Detecção Multilíngue de Serviços Web Duplicados Baseada na Similaridade Textual

Erick Nilsen Pereira de Souza[1], **Daniela Barreiro Claro**[1]

[1]FORMAS - Semantic Formalisms and Applications
LaSiD/MMCC/IM/UFBA
Mestrado Multi-Institucional em Ciência da Computação (MMCC)
Instituto de Matemática
Universidade Federal da Bahia
Av. Adhemar de Barros, s/n, Campus de Ondina, Salvador - Bahia - Brazil 40170-110

ericksouza@dcc.ufba.br, dclaro@dcc.ufba.br

Abstract. *Grouping by similarity represents a significant step in strategies of Web Services discovery and composition. Many clustering methods process the service descriptions in natural language to estimate the degree of correlation between them. However, the use of knowledge bases in specific languages limits the applicability of these methods. In this paper we make an analysis of language independent methods for grouping similar Web Services using their natural language descriptions. In particular, we applied Latent Semantic Indexing (LSI), a language-independent method of Information Retrieval (IR). Moreover, an experimental analysis was performed with three similarity measures in order to determine which one is best suited to duplicated Web Services detection from service's descriptions in two languages.*

Resumo. *O agrupamento por similaridade representa uma etapa relevante nas estratégias de descoberta e composição de serviços web. Muitos métodos de agrupamento processam as descrições dos serviços em linguagem natural para estimar o grau de correlação entre eles. Entretanto, a utilização de bases de conhecimento em idiomas específicos limita a aplicabilidade desses métodos. Neste artigo é proposto um modelo multilíngue para agrupamento de serviços web similares a partir das suas descrições em linguagem natural. Em particular, foi aplicado o Latent Semantic Indexing (LSI), um método de Recuperação da Informação (RI) independente da língua e do domínio. Além disso, foi feita uma análise experimental com três medidas de similaridade, a fim de determinar qual delas é mais adequada à detecção de serviços web duplicados a partir das descrições dos serviços em dois idiomas.*

1. Introdução

As organizações têm competido em ambientes de negócios cada vez mais dinâmicos e heterogêneos, nos quais o intercâmbio de operações entre diferentes Sistemas de Informação deve ser transparente ao usuário. Como resultado, grandes companhias vêm adotando a Arquitetura Orientada a Serviços (SOA) como metodologia para desenvolvimento de Sistemas de Informação em larga escala e disponibilização de serviços entre organizações e usuários [17].

À medida que o número de serviços disponibilizados na Web cresce, a importância das estratégias para descoberta de serviços se torna mais evidente. Uma das estratégias necessárias para tornar a descoberta de serviços mais eficiente é a identificação de duplicatas. Um serviço duplicado é caracterizado quando suas operações, entradas e saídas são similares às de outro serviço já existente [3].

A despeito da diversidade de trabalhos relacionados [3, 10, 4], o problema de descoberta de serviços duplicados demanda soluções mais eficientes. De fato, a identificação baseada em sintaxe, que utiliza essencialmente comparações entre as strings das descrições dos serviços, não garante a precisão adequada para a detecção de similaridade entre eles. De acordo com Qu et al. [12], o paradigma de busca por palavra-chave é insuficiente por dois motivos. Primeiro, palavras-chave não capturam a semântica dos Serviços Web, resultando em perdas de resultados. Por exemplo, em uma busca por um serviço com a palavra 'cep', todos as duplicatas que contenham apenas o termo 'código postal' em sua descrição não serão retornadas. Segundo, porque palavras-chave não capturam a semântica textual de forma flexível: em muitos casos, serviços com entradas e saídas similares podem ser tão úteis quanto os retornados pelo casamento perfeito de strings.

As técnicas atuais para detecção de Serviços Web similares podem ser classificadas em dois tipos: as baseadas em *matching* de documentos e as que exploram a estrutura das descrições dos serviços [12]. Os algoritmos de *matching* de documentos são aplicados nas descrições dos serviços em linguagem natural e tentam agrupá-los semanticamente a partir de algum critério de similaridade [3]. Já a abordagem estrutural trabalha sobre as descrições dos serviços em linguagem padrão, como WSDL [7].

Nesse contexto, trabalhos recentes têm aplicado estratégias de Processamento de Linguagem Natural (PLN) para agrupamento de Serviços Web similares a partir de suas descrições em texto não-estruturado (tais como [8], [7], [15] e [14]), uma vez que este tipo de descrição tem crescido na Web, principalmente em serviços REST [6]. Por conta disso, abordagens de similaridade textual tem sido cada vez mais utilizadas na detecção de serviços duplicados. Além de flexibilizar a busca de serviços por palavras-chave, a similaridade semântica pode ser utilizada na política de substituição de serviços inoperantes por serviços ativos similares [18]. Outra aplicação característica desse problema é o agrupamento de serviços em áreas de contexto, o que acarreta redução do espaço de busca, tornando o processo de descoberta de serviços mais eficiente.

As principais abordagens para extração de similaridade em linguagem natural descritas na literatura utilizam a WordNet [5], uma taxonomia onde palavras na língua inglesa são relacionadas a um conjunto de sinônimos (*synsets*), classes gramaticais, sentidos possíveis e exemplos de utilização. Embora os algoritmos baseados em taxonomias tenham resultados satisfatórios em muitos problemas de processamento linguístico (como a Desambiguação Lexical de Sentido (DLS) [16]), apresentam como desvantagem a dependência do idioma. Ou seja, a detecção de similaridade baseada na WordNet pode ser feita apenas em serviços descritos em inglês.

Apesar disso, existem técnicas de Recuperação de Informação (RI) independentes da língua, que são adequadas ao problema de agrupamento por similaridade. Uma delas é a Indexação Semântica Latente (do inglês, *Latent Semantic Indexing (LSI)*). Em seu tra-

balho precursor, Deerwester [2] define LSI como uma abordagem matemática que utiliza decomposição de matrizes lineares de frequência para agrupar documentos semanticamente associados. Além de ser independente de idioma e domínio, LSI permite recuperar documentos semanticamente relacionados mesmo que não possuam as palavras-chave da busca.

Neste trabalho é proposta uma abordagem para detecção de similaridade textual multilíngue em Serviços Web. Além disso, é feita uma avaliação da abordagem proposta com três medidas de similaridade aplicadas ao LSI em dois idiomas: Inglês e Português do Brasil. O restante deste artigo está organizado como segue. A Seção 2 apresenta os principais conceitos referentes ao problema de detecção de Serviços Web duplcados. Na Seção 3 o problema é contextualizado com base em trabalhos correlatos do estado da arte. A Seção 4 descreve a abordagem proposta neste artigo para detecção de similaridade de Serviços Web. Na Seção 5 são descritos e apresentados os experimentos realizados e resultados obtidos. A Seção 6 conclui este artigo.

2. Fundamentação Teórica

A detecção de similaridade para descoberta de Serviços Web tem sido o foco de diversos trabalho na literatura. Nesta seção, são apresentados os conceitos fundamentais relacionados ao problema da descoberta de Serviços Web, como os componentes estruturais dos Serviços Web e os principais paradigmas de busca por similaridade.

2.1. A Estrutura dos Serviços Web

Um Serviço Web consiste em um conjunto de operações agrupadas para um determinado propósito. Cada serviço possui um arquivo descritor, que no caso dos Serviços Web tradicionais é o WSDL (*Web Service Description Language*), que detalha sua funcionalidade e interface. Os Serviços Web tipicamente contém descrições de nome, operações, entrada e saída [8]. Adicionalmente, um Serviço Web pode conter informações textuais em linguagem natural. Qu et al. [12] definem os metadados de um Serviço Web como uma tripla $W = \{T, B, A\}$, onde T representa a informação de título que inclui nome e comentários do serviço, B denota o nome, comentários e mensagens de entrada e saída das operações e A representa informação adicional, incluindo dados de páginas Web e de containers onde o serviço é disponibilizado. Dessa maneira, cada componente dos metadados pode ser representado como uma coleção de palavras: $T = \{w_1, w_2, ..., w_T\}$, $B = \{w_1, w_2, ..., w_B\}$ e $A = \{w_1, w_2, ..., w_A\}$.

2.2. Busca por Serviços Web Similares

A modelagem dos dados para busca por Serviços Web similares pode levar em conta aspectos estruturais ou descrições em linguagem natural dos serviços. A partir da estrutura formal do Serviço Web, é possível extrair similaridade entre operações e entrada/saída. Dong et al. [3] definem o problema de detecção de similaridade em duas partes:
Matching de Operações: *Dada uma operação de Serviço Web, retorne a lista de operações similares.*
Matching de Entrada/Saída: *Dada a entrada/saída de uma operação de Serviço Web, retorne a lista de operações com entradas/saídas similares.*

As técnicas baseadas na estrutura dos serviços têm como principal desafio a pouca informação semântica presente nos nomes das operações, atributos e entrada/saída dos

serviços. Uma abordagem alternativa para detecção de similaridade pode ser feita através das descrições dos serviços em linguagem natural. Considerando os metadados dos serviços como coleções de palavras, é possível determinar a similaridade entre os serviços levando em conta a similaridade lexical entre as palavras. Formalmente, a similaridade entre dois Serviços Web S_1 e S_2 é definida por Qu el at. [12] como:

$$\begin{aligned} Sim(S_1, S_2) = \\ \alpha * SimSet(S_1.T, S_2.T) + \\ \beta * SimSet(S_1.B, S_2.B) + \\ \gamma * SimSet(S_1.A, S_2.A), \end{aligned}$$

onde S.T denota a coleção de palavras das informações do título do serviço, S.B a coleção de palavras das informações do corpo do serviço e S.A a coleção de palavras das informações adicionais do serviço. Os pesos atribuídos a cada conjunto são ponderados por α, β e γ.

A seguir são descritas as principais abordagens para detecção de similaridade em textos.

2.3. Similaridade Texual

O cálculo de similaridade em textos é fundamental na detecção de Serviços Web duplicados baseada em linguagem natural. As principais abordagens aplicadas a este problema utilizam bases de conhecimento (como corpos de treinamento ou léxicos computacionais) para estimar a similaridade entre os conceitos através de dois tipos de medidas: as baseadas em arestas (*edge-based*) e as que utilizam conteúdo da informação (*information content-based*).

2.3.1. Contagem de Arestas

Intuitivamente, conceitos próximos são mais similares que conceitos afastados em uma ontologia hierárquica. Portanto, o menor caminho entre dois conceitos representa uma medida importante de similaridade entre eles. Essa distância conceitual é definida por Mihalcea e Moldovan [9] como

$dist(c_1, c_2)$ = *número mínimo de arestas separando* c_1 *e* c_2*,*

Onde c_1 e c_2 representam conceitos (nós) na ontologia.

Por outro lado, considerar que as ligações entre os conceitos possuem distância uniforme é uma deficiência das medidas baseadas em arestas, pois esta suposição não reflete a realidade dos relacionamentos semânticos entre conceitos do mundo real [13].

2.3.2. Conteúdo da Informação (IC)

Nas medidas baseadas em Conteúdo da Informação, a frequência com a qual um termo aparece associado a outro indica o grau de similaridade entre eles. Segundo Resnik [13], a associação de probabilidades aos conceitos da taxonomia captura a mesma ideia da similaridade baseada em arestas, mas sem a deficiência da uniformidade de distâncias.

Mais precisamente, seja $p : C \rightarrow [0, 1]$ uma função probabilística aplicada à taxonomia. Para todo $c \in C$, $p(c)$ representa a probabilidade de encontrar uma instância do conceito c. Ou seja, a probabilidade decresce em função da especificidade do conceito. Segundo a Teoria da Informação, $IC(c) = -log[p(c)]$. Intuitivamente, isso significa que quanto mais abstrato for um conceito, menor a sua informação agregada.

Resnik [13] define a similaridade entre dois conceitos como

$$sim_{RES}(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c), \quad (1)$$

onde $S(c_1, c_2)$ consiste no conjunto de todos os pais comuns entre os conceitos c_1 e c_2.

As medidas de similaridade baseadas em léxicos apresentam como desvantagem a dependência do idioma, já que o mapeamento dos conceitos na taxonomia é específico para cada língua. A seguir são descritos os trabalhos correlatos que utilizam os conceitos apresentados para detecção de similaridade em Serviços Web.

3. Trabalhos Relacionados

Alguns trabalhos com foco na detecção de Serviços Web similares podem ser encontrados na literatura. As principais abordagens utilizam técnicas que processam as descrições de serviços em linguagem natural, as estruturas *XML* das descrições WSDL, ou uma combinação das duas.

Qu et al. [12] propuseram uma técnica para mineração de similaridade em Serviços Web chamada WSSM: *A WordNet-Based Web Service Similarity Mining Mechanism*. A WSSM é usada para agrupar Serviços Web semelhantes. O agrupamento permite reduzir o espaço de busca, tornando o processo de descoberta de serviços mais eficiente. As informações dos serviços são obtidas a partir de documentos WSDL, containers de serviços e páginas Web. Em seguida, é utilizado um algoritmo baseado na WordNet para gerar uma matriz de similaridade, que é utilizada como entrada por um algoritmo de clusterização para agrupar serviços similares.

Dong et al. [3] propuseram o *WOOGLE*, um motor de busca para Serviços Web, que suporta pesquisas por Serviços Web similares. Os algoritmos descritos pelos autores combinam diversas características para determinar a similaridade entre um par de operações de serviços. Um algoritmo de clusterização agrupa nomes de parâmetros de operações em conceitos semanticamente significativos, que são usados para determinar a similaridade de entrada e saída das operações.

Stroulia e Wang [18] desenvolveram um método para calcular a similaridade semântica e estrutural entre um dado serviço e um conjunto de serviços anunciados em UDDI. O método proposto combina duas técnicas baseadas na WordNet: primeiro são extraídos os serviços similares de acordo com suas descrições WSDL especificadas em linguagem natural; em seguida, dada a lista de serviços candidatos, é aplicado um algoritmo estrutural que calcula as distâncias semânticas entre os identificadores WSDL.

No trabalho desenvolvido por Martin e Cordy [8], o cálculo de similaridade entre operações de Serviços Web é feito através da detecção de clones (trechos de código

idênticos que se repetem) em descrições WSDL. O maior desafio nessa técnica é reduzir o número de falsos positivos, ocasionados por trechos de código idênticos presentes em contextos distintos. Para resolver esse problema, os autores introduzem a ideia de *contextual clones* - clones que são extraídos a partir da extensão dos fragmentos de código com informações relacionadas ao contexto.

Em [14], é proposto um *framework* para descoberta de *web services* a partir de uma ontologia contendo propriedades específicas a respeito dos serviços, denominado *Web Service Modeling Ontology (WSMO)*. O *framework* permite a busca por serviços a partir de palavras-chave definidas pelo usuário, que é realizada através de técnicas de etiquetação *part-of-speech (POS)*, *lemmatization* e desambiguação lexical de sentido sobre os termos da consulta.

Todos os métodos baseados em linguagem natural pesquisados utilizam a WordNet ou ontologias específicas como base de conhecimento, o que limita a solução do problema a um determinado idioma. A seção seguinte apresenta um modelo multilíngue para detecção de Serviços Web duplicados.

4. Detecção Multilíngue

A principal desvantagem das soluções predecessoras discutidas na seção anterior é a dependência do idioma, uma vez que a WordNet está disponível em poucas línguas, sendo a versão em inglês a que possui maior granularidade de dados [4]. Dessa maneira, as abordagens factíveis em outros idiomas tendem a apresentar resultados inferiores. Além disso, os idiomas para os quais não existe versão da WordNet (como o Português do Brasil) não são contemplados pelas soluções apresentadas.

Com o objetivo de tratar a limitação de idioma das soluções anteriores, este trabalho propõe uma abordagem para detecção de Serviços Web duplicados fundamentada em agrupamento textual multilíngue. Na abordagem proposta, a similaridade é calculada através de um método de Recuperação da Informação (RI) baseado na frequência das palavras contidas nas descrições dos serviços.

A Figura 1 apresenta a abordagem para detecção multilíngue de Serviços Web duplicados proposta neste trabalho. Neste modelo, o processo de descoberta de Serviços Web duplicados é precedido de uma etapa de pré-processamento das informações textuais obtidas de suas descrições em linguagem natural. O pré-processamento neste trabalho é composto por três estratégias: extração de *tokens*, remoção de *stop words* [4] e *stemming* [11].

A extração de *tokens* consiste na transformação de uma sentença em um vetor de termos, eliminando os sinais de pontuação e caracteres especiais do texto. Apesar disso, nem todos os termos obtidos nesse processo são semanticamente significativos. Sendo assim, termos como artigos, preposições e verbos de ligação, denominados *stop words*, devem ser eliminados do vetor. As *stop words* são identificadas em uma lista pré-definida para cada idioma. Além de possuírem pouca informação semântica agregada, as *stop words* podem ser encontradas em qualquer contexto. Essa última característica é especialmente prejudicial a métodos de RI que trabalham com vetores de pesos baseados em frequência de palavras.

Outra característica que pode reduzir a qualidade dos métodos de RI é a derivação

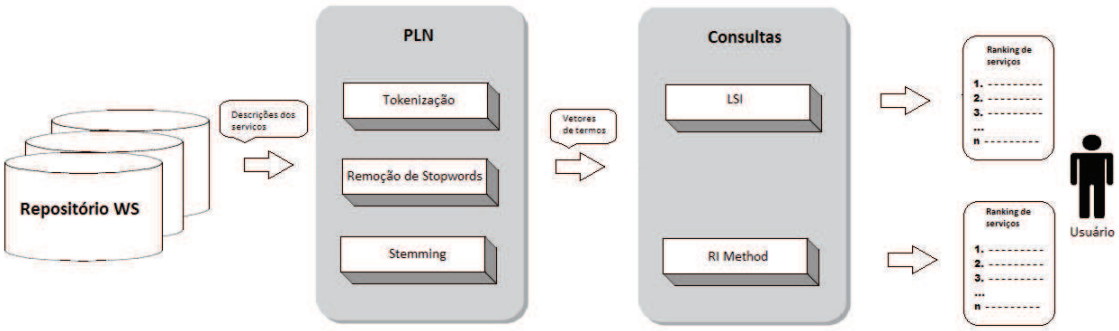

Figura 1. Detecção Multilíngue de Serviços Web Duplicados.

linguística referente ao processo de formação das palavras. Por conta disso, é necessário identificar um radical comum, dentre um conjunto de palavras derivadas, para a obtenção da frequência adequada de determinado termo. Uma das soluções mais populares para a realização de *stemming* é o algoritmo de Porter [11], que é baseado em remoção de sufixos.

Na detecção de Serviços Web duplicados proposta neste trabalho, as técnicas de pré-processamento são aplicadas com objetivo de identificar os termos semanticamente significativos utilizados como entradas dos métodos de RI (Figura 1). Após a fase de pré-processamento, as descrições dos serviços são mapeados em vetores de frequência dos termos significativos, que são utilizados para estimar a similaridade entre os serviços. Em seguida, cada método gera um *ranking* dos serviços mais relevantes sugeridos ao usuário, onde a prioridade é definida pelo grau de correlação entre um dado Serviço Web e os demais serviços contidos no repositório.

Para efeito de validação da abordagem proposta, um método de RI denominado *Latent Semantic Indexing (LSI)* é aplicado e testado no Processador de Consultas (Figura 1) do modelo. As Seções 4.1 e 4.2, a seguir, apresentam a definição formal do LSI e define as medidas usadas para determinar a similaridade entre os vetores gerados pelo LSI neste trabalho.

4.1. Latent Semantic Indexing (LSI)

O LSI é um método que utiliza a estrutura semântica das associações entre termos e documentos, com o objetivo de agrupar documentos relevantes a partir de termos contidos em consultas [2]. Para tanto, utiliza a técnica *Singular Value Decomposition (SVD)*, que decompõe uma matriz em um conjunto de fatores ortogonais, a partir dos quais a matriz original pode ser aproximada por uma combinação linear.

O objetivo dessa abordagem é minimizar a deficiência da busca de documentos por palavras-chave, tratando a ambiguidade da associação entre termos como um problema estatístico. Segundo [2], existe uma estrutura semântica latente nos dados, que permanece parcialmente oculta na forma aleatória de disposição das palavras em um texto. Nesse sentido, o LSI utiliza técnicas estatísticas para estimar essa estrutura latente oculta e eliminar os ruídos causados pela ambiguidade conceitual.

Por ser um método estatístico, o LSI pode ser aplicado a qualquer domínio e idioma sem perda de generalidade. Essa independência do idioma é uma vantagem significativa frente a outros métodos de Extração da Informação em texto não-estruturado. A seguir são descritas as medidas usadas para determinar a similaridade entre os vetores gerados pelo LSI neste trabalho.

4.2. Similaridade em Vetores de Texto

Diversas abordagens de RI necessitam determinar a similaridade entre dois vetores de texto no espaço euclidiano [2]. Formalmente, dados dois vetores de palavras $v_1 = (w_{11}, w_{12}, ..., w_{1n})$ e $v_2 = (w_{21}, w_{22}, ..., w_{2n})$ e seus respectivos vetores de pesos $P_1 = (p_{11}, p_{12}, ..., p_{1n})$ e $P_2 = (p_{21}, p_{22}, ..., p_{2n})$, o objetivo é obter uma medida de similaridade entre v_1 e v_2 em função de P_1 e P_2. Neste trabalho, três medidas de similaridade são estudadas e aplicadas: cosseno, produto escalar e distância euclidiana.

4.2.1. Cosseno

A similaridade entre dois vetores pode ser medida a partir do cosseno do ângulo formado entre eles em um espaço vetorial multidimensional [1]. Na modelagem vetorial, a disposição dos vetores no espaço depende da frequência dos termos contidos na sentença.

Tratando-se de vetores de texto, o valor obtido pode variar no intervalo [0,1], já que a frequência das palavras é sempre um número inteiro positivo. Assim, valores de cosseno que tendem a 0 representam pouca similaridade textual, ao passo que valores próximos de 1 indicam vetores com alta similaridade. Mais precisamente, a similaridade entre v_1 e v_2 é dada por:

$$SIM(v_1, v_2) = \frac{\sum_{i=1}^{n} p_{1,i} p_{2,i}}{\sqrt{(\sum_{i=1}^{n} p_{1,i})^2 \times (\sum_{i=1}^{n} p_{2,i})^2}} \quad (2)$$

4.2.2. Produto Escalar

Outra forma usual de calcular a similaridade entre dois vetores é através do produto escalar [1]. Geometricamente, o produto escalar é definido conforme Equação 3:

$$v_1 \cdot v_2 = ||v_1||\ ||v_2||\ cos\theta \quad (3)$$

Em particular, se v_1 e v_2 são ortogonais, então $v_1 \cdot v_2 = 0$ (baixa similaridade). Por outro lado, se $v_1 = v_2$, então $v_1 \cdot v_2 = 1$ (alta similaridade). Do ponto de vista algébrico, a similaridade por produto escalar é dada por:

$$SIM(v_1, v_2) = v_1 \cdot v_2 = \sum_{i=1}^{n} p_{1,i} p_{2,i} \quad (4)$$

4.2.3. Distância Euclidiana

A Distância Euclidiana é uma medida da distância entre dois pontos no espaço métrico, obtida pela aplicação sucessiva do Teorema de Pitágoras [1]. Assim, a partir da representação textual no espaço euclidiano, é possível obter a similaridade entre dois textos através da Equação 5:

$$SIM(v_1, v_2) = dist(v_1, v_2)^{-1} = \frac{1}{\sqrt{\sum_{i=1}^{n}(p_{1,i} - p_{2,i})^2}} \quad (5)$$

5. Experimentos

Neste trabalho foram realizados experimentos para avaliação da cobertura das medidas de similaridades descritas na seção anterior. Cada medida foi aplicada ao método LSI para detecção de Serviços Web duplicados descritos em dois idiomas: Inglês e Português do Brasil.

As descrições dos serviços foram obtidas do XMethods[1], um repositório de Serviços Web descritos em Língua Inglesa. Para viabilizar a avaliação do método em idiomas distintos, foi realizada a tradução automática das descrições dos serviços do Inglês para o Português do Brasil através da ferramenta de tradução do Google[2].

Nos experimentos foram desconsiderados todos os Serviços Web da base utilizada com descrições inferiores a 130 caracteres. Dessa forma, a base de fato utilizada possui 66 instâncias de teste, cada uma contendo dois serviços duplicados associados. As duplicatas dos serviços foram geradas por meio da substituição automática de um subconjunto de termos significativos por sinônimos extraídos da WordNet. A Tabela 1 apresenta dois exemplos de serviço principal (S_i) e suas respectivas duplicatas (D_{i1} e D_{i2}):

Conjunto	Inglês	Português
S_1	**calculate** saudi shipping prices	**calcula** os preços sauditas de envio
D_{11}	**calculate** arab travel values	**calcula** valores árabes de transporte
D_{12}	arab travel values	valores árabes de transporte
...	...	...
S_n	sends an SMS message to a **mobile** phone	envia uma mensagem SMS para um telefone **móvel**
D_{n1}	invite an SMS text to a **mobile** cellphone	manda um texto SMS para um celular **móvel**
D_{n2}	invite an SMS text to a cellphone	manda um texto SMS para um celular

Tabela 1. Descrições de serviços duplicados

[1] www.xmethods.net
[2] http://translate.google.com.br

Nos exemplos da Tabela 1, D_{i1} representa as descrições de duplicatas que contêm palavras significativas[3] comuns com o serviço principal. Já em D_{i2}, todas as palavras significativas comuns com o serviço principal são removidas. Assim, $\Omega(S_i \cap D_{i1}) \neq \emptyset$ e $\Omega(S_i \cap D_{i2}) = \emptyset$, onde $\Omega(T)$ é uma função que retorna os termos significativos de uma sentença T. Dessa maneira, o termo *calculate* de S_1, que não pode ser indexado na WordNet, foi mantido em D_{11} e removido em D_{12}.

5.1. Resultados e Discussões

As Tabelas 2 e 3 mostram a cobertura (recall) obtida para cada medida de similaridade em função do tamanho do ranking gerado (15 instâncias) nas descrições dos serviços em Inglês e Português, respectivamente. Os resultados são discriminados considerando três conjuntos distintos: apenas as duplicatas D_{i1}, apenas as duplicatas D_{i2} e a união das duplicatas D_{i1} e D_{i2}.

Medida	D_{i1}	D_{i2}	$D_{i1} \cup D_{i2}$
cosseno	0,563	0,211	0,647
produto escalar	0,572	0,083	0,674
distância euclidiana	0,571	0,082	0,679

Tabela 2. Cobertura: serviços descritos em Língua Inglesa

Medida	D_{i1}	D_{i2}	$D_{i1} \cup D_{i2}$
cosseno	0,712	0,197	0,606
produto escalar	0,682	0,076	0,682
distância euclidiana	0,712	0,091	0,689

Tabela 3. Cobertura: serviços descritos em Língua Portuguesa

A metodologia de avaliação dos resultados utiliza o conceito de cobertura adotado nos trabalhos correlatos descritos na Seção 3. Mais precisamente, a cobertura é dada pela razão entre o número de serviços duplicados que o método consegue identificar no ranking e o número total de duplicatas no repositório.

A análise dos resultados permite concluir que a cobertura do LSI tende a ser maior quando as duplicatas possuem palavras comuns com a descrição do serviço principal. De fato, nos contextos que contêm o conjunto D_{i1}, a cobertura média varia de 56-57% em Inglês e 68-71% em Português. Por outro lado, levando em conta apenas duplicatas sem palavras comuns com o serviço principal (conjunto D_{i2}), a cobertura média varia entre 8-21% em Inglês e 7-19% em Português.

Os resultados mostram que o desempenho médio do LSI, considerando todos os conjuntos e todas as medidas, foi superior em 4% para as duplicadas com descrições em Português em relação às duplicatas com descrições em Inglês. Isto se deve ao fato de que, no processo de tradução automática, alguns termos distintos na Língua Inglesa são traduzidos para o mesmo termo em Língua Portuguesa, aumentando assim a quantidade de termos comuns nos conjuntos traduzidos. Por exemplo, ambos os termos *home*

[3]Palavras significativas são os termos obtidos a partir de uma sentença após a remoção de stopwords.

e *house* são traduzidos automaticamente como *casa* em diferentes sentenças da Língua Portuguesa. Dessa maneira, quando os termos comuns pertencem ao serviço principal e às respectivas duplicatas, a cobertura tende a aumentar, já que a similaridade entre eles será maior.

Do ponto de vista da comparação entre as medidas, a similaridade de cosseno foi superior ao produto escalar e à distância euclidiana nos dois idiomas testados. Considerando apenas as duplicatas do conjunto D_{i2}, a medida cosseno apresenta cobertura 12% superior às demais.

6. Conclusões e Trabalhos Futuros

A detecção de Serviços Web similares é fundamental para o problema de descoberta de serviços. Nesse sentido, uma análise do estado da arte revela que diversas técnicas para detecção de serviços similares baseadas em linguagem natural têm sido estudadas e aplicadas. As principais abordagens desse tipo utilizam a WordNet como base de conhecimento para calcular a similaridade semântica entre fragmentos de texto de descrições de serviços. Porém, a dependência do idioma é um fator limitante para a aplicabilidade dessas técnicas. Neste trabalho é proposto um modelo multilíngue para agrupamento de Serviços Web similares. Além disso, é feita uma avaliação do modelo através da aplicação do LSI na detecção de Serviços Web duplicados usando três medidas de similaridade baseadas em vetores de frequência. Constatou-se que o LSI detectou em média mais de 60% dos serviços duplicados a partir de suas descrições textuais no repositório testado. Além disso, cerca de 20% das duplicatas que não possuem palavras comuns com o serviço principal foram devidamente classificadas pelo método.

Em relação às medidas testadas, a similaridade de cosseno apresentou os melhores resultados, principalmente para as duplicatas que não possuem termos comuns com o serviço principal (cerca de 12% superior às demais). Por outro lado, considerando todos os tipos de duplicatas, nenhuma medida se destacou frente às outras.

A avaliação realizada neste trabalho indica que métodos de RI podem ser aplicados para detectar serviços web duplicados a partir de suas descrições em linguagem natural, de forma independente do idioma. Como trabalhos futuros, pretende-se incorporar ao modelo técnicas de detecção de similaridade estrutural a partir dos elementos contidos na linguagem padrão dos serviços, a fim de melhorar o desempenho na detecção de duplicatas.

Agradecimentos

Os autores agradecem à FAPESB pelo apoio financeiro (número do projeto: 19.573.128.2586)

Referências

[1] K. Ando. Latent semantic space: Iterative scaling improves precision of inter-document similarity measurement. *In Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000), pp. 216-223.*, 2000.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society Science*, 1990.

[3] X. Dong, A. Halevy, J. Madhavan, E. Nemes, and J. Zhang. Similarity search for web services. *Proceedings of the 30th VLDB Conference, Toronto, Canada*, 2004.

[4] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches analyzing advanced unstructured data*. New York: CAMBRIDGE UNIVERSITY PRESS, 2007.

[5] C. Fellbaum. Wordnet: An electronic lexical database. language, speech, and communication. *MIT Press, Cambridge, MA.*, 1998.

[6] R. T. Fielding. *Architectural styles and the design of network-based software architectures*. PhD thesis, 2000. AAI9980887.

[7] P. Ly, C. Pedrinaci, and J. Domingue. Automated information extraction from web apis documentation. *In: The 13th International Conference on Web Information System Engineering (WISE 2012), Paphos, Cyprus (Forthcoming)*, 2012.

[8] D. Martin and J. Cordy. Analyzing web service similarity using contextual clones. *IWSC'11, May 23, 2011, Waikiki, Honolulu, HI, USA.*, 2011.

[9] R. Mihalcea and D. Moldovan. Word sense disambiguation based on semantic density. *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1989.

[10] M. Paolucci, T. Kawamura, T. Payne, and K. Sycara. Importing the semantic web in uddi. *Web Services, E-Business, and the Semantic Web, Springer, 2002, 815-821*, 2002.

[11] F. Porter. An algorithm for suffix stripping. *Program, 14(3), 130-137.*, 1980.

[12] X. Qu, H. Sun, X. Li, and W. Lin. Wssm: A wordnet-based web service similarity mining mechanism. *Proceedings of IARIA International Conference on Service Computation*, 2009.

[13] P. Resnik. Using information content to evaluate semantic similarity. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448-453, Montreal.*, 1995.

[14] J. Sangersa, F. Frasincara, F. Hogenbooma, and V. Chepegin. Semantic web service discovery using natural language processing techniques. *Expert Systems with Applications, Vol. 40, Iss. 11, 1 September 2013, Pages 4660-4671*, 2013.

[15] H. Sneed and C. Verhoef. Natural language requirement specification for web service testing. *2013 15th IEEE International Symposium on Web Systems Evolution (WSE)*, 2013.

[16] E. Souza and D. Claro. Evaluation of semantic similarity in wsd: An analysis to incorporate it into the association of terms. *WebMedia'12, October 15-28, São Paulo/SP, Brazil.*, 2012.

[17] K. Souza and M. Fantinato. Explorando a engenharia de requisitos orientada a serviços: Uma revisão sistemática da literatura. *IX Simpósio Brasileiro de Sistemas de Informação - 2013 - João Pessoa, PB, Brasil, 260-271*, 2013.

[18] E. Stroulia and Y. Wang. Structural and semantic matching for assessing web-service similarity. *International Journal of Cooperative Information Systems*, 2005.