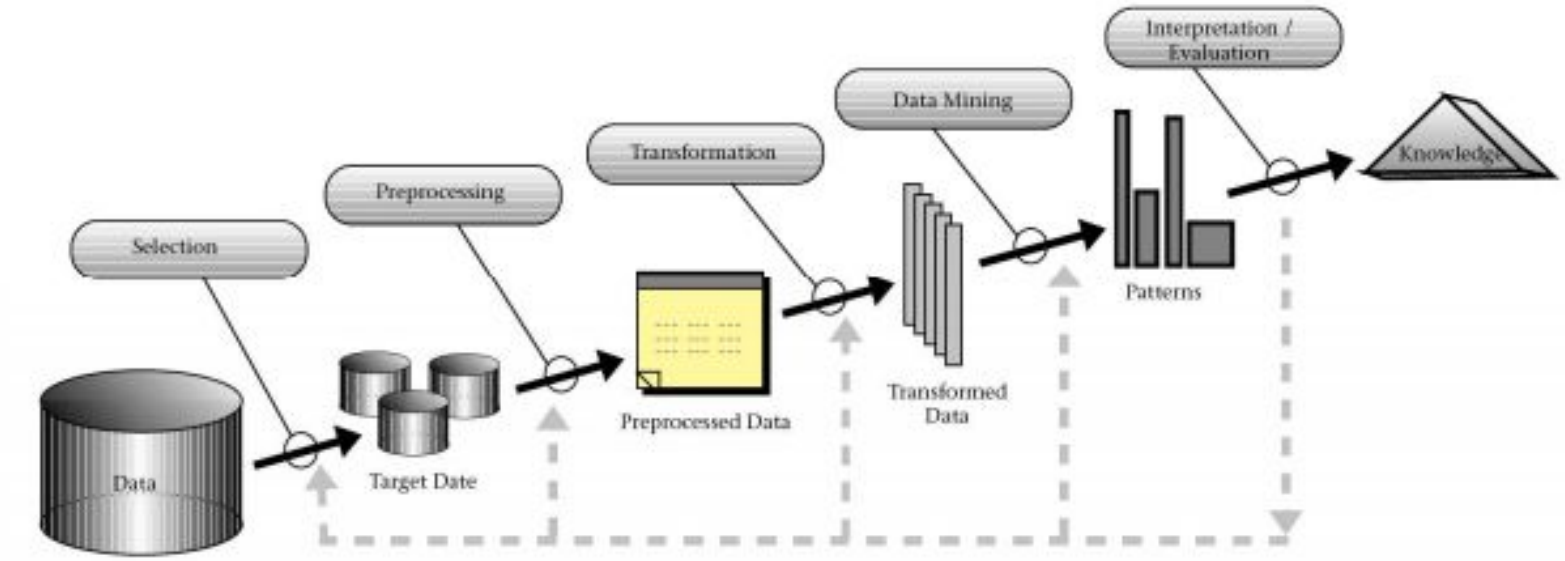
Universidade Federal de Santa Maria – UFSM
Centro de Tecnologia – CT
Curso de Engenharia de Computação
ELC1098 – Data Mining

Data Mining

Luis Felipe de Deus - felipe.deus@ecomp.ufsm.br;
Nathanael Luchetta - nathanael.luchetta@ecomp.ufsm.br;
Tiago Knorst - tiago.knorst@ecomp.ufsm.br;
Yuri Oliveira - yuri.alves@ecomp.ufsm.br.

Novembro/2019

KDD (Knowledge Discovery from Data)

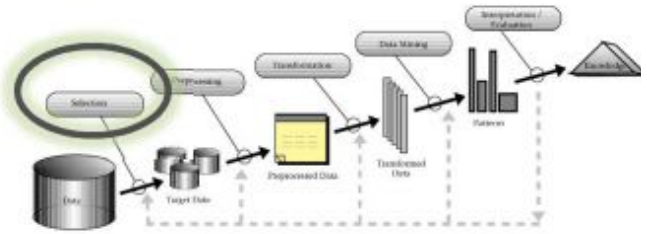


* Figura mostrando o processo de KDD, proposto por Fayyad, 1996.

Data Selection

Três *datasets* .csv:

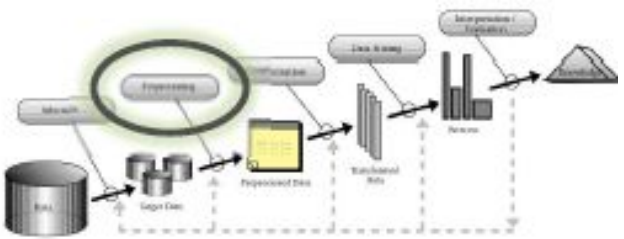
- 1º - ID do *paper*, ID da Conferência, Ano, *Status*;
- 2º - ID do *paper*, subtópico(s);
- 3º - ID do tópico e do subtópico, descrição do tópico e subtópico.



Preprocessing

Tratamento de inconsistências:

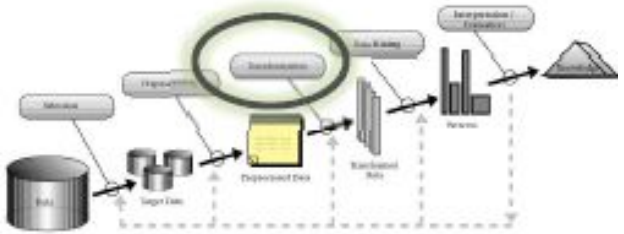
- Anos menores que 2010 removidos (Ex. 1,2,3,4);
- Status incompletos ou em outra idioma ajustados (Ex. acepto, aceito);
- Status incoerentes removidos (Ex. dog, **).



Transformation

Transformação do *dataset* para melhor manipulação dos dados:

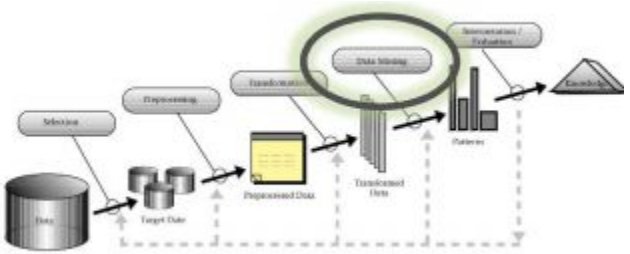
- União dos arquivos em um único *dataset*;
- *Papers* pertencentes a mais de um subtópico foram separados;
- Status *Rejected*/*Accepted* transformado em lógico 0/1;
- Criação de um *sub-dataset* a partir do original contendo apenas informações relevantes para a mineração.



Data Mining

- Utilizadas regras de associação;

- Algoritmo Apriori.



Interpretation

- Baseado no plano de KDD elaborado, é possível o teste de hipóteses para obter-se informação.
- H1-Subáreas com maior e menor taxa de rejeição, há algum padrão ou anomalia ?
- H2-Taxa de rejeição por ano, há algum padrão ou anomalia ?
- H3-Nº de Conferências realizadas por ano ?
- H4-Conferências com maior e menor taxa de rejeição, há algum padrão ou anomalia ?
- H5-Taxa de submissões por conferências ?

Subtópicos com maior índice de aceitação

Subtópicos com maior índice de aceitação		Histórico no tempo							
		2010	2011	2012	2013	2014	2015	2016	2017
Subtópicos	1°	51	17	9	14	27	14	3	33
	2°	16	52	27	10	37	10	31	12
	3°	52	24	16	56	53	47	14	42
	4°	10	14	55	30	42	8	17	27

Tópicos com maior índice de aceitação

- Remapeando subtópicos para seus respectivos tópicos.

Tópicos com maior índice de aceitação		Histórico no tempo							
		2010	2011	2012	2013	2014	2015	2016	2017
Tópicos	1°	7	6	1	2	4	2	1	5
	2°	2	7	4	1	6	1	5	2
	3°	7	4	2	7	7	7	2	6
	4°	1	2	7	5	6	1	6	4

Tópicos com maior índice de aceitação

- Tópico 2 (*Service Management*), presente entre áreas com maior índice de aceitação, ausente somente em 2014;
- Tópico 7 (*Methods*), presente entre áreas com maior índice de aceitação, ausente somente nos anos de 2016 e 2017.

Tópicos com maior índice de aceitação		Histórico no tempo							
		2010	2011	2012	2013	2014	2015	2016	2017
Tópicos	1°	7	6	1	2	4	2	1	5
	2°	2	7	4	1	6	1	5	2
	3°	7	4	2	7	7	7	2	6
	4°	1	2	7	5	6	1	6	4

Subtópicos com menor índice de aceitação

Subtópicos com menor índice de aceitação		Histórico no tempo							
		2010	2011	2012	2013	2014	2015	2016	2017
Subtópicos	1°	20	50	34	54	40	20	20	48
	2°	18	38	24	5	20	38	48	43
	3°	43	46	45	3	15	37	11	29
	4°	45	15	7	43	11	44	24	35

Tópicos com menor índice de aceitação

- Tópico 1 (*Network Management*), presente entre as menores áreas aceitas em quase todos os anos;
- Em 2013 e 2015 houve uma grande rejeição dos tópicos 1 (*Network Management*) e 6 (*Technologies*).

Tópicos com menor índice de aceitação		Histórico no tempo							
		2010	2011	2012	2013	2014	2015	2016	2017
Tópicos	1°	3	7	5	7	6	3	3	7
	2°	5	6	4	1	3	6	7	1
	3°	1	7	6	1	2	6	1	5
	4°	6	2	1	1	1	6	4	5

Anomalias

- Pequena discrepância no percentual de aceitação no ano de 2016;
- Único ano com índice menor de 60% de aceitação.

Year	N° Submission	N° Accepted	Rate (%)
2010	956	577	60.3%
2011	577	368	63.7%
2012	1028	737	71.6%
2013	600	438	73.0%
2014	501	309	61.6%
2015	1102	755	68.5%
2016	725	416	57.3%
2017	411	286	69.5%

Anomalias

- Diferenças no Nº de submissões por Conf. realizadas no ano de 2015;
- Conferência 11 com discrepância no número de submissões;
- Anomalia de baixo índice de aceitação na Conf. 13 realizada em 2016.

Year	Conf.	Subm.	Accep.	Rate (%)
2010	1	601	349	58.0%
	2	355	228	64.2%
2011	3	327	229	67.9%
	4	250	139	55.6%
2012	5	647	457	70.6%
	6	381	280	73.4%
2013	7	600	438	73.0%
2014	8	501	309	61.6%
2015	9	704	523	74.2%
	10	318	178	55.9%
	11	80	54	67.5%
2016	12	485	322	66.3%
	13	240	94	39.1%
2017	14	411	286	69.5%

Knowledge

- Baseado nas informações obtidas, é possível reunir algum conhecimento e potencialmente inferir respostas;
- O tópico 2 esteve entre os mais aceitos exceto em 2014;
- Em 2014 pode ter acontecido uma outra conferência com Qualis maior do que os dados deste *dataset* e os autores optaram pela submissão na de melhor Qualis;
- Simplesmente os *papers* de tópico 2 submetidos em 2014 não eram tão bons.

Knowledge

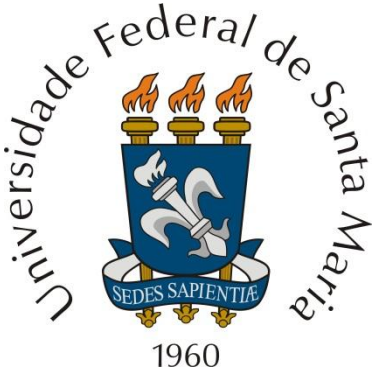
- Em 2015 houveram três conferências (9, 10, 11), foram submetidos respectivamente 704, 318, 80 *papers*.
- Possivelmente duas destas conferências tenham sido em datas próximas/concorrentes.
- Se o ID estiver em ordem cronológica no tempo de forma crescente, separando por exemplo em trimestres do ano, a que contém apenas 80 submissões foi no último trimestre do ano, logo época em que os recursos financeiros são mais escassos.

Conclusão

- As etapas de pré processamento e transformação dos dados são essenciais para uma boa mineração;
- *Datasets* com dados reais (*dataset's* grandes) possuem geralmente suporte baixo;
- Em geral, com dados reais a informação não está explícita, pois é previamente desconhecida, e não é de caráter óbvio.
- Cabe ao analista interpretar os resultados e transformar a informação em conhecimento.

Obrigado pela Atenção !

Perguntas?



Luis Felipe de Deus – felipe.deus@ecomp.ufsm.br;
Nathanael Luchetta – nathanael.luchetta@ecomp.ufsm.br;
Tiago Knorst - tiago.knorst@ecomp.ufsm.br;
Yuri Oliveira - yuri.alves@ecomp.ufsm.br.