Data-Mining for analysis of semi-structured messages from application execution

Oihana Coustié

Data-Mining for analysis of semi-structured messages from application execution

Oihana Coustié
Pursuing a PhD in Computer Science
With the collaboration of:
Airbus Operations
Supervisor : Xavier Baril
and the Université Toulouse I
Supervisors: Professor Josiane Mothe
Professor Olivier Teste

Context

During their execution, information system applications generate a large amount of messages, in textual format, which can be considered as semi-structured data.

Industrial Challenge

#1 : When an error message is found, support teams are asked to provide precise explanation -> manually perform root cause analysis(not trivial, due to high interference between applications)

#2 : Each time the information system is updated -> test new version -> generation of test logs. Integration teams are interested in knowing whether the new version reacts diffenrently to the same test session.

Thesis goal

Use data mining, especially on time series, to :

- automatically perform root cause analysis
- detect changes in test session reactions of the system.

Research question

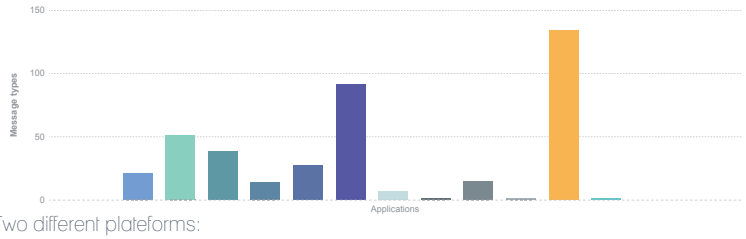
- Extract both numerical series and context information from semi-structured messages
- Apply rule discovery methods on time series or find new time-adapted algorithms (#1)
- Use similarity measure studies to compare two consecutive version of semi-structured dataset (#2)

Data presentation : key figures

350 000
logs each day
for each aircraft

30 000
logs each hour
when the system is on

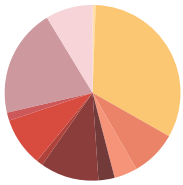
Number of message types for each application



Two different plateforms:

- Both receive messages either from their own services execution or from the applications they host.
- Each application or service can send up to 130 different types of messages : some of those types contain numerical values -> can be converted to time series

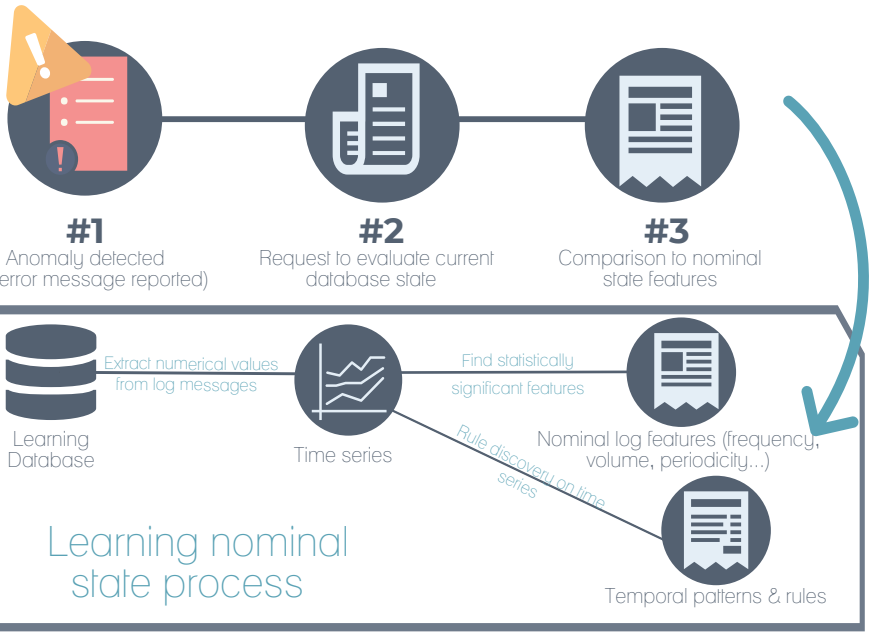
Number of messages recevied per hour application



High imbalance in term of message quantity among the 13 different applications hosted on the system:

- some only send 1 message per hour while others can send up to 7000 messages
- makes it difficult to produce a general automatised solution

Methodology



#1 Automated root cause analysis

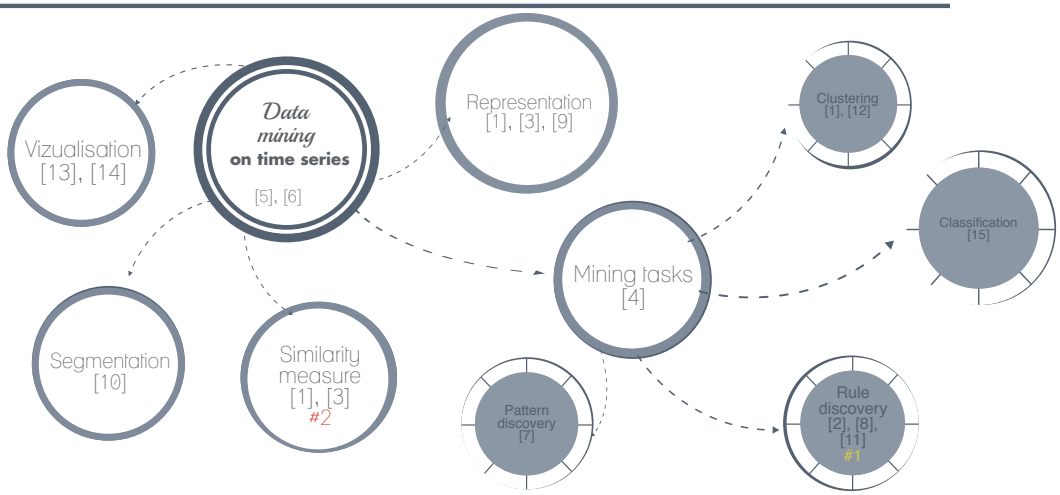
#2 New version changes analysis

#1
Information system version upgrade

#2
Test session (unchanged)

#3
Comparison with previous test logs (generated from the same test session on previous IS version)

State-of-the-art on time series



[1] Aghabazorgi et al., Information Systems (2015)

[2] Das et al., KDD (1998)

[3] Ding et al., Proceedings of the VLDB Endowment (2008)

[4] Esling and Agon, ACM Computing Surveys (CSUR) (2012)

[5] Fu, Engineering Applications of Artificial Intelligence (2011)

[6] Gaber et al., ACM Sigmod Record (2005)

[7] Han et al., IEEE (1999)

[8] Han et al., Proceedings of the 17th international conference on data engineering (2001)

[9] Keogh et al., Knowledge and information Systems (2001)

[10] Keogh et al., World Scientific (2004)

[11] Last et al., IEEE (2001)

[12] Liao, Pattern recognition (2005)

[13] Van Wijk and Van Selow, IEEE (1999)

[14] Weber et al., Infovis (2001)

[15] Xing et al., ACM Sigkdd Explorations Newsletter (2010)