Studies in Big Data 104

Sheetal S. Sonawane
Parikshit N. Mahalle
Archana S. Ghotkar

Information Retrieval and Natural Language Processing

A Graph Theory Approach

Springer

Studies in Big Data

Volume 104

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are reviewed in a single blind peer review process.

Indexed by SCOPUS, EI Compendex, SCIMAGO and zbMATH.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at https://link.springer.com/bookseries/11970

Sheetal S. Sonawane · Parikshit N. Mahalle ·
Archana S. Ghotkar

Information Retrieval and Natural Language Processing

A Graph Theory Approach

Springer

Sheetal S. Sonawane
Department of Computer Engineering
SCTR's Pune Institute of Computer Technology
Pune, India

Archana S. Ghotkar
Department of Computer Engineering
SCTR's Pune Institute of Computer Technology
Pune, India

Parikshit N. Mahalle
Department of Artificial Intelligence and Data Science
Bansilal Ramnath Agarwal Charitable Trust's
Vishwakarma Institute of Information Technology
Pune, India

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-981-16-9994-8 ISBN 978-981-16-9995-5 (eBook)
https://doi.org/10.1007/978-981-16-9995-5

Preface

Calmness, gentleness, silence, self-restraint, and purity: these are the disciplines of the mind

—Bhagvad Gita

This book envisioned to present the precise and summarized contents that put more highlights on a mathematical view using graph theory for natural language processing and information retrieval methods and approaches and how graph theory can make better insights from these use cases. Since last decade, there is much advancement in utilizing graph properties in NLP and IR in preprocessing steps. In addition to this, the traditional BoW approach was found to be not sufficient for major text analytics task. The order and co-occurrence relation are achieved easily with graph. Also, major advanced graph types like hypergraph and semigraph are found to be applicable in various IR tasks. The state-of-the-art methods and techniques provide improvement in the performance of the system.

As the data increasing exponentially, unblocking true potential of this growing data is becoming a real challenge. Due to advancement in semiconductor industry, the space is getting cheaper day by day, and in the sequel storage, extraction, transformation and loading of the data are becoming easier. However, the major challenge faced by all IT leaders today is the information retrieval and making sense out of this big data by drawing meaningful insights. The main focus of this book is to present and discuss information retrieval using graph theory and different approaches. This book also presents the role of natural language processing in informational retrieval using appropriate case studies.

This book gives a comprehensive view of graph theory in IR and NLP. This book provides a number of techniques and examples from state of the art. The book provides understanding of graph theory basics, graph algorithms and networks using graph. The book is divided into three parts and contains nine chapters. The first part gives graph theory basics and graph networks, and the second part provides graph-based information retrieval. The third part covers IR recent applications using graph. This book provides a strong foundation to a beginner. All the technical details that

include tools and technologies used for graph algorithms and implementation are explained in a clear and organized format.

Part I discusses graph theory basics, graph algorithms and network using graph. It lays the groundwork for the remainder of the book by introducing all necessary and important concepts in graph theory, including the notation, graph types, graph properties and graph representations. An overview of wide variety of graph methods, including shortest path algorithm, depth first search and breadth first search, maxflow network, graph clustering, graph coloring and more, is explained with examples and detail illustration. The term/concept network is a naturally occurring relations. This shows its adequacy to model power-law degree distribution. Zips and Heaps law are also explained with code. Further, social network measures like centrality, degree distributions, clustering coefficients are explained using examples.

Part II provides the basic of information retrieval, text document preprocessing using graph theory and graph theory in information retrieval. It gives insight of various steps included in information retrieval starting from preprocessing, indexing and then ranking. In text document, there are different components like words, phrases, sentences and paragraphs. These text components are connected to each other through various relationships which help to find meaning and relatedness in the document. The relationship among these components may be based on their occurrences in the document or similarity based on their context or concept. The graph representation of text along with graph operations for necessary preprocessing algorithms is explained. The graph characteristics, properties and operations useful to solve document preprocessing task is well explained in this part. The task of getting relevant document based on input query is important in information retrieval. The concept of term weighting in graph with generating extractive and abstractive summary by using graph operations is well explained with different graph types and properties. The applications like question answer system, sentiment analysis and recommendation system are also explored in this part. The connection of data elements is provided semantically using knowledge graph. The importance and its applications are discussed in this chapter. The detailing of its implementation using ontology, RDF, SPARQL and Neo4j is introduced in this part. The open source libraries, datasets and tools are provided in the last part of the book.

Part III discusses emerging applications and development in NLP and IR using graph theory. The emerging applications like multimodal summarization and summarization using knowledge graph with graph neural networks are explained in this part. The advanced graph types used in solving the problem of summarization and question answering system are explored. The need and motivation and applications of graph-based neural networks are explained in this part. The open source libraries, datasets and tools are provided in the last part of the book. This part included summary and future scope of the work. It further provides outlook of graph theory in IR, thereby laying a foundation for further research for the readers.

The main characteristics of this book are:

- A concise and summarized description of all the topics.
- Use case and scenarios based descriptions.
- Discusses graph methods that play a prominent role in IR and NLP applications.
- Presents knowledge graph methods and techniques
- Overview of NLP and IR tools and datasets.
- Numerous examples, technical descriptions and real-world scenarios.
- Simple and easy language so that it can be useful to a wide range of stakeholders like a layman to educate users, villages to metros and national to global levels.

NLP and IR are now fundamental course to UG and PG courses in computer engineering, IT, CS and E&TC. This book is useful to provide research opportunities and project development in NLP and IR. This book is also useful to a wider range of researchers and design engineers who are concerned with exploring information retrieval for engineering use cases. Essentially, this book is most useful to all entrepreneurs who are interested to start their start-ups in the field of application of IR to emerging interdisciplinary use cases as well as related product development. The book is useful for undergraduates, postgraduates, industry, researchers and research scholars in ICT, and we are sure that this book will be well-received by all stakeholders.

Pune, India

Dr. Sheetal S. Sonawane
Dr. Parikshit N. Mahalle
Dr. Archana S. Ghotkar

Acknowledgements

We would like to thank many people who encouraged and helped us in various ways throughout this book, namely our colleagues and friends. Special thanks to our family for their support and care. We are very thankful to our students Athang, Saumitra, Aamir, Vaishali and Aakash for their help during writing this book.

We are thankful to Honourable Trustee of SCTR, Mr. R. S. Kothavale, Secretary Mr. Swastik Sirsikar, Dr. P. T. Kulkarni, Director PICT and Principal Dr. (Mrs.) R. Sreemathy and HoD Dr. (Mrs.) Mukta Takalikar for their constant support and continuous encouragement.

We are thankful to BRACT's management, Shri. Bharat Rajkumar Agarwal, Managing Trustee, Dr. Vivek S. Deshpande, Director VIIT, all deans and heads of VIIT for their constant support and continuous encouragement.

We are also very much thankful to all our department colleagues at PICT, VIIT for their continued support, help and keeping us smiling all the time. We are very much thankful to our family members for their love, support and patience during writing of this book.

Last but not least, our acknowledgements would remain incomplete if we do not thank the team of Springer Nature who supported us throughout the development of this book. It has been a pleasure to work with Springer Nature team, and we extend our special thanks to the entire team involved in the publication of this book.

Sheetal S. Sonawane
Parikshit N. Mahalle
Archana S. Ghotkar

Contents

About the Authors

Dr. Sheetal S. Sonawane is an associate professor in the Department of Computer Engineering at Pune institute of Computer Technology (Pune). She received Ph.D. in Computer Engineering from college of Engineering at Savitribai Phule Pune University in 2018. She received her bachelor's degree in computer engineering from Pune University in 2000 and 4th rank holder in Pune University for master's degree in computer engineering in year 2006. She has more than 19 years of teaching and research experience. She has published widely in international journals like Springer, Inderscience and conferences like IEEE, Springer and having more than 615 citations. She is a recipient of Best Paper Award for her IEEE Conference paper. She has written book chapters in books like *Big Data Analytics* by PHI Publication and *Semigraph and Their Applications* by academy of discrete mathematics and applications, India. She is an author of book *Foundations of Data Science Based Healthcare Internet of Things* by Springer. She is reviewer of conferences and journals like IEEE, Elsevier, MDPI Journal, Inderscience, etc. She has published three Indian Patents out of which two are granted. She has delivered talks on national level. She has also remained a technical program committee member and session chair for international conferences. She is involved as coordinator for Data Science Honors course at SPPU and syllabus coordinator for many subjects at University level. Her research interest is in the field of Information retrieval, Natural Language processing and Data Mining. She has focused in the last few years on the research issued in machine learning, handling big unstructured data and graph model for representing and analyzing text documents.

Dr. Parikshit N. Mahalle obtained his B.E. degree in Computer Science and Engineering from Sant Gadge Baba Amravati University, Amravati, India, and M.E. degree in Computer Engineering from Savitribai Phule Pune University, Pune, India. He completed his Ph.D. in Computer Science and Engineering specialization in Wireless Communication from Aalborg University, Aalborg, Denmark. He was a postdoctoral researcher at CMI, Aalborg University, Copenhagen, Denmark. Currently, he worked as a professor and the head in the Department of Computer Engineering at STES's Smt. Kashibai Navale College of Engineering, Pune, India, since 2005 to

June 2021. Currently, he is working as a professor and the head in the Department of Artificial intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune, India. He has more than 21 years of teaching and research experience. He is serving as a subject expert in Computer Engineering, Research and Recognition Committee at several universities like SPPU (Pune), SGBU (Amravati). He is a senior member IEEE, ACM member, Life member CSI and Life member ISTE. Also, he is a member of *IEEE Transaction on Information Forensics and Security*, *IEEE Internet of Things Journal*. He is a reviewer for IGI Global—committee member for international conferences and symposium like IEEE ICC, IEEE INDICON, IEEE GCWSN, IEEE ICCUBEA, etc. He is a reviewer for the Springer *Journal of Wireless Personal Communications*, reviewer for Elsevier *Journal of Applied Computing and Informatics*, member of the Editorial Review Board of IGI Global—*International Journal of Ambient Computing and Intelligence*(IJACI), member of the Editorial Review Board for *Journal of Global Research in Computer Science*. He has published more than 190 research publications having 1900+ citations and H index 18. He has seven edited books to his credit by Springer and CRC Press. He has seven patents to his credit. He has also delivered invited talk on "Identity Management in IoT" to Symantec Research Lab, Mountain View, California. He has delivered more than 100 lectures at national and international level on IoT, big data and digitization. He has authored 15 books on subjects like *Context-aware Pervasive Systems and Application* (Springer Nature Press), *Design and Analysis of Algorithms* (Cambridge University Press), *International Journal of Rough Sets and Data Analysis*(IJRSDA), Associate Editor for IGI Global—*International Journal of Synthetic Emotions* (IJSE), *International Journal of Grid and Utility Computing* (IJGUC). He is a Member-Editorial Review Board for IGI Global—*International Journal of Ambient Computing and Intelligence* (IJACI). He is also working as an Associate Editor for IGI Global—*International Journal of Synthetic Emotions* (IJSE). He has also remained a technical program University), Identity Management for the Internet of Things (River Publications), Data Structure and Algorithms (Cengage Publications), Programming using Python—(Tech-Neo Publications MSBTE). He had worked as Chairman of Board of Studies (Information Technology), SPPU, Pune. He is working as Member—Board of Studies (Computer Engineering), SPPU, Pune. He has been a member of the Board of Studies at several institutions like VIT (Pune), Govt. College (Karad), Sandeep University (Nashik), Vishwakarma University (Pune), Dr. D. Y. Patil International University (Pune), etc. He has also remained a technical program committee member for many international conferences. He is a recognized Ph.D. guide of SSPU, Pune, and guiding seven Ph.D. students in the area of IoT and machine learning. Recently, two students have successfully defended their Ph.D. He is also the recipient of "Best Faculty Award" by Sinhgad Institutes and Cognizant Technologies Solutions. His recent research interests include algorithms, Internet of things, identity management and security. He has visited a few countries like Denmark, France, Sweden, Germany, Austria, Norway, China, Switzerland and Singapore.

Dr. Archana S. Ghotkar obtained her B.E. degree in Computer Science and Engineering from Government Engineering College, Aurangabad, India, and M.E. degree

in Computer Engineering from College of Engineering Pune (COEP), Savitribai Phule Pune University (SPPU), Pune, India. She completed her Ph.D. in Computer Engineering from Savitribai Phule Pune University. She is working as an associate professor in the Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India. She has more than 21 years of teaching and research experience. She has completed a sponsored research Project under Fast Track Young Scientist Scheme, funded by the Department of Science and Technology, Ministry of Science and Technology New Delhi, India. She is a reviewer for *IEEE Transaction on Sensors*, *IEEE Industrial Informatics*, *Springer-Nano-Computing*, *International Journal on Human Computer Interaction*, *IET-Computer Vision*, *IEEE Transactions on Systems, Man and Cybernetics*. She has also remained a technical program committee member, reviewer for many International conferences. She has published more than 25 research publications having 549 citations and H index of 11 and I10 index of 13. She has received two best paper awards at international/national conferences. She has published three Indian Patents out of which two patents are granted. She has worked as a subject expert and an author for development of lab manuals for C++ programming and Question bank development projects under MSBTE Project, India. She is a recognized Ph.D. guide of SPPU, Pune, and currently one Ph.D. student is working under her in the domain of machine learning and natural language processing. Her recent research interests include computer vision, pattern recognition, machine learning, human computer interaction, and natural language processing. Currently, she is focusing her research work on Indian sign language interpretation as well as multilingual sign language translator.

Part I
Basics of Graph Theory

Chapter 1
Graph Theory Basics

1.1 Introduction

Increased size of the data is unanticipated, and the key enablers for this growth in data size are availability of Internet at faster and cheaper rate, reduction in the cost of connection, exponential growth in the scale of devices connected to the internet, more numbers of devices being manufactured with WIFI capabilities, and the market of smart phone is sky rocketing. Due to this, every device which is connected to the Internet is posting some data on the remote storage on Web, i.e. cloud. Due to advancement in the semiconductor industry, the memory is getting cheaper day by day. This has reduced the efforts and issues required for extracting, transforming and loading the data in the form of database management system. In the sequel, information retrieval (IR) is the main issue being faced by all information technology giants and solution providers. Data storage and representations have been always evolving due to the advancements in the programming paradigm and data structures [1]. The data storage and application of IR techniques on this data is one of the important design issues. Selection of appropriate data structure by considering the design and implementation issue plays crucial role in entire IR process.

Organized collection of data items with set of operations defined on these data items is referred as data structures. Data structures are the foundations of any data storage, representation and access. There are numerous data structures presented and discussed in the literature which includes arrays, list, graphs, trees, etc. [2]. Graphs [3–5] are the fundamental data structure to store the data for easy IR. IR and data retrieval are the two terms which are used alternatively in the literature. However, there is clear boundary between these two terms. IR deals with the retrieval of the outcome which is ordered by some relevance/context, whereas outcomes are unordered by relevance/context in the case of data retrieval. In addition to this, IR techniques are generally based on the probabilistic model, whereas data retrieval techniques are using deterministic models for implementation. Algorithmic storage and retrieval of data with the help of computers is referred as IR. Graph-based IR (GBIR) [6, 7] is emerging IR technique which is gaining lot of popularity in these days

S. S. Sonawane et al., *Information Retrieval and Natural Language Processing*,
Studies in Big Data 104, https://doi.org/10.1007/978-981-16-9995-5_1