International Conference on Theory and Practice of Digital Libraries : Second International Conference ; Paphos, Cyprus, September 23-27, 2012 / Panayiotis Zaphiris... (eds.). - Berlin [u.a.] : Springer, 2012. - S. 286-297. - ISBN 978-3-642-33289-0

A Benchmark for Content-Based Retrieval in Bivariate Data Collections

Maximilian Scherer[1], Tatiana von Landesberger[1], and Tobias Schreck[2]

[1] TU Darmstadt, 64283 Darmstadt, Germany
{maximilian.scherer,tatiana.von-landesberger}@gris.tu-darmstadt.de
[2] University of Konstanz, 78457 Konstanz, Germany
tobias.schreck@uni-konstanz.de

Abstract. Huge amounts of various research data are produced and made publicly available in digital libraries. An important category is bivariate data (measurements of one variable versus the other). Examples of bivariate data include observations of temperature and ozone levels (e.g., in environmental observation), domestic production and unemployment (e.g., in economics), or education and income level levels (in the social sciences). For accessing these data, content-based retrieval is an important query modality. It allows researchers to search for specific relationships among data variables (e.g., quadratic dependence of temperature on altitude). However, such retrieval is to date a challenge, as it is not clear which similarity measures to apply. Various approaches have been proposed, yet no benchmarks to compare their retrieval effectiveness have been defined.

In this paper, we construct a benchmark for retrieval of bivariate data. It is based on a large collection of bivariate research data. To define similarity classes, we use category information that was annotated by domain experts. The resulting similarity classes are used to compare several recently proposed content-based retrieval approaches for bivariate data, by means of precision and recall. This study is the first to present an encompassing benchmark data set and compare the performance of respective techniques. We also identify potential research directions based on the results obtained for bivariate data. The benchmark and implementations of similarity functions are made available, to foster research in this emerging area of content-based retrieval.

Keywords: bivariate data, benchmarking, content-based retrieval, feature extraction.

1 Introduction

Scientific disciplines that range from economics and sociology, to medical science, biology, physics, and others heavily rely on empirical research data, that are produced or collected in large amounts on a regular basis. Due to increased efforts in the digital library community, such research data are recently made available in public data repositories. This is important, as effective user access to research

data repositories will eventually lead to a large increase in research productivity and efficiency [14]. In existing data repositories, user access methods are typically based on textual annotations. These are provided by experts who collected the data in the first place, or by data curators. Annotation-based retrieval is however limited to the availability and scope of textual annotations, which often are expensive and ambiguous to obtain. Moreover, annotations do not allow for retrieval of specific data based on its content (e.g., a specific relationship between two variables). Following the general ideas of content-based retrieval in multimedia data [24], first content-based retrieval methods are currently being developed and applied in research data repositories [25,3].

Four large and important categories of research data are univariate (time-series) data, biviarate data, multivariate data and multimedia data (e.g., 2D/3D image data from satellites, microscopes, MRT, etc...). Figure 1 shows an example for each of the first three categories. For collections of time-series data, the task of content-based retrieval has received a significant amount of research attention by the community in the last two decades. Subsequently, several efficient methods for indexing such databases and retrieving the data were established since [9]. For retrieval in collections of bivariate and multivariate data, research was carried out in a rather limited scope. In particular, no exhaustive evaluation of feature extraction techniques that support retrieval in bivariate data collections has been conducted so far. This can be mainly accounted to the fact, that no benchmark to support a quantitative comparison of different techniques has been proposed by now.

We attribute absence of respective data retrieval benchmarks to the difficulty of defining *similarity* for univariate, bivariate or multivariate data. For data like text, images, audio, 3D models or video, the notion of similarity usually follows a straight-forward concept. For example, similar text documents can be about the same subject or similar images show similar scenery or objects. For retrieval in sequential, bivariate or multivariate research data, such similarity concepts are not considered so far to judge relevance of retrieved data objects to a query. A meaningful quantitative evaluation for retrieval in research data collections requires meaningful annotations assigned by humans to the data objects, to allow

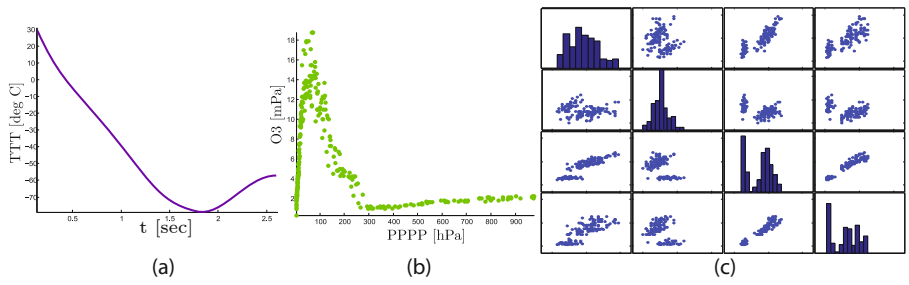


Fig. 1. Examples for time-series data (a), bivariate data (b) and multivariate data (c) (here the four-dimensional Fisher-Iris dataset, visualized as a scatterplot-matrix)

construction of similarity classes or relevance judgments [16]. So far, gathering such annotations was very expensive, as expert users are required to annotate large amounts of rather abstract objects. Due to recent efforts in the digital library community however, research data with accurate annotations by experts became publicly available on a large scale, e.g. in the PANGAEA project [8].

Thus, we use the wealth of publicly available, manually annotated research data to construct a benchmark for retrieval in bivariate data collections. This benchmark is composed of real-world scientific measurements in the domain of earth observation science, that were annotated by domain experts. Based on these metadata annotations – particularly type, location and time of measurement – we automatically group data objects to similarity classes (see Section 3). For example, all measurement data of type *altitude [m] vs. pressure [hPa]* measured at longitude 20.5 and latitude -30.7 in December can be expected to be similar [28] and are therefore assigned to the same similarity class. This results in a labeled collection of research-data, where data objects with similar content have the same label. We extensively evaluate performance of different feature extraction techniques for retrieval via precision and recall on this collection (see Section 4). Since this is the first benchmark of its kind, we also analyze and discuss the composition of similarity classes in the benchmark itself (see Section 5).

2 Related Work

For information retrieval and data mining tasks like regression, clustering, classification and retrieval benchmarking plays an important role. Ideally, such a benchmark consists of many test data sets, covering all aspects of a certain challenge, along with ground truth or a *gold standard* for each of these data sets.

For classical problems like clustering and classification, several established datasets have emerged, serving as benchmarks to compare effectiveness. For example, several datasets in the UCI Machine Learning Repository [12] are widely used to compare results of algorithms for classification and clustering. However, so far no datasets to compare techniques for retrieval are available there.

Several large-scale retrieval challenges exist for text and multimedia retrieval (TREC[1], CLEF [2] and NTCIR [3]), although a track for research-data has not yet been established.

For multimedia information retrieval, many manually annotated datasets exist and are used for benchmarking [19]. The MPEG-7 benchmark is used for 2D shape analysis [18]; the Princeton Shape Benchmark, among others, is used for 3D object retrieval [26]; and several large benchmarks for content-based image retrieval exist [6,7]. For these benchmarks, objects are usually assigned to similarity classes (either manually by humans, or automatically by using *social*

[1] http://trec.nist.gov
[2] http://www.clef-initiative.eu
[3] http://research.nii.ac.jp/ntcir/

tags, e.g., from Flickr), and precision-recall can be computed, to measure effectiveness of feature extraction algorithms for similarity assessment. However, their suitability is sometimes discussed [20], since such automatically designed benchmarks lack specified query sets and relevance judgments of retrieval results.

In 2003, Keogh and Kasetty [16] discussed the need for benchmarking retrieval in time-series data. They empirically show that given a sufficiently large number of datasets to choose from, the superiority of any technique can be shown when only considering numeric similarity of retrieval results. Thus, they argue for the need of similarity concepts to construct a meaningful benchmark. Only recent advances in the digital library community however, led to publicly available, manually annotated research data on a large scale [10,23,11], which is required for such a benchmark construction. Nevertheless, for retrieval in research data, and in particular for retrieval in bivariate data, no such benchmark has been proposed to date.

3 Benchmark Construction for Bivariate Data Retrieval

In this section, we present our approach to construct a benchmark for retrieval in bivariate data collections. Similar to benchmarking in multimedia retrieval, we assign data objects to similarity classes, based on metadata annotations by experts. In our case these data objects are bivariate measurement data in the area of earth observation. The annotations of the measurement data are done by scientists, that describe the type of measurement and the experimental conditions under which it was conducted. So far, such annotations were expensive to obtain, preventing construction of a benchmark large enough. Due to recent efforts in the digital library community however, repositories offering expert-annotated research data became available. We describe how to use this new data source to define similarity classes and subsequently construct a benchmark.

3.1 Data Source

We use earth observation data, which is publicly available from the `PANGAEA` Data Library [8,21]. `PANGAEA` archives, publishes, and distributes geo-referenced primary research data in the domain of earth observation (water, sediment, ice, atmosphere) 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. Most of the data available can be publicly accessed via `http://www.pangaea.de` and can be downloaded under the Creative Commons Attribution License 3.0. For content-based bivariate data retrieval, a subset of these measurement files was recently used as an application example for bivariate data retrieval [25]. Each file consists of a table of multivariate measurements, that include radiation levels, temperature progressions and ozone values, among many more. Each file available through `PANGAEA` is carefully annotated by the scientist who conducted the measurements. Quality control over this annotation process is taken care of by the

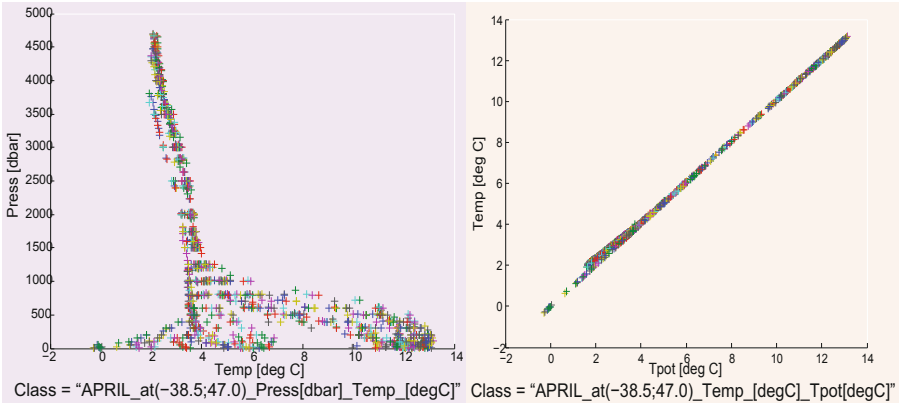


Fig. 2. Exemplary data objects of the two largest similarity classes. For both classes (left and right), data points of all 58 objects are plotted into a single display in separate colors. The unique class labels consist of measurement type, location and time.

PANGAEA data curator. Most importantly for our purposes, these annotations include the type of measurement (standardized names along with base units (SI)[4] for each measurement variable in the data table), as well as the experimental conditions (time and location) under which the measurement was conducted.

To construct a test data set, we downloaded 490 publicly available measurement files from PANGAEA. Each file contains multivariate measurements with 10 to 100 columns each. By extracting every pair-wise variable combination from each of these measurement files, we obtained 24,700 bivariate data objects which form the test data set of our benchmark.

3.2 Definition of Similarity Classes

Based on the expert-annotations, we define similarity classes for the bivariate data objects. In particular, we assume data measuring the same relationship (e.g., *Temperature [deg C] vs Pressure [dbar]*) at the same time of year (e.g., December) at a close-by location (e.g., *longitude* $\approx$ 24, *latitude* $\approx$ 12) to be similar. To compute a unique class identifier, the pair of annotated variable names – already a categorical label – was used directly. The month part of the timestamp was extracted, and the geocode of the location was categorized in a 6x12 grid. By combining these three categorical labels, we assigned data objects to 1,608 different similarity classes. Such a spatio-temporal quantization is biased, as neghboring data points may be assigned to different similarity classes if they are close to the decision border. We discuss such implications on intra- and inter-class variability in Section 5.

This assumption for similarity class definition is based on Tobler's first law of geography: "'Everything is related to everything else, but near things are more

[4] As defined by the International System of Units.

Table 1. Statistics of proposed benchmark. Data point correlation is computed as Pearson's correlation coefficient for each bivariate data object and averaged over all objects of each class.

	Sum	**Mean**	**Std**	**Median**	**Min**	**Max**
objects: total	24,700	-	-	-	-	-
classes: total	1,608	-	-	-	-	-
objects: per class	-	15.36	10.9	11	5	58
data points: per class	-	657.98	1,043	319	51	9,770
data points correlation: avg per class	-	0.66	0.28	0.73	0.001	1.00

related than distant things"' [28]. The decision of the discretization parameters (temporal and spatial resolution) to construct the similarity classes influences similarity of data within a class and similarity of data among classes. We discuss these two benchmark statistics in detail in Section 5.

Table 1 gives a detailed overview of the most important benchmark statistics. Particularly interesting is the high avg data-point correlation per class, which indicates that objects exhibit some (linear) relationship, which can in principle be captured by feature extraction. Figure 2 shows data objects from the two largest similarity classes for illustration. We see that all objects within those two classes are numerically similar.

4 Evaluation

In this section, we evaluate the following nine feature extraction techniques on the benchmark described in the previous section. To measure performance of each technique for bivariate data retrieval, we compute precision-recall on the benchmark.

4.1 Feature Extraction for Retrieval in Bivariate Data

Feature extraction is the process of computing a descriptor, that mathematically represents one or several properties of an object under consideration. Such a descriptor allows to assess pairwise similarity between objects, by computing a distance measure between their respective descriptors. A prominent type of descriptors are feature vectors. As the name implies, we try to capture descriptive and discriminative object features as a vector of numerical values.

The following list provides an overview of the nine techniques, which we adapted to feature extraction of bivariate data and that we propose for evaluation. The techniques are based on time-series analysis, regression and image processing, respectively.

Euclidean Distance (SM). Baseline technique that resamples data by fitting a smoothing spline to the data [5] to allow for measuring Euclidean distance between data objects.

Correlation Coefficients (CORR). Another baseline technique that is composed of Person's sample correlation coefficient r, Kendall's tau rank correlation coefficient τ and Spearman's rank correlation coefficient ρ, to capture how strong a linear relationship exists in a bivariate data object.

Regressional Features (RF). Recently proposed [25] for indexing bivariate data. Based on the goodness-of-fit of data to several, predefined functional models.

Smoothing Splines (Spline). A feature extraction technique based on non-parametric fitting of a smoothing spline to the data [27,13]. The spline's coefficients in pp-form describe the data.

Discrete Fourier Transform (DFT).[5] A technique from signal processing to describe data by its first $k = 0, \ldots, M$ Fourier coefficients [1].

Piece-wise Aggregate Approximation (PAA)[6]**.** Describes sequential data by splitting a sequence of length n into m segments and compute the mean value of all data-points in each segment [29,15].

Symbolic Aggregate Approximation (SAX)[6]**.** A descriptor for time-series based on symbolic representation [17]. The time domain *and* the value domain of time-series data are discretized.

Kernel Density Estimation (KDE). Estimates the kernel function of the probability density of two-dimensional data as a Gaussian kernel [4]. This probability density function is used as a descriptor for the data.

Edge Histogram Descriptor (EHD). Prominent approach in image processing to describe the shapes seen in an image, by computing the distribution of the orientation of *edges* in that image [22]. Also included in the MPEG-7 standard.

4.2 Retrieval Results

For our quantitative evaluation of effectiveness, we compare retrieval performance for each of the nine considered feature extraction algorithms. In particular, we use a query-by-example, leave-one-out evaluation. This means that we use each object as a query and compute precision and recall for the ranking of all other objects in the data set. We compute r-precision (also known as first-tier precision or precision at r, see [2, section 3.2]), which is suitable since our similarity classes have a significantly different number of objects. To compute r-precision, we retrieve $k - 1$ objects from the data set for a given query, where k is the number of objects in the query's similarity class. Then the percentage of relevant objects within these $k - 1$ retrieved objects is the r-precision.

Figure 3 shows the boxplot of the r-precision for retrieval results obtained with each approach on the entire test data set. We see that average r-precision is between 1% and 7%, and that retrieval for several classes does not work at all (r-precision of zero). The difference in r-precision between the techniques is significant nonetheless, as an algorithm that randomly retrieves data objects

[5] Can be applied to bivariate data by sorting the data along either dimension to get a sequential representation.

Table 2. Overview of the considered feature extraction techniques for bivariate data. Average feature extraction time t_{sec} (lower is better) and average r-precision results μ_r (higher is better) indicate performance on the proposed benchmark. The low r-precision of randomized retrieval shows significance of changes in r-precision between the obtained results.

Descriptor	**Abbr.**	**Dim**	t_{sec}	μ_r (%)
Regressional Features	RF	43	2.43	2.5
Smoothing Splines	SPLINE	996	0.115	2.08
Discrete Fourier Transform	DFT	100	$0.1 \cdot 10^{-3}$	1.1
Piecewise Aggregate Approx.	PAA	100	$0.2 \cdot 10^{-3}$	2.3
Symbolic Aggregate Approx.	SAX	100	0.002	3.08
Kernel Density Estimate	KDE	1024	0.07	5.73
Edge Histogram Descriptor	EHD	80	0.26	**6.56**
Correlation Descriptor	CORR	3	0.028	1.7
L_2 of Resampled Data	L_2	100	0.009	1.35
Random Retrieval	-	-	-	0.059

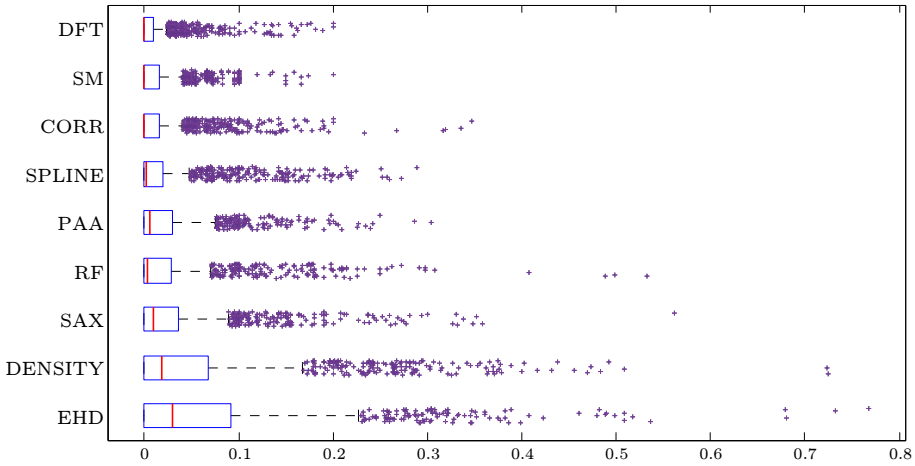
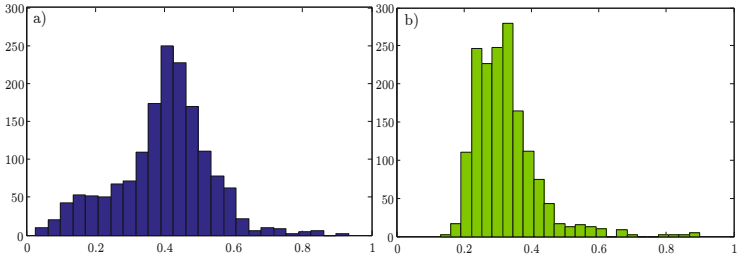


Fig. 3. Boxplot of average r-precision for each studied descriptor for the benchmark data set. The box visualizes the 95% confidence interval around the mean. The red, vertical bar indicates the median and the scattered plus-signs show outliers. Note that the difference in retrieval precision between the techniques is significant, as randomized retrieval only reaches 0.00059 r-precision.

only reaches an average r-precision of 0.059%. Particularly the image-based descriptors `KDE` (density) and `EHD` (shape) perform quite well. The only technique that performs below the two baseline techniques (`CORR` and `SM`) is the discrete Fourier transform based descriptor (`DFT`).



	Mean	Std	Median	Min	Max
a) intra-class divergence: per class	0.39	0.13	0.41	0.03	0.93
b) inter-class divergence: per class	0.32	0.09	0.31	0.13	0.90

Fig. 4. Intra-class (a) and inter-class (b) divergence based on individual data point distribution versus class data point distribution

5 Benchmark Discussion

In this work, we propose a new approach to define similarity among bivariate data objects, by using metadata annotations by experts in research data repositories. After assigning objects to similarity classes, an interesting questions is how numerically dissimilar objects within each class are (intra-class divergence) and how dissimilar objects among classes are (inter-class divergence). Given the motivation for the construction of this benchmark – evaluating feature extraction to retrieve similar objects – measuring these divergences computationally is difficult, as a similarity measure is required.

Looking at the 2D probability density of the bivariate data objects.w e compute intra-class divergence as the Euclidean distance between each data object's individual 2D probability density and the 2D probability density of all objects in the corresponding class. One can visualize this divergence as the difference of the scatter-plot density of a single data object versus the density of the scatter-plot of all bivariate data objects at once (visualized in Figure 2).

To judge inter-class divergence, we select all objects of two random classes and again compute the distance of the 2D probability density. This time, we compute the distance between the distribution of all the data points of these two random classes and the data point distribution of each individual object. By repeating this experiment until convergence for each class, we get an average distance of objects in different similarity classes – thus the inter-class divergence.

The results are presented in Figure 4. We see that on average intra-class divergence is similar to inter-class divergence, which explains the low average r-precision for the evaluated techniques.

We explore the r-precision of individual similarity classes and their respective inter- and intra-class divergence in detail. Figure 5 shows an overview of this relationship. As expected, we see that well performing classes (big, non-blue points) are primarily located in the upper left corner and thus exhibit a low intra-class divergence and a high inter-class divergence. However there are a few

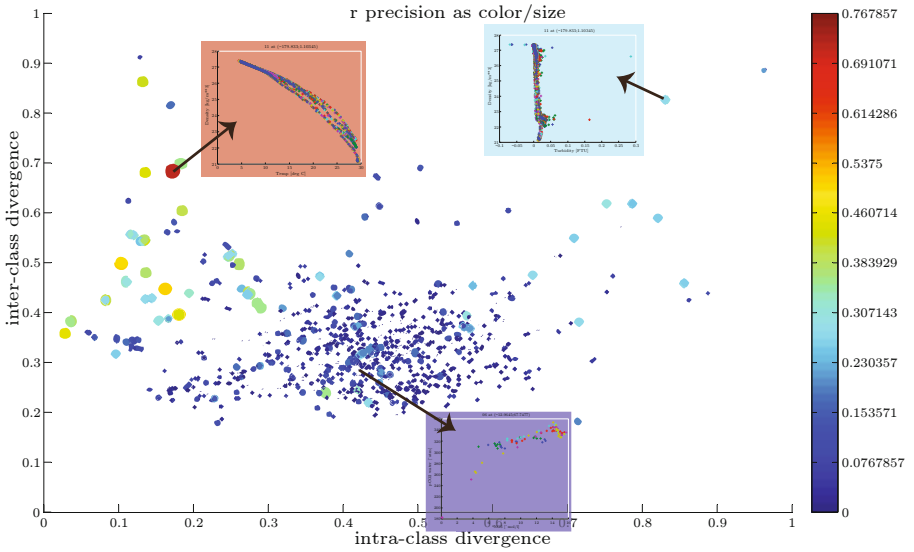


Fig. 5. Intra-class divergence plotted versus inter-class divergence for each similarity class in the benchmark. Color and size of data-points redundantly visualize mean average r-precision for each class. Data-point color in the small sub-images show class-membership.

well performing classes in the upper right corner. These classes exhibit a very high intra-class divergence (which makes retrieval difficult), but at the same a high (but not as high a) inter-class divergence. This indicates that for bivariate retrieval, numerical discrimination from other classes is more important for good retrieval results than numerical similarity within a class. nction, respectively. Retrieved results are shown for each technique detailed in the previous section in a side-by-side comparison.

6 Conclusion and Future Work

In this paper, we constructed a benchmark for retrieval in bivariate data collections, based on metadata annotations by domain experts. To the best of our knowledge, this is the first such benchmark for bivariate data retrieval. We used publicly available earth observation data for this purpose, by defining similarity classes based on available expert annotations. We verified that our definition of similarity classes – type, location and time of measurement – was meaningful, by computationally analyzing inter- and intra-class divergence. We exhaustively evaluated nine different feature extraction techniques for the task of retrieval in bivariate data collections on our benchmark, to give a tenable indication as to their respective retrieval performance. Results show that retrieval performance of all current techniques leaves lots of room for future improvements.

We make the benchmark and all reference implementations available[6] for future research. Our own future work includes the addition of new, more homogeneous datasets to the benchmark. Furthermore we are researching new and refined feature extraction techniques based on the obtained results, to make retrieval techniques available to the digital library community, which yield performance suitable for production-use.

Acknowledgments. We would like to thank the Alfred-Wegener-Institute for Polar and Marine Research in Bremerhaven, and the Center for Marine Environmental Sciences in Bremen for their continued support and collaboration in researching new access modalities to digital data libraries.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient Similarity Search in Sequence Databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730, pp. 69–84. Springer, Heidelberg (1993)
2. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston (1999)
3. Bernard, J., Brase, J., Fellner, D.W., Koepler, O., Kohlhammer, J., Ruppert, T., Schreck, T., Sens, I.: A visual digital library approach for time-oriented scientific primary data. Int. J. on Digital Libraries 11(2), 111–123 (2010)
4. Botev, Z., Grotowski, J., Kroese, D.: Kernel density estimation via diffusion. Annals of Statistics 38(5), 2916–2957 (2010)
5. Cleveland, W.S.: The Elements of Graphing Data. Hobart Press (1985)
6. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) 40(2), 5 (2008)
7. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Information Retrieval 11(2), 77–107 (2008)
8. Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G.: Pangaea–an information system for environmental sciences. Computers & Geosciences 28(10), 1201–1210 (2002)
9. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment 1(2), 1542–1552 (2008)
10. Dryad Digital Repository for Data Underlying Published Works, `http://www.datadryad.org/`
11. ELIXIR European Life Sciences Infrastructure for Biological Information, `http://www.elixir-europe.org/`
12. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
13. Heckman, N., Ramsay, J.: Penalized regression with model-based penalties. Canadian Journal of Statistics 28(2), 241–258 (2000)
14. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington (2009), `http://research.microsoft.com/en-us/collaboration/fourthparadigm/`

[6] `www.gris.informatik.tu-darmstadt.de/%7Emaschere/retrievalBenchmark/`

15. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems 3(3), 263–286 (2001)
16. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and Knowledge Discovery 7(4), 349–371 (2003)
17. Keogh, E., Lin, J., Fu, A.: Hot sax: Efficiently finding the most unusual time series subsequence. In: IEEE International Conference on Data Mining, pp. 226–233 (2005)
18. Latecki, L.J., Lakämper, R., Eckhardt, U.: Shape descriptors for non-rigid shapes with a single closed contour. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 424–429 (2000)
19. Lew, M., Sebe, N., Djeraba, C., Jain, R.: 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)
20. Müller, H., March, S., Pun, T.: The truth about corel - evaluation in image retrieval. In: Proceedings of The Challenge of Image and Video Retrieval (CIVR), pp. 38–49 (2002)
21. PANGAEA Publishing Network for Geoscientific & Environmental Data, `http://www.pangaea.de/`
22. Park, D., Jeon, Y., Won, C.: Efficient use of local edge histogram descriptor. In: Proceedings of the 2000 ACM workshops on Multimedia, pp. 51–54. ACM (2000)
23. PsychData National Repository for Psychological Research Data, `http://psychdata.zpid.de/`
24. Rüger, S.M.: Multimedia Information Retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers (2009)
25. Scherer, M., Bernard, J., Schreck, T.: 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 2011, pp. 363–372. ACM, New York (2011)
26. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Shape Modeling Applications, pp. 167–178. IEEE (2004)
27. Silverman, B.: Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society. Series B (Methodological) 47(1), 1–52 (1985)
28. Tobler, W.: A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240 (1970)
29. Yi, B., Faloutsos, C.: Fast time sequence indexing for arbitrary Lp norms. In: Proceedings of the 26th International Conference on Very Large Data Bases, pp. 385–394 (2000)