Efficient adaptive retrieval and mining in large multimedia databases

Ira Assent
Data Mining and Data Exploration Group
RWTH Aachen University, Germany

Abstract: Multimedia databases are increasingly common in science, business, entertainment and many other applications. Their size and high dimensionality of features are major challenges for efficient and effective retrieval and mining. Effective similarity models are usually computationally far too complex for straightforward usage in large high dimensional multimedia databases.

We propose efficient algorithms for these effective models that show substantially improved scalability. Our index-based methods for efficient query processing and mining restrict the search space to task relevant data. Multistep filter-and-refine approaches using novel filter functions with quality guarantees ensure that fast response times are achieved without any loss of result accuracy.

1 Multimedia Databases

There is tremendous growth in multimedia data in application domains such as medicine, engineering, biology and numerous others. New technologies such as computer tomography or digital photography produce increasingly large volumes of data. At the same time, decreasing storage prices allow archiving large amounts of multimedia data at relatively low cost. For database technology, large multimedia databases require efficient and effective strategies for accessing and processing of multimedia data. Examples of typical multimedia content are illustrated in Figure 1: stock data time series, music, medical magnetic resonance images, shape and color images, as well as gene expression data.

Traditional database management systems have established themselves as reliable and powerful techniques for archiving and retrieving textual or numeric data. Traditional meta data information, however, is typically not adequate for multimedia applications. For example, users typically search for images not with respect to their size on hard disk, but with respect to content related features like color or shape. Consequently, multimedia database systems should provide content-based access. Likewise, aggregating the content of multimedia databases is not straightforward. Traditional reporting tasks query databases for (aggregated) numbers. Aggregating multimedia objects, or, more precisely, their respective feature values, in such a way is not meaningful. Knowledge discovery techniques aggregate multimedia data with respect to type of media and application focus such that expedient summaries are formed.

Multimedia databases thus should provide (1) content-based access, (2) knowledge discov-

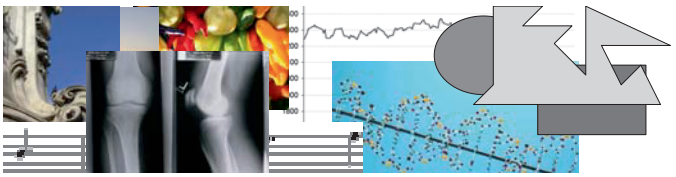

Figure 1: Examples of multimedia data

ery methods, (3) scalability to large data volumes, (4) scalability to high dimensionality of features, (5) good runtime performance.

In this work, we focus on efficient content-based retrieval and mining. Using multistep filter-and-refine architectures, our novel algorithms and lower bounding filters yield substantial runtime improvements without loss of accuracy. The general idea is illustrated in Figure 2. Efficient filters greatly reduce the number of objects relevant to a query to a small set of candidates. These candidates are then refined to the actual result set. Fulfillment of the ICES criteria ensures that the filter is indexable (works on multidimensional index structures), complete (no false dismissals), efficient (fast one-to-one filter comparison) and selective (small candidate sets). Moreover, we devise indexing techniques that greatly reduce database access for similarity search or for data mining.

2 Similarity search and retrieval

Content-based retrieval searches for similar multimedia objects to a query object. Figure 3 gives an example for a small image database. Content-based retrieval requires efficient algorithms on effective similarity models. Similarity models define the relevant characteristics of multimedia objects, for example color distribution in images, or pixel frequencies in shapes. Such histogram features can be compared via distance functions which assign a degree of dissimilarity to any pair of histograms.

The type of distance function is crucial for the effectivity of the similarity search process, as it models which multimedia objects are considered alike. For different types of features, different similarity models have been proposed. One of the simpler models that is commonly used for both histograms and time series is the Euclidean distance, a distance from the family of L_p norms. Based on the differences of corresponding histogram bins or time series values, respectively, the overall dissimilarity between features is computed. Adaptable similarity models allow for incorporation of expert knowledge on application domain.

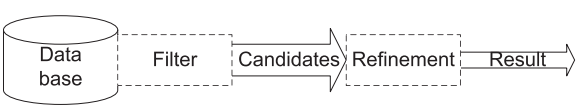

Figure 2: Multistep query processing

Figure 3: Similarity search

The Earth Mover's Distance, introduced in computer vision, computes an overall match on two (histogram) features based on a ground distance in feature space. Likewise, Dynamic Time Warping, originally from speech recognition, allows for stretching and squeezing of time series to match time series features. We propose new efficient retrieval methods for these models.

2.1 Histogram based search under the Earth Mover's Distance

Content-based similarity search requires effective similarity models as well as efficient query processing. Many multimedia features can be represented as histograms, i.e. vectors of attributes. For meaningful comparison of histogram features, distance functions provide dissimilarity values. Comparing image histograms using L_p-norms such as the Euclidean distance is very sensitive to minor shifts in feature value distribution.

A recent approach from computer vision towards human similarity perception, the Earth Mover's Distance (EMD) [RT01] models similarity as the amount of changes necessary to transform one image feature into another one. Formally, the Earth Mover's Distance is defined as a Linear Programming problem which can be solved using the simplex method as in [Dan51, HL90]. Its computation, however, is too complex for usage in interac-

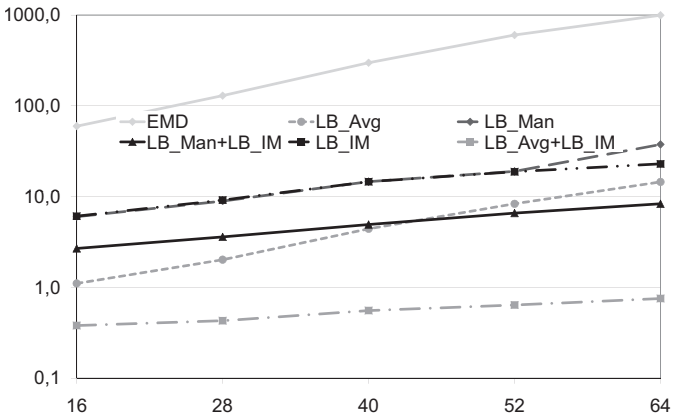

Figure 4: Dimensionality, response time in seconds

tive multimedia database scenarios. In order to enable efficient query processing in large databases, we propose an index-supported multistep algorithm. We therefore develop new lower bounding approximation techniques for the Earth Mover's Distance which satisfy high quality criteria including completeness (no false dismissals), index-suitability and fast computation. High dimensions are problematic for most index structures; we therefore use a dimensionality reduction implicit in the averaging lower bound or explicit in our Manhattan reduction. As this still leaves us with far too many candidates, two-phase filtering that combines the power of the individual filters, ensures an efficient retrieval procedure.

We demonstrate the efficiency of our approach in extensive experiments on large image databases. An important parameter for the performance of lower bounding approximations is the size of the histograms involved. On a database of 200,000 color images, we varied histograms from dimension 16 to 64. Figure 4 shows the corresponding response times. With increasing dimensionality, the computation of the existing LB_{Avg} [RT01] increases in complexity. The response times of our LB_{Man} are more closely related to its high selectivity ratios. The overhead of our novel, more complex, lower bound LB_{IM} is greater than that of the other two lower bounds. Its combination with a low-dimensional LB_{Avg}-index yields the best performance improvements. We include the sequential scan Earth Mover's Distance computation as a baseline comparison. Note that the improvement for 64 dimensions comparing Earth Mover's Distance and the best multistep concept is from 1000 seconds to less than one second, i.e. more than three orders of magnitude. These efficiency gains do not come at the expense of accuracy. We prove completeness of our approach (no false dismissals), and refine using the exact Earth Mover's Distance (no false alarms) [AWS06]. For further approaches to efficient similarity search under the Earth Mover's Distance see also [AWMS08, AKS07, WAKS08].

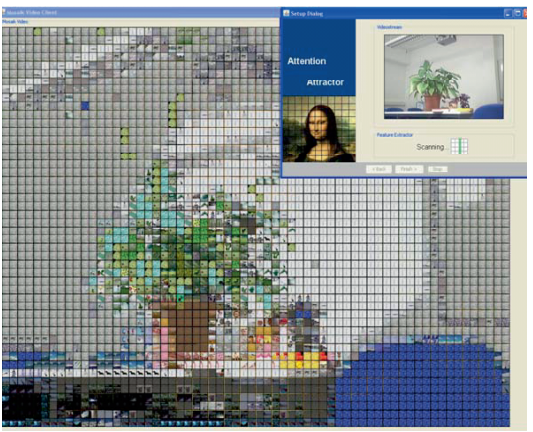

Figure 5: Similarity search video still

As a case study that demonstrates both the efficiency and effectiveness of this setup, we built a video streaming demonstration. A video camera records people passing by, and a monitor shows an alienated version, "mosaic" in real time. This system has been suc-

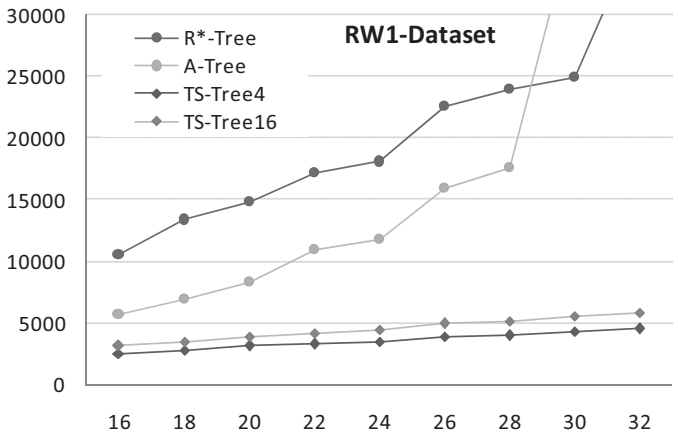

Figure 6: Dimensionality, number of page accesses for RW1 time series

cessfully used on several occasions to demonstrate state-of-the-art similarity search in a practical setting (see Figure 5) [AKS07].

2.2 Time Series Search and Retrieval

Continuous growth in sensor data and other temporal data increases the importance of retrieval and similarity search in time series data. Efficient time series query processing is crucial for interactive applications. Existing multidimensional indexes like the R-tree provide efficient querying for only relatively few dimensions [Keo02]. Time series are typically long which corresponds to extremely high dimensional data in multidimensional indexes. Due to massive overlap of index descriptors, multidimensional indexes degenerate for high dimensions and have to access the entire data by random I/O. Consequently, the efficiency benefits of indexing are lost.

We propose the TS-tree (time series tree), an index structure for efficient time series retrieval and similarity search. Exploiting inherent properties of time series quantization and dimensionality reduction, the TS-tree indexes high-dimensional data in an overlap-free manner. During query processing, powerful pruning via quantized separator and meta data information greatly reduces the number of pages which have to be accessed, resulting in substantial speed-up.

In thorough experiments on synthetic and real world time series data we demonstrate that our TS-tree outperforms existing approaches like the R*-tree or the quantized A-tree. As an example, we investigate the scalability and the query performance of the TS-tree on synthetic and real world data sets using the L_2 norm, i.e. Euclidean distance. Please not that our approach also works very well for Dynamic Time Warping (DTW) [BC94]. To analyze scalability, we report the number of pages read averaged over 25 nearest neighbor queries. The more the dimensionality of a time series data set is reduced the more time

series typically have to be refined. As they are usually very long, scalability in terms of dimensionality is very important. Figures 6 illustrates the results for RW1 data (synthetic random walk data of size 250,000), for time series length 256, reduced to dimensionality 16 to 32 using piecewise aggregate approximation. The TS-tree scales very well since its overlap-free separator split is not impaired by effects of high dimensional data spaces ("curse of dimensionality"). The R*-tree and A-tree, however, degrade with increasing dimensionality. Overall, the TS-tree outperforms both A-tree and R*-tree by nearly one order of magnitude [AKAS08].

3 Knowledge discovery in databases

Knowledge discovery in databases (KDD) extracts novel, potentially useful patterns from databases for generation of knowledge about the database content as a whole. The KDD process, depicted in Figure 7, consists of four steps from the original raw data to the actual knowledge generation [HK01]. In the first step, the data is integrated and cleaned of potential errors, then it is transformed and projected to the task relevant parts of the data. The central step, the data mining step, extracts patterns from this task relevant data, which is then visualized for human evaluation. In this work, we focus on automatic aggregation of the data into meaningful groups, the so-called clustering. Clustering groups objects such that similar ones are within the same group, whereas dissimilar ones are in different groups. We study efficient and effective clustering for high dimensional large databases.

3.1 Subspace clustering

In high-dimensional data, clusters are typically concealed by irrelevant attributes and do not show across the full space. Subspace clustering mines clusters hidden in subspaces of high-dimensional data sets. Density-based approaches have been shown to successfully mine clusters of arbitrary shape even in the presence of noise in full space clustering [EKSX96]. Direct extensions of density-based approaches to subspace clustering, how-

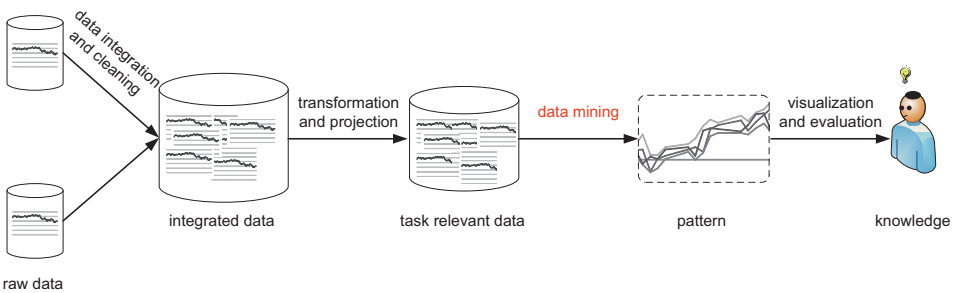

Figure 7: The knowledge discovery process

	Objects	Dimensions	Classes	Source
Pendigits	7494	16	10	[AN07]
Vowel	990	10	11	[AN07]
Glass	214	9	6	[AN07]
Shapes	160	17	9	[KWX+06]

Table 1: Real world data sets

ever, entail runtimes which are infeasible for large high-dimensional data sets, since the number of possible subspace projections is exponential in the dimensionality of the data. Consequently, existing subspace clustering algorithms trade-off efficiency for accuracy.

We propose lossless efficient detection of subspace clusters via a new density-conserving grid data structure which benefits from efficiency gains without losing any clusters. By detecting clusters in a depth-first manner, our EDSC (efficient density-based subspace clustering) algorithm avoids excessive candidate generation.

In thorough experiments on synthetic and real world data sets, we demonstrate that our lossless approach yields accurate clusters and is faster than existing subspace clustering algorithms by orders of magnitude. Real world data used in our experiments is characterized in Table 1. We evaluate the quality of EDSC by determining how accurate the hidden structure of the data given by the classes is identified. For each data set efficiency and accuracy of each algorithm is presented in Table 2. We can see that the EDSC algorithm shows the best quality for all real world data sets. Compared with SUBCLU [KKK04] the EDSC algorithm shows significantly better runtimes. The grid-based SCHISM [SZ04] shows not only worse quality in all data sets but in some also higher runtimes. In the shape data set SCHISM only finds 2-dimensional clusters and thus cannot be compared with the other subspace clustering algorithms which also detect the higher dimensional clusters. Our experiments on large and high-dimensional synthetic and real world data sets show that EDSC outperforms recent subspace clustering algorithms by orders of magnitude while maintaining superior quality [AKMS08].

	EDSC		SCHISM		SUBCLU	
	F1	time [s]	F1	time [s]	F1	time [s]
Pendigits	48%	5743	25%	5278	24%	43825
Vowel	35%	47	21%	88	17%	310
Glass	57%	3	47%	8	52%	10
Shape	51%	34	2%	0.2	40%	1079

Table 2: Quality and runtimes on real world data

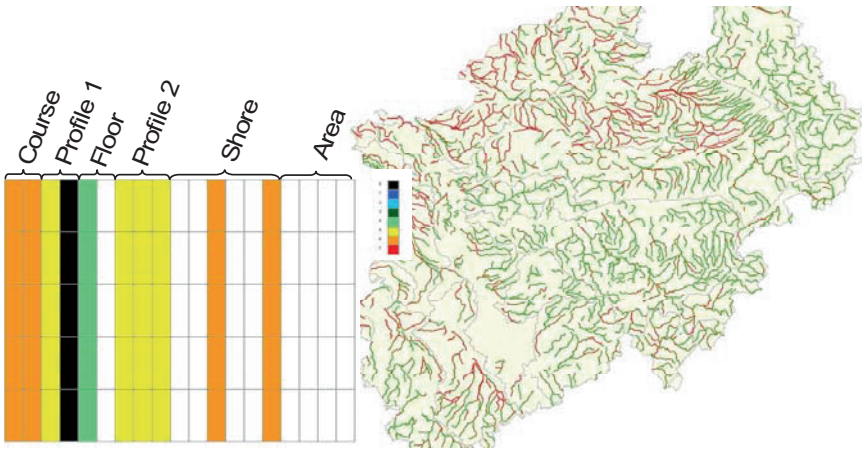

Figure 8: Cluster visualization for river data

3.2 MultiClustering

Environmental sensors produce data streams at successive time points which are often archived for further analysis. Applications like stock market analysis or weather stations gather ordered time series of different values in large temporal databases. Weather stations for example use multiple sensors to measure e.g. barometric pressure, temperature, humidity, rainfall. Many other scientific research fields like observatories and seismographic stations archive similar spatial or spatio-temporal time series.

In a current project of the European Union on renaturation of rivers, the structural quality of river segments is analyzed. For German rivers, about 120,000 one-hundred-meter segments were evaluated according to 19 different structural criteria, e.g. quality of the riverbed [LUA03]. They were mapped to quality categories. The sequence order of the segments is given by the flowing direction of the rivers. Hydrologists have devised packages of measures for different structural damages, formalized in rules specifying the attribute value constraints and the attributes influenced positively. An example constraint might be that a certain segment has good quality riverbed and riverbanks and poor river bending. This could be improved by measures like adding deadwood to positively influence river bending. Finding and analyzing these patterns helps hydrologists summarize the overall state of rivers, give compact representations of typical situations and review the extent to which these situations are covered by measures envisioned. They can identify those patterns which are problematic, i.e. have low quality ratings, but are not yet covered by measures.

From a computer science point of view, finding the intrinsic structure of these multidimensional time series is a two-fold task: (1) detect frequent patterns within time series for all possible pattern lengths, (2) then detect parallel occurrences of these patterns.

Patterns are ranges of values (which correspond to several categories of river quality struc-

ture) found in several (sub-)time series. Pattern analysis has to take into account that the data is subjective and fuzzy, because structural quality of rivers was mapped by different individuals. Our approach is based on weighting by kernel densities in a density-based clustering approach. This effectively detects fuzzy time series patterns. These patterns are clustered efficiently for arbitrary lengths using monotonicity properties. We transform these time series pattern clusters into a cluster position space such that mining parallel patterns can be reduced to efficient FP-tree frequent itemset mining.

Multiclusters are visualized in a geographical information system, as e.g. in Figure 8. Those river segments which are part of the cluster are marked by dark lines. The corresponding cluster summary is depicted on the left side. It indicates the cluster length of five river sections (top to bottom) as well as the cluster range of ten out of nineteen attributes (left to right). A more detailed discussion can be found in [AKGS06, AKGS08].

4 Conclusion

In this work, we propose efficient and effective retrieval and mining of multimedia data for both histogram and time series data. Key techniques of our work are database methods. Indexing organizes the data such that during query processing or mining, only relevant parts of the data have to be accessed to allow for efficient runtimes. Coupled with multistep filter-and-refine architectures, we are able to further enhance runtime performance by fast generation of candidate sets in a filter step that is refined to ensure neither false negatives nor false positives.

References

[AKAS08] Ira Assent, Ralph Krieger, Farzad Afschari, and Thomas Seidl. The TS-Tree: Efficient Time Series Search and Retrieval. In *Proceedings of the 11th International Conference on Extending Data Base Technology (EDBT), Nantes, France*, 2008.

[AKGS06] Ira Assent, Ralph Krieger, Boris Glavic, and Thomas Seidl. Spatial Multidimensional Sequence Clustering. In *Proceedings of the 1st International Workshop on Spatial and Spatio-temporal Data Mining (SSTDM), in conjunction with the 2006 IEEE International Conference on Data Mining (ICDM)*, 2006.

[AKGS08] Ira Assent, Ralph Krieger, Boris Glavic, and Thomas Seidl. Clustering Multidimensional Sequences in Spatial and Temporal Databases. *International Journal on Knowledge and Information Systems (KAIS)*, 2008.

[AKMS08] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. EDSC: Efficient Density-Based Subspace Clustering. In *Proc. ACM 17th Conference on Information and Knowledge Management (CIKM), Napa Valley, USA*, pages 1093–1102, 2008.

[AKS07] Ira Assent, Ralph Krieger, and Thomas Seidl. AttentionAttractor: Efficient video stream similarity query processing in real time. In *Proceedings of the IEEE 2007 International Conference on Data Engineering (ICDE)*, pages 1509–1510, 2007.

[AN07] Arthur Asuncion and David Newman. University of California Machine Learning Repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

[AWMS08] Ira Assent, Marc Wichterich, Tobias Meisen, and Thomas Seidl. Efficient similarity search using the Earth Mover's Distance for large multimedia databases. In *Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE), Cancun, Mexico*, 2008.

[AWS06] Ira Assent, Andrea Wenning, and Thomas Seidl. Approximation Techniques for Indexing the Earth Mover's Distance in Multimedia Databases. In *Proceedings of the 2006 IEEE International Conference on Data Engineering (ICDE)*, 2006.

[BC94] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pages 229–248, 1994.

[Dan51] George Dantzig. *Application of the simplex method to a transportation problem*, pages 359–373. John Wiley and Sons, 1951.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregano, 1996. AAAI Press.

[HK01] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[HL90] Frederick Hillier and Gerald Lieberman. *Introduction to Linear Programming*. McGraw-Hill, 1990.

[Keo02] Eamonn Keogh. Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Databases (VLDB)*, pages 406–417. VLDB Endowment, 2002.

[KKK04] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density-Connected Subspace Clustering for High-Dimensional Data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 246–257, 2004.

[KWX+06] Eamon Keogh, Li Wei, Xiaopeng Xi, Sang-Hee Lee, and Michail Vlachos. LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pages 882–893, 2006.

[LUA03] LUA NRW. River quality data. http://www.lua.nrw.de, 2003.

[RT01] Yossi Rubner and Carlo Tomasi. *Perceptual Metrics for Image Database Navigation*. Kluwer Academic Publishers, 2001.

[SZ04] Karlton Sequeira and Mohammed Zaki. SCHISM: A New Approach for Interesting Subspace Mining. In *Proceedings of the 2004 IEEE International Conference on Data Mining (ICDM)*, pages 186–193, 2004.

[WAKS08] Marc Wichterich, Ira Assent, Philipp Kranen, and Thomas Seidl. Efficient EMD-based Similarity Search in Multimedia Databases via Flexible Dimensionality Reduction. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Vancouver, BC, Canada.*, pages 199–212, 2008.