A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance

 Presented by Youcef REMIL
 Proposed by Infologic R&D
 Advised by Pr. Jean-François BOULICAUT
 Dr. Mehdi KAYTOUE
 Dr. Anes BENDIMERAD

June 04, 2024
A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance

Introduction and Motivation – CIFRE Thesis

1982 - Infologic foundation

2016 – Infologic R&D Initiated

2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

- Significant annual growth
- More than 600 sites
- Over 200K workstations

ERP Software Editor Copilote

- 1982 - Infologic foundation
- 2016 – Infologic R&D Initiated
- 2019 – Preventive Maintenance Project
- 2020 – Ph.D Thesis on AIOps

- Significant annual growth
- More than 600 sites
- Over 200K workstations
Introduction and Motivation – CIFRE Thesis

1982 – Infologic foundation

2016 – Infologic R&D Initiated

2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

Collect and storage of telemetry data

- Boosting efficiency/reliability
- Service quality
- Need for automation

A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance
Introduction and Motivation – CIFRE Thesis

1982 - Infologic foundation

2016 – Infologic R&D Initiated

2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

- Data-centric approach
- Real-time monitoring
- Proactive maintenance
- AIOps* and Automation

Introduction and Motivation – CIFRE Thesis

1982 – Infologic foundation

2016 – Infologic R&D Initiated

2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

- Study of AIOPS field
- Limitations of AIOps
- Development of effective AIOps solutions
- Applicative and Research Contributions

Introduction and Motivation

- **Real pain points of maintenance routines at Infologic**
  - Lack of standardized and automated maintenance routines with higher costs
    - Relying mostly on corrective maintenance
  - Example: A detectable memory leak at a customer’s premises (with +€450m annual revenue) blocked the departure of all delivery trucks from a factory for 30 minutes.
Introduction and Motivation

- Real pain points of maintenance routines at Infologic
  - Lack of standardized and automated maintenance routines with higher costs
    - Higher human and resource costs [statistics by the end of 2019]

<table>
<thead>
<tr>
<th>Code</th>
<th>Libellé</th>
<th>07/19-2000-00:00</th>
<th>07/19-2000-00:00</th>
<th>07/19-2000-00:00</th>
<th>07/19-2000-00:00</th>
<th>07/19-2000-00:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>448719</td>
<td></td>
<td>3h 41m 3s 59h</td>
<td>3h 41m 3s 59h</td>
<td>1h 37m 29s</td>
<td>1h 37m 29s</td>
<td>0.62</td>
</tr>
<tr>
<td>504120</td>
<td></td>
<td>1h 23m 43s</td>
<td>1h 23m 43s</td>
<td>1h 23m 43s</td>
<td>1h 23m 43s</td>
<td>0.74</td>
</tr>
<tr>
<td>100249</td>
<td></td>
<td>1h 34m 1s</td>
<td>1h 34m 1s</td>
<td>1h 34m 1s</td>
<td>1h 34m 1s</td>
<td>0.64</td>
</tr>
<tr>
<td>330110</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>091225</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>091790</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>454270</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>001000</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>840130</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>330557</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>554020</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>724249</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>040334</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>542802</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>200179</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
<tr>
<td>2347301</td>
<td></td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>1h 32m 9s</td>
<td>0.59</td>
</tr>
</tbody>
</table>

- ~5600 days
- ~28 pers
- ~14 %

Maintenance time
Full-time employees
Workforce Percentage
Introduction and Motivation

Real pain points of maintenance routines at Infologic

- **Inefficient incident triage and classification**
  - Need for automatic assigning, ranking and classification
  - Problem of tossing sequence*
  - Presence of recurring similar issues in historical maintenance calls

- **Ineffective root cause analysis and incident correlation**
  - Need for deep fault localization and figure out dependencies among components and services

*Xie et al., Bug Triaging Based on Tossing Sequence Modeling. In Journal of Computer Science and Technology 2019*
Introduction and Motivation

Capabilities of AI for Operating Systems (AIOps)

- **Prevention**
  - Forecast high-severity outages, future events, alerting signals, assessing system health

- **Detection**
  - Detect abnormal conditions, automated pattern discovery, noise reduction in data

- **Location**
  - Root cause analysis, recurrent issues identification, unified topology and contextualization

- **Perception**
  - Data collection and ingestion, data storage, real-time monitoring, querying data

- **Interaction**
  - Human-computer intelligent interaction, interactive analysis and collaboration

- **Action**
  - Reactive triage and routing, prioritization of incidents, set of remediation actions

† Remil et al. AIOps Solutions for Incident Management: Technical Guidelines and A Comprehensive Literature Review, In TOSEM 2023 [Under Submission]
* Dang et al. AIOps: real-world challenges and research innovations. In ICSE 2019
Introduction and Motivation

- Research challenges of AIOps addressed in this thesis
Introduction and Motivation

- Research challenges of AIOps addressed in this thesis

Novel and Unstructured Field

AIOps lacks unified terminology, complete taxonomy, desiderata, technical details
Introduction and Motivation

- Research challenges of AIOps addressed in this thesis

**Novel and Unstructured Field**
- AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

**Data Requirements**
- Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data
Introduction and Motivation

- Research challenges of AIOps addressed in this thesis

**Novel and Unstructured Field**
- AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

**Model Design**
- Impractical supervised methods, overlooking descriptive models

**Data Requirements**
- Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data
Introduction and Motivation

Research challenges of AIOps addressed in this thesis

1. Novel and Unstructured Field
   - AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

2. Data Requirements
   - Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

3. Model Design
   - Impractical supervised methods, overlooking descriptive models

4. Interpretability
   - Best models are black box, transparency is preferred over performance
Introduction and Motivation

Research challenges of AIOps addressed in this thesis

1. Novel and Unstructured Field
   - AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

2. Data Requirements
   - Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

3. Model Design
   - Impractical supervised methods, overlooking descriptive models

4. Interpretability
   - Best models are black box, transparency is preferred over performance

5. Scalability and Robustness
   - Scalability assessment often overlooked, temporal and in-context evaluation
Introduction and Motivation

- Contributions and Key Research Areas

Software Eng.

Data Mining
Introduction and Motivation

- Contributions and Key Research Areas

  AI for Operating Systems

  Study Comprehensively AIOps research and technical area to address challenge 1

- Software Eng.

- Data Mining

---

Remil et al. AIOps Solutions for Incident Management: Technical Guidelines and A Comprehensive Literature Review. In TOSEM 2023 [Under revision, Core 2021, A*]

Bendimerad, Remil et al. On-premise Infrastructure for AIOps in a Software Editor SME: An Experience Report, In ESEC/FSE 2023 [Published, Core 2021, A*]
Introduction and Motivation

Contributions and Key Research Areas

- **AI for Operating Systems**
  - Study Comprehensively AIOps research and technical area to address challenge 1

- **Supervised Rule Discovery**
  - Study SD and EMM to address challenges 2 and 3 related to data quality and models

**Software Eng.**

**Data Mining**

---

Remil et al. What makes my queries slow: Subgroup Discovery for SQL Workload Analysis. In *ASE 2021 [Published, Core 2021, A]*
Remil et al. Interpretable Summaries of Black Box Incident Triaging with Subgroup Discovery, In *DSAA 2021 [Published, Core 2021, A]*
Remil et al. Mining Java Memory Errors using Subjective Interesting Subgroups with Hierarchical Targets, In *ICDMW 2023 [Published, Workshop]*
Introduction and Motivation

Contributions and Key Research Areas

1. **AI for Operating Systems**
   - Study Comprehensively AIOps research and technical area to address challenge 1

2. **Supervised Rule Discovery**
   - Study SD and EMM to address challenges 2 and 3 related to data quality and models

3. **Explainable AI**
   - Study Explainable AI to cope with challenge 4 of interpreting black box models

---

Remil et al. Interpretable Summaries of Black Box Incident Triaging with Subgroup Discovery, In *DSAA 2021* [Published, Core 2021, A]
Remil et al. Découverte de Sous-groupes Interprétables pour le Triage d’Incidents, In *EGC 2022* [Published, National Conf]
Introduction and Motivation

Contributions and Key Research Areas

- **AI for Operating Systems**
  - Study Comprehensively AIOps research and technical area to address challenge 1

- **Locality Sensitive Hashing**
  - Study LSH for fast and efficient similarity search to tackle challenge 5

- **Supervised Rule Discovery**
  - Study SD and EMM to address challenges 2 and 3 related to data quality and models

- **Explainable AI**
  - Study Explainable AI to cope with challenge 4 of interpreting black box models

---


A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance
Subgroup Discovery for SQL Workloads

Huge SQL Workload

- SQL queries
- Execution time
- # rows
- History sessions ...

Database schema

- Query execution plan...

DBAs

Performance analysis

- Slow queries
- Index recommendation
- Concurrency issues
- ...

Database server

A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance
Subgroup Discovery for SQL Workloads

Need for a generic framework to analyse batches of SQL queries and bring answers to the question: How to characterize SQL queries that foster some properties of interest?
Need for a **generic** framework to analyse **batches** of SQL queries and bring answers to the question: **How to characterize SQL queries that foster some properties of interest?**

**Illustrative example of SQL queries**

<table>
<thead>
<tr>
<th>Predicates</th>
<th>Topology</th>
<th>...</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>ik</td>
<td>date</td>
<td>...</td>
<td>time</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>V2</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>V1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>V2</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>V2</td>
<td>25</td>
</tr>
</tbody>
</table>

**Pattern P**: Predicate = \( \text{verrou.date} \land \text{Db. Version} = \text{V2} \) → slow queries

---

*Atzmueller. Subgroup Discovery, in DAMI 2015
Wrobel. An algorithm for multi-relational discovery of subgroups. In PKDD 1997*
Subgroup Discovery building blocks

- Dataset $\mathcal{D}$
- Property of interest
- Target $\mathcal{T}$
- Langage Pattern $\mathcal{L}$
- Enumerated subgroups
- Enumeration Algorithm
  - Exhaustive, heuristic, by sampling, etc.
- Selectors
- Top $k$ subgroups
- Objective, subjective, semantic-based, constraints, etc.
- Interestingness $Q$
- Symbol sets, numerical intervals, subgraphs, subsequences, etc.
- Numerical, binary, multiple complex attributes, etc.
Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis

Example of an SQL query

```
SELECT m.ik
FROM model AS m
JOIN prod AS p
WHERE m.ik = p.ik
  AND m.uex = p1
  AND (m.uex in collection0
       OR m.ik in collection1)
  AND (m.dossier = p3
GROUP BY m.ik
HAVING (COUNT(DISTINCT p.ik) = p2)
  AND (SUM(m.nbembal) = MAX (p.nbembal))
```

Our parser*

```
SELECT_model_ik 1
FROM_model 1
JOIN_prod 1
WHERE_model_ik 2
WHERE_model.uex 1
WHERE_model.dossier 1
WHERE_prod.ik 1
GROUPBY_ik 1
HAVING_prod_ik 1
HAVING_model.nbembal 1
HAVING_prod.nbembal 1
COUNT_prod.ik 1
SUM_model.nbembal 1
MAX_prod.nbembal 1
```

*https://github.com/klahnakoski/mo-sql-parsing/pull/26
Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis

SQL queries
Parsing
Relevant data

Dataset \( \mathcal{D} \)

Selectors

Langage Pattern \( \mathcal{L} \)

Symbol sets, numerical intervals,

Property of interest

DB version

DB version

Blocked sessions

\( V2 \)
\( V3 \)

\( 5 \ 10 \ 15 \ 20 \ 25 \)

\( sel_{1} \land sel_{2} \land \ldots \land sel_{d} \)

\( sel_{1} \rightarrow sel_{2} \rightarrow \ldots \rightarrow sel_{d} \)

\( sel_{1} \land sel_{2} \land \ldots \land sel_{d} \)

\( sel_{i} \land sel_{i+1} \land \ldots \land sel_{d} \)

\( \cup \)

Search Space \( \mathcal{L} \)

Made of

Selectors

Conjunctive combinations

Patterns

- \( sg(P) = ext(P) = \{ c \in \mathcal{O} | P(o) = True \} \)
- \( P_{gen} \subseteq P_{spec} \Rightarrow sg(P_{gen}) \supseteq sg(P_{spec}) \)

\( P_{gen} : blockedSessions \in [15, 25] \)
\( P_{spec} : blockedSessions \in [15, 25] \land dbVersion = V3 \)
Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis

- SQL queries → Parsing → Dataset $\mathcal{D}$ → Property of interest (numerical attributes, e.g., runtime) → Target $T$ → Interestingness $Q$

Objective measures:
- Exceptionality
- Generality

Numerical measures used to evaluate the subgroup patterns:

1. **Mean-based Measure** (sensible to outliers)
   \[
   q_{\text{mean}}^\alpha = i_p^\alpha \cdot (\mu_p - \mu_\emptyset)
   \]

2. **Median-based Measure** (sensible to outliers)
   \[
   q_{\text{med}}^\alpha = i_p^\alpha \cdot |\text{Med}_p - \text{Med}_\emptyset|
   \]

3. **T-Score Measure** (optimize the dispersion)
   \[
   T_{\text{score}} = i_p^{\frac{1}{2}} \cdot \frac{(\mu_p - \mu_\emptyset)}{\sigma_p}
   \]

*Langage Pattern $\mathcal{L}$

Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis

SQL queries
Parsing

Dataset $\mathcal{D}$

Target $T$

Interestingness $Q$

Property of interest
numeric and binary, attributes

Selector

Language Pattern $\mathcal{L}$

Evaluate subgroups

Enumeration Algorithm

Enumerate subgroups

Depth first search
Beam search

Refine

Pruning

Empty pattern

Search Space

Max support

Objective measures:
Exceptionality
Generality

Lemmerich et al. Fast exhaustive subgroup discovery with numerical target concepts. In DAMI 2016
Subgroup Discovery for SQL Workloads

Results on a large workload of Hibernate queries made available by Infologic

<table>
<thead>
<tr>
<th>ID</th>
<th>Target</th>
<th>Measure</th>
<th>Subgroup patterns</th>
<th>Size</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>time</td>
<td>Median</td>
<td>(P_1): WHERE stocks.gestion.modele.lot.prod.ref.auditinfo.etat \geq 1 (P_2): FROM ventes.cumuls.modele.cummulmultiple \geq 1 (P_3): WHERE ventes.cumuls.modele.cummulmultimate.valzvcl1X \geq 1 (P_4): WHERE ventes.cumuls.modele.cummulmultimate.valzvartX \geq 1</td>
<td>8</td>
<td>(161 \cdot q\text{med}(P_2))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(P_5): serverName = ServerX \land system/0 \geq 50</td>
<td>45</td>
<td>21 \cdot q\text{med}(P_2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lift</td>
<td>(P_6): GROUPBY stocks.gestion.modele.mvreallise.refexterne \geq 1</td>
<td>131</td>
<td>(\tau_P \approx 0.99)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(P_7): serverName = ServerX \land system/0 \geq 50</td>
<td>38</td>
<td>(\tau_P \approx 0.99)</td>
</tr>
<tr>
<td>D2</td>
<td>slow (\tau_P \approx 0.6)</td>
<td>Lift</td>
<td>(P_8): WHERE stocks.gestion.modele.mvreallise.etatsynchro \geq 1 \land jdbcMax &lt; 200</td>
<td>20668</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(P_9): WHERE stocks.gestion.modele.mvreallise.auditinfo.datcre \geq 1 \land dbVersion = 2.3</td>
<td>20675</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(P_10): manyActiveSessions = Alarm</td>
<td>44</td>
<td>0.93%</td>
</tr>
</tbody>
</table>
Enhancing Duplicate Crash Report Retrieval

Problem of deduplication*

A bug reported for the ACTI service

But the bug is generic and related to a web feature

*Jiang et al. Igor: Crash Deduplication Through Root-Cause Clustering, In CCS 2021
Enhancing Duplicate Crash Report Retrieval

**Similarity measures** for stack trace comparison embedded in **Clustering** algorithms

- **Complex** similarity measures based
- **Computational** Complexity is very **costly**
- Measures embedded in **clustering** with several **issues**
- It should be handled as **Nearest Neighbours Search** problem

[Dang et al., in ICSE 2012]*

*Dang et al. ReBucket: A Method for Clustering Duplicate Crash Reports Based on Call Stack Similarity. In ICSE 2012
†Wu et al. CrashLocator: Locating Crashing Faults Based on Crash Stacks. In ISSTA 2013
ǂMoroo et al. Reranking-based Crash Report Deduplication. SEKE 2017

---

A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance
Enhancing Duplicate Crash Report Retrieval

**Contribution**

Learn a family of hash functions with a constrained hashing Siamese neural network.

A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance
Enhancing Duplicate Crash Report Retrieval

**Experiments**

Does the model manage to converge to the LSH property?

![Diagram showing the model's convergence to the LSH property for various similarity measures.](image)
Enhancing Duplicate Crash Report Retrieval

Experiments

Is the model fast enough compared to linear scans?

<table>
<thead>
<tr>
<th>Similarity Measure</th>
<th>k-NN</th>
<th>CNNH+LSH</th>
<th>DeepLSH</th>
<th>MinHash</th>
<th>SimHash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>258</td>
<td>30</td>
<td>26</td>
<td>57</td>
<td>-</td>
</tr>
<tr>
<td>Cosine</td>
<td>8288</td>
<td>15</td>
<td>14</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>8510</td>
<td>16</td>
<td>15</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Edit Distance</td>
<td>4911</td>
<td>29</td>
<td>29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PDM</td>
<td>10047</td>
<td>16</td>
<td>16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brodie</td>
<td>Limit</td>
<td>27</td>
<td>27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DURFEX</td>
<td>12160</td>
<td>26</td>
<td>24</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lerch</td>
<td>3118</td>
<td>24</td>
<td>24</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Moroo</td>
<td>15253</td>
<td>25</td>
<td>25</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TraceSim</td>
<td>13050</td>
<td>30</td>
<td>30</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
End

Thanks for your attention
End

Q/A?