Instituto Federal da Paraíba - Programa de Pós Graduação em Tecnologia da Informação;

Relatório Técnico

João Pessoa, 15 de Janeiro de 2020.

Clusterização na indicação de melhor avaliador para resumo de Iniciação Científica

Thiago de Abreu Lima, Damires Yluska Souza,

Instituto Federal da Paraíba (IFPB) - João Pessoa, PB

Na Universidade Federal da Paraíba, todo ano, ocorre o Encontro de Iniciação Científica, onde os resumos dos trabalhos realizados pelos discentes são distribuídos para avaliação dos professores, que podem sugerir mudanças ou rejeitar/aceitar sumariamente o resumo para apresentação. A CAPES define a organização das áreas de conhecimento em quatro níveis de hierarquização. A distribuição atual leva em consideração as áreas cadastradas nos resumos pelos discentes e as áreas cadastradas pelos professores como interesse de avaliação, o que gera algumas incoerências como resumos distribuídos para professores que não são especialistas sobre o assunto do resumo. O presente trabalho apresenta uma avaliação experimental inicial do uso de clusterização na distribuição de resumos para professores.

1. Introdução

A iniciação científica está intrinsecamente ligada à missão das universidades brasileiras. Nesse sentido, a constituinte de 1988 no art. 207 enfatiza a indissociabilidade entre ensino, pesquisa e extensão. A Pró-Reitoria de Pesquisa (PROPESQ) é o órgão auxiliar incumbido de gerenciar e avaliar as políticas de pesquisa científica e tecnológica mantidas pela Universidade Federal da Paraíba (UFPB). Anualmente a PROPESQ promove o Encontro de Iniciação Científica (ENIC), evento destinado aos estudantes participantes dos Programas de Iniciação Científica da UFPB para apresentação dos resultados das pesquisas vinculadas aos projetos desenvolvidos.

Para o ENIC os discentes submetem os resumos do que foi realizado durante o período da sua bolsa. Professores são, então, convidados a avaliar os resumos desses trabalhos científicos. Um dos desafios nesse cenário é distribuir o resumo para o professor avaliador que tenha maior afinidade com o tema da pesquisa. Atualmente, os professores relatam dificuldades com a distribuição realizada, pois os mesmos acabam

recebendo resumos que não levam em consideração seus trabalhos já realizados e área de atuação. Isso ocorre porque as regras de distribuição atual levam em consideração somente a área do resumo e do professor. Por exemplo, um resumo cadastrado com a seguinte hierarquia de áreas: Ciências Exatas e da Terra / Ciência da Computação / Teoria da Computação / Linguagem Formais e Autômatos pode não encontrar professores para avaliação na área mais específica que seria “Linguagem Formais e Autômatos”, logo esse resumo será distribuído para área anterior até que se encontrem professores para avaliação.

Assim sendo, através do presente trabalho, buscou-se responder a seguinte questão: “O uso da técnica de clusterização pode melhorar a distribuição de resumos de iniciação científica?”.

O presente relatório seguirá a seguinte organização: Na Seção 2, serão abordados os conceitos e trabalhos relacionados; Na Seção 3, serão abordados os métodos utilizados para realização da pesquisa; Na Seção 4, será apresentada a implementação e os resultados da pesquisa; Na Seção 5, serão apresentadas as considerações do trabalho e trabalhos futuros.

2. Fundamentação teórica e trabalhos relacionados

As medidas de distâncias e similaridades são usadas para reconhecimento de padrões por meio de classificação, agrupamentos e sistemas de recomendação, por exemplo (a similaridade entre usuários em uma rede social) [Cremonez 2016]. Quando dois objetos possuem distância igual a 0, pode-se dizer que os objetos estão no mesmo lugar, ou seja, ao compararmos a distância entre os atributos dos objetos e, em obtendo a distância igual a 0, podemos dizer que ambos são idênticos [Dubuisson 1994].

A tarefa de clusterização é utilizada para separar as instâncias de dados em subconjuntos ou clusters (agrupamentos), de tal forma que os elementos de um cluster compartilhem propriedades comuns, que servem para distinguir os elementos em outros clusters, tendo como objetivo maximizar similaridade intra-cluster e minimizar similaridade inter-cluster [Galvão 2009].

O algoritmo Simple K-Means é um algoritmo que tem por função principal o agrupamento de dados em k conjuntos diferenciados entre as especificações do grupo de dados. Funciona como uma localização por distância verificando caminhos viáveis minimizados. O algoritmo sugere a utilização de um dado como referência no espaço do conjunto para busca dos padrões. A partir deste momento começa a medição das diferenças de localizações de k conjuntos dos dados perante seu centro [Brum 2019].

Já o algoritmo EM (Expectation-Maximization), descrito por [McLachlan 2007], é um método de agrupamento não supervisionado e tem base nos Modelos de Mistura Gaussiana [Jung 2014]. Funciona em uma abordagem iterativa, subótima, que tenta encontrar os parâmetros da distribuição de probabilidade que tem a máxima probabilidade de seus atributos.

Em [Pimentel 2003] a clusterização foi utilizada para identificar grupos de alunos com características similares em algumas turmas em duas instituições de ensino superior. O trabalho utilizou dois algoritmos de clusterização, o k-means e o SOM. O primeiro algoritmo formou grupos de divisão similar à classificação que um professor da disciplina faria, enquanto que o segundo algoritmo não formou agrupamentos,

possivelmente pela escassez de respostas dos alunos ao questionário aplicado. O resultado apresentado se mostrou útil para formação de grupos homogêneos de aprendizes, mostrando que os agrupamentos formados pelo k-means são similares a que um humano faria utilizando análise de desempenho do aluno em sala de aula.
[Brandão 2003] apresenta uma análise de agrupamento de escolas e Núcleos de Tecnologia Educacional (NTE) do Programa Nacional de Informática na Educação (ProInfo). O algoritmo de clusterização utilizado foi o EM. A estratégia utilizada foi a de testar o algoritmo com vários números de grupos e adotar aquele imediatamente antes de ocorrer um salto no valor que representa perda de verossimilhança.
Foi-se variando o número de grupos como parâmetro do algoritmo, chegando a quatro grupos de escolas e também quatro grupos de NTE do ProInfo. Esses resultados preliminares podem orientar pesquisas futuras no sentido de investigar relações que podem justificar ou explicar os diferentes desempenhos nos grupos encontrados.

3. Desenvolvimento/metodologia

Para essa avaliação experimental, foram usados dados fictícios. Para isso foi criado um arquivo CSV com as palavras-chave de professores e resumos, como mostra a Tabela 1.

Tabela 1. Arquivo CSV usado com as palavras-chave dos professores e resumos

docente/resumo	palavrachave1	palavrachave2	palavrachave3
P1	MACHINE LEARNING	INTELIGENCIA ARTIFICIAL	ALGORITMO GENETICO
P2	SISTEMAS EMBARCADOS	ROBOTICA	ARDUINO
P3	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR
R1	INTELIGENCIA ARTIFICIAL	APRENDIZAGEM PROFUNDA	MACHINE LEARNING
R2	MACHINE LEARNING	ALGORITMO GENETICO	CLUSTERIZACAO
R3	ROBOTICA	SISTEMAS EMBARCADOS	RASPEBERRY
R4	SISTEMAS EMBARCADOS	DRONE	ARDUINO
R5	HUMANO COMPUTADOR	ACESSIBILIDADE	USABILIDADE
R6	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR
R7	MACHINE LEARNING	INTELIGENCIA ARTIFICIAL	ALGORITMO GENETICO
R8	SISTEMAS EMBARCADOS	ROBOTICA	ARDUINO
R9	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR

Para a construção do dataset, foi realizado um processamento manual com o objetivo de selecionar todas as palavras-chave únicas e organizá-las em colunas. Os professores e resumos são as linhas e, como as palavras-chave não se repetem por instância, sua frequência é sempre 0 ou 1, logo será atribuído o valor 1 quando a palavra-chave da coluna coincidir com a palavra-chave do professor/resumo e zero quando não coincidir (Tabela 2). O arquivo gerado contém 12 instancias (3 professores e 9 resumos) e 15 atributos referentes às palavras-chave. Esses dados em CSV foram importados no WEKA e dois algoritmos de clusterização foram utilizados. Com base nos trabalhos relacionados foram selecionados os algoritmos EM e o K-means.

Tabela 2. Reorganização dos dados para uso do algoritmo de clusterização

Professor/Resumo	ACESSIBILIDADE	ALGORITMO GENETICO	APRENDIZAGEM PROFUNDA	ARDUINO	...
P1	0	1	0	0	
P2	0	0	0	1	
P3	1	0	0	0	
R1	0	0	1	0	
R2	0	1	0	0	
R3	0	0	0	0	
R4	0	0	0	1	
R5	1	0	0	0	
R6	1	0	0	0	
R7	0	1	0	0	
R8	0	0	0	1	
R9	1	0	0	0	

4. Resultados e discussão

O ambiente utilizado para aplicação das técnicas de mineração de dados foi o WEKA [Waikato 2004].

Para todas as distribuições foram setados o número de clusters com o valor 3, que é o número de professores no arquivo gerado, já que a intenção é gerar agrupamentos de resumos em volta dos professores. A primeira distribuição utilizou o algoritmo EM gerando 3 clusters com 4 instâncias cada um (Figura 1), a distribuição se mostrou homogênea analisando o cluster 0. O P3 e os resumos R5,R6 e R9 foram agrupados, as palavras-chaves coincidem e se mostra uma boa seleção de resumos para o professor P3 (Tabela 3) .

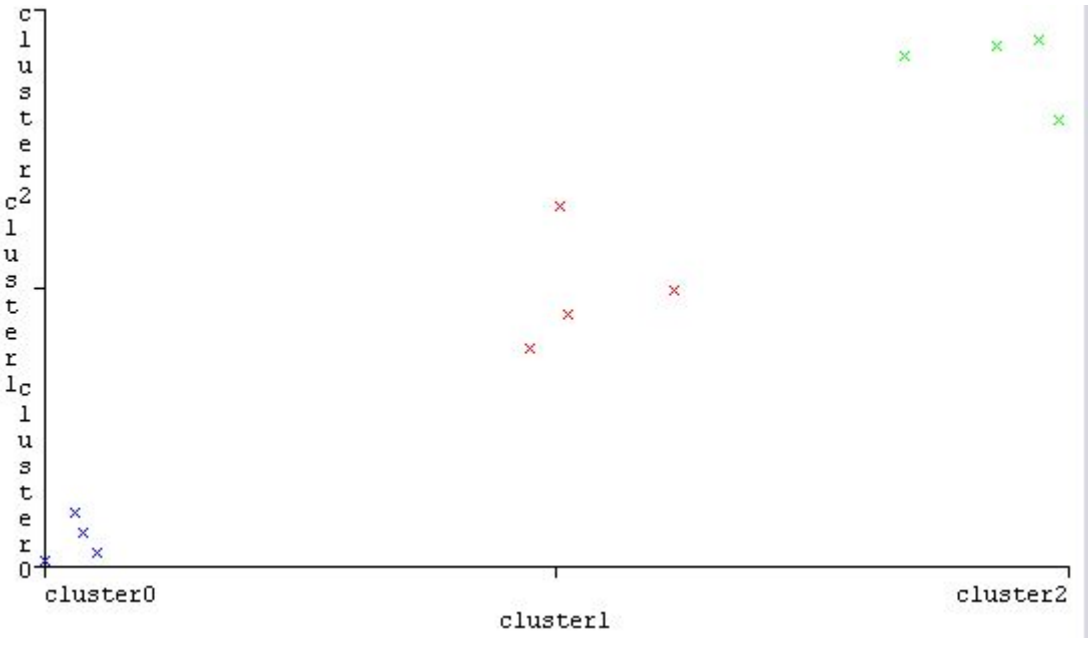


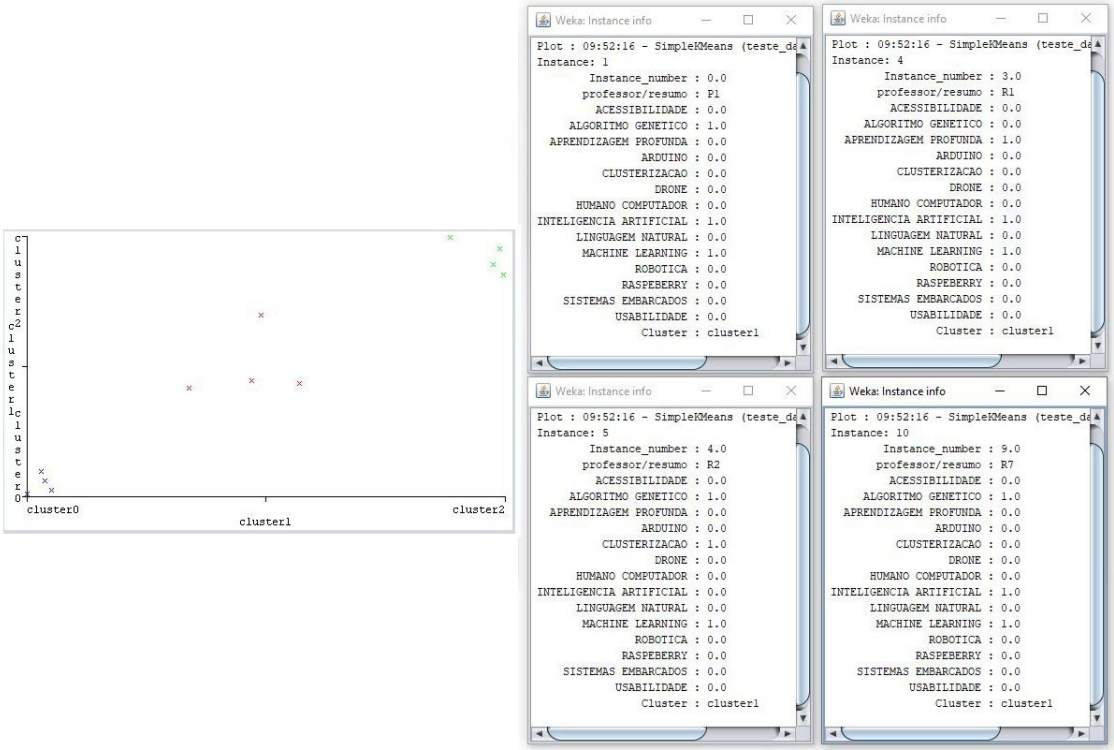
Figura 1. Distribuição gerada pelo algoritmo EM.

Tabela 3. Análise do cluster 0 conforme a distribuição pelo algoritmo EM.

docente/resumo	palavrachave1	palavrachave2	palavrachave3
P3	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR
R5	HUMANO COMPUTADOR	ACESSIBILIDADE	USABILIDADE
R6	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR
R9	ACESSIBILIDADE	LINGUAGEM NATURAL	HUMANO COMPUTADOR

Os outros clusters se mostram tão conscientes quanto o cluster 0, as palavras-chaves coincidem adequadamente e, no geral, a distribuição se mostra coerente.

A próxima distribuição utilizou o algoritmo K-means, a quantidade de clusters foi setada como 3, e a função de distância utilizada foi a euclidiana. A distribuição gerou 3 clusters com 4 instâncias cada. O algoritmo K-means gerou distribuição idêntica ao algoritmo EM (Figura 2). Analisando o cluster 1, percebe-se que as palavras-chaves coincidem e a distribuição se mostra coerente com o esperado.



5. Considerações e trabalhos futuros

O uso de algoritmos de clusterização para distribuição de resumos para avaliação de professores se mostrou promissor. As palavras-chaves se mostram mais adequadas do que o uso das áreas de atuação, já que melhoram o encontro de resumos e professores com temas semelhantes. Os algoritmos utilizados mostraram resultados semelhantes. Somente o uso posterior de dados reais em uma massa maior de informações pode revelar o melhor algoritmo a ser utilizado. Os clusters gerados mostraram-se coerentes agrupando as palavras-chave semelhantes.

Como trabalho futuro pretende-se usar o modelo em dados reais da iniciação científica da UFPB. Para isso deverá ser realizado um pré-processamento para obter as palavras-chave dos resumos, contemplando as palavras com maior frequência no texto do resumo e as palavras relevantes do título. Deverá ser formado um conjunto de palavras-chave a partir dos trabalhos científicos dos professores. Todas essas informações deverão ser hierarquizadas para obter os temas de cada professor e resumo para então aplicar-se alguns algoritmos de clusterização de modo a obter-se uma avaliação experimental mais completa.

6. Referências

Cremonez, Vitor Massaro. Recomendação Entre Usuários De Redes Sociais Por Conteúdo, 2016.

Dubuisson, M.-P.; Jain, A. K. A modified hausdorff distance for object matching. In: IEEE. Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. [S.l.], 1994. v. 1, p. 566–568.

Galvão, Noemi Dreyer; De Fátima Marin, Heimar. Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem, v. 22, n. 5, p. 686-690, 2009.

De Brum, Flávio; Mozzaquatro, Patricia Mariotto; Zanatta, Jocias Maier. Estudo Sobre Os Algoritmos De Clusterização Hierarchical Clusterer E Simple K-means Aplicados No Agrupamento De Padrões Similares. Revista da Universidade Vale do Rio Verde, v. 17, n. 1, 2019.

Mclachlan, Geoffrey J.; Krishnan, Thriyambakam. The EM algorithm and extensions. Wiley-Interscience, 2007

Jung, Yong Gyu; Kang, Min Soo; Heo, Jun. Clustering Performance Comparison Using K-means And Expectation Maximization Algorithms. Biotechnology & Biotechnological Equipment, V. 28, N. Sup1, P. S44-s48, 2014.

Pimentel, Edson P.; De França, Vilma F.; Omar, Nizam. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2003. p. 495-504.

Brandão, Maria de Fátima Ramos; Dos Santos Ramos, Carlos Renato; Tróccoli, Bartholomeu T. Análise de agrupamento de escolas e Núcleos de Tecnologia

Educacional: mineração na base de dados de avaliação do Programa Nacional de Informática na Educação. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2003. p. 366-374.

University of Waikato. (2004) Weka 3 – Machine Learning Software in Java. Disponível em https://www.cs.waikato.ac.nz/ml/weka. Acesso em 30/12/2019.