جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Karima ECHIHABI

soutiendra publiquement sa thèse de Doctorat en Informatique

le Mardi 28 Juillet 2020 à 15h00 au Grand amphi à l'ENSIAS

Intitulé de la thèse

SCALABLE AND ACCURATE HIGH-DIMENSIONAL SIMILARITY SEARCH: FROM DATA SERIES TO DEEP NETWORK EMBEDDINGS

Devant le Jury composé de :

Président :

M. Bouchaib BOUNABAT, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Mme Houda BENBRAHIM, PH, ENSIAS, Université Mohammed V de Rabat

Co-encadrant de thèse :

M. Themis PALPANAS, Professeur, Université de Paris

Rapporteurs :

Mme Zohra BAKKOURY, PES, EMI, Université Mohammed V de Rabat

M. Mohammed RAMDANI, PES, FSTM, Université Hassan II, Casablanca

M. Mohamed LAZAAR, PH, ENSIAS, Université Mohammed V de Rabat

Examinateur :

M. Mostapha ZBAKH, PES, ENSIAS, Université Mohammed V de Rabat

SCALABLE AND ACCURATE HIGH-DIMENSIONAL SIMILARITY SEARCH: FROM DATA SERIES TO DEEP NETWORK EMBEDDINGS

Abstract: The world is drowning in a big data tsunami of high-dimensional objects that need to be analyzed in order to identify useful patterns and extract new knowledge in domains as varied as agriculture, medicine, cybersecurity, seismology, astrophysics, manufacturing, and finance, and others. In response to these needs, it is imperative to build analytical systems that truly support interactive exploration on datasets containing terabytes of high-dimensional objects, with dimensions reaching hundreds to thousands.

A fundamental and challenging operation called similarity search is the main bottleneck of many critical data processing tasks such as data cleaning, data integration and big data analytics (e.g., outlier detection, frequent pattern mining, clustering, and classification). A number of exact and approximate approaches have been proposed in the literature to support similarity search over massive data series collections.

In this thesis, we unify and formally define the terminology used for the different flavors of the similarity search problem. We present a similarity search taxonomy that classifies methods based on the quality guarantees they provide for the search results, and that unifies the varied nomenclature used in the literature. Following this taxonomy, we include a survey of similarity search approaches supporting exact and approximate search, bringing together works from the data series and multidimensional data research communities. We propose extensions to existing data series indexes that can answer approximate

queries with guarantees and that outperform popular state-of-the-art techniques such as LSH, kNN graphs and quantization-based inverted indexes in many scenarios. We also design and conduct the two most exhaustive experimental evaluations in the field cover- ing both exact and approximate techniques. Building upon the deep insights gained from both studies, we propose Hercules, a new

algorithm that outperforms the state-of-the-art similarity search approaches in-memory and on-disk.

Our work has far-reaching fundamental and practical implications. We demonstrate that it is possible to design efficient high-dimensional vector similarity search algorithms with theoretical guarantees on the quality of the answers, and we thus offer a more promising alternative to the two current trends in the literature: (i) LSH-based algorithms that support guarantees, but are relatively slow, and (ii) kNN graphs and inverted indexes, which are relatively fast, but do not provide theoretical guarantees. This finding paves the way for very exciting new developments which will lead to efficient solutions that can support critical analytical tasks such as brain seizure detection, cyber-attack prevention, transportation management and data cleaning automation.