Faculty of Informatics
Masaryk University
Brno

Review of Doctoral Thesis Titled

„Indexing Arbitrary Similarity Models"

The thesis by Tomáš Bartoš is focused on issues of analyzing data in order to state principles used to build indexing structures. In particular, the author proposed general algorithms to discover expressions that lower bound real distance values computed between pairs of data objects.

The thesis is logically structured to nine chapters and guides the reader from fundamentals of similarity searching via a short survey of existing related indexing techniques to the algorithms and their implementations that constitute the SIMDEX framework.

In Chap. 2, I miss a note on the semantic gap phenomenon, which would complete the overview of issues we have to cope with in data management. I also miss a list of important basic terms similar to the one given in Table 4.1, part II, e.g. metric space, object, pivot/reference object.

In Sec. 3.2.8, the author presents techniques that are "alternatives" to SAMs and MAMs. However, some of them are widely applicable in all access methods, e.g. approximate search or so-called early abandoning.

Chap. 4 gives definitions of important terms. In the context of thesis, I do not agree that a distance matrix defines a similarity model. It is the other way around – the distance matrix is computed from the given dataset and distance function. It is not clear whether the selection process of k sample objects may contain the same object repeatedly or not, i.e. whether $C(|S|, k)$ or $V(k, |S|)$ should have been used in the denominator in Def. 4.8. Later in the text (p. 67), the author refers to the combinations. In Sec. 4.4.1, the author states the requirement of having a lower bound evaluated faster than the real distance. It cannot be achieved because the presented grammar does not include any reference to the distance matrix, i.e. a list of precomputed distances, which is crucial here.

The technique of fingerprinting sketched in Sec. 5.1.2 is important for elimination of many duplicate branches of computation, but it has not been given clearly. The example containing $\delta(p_3,p_4)$ and $\delta(p_6,p_5)$ is trivial, since it is a distance evaluation in both cases. Is this the only case handled by this technique?

The description of experimental results is the weakest part of this in overall very good thesis. The proposed algorithms identified some lower bounding expressions but the author does not try to interpret them nor list the most popular "forms". Do they exist? There is an attempt for expression interpretation in Sec. 8.3.1. Next, the datasets used within Chapters 5, 6, 7 and 8 are not consistent. In some figures, it is not clear what data size was used or the measures are not the same (speed-up vs. DC savings), so it makes comparison very difficult.

The experiments focus on comparison between seq. scan and a SIMDEX instance on a preselected distance function. Why did you select them? Is there any rationale behind? Otherwise the results would render to be of theoretical nature only.

Chapter 8 presents a LAESA-based approach to indexing and the results on a real-life data with real distance functions. The outcome is that lower bounds obtained from SIMDEX are capable of improving filtering during query evaluation. The issue here is that small data sizes were used (up to 900 objects). Why did not you use much larger data here?

In Conclusions, you state that PGP-SIMDEX requires a very complex implementation. However, I consider the opposite, since programming MapReduce jobs is quite easy.

The levels of English is very good and the text is comprehensive. However I would suggest using shorter sentences and past tense in description of experimental trials. There are few grammar mistakes only. Sometimes the author presents his own work as it has not been done by him, e.g. Lambda Tuning Algorithm, page 50. From the style point of view, I would not overuse back references to sections, e.g. "(see Section 3.2.7)" on page 51 refers to itself.

Further Remarks / Questions

Features extracted from original (raw) data objects may not always be real-valued vectors and of a fixed number of dimensions. I would not limit the feature space to be R^n only (page 13).

In Sec. 2.2.2, the author lists some examples of non-metric functions. I would suggest to classify them into pseudo-, quasi- and semi-metrics.

Classifying sequential scan algorithm as a metric access method is inappropriate. It is of general purpose and can be applied to any data processing operation.

Fig. 3.13 should depict precision as is stated on page 52. However it depicts average error.

There are some results missing in Figs. 3.12 and 3.13 and it has not been explained why.

Inconsistencies in use of symbols, e.g. "C" vs. "c", "N" vs. "n".

In Fig. 5.1, the individual modules should have been denoted with S1,...,S6 for better comprehension.

In Sec. 5.3, you list datasets and the corresponding distance functions used in experiments. But you did not specify the partial distance of Hausdorff distance. Moreover, I assume you gave wrong distance function for the Corel data.

Table 5.1 gives results of I-SIMDEX evaluation, but there is no key given for selecting the presented expressions. How did you select them amongst 25,000 expressions?

The *PolygonSet* dataset is a good example for manual verification of algorithm results. On page 91, you claim that expr. #5 improves fitness value, but I do consider it as a weaker variant of triangle inequality. In case of expr. #10, it is more or less identical to the triangle ineq. The example of expr. 47 on page 92 is a nice variant of a constant lower bound. Did you try to define constrains to rule out such trivial solutions? E.g. sth. similar to what you discussed later (in Sec. 8.3.1?).

Figs. 6.9 and 6.10 are very hard to compare since different measures were taken.

Tab. 7.1 shows "triangle error" for various distance functions. The errors for metrics L_2 and L_{max} are non-zero. Why?

There are some references that are not very appropriate, e.g. BigData [4,5] (page 7), pivot table [3] (page 92).

Conclusion

The contributions of this thesis were published in seven papers – four SISAP conferences, one EDBT/ICDT conference, one GECCO conference and a vision paper in SIGMOD record. These publication are of very high quality, especially the non-SISAP once. This definitely shows the author's ability to conduct independent research. This area is quite technical and narrow-focused to a particular problem of indexing data, but it is topical with almost direct implications to real applications and their efficiency. The SIMDEX framework is by no means new and contributes to the area of similarity searching with new results. My comments are rather marginal and definitely do not require to have this thesis rewritten. The text is comprehensive. I propose to have this thesis accepted as a doctoral thesis.

July 25, 2014, Brno

doc. RNDr. Vlastislav Dohnal, Ph.D.

Faculty of Informatics
Masaryk University

Tel. +420, 549 49 3360
Email: dohnal@fi.muni.cz