UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

PROGRAMA DE PÓS-GRADUAÇÃO EM PSIQUIATRIA E CIÊNCIAS DO COMPORTAMENTO

PLANO DE ATIVIDADES DE PÓS-DOUTORADO

Análises avançadas de dados brasileiros sobre drogas com inteligência artificial e translação para a clínica

Prof. Dr. Eduardo Nunes Borges
Pós-Doutorando

Prof. Dra. Lisia Von Diemen
Supervisora

Rio Grande, julho de 2018.

1 IDENTIFICAÇÃO DO PROJETO

1.1 Título do Projeto

Análises avançadas de dados brasileiros sobre drogas com inteligência artificial e translação para a clínica

1.2 Número de Projeto na UFRGS

35298

1.3 Programa

Programa de Pós-Graduação em Psiquiatria e Ciências do Comportamento

1.4 Período

06/08/2018 – 05/08/2019

1.5 Finalidade

Pesquisa e ensino.

1.6 Linha de Pesquisa/departamento ao que se vincularão as atividades

Instrumentos de obtenção de dados sobre uso de álcool e drogas

1.7 Local de Realização

Centro de Pesquisa em Álcool e Drogas - HCPA/UFRGS

1.8 Professora Supervisora

Profa. Dra. Lisia Von Diemen

2 IDENTIFICAÇÃO DO BOLSISTA

2.1 Nome completo

Eduardo Nunes Borges

2.2 Formação

Doutorado (2013) e Mestrado (2008) em Computação pela UFRGS; Graduação em Engenharia de Computação pela FURG (2005)

2.3 E-mail

enborges@hcpa.edu.br; eduardoborges@furg.br

2.4 Telefone de contato

(53) 98118-7775; (51) 3359-6488

2.5 Link do currículo Lattes

http://lattes.cnpq.br/5851601274050374

3 QUALIFICAÇÃO DOS PROBLEMAS INVESTIGADOS

Banco de dados são conjuntos de itens de dados organizados de forma a facilitar sua recuperação (SILBERSCHATZ, SUNDARSHAN & KORTH, 2016). Estes itens geralmente são representados como registros identificados por um atributo chave que garante unicidade. Um dos principais problemas na integração de banco de dados são as múltiplas representações da mesma informação proveniente de diversas fontes. Este problema afeta diretamente a qualidade das aplicações e serviços principalmente devido à redundância e desatualização das informações (DATE, 2004).

Considere o seguinte exemplo. Com o crescente uso da internet e da *Web* na sociedade atual, as pessoas tendem a acumular diversas contas de *e-mail* e de redes sociais. Além disso, o avanço da tecnologia tem proporcionado o acesso a todas essas informações de qualquer lugar, a partir de *smartphones* e *tablets*. Ao buscar pelo nome de um contato, uma série de registros são apresentados. Com a enorme quantidade de dados provenientes de múltiplas fontes, o usuário obtém diversas informações repetidas e muitas vezes incompletas, tendo dificuldade em selecionar o registro adequado.

Segundo Lenzerini (2002), uma solução de Integração de Dados deve estabelecer métodos específicos que implementem as seguintes tarefas:

1. importar os registros de dados de diferentes fontes heterogêneas;
2. transformar os dados de forma a obterem uma representação comum, ou seja, um esquema compatível;
3. identificar aqueles registros semanticamente equivalentes, representando o mesmo objeto digital;
4. mesclar as informações provenientes das múltiplas fontes;
5. apresentar ao usuário final o conjunto de registros sem informação duplicada.

A Integração de Dados também envolve a tarefa de limpeza e organização dos dados (RAHM & DO, 2000). Nesta etapa os erros e inconsistências são removidas com o objetivo de incrementar a qualidade dos dados.

A tarefa de identificar em um repositório de dados registros duplicados semanticamente equivalentes, ou seja, que se referem a mesma entidade do mundo real, incluindo variações de grafia e omissão de palavras, é denominada deduplicação (BORGES et al., 2011). Conhecida também como casamento de registros, de objetos ou de instâncias, a deduplicação é a descoberta de registros correspondentes em uma ou mais fontes de dados.

Nos últimos anos, diversos métodos foram propostos para a identificação de registros duplicados no contexto da integração de dados relacionais (WHANG & GARCIA-MOLINA, 2013; KIM & LEE, 2012; CARVALHO et al., 2012; WANG, LI & FENG, 2012; FREITAS et al., 2010; DORNELES et al., 2009). Destacam-se métodos que estimam a similaridade entre os registros através de técnicas de aprendizado de máquina (BIANCO et al., 2013; CARVALHO et al., 2012; BORGES et al., 2011; FREITAS et al. 2010; CHEN, KALASHNIKOV & MEHROTRA, 2009;

VERYKIOS, ELMAGARMID & HOUSTIS, 2000; TEJADA, KNOBLOCK & MINTON, 2000; ZHAO & RAM, 2008; BILENKO & MOONEY, 2003).

O aprendizado de máquina é um subcampo da Inteligência Artificial, intimamente ligado à estatística computacional, que estuda o planejamento e a construção de modelos complexos e algoritmos para análise preditiva e descritiva de dados. O principal objetivo do aprendizado é gerar conhecimento de forma automatizada e apoiar na tomada de decisão. Os modelos de aprendizagem capturam relações entre variáveis, geralmente implícitas, que são usadas para explicar determinados comportamentos nos dados. Aliado a diferentes técnicas de visualização de informações é uma poderosa ferramenta para análise de dados científicos.

No contexto do projeto de pesquisa alvo do presente Plano de Atividades de Pós-Doutorado, o aprendizado de máquina será amplamente utilizado. Por exemplo, através de diferentes algoritmos de regressão, pretende-se estimar o tempo de internação de pacientes usuários de substâncias psicoativas, ou seja, prever a adesão ao tratamento, com base em informações coletadas durante a internação utilizadas no estudo de validação do protocolo *Addiction Severity Index* versão 6 para o Brasil (ASI-6). Também pretende-se calcular a probabilidade de abandono precoce em determinado período de tempo (em dias ou semanas), com base em características sociodemográficas e fatores clínicos, biológicos e psicossociais.

As variáveis preditoras necessárias para a modelagem podem estar distribuídas em diferentes bancos de dados, em formatos distintos. Portanto, pode ser necessária a integração de dados de múltiplas fontes.

O restante do texto está organizado da seguinte forma. Na Seção 4 é descrita em detalhes a metodologia a ser empregada durante o estágio pós-doutoral. A Seção 5 apresenta as atividades de ensino e pesquisa propostas, relacionando os bancos de dados a serem utilizados e a descrição das análises planejadas. Por fim, a cronologia das atividades propostas é apresentada na Seção 6.

4 METODOLOGIA A SER EMPREGADA

Descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases – KDD*) é uma metodologia não trivial de identificar padrões potencialmente úteis e compreensíveis em meio às observações presentes em uma base de dados (FAYYAD, PIATETSKY-SHAPIRO & SMYTH, 1996). Geralmente, esses padrões são extraídos de relacionamentos implícitos entre os dados analisados. Como resultado, os padrões encontrados devem gerar conhecimento inteligível e imediatamente utilizável para o apoio às decisões.

A descoberta de conhecimento é dividida nas seguintes fases:

- Seleção de dados - escolha do conjunto de dados contendo todas as possíveis variáveis (atributos) e observações (registros) que farão parte da análise. Esta fase pode ser bem complexa, uma vez que os dados podem ser extraídos de fontes distintas e heterogêneas (bancos de dados, *data warehouses*, planilhas, textos, páginas Web, etc.) e ainda podem possuir os mais diversos formatos.
- Pré-processamento - limpeza e normalização dos dados. Inclui outras tarefas como remoção de ruído, tratamento de registros incompletos e remoção de redundância.
- Transformação dos dados - adequação dos dados em relação à técnica e algoritmo de mineração a serem utilizados. Esta fase inclui a escolha da representação dos dados e a redução de dimensionalidade (número de atributos).
- Mineração - análise automática dos dados em busca de padrões utilizando algoritmos de mineração de dados.
- Interpretação de resultados - fase final responsável pela geração de conhecimento baseada nos padrões encontrados.

Caso não sejam encontrados resultados relevantes em quaisquer fases, o processo deve retornar a uma das fases anteriores.

A Mineração de Dados é a principal fase da metodologia de descoberta de conhecimento em bases de dados. Pode ser definida como a análise automática dos dados em busca de padrões. É apoiada por algoritmos de aprendizado de máquina, estatísticas e técnicas de visualização.

Dependendo do objetivo da mineração, diversas técnicas podem ser utilizadas:

- Associação - definição de regras do tipo A → B, onde A e B são elementos que coocorrem em diferentes observações da base de dados. Uma das aplicações clássicas da associação consiste na descoberta de produtos que são comprados juntos. A análise dos resultados pode determinar que ações devem ser tomadas para incrementar a venda de B, que produtos são afetados pelo produto A e que promoções podem incluir A para incrementar a venda de B.
- Descoberta de padrões sequenciais - definição de regras do tipo (A)(C) → B onde os pares AB e CB são elementos que coocorrem em diferentes observações

da base de dados, sendo que CB ocorre após AB. Portanto, é possível descobrir que um cliente que comprou o produto A e logo depois comprou o produto C, possivelmente comprará o produto B.

- Agrupamento (*clustering*) - classificação não supervisionada de registros em grupos. É realizado com base na similaridade entre os registros. Deve-se maximizar a similaridade intragrupo e minimizar a similaridade intergrupo. Exemplos de métodos de agrupamento são: particionamento, hierárquico e incremental.
- Classificação - método supervisionado que determina um modelo para um determinado atributo que é função dos valores dos outros atributos. Pode ser utilizado para predizer se uma nova instância fará parte de uma determinada classe.
- Regressão - predição do valor de uma variável contínua baseado no valor de outras variáveis, considerando um modelo de dependência linear ou não linear.
- Detecção de desvios - identificação de desvios significativos em relação ao comportamento normal dos dados. Pode ser utilizada na detecção de fraudes, gargalos de um sistema, etc.

Para cada técnica, diferentes algoritmos de aprendizado de máquina foram propostos e podem ser utilizados para as mais diversas análises. A

Tabela 1 apresenta exemplos dos principais algoritmos da literatura.

Tabela 1: Exemplos de algoritmos de aprendizado de máquina para diferentes tarefas de mineração de dados.

Tipo de tarefa	Tarefa	Algoritmos
Descritiva	Associação	Apriori (AGRAWAL & SRIKANT, 1994) ECLAT (ZAKI, 2000) Fuzzy Apriori (MANGALAMPALLI & PUDI, 2009)
	Descoberta de padrões sequenciais	GSP (SRIKANT & AGRAWAL, 1996)
	Agrupamento	COBWEB (FISHER, 1987) DBSCAN (ESTER, 1996) EM (DEMPSTER, LAIRD & RUBIN, 1977) k-means (ARTHUR & VASSILVITSKII, 2007) OPTICS (ANKERST, 1999) PAM (ROUSSEEUW & KAUFMAN, 1990)
Preditiva	Classificação	C4.5 (QUINLAN, 1993) GLM (MCCULLAGH & NELDER, 1989) k-NN (AHA, KIBLER & ALBERT, 1991) MPL (HAYKIN, 2007) Naïve Bayes (JOHN & LANGLEY, 1995), RIPPER (COHEN, 1995) SVM (BOSER, GUYON & VAPNIK, 1992)
	Regressão	
	Detecção de desvios	

5 PROPOSTA DE ATIVIDADES

As atividades propostas para serem desenvolvidas junto ao Centro de Pesquisa em Álcool e Drogas – CPAD estão distribuídas nas categorias pesquisa e ensino.

5.1 Atividades de ensino

O objetivo principal das atividades de ensino propostas é capacitar recursos humanos alocados no projeto em relação aos temas e conceitos relacionados ao aprendizado de máquina e à descoberta de conhecimento em bases de dados.

As atividades de ensino propostas são sintetizadas a seguir:

1. Ministrar para estudantes e pesquisadores do CPAD envolvidos no projeto eventuais minicursos sobre tarefas de mineração de dados, técnicas específicas de aprendizado, métricas de desempenho de algoritmos e/ou ferramentas de análise (*software*);
2. Coorientação do mestrando Vinicius Serafini Roglio, orientado pelo professor Felix Henrique Paim Kessler, e de futuros estudantes de pós-graduação, principalmente quanto ao planejamento de análises e experimentos computacionais, métodos e melhores práticas da descoberta de conhecimento em bases de dados.

5.2 Atividades de pesquisa

O objetivo principal das atividades de pesquisa propostas neste documento é, através de técnicas avançadas de análise estatística e computacional, principalmente baseadas em aprendizado de máquina, ajustar uma série de modelos que avaliem preditores de adesão ao tratamento em usuários de substâncias; e preditores de risco para beber e dirigir, e conduzir com excesso de velocidade.

Estes modelos serão treinados sobre conjuntos de dados construídos a partir da integração de parte dos seguintes bancos de dados, coletados ao longo dos últimos 10 anos pelo CPAD[1]. Na área de assistência em álcool e drogas:

[1] Todos os bancos de dados são oriundos de projetos aprovados pelo Comitê de Ética em Pesquisa (CEP) do HCPA e incluem apenas participantes que assinaram os respectivos termos de consentimento livre e esclarecido.

1. Estudo de validação do *Addiction Severity Index* (ASI-6), aprovado pelo CEP sob o registro 05-460, inclui 700 participantes com diagnóstico de dependência de cocaína inalada e crack, *cannabis* e álcool. Transtornos psiquiátricos foram mensurados a partir da aplicação do *Mini International Neuropsychiatric Interview* (MINI) e a gravidade da dependência de drogas foi mensurada pelo ASI-6.
2. Avaliação, gerenciamento de caso e seguimento de usuários de crack que se encontram em tratamento em seis estados brasileiros, aprovado pelo CEP sob o registro 10-0193, conta com 750 participantes com diagnóstico de dependência de crack e 216 controles saudáveis, além de um segundo conjunto de 90 adolescentes. Transtornos psiquiátricos e gravidade de dependência foram mensurados com múltiplos instrumentos.
3. Preditores clínicos biológicos e psicossociais da recaída precoce em usuários de crack e álcool, aprovado pelo CEP sob o registro 14-0249, inclui 1038 participantes usuários de diferentes drogas. Os registros são compostos por protocolos ASI 6, Entrevistas Clínicas Estruturadas, questionários de trauma na infância, resiliência e de impulsividade, e amostras de sangue.
4. Ensaio Clínico Randomizado, Duplo-Cego, Controlado com Placebo, para Avaliar o Efeito da N-Acetilcisteína no Tratamento dos Transtornos por Uso de Álcool e Cocaína, aprovado pelo CEP sob o registro 15-0488, inclui mais de 27 pacientes com avaliação de instrumentos psicológicos e amostras biológicas.
5. Cocaínas fumáveis na Argentina, Brasil, Chile e Uruguai. Estudo multicêntrico sobre alterações da função cerebral em usuários de crack, aprovado pelo CEP sob o registro 15-0454, inclui mais de 25 indivíduos com baterias de avaliação psiquiátrica e neuropsicológica além de exames de ressonância magnética funcional.

Já na área de trânsito, álcool e drogas:

6. Projeto *Road Safety* 10 - Vida no Trânsito, contém mais de 460 mil observações de velocidade de veículos e 9 mil entrevistas de motoristas, com informações sobre comportamentos de risco.
7. Tecnologias de *Screening* de SPAs no Trânsito, aprovado pelo CEP sob o registro 14-0685, inclui de 178 indivíduos informações sociodemográficas, perfil do motorista, histórico do uso de drogas, alcoolemia, amostras biológicas e confirmações laboratoriais das tecnologias de *screening*.

A Tabela 2 sumariza os conjuntos de dados a serem analisados com diferentes técnicas de aprendizado de máquina.

Com base nas etapas da metodologia de descoberta de conhecimento em bases de dados, apresentada na seção anterior, são propostas as seguintes atividades:

- Levantamento bibliográfico sobre o assunto investigado, como forma de verificar os trabalhos desenvolvidos ou em desenvolvimento por diferentes grupos de pesquisa;
- Análise exploratória dos diversos bancos de dados utilizados no projeto;
- Limpeza e transformação dos dados oriundos dos diversos bancos;

Tabela 2: Sumário dos bancos de dados utilizados nas atividades de pesquisa propostas.

Banco	Observações	Registro
1	700	05-460
2	1056	10-0193
3	1038	14-0249
4	*27	15-0488
5	*25	15-0454
6	469.000	07-069
7	178	14-0685

* Coleta em andamento, valores atualizados em dezembro de 2017.

- Integrar dados de diferentes fontes com o objetivo de prepará-los para análise;
- Planejamento das seguintes análises;
    - Discriminar perfis de condutores associados aos seguintes comportamentos de risco no trânsito: dirigir acima da velocidade permitida, sem habilitação, após consumo de álcool, sem cinto de segurança, ou usando telefone celular.
    - Relacionar comorbidades psiquiátricas, biomarcadores, características sociodemográficas, perfil do consumo de substâncias, traumas na infância, transtornos do eixo I e escores de impulsividade com os desfechos:
        - perfis de condutores descobertos na modelagem anterior;
        - tipo de alta hospitalar (a pedido do médico ou do paciente), visando determinar um perfil para pacientes com maior probabilidade de adesão ao tratamento;
        - estresse precoce, integrando os dados com outro conjunto coletado por pesquisadores da Pontifícia Universidade Católica do Rio Grande do Sul;
        - prisão, modelando individualmente homens e mulheres, devido às diferenças no envolvimento com atividades criminosas;
        - infecção por HIV;
        - tentativa de suicídio;
        - óbito, integrando os dados com um banco de dados de mortalidade a ser solicitado junto ao HCPA.
    - Modelar os acidentes de trânsito usando georreferenciamento e visualização sobre a superfície da Terra, de modo a prever zonas de maior risco.
    - Predizer o uso de drogas no trânsito com base em comportamentos de risco ao dirigir.

- Análise com modelagem preditiva, usando técnicas de aprendizado de máquina supervisionado como classificação e regressão;
- Análise com modelagem descritiva, usando técnicas de aprendizado de máquina não supervisionado como associação e agrupamento;
- Avaliação e interpretação dos modelos;
- Divulgação da proposta para a comunidade científica.
    - Redação de relatórios técnicos;
    - Produção de resumos e materiais para apresentação em eventos científicos;
    - Submissão de artigos científicos em coautoria com o grupo de pesquisa para periódicos especializados com avaliação adequada segundo o Qualis CAPES, nas áreas Medicina II e/ou Ciência da Computação.

6 CRONOGRAMA DE ATIVIDADES

A Tabela 3 apresenta as atividades descritas na sessão anterior distribuídas ao longo de seis bimestres.

Tabela 3: Cronologia das atividades propostas.

Categoria	Atividade	Ago-Set 2018	Out-Nov 2018	Dez-Jan 2019	Fev-Mar 2019	Abr-Mai 2019	Jun-Jul 2019
Ensino	Capacitação	x			x		
	Coorientação	x	x	x	x	x	x
Pesquisa	Levant. bibliográfico	x		x			
	Análise exploratória	x					
	Limpeza	x		x			
	Integração	x		x	x		
	Planejamento	x		x	x		
	Modelagem preditiva		x		x	x	
	Modelagem descritiva		x		x	x	
	Interpretação			x		x	x
	Divulgação			x		x	x

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: International Conference on Very Large Data Bases, 1994, p. 478-499.

AHA, D.; KIBLER, D.; ALBERT, M. Instance-based learning algorithms. Machine Learning, v.6, n.1, p.37–66, 1991.

ANKERST, Mihael et al. OPTICS: Ordering Points To Identify the Clustering Structure. In: ACM SIGMOD International Conference on Management of Data, 1999, p. 49-60.

ARTHUR, D.; VASSILVITSKII, S. k-means++: the advantages of carefull seeding. In: ACM-SIAM Symposium on Discrete Algorithms, Proceedings… p. 1027-1035, 2007.

BIANCO, G. et al. Tuning large scale deduplication with reduced effort. In: International Conference on Scientific and Statistical Database Management, New York. Proceedings… ACM, 2013. p.18:1–18:12.

BILENKO, M.; MOONEY, R. J. Adaptive duplicate detection using learnable string similarity measures. In Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining, pp. 39–48, 2003.

BORGES, Eduardo N.; BECKER, Karin; HEUSER, Carlos A.; GALANTE, Renata de Matos. A Classification-based Approach for Bibliographic Metadata Deduplication. In: IADIS International Conference WWW/Internet, Rio de Janeiro, Proceedings… IADIS, 2011. p. 221-228.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: Computational Learning Theory, Proceedings…, 1992. p.144–152.

CARVALHO, M. G. et al. A Genetic Programming Approach to Record Deduplication. IEEE Transactions on Knowledge and Data Engineering, Piscataway, NJ, USA, v.24, n.3, p.399–412, 2012.

CHEN, Z.; KALASHNIKOV, D. V.; MEHROTRA, S. Exploiting context analysis for combining multiple entity resolution systems. In: ACM SIGMOD International Conference on Management of Data Conference, New York, NY, USA. Proceedings... ACM, 2009. p.207–218.

COHEN, William W. Fast effective rule induction. In: Machine Learning Proceedings. 1995. p. 115-123.

DATE, Christopher J. Introdução a sistemas de bancos de dados. Elsevier Brasil, 2004.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), p. 1-38, 1977.

DORNELES, C. F. et al. A strategy for allowing meaningful and comparable scores in approximate matching. Information Systems, v.34, n.8, p.673–689, 2009.

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining. 1996. p. 226-231.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, v.39, n.11, p.27–34, 1996.

FISHER, Douglas H. Knowledge acquisition via incremental conceptual clustering. Machine learning, v. 2, n. 2, p. 139-172, 1987.

FREITAS, J. et al. Active Learning Genetic programming for record deduplication. In: IEEE Congress on Evolutionary Computation. Proceedings..., 2010. p.1–8.

HAYKIN, S. Neural Networks: a comprehensive foundation. Upper Saddle River, USA: Prentice-Hall, Inc., 2007.

JOHN, G.; LANGLEY, P. Estimating Continuous Distributions in Bayesian Classifiers. In: Conference in Uncertainty in Artificial Intelligence, Proceedings…, 1995. p.338–345.

KIM, J.; LEE, H. Efficient Exact Similarity Searches Using Multiple Token Orderings. In: IEEE International Conference on Data Engineering. Proceedings..., 2012. p.822–833.

LENZERINI, M. Data integration: a theoretical perspective. In: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Proceedings... New York: ACM, 2002. p. 233-246.

MANGALAMPALLI, A.; PUDI, V. Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets. In: IEEE International Conference on Fuzzy Systems, Proceedings…, 2009. p. 1163-1168.

MCCULLAGH, Peter; NELDER, John A. Generalized linear models. CRC press, 1989.

QUINLAN, J. R. C4.5: programs for machine learning. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1993.

RAHM, E.; DO, H. H. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., v. 23, n. 4, p. 3-13, 2000.

ROUSSEEUW, Peter J.; KAUFMAN, L. Finding groups in data. Series in Probability & Mathematical Statistics, v. 34, n. 1, p. 111-112, 1990.

SILBERSCHATZ, A.; SUNDARSHAN, S.; KORTH, H. F. Sistema de banco de dados. Elsevier Brasil, 2016.

SRIKANT, R.; AGRAWAL, R. Mining sequential patterns: Generalizations and performance improvements. In: International Conference on Extending Database Technology. Springer, Berlin, Heidelberg, 1996. p. 1-17.

TEJADA, S.; KNOBLOCK, C. A. and MINTON, S. Learning object identification rules for information integration. Information Systems, v. 26, n. 8, 2001, 607–633.

VERYKIOS, V. S.; ELMAGARMID, A. K.; HOUSTIS, E. N. Automating the approximate record-matching process. Information Sciences, v.126, n.1-4, p.83–98, 2000.

WANG, J.; LI, G.; FENG, J. Can We Beat the Prefix Filtering? an adaptive framework for similarity join and search. In: ACM SIGMOD International Conference on Management of Data, New York. Proceedings... ACM, 2012. p.85–96.

WHANG, S.; GARCIA-MOLINA, H. Joint entity resolution on multiple datasets. The VLDB Journal, v.22, n.6, p.773–795, 2013.

ZAKI, M. J. Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, v. 12, n. 3, p. 372 – 390, 2000.

ZHAO, H.; RAM, S. Entity matching across heterogeneous data sources: an approach based on constrained cascade generalization. Data & Knowledge Engineering, v.66, n.3, p.368 – 381, 2008.