Parameterized Earth Mover's Distance for Efficient Metric Space Indexing

Jakub Lokoč ○ Christian Beecks ● Thomas Seidl ● Tomáš Skopal ○
○SIRET Research Group, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
○{lokoc,skopal}@ksi.mff.cuni.cz
●Data Management and Data Exploration Group, RWTH Aachen University, Germany
●{beecks,seidl}@cs.rwth-aachen.de

ABSTRACT

The Earth Mover's Distance is a well-known distance measure employed in various domains, especially for content-based retrieval in multimedia databases. However, the distance evaluation is a considerably expensive task and thus for large multimedia databases, efficient query processing becomes a challenging problem. In this paper, we introduce a parameterized version of the Earth Mover's Distance that can be used by database experts to change the distance distribution in the derived distance space in order to improve the *indexability.* We empirically show, that we can significantly improve the indexability of the distance space and that we can tune the retrieval quality by adapting the parameterized Earth Mover's Distance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Search Process; H.3.1 [**Content Analysis and Indexing**]: Indexing Methods

1. INTRODUCTION

The similarity search using metric distances is proved to be a suitable approach applicable to various domains where a database $\mathbb{S}$ consists of objects represented by nontrivial feature descriptors (histograms, signatures) extracted from unstructured raw data. In such databases, a distance function δ provided by the domain experts focusing on effectiveness is often the only available information for the database experts which try to create indexes allowing efficient querying over such data.

During the last two decades, there emerged a class of distance-based indexes considering metric distances, so called metric access methods [4, 9]. These metric access methods utilize metric postulates to partition a database $\mathbb{S}$ into classes of similar objects, which are used during query processing to filter non-relevant objects (avoiding also expensive distance evaluations). However, a provided distance measure satisfying metric postulates is not enough, a good data distribution in the distance space is also necessary for successful metric indexing. Therefore, the intrinsic dimensionality [4] has been introduced as an indexability measure of a distance space.

In general, for distance spaces suffering from high intrinsic dimensionality, e.g., the *Earth Mover's Distance* [7] on feature signatures [7, 3], it is impossible to create an efficient metric index. Hence, in this paper we enter the world of domain experts and provide the *parameterized Earth Mover's Distance*, which can be modified for balancing the trade-off between indexability and retrieval quality. We can consider this new approach as a novel type of approximate scheme where a marginal decrease in retrieval quality is accompanied by a significant increase in indexability.

2. PARAMETERIZED EARTH MOVERS DISTANCE

The *Earth Mover's Distance* [7] is a well-known distance measure originated in the computer vision domain. Its successful utilization gave raise to numerous applications in different domains. This distance describes the cost for transforming one feature signature into another one. Distance is considered to be a transportation problem and thus is the solution to a linear optimization problem which can be solved via a specialized simplex algorithm. The Earth Mover's Distance is defined between two feature signatures S^q and S^o as a minimum cost flow over all possible flows $f_{ij} \in \mathcal{R}$ as:

$$EMD_d(S^q, S^o) = \min_{f_{ij}} \left\{ \frac{\sum_i \sum_j f_{ij} \cdot d(c_i^q, c_j^o)}{\min\{\sum_i w_i^q, \sum_j w_j^o\}} \right\},$$

subjected to the constraints: $\forall i : \sum_j f_{ij} \leq w_i^q$, $\forall j : \sum_i f_{ij} \leq w_j^o$, $\forall i,j : f_{ij} \geq 0$, and $\sum_i \sum_j f_{ij} = \min\{\sum_i w_i^q, \sum_j w_j^o\}$. These constraints guarantee a feasible solution, i.e. all costs are positive and do not exceed the limitations given by the weights in both feature signatures.

However, as there is a minimization problem to solve, the computation time complexity is considerably high. Techniques providing efficient similarity search on large multimedia databases using the Earth Mover's Distance can, for instance, be found in [8, 2, 1]. Another approach for efficient indexing of the Earth Mover's Distance could be the metric space indexing. However, the distance spaces based on the Earth Mover's Distance usually suffer from high intrinsic dimensionality and thus only the approximate search

SISAP '11, June 30-July 1, 2011, Lipari, Italy.
Copyright 2011 ACM 978-1-4503-0795-6/11/06 ...$10.00.

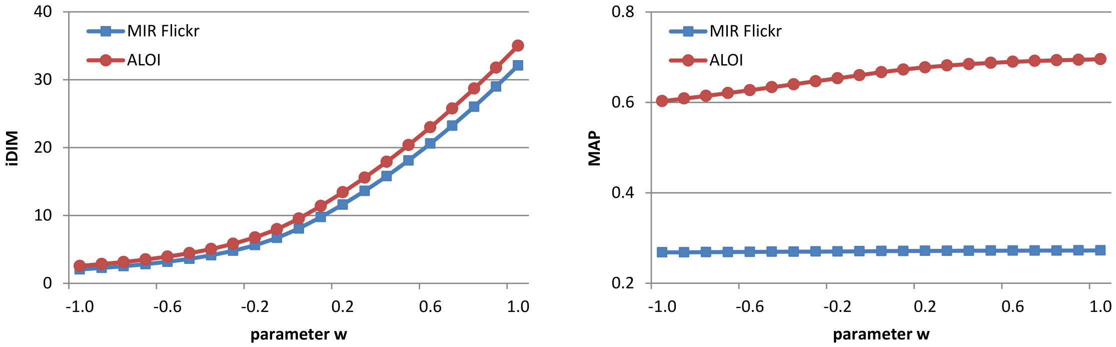

Figure 1: The intrinsic dimensionality (iDIM) and mean average precision (MAP) for the MIR Flickr and ALOI database.

can speedup query processing by metric index.

In this paper, we propose a parameterized version of the Earth Mover's Distance which is defined as:

$$pEMD_d(S^q, S^o, w) = \min_{f_{ij}} \left\{ \frac{\sum_i \sum_j f_{ij} \cdot \mathrm{FP}(d(c_i^q, c_j^o), w)}{\min\{\sum_i w_i^q, \sum_j w_j^o\}} \right\},$$

where FP is the fractional power modifier defined as:

$$\mathrm{FP}(x, w) = \begin{cases} x^{\frac{1}{1+w}} & \text{for } w > 0 \\ x^{1-w} & \text{for } w \leq 0 \end{cases}$$

The intuition behind this modification is quite simple – depending on the parameter w, we can either suppress ($w > 0$) or strengthen ($w < 0$) the transportation costs to outlying bins/centroids when comparing feature histograms/signatures. Hence we can tune the robustness of the measure, i.e., what is the impact of outliers (noise bins or clusters) on the overall distance. We empirically prove in our experiments, that changing the parameter w of the parameterized Earth Mover's Distance affects both mean average precision and the intrinsic dimensionality of the corresponding distance space. However, we have to consider also the topological properties of the modified distance, and unfortunately, the parameterized Earth Mover's Distance employing the FP modification is not a metric for all values of the parameter $w \in \mathbb{R}^-$. Nevertheless, we have inspected in the experiments, that for sufficiently low intrinsic dimensionality (iDIM=5.1), only a negligible fraction of randomly selected triplets (0.1%) do not fulfill triangle inequality, and we can thus use metric access methods for efficient and near-exact search.

3. PRELIMINARY EXPERIMENTS

We conducted the experiments on the MIR Flickr [6] and ALOI [5] databases comprising 25,000 and 72,000 images, respectively, and extracted feature signatures based on color, position, and texture information, similar to [3].

Figure 1 depicts the intrinsic dimensionality (iDIM) and the mean average precision (MAP) values for the aforementioned databases by changing the parameter w of the pEMD. As can be seen in the figure, the intrinsic dimensionality decreases for both databases with decreasing parameter w, while the retrieval quality stays at a considerably high level over 0.6 for the ALOI database and over 0.26 for the MIR Flickr database.

Thus, we have briefly shown that the parameterized Earth Mover's Distance seems to be indexable by metric access methods as the intrinsic dimensionality is considerably low.

4. CONCLUSIONS

We have introduced a parameterized version of the Earth Mover's Distance that can be utilized for efficient flexible similarity search in multimedia databases. In the future, we would like to formally describe the observed behavior and try other modifying functions (not only FP modifier).

5. ACKNOWLEDGMENTS

This research has been supported by Czech Science Foundation projects 201/09/0683, 202/11/0968 (first and fourth author) and by the Deutsche Forschungsgemeinschaft within the Collaborative Research Center SFB 686.

6. REFERENCES

[1] I. Assent, A. Wenning, and T. Seidl. Approximation techniques for indexing the earth mover's distance in multimedia databases. In *Proc. IEEE ICDE*, pages 11–22, 2006.

[2] I. Assent, M. Wichterich, T. Meisen, and T. Seidl. Efficient similarity search using the earth mover's distance for large multimedia databases. In *Proc. IEEE ICDE*, pages 307–316, 2008.

[3] C. Beecks, M. S. Uysal, and T. Seidl. A comparative study of similarity measures for content-based multimedia retrieval. In *Proc. IEEE ICME*, pages 1552–1557, 2010.

[4] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in Metric Spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[5] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam Library of Object Images. *IJCV*, 61(1):103–112, 2005.

[6] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proc. ACM MIR*, pages 39–43, 2008.

[7] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *IJCV*, 40(2):99–121, 2000.

[8] M. Wichterich, I. Assent, P. Kranen, and T. Seidl. Efficient emd-based similarity search in multimedia databases via flexible dimensionality reduction. In *Proc. ACM SIGMOD*, pages 199–212, 2008.

[9] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.