Mineração de Dados - Trabalho Final

Luis Felipe de Deus[1], Nathanael Luchetta[1], Tiago Knorst[1], Yuri Oliveira[1]

[1]Curso de Engenharia de Computação – Universidade Federal de Santa Maria

{felipe.deus@ecomp.ufsm.br, nathanael.luchetta@ecomp.ufsm.br,
tiago.knorst@ecomp.ufsm.br, yuri.alves@ecomp.ufsm.br}

Este trabalho prático trata da análise de um conjunto de dados relacionados ao histórico de submissões de artigos em determinadas conferências, de porte dos dados relativos aos artigos como tópico, subtópico, conferência, ano de publicação, além da situação quanto à aceitação ou rejeição.

A análise dos dados tem por objetivo a descoberta de padrões ou anomalias entre os dados, assim como elencar uma das subáreas que, historicamente, têm maior chance de ser aceita.

Para atingir os objetivos citados, se faz necessário um plano de KDD (descoberta de conhecimento dos dados), onde um passo a passo é estabelecido desde a seleção e pré-processamento dos dados, até a mineração e avaliação dos resultados.

Os tópicos a seguir detalham como cada uma das cinco etapas do processo de KDD foi realizada, assim como a avaliação dos resultados obtidos ao final.

1. Seleção dos dados

Os dados brutos estavam distribuídos em três arquivos no formato tabular (.csv), sendo o 1° arquivo contendo o ID do artigo, ID da conferência, ano e *status* de aprovação, o 2° arquivo contendo ID do artigo e o código dos seus subtópicos, e o 3° arquivo com ID do artigo, código dos subtópicos, descrição do tópico e do subtópico.

Estes três arquivos foram unificados em um único *dataset* utilizando a linguagem R, contendo todas as informações contidas em cada um deles como demonstra a Fig. 1, de modo a facilitar as etapas seguintes do KDD.

	paper	conf	year	status	subtopics	topic	topic_desc	subtopic_desc
1	58571	1	2010	0	2	1	Network Management	Wireless & mobile networks
2	59379	1	2010	1	31	5	Management Approaches	Autonomic and self management
3	60650	1	2010	0	26	4	Functional Areas	Security management
4	60654	1	2010	0	55	7	Methods	Monitoring & Measurements
5	60679	1	2010	0	38	6	Technologies	Mobile agents
6	60748	1	2010	0	22	4	Functional Areas	Fault management
7	60806	1	2010	1	42	6	Technologies	Cloud computing
8	60830	1	2010	1	39	6	Technologies	P2P
9	61053	1	2010	0	32	5	Management Approaches	Policy-based management
10	61054	1	2010	0	36	6	Technologies	Protocols
11	61063	1	2010	0	55	7	Methods	Monitoring & Measurements
12	61088	1	2010	0	6	1	Network Management	Sensor Networks
13	61130	1	2010	0	37	6	Technologies	Middleware
14	61131	1	2010	1	22	4	Functional Areas	Fault management
15	61144	1	2010	1	52	7	Methods	Simulation
16	61151	1	2010	1	36	6	Technologies	Protocols
17	61174	1	2010	1	6	1	Network Management	Sensor Networks
18	61193	1	2010	0	26	4	Functional Areas	Security management
19	61420	1	2010	1	41	6	Technologies	Data, information, and semantic modeling
20	61432	1	2010	1	25	4	Functional Areas	Performance management

Figura 1 - Dataset agrupado.

2. Pré-processamento dos dados

Para a limpeza dos dados também foi elaborado um algoritmo na linguagem R, com objetivo de remover dados que possuíssem inconsistências.

Os dados que possuíam ano de publicação e *status* errôneos foram eliminados, bem como os que possuíam *status* similar ao correto, como por exemplo “aceito” ou “acepto” foram corrigidos, o que evita que dados incorretos sejam processados ao longo do processo de KDD.

Além da limpeza, também se fez uma “redisposição” dos subtópicos, já que se encontravam em uma única coluna separados por hífen e agrupados por tópicos, o que dificultava o desenvolvimento do processo de mineração. Neste caso a linha foi clonada e o subtópico foi dividido.

3. Transformação dos dados

O *status* dos artigos dado em *strings* de aceito ou rejeitado foi transformado operadores lógicos booleanos, seguindo a lógica de 0 (zero) para rejeitado e 1 (um) para aceito, de modo a se adequar aos métodos de mineração da linguagem R.

Por fim, um novo *dataset,* Fig. 2, foi criado, contendo apenas informações relevantes para a aplicação de técnicas de mineração de dados, neste novo dataset foram removidos atributos descritivos como o ID do paper, descrição de tópico e subtópico. Portanto, o novo dataset passou a contar apenas com o *status*, tópico, subtópico, conferência e ano.

	status	topic	subtopic	conf	year
1	0	1	2	1	2010
2	1	5	31	1	2010
3	0	4	26	1	2010
4	0	7	55	1	2010
5	0	6	38	1	2010
6	0	4	22	1	2010
7	1	6	42	1	2010
8	1	6	39	1	2010
9	0	5	32	1	2010
10	0	6	36	1	2010
11	0	7	55	1	2010
12	0	1	6	1	2010
13	0	6	37	1	2010
14	1	4	22	1	2010
15	1	7	52	1	2010
16	1	6	36	1	2010
17	1	1	6	1	2010
18	0	4	26	1	2010
19	1	6	41	1	2010
20	1	4	25	1	2010

Figura 2 - Sub-Dataset para mineração.

4. Mineração dos dados

Regras de associação são utilizadas para a descoberta de elementos que comumente ocorrem em um determinado conjunto de dados. Sua premissa básica é encontrar elementos que implicam na presença de outros em uma mesma transação, e dessa forma evidenciar padrões no conjunto de dados.

O caso estudado neste trabalho tem como transação cada uma das submissões de artigos, onde cada uma das informações inerentes aos artigos equivale a um item presente na transação.

O algoritmo de associação utilizado foi o Apriori, devido à sua característica associativa e de relativo fácil acesso e compreensão de funcionamento.

Nota-se que como o dataset é relativamente grande e diversificado, foi necessário a utilização de um suporte baixo, pois o suporte nada mais é que a quantidade mínima de vezes em que um determinado elemento aparece no dataset para que possa ser considerado nas regras de associação.

5. Padrões e interpretação dos resultados

Para a avaliação dos resultados foram analisadas questões como quantidade de submissões frente à quantidade de artigos aceitos, além da detecção de *outliers*, como por exemplo uma queda abrupta no número de submissões em um dado ano

Em primeiro momento foi realizada a busca dos quatro subtópicos mais aceitos ao longo do tempo, onde os que mais se destacaram são demonstrados na tabela 1.

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

Tabela 1 - Subtópicos com maior índice de aceitação.

Tendo em vista que através do número do subtópico não foi possível a detecção de padrões ou a descoberta de informação, foi aumentado a granularidade, mapeando cada subtópico para seu respectivo tópico como demonstra a tabela 2. Assim, descobrindo quais seriam mais propícios a serem aceitos e evidenciando certos padrões até então ocultos, como segue na tabela 2.

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

Tabela 2 - Tópicos com maior índice de aceitação.

Analisando a tabela 2, é possível concluir que os tópicos 2 (*Service Management)* e 7 (*Methods)*, são os que aparecem com maior frequência ao longo do tempo, pois o tópico 2 é ausente somente no ano de 2014, e 7 nos anos de 2016 e 2017, o que por consequência leva estes a obter um maior nível de aceitação.

Além disso, ocorreu a investigação de quais seriam os subtópicos com menor índice de aceitação, como é possível ver na tabela 3.

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

Tabela 3 - Subtópicos com menor índice de aceitação.

Do mesmo modo em que a tabela 1 foi remapeada, cada subtópico da tabela 3, foi disposto em relação a seu respectivo tópico, descobrindo quais seriam menos propícios a serem aceitos, como segue na tabela 4.

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

Tabela 4 - Tópicos com menor índice de aceitação.

Analisando a tabela 4, é possível concluir que os tópicos 1 (*Network Management)* e 6 (*Technologies)*, são os que aparecem com maior frequência ao longo do tempo, pois o tópico 1 é ausente somente no ano de 2011, e o 6 nos anos de 2016 e 2017, o que por consequência leva estes a obter um maior nível de rejeição.

Após buscar por padrões, ocorreu uma investigação por anomalias nos dados, podendo destacar-se o alto índice de rejeição dos *papers* cujas subáreas pertencem aos tópicos 1 e 6 respectivamente nos anos de 2013 e 2015, como demonstra a Tabela 4.

Por outro lado, foi encontrado uma pequena discrepância no nível de aceitação de papers por ano, sendo 2016, o único ano em que a relação submissão *vs* aceitação, foi abaixo de 60%, como segue na tabela 5.

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%

Tabela 5 - *Rate* de aceitação por ano.

Outras ocorrências de *outliers* ocorreram quando analisados os números de submissões x aceitações por conferência. É possível visualizar que no ano de 2015 existiram três conferências, associado a esta informação está a conferência de número 11, onde houve o menor índice de submissões (80), enquanto para as demais o menor valor foi de 240, como pode ser visto na tabela 6.

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%

Tabela 6 - *Rate* de aceitação por ano e conferência.

Também, analisando a tabela 6, é possível verificar que na conferência número 13 do ano de 2016, o nível de aceitação em relação ao de submissão é muito baixo, em relação a todas as outras conferências, sendo de apenas 39%. Isto significa que, menos da metade dos *papers* submetidos na conferência 13 realizada em 2013 foram aceitos.

6. Conhecimento obtido através das informações mineradas

Em síntese, baseando-se nas informações que foram obtidas através da mineração dos dados, foi possível inferir algumas noções sobre os datasets disponibilizados.

De modo que o tópico 2, esteve entre os mais aceitos, exceto em 2014. Além de que em 2014 pode ter acontecido uma outra conferência com Qualis superior do que os dados deste *dataset*, e então os autores optaram por submeter seus trabalhos na conferência de melhor Qualis. Bem como que em 2014 os trabalhos submetidos, de tópico 2 não eram tão bons quanto os seus predecessores.

No ano de 2015, ocorreram três conferências (9,10,11), onde foram submetidos 704, 318 e 80 *papers* respectivamente, o que leva a pensar que possivelmente as conferências 9 e 10 ocorreram em datas próximas ou concorrentes.

Ainda com relação ao ano de 2015, outra hipótese, é de que levando em consideração que o ID da conferência, tem relação com a data em que ocorreu a conferência, provavelmente a conferência de número 11, tenha ocorrido ao final do ano, logo na época em que os recursos financeiros são mais escassos.

7. Conclusões

Através da utilização do KDD, foi possível constatar que etapas iniciais, tais como pré-processamento e transformação dos dados tornam-se essenciais para uma boa mineração, visto também que *datasets* com dados reais (número volumoso de dados), possuem um suporte baixo quando aplicado o algoritmo Apriori.

Ainda, nota-se o quão importante é analisar os dados em granularidades distintas, visto que em alguns casos, a pode-se inferir resultados diferentes sobre o mesmo elemento ao olhar de outro ponto de vista.

Portanto, geralmente com dados reais, a informação não se encontra explícita, pois não possui características óbvias e bem definidas. Então cabe ao analista de dados interpretar os resultados e transformar a informação em conhecimento, expondo a saída de dados, de um modo claro o suficiente para que outras pessoas também consigam obter algum conhecimento sobre eles.