UNIVERSITY OF WARSAW

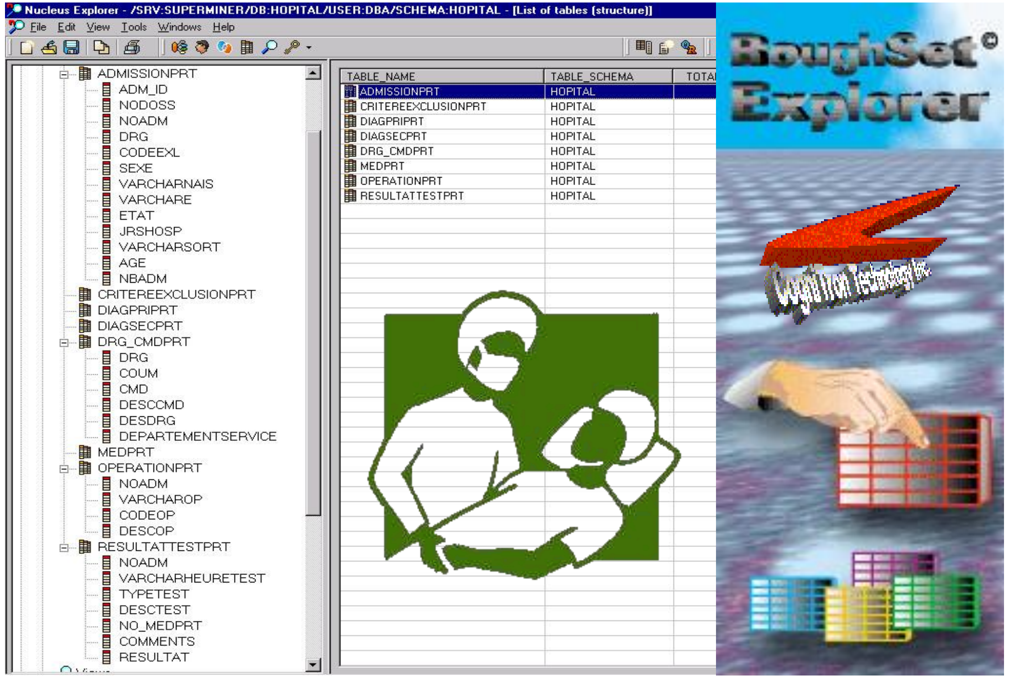
DE SOFTWARE

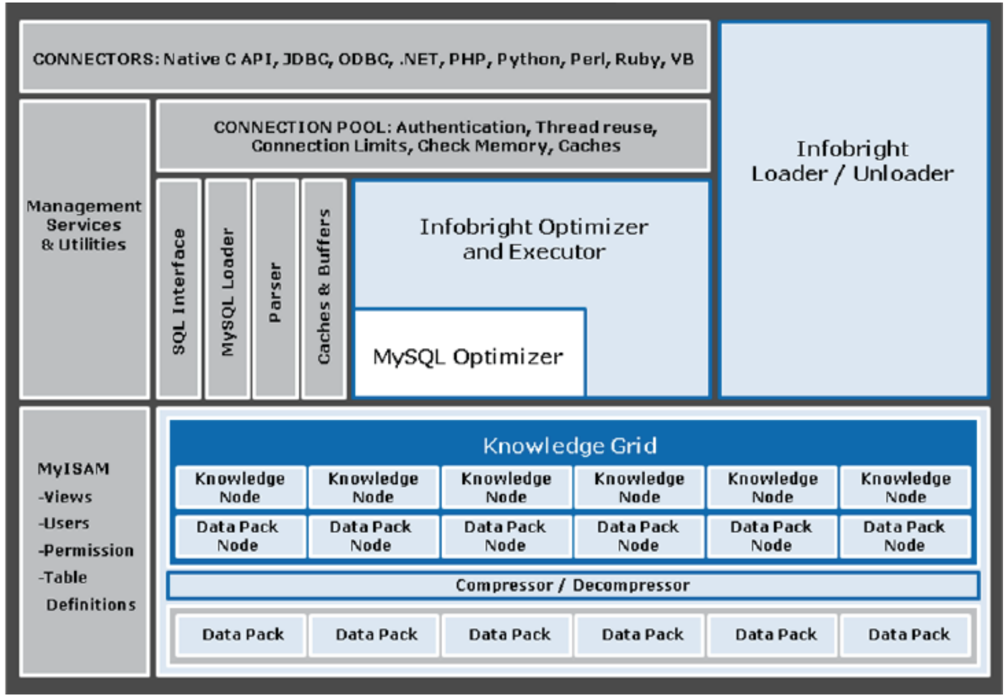
Rough Sets In Industry

Dominik Ślęzak

My (Industry) Background

- Data Analytics 1995-2005
- Data Processing 2005-2015
- *Data Science* 2015-...





INFOBRIGHT

CISCO

SOD SECURITY ON-DEMAND

Information Builders

AI for Pioneers

SOFTWARE

International Journal of Computer and Information Sciences, Vol. 11, No. 5, 1982

Rough Sets

Zdzisław Pawlak[1]

Received June 1981; revised Sempember 1982

We investigate in this paper approximate operations on sets, approximate equality of sets, and approximate inclusion of sets. The presented approach may be considered as an alternative to fuzzy sets theory and tolerance theory. Some applications are outlined.

KEY WORDS: Artificial intelligence; automatic classification; cluster analysis; fuzzy sets; inductive reasoning; learning algorithms; measurement theory; pattern recognition; tolerance theory.

Apart from the known and the unknown, what else is there?
Harold Pinter (The Homecoming)

1. INTRODUCTION

The aim of this paper is to describe some properties of rough sets, introduced in Ref. 7 and investigated in Refs. 1, 2, 4, 5, 6, 8, 9, and 11.

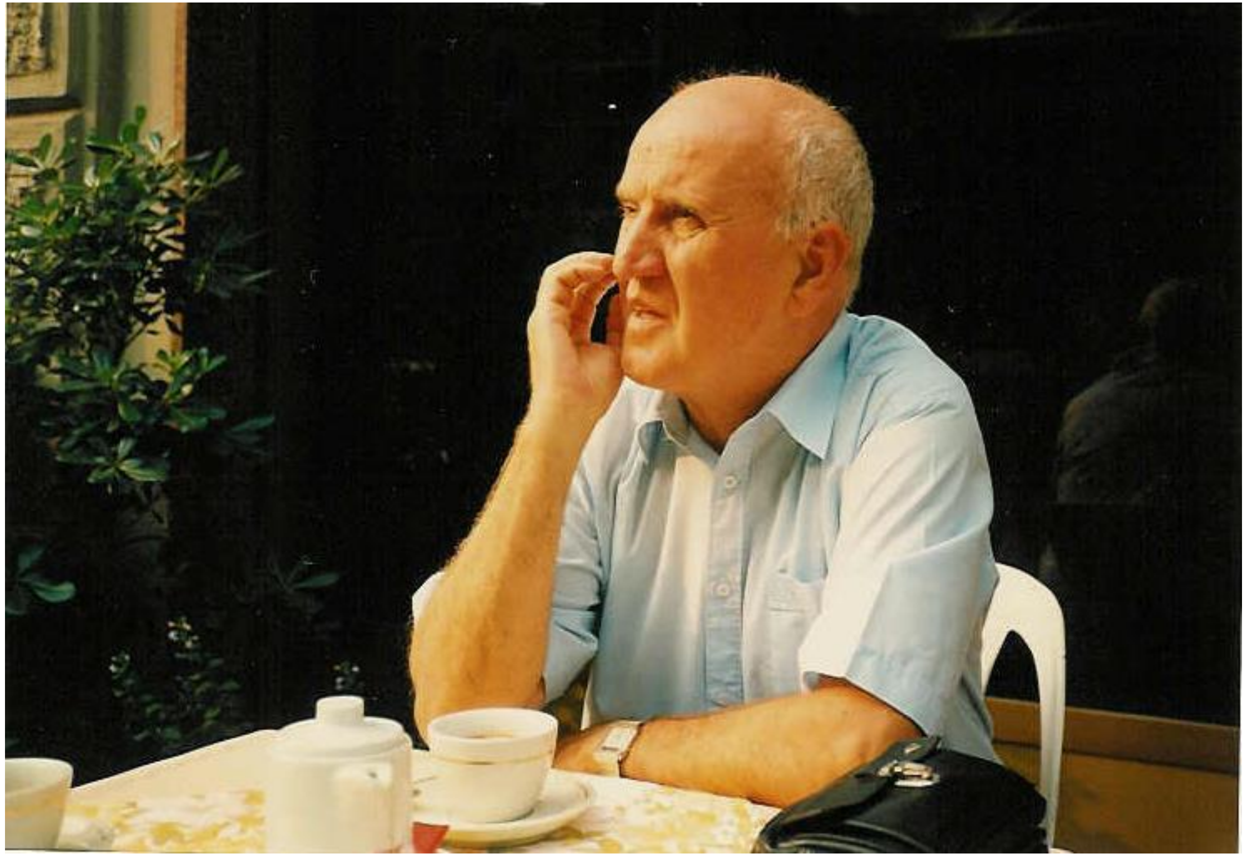
The rough set concept can be of some importance, primarily in some branches of artificial intelligence, such as inductive reasoning, automatic classification, pattern recognition, learning algorithms, etc.

The idea of a rough set could be placed in a more general setting, leading to a fruitful further research and applications in classification theory, cluster analysis, measurement theory, taxonomy, etc.

The key to the presented approach is provided by the exact mathematical formulation of the concept of approximative (rough) equality of sets in a given approximation space; an approximation space is understood as a pair (U, R), where U is a certain set called universe, and $R \subset U \times U$ is an indiscernibility relation. We assume throughout this paper that R is an equivalence relation.

[1] Institute of Computer Sciences, Polish Academy of Sciences, P.O. Box 22, 00-901 Warsaw, PKiN.

0091-7036/82/1000-0341$03.00/0



Zdzisław Pawlak 1926-2006
Rough Sets 1982-2022

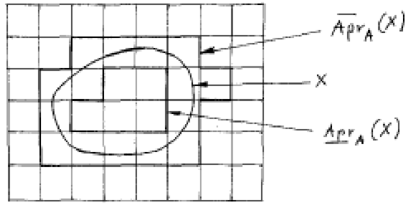


Fig. 1

Sets $\underline{\text{Edg}}_A(X) = X - \underline{\text{Apr}}_A(X)$ (in short $\underline{\text{Edg}}(X)$) and $\overline{\text{Edg}}_A(X) = \overline{\text{Apr}}_A(X) - X$, (in short $\overline{\text{Edg}}(X)$) are referred to as an *internal* and an *external edge* of X in A, respectively.

Of course $\text{Bnd}_A(X) = \overline{\text{Edg}}_A(X) \cup \underline{\text{Edg}}_A(X)$.

Fig. 1 shows the notion of an upper and lower approximation in a two-dimensional approximation space consisting of a rectangle partitioned into elementary squares.

Let us define two membership functions $\underline{\in}_A$, $\overline{\in}_A$ (called *strong* and *weak* membership, respectively), as follows:

$$x \underline{\in}_A X \quad \text{iff} \quad x \in \underline{\text{Apr}}_A(X)$$

$$x \overline{\in}_A X \quad \text{iff} \quad x \in \overline{\text{Apr}}_A(X)$$

If $x \underline{\in}_A X$, we say that "X *surely* belongs to X in A," while $x \overline{\in}_A X$ is to mean that "X *possibly* belongs to X in A." Thus we can interprete approximations as counterparts of necessity and possibility in modal logic.

Of course,

$$\underline{\text{Apr}}_A(X) = \{x : x \underline{\in}_A X\}$$

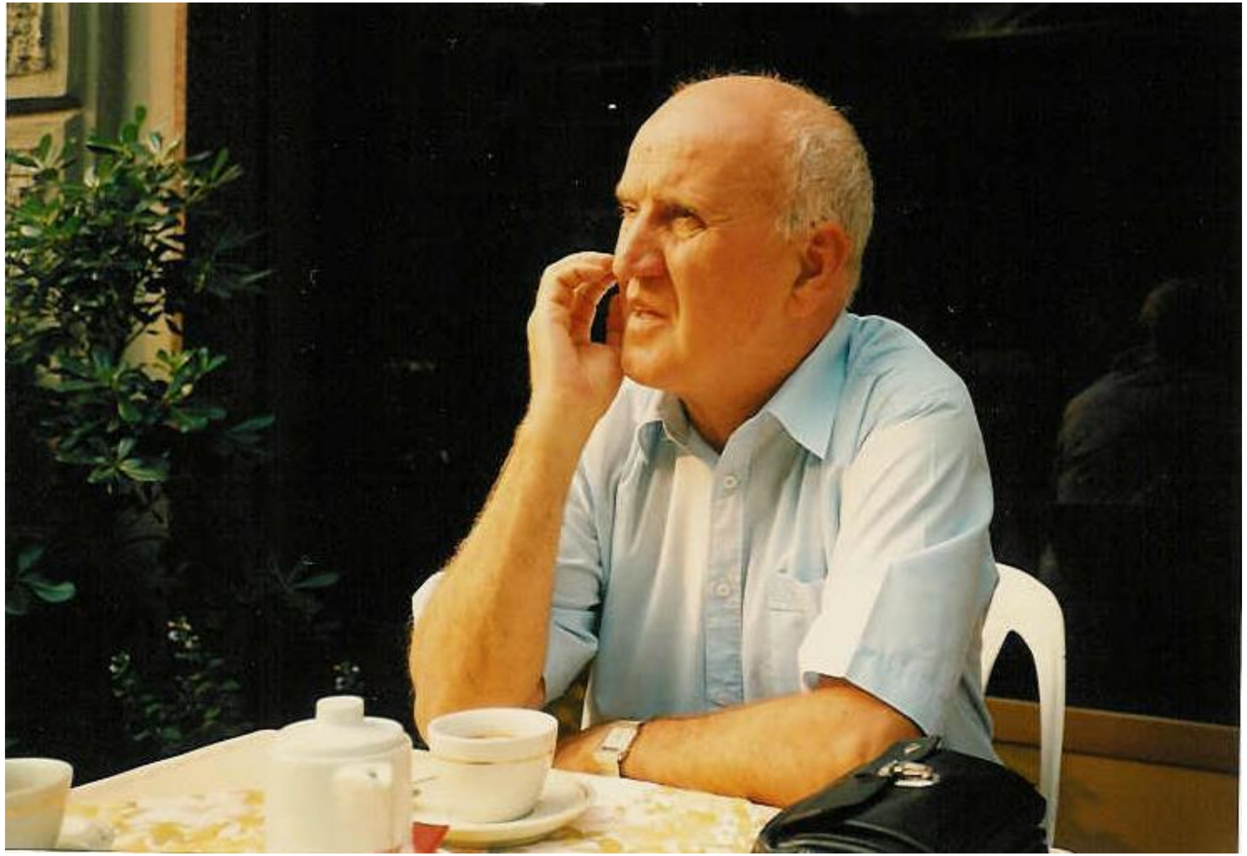
$$\overline{\text{Apr}}_A(X) = \{x : x \overline{\in}_A X\}$$

Thus we can develop our theory in terms of strong and weak membership functions or in terms of approximations. For the sake of simplicity we shall use here the approximational approach.

2.2. Approximation Space and Topological Space

It is easy to check that the approximation space $A = (U, R)$ defines uniquely the topological space $T(A)$ (in short T_A), where $T_A = (U, \text{Com}(A))$, and $\text{Com}(A)$ are the family of all open sets in T_A, and U/R is a base for T_A.

It follows from the definition of (lower and upper) approximations that $\text{Com}(A)$ is both the set of all open and closed sets in T_A. Thus, $\underline{\text{Apr}}_A(X)$ and



Approximations, their calculus and examples

Thus, we can interpret the approximation space $A = (R^+, S)$ as a *measurement system*, where

$$\bar{\mu}_A\langle i, i+1\rangle = \underline{\mu}_A\langle i, i+1\rangle = 1, \quad i = 0, 1, \ldots$$

is the *unit of measurement* in A, and $\eta(0, r)$ is the accuracy of $(0, r)$ in A. For more detail see Ref. 6.

Example 2. Let V be a finite set called a *vocabulary* and let V^* be the set of all finite sequences over V. Any subset of V^* will be called a *language* over V.

Let $R \subset V^* \times V^*$ be an *indiscernibility* relation, and let $A = (V^*, R)$ be an approximation space defined by V^* and R.

A language $L \subset V^*$ is *recognizable* in A if $\underline{\mathrm{Apr}}_A(L) = \overline{\mathrm{Apr}}_A(L)$.

The family of all recognizable languages in A, denoted as $\mathrm{Rec}(A)$, is the topology induced by $A = (V^*, R)$ and the base of the topology is V^*/R.

That is to say that if the language L is not recognizable in A we are able to recognize only the lower and upper approximations in A.

This property can be used in speach recognition, pattern recognition, fault tolerant computers, etc.

Example 3. Let $S = \langle X, A, V, \rho\rangle$ be an information system (see Ref. 10), where

X is the set of *objects*

A is the set of *attributes*

$V = \bigcup V_a$, V_a is the set of values of attribute $a \in A$

$\rho: X \times A \to V$ is an *information function*, $\rho_X: A \to V$

$x \in X$ is called an *information about* x in S, where

$$\rho_x(a) = \rho(x, a)$$

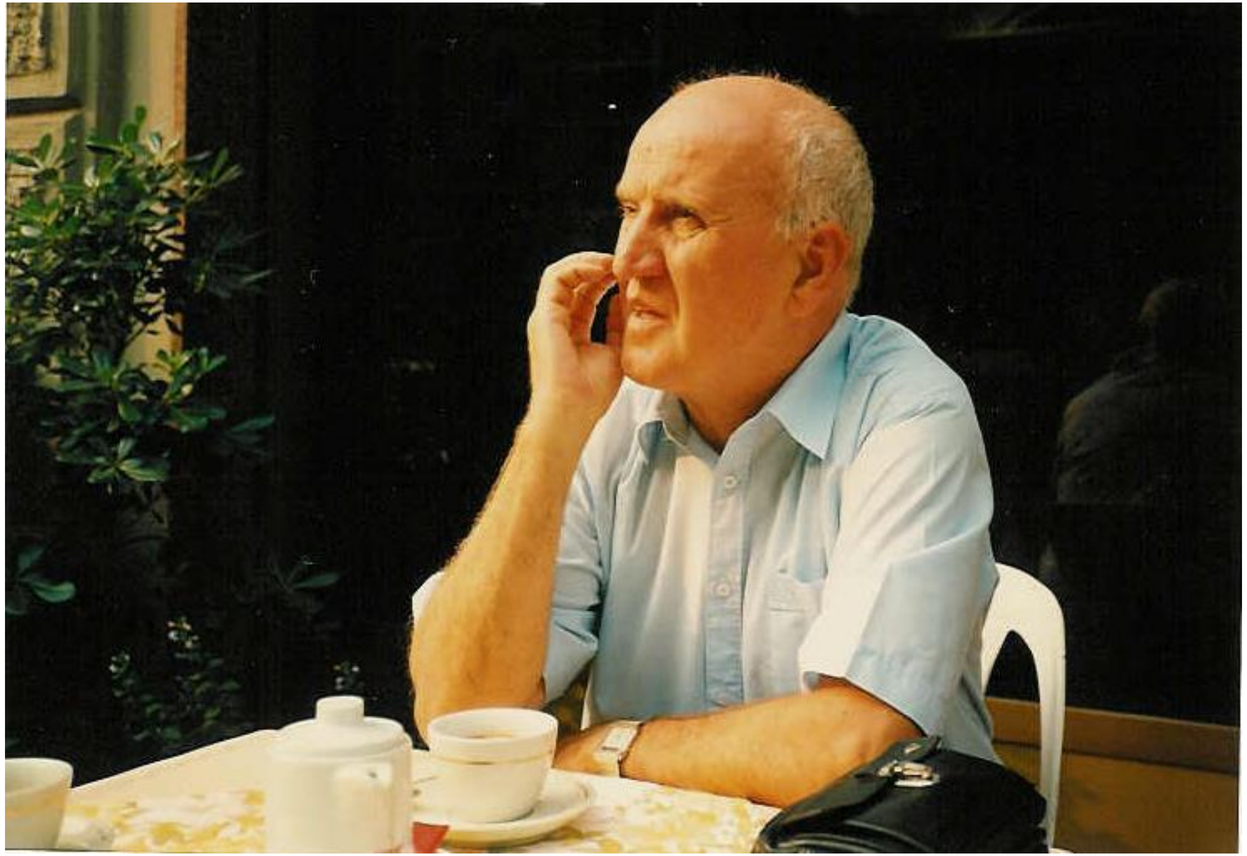
for every $x \in X$ and $a \in A$.

We define the binary relation $\tilde{S}$ over X in the following way:

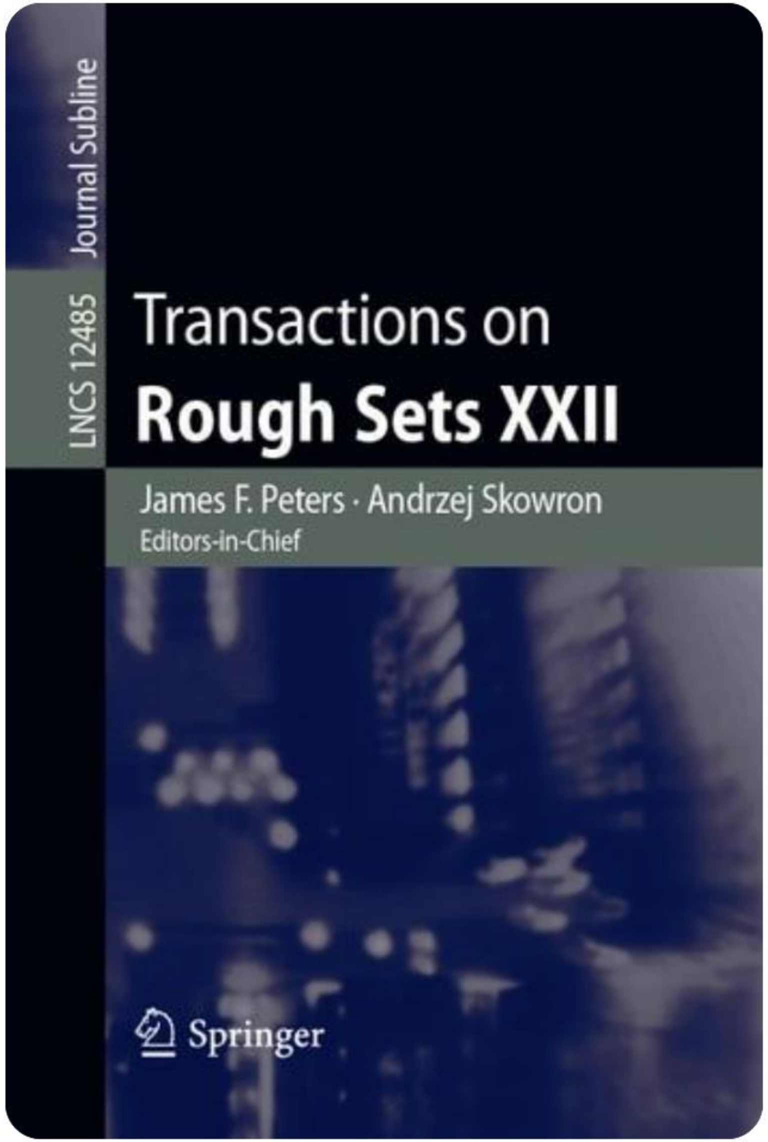
$$x \sim_S y \quad \text{iff} \quad \rho_x = \rho_y$$

Obviously $\tilde{S}$ is an equivalence relation and $A = (X, \tilde{S})$ is the approximation space induced by the information system S.

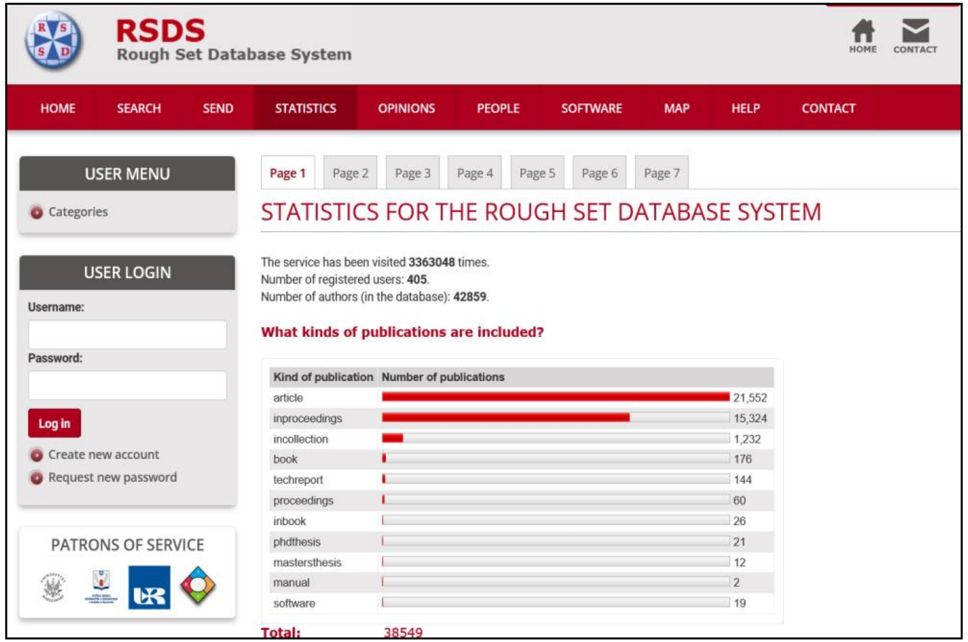
Any subset $Y \subset X$ is called *describable* in S iff $\underline{\mathrm{Apr}}_A(Y) = \overline{\mathrm{Apr}}_A(Y)$. The set of all describable sets in S, denoted as $\mathrm{Des}(S)$, is a topology induced by S on X, and the base of the topology is $X/\tilde{S}$.



Towards decision reducts (how approximations change when we add / remove attributes)



ACM Transactions on Intelligent Systems and Technology
Annals of Pure and Applied Logic
Applied Intelligence
Applied Soft Computing
Artificial Intelligence
Artificial Intelligence Review
BMC Bioinformatics
Communications of the ACM
European Journal of Operational Research
Expert Systems with Applications
Fundamenta Informaticae
Fuzzy Sets and Systems
Group Decision and Negotiation
IEEE Transactions on Computational Social Systems
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Fuzzy Systems
IEEE Transactions on Geoscience and Remote Sensing
IEEE Transactions on Image Processing
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Neural Networks
IEEE Transactions on Systems, Man and Cybernetics
IEEE/ACM Transactions on Audio, Speech, and Language Processing
IEEE/ACM Transactions on Computational Biology and Bioinformatics
Information Sciences
International Journal of Approximate Reasoning
International Journal of Computational Intelligence Systems
International Journal of Machine Learning and Cybernetics
International Journal of Molecular Science
International Journal of Science and Engineering
Journal of Applied Non-Classical Logics
Journal of Biomedical Informatics
Knowledge and Information Systems
Knowledge Based Systems
Neural Computing and Applications
Neural Networks
Neural Processing Letters
Neurocomputing
Pattern Recognition
Pattern Recognition Letters
Pharmaceutics
Sensors
Studia Logica
Theoretical Computer Science
Web Intelligence and Agent Systems...



US011301467B2

(12) **United States Patent**
Slezak et al.

(10) **Patent No.: US 11,301,467 B2**
(45) **Date of Patent: Apr. 12, 2022**

(54) **SYSTEMS AND METHODS FOR INTELLIGENT CAPTURE AND FAST TRANSFORMATIONS OF GRANULATED DATA SUMMARIES IN DATABASE ENGINES**

(71) Applicant: **Security On-Demand, Inc.**, San Diego, CA (US)

(72) Inventors: **Dominik Slezak**, Warsaw (PL); **Richard Glick**, Valley Center, CA (US); **Pawel Betlinski**, Warsaw (PL); **Piotr Synak**, Winterthur (CH); **Jakub Wroblewski**, Lomianki (PL); **Agnieszka Chadzynska-Krasowska**, Sulejowek (PL); **Janusz Borkowski**, Warsaw (PL); **Arkadiusz Wojna**, Warsaw (PL); **Joel Alan Holland**, Encinitas, CA (US)

(73) Assignee: **Security On-Demand, Inc.**, San Diego, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 463 days.

(21) Appl. No.: **16/459,274**

(22) Filed: **Jul. 1, 2019**

(65) **Prior Publication Data**
US 2020/0004749 A1 Jan. 2, 2020

Related U.S. Application Data
(60) Provisional application No. 62/691,751, filed on Jun. 29, 2018.

(51) **Int. Cl.**
G06F 16/2458 (2019.01)
G06F 16/2453 (2019.01)
(Continued)

(52) **U.S. Cl.**
CPC ***G06F 16/24539*** (2019.01); ***G06F 16/2282*** (2019.01); ***G06F 16/2462*** (2019.01); ***G06N 5/003*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 16/24539; G06F 16/2282; G06F 16/2462; G06N 5/003
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,671,772 B1 12/2003 Cousins et al.
8,266,147 B2 9/2012 Slezak et al.
(Continued)

FOREIGN PATENT DOCUMENTS

WO 2008034219 A1 3/2008

OTHER PUBLICATIONS

International Preliminary Report on Patentability for International Patent Application No. PCT/CA2007/001627 dated Jan. 7, 2008.
(Continued)

Primary Examiner — Kris E Mackes
(74) *Attorney, Agent, or Firm* — Smith, Gambrell & Russell LLP

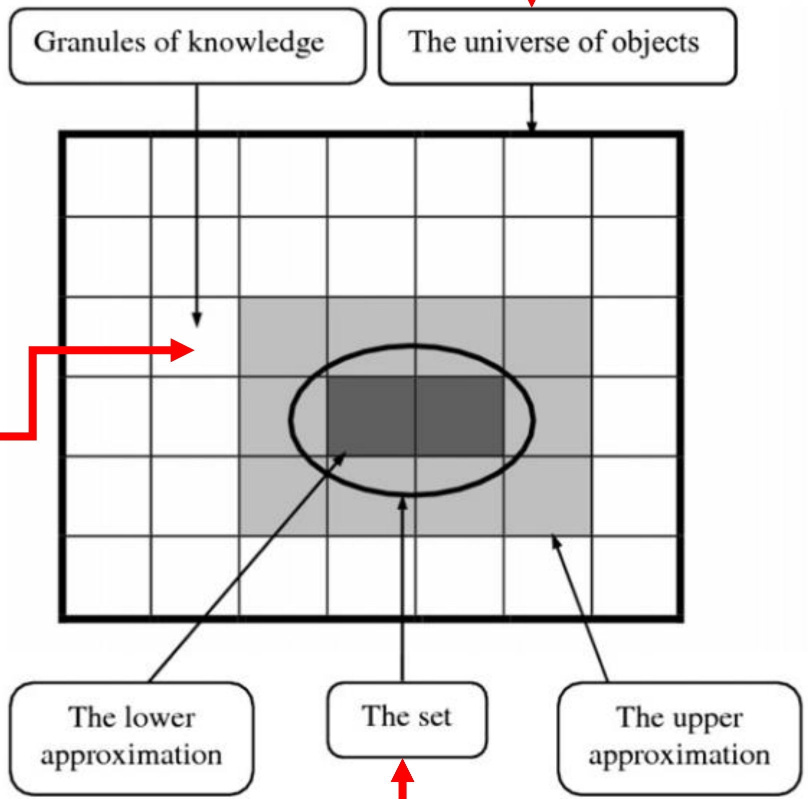
(57) **ABSTRACT**

Embodiments may provide methods and systems for intelligent capture and fast transformation of granulated data summaries. An engine may be used to transform input data summaries into result sets representing query outcomes. The data summaries contain enough knowledge about the original data to accurately perform operations on the summaries without needing to access the original data. In an embodiment, the contents of data summaries are accessible via an SQL approximate engine which retrieves summaries stored on disk and utilizes them for its operations. Alternatively, the contents of data summaries are accessible via virtual tables which give users direct access to the summary contents and
(Continued)

Rows in data table stored in partition blocks

Partition blocks enriched with statistical metadata

SQL select statement to be computed



M Wnuk, S Stawicki, D Ślęzak: Reinventing Infobright's Concept of Rough Calculations on Granulated Tables for the Purpose of Accelerating Modern Data Processing Frameworks. IEEE BigData 2020: 5405-5412

US011301467B2

(12) **United States Patent**
Slezak et al.

(10) **Patent No.: US 11,301,467 B2**
(45) **Date of Patent: Apr. 12, 2022**

(54) **SYSTEMS AND METHODS FOR INTELLIGENT CAPTURE AND FAST TRANSFORMATIONS OF GRANULATED DATA SUMMARIES IN DATABASE ENGINES**

(71) Applicant: **Security On-Demand, Inc.**, San Diego, CA (US)

(72) Inventors: **Dominik Slezak**, Warsaw (PL); **Richard Glick**, Valley Center, CA (US); **Pawel Betlinski**, Warsaw (PL); **Piotr Synak**, Winterthur (CH); **Jakub Wroblewski**, Lomianki (PL); **Agnieszka Chadzynska-Krasowska**, Sulejowek (PL); **Janusz Borkowski**, Warsaw (PL); **Arkadiusz Wojna**, Warsaw (PL); **Joel Alan Holland**, Encinitas, CA (US)

(73) Assignee: **Security On-Demand, Inc.**, San Diego, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 463 days.

(21) Appl. No.: **16/459,274**

(22) Filed: **Jul. 1, 2019**

(65) **Prior Publication Data**

US 2020/0004749 A1 Jan. 2, 2020

Related U.S. Application Data

(60) Provisional application No. 62/691,751, filed on Jun. 29, 2018.

(51) **Int. Cl.**
G06F 16/2458 (2019.01)
G06F 16/2453 (2019.01)
(Continued)

(52) **U.S. Cl.**
CPC ***G06F 16/24539*** (2019.01); ***G06F 16/2282*** (2019.01); ***G06F 16/2462*** (2019.01); ***G06N 5/003*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 16/24539; G06F 16/2282; G06F 16/2462; G06N 5/003
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,671,772 B1 12/2003 Cousins et al.
8,266,147 B2 9/2012 Slezak et al.
(Continued)

FOREIGN PATENT DOCUMENTS

WO 2008034219 A1 3/2008

OTHER PUBLICATIONS

International Preliminary Report on Patentability for International Patent Application No. PCT/CA2007/001627 dated Jan. 7, 2008.
(Continued)

Primary Examiner — Kris E Mackes
(74) *Attorney, Agent, or Firm* — Smith, Gambrell & Russell LLP

(57) **ABSTRACT**

Embodiments may provide methods and systems for intelligent capture and fast transformation of granulated data summaries. An engine may be used to transform input data summaries into result sets representing query outcomes. The data summaries contain enough knowledge about the original data to accurately perform operations on the summaries without needing to access the original data. In an embodiment, the contents of data summaries are accessible via an SQL approximate engine which retrieves summaries stored on disk and utilizes them for its operations. Alternatively, the contents of data summaries are accessible via virtual tables which give users direct access to the summary contents and
(Continued)

SELECT MAX(A) FROM T WHERE B>15;

1 → 2 → 3

Pack A1 Min = 3 Max = 25	Pack B1 Min = 10 Max = 30		S	S	S	E	E
Pack A2 Min = 1 Max = 15	Pack B2 Min = 10 Max = 20		S	I	I	I	I
Pack A3 Min = 18 Max = 22	Pack B3 Min = 5 Max = 50		S	S	S	I/E	I/E
Pack A4 Min = 2 Max = 10	Pack B4 Min = 20 Max = 40		R	I	I	I	I
Pack A5 Min = 7 Max = 26	Pack B5 Min = 5 Max = 10		I	I	I	I	I
Pack A6 Min = 1 Max = 8	Pack B6 Min = 10 Max = 20		S	I	I	I	I

I/S/R denotes irrelevant/suspect/relevant; E – exact computation (decompression)

INFOBRIGHT

SECURITY ON-DEMAND

US011301467B2

(12) **United States Patent**
Slezak et al.

(10) **Patent No.: US 11,301,467 B2**
(45) **Date of Patent: Apr. 12, 2022**

(54) **SYSTEMS AND METHODS FOR INTELLIGENT CAPTURE AND FAST TRANSFORMATIONS OF GRANULATED DATA SUMMARIES IN DATABASE ENGINES**

(71) Applicant: **Security On-Demand, Inc.**, San Diego, CA (US)

(72) Inventors: **Dominik Slezak**, Warsaw (PL); **Richard Glick**, Valley Center, CA (US); **Pawel Betlinski**, Warsaw (PL); **Piotr Synak**, Winterthur (CH); **Jakub Wroblewski**, Lomianki (PL); **Agnieszka Chadzynska-Krasowska**, Sulejowek (PL); **Janusz Borkowski**, Warsaw (PL); **Arkadiusz Wojna**, Warsaw (PL); **Joel Alan Holland**, Encinitas, CA (US)

(73) Assignee: **Security On-Demand, Inc.**, San Diego, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 463 days.

(21) Appl. No.: **16/459,274**

(22) Filed: **Jul. 1, 2019**

(65) **Prior Publication Data**

US 2020/0004749 A1 Jan. 2, 2020

Related U.S. Application Data

(60) Provisional application No. 62/691,751, filed on Jun. 29, 2018.

(51) **Int. Cl.**
G06F 16/2458 (2019.01)
G06F 16/2453 (2019.01)
(Continued)

(52) **U.S. Cl.**
CPC ***G06F 16/24539*** (2019.01); ***G06F 16/2282*** (2019.01); ***G06F 16/2462*** (2019.01); ***G06N 5/003*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 16/24539; G06F 16/2282; G06F 16/2462; G06N 5/003
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,671,772 B1 12/2003 Cousins et al.
8,266,147 B2 9/2012 Slezak et al.
(Continued)

FOREIGN PATENT DOCUMENTS

WO 2008034219 A1 3/2008

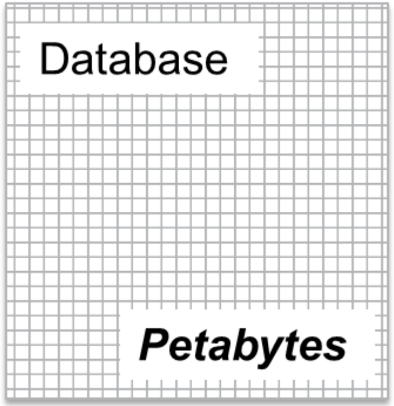
OTHER PUBLICATIONS

International Preliminary Report on Patentability for International Patent Application No. PCT/CA2007/001627 dated Jan. 7, 2008.
(Continued)

Primary Examiner — Kris E Mackes
(74) *Attorney, Agent, or Firm* — Smith, Gambrell & Russell LLP

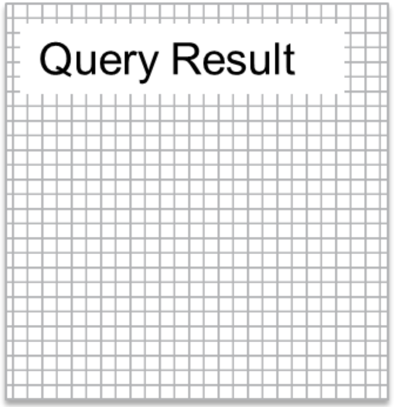
(57) **ABSTRACT**

Embodiments may provide methods and systems for intelligent capture and fast transformation of granulated data summaries. An engine may be used to transform input data summaries into result sets representing query outcomes. The data summaries contain enough knowledge about the original data to accurately perform operations on the summaries without needing to access the original data. In an embodiment, the contents of data summaries are accessible via an SQL approximate engine which retrieves summaries stored on disk and utilizes them for its operations. Alternatively, the contents of data summaries are accessible via virtual tables which give users direct access to the summary contents and
(Continued)



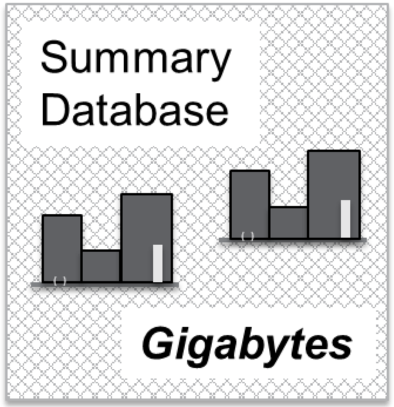
Traditional Query Execution:

- long time to do computations
- lots of disk/memory/processing resources required
- hard to manage in data lake / data cloud environments



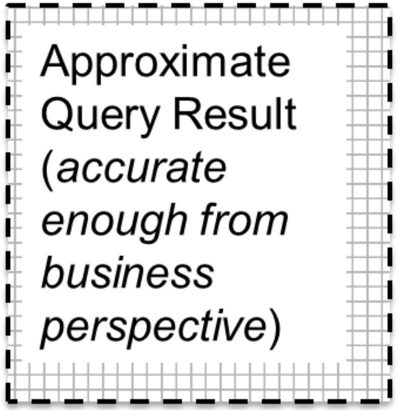
INFOBRIGHT

SECURITY ON-DEMAND



Querying on Data Summaries:

- orders of magnitude faster (*original operations replaced by fast summary transformations*)
- far less resources consumed
- original data remaining in-place



US011301467B2

(12) **United States Patent**
Slezak et al.

(10) **Patent No.:** **US 11,301,467 B2**
(45) **Date of Patent:** **Apr. 12, 2022**

(54) **SYSTEMS AND METHODS FOR INTELLIGENT CAPTURE AND FAST TRANSFORMATIONS OF GRANULATED DATA SUMMARIES IN DATABASE ENGINES**

(71) Applicant: **Security On-Demand, Inc.**, San Diego, CA (US)

(72) Inventors: **Dominik Slezak**, Warsaw (PL); **Richard Glick**, Valley Center, CA (US); **Pawel Betlinski**, Warsaw (PL); **Piotr Synak**, Winterthur (CH); **Jakub Wroblewski**, Lomianki (PL); **Agnieszka Chadzynska-Krasowska**, Sulejowek (PL); **Janusz Borkowski**, Warsaw (PL); **Arkadiusz Wojna**, Warsaw (PL); **Joel Alan Holland**, Encinitas, CA (US)

(73) Assignee: **Security On-Demand, Inc.**, San Diego, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 463 days.

(21) Appl. No.: **16/459,274**

(22) Filed: **Jul. 1, 2019**

(65) **Prior Publication Data**

US 2020/0004749 A1 Jan. 2, 2020

Related U.S. Application Data

(60) Provisional application No. 62/691,751, filed on Jun. 29, 2018.

(51) **Int. Cl.**
G06F 16/2458 (2019.01)
G06F 16/2453 (2019.01)
(Continued)

(52) **U.S. Cl.**
CPC ***G06F 16/24539*** (2019.01); ***G06F 16/2282*** (2019.01); ***G06F 16/2462*** (2019.01); ***G06N 5/003*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 16/24539; G06F 16/2282; G06F 16/2462; G06N 5/003
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,671,772 B1 12/2003 Cousins et al.
8,266,147 B2 9/2012 Slezak et al.
(Continued)

FOREIGN PATENT DOCUMENTS

WO 2008034219 A1 3/2008

OTHER PUBLICATIONS

International Preliminary Report on Patentability for International Patent Application No. PCT/CA2007/001627 dated Jan. 7, 2008.
(Continued)

Primary Examiner — Kris E Mackes
(74) *Attorney, Agent, or Firm* — Smith, Gambrell & Russell LLP

(57) **ABSTRACT**

Embodiments may provide methods and systems for intelligent capture and fast transformation of granulated data summaries. An engine may be used to transform input data summaries into result sets representing query outcomes. The data summaries contain enough knowledge about the original data to accurately perform operations on the summaries without needing to access the original data. In an embodiment, the contents of data summaries are accessible via an SQL approximate engine which retrieves summaries stored on disk and utilizes them for its operations. Alternatively, the contents of data summaries are accessible via virtual tables which give users direct access to the summary contents and
(Continued)

M Bartoszuk, J Litwin, M Wnuk, D Ślęzak:
Tensor-based Approach to Big Data Processing and Machine Learning.
IEEE BigData 2022