EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Programa de Pós-graduação em Sistemas de Informação
PPgSI

VI Workshop de Dissertações de Mestrado do PPgSI
2019

Desenvolvimento de um método exato para a avaliação de tarefas heurísticas de mineração de textos

Autoria de:	André Marinho Valadão Batemarchi			
Orientação de:	Prof. Dr. Alexandre da Silva Freire			
Coorientação de:				
Linha de pesquisa:	[ ] Gestão e Desenvolvimento de Sistemas		[x] Inteligência de Sistemas	
Área de pesquisa:	[ ] Banco de dados [ ] Gestão de tecnologia da informação	[ ] Engenharia de software [ ] Interação humano computador	[ ] Inteligência artificial [ ] Reconhecimento de padrões	[ ] Processamento gráfico [x] Otimização
Área de aplicação:	[ ] Ambientes corporativos / Processos de negócio [ ] Economia [ ] Jogos / Jogos sérios [ ] Saúde	[ ] Bioinformática [ ] Educação / Educação a distância [ ] Linguística / Língua natural [ ] Outra Qual? _____	[ ] Biometria [ ] Governo eletrônico [ ] Quimioinformática	[ ] Dispositivos móveis [ ] Internet / Redes sociais [ ] Robótica [x] Geral*
Semestre no curso (na data do workshop):	[ ] 2º semestre	[x] 3º semestre	[ ] 4º semestre	[ ] 5º semestre
Qualificação:	[ ] Qualificação já realizada em: dd/mm/aaaa		[x] Realização da qualificação planejada para: 01/11/2019	
Defesa:	Prazo máximo para depósito: 01/02/2021		Realização da defesa planejada para: 01/07/2020	
Publicações associadas ao projeto de mestrado:	**Sem publicações até o momento**.			

Resumo do projeto de pesquisa

Contexto:

O contexto desta pesquisa é o desenvolvimento de métodos exatos da otimização combinatória para aplicação na área de mineração de textos. Neste sentido, a identificação de padrões textuais por meio da categorização de textos é uma tarefa caracterizada pela alta dimensionalidade e esparsidade de dados. Trata-se de uma tarefa desafiadora que até o momento não possui uma cota inferior definida. O desenvolvimento de métodos exatos permite realizar a avaliação dos métodos existentes com relação à solução ótima do problema e identificar possíveis melhorias sobre os algoritmos existentes. Portanto, busca-se contribuir com os interessados no desenvolvimento de algoritmos de coagrupamento utilizados para categorização de textos.

Problema de pesquisa:

O método de coagrupamento permite realizar a categorização de textos com soluções sub-ótimas, o que implica em uma complexidade computacional sem cota inferior definida. As principais dificuldades identificadas são decorrentes das características intrínsecas dos dados textuais. Sendo assim, o enfoque desta pesquisa é a não otimalidade do método de coagrupamento e a sua dificuldade em lidar com dados textuais esparsos e de alta dimensionalidade.

Objetivo de pesquisa:

O objetivo de pesquisa é desenvolver um método exato para a avaliação de tarefas heurísticas de mineração de textos utilizando como técnica a programação inteira, de modo a obter a solução ótima do problema de categorização de textos.

Caracterização da solução em desenvolvimento:

Dividiu-se a solução em três etapas. A primeira é o entendimento do problema (método de coagrupamento) e de como ele é abordado na literatura por meio de uma revisão sistemática. A segunda etapa consiste em desenvolver um modelo matemático utilizando a técnica de programação inteira e assim, desenvolver um método exato. Na última etapa realiza-se a avaliação do modelo desenvolvido, a análise dos resultados, a escrita da monografia e a divulgação do trabalho.

Fundamentos:

- Dados textuais do tipo notícia são esparsos, ou seja, possuem grau de ocupação reduzido ao representa-los por meio de matrizes (método de coagrupamento).
- Dados textuais do tipo notícia possuem alta dimensionalidade. Seja um conjunto de dados contendo notícias de diversas categorias como, por exemplo, esporte, entretenimento e política, e dado que há diversas possibilidades de ocorrências de palavras, há uma combinação relativamente elevada de possibilidades.
- O método de coagrupamento obtém cogrupos com base em similaridades parciais. Seja uma matriz de ocorrências de palavras em notícias, obtém-se os padrões relativos às diversas categorias possíveis.
- Algoritmos de coagrupamento são heurísticos, ou seja, não fornecem solução ótima, mas apenas sub-ótimas em determinados casos.
- Um método exato pode fornecer a solução ótima para um problema.
- A programação inteira é a técnica que permite a criação de um modelo matemático para se obter um método exato.
- A programação inteira é um modelo de programação linear que lida com números inteiros, como a ocorrência de palavras em uma notícia.
- O modelo busca maximizar os cogrupos das categorias de notícias correspondentes.
- O algoritmo resultante do método exato é ótimo.

Trabalhos relacionados:

- GREENE, D.; CUNNINGHAM, P. Producing accurate interpretable clusters from high-dimensional data. In*: Proceedings of the 9th European Conference on European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Porto, Portugal: Springer-Verlag, 2005. p. 486–494.
- HUANG, S.; XU, Z.; LV, J. Adaptive local structure learning for document co-clustering. *Knowl.-Based Syst.*, v. 148, n. 1, p. 74–84, 2018.
- LONG, B.; ZHANG, Z. M.; YU, P. S. Co-clustering by block value decomposition. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Chicago, USA: ACM, 2005. p. 635–640.

Validação:

A pesquisa será intrínseca e as avaliações ocorrerão sobre a complexidade computacional do método desenvolvido, e métricas sobre dados padrões, como acurácia, precisão, sensibilidade (recall) e medida-F. Os conjuntos de dados utilizados são o córpus EBC, que contém notícias na forma de texto em português; e o NG20, contendo notícias em inglês.

Portanto, a estratégia geral de validação é analisar assintoticamente o algoritmo desenvolvido e avaliar a qualidade dos agrupamentos utilizando medidas exatas.

Limitações, riscos e ameaças:
A pesquisa está limitada ao estudo da técnica de programação inteira e ao desenvolvimento de um método exato. Sua aplicação será sobre o problema de categorização de textos e assim, problemas equivalentes, como classificação de grafos e expressão gênica, não serão abordados. Há, portanto, o risco de o método não ser replicável sem ser possível avaliar e obter a solução ótima de outros problemas resolvidos utilizando-se coagrupamento.

Contribuição científica:
A pesquisa representará uma contribuição de uma nova abordagem para a área de mineração de textos. Será desenvolvido um método exato de maneira a definir uma cota inferior para o problema de categorização de textos. Os resultados esperados são avançar os estudos da otimização combinatória aplicada à mineração de textos sobre um problema em que pouco se explorou o método do tipo exato fornecendo, assim, os seguintes produtos:

- um relatório técnico dos principais algoritmos heurísticos que realizam coagrupamento para a categorização de textos;
- um modelo matemático para o problema de categorização de textos;
- um método exato utilizando como técnica a programação inteira;
- uma cota inferior para o problema.

Contribuição tecnológica (se pertinente):
A implementação do algoritmo desenvolvido estará disponível para utilização em trabalhos futuros.

Método de pesquisa			
Gênero (escolha UM)	[ ] Pesquisa teórica	[x] Pesquisa prática	[ ] Pesquisa empírica [ ] Pesquisa metodológica
Natureza (escolha UMA)	[ ] Pesquisa básica		[x] Pesquisa aplicada
Abordagem (escolha UMA)	[ ] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[x] Pesquisa mista (quali-quanti)
Revisão de literatura* (você pode escolher mais de uma)	[ ] Revisão narrativa [ ] Revisão descritiva [x] Revisão de escopo	[ ] Meta-análise [x] Revisão sistemática qualitativa [ ] Revisão *guarda-chuva*	[ ] Revisão teórica [ ] Revisão realística [ ] Revisão crítica
Procedimento técnico principal (escolha UM)	[x] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto*	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação	[ ] Pesquisa etnográfica [ ] Teoria fundamentada em dados [ ] Ciência do projeto [ ] Outra Qual? _____
Análise de dados (você pode escolher mais de uma)	[x] Estatística descritiva [ ] Estatística inferencial	[ ] Teste estatístico [x] Análise de conteúdo	[ ] Análise do discurso [ ] Outros: _____

* Definição de tipos de revisões de literatura estabelecida por Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Próximas atividades:

1. Finalização da revisão sistemática;
2. Preparação dos dados do problema;
3. Elaboração de um modelo matemático;
4. Desenvolvimento de um método exato;
5. Avaliação do método desenvolvido;
6. Análise dos resultados;
7. Redação da monografia;
8. Divulgação do trabalho.