Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 12, n. 3
set-dez 2013

Editorial

EDITORIAL
Alexandre Reis Graeml

Foco nas organizações

CRIAÇÃO COLETIVA NA WEB 2.0: UM ESTUDO DE CASO EM UMA EMPRESA BRASILEIRA DE CROWDSOURCING
Leticia Ribeiro Eboli, Luís Antônio da Rocha Dib

OS FATORES QUE EXPLICAM O GRAU DE ACEITAÇÃO DE UM SISTEMA DE INFORMAÇÃO ACADÊMICA: UM ESTUDO DE CASO COM DOCENTES DE UMA IES PRIVADA
Patrícia Nunes Costa Reis, Claudio Pitassi, Marco Aurélio Bouzada

GESTÃO DE TECNOLOGIA DE INFORMAÇÃO: UM MÉTODO DE AVALIAÇÃO DO WMS
Priscilla Cristina Cabral Ribeiro, Nayara Louise Alves de Carvalho

Foco na tecnologia

REDUZINDO O ESFORÇO NA PREPARAÇÃO DE METADADOS: USO DE SOFTWARE LIVRE PARA DOCUMENTAR DADOS ESPACIAIS NO PERFIL MGB
Wagner Dias de Souza, Rafaella da Silva Nogueira, Angélica Aparecida de Almeida Ribeiro, Jarbas Nunes Vidal Filho, Alex da Silva Santos, Jaqueline Alvarenga Silveira, Daniel Camilo de Oliveira Duarte, Jugurta Lisboa Filho

Ensaios

REALIZING EMANCIPATORY IDEALS IN PHENOMENOLOGICAL IS RESEARCH
Valter Moreno, Jr.

Fast track SBSI

INSIDERS: ANÁLISE E POSSIBILIDADES DE MITIGAÇÃO DE AMEAÇAS INTERNAS
Gliner Dias Alencar, Anderson Apolonio Lira Queiroz, Ruy José Guerra Barretto de Queiroz

UMA METODOLOGIA PARA O APRENDIZADO DE UM MODELO CLASSIFICADOR PARA O ALINHAMENTO DE ONTOLOGIAS
Alex Alves, Anselmo Guedes, Kate Revoredo, Fernanda Baiao

A PERSPECTIVA DE ANÁLISE COMPORTAMENTAL COMO FORMA DE COMBATE À ENGENHARIA SOCIAL E PHISHING
Gliner Dias Alencar, Marcelo Ferreira de Lima, André Caetano Alves Firmo

Nominata de avaliadores

Avaliadores ad hoc - 2013

ISSN: 1677-3071

Revista hospedada em: http://revistas.facecla.com.br/index.php/reinfo
Forma de avaliação: *double blind review*

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

UMA METODOLOGIA PARA O APRENDIZADO DE UM MODELO CLASSIFICADOR PARA O ALINHAMENTO DE ONTOLOGIAS

A LEARNING METHODOLOGY FOR A CLASSIFYING MODEL FOR ONTOLOGY ALIGNMENT

(artigo submetido em dezembro de 2013)

Alex Alves
Bacharel em Ciência da Computação pelo Centro Universitário da Cidade e Mestre em Informática pelo Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
alex.alves@uniriotec.br

Anselmo Guedes
Licenciado em Computação pelo Centro Universitário Geraldo Di Biase e aluno de mestrado no Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
anselmo.guedes@uniriotec.br

Kate Cerqueira Revoredo
Doutora em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ) e Professora Adjunta da Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
katerevoredo@uniriotec.br

Fernanda Baião
Doutora em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ) e Professora Associada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
fernanda.baiao@uniriotec.br

ABSTRACT

Ontology alignment is a common and successful way to reduce the semantic heterogeneity among ontologies, relying on the application of similarity functions to decide whether a pair of entities from two input ontologies corresponds to each other. There are several similarity functions proposed in the literature capturing distinct and complementary perspectives, but the challenge is on how to combine their use. This paper presents a methodology to automatically learn a classifier that combines distinct string-based similarity functions for the ontology alignment task, through machine learning. The proposed approach was evaluated experimentally on sixteen scenarios defined on top of the Ontology Alignment Evaluation Initiative (OAEI).

Key-words: ontologies; ontology matching; machine learning; classifier.

RESUMO

O alinhamento de ontologias é uma estratégia comum e que tem sido aplicada com sucesso para reduzir a heterogeneidade semântica entre ontologias de um mesmo domínio. Durante o processo de alinhamento são consideradas diferentes funções de similaridade a fim de selecionar corretamente os pares de entidades correspondentes entre as duas ontologias sendo alinhadas. Existem diversas funções de similaridade, mas o desafio atual está em como combiná-las para gerar alinhamentos de melhor qualidade. Este trabalho apresenta uma metodologia para gerar um modelo classificador, que combina diferentes funções de similaridade baseadas em string no alinhamento de ontologias, por meio de aprendizado de máquina. A abordagem proposta foi avaliada experimentalmente em dezesseis cenários definidos sobre a Iniciativa de Avaliação de Alinhamento de Ontologias (OAEI).

Palavras-chave: ontologias; alinhamento de ontologias, aprendizado de máquina; classificador.

1 INTRODUÇÃO

Ontologia de domínio é uma representação de consenso (ou modelo de referência) de uma conceituação compartilhada (GRUBER, 1993). Ela define os conceitos dos domínios assim como os relacionamentos entre esses e as suas instâncias. O conjunto formado pelos conceitos, relacionamentos e instâncias define as entidades de uma ontologia. Um dos propósitos das ontologias de domínio é minimizar a heterogeneidade conceitual por meio da representação formal do domínio de discurso. Contudo, em sistemas abertos e em constante evolução, como a web semântica, a heterogeneidade não pode ser evitada. Diferentes comunidades utilizando aplicações distintas, com níveis de detalhamento variados, ao contrário, elevam o problema da heterogeneidade para níveis mais altos (EUZENAT e SHVAIKO, 2007).

Para superar esse problema de heterogeneidade semântica, uma das etapas necessárias é a identificação de correspondências entre as entidades das duas ontologias em questão. Essa identificação é feita com o uso de funções de similaridade, que são capazes de medir a força da similaridade entre duas entidades. Geralmente essa força é um valor entre 0 e 1, onde quanto mais próximo de 1, mais similares são as duas entidades. O conjunto das correspondências encontradas define o alinhamento das ontologias (SHVAIKO e EUZENAT, 2005). Diversas soluções de alinhamento de ontologias vêm sendo propostas nos últimos anos, dentre elas técnicas que utilizam aprendizado de máquina (MITCHELL, 1997).

Este trabalho tem como objetivo propor uma metodologia, baseada em aprendizado de máquina, para o alinhamento de ontologias. A metodologia proposta estende nosso trabalho anterior (ALVES *et al.*, 2013), expandindo substancialmente a análise da literatura e a fundamentação teórica, detalhando as técnicas e os passos da metodologia proposta e apresentando resultados utilizando novas métricas de avaliação, para tornar mais clara a análise da qualidade dos classificadores obtidos. A metodologia proposta é avaliada por meio de um experimento exploratório que analisa a importância dos conceitos para o alinhamento (quando comparado às demais entidades da ontologia) e a relevância de diferentes funções de similaridade baseadas em *strings*.

O artigo está estruturado da seguinte forma: Na Seção 2 é apresentada a fundamentação teórica; na Seção 3 a abordagem para geração do classificador; na Seção 4, são descritos os experimentos executados e os resultados obtidos; na Seção 5 são abordados alguns trabalhos relacionados; e na Seção 6 é concluído o trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta conceitos fundamentais sobre a tarefa de alinhamento de ontologias e sobre o processo de descoberta de conhecimento em bases de dados, necessários para a compreensão da abordagem proposta.

doi:10.5329/RESI.2013.1203007

2.1 ALINHAMENTO DE ONTOLOGIAS

Em sistemas abertos e distribuídos, como a web semântica, a heterogeneidade dos dados não pode ser evitada. Diferentes pessoas têm interesses e hábitos diferentes, utilizam ferramentas diferentes e possuem conhecimento, na maioria das vezes, em diferentes níveis de detalhe (SHVAIKO e EUZENAT, 2005; ALVES *et al.*, 2012). Estas várias razões levam a diversas formas de heterogeneidade e, portanto, devem ser cuidadosamente levadas em consideração.

A questão da heterogeneidade também pode ser percebida quando considerando a modelagem de um domínio com base em ontologias (GRUBER, 1993). Uma ontologia muitas vezes pode ser utilizada para definir o vocabulário utilizado por alguma aplicação em particular.

O objetivo do alinhamento de duas ou mais ontologias é reduzir a heterogeneidade existente entre elas. A heterogeneidade não reside exclusivamente nas diferenças entre os objetivos das aplicações, de acordo com o fim para o qual foram concebidas ou nos formalismos expressos nas ontologias nas quais foram codificadas. O alinhamento de ontologias ocorre no sentido de identificar as correspondências entre as entidades individuais de múltiplas ontologias e é uma condição necessária para estabelecer a interoperabilidade entre elas (EHRIG, 2007).

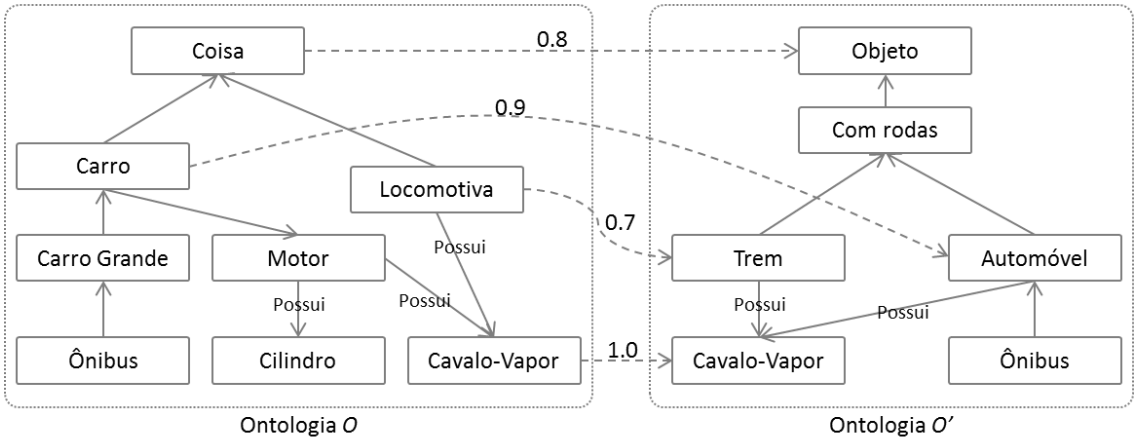

Figura 1. Exemplo de alinhamento entre duas ontologias.
Fonte: adaptado de (ABOLHASSANI *et al.*, 2006).

Por exemplo, considerem-se as duas ontologias (*O* e *O'*) ilustradas na Figura 1, onde as arestas pontilhadas indicam correspondências entre entidades de *O* e *O'*. Neste exemplo, existe uma correspondência entre as entidades *carro* e *automóvel* das ontologias *O* e *O'*, respectivamente. Além disso, é possível indicar o grau de confiança (também chamado de força) dessa correspondência por meio da atribuição de peso às arestas. A confiança da correspondência entre *carro* e *automóvel* é de 0,9.

Segundo EUZENAT e SHVAIKO (2007), tecnicamente o processo de alinhamento de ontologias ocorre a partir de duas ontologias *O* e *O'*, podendo ser adicionado um conjunto de recursos *r*, estabelecido um conjunto de parâmetros *p* e adicionada a entrada de um alinhamento *A*. O

resultado deste processo é um alinhamento *A'* entre as ontologias *O* e *O'*, conforme apresentado na Figura 2, podendo ser representado como *A' = f (O, O', A, p, r)*.

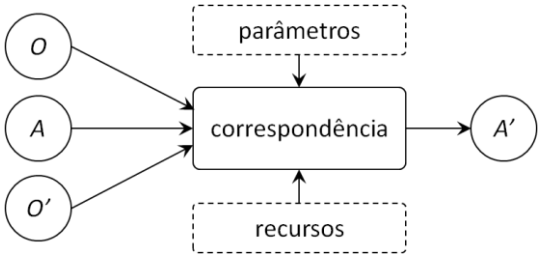

Figura 2. Processo de alinhamento de duas ontologias
Fonte: Euzenat e Shvaiko (2007).

Basicamente, alinhamento de ontologias é um processo em que ligações semânticas entre entidades de ontologias são estabelecidas. Essas ligações semânticas podem ser estabelecidas a partir de funções de similaridade (ver Seção 2.1.1). Como resultado desse processo, é obtido um conjunto dessas ligações semânticas, que são denominadas correspondências. Formalmente, uma correspondência entre entidades de duas ontologias O e O' pode ser definida como uma quádrupla $\langle e, e', r, n \rangle$, onde $e \in O$, $e' \in O'$, r indica o tipo de correspondência, que pode ser de equivalência ou subsunção e $n \in [0,1]$ indica o valor da confiança na correspondência. Quando uma função de similaridade é utilizada esse valor é determinado por essa função.

2.1.1 Funções de similaridade

Funções de similaridade são funções que, dadas duas entidades, indicam a força da correspondência entre elas. Algumas dessas funções são baseadas em *string*, ou seja, comparam as entidades (termos) considerando a sequência de caracteres que os nomeia. Estas funções consideram que a semelhança entre dois termos aumenta quando a semelhança entre as suas sequências de caracteres correspondentes também aumenta. Essas funções se aproveitam da estrutura da *string*. Alguns exemplos são:

- *NameEqAlignment* (JÉRÔME *et al.*, 2011) – realiza o alinhamento de duas ontologias baseado na igualdade do nome de suas entidades, associando 1 quando os nomes forem iguais e 0 caso contrário.
- *Smith e Waterman* (SMITH e WATERMAN, 1981) – determina as regiões semelhantes entre duas sequências de caracteres. Em vez de olhar para a sequência total, compara segmentos de todos os comprimentos possíveis e otimiza a medida de similaridade.
- *Jaro* e *Winkler* (JARO, 1989) – determina a similaridade baseada na distância d_j entre as duas *strings* s1 e s2, da seguinte forma:

$$d_j = \begin{cases} 0, & se\ m = \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{3}}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) & caso\ contrário \end{cases}$$

Onde:

  - m é o número de correlações entre caracteres;
  - |s1| e |s2| são os comprimentos de s1 e s2, respectivamente;
  - t é o número de transposições (alteração na ordem das palavras).
- *Q-GramsDistance* (UKKONEN, 1992) – compara as entidades contando o número de ocorrências de diferentes q-grams entre as *strings* de entrada, onde um q-gram é uma sequência de q caracteres. As *strings* serão mais distantes entre si à medida que o número de q-grams distintos aumentar.

Tais funções, de acordo com a classificação das abordagens de alinhamento proposta por Euzenat e Shvaiko (2007), são técnicas no nível sintático e/ou terminológico.

2.1.2 Métricas para avaliação do alinhamento

A qualidade do alinhamento encontrado considerando uma função de similaridade, uma combinação de funções de similaridade ou um determinado sistema de alinhamento pode ser medida considerando um alinhamento de referência (alinhamento gabarito) (EHRIG, 2007), ou seja, um alinhamento esperado entre as duas ontologias sendo alinhadas. Ao comparar o alinhamento A encontrado com o alinhamento de referência R os seguintes conjuntos podem ser gerados:

- verdadeiros positivos (VP): define o conjunto de correspondências identificadas em A, que realmente são correspondências, ou seja, que fazem parte do conjunto de correspondências de R;
- falsos positivos (FP): define o conjunto de correspondências identificadas em A, que não são correspondências, ou seja, não fazem parte do conjunto de correspondências de R;
- falsos negativos (FN): define o conjunto de pares de entidades que não foram identificados como correspondências em A, mas que deveriam, pois fazem parte do conjunto de correspondências de R.

A partir desses conjuntos, algumas métricas de avaliação de alinhamento podem ser consideradas. São elas:

- precisão (*precision (P)*): proporção de correspondências encontradas que realmente são correspondências.

$$P(A,R) = \frac{[VP]}{[VP+FP]}$$

- cobertura (*recall (R)*): proporção de correspondências existentes que foram encontradas.

$$R(A,R) = \frac{[VP]}{[VP+FN]}$$

- medida-F (*F-measure (F)*): é uma harmonização entre a precisão e a cobertura e tem sido utilizada como medida principal para avaliar a

qualidade de um alinhamento. Pode ser calculada da seguinte forma:

$$F(A,R) = \frac{(b^2+1).P(A,R).R(A,R)}{b^2.P(A,R)+R(A,R)}$$

Com b=1 sendo um fator de peso padrão, chega-se a:

$$F_1(A,R) = \frac{2.P(A,R).R(A,R)}{P(A,R)+R(A,R)}$$

2.2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O processo de descoberta de conhecimento em bases de dados (*knowledge discovery in databases* - KDD) visa a buscar padrões desconhecidos existentes nas bases de dados e envolve diversas áreas de conhecimento, tais como a estatística, inteligência artificial, aprendizado de máquina, banco de dados, reconhecimento de padrões, armazém de dados (*data warehousing*), visualização de dados entre outras (HAN e KAMBER, 2006; GOLDSCHMIDT e PASSOS, 2005). Por meio da exploração das bases de dados, consideradas verdadeiros "depósitos" de conhecimento em potencial, novos conceitos são extraídos gerando um novo conhecimento consistente, útil e compreensível, a ser incorporado ao conhecimento já existente. Segundo Fayyad *et al.* (1996), KDD é um processo composto por várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. A Figura 3 mostra a sequência de passos.

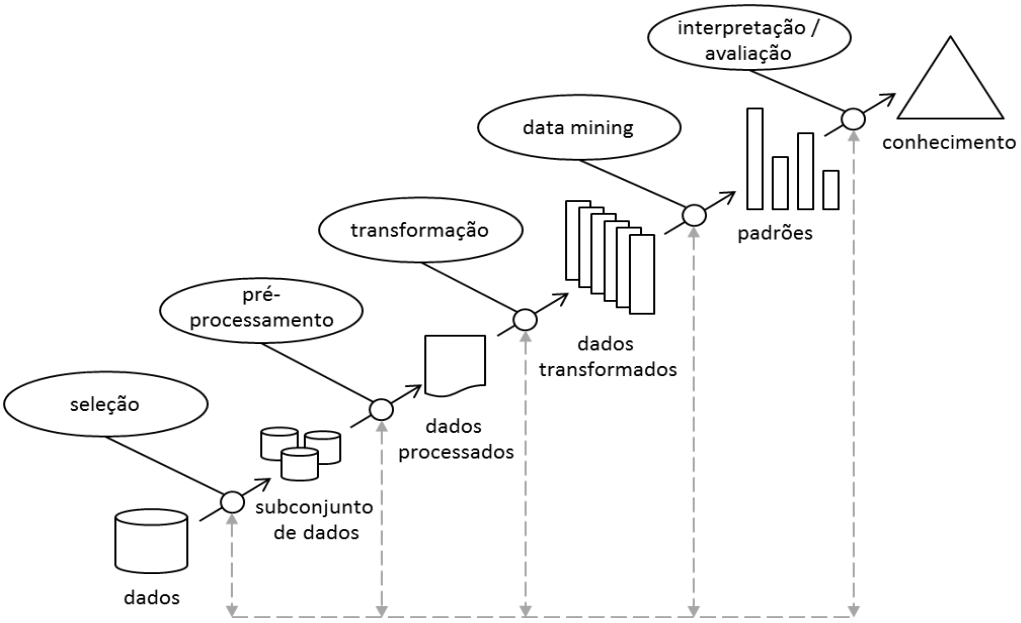

Figura 3. Processo KDD proposto por Fayyad
Fonte: Fayyad *et al.* (1996)

A partir de um repositório de dados que reúne informações históricas sobre o domínio em que se quer tratar (tais informações são denominadas

dados de treinamento), são aplicadas técnicas de pré-processamento, como a limpeza, tratamento de valores ausentes, detecção de ruído dos dados e *outliers*. Também é nessa etapa que podem ser selecionados os dados mais relevantes e que atributos contínuos são normalizados, se necessário. Normalização evita que os atributos com grandes intervalos de valores sejam preteridos em detrimento de outros (HAN e KAMBER, 2006). A próxima etapa, denominada mineração de dados, é quando os algoritmos de aprendizado de máquina (MITCHELL, 1997) são utilizados para extração de padrões existentes nos dados, de forma a aprender um modelo que reflita o conjunto de dados de entrada, explicitando os padrões escondidos. Finalmente, o modelo aprendido passa por um passo de pós-processamento, onde os padrões interessantes são filtrados, apresentados visualmente e interpretados. Sendo um processo iterativo, pode-se voltar ao passo anterior sempre que necessário. O processo é concluído quando o conhecimento for encontrado.

Uma das tarefas mais populares em que algoritmos de aprendizado de máquina são aplicados é a classificação. Na tarefa de classificação, cada elemento do conjunto de dados de entrada (denominado instância) é caracterizado por um conjunto fixo de atributos. Um dos atributos é denominado classe e é o foco do aprendizado. O modelo aprendido tem como objetivo atribuir um valor de classe correto, dado o conjunto de valores dos demais atributos de uma instância.

Na comunidade de Sistemas de Informação, técnicas de mineração de dados já foram aplicadas nos últimos anos em diversos cenários, incluindo a descoberta automática de características espaço-temporais de trajetórias de objetos reais (BRAZ, 2008), sistemas agrícolas na classificação automática do acabamento de gordura de carcaças bovinas em imagens digitais (BITTENCOURT *et al.*, 2008), na detecção automática de cartéis em licitações públicas (SILVA e RALHA, 2011), na detecção do estágio do ciclo de vida de um projeto de inovação (PEDROSO *et al.*, 2013) e no domínio acadêmico (CORUMBA e MACEDO, 2011).

No contexto do presente artigo, o alinhamento de ontologias é tratado como uma tarefa de classificação em que se pretende atribuir um valor de classe binário (1/0, sim/não) a partir das características (atributos) de um par de entidades das ontologias, denotando se tais entidades são correspondentes ou não. Cada característica representa o resultado de uma função de similaridade. Os algoritmos de classificação que foram considerados são descritos a seguir.

Regressão Linear. Quando os atributos são numéricos, a regressão linear é uma técnica natural a considerar. Este é um método básico na estatística. A ideia é expressar a classe como uma combinação linear ponderada entre os atributos (WITTEN *et al.*, 2011).

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

Onde: x é a classe; a_1, a_2, ..., a_k são os valores de atributos, e w_0, w_1, ..., w_k são pesos. Os pesos são calculados a partir dos dados de treina-

mento. A notação fica um pouco pesada, porque é necessária uma forma de expressar os valores de atributos para cada instância de treinamento. A primeira instância terá uma classe, por exemplo, $x^{(1)}$, e valores de atributo $a_1^{(1)}$, $a_2^{(1)}$, ..., $a_k^{(1)}$, onde o expoente indica que se trata do primeiro exemplo. Uma vez que os cálculos tenham sido realizados, o resultado é um conjunto de pesos numéricos, com base nos dados de treinamento, o qual pode ser usado para predizer a classe de novos casos.

Máquina de vetores de suporte. Máquina de vetores de suporte (*support vector machine* - SVM) possui seus fundamentos na teoria de aprendizagem estatística e tem mostrado resultados empíricos promissores em muitas aplicações práticas (WITTEN *et al.*, 2011). Consiste em um método de aprendizado que tenta encontrar um hiperplano ótimo de modo que ele possa separar diferentes classes de dados com a maior margem possível (PANG-NING *et al.*, 2009), chamada *soft margin* como mostra a Figura 4.

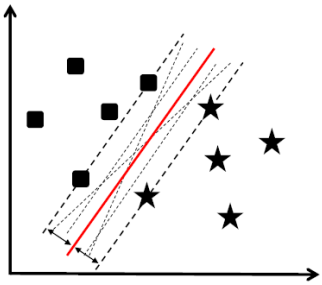

Figura 4. Hiperplano ótimo
Fonte: Lima *et al.* (2009).

O SVM foi originalmente concebido para lidar com classificações binárias, entretanto a maior parte dos problemas reais requer múltiplas classes. Para se utilizar uma SVM para classificar múltiplas classes é necessário transformar o problema multiclasse em vários problemas de classes binárias. Outro fator importante a considerar é que em muitos problemas reais as classes não são linearmente separáveis mesmo utilizando a margem de folga. Nestes casos, SVM mapeia os dados para um espaço de dimensão maior para então procurar pelo hiperplano ótimo.

3 METODOLOGIA PROPOSTA

A metodologia proposta no presente trabalho aprende um modelo classificador que combina funções de similaridade para a obtenção de um alinhamento correto. Este trabalho não se propõe a avaliar nenhuma ferramenta de alinhamento ou função de similaridade individualmente, mas servir como um recurso externo a essas ferramentas para aumentar a qualidade do alinhamento por elas gerado, tomando como base o classificador obtido. É possível ainda perceber a relevância de cada função de similaridade para o alinhamento final.

Para a geração da base de dados que será utilizada para obtenção do classificador, é necessário que alguns passos sejam executados de forma

sistemática. Para tal, propõe-se que sejam executados os passos ilustrados na Figura 5.

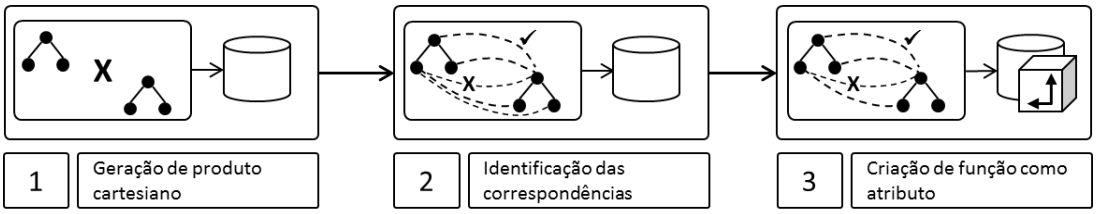

Figura 5. Passos da abordagem proposta
Fonte: elaborada pelos autores

1. Geração de produto cartesiano: é comum que as ferramentas de alinhamento de ontologias retornam apenas as correspondências encontradas entre as ontologias alinhadas. Para o processo de mineração de dados, é importante ter dados que reflitam situações de alinhamento e situações de não alinhamento para que um classificador que distingue as duas situações possa ser aprendido. Assim, as instâncias da base de dados sendo gerada correspondem aos pares do produto cartesiano das entidades das duas ontologias, a fim de gerar uma base com todas as associações possíveis entre as entidades das ontologias *O* e *O'*. O primeiro atributo criado na base é o atributo identificador, que indica que par de entidades a instância representa.

2. Identificação das correspondências: Para o conjunto de pares de entidades (gerado a partir do produto cartesiano entre as ontologias *O* e *O'*), é necessário distinguir os pares que representam uma correspondência. Tal distinção é indicada no atributo de classe da base de dados sendo gerada. Os pares de entidades presentes no alinhamento de referência (gabarito do alinhamento) das ontologias *O* e *O'* consideradas indicam correspondências e as instâncias da base de dados relacionadas a elas terão o valor 1 associado ao atributo de classe. As instâncias dos pares de entidades não observados no alinhamento de referência recebem o valor 0 para o atributo de classe.

3. Criação de função como atributo: cada função de similaridade utilizada para determinar a correspondência entre os pares de entidades é considerada como um atributo na base de dados sendo gerada. Assim, a base de dados é populada com o retorno de cada função, associando o valor retornado ao atributo correspondente para a instância que representa o par de entidades sendo avaliado.

A Tabela 1 ilustra a estrutura final da base de dados gerada considerando as ontologias ilustradas na Figura 1.

Tabela 1. Estrutura do conjunto de dados gerado

Par de entidades	Função 1	Função 2	...	Função N	Classe
(Coisa :: Objeto)	0,00	0,88	...	0,62	1
(Coisa :: Com rodas)	1,00	1,00	...	1,00	0
(Coisa :: Trem)	0,00	0,96	...	1,00	0
(Coisa :: Automóvel)	0,00	0,95	...	1,00	1
(Coisa :: Cavalo-vapor)	0,00	0,96	...	1,00	0
(Coisa :: Ônibus)	0,00	0,40	...	0,95	1
(Carro :: Objeto)	0,00	0,96	...	1,00	0
(Carro :: Com rodas)	0,00	0,96	...	1,00	0
(Carro :: Trem)	0,00	0,95	...	0,67	0
(Carro :: Automóvel)	0,00	0,93	...	0,73	1
(Carro :: Cavalo-vapor)	0,00	0,48	...	0,30	0
(Carro :: Ônibus)	0,00	0,95	...	1,00	1
(Motor :: Objeto)	0,00	0,96	...	1,00	0
(Motor :: Com rodas)	0,00	0,72	...	0,38	0
(Motor :: Automóvel)	0,00	0,79	...	0,60	0
(Motor :: Trem)	0,00	0,30	...	0,67	0
(Motor :: Cavalo-vapor)	1,00	1,00	...	1,00	0

Fonte: elaborada pelos autores

Após a geração da base de dados de acordo com os passos de 1 a 3, o processo de descoberta de conhecimento em bases de dados é executado com o objetivo de aprender um modelo classificador. O atributo identificador é retirado da base, já que não é possível encontrar nenhum padrão a partir de um atributo com valores todos distintos.

4 AVALIAÇÃO EXPERIMENTAL

Para a realização do experimento, foram utilizados dados disponibilizados publicamente pela OAEI[1] (*Ontology Alignment Evaluation Initiative*) (EUZENAT *et al.,* 2011). A OAEI é uma iniciativa internacional coordenada cujo objetivo é avaliar os pontos fortes e fracos das ferramentas de alinhamento, comparar o desempenho de técnicas e melhorar as técnicas de avaliação. Possui como principal meta fomentar a melhoria continua das ferramentas de alinhamento de ontologias. Os conjuntos de dados disponibilizados pela OAEI estão em observância com os critérios necessários para a realização de um projeto de avaliação de alinhamento de ontologias e possuem as características necessárias para a execução do processo apresentado na Seção 3.

[1] http://oaei.ontologymatching.org/

Diante disso, o conjunto de dados *benchmark*, disponibilizado pela campanha da OAEI de 2012, foi selecionado para a execução do experimento. Este conjunto de dados é composto de 51 ontologias, sendo uma base e as restantes variações desta para a execução do alinhamento entre elas. Foram considerados cinquenta pares de ontologias para testar a abordagem proposta. Para cada alinhamento é fornecido um alinhamento de referência, por meio do qual é possível calcular as métricas de avaliação e executar o processo de aprendizado supervisionado.

4.1 METODOLOGIA EXPERIMENTAL

Na metodologia experimental adotada para avaliação deste trabalho, foram variados três eixos: os algoritmos de aprendizado de máquina, o conjunto de atributos selecionados e o conjunto de instâncias selecionadas.

4.1.1 Algoritmos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina utilizados foram os algoritmos para regressão linear e *support vector machines* (SVM), já que são algoritmos que trabalham bem com dados numéricos, como é o caso nesse trabalho.

Um dos objetivos desse experimento é verificar o potencial desses dois algoritmos na tarefa de alinhamento de ontologias, indicando qual deles é o melhor nos cenários considerados.

Ambos os algoritmos foram executados utilizando a ferramenta de mineração de dados Weka[2]. A escolha dessa ferramenta se deve ao fato de estar disponível para uso (licença GPL) e ser amplamente utilizada em trabalhos científicos (WITTEN *et al.*, 2011; THORNTON *et al.*, 2013). O software foi escrito na linguagem Java e contém uma GUI (*Graphical User Interface*) para interagir com arquivos de dados e produzir resultados visuais, o que facilitou o nosso experimento.

4.1.2 Seleção dos atributos

As funções de similaridade consideradas foram as baseadas em *string*, mais especificamente:

1) *NameEqAlignment*
2) *Smith e Waterman*
3) *Jaro* e *Winkler*
4) *QGramsDistance*
5) *StringDistAlignment*
6) *EditDistNameAlignment*
7) *NameAndPropertyAlignment*
8) *SMOANameAlignment*

[2] http://www.cs.waikato.ac.nz/ml/weka/

9) *SubsDistNameAlignment* (EUZENAT *et al.*, 2011)
10) *Trigram* (MASSMANN *et al.*, 2011)
11) *Levenshtein* (LEVENSHTEIN, 1966)
12) *MongeElkan* (MONGE e ELKAN, 1996)

Para a geração das correspondências a partir das funções de 1 a 6, foi utilizada a *Alignment API and Server 4.4*[3], que é uma plataforma extensível de alinhamento de ontologias desenvolvida pelos organizadores do OAEI, caracterizada por um conjunto de abstrações para expressar, acessar e compartilhar alinhamentos de ontologias. A função 7 foi utiliza a partir da ferramenta *COMA Community Edition 3.0* [4], desenvolvida pela Universidade de Leipzig, Alemanha, que é uma ferramenta baseada em *workflow* com a possibilidade de combinação de funções. Por fim, as funções 8 a 12 estão disponíveis no projeto Simmetrics[5], que é uma biblioteca fornecida pela Universidade de Sheffield, no Reino Unido.

Como descrito na Seção 3, cada função corresponde a um atributo na base de dados. Logo, na etapa de pré-processamento do processo de KDD, ao aplicarmos uma heurística para a seleção de atributos, está-se na verdade selecionando conjuntos de funções de similaridade. Além da heurística que seleciona todas as funções de similaridade que estão na base de dados, as seguintes heurísticas foram consideradas para a seleção de atributos: (i) as três funções de similaridade que apresentaram melhores resultados para a métrica de precisão, (ii) as três funções de similaridade que apresentaram melhores resultados para a métrica de cobertura e (iii) as três funções de similaridade que apresentaram melhores resultados da métrica medida-F. Para todas as heurísticas as funções de similaridade foram escolhidas de acordo com a sua avaliação individual para cada uma das três métricas de avaliação.

Ao variar o grupo de funções de similaridade pode-se avaliar o potencial de cada um desses grupos, mostrando o impacto deles no processo de alinhamento. Além disso, pode-se avaliar, no cenário considerado, qual o melhor grupo.

4.1.3 Seleção das instâncias

Como descrito na Seção 3, o conjunto de dados foi gerado considerando o produto cartesiano das entidades das duas ontologias sendo alinhadas. Como as entidades de uma ontologia podem ser os seus conceitos, relacionamentos e instâncias, esse produto cartesiano considera entidades destas três naturezas, diferentemente de outras abordagens em que apenas os conceitos são considerados. Sendo assim, foram especificados cenários de avaliação considerando a base de dados contendo todas as instâncias e a base de dados contendo apenas as instâncias que correspondiam a pares de conceitos. Essa variação permite avaliar o

[3] http://alignapi.gforge.inria.fr/

[4] http://dbs.uni-leipzig.de/Research/coma.html

[5] http://sourceforge.net/projects/simmetrics/

potencial de considerar todas as entidades da ontologia quando comparado a apenas considerar os conceitos.

Portanto, foram executados um total de dezesseis cenários de avaliação, combinando-se os dois algoritmos de aprendizado utilizados, as quatro heurísticas para selecionar o grupo de funções de similaridade e as duas bases de dados utilizadas. Os cenários estão descritos na Tabela 2 e foram considerados aplicando os cinquenta pares de ontologias do *Benchmark 2012* do *OAEI*.

Tabela 2. Cenários executados na metodologia de avaliação da abordagem proposta

Cenário	Descrição
1	Regressão linear considerando todas as funções de similaridade utilizando todas as entidades das ontologias
2	Regressão linear considerando todas as funções de similaridade utilizando somente os conceitos das ontologias
3	SVM considerando todas as funções de similaridade utilizando todas as entidades das ontologias
4	SVM considerando todas as funções de similaridade utilizando somente os conceitos das ontologias
5	Regressão linear considerando as três funções que apresentaram melhores resultados de cobertura utilizando todas as entidades das ontologias
6	Regressão linear considerando as três funções que apresentaram melhores resultados de cobertura utilizando somente os conceitos das ontologias
7	SVM considerando as três funções que apresentaram melhores resultados de cobertura utilizando todas as entidades das ontologias
8	SVM considerando as três funções que apresentaram melhores resultados de cobertura utilizando somente os conceitos das ontologias
9	Regressão linear considerando as três funções que apresentaram melhores resultados de precisão utilizando todas as entidades das ontologias
10	Regressão linear considerando as três funções que apresentaram melhores resultados de precisão utilizando somente os conceitos das ontologias
11	SVM considerando as três funções que apresentaram melhores resultados de precisão utilizando todas as entidades das ontologias
12	SVM considerando as três funções que apresentaram melhores resultados de precisão utilizando somente os conceitos das ontologias
13	Regressão linear considerando as três funções que apresentaram melhores resultados de medida-F utilizando todas as entidades das ontologias
14	Regressão linear considerando as três funções que apresentaram melhores resultados de medida-F utilizando somente os conceitos das ontologias
15	SVM considerando as três funções que apresentaram melhores resultados de medida-F utilizado todas as entidades das ontologias
16	SVM considerando as três funções que apresentaram melhores resultados de medida-F utilizado somente os conceitos das ontologias

Fonte: elaborada pelos autores

O método utilizado para avaliação dos resultados foi a validação cruzada, que é utilizado para a validação do modelo. Em ambos os experimentos ele foi configurado em 10 *folds*, ou seja, os dados foram igual-

mente particionados (10 conjuntos) e cada *fold* foi separado dos demais onde serviu para teste do modelo gerado.

Para avaliar os classificadores, utilizou-se o erro médio absoluto (*mean absolute error – MAE*), o qual mede o quão perto a predição do alinhamento está do alinhamento de referência. Ele é definido da seguinte forma:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|p_i - r_i\right|,$$

onde n é o número de instâncias, p_i é o valor predito pelo classificador para a instância i, e r_i é o valor de classe (proveniente do alinhamento de referência) para a instância i.

4.2 ANÁLISE DOS RESULTADOS

Todos os modelos gerados apresentaram bons resultados, considerando a métrica do erro médio absoluto. É interessante observar que os modelos aprendidos a partir de dados de treinamento considerando todas as entidades da ontologia (conceitos e relacionamentos) apresentaram resultados melhores do que os modelos aprendidos a partir de dados de treinamento considerando apenas conceitos, o que seria a princípio mais intuitivo. Esse fenômeno pode ter ocorrido pelo tamanho do conjunto de dados, onde o conjunto de dados considerando pares entre todas as entidades possui 34.225 instâncias, ao passo que o conjunto de dados formado apenas por pares entre conceitos possui 2.618 instâncias.

Finalmente, o algoritmo SVM apresentou erro médio absoluto menor do que a regressão linear, dando destaque à uniformidade da taxa de erros, que pouco varia entre os conjuntos de dados utilizados.

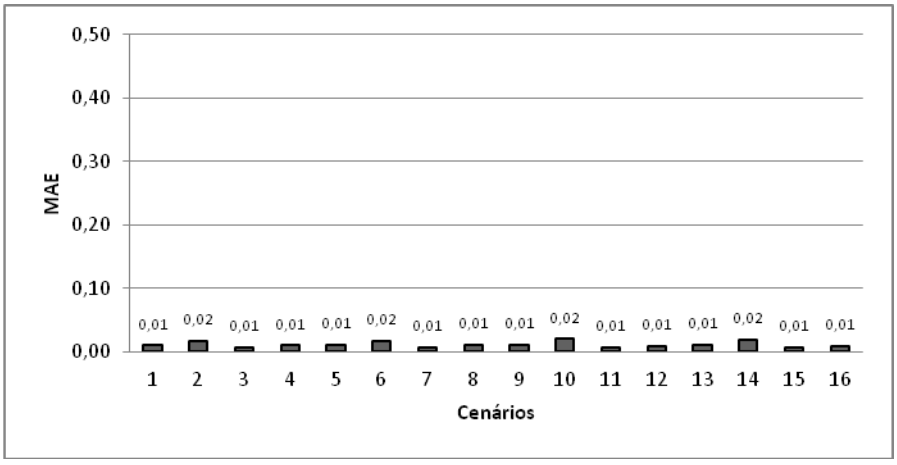

Figura 5. Erro médio absoluto (MAE) por cenário avaliado
Fonte: elaborada pelos autores

Além disso, as maiores taxas de erros foram encontradas em cenários em que nem todas as funções de similaridade foram consideradas, o que indica que, apesar de todas as funções serem baseadas em *string*, a

combinação delas pode trazer ganhos para a identificação das correspondências.

5 TRABALHOS RELACIONADOS

YAM ++ (DUYHOA *et al.*, 2011) é uma ferramenta de alinhamento de ontologias que, de forma similar à presente proposta, aplica técnicas de aprendizado de máquina para o aprendizado de um classificador. A principal diferença recai nas funções de similaridades utilizadas, já que YAM++ considera funções baseadas em recursos linguísticos (Wordnet e dicionário) além das funções de similaridade baseadas em *strings*.

Ehrig (2007) propõe um processo para otimizar um alinhamento de forma parametrizável, que foi chamado de APFEL (*Alignment Process Feature Estimation and Learning*). Este processo consiste basicamente no treinamento supervisionado de uma base de alinhamentos para a classificação do alinhamento, com a entrada de dados parametrizada e o alinhamento obtido validado pelo usuário. No presente trabalho, o processo ocorre de forma automática.

Glue (DOAN *et al.*, 2003) é um sistema de alinhamento de ontologias que explora a aprendizagem de máquina para calcular mapeamentos semânticos entre os conceitos. Considerando duas ontologias distintas, o processo de descoberta do mapeamento entre os conceitos é baseado em medida de similaridade definida pela distribuição de probabilidade conjunta, a qual é calculada por meio de duas técnicas de aprendizagem base aplicadas às instâncias ontologia. Neste trabalho há a necessidade da intervenção do usuário no processo de alinhamento, que é semiautomático.

6 CONCLUSÃO

O alinhamento de ontologias tem sido amplamente utilizado para tratar a heterogeneidade semântica nos mais diversos cenários. No entanto, as abordagens e ferramentas atuais de alinhamento se baseiam em funções de similaridade que ainda resultam em baixa precisão e cobertura. Neste sentido, este trabalho propôs uma abordagem para o aprendizado automático de um modelo classificador que combina diversas funções de similaridade baseadas em *string*, resultando em alinhamentos de melhor qualidade. O modelo gerado pode ser utilizado como um recurso externo para melhorar os resultados de ferramentas atuais de alinhamento de ontologias.

Como trabalhos futuros, estão sendo investigadas a avaliação de outras funções de similaridade (especialmente de outros tipos, como por exemplo baseadas em recursos linguísticos e estruturais) e a avaliação de outros domínios da OAEI.

REFERÊNCIAS

ABOLHASSANI, Hassan; HARIRI, Babak B.; HAERI, Seyed H. On ontology alignment experiments. Disponível em: Webology. 2006. Disponível em: http://www.webology.org/2006/v3n3/a28.html. Acesso em: 09/01/2013.

ALVES, Alex; PADILHA, Natalia; SIQUEIRA, Sean; BAIÃO, Fernanda; REVOREDO, Kate. Using concept maps and ontology alignment for learning assessment. *IEEE Technology and Engineering Education (ITEE),* 1558-7908, v. 7, p. 33-40, September, 2012.

ALVES, Alex; GUEDES, Anselmo; REVOREDO, Kate; BAIÃO, Fernanda. Classificador de alinhamento de ontologias utilizando técnicas de aprendizado de máquina. Simpósio Brasileiro de Sistemas de Informação (SBSI), João Pessoa, p. 25-36, 2013.

BITTENCOURT, Carmem; LADEIRA, Marcelo; SILVA, Saul; BITTENCOURT, Anderson; BORGES, Díbio. Sistema de classificação automática de carcaças bovinas, Simpósio Brasileiro de Sistemas de Informação (SBSI), 2008.

BRAZ, Fernando. Knowledge discovery on trajectory data warehouses: possible usage of the data mining techniques. In: *IV Simpósio Brasileiro de Sistemas de Informação*, Rio de Janeiro, SBC, 2008.

CORUMBA, Daniela; MACEDO, Hendrik. Categorização automática de mensagens de call-for-papers. *Revista Eletrônica de Sistemas de Informação*, v. 10, n. 2, artigo 6, dez 2011.

DOAN, Anhai; MADHAVAN, Jayant; DOMINGOS, Pedro; HALEVY, Aloan. Ontology matching: a machine learning approach. In*: Handbook on Ontologies in Information Systems*. New York: Information Science Reference, p. 397–416, 2003.

DUYHOA, Ngo; ZOHRA, Bellahsene; REMI Coletta. A flexible system for ontology matching. In: Proceedings In Caise 2011 Forum, 2011.

EHRIG, Marc. *Ontology alignment*: bridging the semantic gap, Springer. 2007.

EUZENAT, Jérôme; MEILICKE, Christian; STUCKENSCHMIDT, Heiner; SHVAIKO, Pavel; TROJAHN, Cassia. Ontology alignment evaluation initiative: six years of experience. *Journal on Data Semantics XV*. Springer Berlin Heidelberg, p. 158-192, 2011.

EUZENAT, Jérôme; SHVAIKO, Pavel. *Ontology matching*, Springer-Verlag. Berlin Heidelberg, 2007.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. *AI Magazine, American Association for Artificial Intelligence*, p. 37-54, 1996.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. *Data mining:* um guia prático. Rio de Janeiro: Elsevier, 2005.

GRUBER, Thomas R. A translation approach to portable ontologies. *Knowledge Acquisition*. v. 5, n. 2, p. 199-220, 1993. http://dx.doi.org/10.1006/knac.1993.1008

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data mining:* concepts and techniques. Morgan Kaufmann, 2006.

JARO, Matthew A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84.406, p. 414-420, 1989. http://dx.doi.org/10.1080/01621459.1989.10478785

JÉRÔME, David; EUZENAT, Jérôme; SCHARFFE, François; SANTOS, Cássia Trojahn. The Alignment API 4.0*. Semantic Web Journal*, v. 2, n. 1, p. 3-10, 2011.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, v. 10, p. 707, 1966.

LIMA, Edirlei E. Soares de; POZZER, Cesar T.; D'ORNELLAS, Marcos C.; CIARLINI, Angelo E. M.; FEIJO, Bruno; FURTADO, Antonio L. Support vector machines for cinematography real-time camera control in storytelling environments. In: *VIII Brazilian Symposium on Games and Digital Entertainment*, Rio de Janeiro, Brazil, p. 44-51, 2009.

MASSMANN, Sabine; RAUNICH, Salvatore; AUMUELLER, David; ARNOLD, Patrick; RAHM, Erhard. Evolution of the COMA match system, In: *OM-2011 The Sixth International Workshop on Ontology Matching*, p. 49, 2011.

MITCHELL, Tom M. *Machine Learning*. Burr Ridge, IL: McGraw-Hill, 1997.

MONGE, Alvaro; ELKAN, Charles. The field-matching problem: algorithm and applications. In: *Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, p. 267-270, 1996.

PANG-NING, Tan; STEINBACH, Michael; KUMAR, Vipin. *Introdução ao "Data Mining" - Mineração de Dados*. Rio de Janeiro: Editora Ciência Moderna, 2009.

PEDROSO, Louise; REVOREDO, Kate; BAIÃO, Fernanda. Uma abordagem baseada em mineração de dados para apoio ao ciclo de vida de projetos de pesquisa e inovação. Simpósio Brasileiro de Sistemas de Informação (SBSI), 2013.

SHVAIKO, Pavel; EUZENAT, Jérôme. A survey of schema-based matching approaches. *Journal on Data Semantics IV*. Springer Berlin Heidelberg, p. 146-171, 2005.

SILVA, Carlos; RALHA, Célia. Detecção de cartéis em licitações públicas com agentes de mineração de dados. *Revista Eletrônica de Sistemas de Informação*, v. 10, n. 1, artigo 8, jun 2011.

SMITH, Temple F.; WATERMAN, Michael S. Identification of common molecular subsequences. *Journal of Molecular Biology*, v. 147, n. 1, p. 195–197, 1981. http://dx.doi.org/10.1016/0022-2836(81)90087-5

THORNTON, Chris; HUTTER, Frank; HOOS, Holger H.; LEYTON-BROWN, Kevin. Auto-WEKA: combined selection and hyper-parameter optimization of classification algorithms. *In: Proceedings of the 19th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, August, p. 847-855, ACM, 2013.

UKKONEN, Esko. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, v. 92, p. 191–211. 1992. http://dx.doi.org/10.1016/0304-3975(92)90143-4

WITTEN, Ian H.; EIBE, Frank; MARK, Hall. *Data mining: practical machine learning tools and techniques*. The Morgan Kaufmann Series in Data Management Systems. 3rd edition. 2011.