Data-driven Learned Metric Index: an Unsupervised Approach

Terézia Slanináková[1] and **Matej Antol**[1] and **Jaroslav Oľha**[1] and **Vojtěch Kaňa**[1] and **Vlastislav Dohnal**[1]

[1]Faculty of Informatics, Masaryk University

Abstract

Metric indexes are traditionally used for organizing unstructured or complex data to speed up similarity queries. The most widely-used indexes cluster data or divide space using hyper-planes. While searching, the mutual distances between objects and the metric properties allow for the pruning of branches with irrelevant data – this is usually implemented by utilizing selected anchor objects called pivots. Recently, we have introduced an alternative to this approach called **Learned Metric Index**. In this method, **a series of machine learning models** substitute decisions performed on pivots – the query evaluation is then determined by the **predictions of these models**. This technique relies upon a traditional metric index as a template for its own structure – this dependence on a pre-existing index and the related overhead is the main drawback of the approach.
In our paper [6], we propose a data-driven variant of the Learned Metric Index, which organizes the data using their descriptors directly, thus eliminating the need for a template. The proposed learned index shows significant gains in performance over its earlier version, as well as the established indexing structure M-index.

Data-driven LMI

Learned Metric Index (LMI), as introduced in [1], is a hierarchical tree index structure of nodes containing machine learning models. In general, the concept of LMI can be realized in two distinct ways. The first one involves using a pre-existing index and its data partitioning as labels for *supervised* training. We have examined this variant in [1] and demonstrated that it can achieve more than competitive performance with state-of-the-art methods. The other option is to **assemble LMI "from scratch" by letting it create its own meaningful divisions of the data**. Such approach exploits the information embedded in the descriptors of data objects to emulate the similarity function. This constitutes an *unsupervised* learning problem, which is the subject of our paper [6].

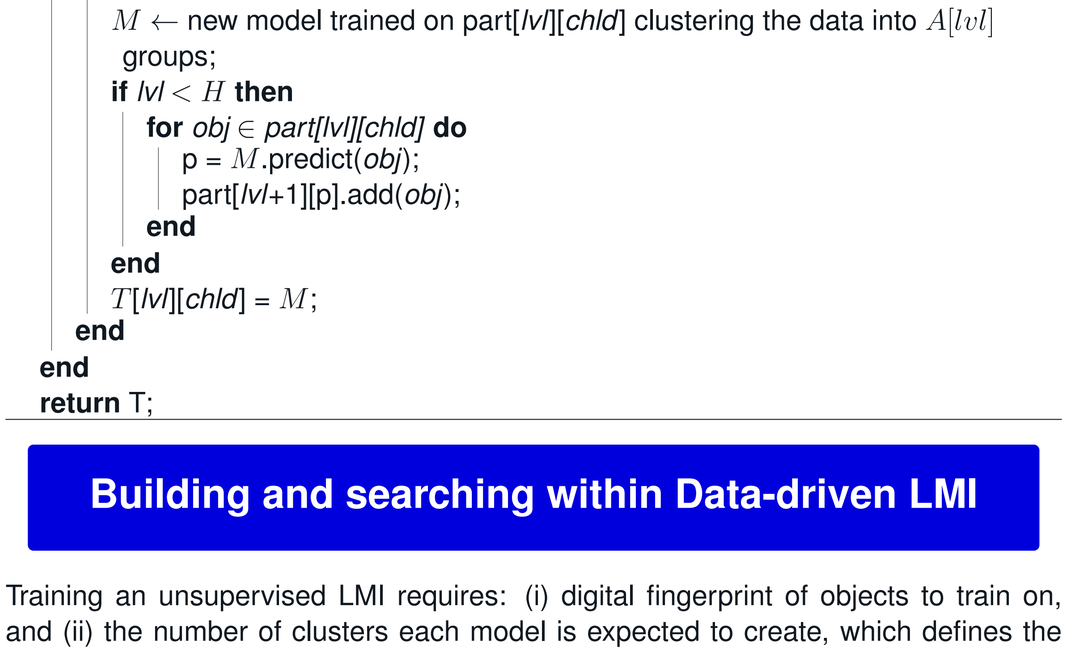
Algorithm 1: Data-driven Learned Metric Index training

Input: a data-set X, max. depth H (tree height), max. number of children per level $A[]$
Output: a tree of trained models $T[][]$

```
part[1][1] = X;
for lvl ← 1 to H do
    for chld ← 1 to A[lvl] do
        if part[lvl][chld] = ∅ then
            continue
        end
        M ← new model trained on part[lvl][chld] clustering the data into A[lvl]
         groups;
        if lvl < H then
            for obj ∈ part[lvl][chld] do
                p = M.predict(obj);
                part[lvl+1][p].add(obj);
            end
        end
        T[lvl][chld] = M;
    end
end
return T;
```

Building and searching within Data-driven LMI

Training an unsupervised LMI requires: (i) digital fingerprint of objects to train on, and (ii) the number of clusters each model is expected to create, which defines the shape of the learned index structure. The training procedure of the whole LMI then starts with the root model, which is trained on the entirety of the given data-set, while its descendants are trained on smaller and smaller portions of the data as we dive deeper into the structure.
During the training, each model is presented with a clustering problem. The objective is to organize the data into a pre-specified number of groups according to their mutual similarity obtained from the descriptors. Each training epoch re-organizes the data to allow mutually similar objects to end up in the same cluster.



Experiments

We have executed a wide range of experiments with three different multimedia data-sets: *CoPhIR*, *Profiset* and *MoCap*. CoPhIR [2] is a data collection of 282-dimensional vectors derived from five visual descriptors of images. Profiset [5] is a series of 4096-dimensional vectors extracted from Photo-stock images using a convolutional network. Finally, MoCap is HDM05 data-set [4] that consists of sequences of 3D skeleton poses, which were segmented to extract 4096-dimensional descriptors using AlexNet [3]. The data-set sizes were fixed at 1-million objects for CoPhIR and Profiset. MoCap contains 354,893 segments.

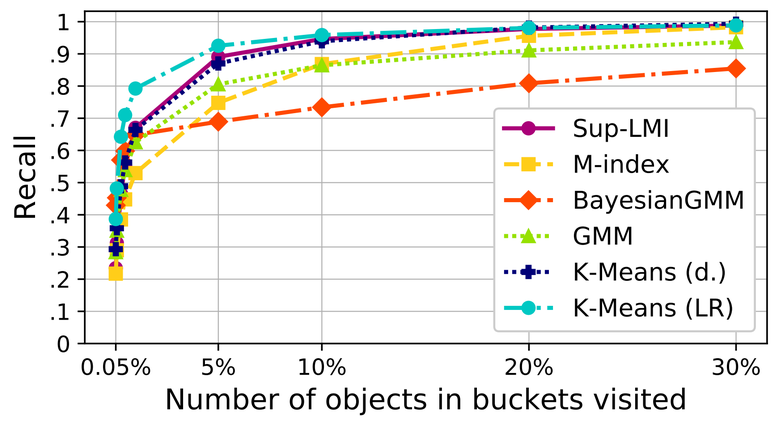


Fig. 2: Comparison between the recall of Data-driven LMI models, and the benchmarks – the best supervised setup from Antol et al. [1] (Sup-LMI), and M-index in the Profimedia dataset.

Conclusion

In our paper [6], we extend the capabilities of Learned Metric Index – a novel, machine-learning-based indexing paradigm introduced in [1]. We present a new means of LMI construction that builds the index from scratch – no pre-existing index is needed to guide the building process. Our experiments confirm that **building a data-driven LMI is a viable approach**, and **clustering algorithms within LMI create meaningful divisions of the data**. In comparison to the formerly introduced supervised LMI, the building costs are significantly lower. By far the **most significant benefit** of the data-driven LMI **is the overall search performance measured as recall in time** – our new approach **managed to beat both benchmarks (M-index and supervised LMI) by at least one order of magnitude in all cases**. If we measured performance as recall per portion of the index structure visited (navigation), the **data-driven LMI was superior to both benchmarks by approximately 10% in two out of the three tested data-sets**. On the third data-set, the unsupervised methods fell behind when searching a larger portion of the structures. However, even in these cases, the computation speed of the data-driven LMI outweighs the navigation deficit and reaches all accuracy thresholds in shorter time.

Acknowledgements

This research has been supported by the Czech Science Foundation project No. GA19-02033S. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] Matej Antol et al. "Learned Metric Index — Proposition of learned indexing for unstructured data". In: *Information Systems* 100 (2021). ISSN: 0306-4379.

[2] Michal Batko et al. "Building a web-scale image similarity search system". In: *Multimedia Tools and Applications* 47.3 (2009), pp. 599–629. ISSN: 1573-7721.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[4] M. Müller et al. *Documentation Mocap Database HDM05*. Tech. rep. CG-2007-2. Universität Bonn, 2007.

[5] David Novak, Michal Batko, and Pavel Zezula. "Large-scale Image Retrieval Using Neural Net Descriptors". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Santiago, Chile: ACM, 2015, pp. 1039–1040. ISBN: 978-1-4503-3621-5.

[6] Terézia Slanináková et al. "Data-driven Learned Metric Index: an Unsupervised Approach". In: *14th International Conference on Similarity Search and Applications*. Springer. 2021, pp. 1–14.