Preface

With the growing variety of entities that have their presence on the web, retrieving relevant entities for various user requirements becomes an important problem. The area of Similarity Search that addresses this problem has received a lot of attention in the last fifteen years. Increasingly sophisticated data representations, query specifications, indexing mechanisms and algorithms to retrieve relevant entities to a query are being devised. Of these, developing indexes tailored to new kinds of data and devising algorithms to use such indexes to reduce the turnaround time for similarity search has attracted attention from the database community, resulting in several focused surveys and a few books that educate the audience about the field. Though relatively less discussed, another dimension in retrieval that has recorded tremendous progress over the years has been the development of mechanisms to enhance expressivity in specifying information needs. Similarity operators seek to advance the utility of similarity search systems from the user side by allowing the user to express her needs better by providing a richer set of querying options. In this book, we focus on the vocabulary of similarity operators that has grown vastly from just a set of two operators, top-k and skyline search, as it stood in the early 2000s. Today, there are ways to express complicated needs such as finding the top-k customers for a product wherein the customers are to be sorted based on the rank of the chosen product in their preference list. Some representative operators that have been proposed recently include K-N-Match, Reverse Furthest Neighbor, KN Diverse Neighbors, and Reverse kNN/Skyline operators. Arguably due to the complexity in the specification of new operators such as the above, uptake of such similarity operators has been low even though emergence of complex entities such as social media profiles warrants significant expansion in querying expressivity. To address this gap, we systematically survey the set of similarity operators, primarily focusing on their semantics, while also touching upon mechanisms to process them effectively. The aims of this book are to cover the following:

- A gentle introduction to the field of similarity operators starting from the fundamentals of similarity search systems.

- A comprehensive survey of the various similarity operators that have been proposed so far, in a structured manner to allow for easy assimilation.
- Positioning of the state-of-the-art in similarity operators with respect to the variety and complexity of entities that similarity search systems of today deal with, highlighting new directions and potential research gaps.
- A high-level overview of the indexing techniques and algorithms used for various types of data and similarity operators.

In this book, we expect to cover most of the important research advances in the area of similarity operators over the last fifteen years. To the best of our knowledge, this would be the first book focusing on the area of similarity operators.

The main emphasis of the book would be on providing a detailed tutorial on the area of operators for similarity search. The book will start off by providing introductory material on similarity search systems, highlighting the central role of similarity operators in such systems. This will include the insights gained from psychology and cognitive research that sheds light on the way the brain processes similarities and sets the stage for defining appropriate similarity measures and operators. This will be followed by a systematic categorized overview of the variety of similarity operators that have been proposed in literature over the last two decades. Indexing is a core technology to aid practical implementation of similarity operators; we will introduce and describe some of the indexing mechanisms that have been proposed in literature. Lastly, we will outline the research challenges in this area, so as to enable the interested researcher to identify potential directions of exploration. In summary, this book would provide a comprehensive overview of the field of similarity search operators, and cover the entire spectrum of technical issues related to the area.

We expect that this book would be useful for people across a wide variety of profiles such as students, educators and researchers. For students, we expect this would provide enough background to undertake research and implementation projects. In particular, for students who would like to build systems to illustrate the applicability of specific operator in the context of a specific domain or application, this book would provide a self-contained reference material. For educators who design and offer advanced graduate level courses on similarity search and recommender systems and would like to incorporate a segment on similarity operators, our book would be useful reference material to use as a platform for teaching and also to suggest as reading material for students. Researchers who are interested in advancing the state-of-the-art in similarity search would find this book useful to get up to speed to start working in the area. We have also included potential research directions for advancing this area keeping a research audience in mind.

We intend to start descriptions from the ground-up in a manner that would make the book accessible to undergraduate students with some familiarity with similarity search systems. Since similarity search systems and recommender systems are used by most people on a day to day basis, be it in the context of product search or social media browsing, we expect that most people with an interest in computer science would be able to understand this book. The chapter of indexing, however, expects the reader to have some background in information management; a person who has

taken an undergraduate level course in databases should be able to easily grasp the contents of that chapter.

We hope you do enjoy and benefit from this book, and look forward to receiving any suggestions or comments that you might have at our respective email addresses.

Usage in Courses: A detailed treatment of similarity operators covering all the content in this book would form a full *one-semester course*. Similarity Search, despite being a fairly active field in itself, is intimately related to various other disciplines such as recommender systems, information retrieval and case-based reasoning. Thus, selected parts of this book may be used as segments within courses focused on any of the above areas. We now outline some possible segments that could be carved out of this book:

- **Psychological Notion of Similarity:** Section 1.1 can be used as an introduction to the mind's notion of similarity, with a focus on those aspects that would potentially matter to similarity search systems.
- **Introduction to Similarity Search Operators:** Contents across Chapters 2 and 3 form an introductory segment to similarity operators.
- **Understanding Similarity Search Operators:** A segment covering Chapters 2, 3 and 4 can be covered to provide a reasonably detailed overview of similarity search operators to learners.
- **Deep-dive into Similarity Search Operators:** The meat of the book, comprising Chapters 2, 3, 4 and 5 form a comprehensive overview of the discipline of similarity search operators.
- **Indexing for Similarity Search:** Chapters 2, 3 and 6 contain material suitable for an introductory overview of similarity search with a focus on indexing.

Relevant sections of Chapter 7 may be added to each of the above segments for a graduate course to provide learners with a flavor for the kinds of research issues in similarity search systems that they may consider exploring in a course project or a larger research effort.

Online Resources: The website at *https://sites.google.com/site/ofssbook/* will host supporting content for this book such as powerpoint slides, and web links for additional reading.

Acknowledgements: The authors would like to acknowledge the support that they received from their friends and colleagues in the course of the planning and preparation of the material for this book. Deepak takes this opportunity to acknowledge the support and patience from his wife, Amrutha Jyothi, especially since most of the effort towards authoring this book was concentrated on the evenings and weekends. Prasad would like to express special thanks to his family, specifically his wife and kids for giving him the liberty to spend a signficant amout of family time on this project.

Bangalore,
May 2015

Deepak P
Prasad M. Deshpande

http://www.springer.com/978-3-319-21256-2

Operators for Similarity Search
Semantics, Techniques and Usage Scenarios
P, D.; Deshpande, P.M.
2015, XI, 115 p. 44 illus., Softcover
ISBN: 978-3-319-21256-2