Similarity vs. Relevance: From Simple Searches to Complex Discovery

Tomáš Skopal ✉[1][0000−0002−6591−0879], David Bernhauer[1,2][0000−0003−2368−7506], Petr Škoda[1][0000−0002−2732−9370], Jakub Klímek[1][0000−0001−7234−3051], Martin Nečaský[1][0000−0002−5186−7734]

[1]SIRET research group, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
⟨name⟩.⟨surname⟩@matfyz.cuni.cz
[2]Faculty of Information Technology, Czech Technical University in Prague, Czech Republic

Abstract. Similarity queries play the crucial role in content-based retrieval. The similarity function itself is regarded as the function of relevance between a query object and objects from database; the most similar objects are understood as the most relevant. However, such an automatic adoption of similarity as relevance leads to limited applicability of similarity search in domains like entity discovery, where relevant objects are not supposed to be similar in the traditional meaning. In this paper, we propose the meta-model of data-transitive similarity operating on top of a particular similarity model and a database. This meta-model enables to treat directly non-similar objects $\mathbf{x}$,$\mathbf{y}$ as similar if there exists a chain of objects $\mathbf{x}$,i_1,...,i_n,$\mathbf{y}$ having the neighboring members similar enough. Hence, this approach places the similarity in the role of relevance, where objects do not need to be directly similar but still remain relevant to each other (transitively similar). The data-transitive similarity concept allows to use standard similarity-search methods (queries, joins, rankings, analytics) in more complex tasks, like the entity discovery, where relevant results are often complementary or orthogonal to the query, rather than directly similar. Moreover, we show the data-transitive similarity is inherently self-explainable and non-metric. We discuss the approach in the domain of open dataset discovery.

1 Introduction

When searching data, we can choose from a multitude of available models and paradigms. Some models assume exact data structure and semantics, such as the relational database model (and SQL) or graph database model (RDF+SPARQL, XML+XQuery). In such models, the relevance of a data entity to a particular query is binary (relevant/not relevant); specified by a binary predicate. The precision and recall in retrieval of structured data is always 100% as there is no uncertainty expected. Also, structured query languages offer high expressive power that allows the user to specify the relevance of data in many ways.

On the other side of the data universe, when searching in unstructured or loosely structured data (like multimedia, text, time series), we do not have enough a-priori information on how to model the data features for exact search. In such situation the similarity search models could be used, representing a universal way of content-based retrieval in unstructured data. Instead of formulating a structured query aiming at binary relevance, in similarity search we use a ranking of the database objects determined by their similarity score to a query example (the query-by-example paradigm). Hence, the relevance is relaxed from binary to multiple-value. When compared to retrieval of structured data, the similarity search is more like an "emergency solution" for unstructured data. The expressive power of similarity queries is limited to a ranking induced by numeric aggregation of differences between the query example and the database objects; keeping it a black-box search for the user. The low expressive power of the query-by-example paradigm leads to a paradox – we search for what we already have. Specifically, we query for as good results as possible, having the best result already at hand – the query example. Of course, in practical applications the query-by-example paradigm makes sense, because the query example itself does not contain the whole information we search for. For instance, searching by the photo of Eiffel tower we not only get another Eiffel tower image, but also some context (the Wikipedia web page the result image was embedded in). Nevertheless, the context (external information attached to data) does not remove the essence of the paradox – based purely on the similarity of results, the query example itself is always the best result[1].

Historically, the low expressive power of similarity search has been accepted in the major application area – the multimedia retrieval. Here the semantics to be captured in multimedia objects (the descriptors) is rather vague, general and bound to human common knowledge. The similarity search is thus a perfect method for multimedia retrieval as the similarity concept itself is vague and general (and so is the human cognition – the inspiration for similarity search). When combined with descriptor models employing high-level "canonized" semantics, such as the bag of words using the vocabulary of deep features [11], then even the cosine similarity can perform well. Unfortunately, the domain experts are not always so lucky to work with nicely shaped semantic descriptors, while then the low expressive power of similarity search is fully revealed. A solution to this could be a proposal of similarity-aware relevance of data objects to an example object (query) that enables much more complex aggregation than just evaluating the direct similarity (the "exampleness" of the results). If we find a way of how to extend the concept of similarity into a relevance, we would be able to use the existing similarity search methods in more expressive retrieval scenarios. For example, consider a fashion e-shop where a user searches for a product by an example image, e.g., shoes. The result could not only consist of similar shoes, but it could also return related accessories (handbag, belt) sharing some design features with the shoes [14].

[1] Let's omit another problem; where to acquire such a "holy grail" example in real-world problems.

In the following, we continue the discussion in the specific domain of open datasets discovery. Unlike in multimedia retrieval, where direct audiovisual similarity to a query usually leads to good results, in open datasets with sparse descriptors we often do not find anything directly (non-trivially) similar. Here the similarity extended towards more general relevance could improve the retrieval effectiveness in a fundamental way.

1.1 Discovery of Open Datasets by Similarity

The similarity search models can utilize not only content features but also metadata (if available). The focus on metadata can be efficient and effective in domains where the content of the objects is too heterogeneous so that it is hard to extract features for measuring similarity (or relevance). On the other hand, such objects could be catalogued by a community to enable search of the objects by metadata.

This is the case of the domain of open datasets search and discovery [12]. There are various datasets published on the internet which are catalogued in open data catalogs [18]. They are extremely heterogeneous in structure and semantics so that modeling them by content is nearly impossible (consider tables and spreadsheets without schema, full-text reports, database dumps, geographical and map data, logs, etc.). Open data catalogs provide descriptive metadata about the datasets in a single place where potential consumers can search for datasets. However, the problem of metadata is that they are often sparse and poor. In the open data domain, dataset publishers usually limit their descriptive metadata to briefly describe the core semantics of their datasets (by title, keywords, text description). No broader context of a dataset including some description of its relationships to other datasets is specified in the metadata. Using such sparse metadata for similarity retrieval is therefore limited. We confirmed this in our previous work [26] where we showed that various similarity methods do not perform very well when applied to the descriptive metadata of open datasets.

In our experiments, we noticed situations where two datasets are relevant to each other but none of the similarity models is able to identify this relevance. Let us demonstrate this on a concrete example of open data published by public authorities in Czechia. The datasets are catalogued in the National Open Data Catalog (NODC)[2]. There are two datasets entitled *IDOL Integrated Transport System Tariff Zones* and *Traffic intensity on sections of motorways*. The similarity of both datasets based on their metadata descriptions is low according to various similarity models presented in [26]. However, when we reviewed the datasets manually we found out that they are very relevant to each other. The first one is related to public transport. The second one is related to transport on motorways. So when users find one of the datasets, they would like to get also the other dataset as well. What makes them relevant to each other is the background semantics which is not directly expressed in the descriptive metadata. Since it is not expressed in the metadata, no similarity model can work with

[2] `https://data.gov.cz/english/`

this. However, there is a third dataset in NODC titled *BKOM transport yearbook*. The similarity models identify its similarity with the original two datasets on the base of available metadata. So using the third dataset we could say that the two original datasets are relevant to each other because they are both similar to the third one. In other words, they are *transitively similar* when using other datasets as a context. What is also interesting in the example is that metadata about the third dataset express explicitly the concept of transport. So, the third dataset is not just an intermediary dataset between the two. It explains why they are relevant, contributing thus to the discussion on *explainability* of similarity search.

2 Related Work

Before presenting the meta-model of data-transitive similarity, we discuss several related points.

2.1 Similarity Modeling

The research in the similarity search area had intensified some three decades ago by setting the metric space model as the golden standard [25]. The metric distances in place of (dis)similarity functions were introduced purely for database indexing reasons (i.e., for fast search). Though a good trade-off for many problems, the metric space model remains quite restrictive for modeling similarity. The restrictions are even more strict in follow-up models aiming at improving search efficiency, such as the ptolemaic [15] or supermetric [9] models. As mentioned in the previous section, this might not be a problem in case the descriptors are canonized and semantic (such as histograms referring to a vocabulary of deep features). However, for the lower-semantic cases there were alternative approaches to indexing similarity proposed in the past 15 years, ranging from dynamic combinations of multiple metrics [5] for multi-modal retrieval to completely unrestricted, non-metric approaches [23]. The rationale for their introduction was to increase the expressive power of similarity search (and effectiveness) and still provide an acceptable retrieval efficiency.

2.2 Retrieval Mechanisms

No matter if we choose metric or non-metric similarity, the expressive power of retrieval is also affected by the retrieval mechanism used. The query-by-example paradigm constitutes the basic functionality of similarity search in form of kNN or range queries. The similarity joins enable the use of similarity within the database JOIN operators [22]. The similarity queries could be also used with additional post-processing techniques for multi-modal retrieval and analytics, such as the late fusion [21] and content-based recommender systems [1]. Last but not least, there appear proposals and frameworks helping with the integration of similarity search constructs into query languages, such as SimilarQL [24],

or MSQL [19]. The ultimate goal is to establish higher-level declarative query models for similarity search [3].

2.3 Dataset Discovery

Finding related datasets, also known as dataset discovery, is one of the important tasks in data integration [20]. Large companies such as Google have developed their own dataset search techniques and solutions [4]. New solutions for dataset search in specific domains started to appear recently. For example, *Datamed* [8] is an open source discovery index for finding biomedical datasets. The existing works emphasize the role of quality metadata for dataset findability while [6] points out that available metadata does not always describe what is actually in a dataset and whether a described dataset fits for a given task. Other studies [12,13,16] confirm that dataset discovery is highly contextual depending on the current user's task. The studies show that this contextual dependency must be reflected by the dataset search engines. This makes the task of dataset discovery harder as it may not be sufficient to search for datasets only by classical keyword-based search. More sophisticated approaches being able to search for similar or related datasets could be helpful in these scenarios. As shown by [6,20] many existing dataset discovery solutions are based on simple keyword search. Discovery of datasets by similarity is discussed in the recent survey [6]. Several papers propose dataset retrieval techniques based on metadata similarity. In [2] a method is described which enables to measure similarity between datasets on the base of papers citing the datasets and a citation network between datasets. In [10] four different metadata-based models are evaluated for searching spatially related datasets, i.e., datasets which are related because of the same or similar spatial area covered. To the best of our knowledge, none of the approaches does apply the following technique of data-transitive similarity in dataset discovery.

3 Data-Transitive Similarity

In this section, we introduce the meta-model of data-transitive similarity. The original inspiration was the omnipresent database operation JOIN, used in many data management use cases for interconnecting relevant pieces of information. In relational databases the join operations allow to connect data records by means of shared attribute(s). In an extensive interpretation, the mechanism in database joins has roots in an identification of relevant entities by partial matches (equality predicate) or by partial similarity (inequality predicate). Analogously, by introducing data-transitive similarity we aim at consecutively joining similar objects and evaluating the overall relevance as an aggregation over the partial similarity scores.

The basic assumption of data-transitive similarity is thus a chain of objects from the database that are similar to each other, but the beginning and end of the chain could be quite dissimilar (yet relevant). Remember the well-known example with the human and the horse, illustrating the violation of the triangle

inequality [23]. These two creatures tend to be quite dissimilar, yet they can be relevant (transitively similar). The relevance here can be ensured by a connecting object in the middle of the chain – a horseman, or more poetically a centaur, creature that is half man and half horse. The data-transitive similarity itself, however, can be more complex; the connecting agent may not be a single object, but a whole chain of objects. This chain also serves as an explanation of why the two objects are relevant and in what context (addressing the explainability issue).

The connection itself can be formalized as an aggregation of several consecutive ground distances. The Equation 1 defines general form of data-transitive distance function $\hat{d}$, where $\mathcal{D}$ is a set of objects (the database in practical applications), d is a ground distance (the direct similarity), n is the length of the chain. Operator $\bigodot$ is an outer aggregation over all permutations of length n over elements of database $\mathcal{D}$ (e.g., min, max, avg). Operator $\biguplus$ is an inner aggregation over the individual direct distances within a particular chain. Table 1 shows examples of various inner aggregation functions. They are also the aggregation functions we worked with in our preliminary experiments. A more complex alternative may be a combination of several kinds of aggregations or distances.

$$\hat{d}_{\biguplus}^{\bigodot,n}(\mathbf{x},\mathbf{y}) = \bigodot_{(i_1,\ldots,i_n)\in\mathcal{D}^n} \biguplus \left(d(\mathbf{x},i_1), d(i_1,i_2), \ldots, d(i_n,\mathbf{y})\right) \tag{1}$$

$\mathrm{sum}(\delta_0,\delta_1,\ldots,\delta_n)$	$= \sum_{j=0}^{n} \delta_j$
$\min(\delta_0,\delta_1,\ldots,\delta_n)$	$= \min\{\delta_0,\delta_1,\ldots,\delta_n\}$
$\max(\delta_0,\delta_1,\ldots,\delta_n)$	$= \max\{\delta_0,\delta_1,\ldots,\delta_n\}$
$\mathrm{prod}(\delta_0,\delta_1,\ldots,\delta_n)$	$= \prod_{j=0}^{n} \delta_j$
$\mathrm{iprod}(\delta_0,\delta_1,\ldots,\delta_n)$	$= 1 - \prod_{j=0}^{n} (1-\delta_j)$

Table 1: Examples of inner aggregation $\biguplus$.

To summarize, we define the data-transitive similarity $\hat{d}$ as a meta-model operating on top of a ground similarity model d and a particular database $\mathcal{D}$. The computation of a single data-transitive distance involves a series of similarity queries over the database. The computational complexity of the data-transitive similarity thus involves not just the complexity of d but also the size of the database $|\mathcal{D}|$. Depending on the implementation, the worst-case time complexity $O(\hat{d})$ can vary from $O(d)$ to $O(d)O(|\mathcal{D}|^n)$, assuming n as a constant or $n \ll |\mathcal{D}|$.

From the definitions above it immediately follows that data-transitive distances are not metric distances – not only due to the possibly non-linear combination of the particular ground distances, but mainly due to the database-

dependent nature of the distance topology (non-uniform distribution of points in the data universe and its impact on the chain members).

One might say that such advanced relevance constructions should not be modeled at the level of similarity, as they are part of higher retrieval models closer to the application level (e.g., a part of content-based recommender system). However, we want to stress that we intentionally included the data-transitive similarity into the family of generic pair-wise non-metric similarities. As such, it can be plugged into any search engine that supports non-metric similarities. This would not be possible if designed as a proprietary late-fusion retrieval model.

3.1 Implementation

The fundamental problem we have addressed in the data-transitive similarity design was determining the number of intermediaries (the chain length n) to form a transitive similarity. Although our model assumes an arbitrary n, determining the specific value is not a straightforward problem itself. A significant issue may be that for some objects, there is no intermediary to form transitive similarity. In general, the number of intermediaries may not be constant, and for different objects this value needs to be chosen dynamically.

Thus, for our experiment, we have applied a simplification in this regard and assume that data-transitive similarity has at most one intermediary (i.e., $n = 1$). Therefore, we always have a triplet: a query, an intermediary, and a result. This decision reduces the number of hyperparameters with respect to longer chains (e.g., number of intermediaries, different aggregation functions). This approach also has the advantage of a higher level of explainability. For longer chains of intermediaries, we need to discuss whether each part of the sequence makes sense for given transitivity. Whereas in the case of a single intermediary, we can argue with a reasonable certainty whether the query and result are relevant from the perspective of the intermediary explanation.

The second problem is the transitivity involving duplicates or near-duplicates in the chain – intermediaries very d-close to the query or to the result. Such duplicate intermediaries usually do not add any value. Therefore, small distances d (the first 5% of distance distribution) are not considered (in fact, all such distances are set to infinity to become disqualified in $\hat{d}$).

Third, all ground distances are required to be normalized to $0 - 1$ because some aggregations ($\biguplus$ = prod, $\biguplus$ = iprod) require a bounded distance. In our implementation, we do not implement any optimizations, while to compute the data-transitive similarity we need to iterate over all database objects in the role of an intermediary. At the moment, optimizations for reduction of the set of intermediaries are beyond the subject of our research.

3.2 Open Dataset Testbed

For the open dataset testbed presented in Section 1.1, we considered title, description, and keywords metadata. Since the original data provided by the National Open Data Catalog are in the Czech language, we used the automatic

English translation [17], followed by the words lemmatization and filtering non-meaningful words (we consider only nouns, adjectives, verbs, and adverbs). In addition, we ignored several experimentally detected stop-words (data, dial, export, etc.). The metadata descriptors were represented in the bag of words model (BoW) with tf-idf weights.

Over these descriptors, the ground cosine distance was computed as $d_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - s_{\cos}(\mathbf{x}, \mathbf{y})$ (where $s_{\cos}$ is cosine similarity) for all pairs of objects (all pairs of datasets in our case). Figure 1 shows the distribution of distances $d_{\cos}$ over this testbed. We can see that most of the datasets are not $d_{\cos}$-similar, and the testbed exhibits high intrinsic dimensionality [7]. This is due to the relatively sparse metadata (average about 20 words). For some datasets, some parts, such as description or keywords are empty; there is only the title description.

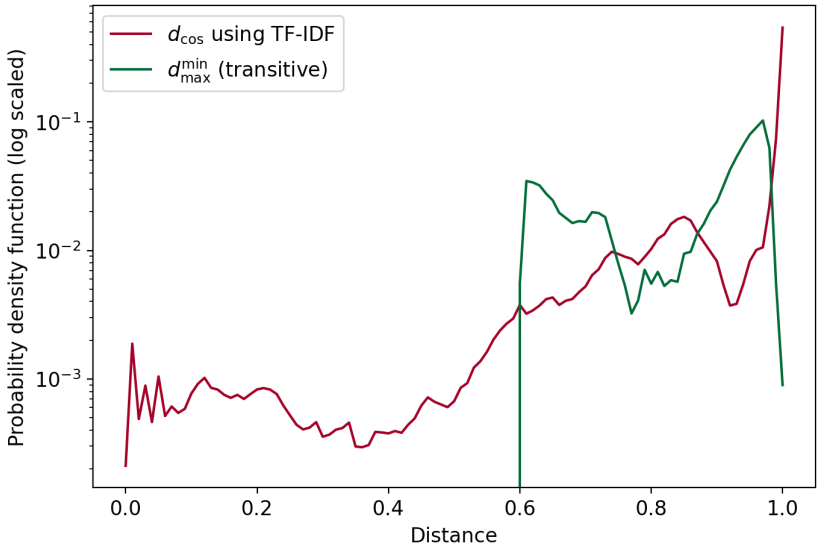

Fig. 1: Distance distribution of d_{cos} and $\hat{d}^{\min}_{\max}$ transitive similarity

In our experiment, we took only one intermediary, while $\hat{d}^{\min}_{\max}$ (Formula 2) was chosen as the data-transitive similarity function, since it exhibited the most robust aggregation in our preliminary experiments. Figure 1 shows how the distribution of $\hat{d}^{\min}_{\max}$-distances is different when compared to $d_{\cos}$. Smaller distances (below approx. 0.6) are eliminated due to the removal of near-duplicate dataset pairs (set to 5% closest datasets), as mentioned in the previous subsection. The rest of the $\hat{d}^{\min}_{\max}$-distance domain is split into two categories representing rel-

evance (more relevant around 0.7, less relevant around 0.9), with many $d_{\cos}$-dissimilar datasets moving into the category of more $\hat{d}^{\min}_{\max}$-relevant datasets.

$$\hat{d}^{\min}_{\max}(\mathbf{x}, \mathbf{y}) = \min_{\forall i \in \mathcal{D}} \max \{d(\mathbf{x}, i), d(i, \mathbf{y})\} \tag{2}$$

4 Evaluation

As we have already discussed in [26], the findability evaluation in the open dataset discovery is complicated from several points of view. The database contains a relatively large number of datasets, but there is no sufficient ground truth for dataset similarity. To overcome the lack of ground truth, in this paper we evaluate the concept of relevance which is closer to dataset discovery, rather than direct context-independent similarity of datasets.

4.1 Methodology

Our evaluation targets the additional value of data-transitive similarity search over the standard (direct d_{cos}) similarity search. First, the search for similar datasets using standard d_{cos}-similarity search is performed. Let us represent this search as a k_dNN query, where k_d is the number of results. Then, there are k_t results displayed to the user using data-transitive similarity based k_tNN query, while filtering out results of the previous k_dNN query. For our experiment, we assume $k_d = 100$ and $k_t = 20$.

The user (evaluator) is given a list of triplets (query, intermediary, result) and then evaluates each such triplet as relevant or non-relevant. A triplet is relevant if the user finds a possible use case for the query dataset and the result dataset and, at the same time, the intermediary dataset reasonably connects the two datasets. Let us repeat that the user is only confronted with results that were not findable by standard (direct) similarity search. A total of 5 users (evaluators) participated in the evaluation.

During the evaluation, we encountered the problem that some pairs of datasets are only relevant if we ignore specific fine-grained attributes of the datasets. The first observed attribute is the information about the publisher, e.g., contracts of the Ministry of Finance and invoices of the Ministry of Finance. The second attribute is the time or date of repeatedly published datasets, e.g., the list of companies for the year 2020. The third attribute is the localization specified in the datasets, e.g., hospitals in Prague vs. hospitals in Brno. For the evaluation, we decided to ignore these attributes as they only contribute to fragmentation of the datasets that are otherwise relevant to each other. However, this problem might disappear if we consider more than just one intermediary in the data-transitivity model (subject of future evaluations).

As part of the experiment, we evaluated the relevance of the results for a set of prepared queries. This set was created based on previous experiments presented in [26]. A total of 64 transitive results were found for 11 different queries.

4.2 Results

During the evaluation, we looked at two main criteria: consistency and effectiveness. For every triplet, we have computed its score as sum of 0 (non-relevant) and 1 (relevant) ratings of all evaluators. In our case, the score ranges from 0 (all evaluators claim the triplet is non-relevant) to 5 (all evaluators claim the triplet is relevant). Figure 2 (left) shows the number of triplets with particular score, Figure 2 (right) shows the number of triplets per data-transitive distance ranges and distribution of scores inside these ranges.

The consistency is validated based on the evaluators' agreement on the relevance of the evaluated triplets. Figure 2 (left) shows that in almost 78.13% of the cases, majority of evaluators (scores $0-1$ and $4-5$) agreed on the triplets' relevances. This observation confirms that the overall evaluation results are not just random noise.

Effectiveness is measured as the ratio of relevant datasets to all returned results. This gives us a measure of how much data-transitive similarity can improve the standard search. At Figure 2 (left), we see that in 57.81% of the cases, the triplet was marked as relevant by a majority of evaluators (score $4-5$).

Although the overall effectiveness may not seem significant, we must stress that all the relevant results found were not achievable by the direct similarity search (as already mentioned in Section 4.1). For 65.63% of the datasets, d_{cos} distances to query are maximal. We can also notice in Figure 2 (right) that the data-transitive similarity model complies with the general thesis of similarity search (more distant datasets are less relevant and vice versa).

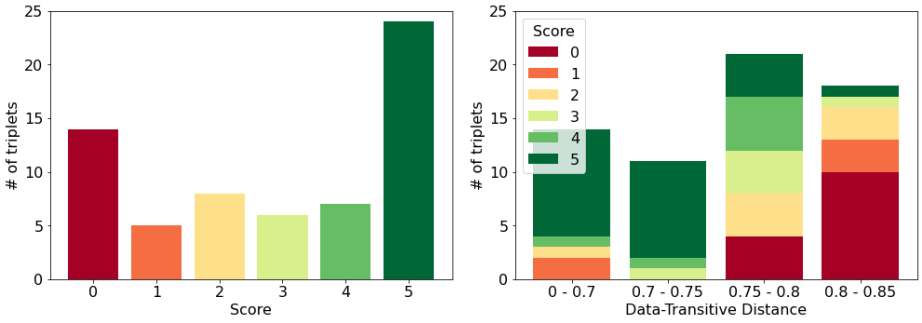

Fig. 2: The left figure shows the distribution of triplet ratings (how many triplets were rated by a particular relevancy score). For example, the score = 3 means that 3 evaluators thought the triplet was relevant (they rated it 1) and 2 evaluators thought the triplet was not relevant (they rated it 0). The right figure shows the distribution of ratings according to each data-transitive distance interval.

4.3 Qualitative Analysis

In Table 2 we see an example of triplet (Q, I, R) that was evaluated as relevant in our experiment (small data-transitive distance $\hat{d}^{\min}_{\max}(Q, R)$ through I). If we analyze the distance structure, the query dataset (Q) "Floods in the 19th century" does not have the "water" keyword in the metadata. However, thanks to the intermediary "5-year water" dataset (I), we have both "water" and "flood" in metadata and so the query dataset is transitively similar to the result dataset (R) "Water reservoirs". In the original similarity (the direct ground distance d_{cos}), the query Q and the result R datasets have maximum distance; they have nothing in common. In the data-transitive similarity search, however, the dataset R is within the first 20 results thanks to the connection with I. The relevance here can be explained by the fact that reservoirs can affect flooding and so the dataset R might be useful in flood prevention planning.

	Title **Keywords**	**Description**
Q	*Floods in the 19th century* **Floods, Environment, GIS**	Flooded areas in a 19th century flood in the Pilsen region.
I	*5-year water* **GIS, Floods, Environment**	Flooding areas of n-year water in the Pilsen region.
R	*Water reservoirs under the management of the river basin and the forest of the Czech Republic under the territorial jurisdiction of the river Vltava* **water tanks, water management**	The shp file contains points representing water reservoirs whose permitted volume of buoyant or accumulated water exceeds 1 000 000 m3 or to which the Forests of the Czech Republic, p. The registers are updated continuously, the dataset only once a year. The current data can be viewed on the water information portal VODA – www.voda.gov.cz.

Table 2: Example of Query, Intermediary, Result triplet: floods vs water. Title, keywords and description metadata are provided for each dataset.

The second example (Table 3) shows the imbalance of some descriptions, where the query dataset "Housing Young 2017" description has 3 paragraphs of text and the result dataset "BUG[3] - Economy and Labour Market" description has only one sentence. Although these datasets share some keywords, the resulting position in ranking is too far when using the direct distance d_{cos}, so that the user cannot find the dataset. With the data-transitive similarity using the intermediary "BUG - people and housing" dataset the problem is mitigated.

[3] BUG = Brno Urban Grid

	Title Keywords	Description
	Housing Young 2017	
Q	**sociology, housing research, housing young, housing, Brno**	The main objective of the Youth Housing survey conducted in 2017 was to identify and describe the housing needs of young people living in Brno, as well as their preferences in this area. . . . *3 paragraphs of text here* . . .
	BUG - people and housing	
I	**Brno urban Grid, housing, people, BUG**	Datasets from the Brno Urban Grid - theme people and housing
	BUG - Economy and Labour Market	
R	**BUG, labour market, economy, Brno Urban Grid**	Datasets from the Brno Urban Grid application - theme of economy and labour market

Table 3: Example of Query, Intermediary, Result triplet: housing vs labour. Title, keywords and description metadata are provided for each dataset.

In this case, we are able to explain the relevance between the housing of young people and the state of the labour market.

5 Conclusion and Future Work

We proposed an extended concept of similarity search by introducing the metamodel of data-transitive similarity operating on top of a particular similarity model. In the evaluation focused on the open data domain, we have demonstrated that the user is able to find relevant datasets that were not findable using standard (direct) similarity search. Moreover, as the data-transitive similarity is a variant of pair-wise non-metric similarity, it can be plugged into any search engine that supports non-metric similarities. It also confirms the necessity of non-metric approaches in complex retrieval tasks, such as the entity discovery.

In the future we plan to investigate more general chains of intermediaries, as well as internal indexing techniques for the data-transitive similarity computation itself. We also plan to experiment with other domains that require more complex explainable similarity approaches.

Acknowledgments

This work was supported by the Czech Science Foundation (GAČR), grant number 19-01641S.

References

1. Aggarwal, C.C.: Recommender Systems - The Textbook. Springer (2016)

2. Altaf, B., Akujuobi, U., Yu, L., Zhang, X.: Dataset recommendation via variational graph autoencoder. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 11–20 (2019). https://doi.org/10.1109/ICDM.2019.00011
3. Augsten, N.: A roadmap towards declarative similarity queries. In: Bohlen, M., Pichler, R., May, N., Rahm, E., Wu, S.H., Hose, K. (eds.) Advances in Database Technology - EDBT 2018. pp. 509–512. Advances in Database Technology - EDBT, OpenProceedings.org (Jan 2018). https://doi.org/10.5441/002/edbt.2018.59
4. Brickley, D., Burgess, M., Noy, N.F.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019. pp. 1365–1375. ACM (2019). https://doi.org/10.1145/3308558.3313685
5. Bustos, B., Kreft, S., Skopal, T.: Adapting metric indexes for searching in multi-metric spaces. Multimedia Tools and Applications **58**, 1–30 (06 2012). https://doi.org/10.1007/s11042-011-0731-3
6. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. VLDB J. **29**(1), 251–272 (2020). https://doi.org/10.1007/s00778-019-00564-x, https://doi.org/10.1007/s00778-019-00564-x
7. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Comput. Surv. **33**(3), 273–321 (Sep 2001). https://doi.org/10.1145/502807.502808, https://doi.org/10.1145/502807.502808
8. Chen, X., Gururaj, A.E., Ozyurt, B., Liu, R., Soysal, E., Cohen, T., Tiryaki, F., Li, Y., Zong, N., Jiang, M., Rogith, D., Salimi, M., Kim, H.e., Rocca-Serra, P., Gonzalez-Beltran, A., Farcas, C., Johnson, T., Margolis, R., Alter, G., Sansone, S.A., Fore, I.M., Ohno-Machado, L., Grethe, J.S., Xu, H.: DataMed – an open source discovery index for finding biomedical datasets. Journal of the American Medical Informatics Association **25**(3), 300–308 (01 2018). https://doi.org/10.1093/jamia/ocx121, https://doi.org/10.1093/jamia/ocx121
9. Connor, R., Vadicamo, L., Cardillo, F.A., Rabitti, F.: Supermetric search. Information Systems **80**, 108–123 (2019). https://doi.org/https://doi.org/10.1016/j.is.2018.01.002, https://www.sciencedirect.com/science/article/pii/S0306437917301588
10. Degbelo, A., Teka, B.B.: Spatial search strategies for open government data: A systematic comparison. CoRR **abs/1911.01097** (2019), https://arxiv.org/abs/1911.01097
11. Gkelios, S., Sophokleous, A., Plakias, S., Boutalis, Y., Chatzichristofis, S.A.: Deep convolutional features for image retrieval. Expert Systems with Applications **177**, 114940 (2021). https://doi.org/https://doi.org/10.1016/j.eswa.2021.114940, https://www.sciencedirect.com/science/article/pii/S095741742100381X
12. Gregory, K., Groth, P., Scharnhorst, A., Wyatt, S.: Lost or found? discovering data needed for research. Harvard Data Science Review **2**(2) (4 2020). https://doi.org/10.1162/99608f92.e38165eb, https://hdsr.mitpress.mit.edu/pub/gw3r97ht
13. Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A., Wyatt, S.: Understanding data search as a socio-technical practice. Journal of Information Science **46**(4), 459–475 (2020). https://doi.org/10.1177/0165551519837182, https://doi.org/10.1177/0165551519837182
14. Grosup, T., Peska, L., Skopal, T.: Towards augmented database schemes by discovery of latent visual attributes. In: Herschel, M., Galhardas, H., Reinwald, B.,

Fundulaki, I., Binnig, C., Kaoudi, Z. (eds.) Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019. pp. 670–673. OpenProceedings.org (2019). https://doi.org/10.5441/002/edbt.2019.83, https://doi.org/10.5441/002/edbt.2019.83

15. Hetland, M.L., Skopal, T., Lokoc, J., Beecks, C.: Ptolemaic access methods: Challenging the reign of the metric space model. Inf. Syst. **38**(7), 989–1006 (2013). https://doi.org/10.1016/j.is.2012.05.011, https://doi.org/10.1016/j.is.2012.05.011
16. Koesten, L.: A user centred perspective on structured data discovery. In: Companion Proceedings of the The Web Conference 2018. p. 849–853. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3184558.3186574
17. Košarko, O., Variš, D., Popel, M.: LINDAT translation service (2019), http://hdl.handle.net/11234/1-2922, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
18. Kučera, J., Chlapek, D., Nečaský, M.: Open government data catalogs: Current approaches and quality perspective. In: Kő, A., Leitner, C., Leitold, H., Prosser, A. (eds.) Technology-Enabled Innovation for Democracy, Government and Governance. pp. 152–166. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
19. Lu, W., Hou, J., Yan, Y., Zhang, M., Du, X., Moscibroda, T.: Msql: efficient similarity search in metric spaces using sql. The VLDB Journal pp. 3–26 (October 2017), https://www.microsoft.com/en-us/research/publication/msql-efficient-similarity-search-metric-spaces-using-sql/
20. Miller, R.J., Nargesian, F., Zhu, E., Christodoulakis, C., Pu, K.Q., Andritsos, P.: Making open data transparent: Data discovery on open data. IEEE Data Eng. Bull. **41**(2), 59–70 (2018), http://sites.computer.org/debull/A18june/p59.pdf
21. Novak, D., Zezula, P., Budikova, P., Batko, M.: Inherent fusion: Towards scalable multi-modal similarity search. J. Database Manage. **27**(4), 1–23 (Oct 2016). https://doi.org/10.4018/JDM.2016100101, https://doi.org/10.4018/JDM.2016100101
22. Silva, Y.N., Pearson, S.S., Chon, J., Roberts, R.: Similarity joins: Their implementation and interactions with other database operators. Information Systems **52**, 149–162 (2015). https://doi.org/10.1016/j.is.2015.01.008, special Issue on Selected Papers from SISAP 2013
23. Skopal, T., Bustos, B.: On nonmetric similarity search problems in complex domains. ACM Comput. Surv. **43**(4) (Oct 2011). https://doi.org/10.1145/1978802.1978813
24. Traina, C., Moriyama, A., da Rocha, G.M., Cordeiro, R.L.F., de Aguiar Ciferri, C.D., Traina, A.J.M.: The similarql framework: similarity queries in plain SQL. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, Limassol, Cyprus, April 8-12, 2019. pp. 468–471. ACM (2019). https://doi.org/10.1145/3297280.3299736
25. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach, Advances in Database Systems, vol. 32. Springer (2006)
26. Škoda, P., Bernhauer, D., Nečaský, M., Klímek, J., Skopal, T.: Evaluation Framework for Search Methods Focused on Dataset Findability in Open Data Catalogs. In: Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services. pp. 200–209 (2020)