5: Search Result Highlighting on our front-end: For each of the retrieved multivariate documents, the scatter-plots that contain the search terms are highlighted by coloring the axis labels and/or the data points in yellow. We searched for multivariate documents that contain a sigmoid like relationship between water density (sigma-theta) and water depth, as well as an arbitrary relationship of water depth versus temperature and salinity versus pressure. Note that the retrieved documents contain all of these search terms; we highlight the results accordingly to show the user, why these documents were retrieved.

3.2 Retrieval

We use the bag-of-words index described in the previous subsection for content-based retrieval. A user can search for arbitrary meta-data, parameters, parameter combinations, data pattern id or a specific relationship of a versus b with pattern c. Figure 4 shows an example search query. In this example, we searched for a complete tri-gram by specifying both axis labels (water depth versus pressure) and a data pattern (a linear relationship between those two variables).

As with any full-text index that is based on inverted lists, we can efficiently combine several search terms by intersecting the associated lists. Thus, the default behavior of our approach is to look for all search terms, and only return those documents, that contain every search term and rank them according to their aggregated term weight as described above.

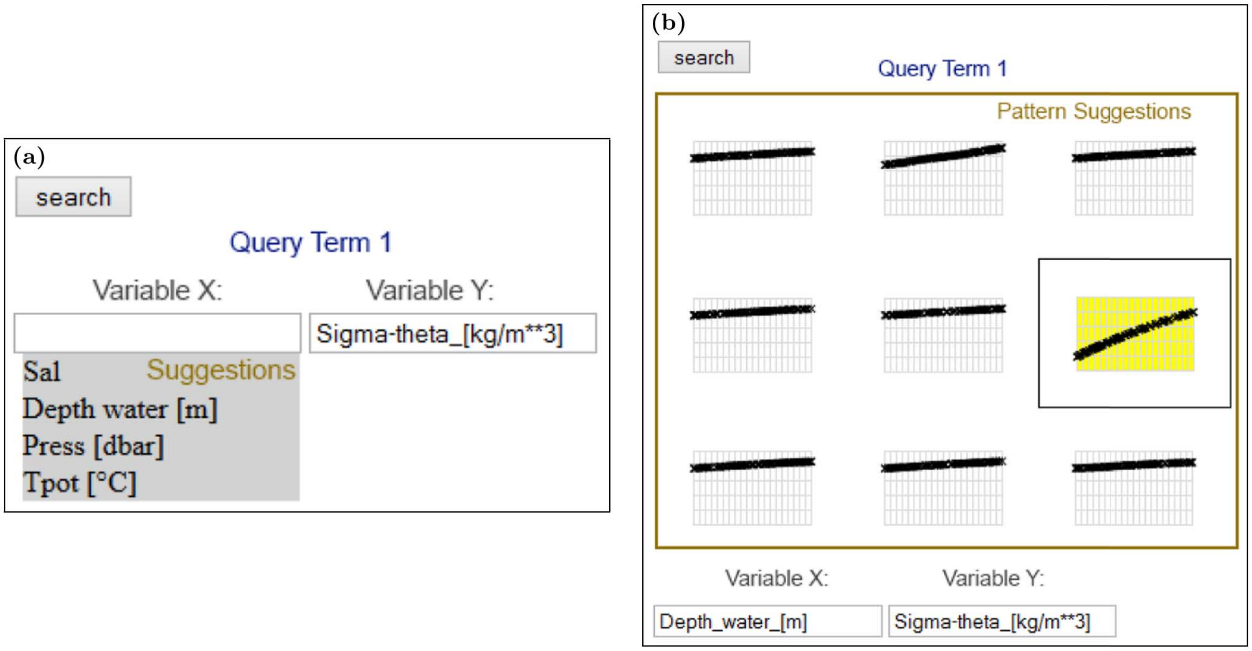
3.3 Instant Results

Due to the full-text-like indexing of our bag-of-words approach, we are able to perform search queries in less than 300 milliseconds, which is generally accepted as "instantaneous" in retrieval applications (see 4.1 for our test-setup). Thus, while the user is still formulating her query, we provide her with immediate results as this has been shown to speed up the retrieval process.

As long as the full-text-index and the primary key index of the database fully reside in the system's main memory, the look-up part of the query time is independent of the number of documents and is dominated by the time required to read the result data from the hard disk.

3.4 Result Highlighting

Highlighting of search results is very important to explain to the user, why a particular document is being returned.



(a) Variable X:
Variable Y: Sigma-theta_[kg/m**3]

(b) Variable X: Depth_water_[m]
Variable Y: Sigma-theta_[kg/m**3]

Figure 6: Search Suggestions on our front-end: (a) After specifying one axis label for our search term, the system presents search terms for the other axis. In this example we specified water density as the y axis. The system suggests to search for salinity, water depth, pressure or temperature – precisely those four variables that water density is functionally dependent of [29]. (b) After selecting one of the suggested x-axis labels, the system suggests search terms for the curve progression by visualizing small scatter-plots to the user.

For text retrieval, highlighting the search terms and showing a few surrounding sentences is a very suitable way to do so. We adapt this to our retrieval scenario. For each retrieved multivariate document, we show up to five plots of its scatter-plot-matrix. These scatter-plots visualize those bivariate patterns in each document that matched the user query. We further highlight those scatter-plots by coloring the axis labels and / or the data-points, depending on their match with the search term. Figure 5 shows an example query, where scatter-plots of each returned document highlight the user's query matches.

3.5 Search Suggestions

We provide users with search suggestions as they provide a major increase in useability for most retrieval systems. Search suggestions and auto-completions gained particular popularity due to their introduction into Google's and Amazon's search front-ends. Since then, search suggestions have also become central to the user's expectations (or mental model) how a search engine works [15]. As such, failure to provide the user with this functionality often leads to queries with no results due to the search for non-existent patters. In other cases, users do not have a precise pattern in mind to search for (or to continue / refine their search with). In these cases, search suggestions provide the user with a much needed overview of patterns she can search for next.

Our approach for search suggestions works as follows:

- retrieve d documents that contain all query terms the user searched for so far
- sum up the ranking scores for all tri-gram terms in the result set
- restrict possible tri-grams according to a partial search term (e.g., a partial axis label) the user supplied
- return the h tri-grams with the highest score as search suggestions and visualize them accordingly

The visualization shows a pop-up of potential x- and y-axis labels, as well as up to nine scatter-plots as small preview images the user can select from (see Figure 6).

The parameter d, the number of documents retrieved to select search suggestions from, influences the quality of the suggestions (higher d is better) but also the computational cost of the suggestions (lower d is better). We found $d = 50$ to be a good compromise between run-time and search suggestion accuracy.

4. APPLICATION

4.1 Data Source and Setup

We show the applicability and scalability our proposed approach to real-world data from *PANGAEA* Data Library [6, 21]. *PANGAEA* is a digital library for environmental sciences and it archives, publishes, and distributes geo-referenced primary research data from scientists all over the world. It is operated by the Alfred-Wegener-Institute for Polar and Marine Research in Bremerhaven, and the Center for Marine Environmental Sciences in Bremen, Germany.

For this use-case, we considered every document that is currently available under the Creative Commons Attribution

License 3.0 and downloadable from http://www.pangaea.de. In total, we were able to obtain and index 98,416 such documents. The raw uncompressed data of these documents requires approximately 35 GB of disk space. Using our approach, we computed and indexed approximately 2.5 million terms. This requires about 2 GB of RAM to keep the index fully within the main memory of our test setup.

Each document is uniquely identified with a DOI (digital object identifier) and consists of a table of multivariate measurements, that include radiation levels, temperature progressions and ozone values, among many more. Each document available at *PANGAEA* is carefully annotated by the scientist who conducted the measurements. A data curator controls the quality of this annotation process. These meta-data annotations include standardized names along with base units for each measurement variable in the data table, which we use for tokenization as described in Section 3.

4.2 Case Study

We use this wealth of environmental data indexed with our approach for a case study. Assuming no prior knowledge of the contents of this data repository, we look for data sets that show similar measurement patterns as part of an explorative search process. It can be the basis to hypothesize about the reason for the observed similarities. First, we enter two intuitive variables, namely water temperature and water depth. Since we do not know what kind of pattern a measurement between these two variables should look like, we let the system provide us with an overview of important patterns (see upper left part of Figure 1). We initially assumed temperature to either drop or rise with water depth (depending on whether the environment is warm or cold). However, we were surprised the system suggested a pattern that indicates a temperature drop up until a certain water depth, and then an increase (see scatter-plot "'Suggestions"' in upper left of Figure 1). Thus, we searched for this pattern. Moreover, we wanted to highlight water density versus water depth in the result set, as we assumed this to be similar to each other as well. The result of this query can be seen in Figure 1.

The result set that was retrieved did exhibit the pattern we queried for and was highlighted accordingly. On top of that, the relationship of density versus depth was also similar. Looking at the locations where those measurements were taken, we were surprised to see such different locations as the Norwegian sea and the antarctic southern ocean. However, looking into some details about the Norwegian sea did reveal that it is a maritime subarctic climatic zone, explaining the similarity of the temperature patterns.

Further refinement of our search by selecting a specific pattern for the relationship between water density and water depth as well, did reduce the result set to a more homogeneous region as we expected at that point (see Figure 7). Using that query, all retrieved documents were measured in the antarctic southern ocean, approximately at the longitude of New Zealand (169°).

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach for content-based indexing of multivariate data by using a bag-of-words approach. On this basis, we developed visual-interactive query modalities that were not available for this kind of document before. In particular, our approach provides the user with result highlighting and search suggestions. We showed the applicability and scalability of our approach by indexing the complete collection of multivariate research data that is publicly available from a data library for the environmental sciences. We provided an exemplary use-case on this repository by retrieving data documents measured in similar maritime climate zones, which would not have been possible using meta-data alone.

For future work, we plan to improve the similarity functions. One extension is to consider hierarchical relationships, which may exist between the observation variables. These could be included to provide approximate matches for variable queries in cases where queried variables do not match, but other related variables exist: An example, that has also been discussed in the application, is the variable of water pressure, which may be closely related to water depth. Another specific task for future work is to explore the parameter space for the algorithms used in this approach, such as the number of clusters for the k-means algorithm.

More generally, future work will include defining a suitable similarity measure between multivariate documents by using the bag-of-words index. Such a similarity measure will allow for query-by-example of complete multivariate data documents, as well as new interactive tools for exploratory search by layouting multivariate documents based on similar content. Finally, evaluation of the effectiveness of the expansion- and highlighting-based search proposed from a user perspective would be interesting. We note that ground-truth benchmarking is not directly applicable, as our approach supports explorative search. Ultimately, we need to measure the degree of insight or information increase, that is brought about by the approach, in a given domain problem.

Acknowledgments

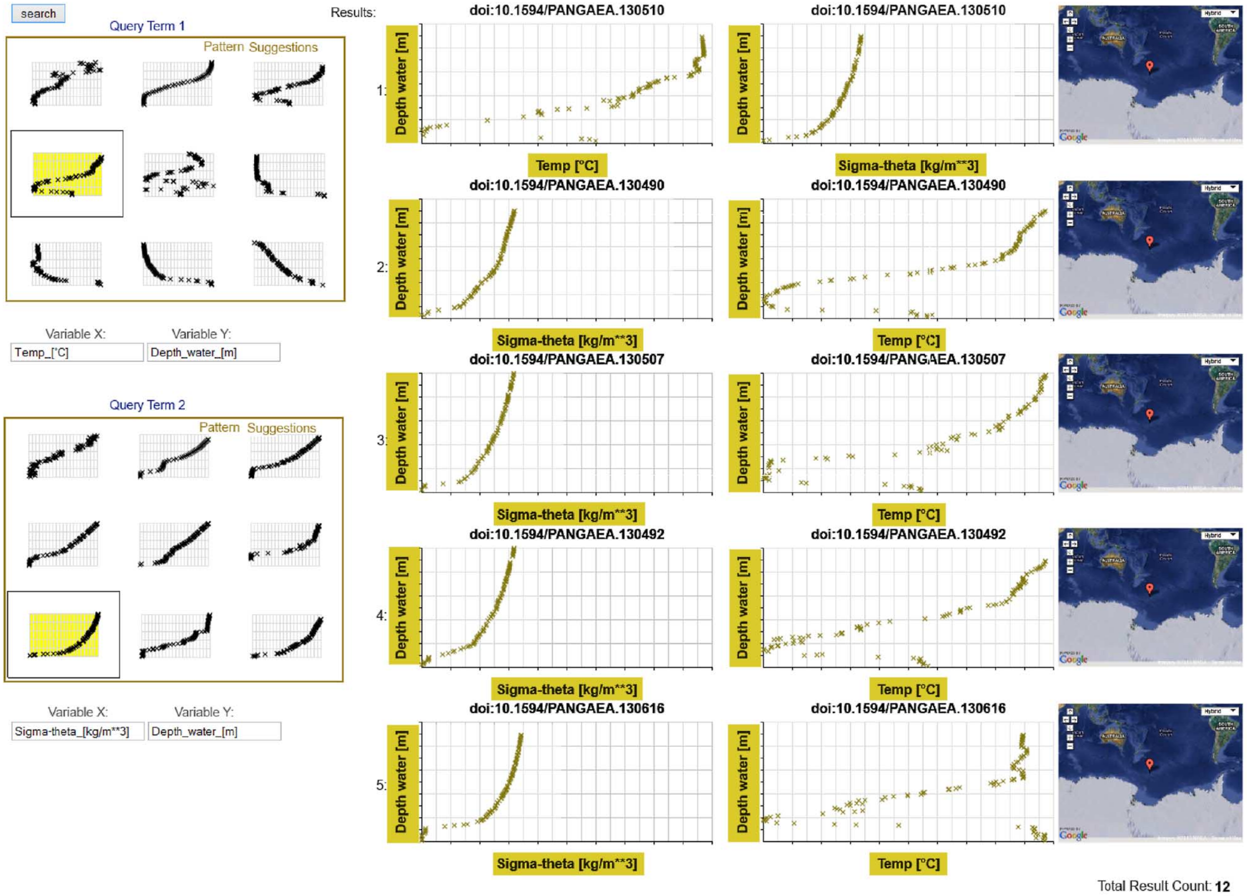
We thank the Alfred-Wegener-Institute (AWI) in Bremerhaven, particularly Rainer Sieger, Hannes Grobe and Gert König-Langlo, and everyone involved with PANGAEA for supporting this research effort. We are especially grateful to the many scientists that contributed the data available through BSRN and other research projects.

6. REFERENCES

[1] J. Bernard, J. Brase, D. W. Fellner, O. Koepler, J. Kohlhammer, T. Ruppert, T. Schreck, and I. Sens. A visual digital library approach for time-oriented scientific primary data. *Int. J. on Digital Libraries*, 11(2):111–123, 2010.

[2] J. Bernard, T. Ruppert, M. Scherer, J. Kohlhammer, and T. Schreck. Content-based layouts for exploratory metadata search in scientific research data. In K. B. Boughida, B. Howard, M. L. Nelson, H. V. de Sompel, and I. Sølvberg, editors, *JCDL*, pages 139–148. ACM, 2012.

[3] R. Berndt, I. Blümel, M. Clausen, D. Damm, J. Diet, D. W. Fellner, C. Fremerey, R. Klein, F. Krahl, M. Scherer, T. Schreck, I. Sens, V. Thomas, and R. Wessel. The probado project - approach and lessons learned in building a digital library system for heterogeneous non-textual documents. In *ECDL*, pages 376–383, Sept. 2010.



Variable X: Temp_[°C]

Variable Y: Depth_water_[m]

Variable X: Sigma-theta_[kg/m**3]

Variable Y: Depth_water_[m]

Total Result Count: 12

Figure 7: Case Study: We refined our original search query (see Figure 1) by also selecting a specific relationship between water density and water depth that was suggested by system. We see that this combination of patterns only occurs in documents that originate from the antarctic southern ocean at longitude 169°. Map Data is attributed to Google Maps.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *In Proceedings ACM International Workshop on Multimedia Information Retrieval*, pages 253–262. ACM Press, 2005.

[6] M. Diepenbroek, H. Grobe, M. Reinke, U. Schindler, R. Schlitzer, R. Sieger, and G. Wefer. Pangaea–an information system for environmental sciences. *Computers & Geosciences*, 28(10):1201–1210, 2002.

[7] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[8] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa. Sketch-based shape retrieval. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):31:1–31:10, 2012.

[9] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 1 edition, 2009.

[10] G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomput.*, 73(7-9):1125–1141, 2010.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[12] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[13] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[14] E. Keogh, J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *IEEE International Conference on Data Mining*, pages 226–233, 2005.

[15] M. Khoo and C. Hall. What would 'google' do? users' mental models of a digital library search engine. In P. Zaphiris, G. Buchanan, E. Rasmussen, and

F. Loizides, editors, *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2012.

[16] B. Köhncke, S. Tönnies, and W.-T. Balke. Catching the drift – indexing implicit knowledge in chemical digital libraries. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 383–395. Springer, 2012.

[17] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

[18] J. Lin, R. Khade, and Y. Li. Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, pages 1–29, 2011.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[20] C. A. Lynch. Jim gray's fourth paradigm and the construction of the scientific record. In T. Hey, S. Tansley, and K. M. Tolle, editors, *The Fourth Paradigm*, pages 177–183. Microsoft Research, 2009.

[21] PANGAEA Publishing Network for Geoscientific & Environmental Data. http://www.pangaea.de/.

[22] M. Riley, E. Heinen, and J. Ghosh. A text retrieval approach to content-based audio retrieval. In *Int. Symp. on Music Information Retrieval (ISMIR)*, pages 295–300, 2008.

[23] R. Rowley-Brooke, F. Pitié, and A. Kokaram. A ground truth bleed-through document image database. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Theory and Practice of Digital Libraries*, volume 7489 of *Lecture Notes in Computer Science*, pages 185–196. Springer, 2012.

[24] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[25] M. Scherer, J. Bernard, and T. Schreck. Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 363–372, New York, NY, USA, 2011. ACM.

[26] M. Scherer, T. von Landesberger, and T. Schreck. A benchmark for content-based retrieval in bivariate data collections. In *Proceedings of the Second international conference on Theory and Practice of Digital Libraries*, TPDL'12, pages 286–297, Berlin, Heidelberg, 2012. Springer-Verlag.

[27] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, number UMCP-CSD CS-TR-3665, pages 336–343, College Park, Maryland 20742, U.S.A., 1996.

[28] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[29] R. Stewart. *Introduction to physical oceanography*. Texas A & M University, 2004.

[30] R. W. White and R. A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[31] B. Wong, S. Choudhury, C. Rooney, R. Chen, and K. Xu. Invisque: technology and methodologies for interactive information visualization and analytics in large library collections. *Research and Advanced Technology for Digital Libraries*, pages 227–235, 2011.

[32] R. Xu, D. Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[33] J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.