Witold Pedrycz
Shyi-Ming Chen *Editors*

Information Granularity, Big Data, and Computational Intelligence

Springer

Witold Pedrycz · Shyi-Ming Chen
Editors

Information Granularity, Big Data, and Computational Intelligence

Springer

Editors
Witold Pedrycz
Department of Electrical and Computer Engineering
University of Alberta
Edmonton, AB
Canada

Shyi-Ming Chen
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei
Taiwan

ISSN 2197-6503 ISSN 2197-6511 (electronic)
ISBN 978-3-319-08253-0 ISBN 978-3-319-08254-7 (eBook)
DOI 10.1007/978-3-319-08254-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014944331

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Studies in Big Data

Volume 8

For further volumes:
http://www.springer.com/series/11970

About this Series

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 incl. 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.

Preface

The recent pursuits emerging in big data processing, interpretation, collection, and organization have emerged in numerous sectors including business, industry, and not-for-profit organizations. Data sets such as customer transactions for a mega-retailer, weather monitoring, intelligence gathering can quickly outpace the capacity of traditional techniques and tools of data analysis. We have been witnessing an emergence of new techniques and tools including NoSQL databases, MapReduce, Natural Language Processing, Machine Learning, visualization, acquisition, and serialization.

It becomes imperative to fully become aware what happens when big data grows up: how they are being applied and where they start playing a crucial role. We also need to become fully become aware of implications and requirements imposed on the existing techniques and various methods under development.

Soft Computing regarded as a plethora of technologies of fuzzy sets (or Granular Computing, in general), neurocomputing, and evolutionary optimization brings forward a number of unique features that might be instrumental to the development of concepts and algorithms to deal with big data. In particular, setting up a suitable and fully legitimate level of abstraction by forming semantically meaningful information granules is of paramount relevance. In light of their sheer volume, big data may call for distributed processing, where results of intensive data mining realized locally are afterwards reconciled leading to information granules of higher type. Neurocomputing operating at information granules leads to more tractable learning tasks. Evolutionary computing delivers an essential framework supporting global optimization.

In light of the inherent human-centric facet of Granular Computing the principles and practice of Computational Intelligence have been poised to play a vital role in the analysis, design, and interpretation of the architectures and functioning of mechanisms of big data.

Our ultimate objectives of this edited volume is to provide the reader with an updated, in-depth material on the emerging principles, conceptual underpinnings, algorithms, and practice of Computational Intelligence in the realization of concepts and implementation of big data architectures, analysis, and interpretation as well as data analytics.

An overall concise characterization of the objectives of the edited volume is expressed by highlighting several focal points:

- Systematic exposure of the concepts, design methodology, and detailed algorithms. In general, the volume adheres to the top-down strategy starting with the concepts and motivation and then proceeding with the detailed design that materializes in specific algorithms and representative applications.
- Individual chapters with clearly delineated agenda and well-defined focus and additional reading material available via carefully selected references.
- A wealth of carefully structured and organized illustrative material. The volume includes a series of brief illustrative numeric experiments, detailed schemes, and more advanced problems. They make the material more readable and appealing.
- Self-containment. Given the emerging character of the area of big data, our ultimate intent is to deliver a material that is self-contained and provides the reader with all necessary prerequisites and, if necessary, augments some parts with a step-by-step explanation of more advanced concepts supported by a significant amount of illustrative numeric material and some application scenarios to motivate the reader and make some abstract concepts more tangible.

The area of big data is highly diversified and this volume offers a quite representative view of the area. The contributions published here can be organized into three main parts. The first part, Fundamentals, which comprises chapters "Nearest Neighbor Queries on Big Data" to "Building Fuzzy Robust Regression Model Based on Granularity and Possibility Distribution" is focused on the methodological issues covering a broad spectrum of the approaches and detailed algorithmic pursuits including essential topics of forming cliques in big data, exploiting robust regression and its variants, constructing and optimizing rule-based models, Latent Semantic Indexing, information granulation, and Nearest Neighbor Querying. Part II entitled Architectures consisting of chapters "The Role of Cloud Computing Architectures in Big Data" to "The Web Know ARR Framework: Orchestrating Computational Intelligence with Graph Databases" is aimed at looking at the dedicated computing architectures such as cloud computing and the use of data storage techniques. Part III (case studies) includes chapters "Customer Relationship Management and Big Data Mining" to "Application of Computational Intelligence on Analysis of Air Quality Monitoring Big Data" which offer a suite of studies serving as a testimony to a wealth of promising applications including among others Customer Relationship management, market movements, weather forecasting, and air quality monitoring.

Given the theme of this project, this book is aimed at a broad audience of researchers and practitioners. Owing to the nature of the material being covered and the way it is organized, one can project with high confidence that it will appeal to the well-established communities including those active in various disciplines in which big data, their analysis, and optimization are of genuine relevance. Those involved in data mining, data analysis, management, various branches of engineering, and economics will benefit from the exposure to the subject matter.

Considering a way in which the edited volume is structured, this book could serve as a highly useful reference material for graduate students and senior undergraduate students in courses such as those on intelligent system, data mining, pattern recognition, decision-making, Internet engineering, Computational Intelligence, management, operations research, and knowledge-based systems.

We would like to take this opportunity to express our sincere thanks to the authors for sharing the results of their innovative research and delivering their insights into the area. The reviewers deserve our thanks for their constructive and timely input. We greatly appreciate a continuous support and encouragement coming from the Editor-in-Chief, Prof. Janusz Kacprzyk whose leadership and vision makes this book series a unique vehicle to disseminate the most recent, highly relevant and far-reaching publications in the domain of Computational Intelligence and its various applications.

We hope that the readers will find this volume of genuine interest and the research reported here will help foster further progress in research, education, and numerous practical endeavors.

Witold Pedrycz
Shyi-Ming Chen

Contents

Part I Fundamentals

Part II Architectures

Part III Case Studies

Part I
Fundamentals

Nearest Neighbor Queries on Big Data

Georgios Chatzimilioudis, Andreas Konstantinidis and Demetrios Zeinalipour-Yazti

Abstract *k Nearest Neighbor (kNN)* search is one of the simplest non-parametric learning approaches, mainly used for classification and regression. *kNN* identifies the *k* nearest neighbors to a given node given a distance metric. A new challenging *kNN* task is to identify the *k* nearest neighbors for all nodes simultaneously; also known as *All kNN (AkNN)* search. Similarly, the *Continuous All kNN (CAkNN)* search answers an *AkNN* search in real-time on streaming data. Although such techniques find immediate application in computational intelligence tasks, among others, they have not been efficiently optimized to this date. We study specialized scalable solutions for *AkNN* and *CAkNN* processing as demanded by the volume–velocity-variety of data in the Big Data era. We present an algorithm, coined *Proximity*, which does not require any additional infrastructure or specialized hardware, and its efficiency is mainly attributed to our smart search space sharing technique. Its implementation is based on a novel data structure, coined k^+-heap. *Proximity*, being parameter-free, performs efficiently in the face of high velocity and skewed data. In our analytical studies, we found that *Proximity* provides better time complexity compared to existing approaches and is very well suited for large scale scenarios.

Keywords k Nearest neighbors · Big data · Computational intelligence · Smartphones

G. Chatzimilioudis (✉) · A. Konstantinidis · D. Zeinalipour-Yazti
Department of Computer Science, University of Cyprus, 1 University Avenue,
P.O. Box 20537, 1678 Nicosia, Cyprus
e-mail: gchatzim@cs.ucy.ac.cy

A. Konstantinidis
e-mail: akonstan@cs.ucy.ac.cy

D. Zeinalipour-Yazti
e-mail: dzeina@cs.ucy.ac.cy

W. Pedrycz and S.-M. Chen (eds.), *Information Granularity, Big Data, and Computational Intelligence*, Studies in Big Data 8,
DOI: 10.1007/978-3-319-08254-7_1,

1 Introduction

The k Nearest Neighbor (kNN) search [1] is one of the simplest non-parametric learning approaches, mainly used for classification [2] and regression [3]. The kNN of an object o_a from some dataset O, denoted as $kNN(o_a, O)$, are the k objects whose attributes are most similar to the attributes of o_a [1]. Formally, $\forall o_b \in kNN(o_a, O)$ and $\forall o_c \in O - kNN(o_a, O)$ given $o_a \neq o_b \neq o_c$, it always holds that $dist(o_a, o_b) \leq dist(o_a, o_c)$, where $dist$ can be any L_p-norm metric, such as Manhattan (L_1), Euclidean (L_2) or Chebyshev (L_∞).

kNN search is a classical Computational Intelligence problem with extensions that include the Condensed nearest neighbor (CNN, the Hart algorithm) algorithm that reduces the data set for kNN classification [4] and the fuzzy-kNN [5] that deals with uneven and dense training datasets. kNN further finds applicabilities in several domains such as computational geometry [6–8], image processing [9, 10], spatial databases [11, 12], and recently in social networks [13].

A new challenging *kNN* task is to identify the *k* nearest neighbors for all nodes simultaneously; also known as *All kNN* (*AkNN*) search. An AkNN search is viewed as a generalization of the basic kNN search that computes the $kNN(o, O) \forall o \in O$. In temporal and streaming data a similar task of interest is the *Continuous All kNN* (*CAkNN*) search, which answers an AkNN query in real-time.

AkNN and CAkNN are new and challenging computational intelligence problems, which cannot be efficiently tackled using existing techniques. Thus, they may serve as both a real-world benchmark and an improvement technique of computational intelligence methods. For example, CAkNN can be used for both real-time classification and regression in Big Data cases where the volume-velocity-variety of data is high and cannot be handled by conventional techniques. Consider classifying tweets provided by Twitter users in real time (with 100 billion daily active users and 143,199 tweets per second in 2013). Furthermore, it can be combined with Neural Network [14] to improve its real-time predictability when new data arrive again with high velocity-volume-variety, it can also be used with minor modifications to extend the well-known Variable Neighborhood search approach [15] for tackling combinatorial and global optimization problems. Finally, AkNN and CAkNN can be combined with Multi-Objective Evolutionary Computation approaches [13] for improving their performance in terms of speed and efficiency when dealing with Multi-objective Optimization Problems (MOPs). For example, it can be combined with MOEA based on Decomposition (MOEA/D) [16] for finding neighbors of each solution in the weight space faster, or it can be combined with the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [17] for improving the crowding-distance estimation (i.e., a techniques utilized for improving the diversity of the final result).

The advances in technology and the automatization of many processes in numerous sectors, including business, industry, and not-for-profit organizations, has led to the pursuit of big data processing, interpretation, collection and organization. For example, the proliferation of smart devices and sensors with the trend to share, communicate and store data has brought an information explosion in spatio-temporal applications, with ever increasing amounts of data that need to be efficiently managed. The CAkNN task is of great interest since it offers a new dimension of neighborhood "sensing". Applications of this neighborhood "sensing" capability could enhance public emergency services like E9-1-1 [18] and NG9-1-1 [19], and facilitate the uptake of location-based social networks (e.g., Rayzit [20], Waze [21]).

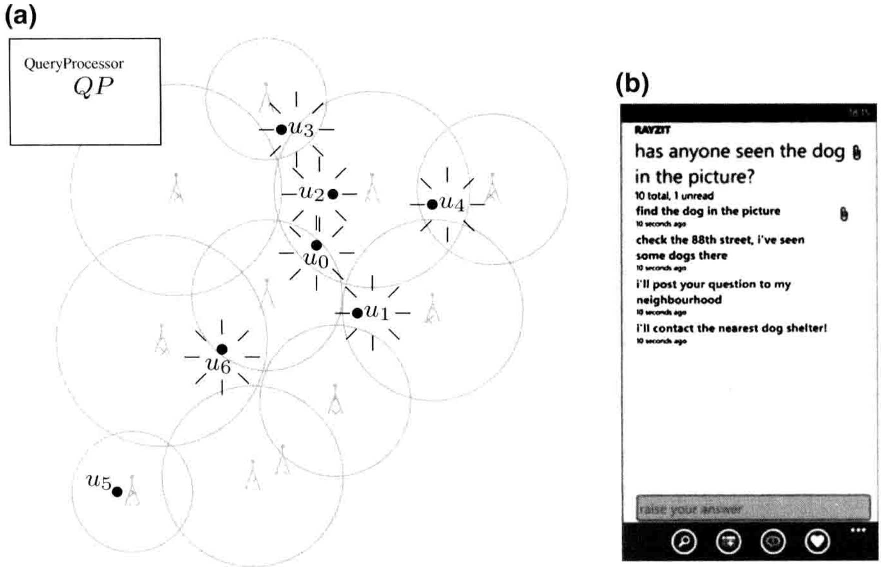
In this chapter, we study the problem of efficiently processing a CAkNN search in a cellular or WiFi network, both of which are ubiquitous. We present an algorithm, coined *Proximity*, which does not require any additional infrastructure or specialized hardware and its efficiency is mainly attributed to a smart *search space sharing* technique we present and analyze. Its implementation is based on a novel data structure, coined k^+-heap. *Proximity*, being parameter-free, performs efficiently in the face of high mobility and skewed distribution of users (e.g., the service works equally well in downtown, suburban, or rural areas).

Consider a set of smartphone users moving in the plane of a geographic region. Let such an area be covered by a set of *Network Connectivity Points* (*NCP*) (e.g., cellular towers of cellular networks, WiFi access points of wireless 802.11 networks etc.) Each *NCP* inherently creates the notion of a *cell*. Without loss of generality, let the cell be represented by a circular area[1] with an arbitrary radius. A mobile user u is serviced at any given time point by one *NCP*, but is also aware of the other *NCP*s in the vicinity whose communication range reach u (e.g., cell-ids of different providers in an area, or MAC addresses of WiFi hot-spots in an area).

To illustrate our abstraction, consider the example network shown in Fig. 1, where we provide a micro-blogging chat channel [20] between each user u and its $k = 2$ nearest neighbors. In the given scenario, each user concurrently requires a different answer-set to a globally executed query, as shown in the caption of Fig. 1. Notice that the answer-set for each user u is not limited within its own *NCP* and that each *NCP* has its own communication range. Additionally, there might be areas with dense user population and others with sparse user population. Consequently, finding the k-nearest neighbors of some arbitrary user u could naively involve from a simple lookup in the *NCP* of u to a complex iterative deepening into neighboring *NCP*s, as we will show in Fig. 3b.

The remaining of this chapter is organized as follows: Sect. 2 introduces fields for which the proposed framework caters for. Section 3 defines our system model and the problem. Section 4 provides the related work necessary for understanding the foundations of this work. Section 5 presents the *Proximity* framework and a breakdown of our data structures and algorithms. Section 6 finally summarizes and concludes the knowledge acquired from existing research and discusses our future plans.

[1] Using other geometric shapes (e.g., hexagons, Voronoi polygons, grid-rectangles, etc.) for space partitioning is outside the scope of this paper.



has anyone seen the dog in the picture?

10 total, 1 unread

find the dog in the picture
10 seconds ago

check the 88th street, i've seen some dogs there
10 seconds ago

i'll post your question to my neighbourhood
10 seconds ago

i'll contact the nearest dog shelter!
10 seconds ago

raise your answer

Fig. 1 **a** A snapshot of a cellular network instance, where the 2-nearest neighbors for u_0 are $\{u_1, u_2\}$. Similarly for the other users: $u_1 \rightarrow \{u_0, u_2\}, u_2 \rightarrow \{u_3, u_0\}$, $u_3 \rightarrow \{u_2, u_0\}$, $u_4 \rightarrow \{u_2, u_3\}$, $u_6 \rightarrow \{u_0, u_1\}$. **b** Rayzit [20], an example application of a proximity-based micro-blogging chat

2 Background

Big data refers to data sets whose size and structure strains the ability of commonly used relational DBMSs to capture, manage, and process the data within a tolerable elapsed time [22]. The *volume–velocity-variety* of information in this kind of datasets give rise to the big data challenge, which is also known as the 3V challenge.

The *volume* of such datasets is in the order of few terabytes (TB) to petabytes (PB) that are often of high *information granularity*. Examples of such volumes are the U.S. Library of Congress that in April 2011 had more than 235 TB of data stored and the World of Warcraft online game using 1.3 PB of storage to maintain its game, the German Climate Computing Center (DKRZ) storing 60 PB of climate data.

The *velocity* of information in social media applications (such as photovoltaic, traffic and other monitoring apps) can grow exponentially as users join the community. Such growth can produce unprecedented volumes of data streams. For example, Ontario's Meter Data Management and Repository (MDM/R) [23] stores, processes and manages data from 4.6 million smart meters in Ontario, Canada and provides hourly billing quantity and extensive reports counting 110 million meter reads per day on an annual basis that exceeds the number of debit card transactions processed in Canada.