CHARLES UNIVERSITY PRAGUE

faculty of mathematics and physics

To:

The Board of Doctoral Study
Faculty of Mathematics and Physics
Charles University in Prague

July 30th, 2014

Report on the PhD thesis by Tomáš Bartoš

The research of similarity search in large databases underwent a long way towards efficient management. The advances were mostly based on leveraging the metric space model of similarity. However, in terms of effectiveness, the metric space model is way too restrictive for domain experts and practitioners. Hence, it is of immense importance to research and develop database methods for non-metric (i.e., arbitrary) similarity models. The topic of this doctoral thesis is focused on indexing arbitrary similarity models, not assuming any topological properties such as the metric space axioms.

When starting this topic, the task for the candidate was extremely hard. It stated: “Let’s develop a framework that mines analytical distance lower-bounding inequalities that comply with the underlying similarity model and, at the same time, provide fast similarity search in that model”. The general description of the framework (called SIMDEX) is given in Chapter 4. The first attempt to the solution was a grammar-driven approach, where a set of lower-bounding forms was derived using a specific grammar as the Iterative SIMDEX (Chapter 5). While this method was systematic and deterministic, it turned out it couldn’t reach a practical usability due to the extreme time complexity. Therefore, a completely different implementation of SIMDEX was proposed, called the GP-SIMDEX that generates the axiom set non-deterministically using genetic programming (Chapter 6). A parallel version was also proposed (Chapter 7), designed for multi-core as well as massively parallel platforms (MapReduce). Finally, an access method called Smart Pivot Table was proposed (Chapter 8), incorporating the GP-SIMDEX into the indexing routines. Moreover, a hybrid strategy Triangle+ was introduced, enhancing SIMDEX with standard metric space filtering (either natively or by modification using the TriGen algorithm). This ultimate combination finally proved the practical indexing abilities of the explored lower-bounding axioms, hence, the thesis task was completed.

The thesis demonstrates a comprehensive insight of the candidate into the problem, while the proposed contributions to the research area are significant and original. The candidate also proved his ability to independently recognize, formulate and develop novel approaches to fast similarity search. Besides the main contribution of the thesis, the Chapters 1-3 are valuable to an IT reader for a quick introduction into (non-)metric similarity search.

The results presented in this thesis have been published in proceedings on a number of representative international conferences (4xSISAP, EDBT, GECCO, VLDB PhD workshop) published by ACM/Springer, and in SIGMOD Record journal. In addition to these, the candidate was a co-author of several other conference publications (ADC, ISD, NDT).

Based on the evaluation above, I **recommend** the candidate Tomáš Bartoš to obtain the PhD degree.

Doc. RNDr. Tomáš Skopal, Ph.D.
supervisor