

II Workshop de Dissertações de Mestrado do PPgSI

FICHA DA PESQUISAⁱ

| DADOS GERAIS | | | | | | | | | | | | | | | | | |
|---|--|---|--|--|--|---|---|--------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|---|--------------------------------|-----------------------------------|--------------------------------|----------------------------------|---|
| Título do projeto de pesquisa | Resolução do problema de Coclusterização em Matrizes de Dados Esparsas usando Fatoração de Matrizes | | | | | | | | | | | | | | | | |
| Orientando | Lucas Fernandes Brunialti | | | | | | | | | | | | | | | | |
| Orientador(es) | Profa. Dra. Sarajane Marques Peres (orientadora), Prof. Dr. Valdinei Freire Silva (co-orientador) | | | | | | | | | | | | | | | | |
| Momento atual | <input type="checkbox"/> 3º semestre <input type="checkbox"/> 4º semestre <input checked="" type="checkbox"/> 5º semestre <input type="checkbox"/> 6º semestre | | | | | | | | | | | | | | | | |
| Qualificação | <input checked="" type="checkbox"/> Qualificação já realizada em: 11/03/2015 <input type="checkbox"/> Qualificação planejada para: | | | | | | | | | | | | | | | | |
| Defesa | Prazo máximo para depósito: 18/07/2016 Depósito planejado para: 10/03/2016 | | | | | | | | | | | | | | | | |
| Linha e Área de pesquisa | <table style="width: 100%; border: none;"> <tr> <td style="width: 50%; border: none;"> <input type="checkbox"/> Gestão e desenvolvimento de Sistemas: <input type="checkbox"/> BD <input type="checkbox"/> Gestão de SI <input type="checkbox"/> Eng. de Software <input type="checkbox"/> IHC </td> <td style="width: 50%; border: none;"> <input checked="" type="checkbox"/> Inteligência de Sistemas: <input type="checkbox"/> IA <input checked="" type="checkbox"/> Rec. de Padrões <input type="checkbox"/> Proc. Gráfico </td> </tr> </table> | <input type="checkbox"/> Gestão e desenvolvimento de Sistemas: <input type="checkbox"/> BD <input type="checkbox"/> Gestão de SI <input type="checkbox"/> Eng. de Software <input type="checkbox"/> IHC | <input checked="" type="checkbox"/> Inteligência de Sistemas: <input type="checkbox"/> IA <input checked="" type="checkbox"/> Rec. de Padrões <input type="checkbox"/> Proc. Gráfico | | | | | | | | | | | | | | |
| <input type="checkbox"/> Gestão e desenvolvimento de Sistemas: <input type="checkbox"/> BD <input type="checkbox"/> Gestão de SI <input type="checkbox"/> Eng. de Software <input type="checkbox"/> IHC | <input checked="" type="checkbox"/> Inteligência de Sistemas: <input type="checkbox"/> IA <input checked="" type="checkbox"/> Rec. de Padrões <input type="checkbox"/> Proc. Gráfico | | | | | | | | | | | | | | | | |
| Área de aplicação | <table style="width: 100%; border: none;"> <tr> <td style="width: 25%; border: none;"><input type="checkbox"/> Ambientes Corporativos</td> <td style="width: 25%; border: none;"><input type="checkbox"/> Educação</td> <td style="width: 25%; border: none;"><input type="checkbox"/> Linguagem Natural</td> <td style="width: 25%; border: none;"><input type="checkbox"/> Redes Sociais</td> </tr> <tr> <td style="border: none;"><input type="checkbox"/> Bioinformática</td> <td style="border: none;"><input type="checkbox"/> Educação a Distância</td> <td style="border: none;"><input type="checkbox"/> Linguística</td> <td style="border: none;"><input type="checkbox"/> Robótica</td> </tr> <tr> <td style="border: none;"><input type="checkbox"/> Biometria</td> <td style="border: none;"><input type="checkbox"/> Internet</td> <td style="border: none;"><input type="checkbox"/> Processos de Negócio</td> <td style="border: none;"><input type="checkbox"/> Saúde</td> </tr> <tr> <td style="border: none;"><input type="checkbox"/> Economia</td> <td style="border: none;"><input type="checkbox"/> Jogos</td> <td style="border: none;"><input type="checkbox"/> Química</td> <td style="border: none;"><input checked="" type="checkbox"/> [outro – Geral]</td> </tr> </table> | <input type="checkbox"/> Ambientes Corporativos | <input type="checkbox"/> Educação | <input type="checkbox"/> Linguagem Natural | <input type="checkbox"/> Redes Sociais | <input type="checkbox"/> Bioinformática | <input type="checkbox"/> Educação a Distância | <input type="checkbox"/> Linguística | <input type="checkbox"/> Robótica | <input type="checkbox"/> Biometria | <input type="checkbox"/> Internet | <input type="checkbox"/> Processos de Negócio | <input type="checkbox"/> Saúde | <input type="checkbox"/> Economia | <input type="checkbox"/> Jogos | <input type="checkbox"/> Química | <input checked="" type="checkbox"/> [outro – Geral] |
| <input type="checkbox"/> Ambientes Corporativos | <input type="checkbox"/> Educação | <input type="checkbox"/> Linguagem Natural | <input type="checkbox"/> Redes Sociais | | | | | | | | | | | | | | |
| <input type="checkbox"/> Bioinformática | <input type="checkbox"/> Educação a Distância | <input type="checkbox"/> Linguística | <input type="checkbox"/> Robótica | | | | | | | | | | | | | | |
| <input type="checkbox"/> Biometria | <input type="checkbox"/> Internet | <input type="checkbox"/> Processos de Negócio | <input type="checkbox"/> Saúde | | | | | | | | | | | | | | |
| <input type="checkbox"/> Economia | <input type="checkbox"/> Jogos | <input type="checkbox"/> Química | <input checked="" type="checkbox"/> [outro – Geral] | | | | | | | | | | | | | | |

DESCRIÇÃO DO PROJETO DE PESQUISA

| | |
|-------------------------------------|--|
| Contextualização / motivação | Algoritmos de coclusterização tem o objetivo de encontrar grupos em uma matriz de dados. Tais grupos são chamados coclusters por se caracterizarem como um subconjunto de linhas e colunas, e podem apresentar sob diferentes tipos de organização. Especialmente, organizações nas quais diferentes coclusters possuem linhas e/ou colunas em comum (organização com sobreposição) podem ser úteis para diversas aplicações, como mineração de textos, filtro colaborativo, recuperação de informação, etc. |
| Problema de pesquisa | Encontrar coclusters organizados com sobreposição em uma matriz de dados esparsa contendo valores reais positivos que representam dados de um domínio de aplicação. |
| Objetivo geral da pesquisa | Propor estratégias algorítmicas baseadas em Fatoração de Matrizes (FM) que sejam capazes de encontrar coclusters organizados com sobreposição em uma matriz esparsa de valores reais positivos. |
| Trabalhos relacionados | <p>Biclustering algorithms for Biological Data Analysis: A Survey - Survey em algoritmos de coclusterização/biclusterização. Discute diferentes organizações de coclusters e algoritmos para cada uma delas.</p> <p>Fast Nonnegative Matrix Tri-Factorization for Large-Scale Data Co-clustering - Apresenta algoritmos de FM para coclusterização em dados binários. Não trata coclusters organizados com sobreposição.</p> <p>Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds - Apresenta uma survey com algoritmos baseados em FM juntamente com um novo algoritmo que encontra coclusters ortogonais. Estes algoritmos não encontram coclusters organizados com sobreposição.</p> |
| Justificativa e relevância | Problemas reais se manifestam em domínios que podem apresentar coclusters organizados com sobreposição sobre matrizes de dados esparsas. Um exemplo é a mineração de textos aplicada à agrupamento de notícias. A representação vetorial de textos (as notícias), conhecidamente, gera matrizes esparsas. Notícias podem ser agrupadas em diferentes grupos a depender de quais de suas características são analisadas. Tais fatores caracterizam o problema sob estudo neste projeto. |
| Proposta para Solução | Estão sendo desenvolvidos algoritmos utilizando técnicas de FM com o intuito de resolver o problema de encontrar coclusters organizados com sobreposição em domínio nos quais os dados se apresentam como uma matriz esparsa. |
| Dados | Estão sendo usados: dados sintéticos; dados de notícias da base de dados Reuters; dados de trabalhos acadêmicos da Rochester University; dados de notícias de um grande portal brasileiro; e um conjunto de dados (movie lens) para teste também na área de filtro colaborativo. |
| Forma de validação | Provas de correção e convergência dos algoritmos desenvolvidos; comparação e experimentação com os algoritmos presentes nos trabalhos correlatos. Avaliação via índices de validação de clustering e coclusterização. |
| Limitações | FM supõe que os dados e características são separáveis linearmente. Esta restrição não é garantida em problemas reais, mas não impede que soluções baseadas em FM gerem resultados úteis. |

| | |
|-----------------------------|---|
| Resultados esperados | Contribuições científicas: desenvolvimento de algoritmos inéditos baseados em FM para coclusterização organizados com sobreposição em matrizes esparsas positivas |
| | Contribuições tecnológicas: - |

| MÉTODO DE PESQUISA | | | |
|---|---|---|--|
| Gênero | <input checked="" type="checkbox"/> Pesquisa teórica | <input type="checkbox"/> Pesquisa prática | <input type="checkbox"/> Pesquisa empírica <input type="checkbox"/> Pesquisa metodológica |
| Natureza | <input type="checkbox"/> Pesquisa básica/pura | <input checked="" type="checkbox"/> Pesquisa aplicada | |
| Objetivo | <input type="checkbox"/> Pesquisa descritiva | <input checked="" type="checkbox"/> Pesquisa exploratória | <input type="checkbox"/> Pesquisa explicativa <input type="checkbox"/> [outro] |
| Abordagem | <input checked="" type="checkbox"/> Pesquisa quantitativa | <input type="checkbox"/> Pesquisa qualitativa | <input type="checkbox"/> Pesquisa mista (quali-quant) |
| Procedimento(s) técnico(s) | <input checked="" type="checkbox"/> Pesquisa experimental <input type="checkbox"/> Pesquisa bibliográfica <input type="checkbox"/> Pesquisa documental <input type="checkbox"/> Pesquisa <i>ex-post-facto</i> <input type="checkbox"/> Pesquisa de levantamento | <input type="checkbox"/> Pesquisa com <i>survey</i> <input type="checkbox"/> Estudo de caso <input type="checkbox"/> Pesquisa participante <input type="checkbox"/> Pesquisa-ação <input type="checkbox"/> Pesquisa etnográfica | <input type="checkbox"/> Pesquisa netnográfica <input type="checkbox"/> Teoria fundamentada em dados (<i>grounded theory</i>) <input type="checkbox"/> Ciência do projeto (<i>Design science research</i>) <input type="checkbox"/> [outro] |
| Fonte(s) de dados | <input checked="" type="checkbox"/> pesquisa de laboratório | <input type="checkbox"/> pesquisa de campo | <input type="checkbox"/> pesquisa bibliográfica <input type="checkbox"/> [outro] |
| Técnica(s) / Instrumento(s) de coleta de dados | <input type="checkbox"/> medição <input type="checkbox"/> questionário <input type="checkbox"/> entrevista <input type="checkbox"/> grupos focais | <input type="checkbox"/> formulário <input type="checkbox"/> <i>benchmark</i> | <input type="checkbox"/> observação (direta / participante) <input type="checkbox"/> diário de campo / notas de campo <input type="checkbox"/> análise documental (ou de artefatos) <input type="checkbox"/> [outro] |
| Técnica(s) de análise de dados | Análise quantitativa: <input type="checkbox"/> Estatística descritiva <input type="checkbox"/> Estatística inferencial <input checked="" type="checkbox"/> [outro – índices de validação de clustering e co-clustering] | Análise qualitativa: <input type="checkbox"/> Análise de conteúdo <input type="checkbox"/> Análise do discurso <input type="checkbox"/> [outro] | |

| CRONOGRAMA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--------------------------|------|---|---|---|---|---|---|---|---|----|----|----|------|---|---|---|---|---|---|---|---|----|----|----|------|---|---|---|---|---|---|---|---|----|----|----|--|--|--|
| | 2014 | | | | | | | | | | | | 2015 | | | | | | | | | | | | 2016 | | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | | | |
| Estudo da literatura | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Revisão sistemática | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Definição da proposta | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Execução da proposta | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Coleta/obtenção de dados | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Análise dos dados | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Validação da proposta | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Preparo da qualificação | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Exame de qualificação | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Escrita da dissertação | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Escrita de artigo | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Depósito da dissertação | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| (Outros) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

ⁱ Esta ficha é uma adaptação da usada no “VIII Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI 2015)” realizado como parte do “XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)”