AGRUPAMENTO DE TEMAS NO LEVANTAMENTO BIBLIOGRAFICO SOBRE MINERAÇÃO DE PROCESSOS

Ferreira, Mirley Bitencourt
Borges, Felipe
Kohler , Manoella

Curso de Pós Graduação BI Master

Pontificia Universidade Católica,
Rio de Janeiro, Brasil.

---

1. Introdução

Para iniciar pesquisas científicas ou até mesmo em busca de novas tecnologias para serem aplicadas em organizações públicas ou privadas, se faz necessário o levantamento de trabalhos relacionados sobre o tema a ser explorado. Os levantamentos feitos em sites de busca ou em bases científicas muitas vezes retornam milhares de artigos o que fica inviável para que se ler e selecionar aqueles que tratam do assunto específico a ser estudado. Muitas exportações dessas bases de artigos retornam informações como área do estudo e até as palavras chaves, porém muitas vezes são vagas, e não se conhece rapidamente o agrupamento de temas.

A fim de auxiliar a autora em sua pesquisa sobre “mineração de processo” e descobrir os assuntos que circundam esta subárea da Ciência de Dados, este trabalho tem como objetivo explorar os artigos que estão sendo desenvolvidos e aplicados para auxiliá-la em sua revisão bibliográfica e trabalhos relacionados. E um objetivo mais específico é encontrar os artigos relacionados à descoberta de estruturas organizacionais com técnicas de mineração de processos.

Para isso utilizou-se de Mineração de Textos com a atividade de agrupamento (clusterização) a fim de encontrar artigo relacionados por assuntos. Esta atividade é do tipo não supervisionado. Uma aplicação dos algoritmos k-means e cluster hierárquico para agrupamento e também TSNE que auxiliou principalemente na visualização 2D destes agrupamentos. Na figura 1 são apresentadas as etapas desde a coleta até a avaliação dos grupos encontrados.

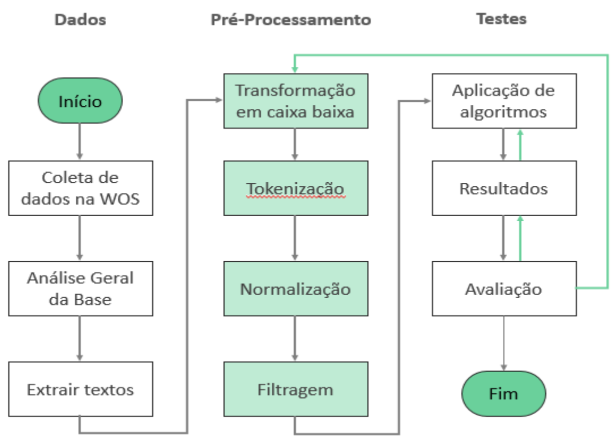

Figura 1 – Etapas de elaboração do estudo

O software utilizado foi o Orange, sendo gratuito e com todas as funcionalidades disponíveis para aplicação direta das etapas deste trabalho. E sua forma visual de se parametrizar os algoritmos facilitou muito.A base de dados foi extraída através de uma pesquisa na base *Web of Science,* acessada através do Portal de Periódicos da CAPES e tendo como termo de pesquisa “process mining” no campo tópico. Esta Consulta foi realizada em 31 de outubro de 2020 e retornou 1.969 documentos científicos abrangendo textos de todo o mundo, desde o ano de 1999.

Após esse tratamento, na lista de 1969 documento criou-se 220.846 tokens 1 a gerou-se a nuvem de palavras conforme figura 2. E nesta figura percebe-se um destaque em “data, model, system, log” que é a base de uma mineração de processos e ao seu redor todas as demais expressões em destaque.

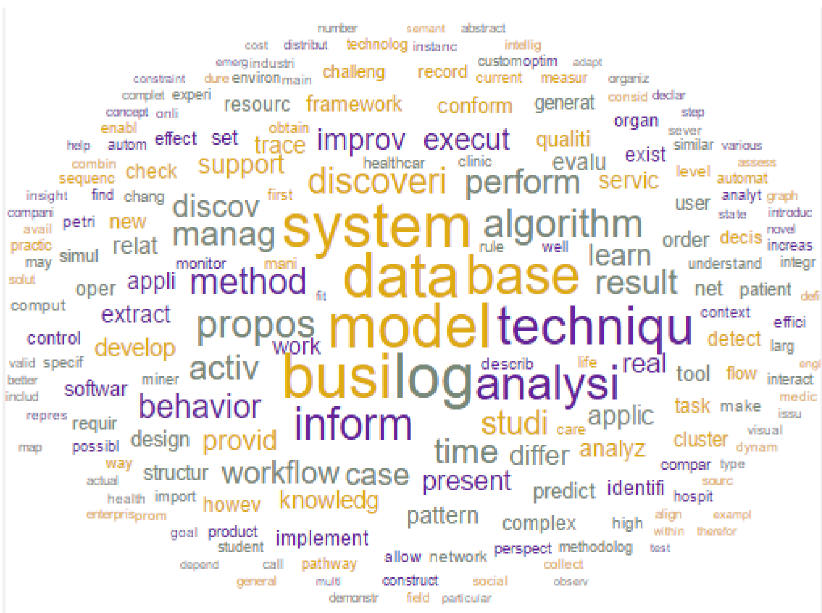

Figura 2 – Nuvem de palavra após pré-processamento de textos e sem mais frequentes

Para aplicação dos algoritmos de *clusterização* e análise dos agrupamentos formados no Orange, foram seguidas as etapas conforme a figura 3, disponibilizada diretamente do software escolhido.

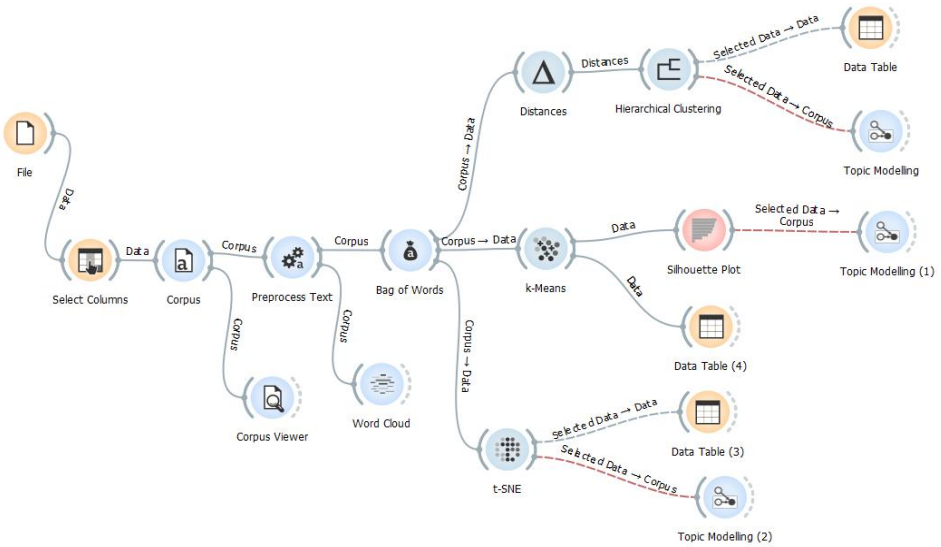

Figura 3 – Etapas no Orange

2. Resultados

Percebe-se que na tabela 1 com 875 documentos, ou seja 44% do total, encontra-se no cluster 14. E outros 491, 25% do total, encontram-se no cluster C7. Além disso, os grupos C3, C4, C5 e C6 ainda tratam do tema de fluxo e descobrimento de processo totalizando outros 131. Um grupo em destaque para o tema de mineração de processo com estudos de caso na área da saúde é o C1 com 144 artigos. Já voltado para o aprendizado está o grupo C2. No cluster C8 com 42 artigos focados em predição. O C9 e C10 são voltados para desenvolvimento de software e simulações. Tratando de conformidade no cluster 12. Focados em estudos de casos em serviços cluster 13. E, focado em recursos e rede social o cluster c15.

Tabela 1 – Tópicos de cada cluster agrupados pelo cluster hierárquico da fase 3

Topic Fase 3	Cluster	Total doc.
patient, pathway,clinic,data, model, care, medic, healthcar, hospital, health	C1	144
learn, student, data, model, studi, regul, cour, self, activ, permof	C2	67
workflow, model, log, system, data, netm busi, algoritm, execut, inform	C3	45
declar, model, constraint, log, discoveri, techniqu, base, discov, set, check	C4	28

trace, cluster, model, log, techniq, discoveri, base, complex, result, real	C5	30
petri, model, repair, net, log, base, method, techiqu, choic, system	C6	28
model, log, busi, algotim, discoveri, data, techniq, base, system, discovery	C7	491
predict, model, busi, time, monitor, data, log, case, perfom, prospo	C8	42
software, develop, model, data, behavior, log, system, execut, analysi, tech	C9	45
simul, model, busi, system, agent, log, base, analysi, data, behavior	C10	45
detect, drift, concept, change, busi, time, analysi, data, stream, method	C11	13
conform, audit, compliance, busi, data, check, log, inform, system, rule	C12	37
service, custom, behavior, web, call, base, model, system, composit, journey	C13	34
model, data, busi, log, system, analys, base, technique, inform, propos	C14	875
resource, busi, network, organiz, alloc, model, log, social, data, inform	C15	45
Grand Total		**1969**

Na figura 4 tem-se o T-SNE com os grupos gerados pelo cluster hierárquico. Pode-se ver uma aproximação entre entre os clusters gerados também pelo TSNE.

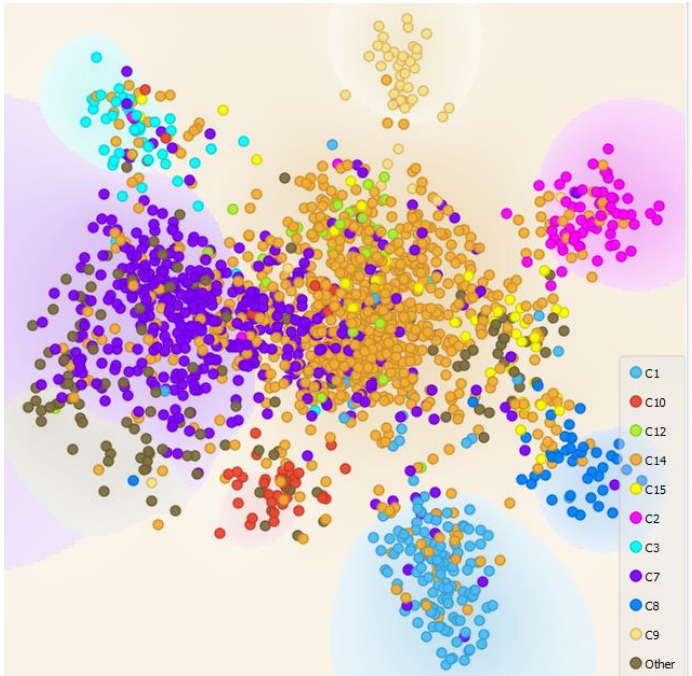

Figura 4 – Visão T-SNE com agrupamento do algoritmo do cluster hierárquico

Na tabela 2 tem-se a distribuição da quantidade dos clusters classificados pelo algoritmo k-means. Com o aumento de artigos na base o resultado não está satisfatório, pois em apenas um grupo foi inserido 1771 documentos.

Tabela 2 – Tópicos de cada cluster agrupados pelo cluster k-means

Topic	Cluster	Total doc.
busi, thrshold, reacl, preci, marbl, effect, studi, obtain, valu, retriev	C1	1
workflow, model, log, base, net, system, data, algoritim, busi, manag	C2	79
applic, helpdesk, acadam, model, activ, system, time, inform, gracia, find	C3	9
behavior, patterm, game, pm, social, player, dure, fucntion, naturalist, semi	C4	11
model, busi, log, behavior, data, sustem, base, inform, techniq, analysi	C5	1771
video, student, moos, difficult, cluster, propos, base, import, knowlog, leve	C6	1
lake, sediment, factor, anthropogen, nin sourc, artici, collect	C7	2
software, model, log, system, develop, product, algoritm, data, inform, behavior,	C8	62
malwar, applic, famil, detect, malici, mobil, phylogeni, sutdi, obtain, becom,	C9	5
lear, student, regul, base, course, self, activ	C10	1
data, patient, heath, care, heathcar, hospital, model, analysi, studi, base	C11	1
cloud, hybrid, schedul, task, compliance, busi, rule, comput, system, challeng,	C12	2
queu, predict, base, time, predictor, wait, technique, inform, variable	C13	1
depoist, document, year, reserv, copper, leas, are, lose, ore, extract	C14	1
patient, pathway, clinic, model, data, medic, topic, method, treatment, log	C15	22
Grand Total		**1969**

E também foi gerada a aplicação deste TSNE, figura 5, com os grupos do K-Means ficou com esta distribuição.

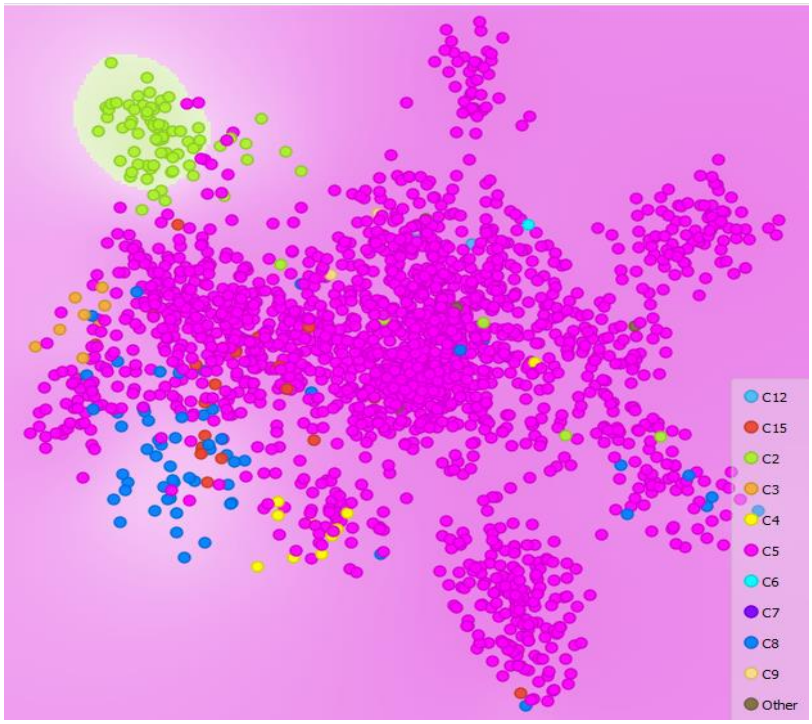

Figura 5 – Visão T-SNE com agrupamento do algoritmo do cluster k-means

O k-means por ser mais simples, não ajudou muito pois é preciso encontrar melhor os grupos de cada assunto ou temas que têm sido publicados.

Discussão

Foi feito uma análise pontual e manual com uma amostra escolhida aleatoriamente de 240 (para um grau de confiabilidade de 95%) conforme apresentado na tabela 9. A coluna “NOK”, onde detectou-se que um artigo poderia estar melhor classificado em algum outro grupo criado. Já o “OK” diz respeito ao agrupamento aceitável. Observou-se que quando um documento está classificado em um dos grupos com menos quantidade de artigos, na sua grande maioria, estes estão classificados corretamente. Observou-se também que estes mesmo podem ter uma duplicidade de temas podendo ser sobre o tópico que trata de "resource" e/ou "heath", podendo ser do cluster C15 ou C1. Já quando o artigo foi classificado pelo algoritmo de *clusterização* para ficar no cluster maior C14 ou no C7, temos falhas de classificação que poderiam ser melhor alocadas num dos outros clusters. Mas mesmo nestes clusters foi possível encontrar classificação correta ou até duplicidade de temas também. Foi feita uma avaliação entre os clusters menores e há percentual alto de aderência perante o esperado, porém, os clusters maiores tem uma aderência menor pelos motivos apresentados. Com esta análise pode-se afirmar que a classificação está aderente em 74% dos casos.

Cluter e seus tópicos	NOK	OK	% OK
C1 - patient, pathway,clinic,data, model, care, medic, healthcar, hospital, health		18	**100%**
C2 - learn, student, data, model, studi, regul, cour, self, activ, permof		8	**100%**
C3 - workflow, model, log, system, data, netm busi, algoritm, execut, inform	2	3	**60%**
C4 - declar, model, constraint, log, discoveri, techniqu, base, discov, set, check		5	**100%**
C5 - trace, cluster, model, log, techniq, discoveri, base, complex, result, real		2	**100%**
C6 - petri, model, repair, net, log, base, method, techiqu, choic, system		3	**100%**
C7 - model, log, busi, algotim, discoveri, data, techniq, base, system, discovery	17	41	**71%**
C8 - predict, model, busi, time, monitor, data, log, case, perfom, prospo		3	**100%**
C9 - software, develop, model, data, behavior, log, system, execut, analysi, tech	1	3	**75%**
C10 - simul, model, busi, system, agent, log, base, analysi, data, behavior		5	**100%**
C11 - detect, drift, concept, change, busi, time, analysi, data, stream, method		3	**100%**
C12 - conform, audit, compliance, busi, data, check, log, inform, system, rule		3	**100%**
C13 - service, custom, behavior, web, call, base, model, system, composit, journey		1	**100%**
C14 - model, data, busi, log, system, analys, base, technique, inform, propos	43	74	**63%**
C15 - resource, busi, network, prganiz, alloc, model, log, social, data, inform		5	**100%**
Total	**63**	**177**	**74%**

Conclusão

O estudo mostrou um levantamento bibliométrico do tema de Mineração de Processos, através de aplicações das técnicas de mineração de texto. Tratou-se de um estudo baseado em 1969 artigos encontrados na WOS de modo a identificar padrões referentes aos seus assuntos abordados. Especificamente, buscou-se identificar os grupos de artigos com a técnica de agrupamento de temas dos artigos por mineração de texto.

Há um crescente interesse sobre mineração de processos e que estes encontramos 15 clusters os quais, na maioria dos levantamentos, há 2 grandes grupos que tratam de descobrimento e mapeamento que é a grande base de descobrimento de processo. Apresentou também essas grandes áreas em tipos de abrangência de mineração de processos e suas perspectivas. O objetivo específico também foi atingido ao encontrar 22 artigos que tratam especificamente deste sub tema em mineração de processo. Dentre os grupos criados houveram temas que representam cada um destes. Porém destacou-se um grupo de temas bem específicos como “learn” que são mineração de processos para análise e/ou melhoria de aprendizado de alunos. Essa foi uma descoberta nova para a autora. É de interesse da autora aprofundar em estudos sobre recursos de um processo, e com este levantamento há um cluster específico sobre este tema, o qual facilitará o seu primeiro levantamento sobre este tema.

Conforme apresentado na estrutura de trabalho, houveram vários testes e avaliações dos algoritmos até que chegassem num modelo com melhor e maior clusterização possível. É comum em avaliações de clusters estes vários testes

e parâmetros de aceitação para o clusters, pois não há o que comparar para evidenciar alguma métrica. Observações que a atualização de parâmetro influencia muito nos agrupamentos. Assim como a quantidade de clusters selecionados, como foi o caso do k-means na fase 3, que ao aumentar o número de clusters ele generalizou e a maioria dos artigos ficou em apenas um cluster.

Para projetos futuros é possível criar métricas e usar algoritmos mais avançados para melhorar o agrupamento atual e buscar minimizar e enquadrar melhor os grandes grupos. E assim também, entender o comportamento deste algoritmo em quantidade de cluster diferente. Pretende-se também entender o porquê dessa não classificação "correta" de certos textos que poderiam estar em cluster mais específicos.