



Isnard Thomas Martins

**Descoberta de Conhecimento em Históricos Criminais:
Algoritmos e Sistemas**

Tese de Doutorado

Tese apresentada no Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia Industrial da PUC-Rio como parte dos requisitos parciais para obtenção do título de Doutor em Engenharia de Produção.

Orientador: Prof. Silvio Hamacher

Rio de Janeiro
Março de 2009



Isnard Thomas Martins

**Descoberta de Conhecimento em Históricos Criminais:
algoritmos e sistemas**

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia de Produção da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Silvio Hamacher

Orientador

Departamento de Engenharia Industrial - PUC-Rio

Prof. Mozart B. C. Menezes

MIT - Zaragoza

Prof. Cláudio Roberto Contador

Funenseg

Prof. Luiz Eduardo Bento de Mello Soares

UERJ

Prof^a. Laura Bahiense da Silva Leite

UFRJ

Prof. Nélio Domingues Pizzolato

Departamento de Engenharia Industrial - PUC-Rio

Prof. José Eugenio Leal

Coordenador Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 05 de março de 2009

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Isnard Thomas Martins

Professor (PUC-Rio), Economista, Pesquisador de Atividades em Segurança Pública, Defendeu dissertação de mestrado na PUC-Rio, intitulada Ergonomia de Ambientes Instrucionais de Educação à Distância, Desenvolvedor de Tecnologia de Inteligência aplicada, destacando pesquisas em bases criminais. Apresenta experiência acadêmica e empresarial, tendo desempenhado funções executivas e de consultoria na área de Segurança Pública em diversos Estados Brasileiros, Ministério Público e SENASP em Brasília. Tem diversos projetos desenvolvidos e implantados na área de Segurança Pública e Sistemas comerciais, participando como mentor, analista e programador de aplicações.

Ficha Catalográfica

Martins, Isnard Thomas

Descoberta de conhecimento em históricos criminais: algoritmos e sistemas / Isnard Thomas Martins ; orientador: Silvio Hamacher. – 2009.

201 f.; 30 cm

Tese (Doutorado em Engenharia Industrial) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Análise criminal. 3. Grafos. 4. Redes sociais. 5. Redes neurais. 6. Pesquisa operacional. 7. Mineração de dados. I. Hamacher, Silvio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. III. Título.

CDD: 658.5

Agradecimentos

Aos meus filhos e à minha mulher Vânia, pelo apoio, paciência e incentivo para conclusão deste trabalho.

Ao professor Silvio Hamacher, pela sabedoria, paciência e contribuições fornecidas por seu trabalho de orientação para materialização desta Tese, a quem dedico especiais agradecimentos.

Aos professores Nélio Domingues Pizzolato e Marley Bernardes Rebuszi Velasco pelo apoio e contribuições para realização deste trabalho.

Aos professores Luiz Eduardo Soares, Mozart B.C. Menezes, Laura Bahiense da Silva Leite, Cláudio Roberto Contador pela participação na Banca Examinadora.

A FENASEG - Federação Nacional das Empresas de Seguros Privados e da Capitalização, professor Cláudio Contador, professor Luiz Roberto Cunha e IAPUC pelo apoio para concessão da Bolsa de estudos concedida para conclusão da Tese.

Ao professor Madiagne Diallo pelo apoio e contribuições fornecidas.

Aos demais professores da PUC-Rio dos Programas de Engenharia de Produção, I.A.G. e Engenharia Elétrica pelas inestimáveis contribuições e conhecimentos transmitidos.

Aos profissionais da Polícia do Rio de Janeiro pelas contribuições e críticas desenvolvidas para materialização deste trabalho.

À Janduy Coutinho e Luciano Coutinho da UNIMIX de Brasília pelo apoio e confiança empenhados ao longo da fase de conclusão deste trabalho.

Resumo

Isnard, Thomas Martins; Hamacher Silvio (Orientador). **Descoberta de Conhecimento em Históricos Criminais: Algoritmos e Sistemas.** Rio de Janeiro, 2009. 201p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Esta Tese propõe uma metodologia para extração de conhecimento em bases de históricos criminais. A abrangência da metodologia proposta envolve todo o ciclo de tratamento dos históricos criminais, desde a extração de radicais temáticos, passando pela construção de dicionários especializados para apoio à extração de entidades até o desenvolvimento de cenários criminais em formato de uma matriz de relacionamentos. Os cenários são convertidos em Mapas de Inteligência destinados à análise de vínculos criminais e descoberta de conhecimento para investigação e elucidação de delitos. Os Mapas de Inteligência extraídos são representados por redes de vínculos, posteriormente tratados como um grafo capacitado. Análises de associações extraídas serão desenvolvidas, utilizando métodos de caminho mais curto em grafos, mapas neurais auto-organizáveis e indicadores de relacionamentos sociais. O método proposto nesta pesquisa permite a visão de indícios ocultos pela complexidade das informações textuais e a descoberta de conhecimento entre associações criminais aplicando-se algoritmos híbridos. A metodologia proposta foi testada utilizando bases de documentos criminais referentes à quadrilhas de narcotraficantes e casos de crimes de maior comoção social ocorridos no Rio de Janeiro entre 1999 e 2003.

Palavras-chave

Análise Criminal, Grafos, Redes Sociais, Redes Neurais, Pesquisa Operacional, Mineração de Dados.

Abstract

Isnard, Thomas Martins; Hamacher Silvio (Advisor). **Knowledge Discovery in Police Criminal Records: Algorithms and Systems**. Rio de Janeiro, 2009. 201p. Doctorate Thesis – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

This Dissertation proposes a methodology to extract knowledge from databases of police criminal records. The scope of the proposed methodology comprises the full cycle for treatment of the criminal records, from the extraction of word radicals, including the construction of specialized dictionaries to support entity extraction, up to the development of criminal scenarios shaped into a relationship matrix. The scenarios are converted into intelligence maps for the analysis of criminal connections and the discovery of knowledge aimed at investigating and clarifying crimes. The intelligence maps extracted are represented by grids which are subsequently treated as capacitated graphs. Analyses of the connections extracted are carried out using the shortest path method in graphs, self-organizing neural maps, and indicators of social relationships. The method proposed in this study helps revealing evidence that was concealed by the complexity of textual information, and discovering knowledge based on criminal connections by applying hybrid algorithms. The proposed methodology was tested using databases of criminal police records related to drug traffic organizations and crimes that caused major social disturbances in Rio de Janeiro, Brazil, from 1999 to 2003.

Keywords

Criminal Analysis, Graphs, Social Networks, Neural Networks, Operations Research, Datamining.

Sumário

1. INTRODUÇÃO	15
1.1- O Problema	16
1.2 - Objetivos da Tese	17
1.3 - O Método	18
1.4 - Resultados	18
1.5 - Organização da Tese	19
2. VISÃO GERAL DO MODELO	20
2.1 - A Utilização de Mapas de Inteligência	20
2.2 - Exemplo do modelo de relacionamentos do Grupo Onze de Setembro	23
2.3 - Dinâmica do Crime - A Importância dos Padrões de Procedimentos	26
2.4 - Procedimentos Manuais na Análise Criminal	27
2.5 - Estrutura da Pesquisa	29
3. DESENVOLVIMENTO DE DICIONÁRIOS TEMÁTICOS	32
3.1- Referencial Teórico	32
3.1.1 - Preparação dos Dicionários	34
3.1.2 - Métodos de Construção	35
3.1.3 - Categorias e classificações de entidades	36
3.1.4 - Restrições de linguagem e regras para extração de palavras-chave	37
3.1.5 - Extração de palavras-chave em língua portuguesa	39
3.1.6 - Extração de Radicais (Stemming)	40
3.2 - Algoritmos Pesquisados	42
3.2.1 - Algoritmo PORTER	42
3.2.2 - Algoritmo ORENGO e HUYCK	44
3.2.3 - Algoritmo KEA	45
3.2.4 - Algoritmo LOVINS	46
3.2.5 - Limitações dos algoritmos	47
3.2.6 - Uso de Dicionários para descoberta de Interseções	50
3.3 - Descrição do Método para Construção de Dicionários	51
3.4 - Algoritmos utilizados para construção do dicionário especialista	54
3.4.1 - Algoritmo Anaphora RV para Extração de Radicais	55
3.4.2 - Anaphora PCh para extração de palavras-chave	56
3.5 - Sistemas para apoio à extração de dicionários	60
3.5.1 - Descrição do Sistema Stemmer Anaphora	60
3.5.2 - Sistema DicTools	65
3.6 - Avaliação do método para extração de palavras chave	67
3.6.1 - Comparação entre o método manual e automático	67
3.6.2 - Testes Comparativos entre algoritmos extratores de Radicais	69
3.6.3 - Extração e treinamento de um dicionário especialista temático	73
3.6.4 - Dicionários Complementares de Apoio à Extração	74
3.7 - Conclusões deste Capítulo	75
4. EXTRAÇÃO DE ENTIDADES E MODELAGEM DE ASSOCIAÇÕES	76
4.1 - Referencial Teórico	76

4.1.1 - Similaridade e Co-ocorrência	78
4.1.2 - Representação da Rede Semântica	81
4.1.3 - Nível das Representações	81
4.1.4 - Problemas na extração de entidades de históricos policiais	83
4.1.5 - Problemas de consolidação de referências e indicativos de fraudes	85
4.2 - Construção de cenários criminais representados por grafos	87
4.2.1 - Extração de entidades	88
4.3 - Algoritmos pesquisados	89
4.3.1 - Algoritmo Chen & Lynch	89
4.3.2 - Descrição do algoritmo Hauck para cálculo de co-ocorrências	91
4.3.3 - Limitações e problemas relatados com algoritmos para modelagem de associações	94
4.4 - Método para Construção da Matriz de Relacionamentos Criminais	94
4.5 - Alternativas para Modelagem de Cenários	96
4.6 - Algoritmos utilizados para construção da Matriz de Relacionamentos Criminais	98
4.6.1 - Descrição do algoritmo AnaphoraNET para cálculo de co-ocorrências	99
4.6.2 - Algoritmo para Consolidação de Chaves de Acesso	102
4.7 - Sistemas	102
4.7.1 - Organização dos dados para tratamento das associações	103
4.7.2 - Visão funcional do sistema DataAssociations	104
4.8 - Resultados do Processamento do Sistema DataAssociations	106
4.8.1 - Resultados 1	107
4.8.2 - Resultados 2	107
4.8.3 - Resultados 3	107
4.8.4 - Resultados 4	108
4.8.5 - Saída dos dados para análise	108
4.9 - Funções complementares do Sistema DataAssociations	108
5 - Conclusões deste capítulo	110
5. ANÁLISE DAS CONEXÕES	111
5.1 - Referencial Teórico	111
5.1.1 - Organizações e Estruturas	114
5.1.2 - Aplicações dos princípios de centralidade em Investigações	116
5.1.3 - Identificação de Centralidade de um subgrupo	117
5.1.4 - Centralização em grafos Direcionados	123
5.1.5 - Centralidade Máxima em uma Rede Direcionada	124
5.1.6 - Descoberta de subgrupos	125
5.1.7 - SOM - Mapas Auto organizáveis	126
5.1.8 - Ferramentas existentes para Descoberta de Vínculos	129
5.2 - Algoritmos utilizados	130
5.2.1 - Introdução	130
5.2.2 - Algoritmo PFS Modificado (Xu & Chen, 2004)	133
5.3 - Método Proposto	133
5.3.1 - Preparação do Mapa de Inteligência	134
5.3.2 - Método para extração das mais fortes conexões entre as entidades do grafo.	136
5.3.3 - Método para extração de árvores de relacionamentos	137

5.3.4 - Método para identificação de configurações organizacionais em subgrupos criminais	138
5.3.5 - Método para classificação de subgrupos criminais	139
5.3.6 - Método para identificação de Densidade entre subgrupos criminais	140
5.3.7 - Identificação de Centralidades	142
5.4 - Sistemas	143
5.4.1 - AnaphoraVisual	143
5.4.2 - Visual SOM	145
6 - Resumo	149
6. ESTUDOS DE CASOS	151
6.1 - Casos selecionados para experimentação do método	151
6.2 - Desenvolvimento dos Casos Propostos	152
6.3 - Extração de Entidades úteis e modelagem de Mapas de Inteligência	152
6.4 - Análise do caso Narcotraficantes	155
6.5 - Análise do Caso Tim Lopes	164
6.6 - O caso Seqüestro	168
7. CONCLUSÕES	172
7.1 - Visão Geral dos Trabalhos Desenvolvidos	172
7.2 - Resumo dos Objetivos Propostos e Alcançados	174
7.3 - Contribuições oferecidas através da Pesquisa Desenvolvida	175
7.4 - Recomendações para trabalhos Futuros	177
7.4.1 - Ampliação da capacidade de extração em textos livres	177
7.4.2 - Identificação de nomes próprios	177
7.4.3 - Algoritmos para redução do tempo da Extração	178
7.4.4 - Testes exaustivos para cálculos de Co-Ocorrências	179
8. REFERÊNCIAS BIBLIOGRÁFICAS	180
APÊNDICES	187
Apêndice A - Algoritmo Stemming para o idioma Inglês	187
Apêndice B - Algoritmo Stemming para o idioma Português	190
Apêndice C - Algoritmo para extração das mais fortes conexões	192
Apêndice D - Algoritmo para Cálculo de Densidade entre SubGrupos	194
Apêndice E - Algoritmo AnaphoraSom - Mapa Auto Organizado	196
Apêndice F - PFS modificado para cálculo do caminho mínimo entre um par de os nós de um grafo	198
Apêndice G - Exemplo de utilização da Matriz de Relacionamentos	200

Lista de figuras

Figura 2.1 - Mapa de vínculos e conexões entre os terroristas da rede Al-Qaeda, 11 de setembro, USA. (Krebs, 2001; Xu & Chen, 2004)	23
Figura 2.2 - Adaptação do Mapa de Inteligência desenvolvido por Krebs (2001) para o ataque do Onze de Setembro	24
Figura 2.3 - Produção manual de processos para análise criminal	29
Figura 2.4 - Diagrama sintético do sistema de Extração de Entidades, Modelagem de Mapas de Inteligência e Análise	30
Figura 3.1 - Exemplos de regiões usadas na extração de sufixos R1, R2, RV (Porter, 2008)	43
Figura 3.2 - Algoritmo Lovins para o idioma inglês, (Lovins, 1998).	47
Figura 3.3 - Fluxo do processamento para extração de palavras-chave do dicionário especialista	54
Figura 3.4 - Algoritmo Anaphora RV de extração de Radicais	56
Figura 3.5 - Algoritmo Anaphora PCh para construção do Dicionário Especialista	58
Figura 3.6 - Exemplo de regras, aplicadas ao algoritmo Orengo e Huyck, aplicativo Anaphora	62
Figura 3.7 - Exemplo de exceções aplicadas ao algoritmo Orengo e Huyck, aplicativo Anaphora	62
Figura 3.8 - Painel do Sistema Stemmer Anaphora	63
Figura 3.9 - Sequência de redução de palavras- algoritmo Orengo e Huyck	64
Figura 3.10 - Sequência de redução de palavras- algoritmo Anaphora RV	64
Figura 3.11 - Sequência de treinamento de radicais no Sistema Anaphora	65
Figura 3.12 - Painel do Sistema DicTools	67
Figura 3.13 - Resumo de um ciclo de processamento para extração de palavras-chave	73
Figura 4.1 - Co-Ocorrências Entidade-Documento - Hauck	92
Figura 4.2 - Cálculo do peso relativo entre Entidades (Hauck,2002)	92
Figura 4.3 - Pesos combinados das entidades j_k / k_j no documento i (Hauck, 2002)	93
Figura 4.4 - Função WeightingFactor (fator de amortecimento)	93

Figura 4.5 - Algoritmo AnaphoraNET para extração de Entidades e Modelagem de Associações	96
Figura 4.6 - Co-Ocorrência Entidade-Documento - AnaphoraNet	100
Figura 4.7 - Fator intermediário WeightFactor na Função AnaphoraNET	100
Figura 4.8 - Co-ocorrência Entidade-Entidade algoritmo AnaphoraNET	101
Figura 4.9 - Processamento do Algoritmo AnaphoraNET	101
Figura 4.10 - Caixa de seleção de arquivos para pesquisa - sistema DataAssociations	105
Figura 4.11 - Caixa de diálogo para ativação de dicionários Sistema DataAssociations	105
Figura 4.12 - Extração consolidada de entidades extraídas pelo sistema DataAsociations	106
Figura 4.13 - Painel de opções de algoritmos de extração para o sistema DataAssociations	106
Figura 4.14 - Matriz de frequência das entidades nos documentos - sistema DataAssociations	107
Figura 4.15 - Matriz de frequências consolidadas entre entidades - sistema DataAssociations	107
Figura 4.16 – Dados normalizados das frequências entre entidades - sistema DataAssociations	108
Figura 4.17 - Seleção de relacionamentos entre entidades sistema Associations - Formato tabela.	109
Figura 4.18 - Seleção de relacionamentos entre entidades sistema Associations - Formato estrela	110
Figura 5.1- Representação de relacionamentos: formato lista e formato arvore hiperbólica (Xiang et al., 2005)	114
Figura 5.2 - Tipos de estruturas de redes criminais (Arquilla & Ronfeldt, 1996).	115
Figura 5.3 - Grafo de cinco arcos de Freeman	118
Figura 5.4 - Exemplos de Centralidades de Grafos em Estrela . Freeman (1977)	120
Figura 5.5 - Exemplos de Centralidade não representada por maior grau de conectividade. Moody (2003)	121
Figura 5.6 - Indicadores de Centralidade e dispersão em Grafos	122

Figura 5.7 - Centralidade em grafos Direcionados (White & Borgatti, 1994)	125
Figura 5.8 - Clusterização através de divisão e árvore expandida (Kohonen, 1997)	126
Figura 5.9 - Vista parcial de uma pré-competição SOM	128
Figura 5.10 - Dois nós indiretamente conectados (A e C) (Xu & Chen, 2004)	131
Figura 5.11 - Método proposto para extração de recursos para análise	134
Figura 5.12 - Matriz de relacionamentos contendo associações pré-calculadas	135
Figura 5.13 - Extração de árvore de relacionamentos	138
Figura 5.14 – Painel principal do Sistema AnaphoraVisual	144
Figura 5.15 – Painel Gráfico de sistema Visual SOM	146
Figura 5.16 – Painel de Funções Visual SOM	147
Figura 5.17 - Vetores Normalizados Visual SOM	148
Figura 6.1 - Matrizes preparatórias para geração de um arquivo cenário (caso Narcotráfico)	154
Figura 6.2 - Matriz de relacionamentos representativa do Mapa de Inteligência	155
Figura 6.3 - Análise de configurações usando arvores de relacionamentos	157
Figura 6.4- Configurações relacionadas com clusters criminais	158
Figura 6.5 - Classificação de subgrupos criminais usando SOM	159
Figura 6.6 - Caminhos mais próximos entre armamentos	161
Figura 6.7 - Caminhos mais próximos entre entidades	162
Figura 6.8 - Análise de Centralidades	163
Figura 6.9 - Análise das mais fortes conexões entre Tim Lopes e Entidades Nominais	165
Figura 6.10 - Identificação de autoria do assassinato de Tim Lopes	165
Figura 6.11 - Identificação de pessoas através da função de densidade	166
Figura 6.12 - Interseção entre entidades na busca por logradouros	167
Figura 6.13 - Configuração organizacional do subgrupo criminal de Elias Pereira	167
Figura 6.14 - Identificação da entidade de mais forte conexão com uma vítima	168

Figura 6.15 - Identificação de cumplicidade usando função de interseção (I)	169
Figura 6.16 - Identificação de cumplicidade usando função de interseção (II)	170
Figura 6.17 - Identificação de cumplicidade usando função de interseção (III)	170
Figura 9.1 Algoritmo Stemming para o idioma Inglês	187
Figura 9.2 - Desmembramento de palavras em Português, segundo as regiões de substituição de Porter	189
Figura 9.3 - Algoritmo Stemming para o idioma Português	190
Figura 9.4 - Matriz de Alcance contendo trilha reversa de caminhos entre entidades	201
Figura 9.5- Exemplo de uma trilha para cálculo dos mais fortes vínculos entre duas entidades selecionadas	201

Lista de tabelas

Tabela 3.1 - Exemplos de palavras chaves especializadas em domínios temáticos	37
Tabela 3.2 - Regras de reduções	41
Tabela 3.3 - Lista parcial de regras aplicadas ao algoritmo Lovins	47
Tabela 3.4 - Comparação entre os algoritmos Porter (1997) e Lovins (1998)	48
Tabela 3.5 - Resultados computados no experimento de aferição do algoritmo	49
Tabela 3.6 - Regras associadas ao algoritmo Orengo e Huyck .	61
Tabela 3.7 - Resultados apurados no experimento da comparação entre Algoritmo automatizado e extração manual	68
Tabela 3.8 - Resultados do teste comparativo entre algoritmos Orengo e Huyck e Anaphora RV	71
Tabela 4.1 - Exemplo de propriedades associadas a uma pessoa	83
Tabela 4.2 - Chaves convergentes - Dicionário de nomes	102
Tabela 5.1 - Tabela de Vetores classificados saída texto	149
Tabela 6.1 - Configuração dos Dicionários Temáticos para os estudos de casos	153
Tabela 6.2 - Conteúdo parcial de um arquivo cenário	154
Tabela 6.3 - Listagem parcial de entidades vinculadas com comando vermelho	156
Tabela 6.4 - Listagem parcial de áreas vinculadas a organizações criminais	160
Tabela 6.5 - Lista parcial do armamento utilizado	160