The Dataset-Similarity-Based Approach to Select Datasets for Evaluation in Similarity Retrieval

Matheus A. L. Matiazzo[1], Vitor de Castro-Silva[1], Rafael S. Oyamada[2], and Daniel S. Kaster[1](✉)

[1] University of Londrina, Londrina, Brazil
{vitor.castro.silva,dskaster}@uel.br
[2] University of Milan, Milan, Italy
rafael.oyamada@unimi.it

Abstract. Most papers on similarity retrieval present experiments executed on an assortion of complex datasets. However, no work focuses on analyzing the selection of datasets to evaluate the techniques proposed in the related literature. Ideally, the datasets chosen for experimental analysis should cover a variety of properties to ensure a proper evaluation; however, this is not always the case. This paper introduces the dataset-similarity-based approach, a new conceptual view of datasets that explores how they vary according to their characteristics. The approach is based on extracting a set of features from the datasets to represent them in a similarity space and analyze their distribution in this space. We present an instantiation of our approach using datasets gathered by surveying the dataset usage in papers published in relevant conferences on similarity retrieval and sample analyses. Our analyses show that datasets often used together in experiments are more similar than they seem to be at first glance, reducing the variability. The proposed representation of datasets in a similarity space allows future works to improve the choice of datasets for running experiments in similarity retrieval.

Keywords: Similarity Retrieval · Datasets · Experimental Analysis · Similarity Space of Datasets

1 Introduction

Similarity search enables retrieving similar elements from a database given one or more reference elements according to their features. An important aspect to consider when developing new methods for similarity searching is the choice of datasets to evaluate the proposed methods. There are several datasets in the field of similarity search for the most varied applications [3,9,12]. However, in

This work has been financially supported by the Brazilian funding agencies CNPq and Araucaria Foundation, and by the University of Milan, Italy.

O. Pedreira and V. Estivill-Castro (Eds.): SISAP 2023, LNCS 14289, pp. 125–132, 2023.
https://doi.org/10.1007/978-3-031-46994-7_11

most works in similarity searching, the selection of datasets for evaluating the proposals is solely based on experience, without a deeper variability analysis.

In the literature, there are several papers describing the underlying characteristics of datasets that make them more or less difficult for similarity search purposes. Particularly, dataset properties that might affect the performance of similarity search methods include intrinsic dimensionality [2], relative contrast [7], fractal correlation [4], and the concentration of distances [6]. However, works in the field rely on properties like these to select the datasets for experimentation without analyzing the interconnection between properties that affect the dataset variability. To the best of our knowledge, no previous research covers a wide range of these datasets, focusing on jointly analyzing their underlying characteristics for similarity search purposes.

In this paper, our goal is to survey and analyze the properties of datasets employed by researchers in the field of similarity search in order to provide a better understanding of the diversity of datasets studied. We accomplish that by gathering several datasets used in the literature and extracting an assortion of their features. The relevance of this contribution relies on considering joint properties to represent the datasets using a new conceptual view, called the *dataset-similarity-based* approach. Employing meaningful features allows us to create a similarity space of datasets, which can be used to analyze the diversity of datasets for many purposes, including selecting datasets to conduct experiments on similarity retrieval methods. We present an instantiation of proposed conceptual view based on a survey on dataset usage in works in the field over the last decade and analyses illustrating the potential of the approach. The methods used to extract such features and the surveyed datasets are publicly available.

This paper is organized as follows. Section 2 presents the background to understand the paper and related work. Section 3 describes the proposed approach, including how we define the dataset similarity space and the methodology we employed to survey the dataset usage in papers on similarity retrieval in the last decade to instantiate our approach. Section 4 presents sample analyses using our proposed dataset-similarity-space view, and in Sect. 5 we conclude.

2 Background and Related Work

The most intuitive strategy to evaluate searching algorithms is to compare their performances across datasets with different characteristics, such as cardinality and dimensionality. However, the choice of the datasets employed in the evaluation is fundamentally based on experience, which can create a bias in the choice, such as selecting a dataset only based on popularity.

Several metrics can be extracted from datasets to measure the complexity of a similarity search problem [2,10]. Existing metrics include the Relative Contrast [7], intrinsic dimensionality [2], fractal correlation [4], and the concentration of distances [6]. These metrics can characterize datasets providing valuable insights for choosing similarity search algorithms and their parameters.

For instance, Aumller and Ceccarello [2] describe an application of the local intrinsic dimensionality to measure the complexity of a dataset. The authors show that the so-called *curse of dimensionality* is not the leading cause to degrade the performance of similarity queries and that the intrinsic dimensionality of a dataset can be used to obtain better insights. On the other hand, in [14], we proposed a learned approach to map the combination of dataset metrics to the performance of proximity graph algorithms. This way, it becomes possible to predict the performance of different algorithms for a given dataset and to choose the best algorithm for a given dataset without having to run a greedy search to find a suitable algorithm configuration [16].

Understanding the complexity of datasets is also a relevant subject in other fields. In [15], Šikonja presents a study comparing complex datasets to determine if datasets and subsets are similar enough to be used together in a data mining task. The proposal consists of a methodology to compare the statistical properties of the attributes and the similarity between clusters of elements. In the machine learning field, [11] presents a study on the properties of datasets aiming at understanding the impact of these properties on the performance of classification algorithms. These properties include generic properties (e.g., dimensionality and cardinality), classification properties (e.g., class imbalance and number of classes), and neighborhood properties (e.g., number of clusters and hub score). In a similar direction, a new research field called dataset discovery has recently emerged in the literature [5]. The problem consists of finding datasets that are relevant to a given query. However, most works in this field are keyword-based [8], metadata-based [13], or context-based [1].

Therefore, to the best of our knowledge, there is no survey in the literature with an analysis of the datasets used by researchers in similarity search and no approach that considers the similarity between datasets as we propose in this work. Defining a dataset similarity space is challenging as it requires gathering a comprehensive amount of datasets and identifying a set of features to extract from them concerning a variety of aspects. This is a laborious task as datasets commonly used come from different sources scattered over several repositories, most poorly structured and often storing inconsistent versions of the same dataset. Moreover, defining a suitable set of features to extract from the datasets still demands investigation.

3 A Dataset-similarity-Based Proposal to Support Experimental Analysis

In this work, we introduce a novel view of datasets employed for experimental analysis in similarity retrieval. The choice of a set of datasets to evaluate a new method is challenging, both to cover a variety of properties of the underlying problem and to discuss the results.

Our proposal is to define a similarity space of datasets and observe how they are spread across this space to gain an understanding of their variability. We call the *dataset-similarity-based* approach such a conceptual view, which has the potential to provide a broader knowledge of datasets based on their intrinsic properties. Our proposal is based on the fact that datasets vary according to a multitude of properties that may contribute to the selection criterion for experimentation and to interpret the results. Using an assortion of features to represent datasets in a similarity space provides a joint view of how the datasets are distributed according to their properties.

3.1 Definition of the Dataset Similarity Space

This section details how we define the datasets' similarity space. We employed features comprising three categories.

1. Statistical and information-theoretical measures on the dataset attributes.
2. Measures commonly used in similarity retrieval analysis, such as the dataset cardinality and embedding dimensionality.
3. Measures of the hardness of similarity searching, including local intrinsic dimensionality, relative variance, and features derived from these complexity metrics (e.g., histograms of local intrinsic dimensionality).

For the complete list of features, refer to the implementation we developed to extract dataset features, available online[1]. We highlight that the most important point is the dataset-similarity-based conceptual view we propose. The instantiation we present herein considers a wide number of features to describe the datasets, but it is not intended to be exhaustive. Certainly, there is room for improvement following our conceptual view.

3.2 Survey of Dataset Usage in Similarity Retrieval Evaluation

To instantiate our proposal, we surveyed the datasets used by papers related to similarity search and gathered those publicly available in a centralized repository. We limited the survey scope to papers published in the International Conference on Similarity Search and Applications (SISAP), which is the principal forum on similarity search, and in the two most prestigious conferences on databases, namely the International Conference on Management of Data (SIGMOD) and the International Conference on Very Large Data Bases (VLDB). We justify such a limitation due to the need to survey the datasets used in the papers, which multiplies the effort.

The analysis covered more than one decade, comprising the years between 2008, which was the first edition of SISAP, and 2020. Every article published in any of the three conferences in the period was manually checked to verify if the main contribution is on similarity searching (e.g., an index method, a search algorithm, etc.), adding up to 146 articles. Other topics related to similarity were kept out of our scope.

Then, we searched for the dataset information presented in the papers, collecting features like the dataset's common name, the raw data type (e.g., image, text, etc.), the feature vector extracted (e.g., for images, color histogram, texture features, SIFT, etc.), the cardinality, the dimensionality, and others. We

[1] https://github.com/raseidi/annmf.

identified 461 dataset mentions in the papers, comprising 237 distinct dataset instances. We performed a manual data integration of the instances, grouping datasets we understand as the same dataset by the context provided in the paper. Finally, we imputed missing dataset features with information about the datasets from external sources, being the original source whenever available.

From the datasets surveyed, we found 49 available for download. We generated several subsets of many of these datasets through random sampling to explore the impact of cardinality in the dataset properties, as this is a key performance factor for searching. In total, the repository built in this work is composed of 198 datasets. Then, we extracted features for all datasets in the repository.

4 Analyses in the Dataset Similarity Space

This section presents sample analyses to illustrate what can be done using our dataset-similarity-based conceptual view. We performed a Principal Component Analysis (PCA) on the features of all datasets in the built repository and generated visualizations using the two first principal components to show how they relate to each other in the similarity space.

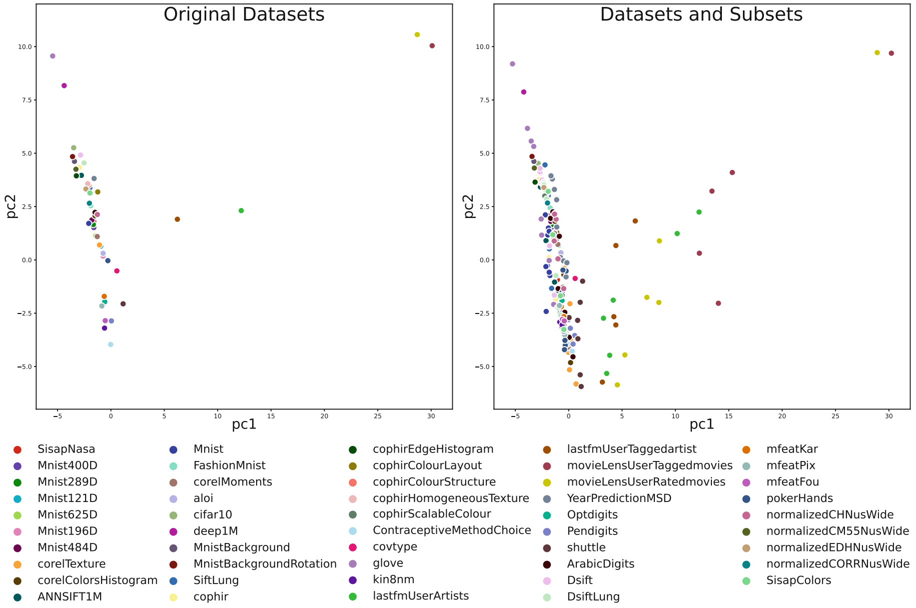

Fig. 1. Datasets and sub-datasets spreaded out in the similarity space

Figure 1 shows on the left the distribution of the 46 original (complete) datasets, and on the right, the distribution of all datasets (including the subsets). In this paper's figures, subsets of the same dataset have the same color.

It is clear that subsets promoted a significant variability in most datasets, highlighting the significance of cardinality for dataset characterization. The fact that most datasets got clustered parallel to the PC2 axis in the visualization does not indicate that they are redundant but that the PCA is highly impacted by the “outliers” *movieLensUser* and *lastfmUser* datasets.

Figure 2 shows the dataset distribution classified by ranges of cardinality, embedding dimensionality, and intrinsic dimensionality. As shown, datasets with small cardinality tend to cluster on the lower area, datasets with average and high cardinality in the middle, and datasets with very high cardinality are mainly on the upper part of the graph. Regarding dimensionality, it is clear that intrinsic dimensionality is more relevant for dataset distribution than embedding dimensionality, since intrinsic dimensionality more clearly spreads datasets. The datasets out of the main cluster are challenging datasets, such glove (Global Vectors for Word Representation) and deep1m (features derived from convolution neural networks), with high to very high cardinality and intrinsic dimensionality.

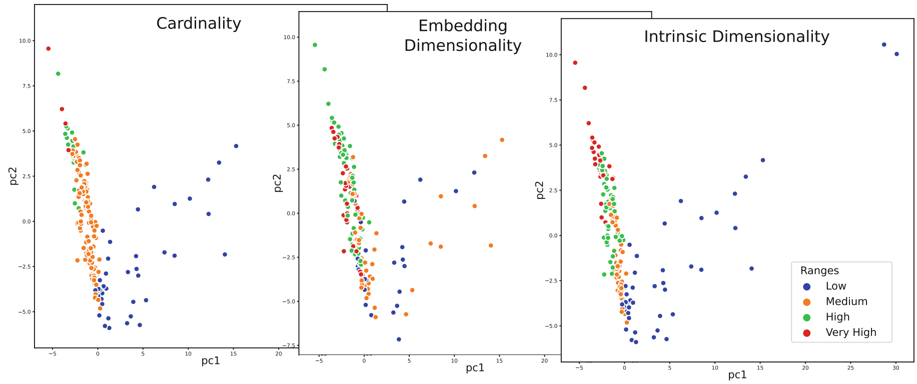

Fig. 2. Dataset distribution according to ranges of cardinality, embedding dimensionality and intrinsic dimensionality

An example of feature analysis that helps complementary dataset selection considers the MNIST dataset. Figure 3(a) shows all datasets used with MNIST in at least one paper surveyed. From this figure, it is possible to see the variance of datasets used with MNIST. With this analysis, a work could gather a diverse number of datasets by making a choice that maximizes the feature space of this representation. A viable choice would be the datasets CoPhIR or pokerHands, since the feature space would be relatively large. On the other hand, FashionMNIST would result in a less diverse dataset selection.

Another use case for the tools made available by this work is the ability to create experiments with any other datasets that other studies may use. To illustrate this usage, a comparison was made between datasets and sub-datasets created by the SIFT and Dense SIFT extractors when applied to an image dataset. While the SIFT extractor picks central points in an image and generates feature vectors based on these points, Dense SIFT leverages points over the whole image. Both extractors were applied to two datasets of fairly distinct image types: images from lung medical exams and random images from Flickr. Figure 3(b) shows the distribution of the datasets and sub-datasets created. It is clear that the most significant difference is between extractors and not between the image types, clustering the datasets by the feature regardless of the image type.

We showed sample analyses that could be done using our similarity-based conceptual view of datasets. It allows for gaining a broader knowledge of the distribution and variability of datasets employed for evaluating similarity search methods according to the datasets' intrinsic features. Potential applications of this approach include the selection of datasets for experimentation based on a varied number of properties, enhancing the variability among the chosen ones, and taking advantage of dataset similarities to guide decisions, such as parameter recommendation for indexing structures, as we did in a previous work [14].

5 Conclusion

In this work, we introduced the new dataset-similarity-based approach to handle datasets for different purposes. Our approach considers joint properties to represent the datasets as a similarity space composed of extracted features. We presented an instantiation of the approach based on a survey on how similarity search researchers have used datasets and showed analyses highlighting that it is possible to enhance data set choices when feature variance is a relevant selection criterion.

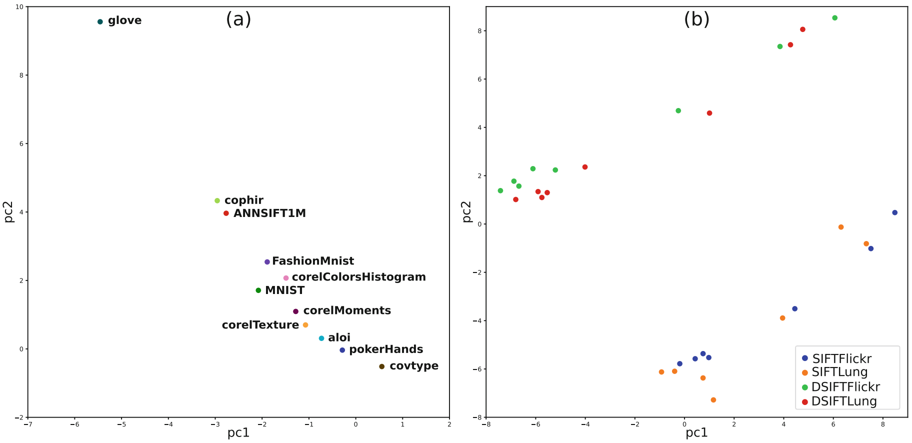

Fig. 3. MNIST and datasets used at least once with MNIST (no sub-datasets) and features SIFT and Dense SIFT extracted from different image types

Future work includes advancing the approach, for instance, by adding more features to describe the datasets and expanding the repository since the more datasets are included in the feature space, the richer the data set choice for other studies becomes. We also plan to employ the dataset-similarity-based conceptual view to support decisions or applications demanding dataset analysis.

References

1. Altaf, B., Akujuobi, U., Yu, L., Zhang, X.: Dataset recommendation via variational graph autoencoder. In: ICDM, pp. 11–20. IEEE (2019)
2. Aumüller, M., Ceccarello, M.: Benchmarking nearest neighbor search: influence of local intrinsic dimensionality and result diversity in real-world datasets. In: EDML SDM. CEUR Workshop Proceedings, vol. 2436, pp. 14–23. CEUR-WS.org (2019)
3. Bolettieri, P., et al.: CoPhIR: a test collection for content-based image retrieval. CoRR abs/0905.4627 (2009)
4. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method. IEEE Trans. Pattern Anal. Mach. Intell. **24**(10), 1404–1407 (2002)
5. Chapman, A., et al.: Dataset search: a survey. VLDB J. **29**(1), 251–272 (2020)
6. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. **19**(7), 873–886 (2007)
7. He, J., Kumar, S., Chang, S.: On the difficulty of nearest neighbor search. In: ICML. icml.cc/Omnipress (2012)
8. Hendler, J.A., Holm, J., Musialek, C., Thomas, G.: US government linked open data: semantic.data.gov. IEEE Intell. Syst. **27**(3), 25–31 (2012)
9. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
10. Li, W., et al.: Approximate nearest neighbor search on high dimensional data - experiments, analyses, and improvement. IEEE Trans. Knowl. Data Eng. **32**(8), 1475–1488 (2020)
11. Lorena, A.C., Garcia, L.P.F., Lehmann, J., de Souto, M.C.P., Ho, T.K.: How complex is your classification problem?: A survey on measuring classification complexity. ACM Comput. Surv. **52**(5), 1–34 (2019)
12. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157. IEEE Computer Society (1999)
13. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. ACM J. Data Inf. Qual. **8**(1), 1–29 (2016)
14. Oyamada, R.S., Shimomura, L.C., Barbon, S., Jr., Kaster, D.S.: A meta-learning configuration framework for graph-based similarity search indexes. Inf. Syst. **112**, 102123 (2023)
15. Robnik-Šikonja, M.: Dataset comparison workflows. Int. J. Data Sci. **3**(2), 126–145 (2018)
16. Shimomura, L.C., Oyamada, R.S., Vieira, M.R., Kaster, D.S.: A survey on graph-based methods for similarity searches in metric spaces. Inf. Syst. **95**, 101507 (2021)