Masarykova univerzita
Fakulta informatiky
Datum: 12 -12- 2019
Č.j./E.č.: MU-15/158249/2019/
Počet listů dokumentu: 99/890/F1-4
Počet příloh a listů/sv.: /
Počet a druh nelist. příloh: /

Review of dissertation thesis

PhD candidate:	RNDr. Vladimír Míč
Thesis topic:	Binary Sketches for Similarity Search
Reviewer:	prof. RNDr. Tomáš Skopal, PhD.
Institution:	Charles university, Faculty of Mathematics and Physics

Thesis structure:

The thesis format is a commentary to eight research papers, seven published in conference proceedings and one in a journal, included as appendices. The first section is an introduction, the second and third briefly overview the problem of similarity search and in fourth section the related work and background is discussed. The core of the commentary is the fifth section, thesis contribution, that in short comments on the results achieved and described in detail in research papers (included as appendices A-H). In subsection 5.1 the sketching technique is defined and various properties are discussed (sketch length, idim, bit correlations), referred to in detail in papers A, B, C, G. Subsection 5.2 discusses the filtering with sketches (papers D, E), subsection 5.3 is dealing with selection of sketching technique (paper F), and finally subsection 5.4 is addressed to the indexing of sketches using weights (paper G). Section 6 concludes the paper. Besides the cited papers (appendices A-G), the appendix H is a paper not related to binary sketches that was included just as a motivation for efficient and effective similarity search.

The results of the work:

The thesis deals with efficient approximate similarity search using binary sketches in Hamming space. The author investigates theoretical questions concerning the binary sketch construction and usage as well as develops practical retrieval methods and experiments to confirm the theoretical analysis. The results of this thesis were published in proceedings of good international conferences, and especially in the ACM TOIS journal (28 pages paper, IF 2.3).

The results achieved in the thesis (in the papers, actually) significantly contribute to the state of the art in the domain of similarity search. The particular results form a coherent piece of work focused on different aspects of the binary sketches (with Hamming distance) under the application of approximate similarity search. Especially the theoretical findings are at very good level.

Language and graphics level:

The language level is very good, as well as the formating, figures, and graphics level in general.

Text comments:

The discussion in section 5.1.2 + figure 5.3 is somewhat misleading. I understand this section abstracts from pivots and indexing, however, then it's bad to interpret the facts. The parameter is not the sketch length, but bit correlation. I would expect that natural parameter is the sketch length (also the section title suggests so) where the correlation is an observed statistics. I miss here the related problem of pivots selection. Simply said, beyond a certain sketch length, i.e., a certain number of pivots (or pivot pairs for GHT), the bit correlation is no longer changing much and therefore adding new pivots only leads to redundancy. Is that true?

Defense questions:

1) How are binary sketches related to permutation indexes? Both use "compressed" approximate representations, but the coding is different. However, also permutations are coded as binary strings related to pivot order.
2) The motivation for binary sketches was a more compact descriptor representation and cheap distance, i.e., real-valued pivot space vs. binary Hamming space. On the other hand, isn't the relaxing to unbalanced and longer sketches going against the original intent, where the lengths of binary sketches approach the bit lengths of floats?
3) Unlike pivot tables or permutation indexes, where the cheap lower bounds are utilized in the filter phase, the author suggests to use the binary sketches together with another metric index within its refine phase. I understand this as an evidence, that binary sketches exhibit properties for high precision but low recall. Could you comment on this?

Conclusions:

The author in his dissertation thesis has demonstrated the ability to work independently and creatively in the specified field and the thesis meets the standard requirements imposed on dissertation thesis in the specified field. I recommend to **accept** the thesis and grade it as **A „excellent"**.

Prague, December 4, 2019

prof. RNDr. Tomáš Skopal, Ph.D.