UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CARLOS EDUARDO MANZONI MOREIRA

Descoberta de *Cross-Language Links* Ausentes na Wikipédia

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof[a]. Dra. Viviane Pereira Moreira
Orientador

Porto Alegre, abril de 2014

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Moreira, Carlos Eduardo Manzoni

Descoberta de *Cross-Language Links* Ausentes na Wikipédia / Carlos Eduardo Manzoni Moreira. – Porto Alegre: PPGC da UFRGS, 2014.

67 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2014. Orientador: Viviane Pereira Moreira.

1. Classification. 2. Cross-language links. 3. Similarity functions. 4. Wikipedia. I. Moreira, Viviane Pereira. II. Título.

“E sem saber que era impossível
foi lá e fez.”
— JEAN COCTEAU

AGRADECIMENTOS

Além do esforço e empenho necessários para a construção deste trabalho, esta dissertação é o resultado de todo o apoio fornecido pelas pessoas com quem convivi durante essa etapa da minha vida.

Em especial gostaria de agradecer:

A Professora Viviane pela dedicação na orientação conduzida. A sua experiência, conhecimento e conselhos durante toda a orientação foram fundamentais para a construção deste trabalho.

Ao Instituto de Informática da UFRGS por toda a infraestrutura oferecida e, ao PPGC e seus funcionários.

Aos professores da UFRGS, que também foram meus professores do curso de graduação de Ciência da Computação, agradeço pelos ensinamentos transmitidos que tem sido fundamental para a construção da minha carreira profissional.

A toda a minha família e aos meus amigos que acompanharam, de perto ou de longe, a realização deste trabalho e que torceram para que o mesmo fosse concluído com êxito.

Em especial, agradeço: ao meu pai, Luiz Carlos Gomes Moreira, por toda a inspiração e motivação fornecida em construir uma carreira como profissional de TI; a minha mãe, Rosane Manzoni Moreira, por todo o amor e carinho dedicado a mim e aos meus irmãos; aos meus irmãos, Ricardo Manzoni Moreira e Gabriel Manzoni Moreira, pelo companheirismo e amizade.

Por fim, agradeço à Rachel Kerber Gonçalves, meu eterno amor, pela paciência, companheirismo e amor ao longo destes 9 anos de convivência. Agradeço por entender as minhas ausências e por ter me propiciado e incentivado alcançar mais esta meta.

RESUMO

A Wikipédia é uma enciclopédia pública composta por milhões de artigos editados diariamente por uma comunidade de autores de diferentes regiões do mundo. Os artigos que constituem a Wikipédia possuem um tipo de *link* chamado de *Cross-language Link* que relaciona artigos correspondentes em idiomas diferentes. O objetivo principal dessa estrutura é permitir a navegação dos usuários por diferentes versões de um mesmo artigo em busca da informação desejada. Além disso, por permitir a obtenção de *corpora comparáveis*, os *Cross-language Links* são extremamente importantes para aplicações que trabalham com tradução automática e recuperação de informações multilíngues. Visto que os *Cross-language Links* são inseridos manualmente pelos autores dos artigos, quando o autor não reconhece o seu correspondente em determinado idioma ocorre uma situação de *Cross-language Links* ausente. Sendo assim, é importante o desenvolvimento de uma abordagem que realize a descoberta de *Cross-language Links* entre artigos que são correspondentes, porém, não estão conectados por esse tipo *link*. Nesta dissertação, é apresentado o CLLFinder, uma abordagem para a descoberta de *Cross-language Links* ausentes. A nossa abordagem utiliza o relacionamento entre as categorias e a indexação e consulta do conteúdo dos artigos para realizar a seleção do conjunto de candidatos. Para a identificação do artigo correspondente, são utilizados atributos que exploram a transitividade de *Cross-language Links* entre outros idiomas bem como características textuais dos artigos. Os resultados demonstram a criação de um conjunto de candidatos com 84,3% de presença do artigo correspondente, superando o trabalho utilizado como *baseline*. A avaliação experimental com mais de dois milhões de pares de artigos aponta uma precisão de 99,2% e uma revocação geral de 78,9%, superando, também, o *baseline*. Uma inspeção manual dos resultados do CLLFinder aplicado em um cenário real indica que 73,6% dos novos *Cross-language Links* sugeridos pela nossa abordagem eram de fato correspondentes.

Palavras-chave: Classification, cross-language links, similarity functions, wikipedia.

Identifying Missing Cross-Language Links in Wikipedia

ABSTRACT

Wikipedia is a public encyclopedia composed of millions of articles written daily by volunteer authors from different regions of the world. The articles contain links called Cross-language Links which relate corresponding articles across different languages. This feature is extremely useful for applications that work with automatic translation and multilingual information retrieval as it allows the assembly of comparable corpora. Since these links are created manually, in many occasions, the authors fail to do so. Thus, it is important to have a mechanism that automatically creates such links. This has been motivating the development of techniques to identify missing cross-language links. In this work, we present CLLFinder, an approach for finding missing cross-language links. The approach makes use of the links between categories and an index of the content of the articles to select candidates. In order to identify corresponding articles, the method uses the transitivity between existing cross-language links in other languages as well as textual features extracted from the articles. Experiments on over two million pairs of articles from the English and Portuguese Wikipedias show that our approach has a recall of 78.9% and a precision of 99.2%, outperforming the baseline system.A manual inspection of the results of CLLFinder applied to a real situation indicates that our approach was able to identify the Cross-language Link correctly 73.6% of the time.

Keywords: Classification, cross-language links, similarity functions, wikipedia.

LISTA DE ABREVIATURAS E SIGLAS

CLL	Cross-Language Link
RI	Recuperação de Informações
RI-ML	Recuperação de Informações Multilíngues
SRI	Sistemas de Recuperação de Informações
TF-IDF	Term Frequency - Inverse Document Frequency
WLM	Wikipedia Link-based Measure

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Wikipédia é uma enciclopédia multilíngue disponível na Internet e criada sob uma licença livre cuja a escrita ocorre de forma colaborativa. Sua missão, conforme seu criadores Larry Sanger e Jimmy Wales, é coletar e desenvolver conteúdo educacional em um domínio público de forma a disseminá-lo globalmente. Atualmente, possui mais de 30 milhões de artigos (janeiro de 2014) que são escritos, modificados e relacionados diariamente por uma comunidade voluntária de autores e editores ao redor do mundo (WIKIPEDIA, 2014). Visto que seus artigos são criados e publicados em vários idiomas, existem muitos artigos correspondentes (também chamados de equivalentes), porém em idiomas diferentes. Tal característica torna a Wikipédia um valioso repositório de informações multilíngues (ADAR; SKINNER; WELD, 2009).

Conceitualmente, um artigo da Wikipédia é representado por uma página que possui informações sobre o assunto do qual ele trata. O chamado *Cross-language Link* (CLL) é um recurso muito interessante presente nos artigos que possibilita a navegação entre artigos correspondentes escritos em outros idiomas. Por tratar-se de um mecanismo de ligação entre artigos que possuem conteúdos equivalentes, os CLLs possuem um valor extraordinário para aplicações multilíngues (NASTASE; STRUBE, 2013).

Além de possibilitar a navegação dos usuários por diferentes versões de um mesmo artigo, os CLLs também podem ser utilizados para vários objetivos dentro da área de Recuperação de Informações (RI). NGUYEN et al. (2011), OH et al. (2008), e ERDMANN et al. (2009) utilizaram os CLLs entre artigos para criar um dicionário bilíngue, enquanto ADAFRE; RIJKE (2006) usaram os CLLs para encontrar similaridades entre sentenças de diferentes idiomas. Outros trabalhos confiam nos CLLs como componente fundamental para obter *corpora comparáveis*[1] (ADAFRE; RIJKE, 2006; POTTHAST; STEIN; ANDERKA, 2008; SORG; CIMIANO, 2008b).

Para o usuário da Wikipédia, o aspecto mais interessante do CLL é a navegação entre os idiomas de um determinado artigo com o objetivo de encontrar o que está mais completo e/ou possui mais referências. Isso contribui com uma maior credibilidade e cobertura de informações sobre um determinado artigo. Além disso, os CLLs permitem o mapeamento de *infoboxes* que, por conterem informações estruturadas de um artigo, são utilizadas em pesquisas que tratam da detecção de inconsistências nas informações. *Infobox* é o nome que se dá a um conjunto de dados do tipo atributo-valor que resume informações da entidade descrita no artigo e localiza-se, em geral, no canto superior direito. Sendo assim, conforme o tipo de artigo a ser escrito, o autor deve preencher esses dados estruturados que descrevem as principais características do artigo. Com o uso do

[1] Corpora comparáveis são coleções de textos em duas ou mais línguas onde todos os textos descrevem um mesmo tópico.

CLL para o mapeamento de atributos e valores das infoboxes de artigos correspondentes, é possível detectar e corrigir inconsistências em diferentes Wikipédias, ajudando a comunidade Wiki no aumento da consistência dos dados (RINSER; LANGE; NAUMANN, 2013).

Figura 1.1: Exemplos de artigos Correspondentes com (a) e sem CLL (b)

(a) Cross-Language Link

(b) Cross-Language Link Ausente

Para que seja possível extrair informação multilíngue é necessário que se tenha uma boa cobertura e precisão em termos de CLLs. No entanto, as duas situações abaixo podem ocorrer diminuindo a cobertura de CLLs:

- CLLs conectados de forma incorreta: quando a ligação é entre artigos que não são correspondentes, e
- CLLs inexistentes entre dois artigos correspondentes.

A situação que buscamos solucionar ao desenvolver esse trabalho é a última: CLLs inexistentes entre dois artigos correspondentes. Para ela, damos o nome de Cross-language Links ausentes (*CLLs ausentes*), ou seja, a falta de um CLL entre dois artigos equivalentes.

Conforme (OH et al., 2008), os CLLs são tipicamente adicionados pelos autores dos artigos. Quando o autor de um artigo não reconhece os seus equivalentes em outros idiomas ocorre uma situação de CLL ausente. Nesse caso, o CLL acaba não sendo inserido, inviabilizando a navegação do usuário pelos diferentes idiomas de um artigo e impedindo que aplicações explorem todo o potencial multilíngue da Wikipédia. Com o objetivo de enriquecer a característica multilíngue da Wikipédia, é fortemente desejável uma abordagem que realize a descoberta de CLL ausentes.

A Figura 1.1(a) mostra um exemplo de CLL (dentro dos retângulos vermelhos) conectando o artigo *Parque Farroupilha* entre as Wikipédias em inglês e português. Já a Figura 1.1(b) mostra um exemplo de CLL ausente. O artigo *Dente de Ovo* da Wikipédia em português não possui CLL com o artigo *Egg Tooth* que é seu correspondente em inglês. Da mesma forma, o artigo em inglês também não possui CLL para o em português.

1.1 Motivação

A possibilidade de complementar informações entre artigos equivalentes abre um campo enorme para pequisa de algoritmos que descubram tal equivalência. Corroborando com tal estudo, verifica-se que a consulta de artigos em outros idiomas aumenta consideravelmente as fontes de pesquisa do indivíduo que utiliza a Wikipédia. Um exemplo de pequisa que seria beneficiada através da equivalência multilíngue é sobre artigos que tratam de assuntos específicos da cultura de um país ou de uma região (BOUMA; DUARTE; ISLAM, 2009). Por exemplo, informações sobre turismo e cultura brasileira, estarão mais completas na Wikipédia do Brasil do que em outras Wikipédias.

Todos esses benefícios resultantes da presença de CLLs nos motivou a desenvolver uma abordagem para a descoberta de CLLs ausentes que possa contribuir de forma a aumentar a cobertura desses *links* entre artigos correspondentes. Tal abordagem, ao ser alimentada com os dados da Wikipédia, deve ser capaz de realizar a análise de propriedades dos artigos que indiquem a existência, ou não, de similaridade entre um par de artigos. Se o nosso classificador indicar, através dos escores de similaridade gerados, que tratam-se de artigos correspondentes, estaremos contribuindo com possibilidade de complementar a Wikipédia em termos de CLLs. Tal fato irá auxiliar tanto a navegação do usuário final quanto os trabalhos que usam esse mecanismo para a descoberta de informação.

Na análise realizada nesse trabalho com o *dump* das Wikipédias de maio de 2011 dos idiomas inglês e português foram verificadas as estatísticas em relação a presença de CLLs para os artigos desses idiomas. Conforme a Tabela 1.1, verifica-se que somente uma pequena fração de 12,3% dos artigos em inglês da Wikipédia são mapeados por

CLLs para os artigos em português. Já a fração do número de artigos em português que possuem CLLs para artigos em inglês é muito maior, representando 65,9% do total da Wikipédia em português. Essa diferença de percentual de cobertura de CLLs deve-se principalmente aos tamanhos distintos das Wikipédias trabalhadas, pois a Wikipédia em português representa aproximadamente 18% do tamanho da Wikipédia em inglês.

Tabela 1.1: Estatísticas das Wikipédias em inglês e português (Maio de 2011)

		Cross-language Links		
Wikipédia	Núm. Artigos	Direção	Núm. CLLs	% CLLs vs Núm. Artigos
Inglês	3.632.660	EN $\xrightarrow{CLL}$ PT	447.372	12,3%
Português	681.499	PT $\xrightarrow{CLL}$ EN	449.305	65,9%

A Tabela 1.2 nos mostra o número aproximado de CLLs ausentes que podem ser descobertos. Visto que estamos trabalhando com as Wikipédias em português e inglês, tomamos como base para o número máximo possível de CLLs o tamanho da Wikipédia em português. Em números absolutos, existem 234.127 CLLs ausentes do inglês para o português e 232.194 CLLs ausentes do português para o inglês. Tais valores representam respectivamente 34,4% e 34,1% da Wikipédia em português. Nesse ponto, é necessário fazer uma observação: para alguns artigos provavelmente não exista o seu correspondente no outro idioma devido ao contexto local e/ou específico do país. No entanto, esse não é o caso para que 34,1% da Wikipédia em português não possua CLLs para a em inglês. Sendo assim, mesmo descontando tais casos, ainda existe uma quantidade imensa de CLLs ausentes que podem ser descobertos entre os idiomas português e inglês. Tal análise nos motivou a desenvolver uma abordagem que seja capaz de realizar a descoberta de CLLs ausentes entre artigos correspondentes e contribuir para o aumento da cobertura desses *links* na Wikipédia.

Tabela 1.2: Número de CLLs ausentes

		Cross-language Links					
Wikipédia	Núm. Artigos	Direção	CLLs Existentes	CLLs Possíveis	% CLLs Existentes	% CLLs ausentes	CLLs ausentes
Inglês	3.632.660	EN $\xrightarrow{CLL}$ PT	447.372	681.499	65,6%	34,4%	234.127
Português	681.499	PT $\xrightarrow{CLL}$ EN	449.305	681.499	65,9%	34,1%	232.194

1.2 Objetivo e Estrutura da Abordagem

O objetivo desse trabalho é propor uma abordagem chamada Cross-language Link Finder (CLLFinder) que possui um conjunto de técnicas que foram implementadas de forma a realizar a descoberta de CLLs ausentes entre artigos da Wikipédia. Na nossa abordagem, existem duas grandes etapas envolvidas: a seleção de candidatos a serem comparados com um dado artigo e a aplicação de atributos de similaridade necessários para a identificação do candidato correspondente ao artigo em questão. Todo par de artigos que for considerado equivalente pela análise dos atributos realizada pelo classificador poderá ser ligado por um CLL.

Em relação à primeira etapa, é necessária a geração de um subconjunto pequeno que contenha os artigos a serem comparados, na segunda etapa, com o artigo para o qual busca-se o CLL. Isso ocorre visto que não é computacionalmente possível comparar cada par de artigos existentes na Wikipédia entre dois determinados idiomas, pois conforme os dados da Tabela 1.1, para um artigo em português seriam mais de 3.600.000 comparações para encontrar o seu equivalente em inglês. No que diz respeito à geração desse subconjunto de artigos, foram desenvolvidos nesse trabalho dois métodos: o *CategoryLink* e o *ArticleContentIndex*. O primeiro leva em consideração as categorias as quais um artigo pertence para determinar seus candidatos. Já o segundo trabalha com a Recuperação de Informações Multilíngue, realizando a indexação do conteúdo dos artigos e consultas ao índice que leva em consideração a relevância do texto. Além desses dois primeiros métodos, cuja autoria pertence ao nosso trabalho, foi implementado e acrescentado à primeira etapa da nossa abordagem o *Chain Link Hypothesis* do SORG; CIMIANO (2008a). Ressaltamos que o método *Chain Link Hypothesis* não é proposto pelo nosso trabalho, no entanto, foi utilizado em complemento aos nossos dois métodos para a seleção de candidatos (*CategoryLink* e *ArticleContentIndex*) devido aos bons resultados reportados por SORG; CIMIANO (2008a).

Para a segunda etapa, foram selecionados atributos que refletem a similaridade entre os artigos. O nosso principal atributo é o `CLLTransitivity`, que leva em consideração a transitividade de CLLs entre idiomas intermediários na geração do coeficiente de similaridade entre dois artigos. Além dele, foram utilizados os seguintes atributos: `Similaridade de Títulos`, que realiza um pré-processamento nos títulos antes de medir a similaridade, `Distância de Edição`, que é aplicada sobre os títulos dos artigos, e o `Termos em Comum`, que realiza a contagem do número de palavras iguais no texto. Os coeficientes de similaridade gerados pelos atributos acima foram submetidos a um classificador para criar um modelo de decisão. Posteriormente, foram submetidos conjuntos de artigos para a verificação dos resultados e validação do modelo gerado.

1.3 Contribuições

O principal resultado deste trabalho é o desenvolvimento da abordagem CLLFinder para encontrar CLLs ausentes na Wikipédia que possui alta precisão e revocação. Além disso, conforme a análise dos trabalhos relacionados, o nosso estudo é o primeiro a realizar análise da cobertura de CLLs entre as Wikipédias em inglês e português. De forma mais específica, podemos citar os seguintes diferenciais:

- o uso de categorias para seleção do conjunto de candidatos (método *CategoryLink* na Seção 3.2.2) apresentando ganhos significativos quando comparado a um *baseline*,
- o método *ArticleContentIndex* (Seção 3.2.3) que indexa o conteúdo de todos os artigos e realiza consultas baseadas no texto dos mesmos para formar o conjunto de candidatos, alcançou melhoras significativas na seleção de artigos candidatos, e
- o atributo `CLLTransitivity` que pontua um escore de similaridade de acordo com a transitividade de CLLs e apresentou ótima precisão para atribuição de CLLs entre os pares de artigos.

Salientamos que foram realizados esforços em termos de otimizações de banco de dados e aplicação para conseguir trabalhar com a grande quantidade de dados da Wikipédia.

A implementação do CLLFinder foi feita em um computador doméstico de performance mediana adquirido em 2010 (Athlon X2 com 4Gb de RAM).

Parte do trabalho desenvolvido nesta dissertação foi publicada como artigo completo no Journal of Information and Data Management (JIDM)[2] e apresentada no SBBD 2013 (MOREIRA; MOREIRA, 2013). O trabalho intitula-se "*Finding Missing Cross-Language Links in Wikipedia*".

1.4 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: no Capítulo 2 descrevemos os principais trabalhos que orientaram o desenvolvimento da nossa abordagem. No Capítulo 3, descrevemos os métodos propostos para a seleção de candidatos, a organização do conjunto de candidatos, os atributos para a identificação de artigos correspondentes e o classificador utilizado. No Capítulo 4 apresentamos os experimentos realizados, os resultados alcançados e a comparação com o trabalho *baseline*. Por fim, no Capítulo 5 apresentamos as conclusões e possibilidades para trabalhos futuros.

[2]http://seer.lcc.ufmg.br/index.php/jidm

2 TRABALHOS RELACIONADOS

Este capítulo apresenta os principais trabalhos que estão relacionados aos assuntos tratados na dissertação. Na Seção 2.1, abordamos os trabalhos que, como o nosso, estão preocupados com a descoberta de CLLs ausentes para enriquecer a estrutura de links da Wikipédia. Na Seção 2.2, analisamos os trabalhos que utilizam os CLLs para alinhamento de infoboxes e complementação de informações. A Seção 2.4 apresenta as considerações finais do capítulo.

2.1 Descoberta de Cross-language Links ausentes

A seguir seguem os três principais trabalhos relacionados à abordagem proposta.

2.1.1 Proposta de OH et al. (2008)

OH et al. (2008) é o primeiro trabalho a propor uma abordagem para a descoberta de CLLs ausentes entre as Wikipédias em japonês e inglês. A abordagem deles para a seleção do conjunto de candidatos leva em consideração a análise morfológica dos artigos envolvidos. Para realizar o cálculo de similaridade, dado um artigo a do idioma α é gerado um vetor de atributos $V(a)$ que representa as características de cada artigo a. $V(a)$ é composto de outros três outros vetores ($V(a)$ = $<V_L(a)$, $V_T(a)$, $V_C(a)>$) definidos como segue:

- $V_L(a)$: conjunto de títulos de $L(a)$ onde $L(a)$ é o conjunto de artigos diretamente conectados em a tanto por *out-links* quanto por *in-links*. *Out-links* são os links com os artigos que são referenciados no texto de a, enquanto *in-links* são os links dos artigos que referenciam a no seu texto,

- $V_T(a)$: conjunto de termos de $T(a)$ onde cada sentença é analisada morfologicamente para a identificação de substantivos e de sintagmas nominais, e

- $V_C(a)$: conjunto de termos de $C(a)$ contextualmente relacionados a a. Tal conjunto possui todos os termos que ocorrem em uma janela de 5 palavras de a.

O vetor $V(a)$, portanto, é comparado com os vetores $V(b)$ no idioma β e um escore de similaridade entre $V(a)$ e $V(b)$ é atribuído. Se o par $\langle V(a), V(b)\rangle$ possuir uma similaridade superior a um determinado limiar, então b fará parte do conjunto de candidatos de a.

Uma vez que tem-se o conjunto de candidatos, o mesmo é submetido a um classificador que emprega 14 atributos. Esses atributos estão divididos em quatro categorias: *(i)* estrutura de links (3 atributos), *(ii)* característica do texto (4 atributos), *(iii)* estrutura do

texto (4 atributos) e *(iv)* título (3 atributos). Na avaliação da abordagem, foram reportados valores de 93,4% de precisão e 79,7% de revocação.

É importante destacar que a abordagem utilizada por OH et al. (2008) depende de atributos que são específicos para o par de línguas inglês-japonês. Um exemplo seria o primeiro atributo da categoria *(iii)* "estrutura do texto" denominado por "TS1": tal atributo utiliza a propriedade de que, em muitos casos, o artigo em japonês possui na primeira sentença do texto a tradução do título, entre colchetes, para o artigo em inglês. Além disso, realizar análise morfológica para cada artigo, tanto na descoberta de candidatos quanto para a aplicação dos atributos de similaridade, torna a abordagem apresentada muito custosa.

2.1.2 Proposta de SORG; CIMIANO (2008a)

O trabalho de SORG; CIMIANO (2008a) propõe uma abordagem para encontrar CLLs ausentes entre duas Wikipédias. Essa abordagem foi utilizado como *baseline* de comparação com a nossa abordagem (CLLFinder) devido aos bons resultados reportados por SORG; CIMIANO (2008a). Além disso, a clareza com a qual descreve as técnicas implementadas facilitou a comparação que realizamos com o CLLFinder.

As implementações necessárias bem como a metodologia de comparação e os resultados estão descritos na Seção 4.3.5. Da mesma forma que OH et al. (2008), a abordagem é dividido em duas etapas: seleção de candidatos e aplicação dos atributos de similaridade para descoberta de CLLs ausentes.

Para a primeira, SORG; CIMIANO (2008a) apresenta a *Chain Link Hypothesis* que busca selecionar candidatos para um dado artigo a. Fundamentalmente, essa hipótese supõe que todos artigos correspondentes estão ligados por um caminho formado por um conjunto de links internos (*pagelinks*) e externos (CLLs) ao idioma. Sendo assim, a busca de candidatos se dá pela coleta de todos artigos do idioma de destino que possuam esse caminho com o artigo para o qual se deseja encontrar o correspondente. Visto que a nossa abordagem implementou tal algoritmo, ele está detalhado na Seção 3.2.1.

Os atributos desenvolvidos por SORG; CIMIANO (2008a) para descoberta de CLLs ausentes na segunda etapa podem ser categorizados da seguinte forma:

- baseados em grafos: o artigo em questão é um dos vértices e seus links com outros artigos são as arestas ligando outros vértices. Para essa categoria, existem os seguintes atributos: *Chain Link Count*, *Normalized Chain Link Count*, *Chain Link Inlink Intervals*, *Common Categories* e *CLIA Graph*.

- baseados no texto: são analisados características do título e do texto do artigo para atribuição de escore de similaridade. Os atributos *Distância de Edição* e *Termos em Comum* enquadram-se nessa categoria.

Esses atributos foram implementados para podermos realizar a comparação do CLLFinder com o *baseline* e, portanto, estão detalhados na Seção 4.3.5.

Com base na pontuação de similaridade gerada pelos sete atributos, SORG; CIMIANO (2008a) utilizaram o SVMlight (JOACHIMS, 1999) para gerar o modelo de classificação. Os resultados de experimentos com as Wikipédias em inglês e alemão mostram uma precisão de 93.5% e uma revocação de 69.6%.

2.1.3 Proposta de PENTA et al. (2012)

Mais recentemente, PENTA et al. (2012) propuseram o método *WikiCL* para a descoberta de CLLs ausentes. Tal método modela as Wikipédias de origem (W_s) e destino (W_t)

como grafos direcionados onde os artigos são vértices e as arestas são *links* (CLLs) entre os artigos. Os nodos do grafo que representa determinada Wikipédia podem ser classificados em três classes: (i) nodos que se referem a entidades não geográficas, (ii) nodos que se referem a entidades geográficas e (iii) nodos que não se referem a entidades.

Baseado na classe do artigo para o qual se busca o CLL ausente, três estratégias diferentes podem ser adotadas:

- artigo refere-se à entidade não geográfica: a ideia é realizar a comparação baseada em heurísticas somente entre os artigos que não tratam de entidades geográficas. O método fala em relacionar os artigos candidatos a partir do título, no entanto, adverte sobre problemas quando os alfabetos das Wikipédias não são os mesmos.

- artigo refere-se à entidade geográfica: usualmente, artigos que referem-se a entidade geográficas possuem a informação de latitude e longitude. O método busca, portanto, selecionar os candidatos baseado nas coordenadas geográficas do artigo em questão. Para selecionar candidatos com erros na sua localização, é usado um raio de tolerância a partir das coordenadas do artigo de origem.

- artigo não é uma entidade nomeada: quando o artigo não se refere a uma entidade, um método muito parecido com o *Chain Link Hypothesis* do SORG; CIMIANO (2008a) é utilizado.

A partir do conjunto de candidatos, a identificação do artigo correspondente é feita através de uma medida de relação semântica denominada de *Wikipedia Link-based Measure* (WLM) (Eq. 2.1) apresentada por MILNE; WITTEN (2008). Essa medida é baseada na *Normalized Google Distance*:

$$WLM(v,t) = \frac{max\{logf(v), logf(i)\} - logf(v,i)}{log|W| - min\{logf(v), logf(i)\}} \tag{2.1}$$

onde v e i são dois artigos da Wikipédia, $f(v)$ e $f(i)$ denotam o número de artigos que possuem, respectivamente, *link* com v e i, $f(v,i)$ é o número de artigos que possuem *link* com v e i, e $|W|$ é o número total de artigos da Wikipédia.

Nos experimentos, PENTA et al. (2012) utilizaram os artigos da Wikipédia em inglês para encontrar correspondentes nos idiomas italiano, alemão e francês. Os resultados apresentados atingem precisão entre 89% e 94% e revocação entre 89% e 93%. Os autores compararam a abordagem deles com a de SORG; CIMIANO (2008a) e destacaram que o *WikiCL* possui uma maior revocação, porém uma menor precisão. É importante destacar que a precisão é a medida mais importante para avaliar a descoberta de CLLs ausentes.

2.2 Cross-language Links como suporte à Descoberta de Informação

Além de trabalhos que fazem a descoberta de CLLs ausentes, existem muitos outros trabalhos que fazem uso da funcionalidade de ligação fornecida pelos CLLs. Abaixo, relatamos trabalhos recentes que se enquadram nessa categoria.

O trabalho de (ADAR; SKINNER; WELD, 2009) introduz um mecanismo de detecção e exploração das informações presentes nos campos das infoboxes dos artigos de forma a complementá-las com dados provenientes de idiomas diferentes. Por serem estruturadas, torna-se mais fácil a comparação entre essas estruturas do que, por exemplo, do texto dos artigos da Wikipédia. O método proposto é chamado de *Ziggurat* e tenta

resolver o seguinte problema: para um dado artigo em um idioma contendo uma infobox que não possua o valor para um dado campo, encontrar o valor mais apropriado que quando traduzido produziria uma substituição correta. Visto que a substituição normalmente ocorre de um artigo correspondente em outro idioma, o *Ziggurat* constrói *clusters* que agrupam os artigos correspondentes. A primeira etapa do algoritmo é a o alinhamento de tais artigos (através do seu CLL), onde é gerado um *cluster* contendo artigos equivalentes para cada tópico existente na Wikipédia. Na próxima etapa, são identificados e mapeados os pares correspondentes de cada infobox. Por fim, é estimada a probabilidade desse mapeamento ter sido feito de forma correta e definido um limiar de corte. Segundo o artigo, o classificador alcançou 90,7% de precisão na definição de pares de valores.

O trabalho de (BOUMA; DUARTE; ISLAM, 2009) é semelhante ao de (ADAR; SKINNER; WELD, 2009), e apresenta uma abordagem de alinhamento de informações na Wikipédia através dos CLLs. Tal abordagem completa infoboxes de um idioma a partir de informações contidas no conceito de *templates* de outro idioma. A ideia é expandir automaticamente a quantidade de informações presentes nos *templates* de artigos em um idioma destino com base nas informações contidas nos artigos equivalentes do idioma origem. Nos experimentos, dada uma página em inglês e sua página equivalente em holandês, primeiro são encontradas todas as tuplas inglês-holandês na forma de <atributo,valor> que têm o mesmo valor. Baseado na frequência dessa correspondência é criado um mapeamento bidirecional entre os atributos ingleses e holandeses. Assim é possível usar esse conjunto para expandir o número de atributos em holandês, baseado nas páginas em inglês. Além disso, é possível normalizar o nome dos atributos e detectar inconsistências nos valores. Segundo (BOUMA; DUARTE; ISLAM, 2009), foi possível expandir o número de *templates* holandeses em torno de 50%, além de normalizar informações e detectar possíveis inconsistências.

O trabalho de NGUYEN et al. (2011) apresenta uma abordagem para identificação de mapeamento entre os atributos das infoboxes de páginas em diferentes idiomas. Conforme NGUYEN et al. (2011), existem vários desafios envolvendo tal mapeamento de atributos. Um deles são as diferentes estruturas que as infoboxes podem ter. Por serem feitas por autores diferentes, elas contém dados diferentes que não podem ser sempre mapeados de forma idêntica ao seu correspondente. É proposto, portanto, o *WikiMatch*, um método de mapeamento de esquemas multilíngue através dos CLLs que pega evidências de similaridades de diferentes fontes: valor de atributos, estrutura de links, estatísticas de co-ocorrência entre idiomas e derivação automática de dicionário bilíngue. O *WikiMatch* possui três passos: identificação do mapeamento entre entidades com tipos, a verificação da similaridade de cada par mapeado e a identificação de correspondências adicionais de atributos não mapeados para melhorar a revocação.

2.3 Recuperação de Informações Multilíngue

Conforme GEY; KANDO; PETERS (2005), a Recuperação de Informações Multilíngue (RI-ML) preocupa-se com a busca de documentos em um idioma através de uma consulta realizada em outro idioma. A RI-ML difere-se, portanto, da Recuperação de Informações (RI) tradicional visto que nessa tanto os documentos quanto a consulta encontram-se no mesmo idioma. Já a consulta em RI-ML é criada com palavras-chave de um idioma e será aplicada a um conjunto de documentos que estão em outro idioma (NIE, 2010).

Abaixo, GREFENSTETTE; GREGORY (1998) define os três principais problemas associados à RI-ML:

- como o termo de um idioma será escrito no outro idioma,
- quais as possíveis traduções deverão ser mantidas, e
- como pesar corretamente as traduções quando mais de uma é selecionada.

As estratégias utilizadas para possibilitar a RI-ML são dividas conforme segue:

- Tradução Automática: consiste em traduzir automaticamente a consulta para o idioma da coleção de documentos.
- Tesauro: sistema procura os termos da consulta informados pelo usuário e realiza a substituição dos termos originais pelos sinônimos encontrados no tesauro.
- Dicionário Eletrônico: termos das consultas são substituídos pelos termos traduzidos pelo dicionário.
- Baseado em Corpus: análise de coleções de textos em vários idiomas de forma a extrair de maneira automática a informação para saber se um termo pode ser mapeado para outro idioma.

Segundo STOKOE; OAKES; TAIT (2003), os trabalhos que utilizam RI-ML mesclam diferentes técnicas de forma a obter o resultado desejado. Tais técnicas podem ser aplicadas tanto sobre a consulta quanto sobre a coleção. Em geral, prefere-se a a aplicação de técnicas sobre as consultas visto que essas possuem um tamanho muito menor do que a coleção, facilitando tal operação.

2.4 Sumário

Este Capítulo apresentou os principais trabalhos relacionados à abordagem proposta. Na Seção 2.1, foram descritos os trabalhos de OH et al. (2008), SORG; CIMIANO (2008a) e PENTA et al. (2012). Na Seção 2.2, foram relatados trabalhos que utilizam os CLLs para a exploração, descoberta e correção de informações. Na Seção 2.3, foi apresentado um resumo sobre a Recuperação de Informações Multilíngue que é um dos métodos utilizados neste trabalho para realizar a seleção de artigos candidatos..

A abordagem proposta nessa dissertação, que é descrita no próximo Capítulo, difere dos trabalhos apresentados neste Capítulo em alguns pontos. Diferentemente da abordagem de (OH et al., 2008), a nossa foi implementada de forma a não utilizar nenhuma propriedade específica de determinado idioma. Comparado com as abordagem existentes, o CLLFinder é o único que utiliza a Recuperação de Informações Multilíngue através da indexação e consulta de artigos para a seleção do conjunto de candidatos. Além disso, o nosso trabalho simplifica o processo de análise de similaridade utilizando uma menor quantidade de atributos. De acordo com os nosso experimentos detalhados no Capítulo 4, o CLLFinder alcançou níveis mais altos de precisão e revocação quando comparado com o *baseline* de SORG; CIMIANO (2008a).

3 CLLFINDER PARA DESCOBERTA DE *CROSS-LANGUAGE LINKS* AUSENTES

Neste trabalho, a descoberta de CLLs ausentes é realizada em duas etapas principais:

(i) seleção de um conjunto de candidatos para um determinado artigo, e

(ii) identificação de artigos equivalentes através da pontuação de evidências de similaridade entre os mesmos.

Ainda na última etapa, é utilizado um classificador que identifica quais pares de artigos são correspondentes conforme o coeficiente de similaridade das características analisadas. Dessa forma, esse capítulo descreve uma visão geral da *abordagem* (Seção 3.1) composta pelos três módulos do CLLFinder: *Seleção do Conjunto de Candidatos* (Seção 3.2), *Identificação de Artigos Equivalentes* (Seção 3.3) e *Classificador* (Seção 3.4). O resumo do capítulo está descrito na Seção 3.5.

3.1 Visão geral da abordagem para a Identificação de CLLs Ausentes

A abordagem utilizada para o desenvolvimento e implementação do CLLFinder foi dividida nos seguintes módulos:

(i) seleção do conjunto de candidatos,

(ii) verificação de evidências de similaridade entre um determinado artigo e seus candidatos, e

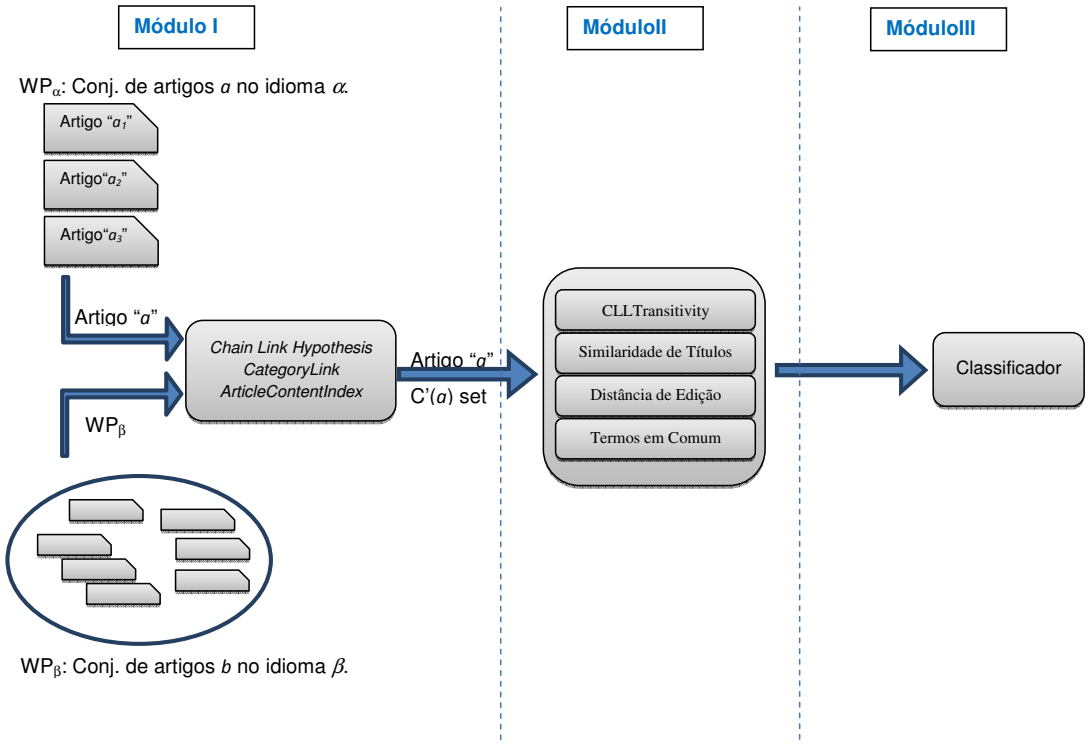
(iii) submissão desse conjunto para um classificador.

O primeiro módulo tem por objetivo restringir o conjunto de artigos candidatos a serem comparados na segunda etapa com o artigo para o qual se busca o CLL ausente. Sendo assim, para um dado artigo a de um idioma de origem α, o primeiro módulo é responsável por reduzir o número de candidatos do idioma de destino β que serão comparados com o artigo a na busca pelo CLL de a. Tal etapa é necessária visto que não é viável comparar todos os pares possíveis de artigos da Wikipedia entre os idiomas α e β. De acordo com os números da tabela 1.1, a Wikipédia em inglês possui aproximadamente 3,6 milhões de artigos. Dessa forma, não seria factível comparar cada artigo em português sem CLL com todos artigos em inglês devido ao tempo de processamento necessário para realizar tal operação. Portanto, dado que WP_α é o conjunto de artigos do idioma α e WP_β é o conjunto de artigos do idioma β, para cada artigo $a \in WP_\alpha$, esse módulo gera um conjunto restrito de candidatos $C'(a)|C'(a) \subset WP_\beta$.

O segundo módulo é responsável pela análise de atributos que refletem a similaridade entre um par de artigos da Wikipédia. Dessa forma, para cada par de artigos $\langle a, b \rangle$ onde $a \in WP_\alpha$ e $b \in C'(a)$, que foi gerado no Módulo I, é realizada a análise de correspondência entre a e b. Nessa etapa, foram desenvolvidos e aplicados os seguintes atributos para a identificação de similaridade entre artigos: `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`.

No terceiro e último módulo, os coeficientes de similaridade calculados pelo Módulo II são submetidos ao classificador com o objetivo de gerar um modelo de classificação. A figura 3.1 mostra de que forma esses três módulos se relacionam, bem como suas entradas e saídas.

Figura 3.1: Módulos do CLLFinder



3.2 Seleção do Conjunto de Candidatos

A seleção de um conjunto de candidatos para cada artigo para o qual se deseja encontrar CLLs ausentes é um problema comum para abordagens que lidam com tal questão. O objetivo deste módulo é alcançar um alto percentual de presença do artigo correspondente dentro do conjunto de candidatos. Tal etapa é fundamental para o nosso trabalho, pois de nada adianta possuirmos bons atributos de similaridade se o artigo correspondente não se encontra dentro do conjunto de candidatos.

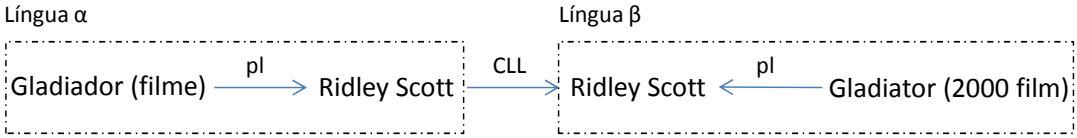
Nas Seções 3.2.1, 3.2.2 e 3.2.3, apresentamos, respectivamente, os seguintes algoritmos para a seleção de candidatos: *Chain Link Hypothesis*, *CategoryLink* e *ArticleContentIndex*. Na Seção 3.2.4 realizamos a análise da organização do conjunto de candidatos bem como a forma com que os limiares são aplicados.

3.2.1 Chain Link Hypothesis

O trabalho de SORG; CIMIANO (2008a) propõe um algoritmo baseado na hipótese denominada de *Chain Link Hypothesis* que busca encontrar candidatos para um determinado artigo através de um caminho chamado de *Chain Link*. O *Chain Link* é formado por dois tipos de *links*: *pagelinks*, que são *links* entre artigos de um mesmo idioma, e CLLs. A definição de um *Chain Link* é como segue: para duas Wikipédias distintas WP_α e WP_β, existe um *Chain Link* entre dois artigos $a_\alpha \in WP_\alpha$ e $a_\beta \in WP_\beta$ se $a_\alpha \xrightarrow{pl} b_\alpha \xrightarrow{CLL} b_\beta \xleftarrow{pl} a_\beta$, onde pl é a sigla para um *pagelinks*. Se existir um *Chain Link* entre a_α e a_β, então a_β passa a fazer parte do conjunto de artigos candidatos de a_α. Sendo assim, o *Chain Link Hypothesis* assume a hipótese de que um artigo possui pelo menos um *Chain Link* com o seu correspondente em outro idioma e, portanto, explora tal propriedade para coletar os possíveis candidatos de um artigo.

A Figura 3.2 mostra um exemplo do *Chain Link Hypothesis* para uma situação onde há artigos correspondentes. Dado que α e β são os idiomas em português e inglês, temos que o artigo *Gladiador (filme)* possui um *Chain Link* para o seu artigo correspondente *Gladiator (2000 film)* através da relação de *pagelinks* que ambos artigos possuem com os artigos *Ridley Scott*, dos seus respectivos idiomas, e através do CLL existente entre esses dois últimos.

Figura 3.2: Funcionamento do *Chain Link Hypothesis*



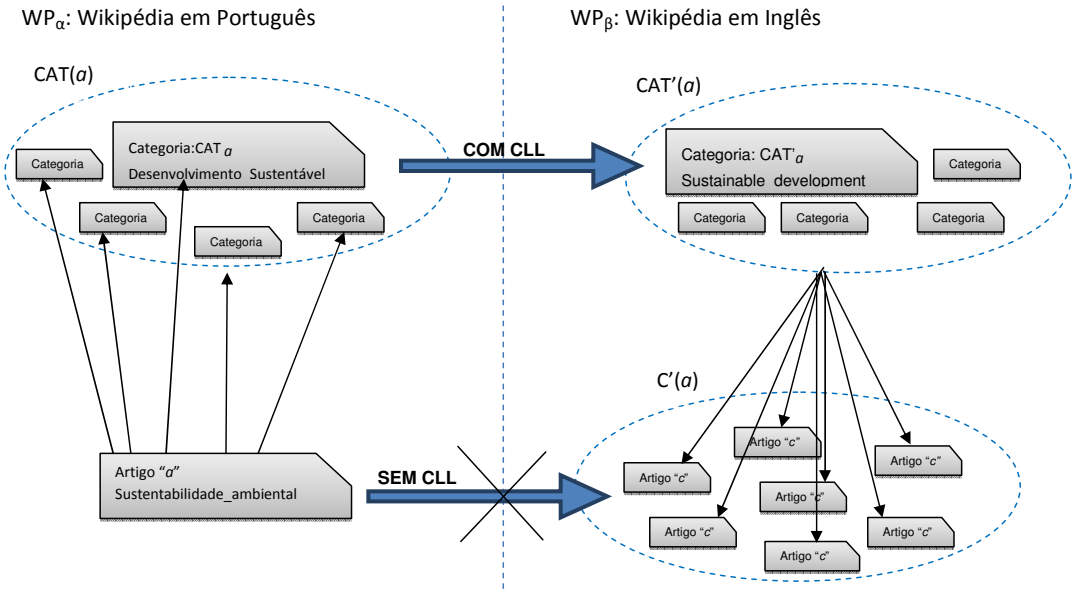
3.2.2 CategoryLink

A categorização de artigos é uma funcionalidade da Wikipédia que permite que artigos sejam colocadas em categorias, possibilitando que os leitores naveguem entre artigos relacionados. Dessa forma, as *categorias* na Wikipédia são conceituadas como páginas especiais que agrupam todos os artigos que pertencem a determinada categoria. A representação do relacionamento entre as categorias de um mesmo idioma é dada em forma de grafo (e não árvore), visto que categorias podem ser definidas como subcategorias de uma ou mais categorias. Tal propriedade ajuda os leitores a encontrar artigos sobre temas relacionados, mesmo sem saber sobre a existência de determinado artigo.

Um artigo também pode ser atribuído a uma ou mais categorias. O artigo da atriz *Julia Roberts*, por exemplo, pertence às seguintes categorias: *Nascidos em 1967*, *Atores dos Estados Unidos*, *Atrizes premiadas com o Oscar*, *Atrizes premiadas com o Globo de Ouro*, dentre outros. Quando verificamos, por exemplo, os artigos que estão associados a categoria *Atores dos Estados Unidos*, encontramos artigos que estão relacionados por falarem sobre atores dos Estados Unidos.

Valendo-se da propriedade de agrupar artigos relacionados em categorias, é possível supor que existe uma probabilidade de que artigos de categorias correspondentes em idiomas diferentes possam ser também correspondentes. Sendo assim, foi desenvolvido o método chamado *CategoryLink* para a seleção do conjunto de candidatos. Tal método considera tanto as categorias do artigo do idioma de origem quanto as categorias correspondentes do idioma de destino para construir o conjunto de candidatos do artigo de origem. Este mecanismo está exemplificado na Figura 3.3.

Figura 3.3: Funcionamento do *CategoryLink*



O objetivo do *CategoryLink* é encontrar candidatos para o artigo $a \mid a \in WP_\alpha$, cujo título é *Sustentabilidade Ambiental*. Sendo assim, é selecionado todo o conjunto de categorias de a representada por $CAT(a)$. Para cada categoria de $CAT(a)$, verifica-se quais possuem CLL de forma que $CAT_a \xrightarrow{CLL} CAT'_a$, formando o conjunto $CAT'(a)$. Por fim, o conjunto de candidatos de a representado por $C'(a)$, que está no idioma β, será formado por todos artigos $c|c \xrightarrow{pertence\ a\ categoria} CAT'(a)$ e $c \in WP_\beta$.

3.2.3 ArticleContentIndex

Tanto o método *Chain Link Hypothesis* (Seção 3.2.1) quanto o *CategoryLink* (Seção 3.2.2) utilizam, respectivamente, propriedades de *links* entre artigos e *links* entre categorias para realizar a seleção do conjunto de candidatos. Ainda que tais propriedades contribuam para a determinação dos possíveis candidatos, o conteúdo do próprio artigo não está sendo utilizado, apenas os seus relacionamentos com outros artigos do mesmo idioma e as categorias as quais ele pertence. Visto que artigos correspondentes possuem conteúdo semelhante, torna-se interessante explorar tal propriedade para a seleção de um conjunto de candidatos.

Sendo assim, foi desenvolvido o método *ArticleContentIndex* para a seleção do conjunto de candidatos através da aplicação da RI-ML, descrita na Seção 2.3. Fundamentalmente, o *ArticleContentIndex* trabalha com a indexação e consulta de documentos. A consulta a índices de documentos consiste em comparar os termos da consulta com os termos do índice e retornar os documentos relevantes à pesquisa que contêm os termos procurados. O resultado é uma lista de documentos ordenados de acordo com o seu escore de similaridade em relação à consulta realizada (MANNING; RAGHAVAN; SCHüTZE, 2008). Na Figura 3.4, está representado o *ArticleContentIndex* que consiste nas seguintes etapas:

(i) obtenção e organização do conteúdo de todos artigos b do idioma de destino β em documentos individuais,

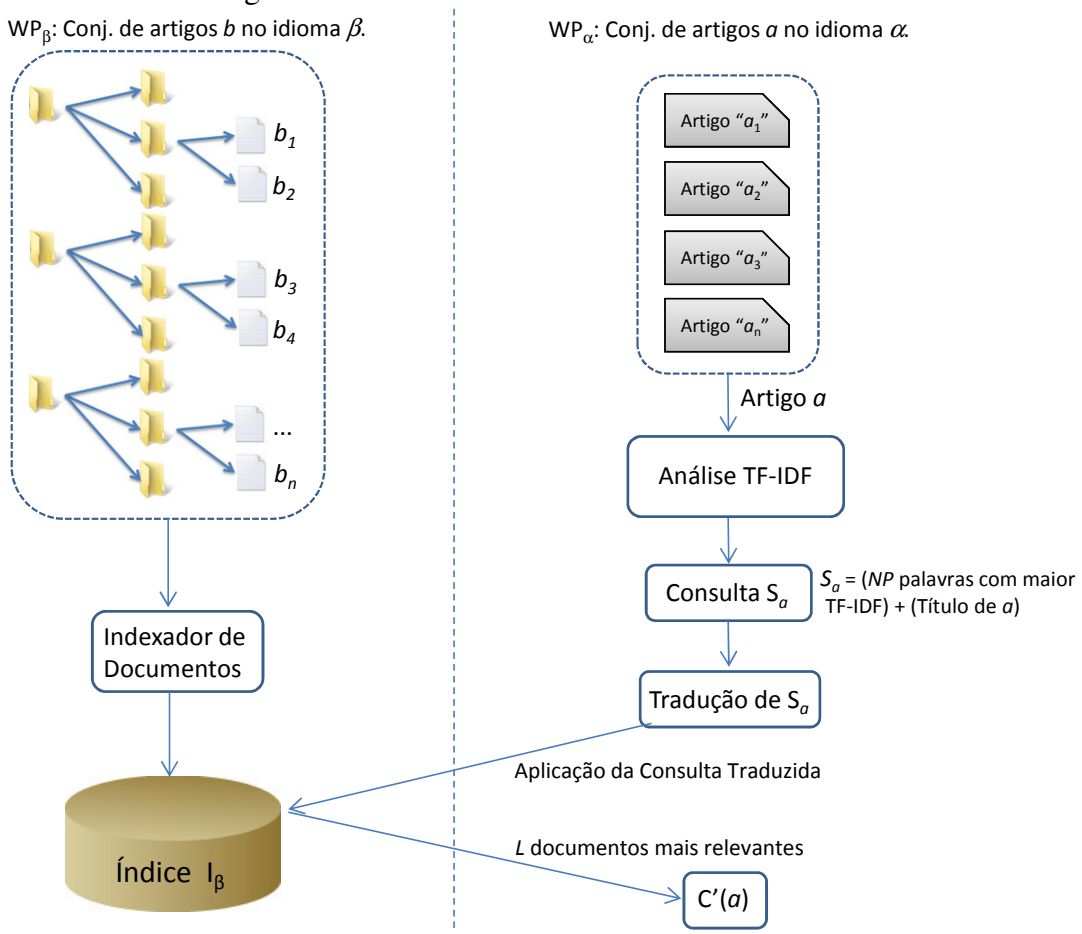
(ii) criação do índice I_β utilizando uma ferramenta de indexação,

(iii) preparação da consulta S_a conforme conteúdo do artigo a do idioma de origem α,

(iv) tradução da consulta S_a

(v) aplicação da consulta S_a traduzida sobre o índice I_β.

Figura 3.4: Funcionamento do *ArticleContentIndex*



Conforme as etapas acima, a definição do *ArticleContentIndex* segue da seguinte forma: dado que a é artigo do idioma de origem α, β é o idioma de destino e I_β é o índice formado pelos artigos do idioma de destino β, S_a é a consulta formada a partir do conteúdo de a a ser aplicada sobre o I_β. Já a consulta S_a é formada pelas NP palavras que possuem o maior *TF-IDF* (*Term Frequency - Inverse Document Frequency*) de a acrescidas do título de a. Na Seção 4.2.2 explicamos a escolha do limiar NP. Antes de ser executada sobre o índice cujos documentos estão no idioma β, a consulta é traduzida para o idioma de destino (β) usando o dicionário do Microsoft Developer Network[1]. Tal operação nos permite a efetiva realização da Recuperação de Informações Multilíngue (RI-ML).

Segundo MANNING; RAGHAVAN; SCHüTZE (2008), o esquema *TF-IDF* (Eq. 3.1) é bastante utilizado na área de Recuperação de Informações e mede, respectivamente, a frequência do termo no documento e a frequência do termo no conjunto de documentos da coleção. Termos que ocorrem com maior frequência em um documento são bons discriminadores do mesmo e, portanto, possuem maior peso. Por outro lado, termos que ocorrem

[1] `http://msdn.microsoft.com`

em muitos documentos não são capazes de diferenciar um documento do outro. Portanto, os melhores termos de indexação são aqueles que aparecem com maior frequência em um documento, possuindo alto valor de TF, e aparecem com pouca frequência dentro da coleção, resultando em alto valor de IDF.

O *TF-IDF*, portanto, resulta em um score que reflete o quão importante é um termo de um documento dentro de uma coleção de documentos. Visto que estamos buscando documentos do idioma de destino β passíveis de serem equivalentes ao artigo a, realizar uma consulta com os termos de mais peso no texto de a nos parece uma boa estratégia para alcançar o resultado desejado. Sendo assim, para cada termo do texto de a, foi feito o cálculo do seu *TF-IDF*, que é o resultado da multiplicação do $tf(t, d)$, que é o número de vezes que o termo t ocorre no documento d, pelo $idf(t, D)$ (Eq. 3.2).

$$tfidf(t, d, D) = tf(t, d)Xidf(t, D) \tag{3.1}$$

onde t é uma termo de d, d é o artigo a em questão e D é o conjunto de documentos de I_β.

$$df(t, D) = \frac{|D|}{|d \in D : t \in d|} \tag{3.2}$$

onde |D| é o número total de documentos de I_β e $|d \in D : t \in d|$ é o número de documentos onde t aparece.

Para complementar a consulta a ser submetida ao índice I_β, o título do artigo a é adicionado à consulta e a consulta é traduzida para o idioma β. Através de testes realizados para um conjunto restrito de artigos, verificou-se que muitos dos artigos correspondentes possuem títulos similares. Visto que os termos do título de um artigo são extremamente representativos do mesmo, devem necessariamente constar na consulta ao índice, o que justifica a inclusão do título de a na consulta S_a.

Após a aplicação da consulta, tem-se o retorno dos documentos cujo os artigos pertencem ao conjunto total de artigos do idioma de destino β. A esse subconjunto de artigos retornados, aplica-se o limiar L para selecionar os L artigos mais relevantes, ou seja, com os termos mais representativos em relação à consulta, e incluí-los no conjunto de artigos candidatos de a representado por $C'(a)$. O *ArticleContentIndex*, portanto, leva em consideração a ideia de que artigos equivalentes possuem conteúdo semelhante. Desta maneira, são acrescentados ao conjunto de candidatos os L artigos mais similares à consulta aplicada.

3.2.4 Organização do Conjunto de Candidatos

O conjunto de candidatos é formado pelo resultado da aplicação dos três métodos acima: *Chain Link Hypothesis* de SORG; CIMIANO (2008a), *CategoryLink* e *ArticleContentIndex*. Nesse ponto, é importante observar que pelo funcionamento do *CategoryLink*, determinado artigo candidato c pode repetir-se muitas vezes dentro do $C'(a)$, pois o mesmo pode fazer parte de mais de uma categoria do conjunto $CAT'(a)$. Da mesma forma, o conjunto de candidatos gerados pelo *Chain Link Hypothesis* pode possuir repetição visto que a partir do artigo a podem existir vários *Chains Links* para um candidato c. Em ambos os casos, quanto maior for o número de repetições de um determinado candidato c maior é a sua chance de ser o correspondente do artigo a.

A Tabela 3.1 mostra a ordenação conforme a repetição dos candidatos do conjunto $C'(a)$ dos artigos em inglês gerados a partir do artigo *Aves* em português. No exemplo

mostrado, o candidato *Bird*, que efetivamente é o artigo correspondente de *Aves*, ocupa a primeira posição no rank visto que é o artigo dentro de $C'(a)$ que mais possui repetições.

Tabela 3.1: Candidatos para o artigo *Aves* ordenados pelo seu número de ocorrências

Artigo	**Artigo Candidato**	**Núm. de Ocorrências**	**Rank**
Aves	Bird	288	1
Aves	List_of_birds	184	2
Aves	Archaeopteryx	172	3
Aves	Palaeognathae	168	4
Aves	...	...	...
Aves	Odonata	10	1913
Aves	...	...	...
Aves	Pesticide	6	7619

Em experimentos realizados com a aplicação do *Chain Link Hypothesis* e do *CategoryLink*, verificamos que o conjunto de candidatos gerados para um artigo pode conter, em média, 150 mil artigos. Esse número, ainda que represente uma restrição em relação ao conjunto total de artigos da Wikipédia para o idioma de destino β, continua sendo muito grande para aplicar os coeficientes de similaridade. Visto que possuímos um conjunto ordenado pelas evidências de similaridade coletadas, o mesmo deve ser restringido por um limiar N que limita o conjunto aos N primeiros candidatos.

Dessa forma, para o *Chain Link Hypothesis* e o *CategoryLink*, foi implementado um mecanismo de restrição dentro do conjunto de candidatos que faz uma diferenciação devido ao número de ocorrências de cada candidato selecionado por esses dois métodos. Já para o método *ArticleContentIndex*, não existe esse conceito de relevância relacionada à repetição do candidato, pois o conjunto de artigos retornado pelo mesmo já está ordenado conforme a similaridade dos termos do texto em relação à consulta aplicada ao índice.

Sendo assim, para os artigos pertencentes ao conjunto de candidatos $C'(a)$ encontrados pelos métodos *Chain Link Hypothesis*, *CategoryLink* e *ArticleContentIndex*, são selecionados efetivamente como candidatos os que estiverem dentro dos seguintes critérios de restrição:

- dentro do limiar L, que já foi definido na consulta ao índice, para o método *ArticleContentIndex*, e
- os N primeiros candidatos dos métodos *Chain Link Hypothesis* e *CategoryLink*.

Na Seção 4.2.2 explicamos a escolha dos limiares L e N. Na Tabela 4.3 do Capítulo 4 é possível verificar a revocação para cada um dos limiares N utilizados.

3.3 Identificação de Artigos Equivalentes

O módulo de identificação de equivalência entre artigos possui duas entradas:

(i) o artigo a do idioma de origem α para o qual deseja-se saber o artigo correspondente b do idioma de destino β, e

(ii) o conjunto de candidatos $C'(a)$ gerado no módulo anterior.

Portanto, para cada par $\langle a, c\rangle \mid c \in C'(a)$, onde $C'(a)$ é conjunto de candidatos de a, quatro atributos de similaridade são calculados e utilizados com o objetivo de treinar o classificador: `CLLTransitivity` (Seção 3.3.1), `Similaridade de Títulos` (Seção 3.3.2), `Distância de Edição` (Seção 3.3.3) e `Termos em Comum` (Seção 3.3.4).

3.3.1 Cross-language Link Transitivity

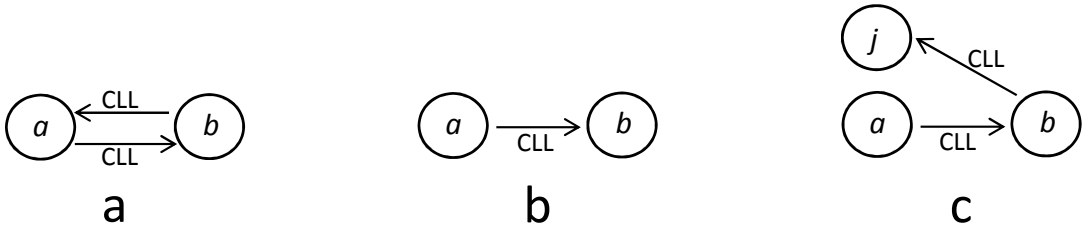
O `CLLTransitivity` é um atributo proposto nesse trabalho que explora a transitividade de CLLs entre outros idiomas da Wikipédia. Ele baseia-se na ideia de que um CLL ausente pode ter ocorrido devido ao fato de um autor de determinado artigo a do idioma α não ter reconhecido o artigo b, do idioma β, como equivalente de a, porém ter atribuído CLL corretamente para o artigo g de um terceiro idioma γ. Por sua vez, no idioma γ o autor pode ter reconhecido que b é o artigo equivalente de g e colocado CLL de g para b. Essa transitividade de CLLs ilustrada acima por $a \xrightarrow{CLL} g \xrightarrow{CLL} b$ caracteriza-se como uma evidência de similaridade entre a e b justificando o estudo e implementação de uma ferramenta que explore tal propriedade.

Dado que α, β e γ são três idiomas distintos, para cada candidato c do par $\langle a, c\rangle$, onde $a \in WP_\alpha$, $c \in C'(a)$ e $C'(a) \subset WP_\beta$, o algoritmo verifica a existência de dois CLL $a \xrightarrow{CLL} g \mid g \in WP_\gamma$ e $g \xrightarrow{CLL} j \mid j \in WP_\beta$ e $j = c$. Se essas três condições forem satisfeitas, significa que é possível navegar a partir do artigo de origem a até seu artigo correspondente c através de outro idioma representado por γ.

Com o objetivo de adicionar confiabilidade e precisão ao atributo, levamos em consideração, em ambas relações ($a \xrightarrow{CLL} g$ e $g \xrightarrow{CLL} j$), o tipo de CLL empregado. Segundo RINSER; LANGE; NAUMANN (2013), existem três possíveis situações para CLLs representadas na Figura 3.5:

- bi-direcionais: conforme a Figura 3.5 (a), para um determinado par de artigos $\langle a, b\rangle | a \in WP_\alpha, b \in WP_\beta \; e \; \exists \; a \xrightarrow{CLL} b$ deve necessariamente existir o CLL $b \xrightarrow{CLL} a$,
- unidirecionais: conforme a Figura 3.5 (b), para um determinado par de artigos $\langle a, b\rangle | a \in WP_\alpha, b \in WP_\beta \; e \; \exists \; a \xrightarrow{CLL} b$, porém não existe o CLL $b \xrightarrow{CLL} a$, e
- conflitantes: conforme a Figura 3.5 (c), para um determinado par de artigos $\langle a, b\rangle | a \in WP_\alpha, b \in WP_\beta, \; \exists \; a \xrightarrow{CLL} b$ e existe CLL $b \xrightarrow{CLL} j | j <> a$.

Figura 3.5: Possíveis situações de CLL



Para o `CLLTransitivity`, trabalhamos apenas com as duas primeiras situações de CLL. Sendo assim, para um conjunto de transitividade dado por $a \xrightarrow{CLL} g$ e $g \xrightarrow{CLL} j$

tanto a relação $a \xrightarrow{CLL} g$ quanto a $g \xrightarrow{CLL} j$ são pontuadas conforme o tipo de situação de CLL que elas representam. O peso dado é conforme a tabela 3.2.

Tabela 3.2: Pesos para as relações de CLLs

Relação	Peso
$\alpha \xrightarrow{CLL} \gamma$ e $\gamma \xrightarrow{CLL} \beta$ (existe caminho)	0.9
$\alpha \xleftarrow{CLL} \gamma$ (primeiro CLL é bidirecional)	0.05
$\gamma \xleftarrow{CLL} \beta$ (segundo CLL é bidirecional)	0.05
Ambos CLLs são bidirecionais	1.0

Esses valores foram manualmente definidos através de experimentos que analisaram a proporção em que ocorrem os *links* bi-direcionais e unidirecionais na nossa base de dados. Sendo assim, os números da tabela refletem a suposição de que quando é possível navegar entre artigos correspondentes sobre os idiomas $\alpha \xrightarrow{CLL} \gamma \xrightarrow{CLL} \beta$ existe 90% de probabilidade de haver um CLL entre os artigos do idioma $\alpha \xrightarrow{CLL} \beta$. Além disso, se a relação $\alpha \xrightarrow{CLL} \gamma$ ou $\gamma \xrightarrow{CLL} \beta$ for bidirecional, é acrescentado 5% de confiança para cada *link*.

No exemplo acima, foi empregado apenas um idioma representada por γ. No entanto, na implementação do `CLLTransitivity` foi utilizada a estratégia de levar em consideração mais de um idioma intermediário. Como poderá ser visto adiante (Seção 4.1.1), três idiomas intermediários foram utilizados. A ideia é que quanto maior o número de idiomas, mais confiável será a pontuação de similaridada gerada. Quando mais de um idioma intermediário é utilizado, torna-se necessário realizar a ponderação do escore conforme o número de idiomas empregados.

O Algoritmo 1 mostra o funcionamento do atributo `CLLTransitivity`. A entrada é composta pelo artigo do idioma de origem a, o conjunto de candidatos $C'(a)$, o idioma de destino dos candidatos β, o conjunto I dos idiomas utilizados e o conjunto de pesos para pontuação da relação de CLL dado na tabela 3.2. Fundamentalmente, para cada candidato $c_n \in C'(a)$ e para cada idioma intermediário $i_n \in I$ ele verifica se existe um artigo intermediário no idioma i_n de forma que $a \xrightarrow{CLL} ArtigoIntermediario$ e $ArtigoIntermediario \xrightarrow{CLL} c_n$. Se existir, ele realiza a pontuação dos CLLs e depois pondera o score baseado no número de idiomas intermediários empregados.

3.3.2 Similaridade de Títulos

A `Similaridade de Títulos` tem o objetivo de realizar o cálculo de similaridade entre o título de determinado artigo a do idioma de origem α e seu candidato $c \mid c \in C'(a)$ do idioma de destino β. A ideia por trás do atributo é que artigos correspondentes, por tratarem do mesmo assunto, podem possuir títulos semelhantes quando comparados em um mesmo idioma. Essa estratégia já foi utilizada em trabalhos relacionados tais como OH et al. (2008); SORG; CIMIANO (2008a); ADAR; SKINNER; WELD (2009).

A execução do atributo `Similaridade de Títulos` envolve os seguintes processos: tokenização do título, remoção das *stopwords* e realização do *stemming* dos termos do título. A tokenização é o processo de extração de termos de um texto onde cada termo é denominado de *token*. Geralmente, nesse processo são descartados caracteres especiais tais como pontuação e contrações. A remoção de *stopwords* é feita para eliminar

Algorithm 1 Cross-language Link Transitivity

```
1: function CALCULO_CLLTRANSITIVITY(a, C'(a), β, I,P)
2:     CLLTransitivity = null
3:     for C_n ∈ C'(a) do
4:         for I_n ∈ I do
5:             VetorTransitividade = null
6:             if existeCLL(a,I_n) then
7:                 ArtigoIntermediario = retornaArtigoCLL(a,I_n)
8:                 if existeCLL(ArtigoIntermediario,β) then
9:                     ArtigoDestino = retornaArtigoCLL(ArtigoIntermediario,β)
10:                    if (ArtigoDestino == C_n) then
11:                        VetorTransitividade[0] = 1
12:                        if (biDirecional(a,ArtigoIntermediario) then
13:                            VetorTransitividade[1] = 1
14:                        end if
15:                        if (biDirecional(ArtigoIntermediario,ArtigoDestino) then
16:                            VetorTransitividade[2] = 1
17:                        end if
18:                    end if
19:                end if
20:            end if
21:            CLLTransitivity+=(VetorTransitividade[0]*P[0] +
    VetorTransitividade[1]*P[1] + VetorTransitividade[2]*P[2])*I/sizeOf(I);
22:        end for
23:    end for
24:    return CLLTransitivity
25: end function
```

palavras comuns dentro de um idioma. Em geral, *stopwords* são compostas por conectores linguísticos tais como artigos, preposições, conjunções, pronomes e advérbios. Visto que possuem pouco valor semântico, esses termos não participam da comparação de títulos. Já o *stemming* tem o objetivo de reduzir uma palavra a sua raiz morfológica. A raiz morfológica é a parte que está presente em todas as derivações da palavra, sendo assim, é uma representação única de todas as palavras que apontam para o mesmo conceito. Isso facilita quando busca-se verificar a relação de similaridade entre duas palavras, pois garante-se que as suas derivações não serão levada sem conta durante a comparação.

O algoritmo escolhido para a realização de *stemming* nesse trabalho é o PorterStemmer (PORTER, 1980), pois o mesmo é comumente aplicado em sistemas de Recuperação de Informações. O uso do PorterStemmer facilita a comparação entre as palavras visto que, conforme a definição acima, ele retira características morfológicas que poderiam diferenciá-las. Um exemplo é a aplicação desse algoritmo sobre a palavra *Cars*. O último *s* significa que tal palavra está no plural, o que pode ser uma característica não desejada na hora de aplicar o Coeficiente de Dice (Eq. 3.3) para a contagem de palavras iguais. Nesse exemplo, o PorterStemmer removeria o plural fazendo com que a palavra ficasse normalizada.

Sendo assim, a `Similaridade de Títulos` é formada pelas seguintes etapas:

(i) os títulos dos artigos dos idiomas α e β são tokenizados,

(ii) os *tokens* do título do idioma α são traduzidos para o idioma de destino β usando o dicionário do Microsoft Developer Network[2],

(iii) com ambos os títulos no mesmo idioma (β), *stopwords* são removidas e é realizado o *stemming*, e

(iv) a similaridade é computada utilizando o Coeficiente de Dice (Eq. 3.3).

$$CoeficientedeDice(DC) = \frac{T_a \cap T_c}{max(T_a, T_c)} \tag{3.3}$$

onde T_a e T_c são o número de *tokens* dos artigos a e c, respectivamente.

É válido ressaltar que somente os títulos dos artigos são traduzidos. Visto que, em geral, o tamanho dos mesmos fica entre uma e três palavras, tal tradução não causa grande impacto em termos de desempenho geral do CLLFinder. Além disso, esse é o único atributo que depende da utilização de ferramentas externas tais como o *stemmer* e o tradutor do Microsoft Developer Network[3].

Nesto ponto, é importante observar que a qualidade do tradutor utilizado influencia na diretamente eficácia do atributo `Similaridade de Títulos`. Uma tradução incorreta de um termo, por exemplo, pode comprometer a comparação de termos iguais em um mesmo idioma realizada pelo Coeficiente de Dice. Por outro lado, quanto melhor for o dicionário empregado, maior a chance dos termos traduzidos efetivamente serem iguais quando comparados em um mesmo idioma. Neste trabalho foi utilizado o tradutor do Microsoft Developer Network, pois foi o único tradutor online gratuito que encontramos. Além disso, ele permitia várias conexões em um curto período de tempo sem realizar o bloqueio das mesmas.

Podemos observar que o cálculo da similaridade entre títulos é dependente de linguagem devido à necessidade de tradução dos títulos. Sendo assim, na Seção 4.3.6, realizamos testes sem o atributo `Similaridade de Títulos` para verificar o comportamento do CLLFinder. Naturalmente, ocorre uma pequena queda de precisão e revocação, porém, a nossa abordagem mantém bons índices dos mesmos devido à complementação dos outros atributos de similaridade implementados. Dessa forma, em uma situação na qual se necessite que o CLLFinder não dependa de ferramentas externas, é possível retirar o atributo `Similaridade de Títulos` e ainda teremos um excelente resultado conforme mostra a Tabela 4.4.

3.3.3 Distância de Edição

O atributo `Distância de Edição`, também conhecido como *Distância de Levenshtein* (LEVENSHTEIN, 1966), foi aplicado diretamente entre os títulos do artigo e seu candidato. Para manter esse atributo independente de linguagem, não realizamos a tradução nem os processos de remoção de *stopwords* e *stemming*. Em testes realizados, tal atributo mostrou bons resultados quando o artigo diz respeito a pessoas, lugares, entidades ou objetos universais, isto é, quando tratam-se de palavras iguais independente do idioma. Isso ocorre porque nessas situações os títulos são iguais ou muito similares em ambos idiomas. Visto que a pontuação dada pela `Distância de Edição` está relacionada ao número mínimo de operações necessárias para transformar um *string* no

[2] `http://msdn.microsoft.com`
[3] `http://msdn.microsoft.com`

outro, será atribuído uma pontuação alta para títulos idênticos ou que diferem em poucas letras.

3.3.4 Termos em Comum

Para um dado artigo a e o seu candidato $c \in C'(a)$, o atributo `Termos em Comum` calcula uma pontuação de similaridade relacionada ao número de palavras em comum entre os textos dos artigos a e c. Da mesma forma que a `Distância de Edição`, com o objetivo de manter esse atributo independente de linguagem, não realizamos a tradução das palavras nem os processos de remoção de *stopwords* e *stemming*. O coeficiente de similaridade é calculado usando o Coeficiente de Dice (Eq. 3.3), ou seja, é a divisão do número de palavras em comum em ambos os textos pelo número total de palavras do artigo que contém mais palavras.

Visto que não existe tradução neste processo, o cálculo dos termos em comum irá gerar uma pontuação mais alta para os idiomas que forem morfologicamente similares do que para os que são diferentes nesse aspecto. No entanto, mesmo idiomas morfologicamente distintos compartilham muitos nomes próprios, número, datas e nomes de localizações, ou seja, possuem em comum palavras que não são passíveis de tradução. Dessa forma, esse evidência será útil para identificar a equivalência entre artigos que compartilham termos citados acima. Esse tipo de atributo de similaridade também já foi utilizado por SORG; CIMIANO (2008a).

3.4 Classificador

A saída do segundo módulo é um conjunto de pares de artigos $\langle a, c \rangle$ seguidos de quatro coeficientes de similaridade dados pelos seguintes atributos: `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`. A pontuação atribuída por cada um desses atributos visa quantificar a probabilidade de equivalência entre o artigo a e seu candidato c. Sendo assim, nesse terceiro módulo, esses coeficientes são submetidos à análise de um classificador que irá determinar quando um par de artigos é ou não equivalente.

Com o objetivo de treinar o classificador, artigos do idioma de origem α para os quais sabe-se quem é o correspondente no idioma destino β serão utilizados como exemplos positivos, ou seja, o atributo de classe terá valor "1". Já os artigos que não são correspondentes serão os exemplos negativos cujo atributo de classe terá valor "0".

Visto que no nosso conjunto total de dados existem menos exemplos positivos, foi utilizado um conjunto menor para o treino de forma que o número de exemplos positivos fosse igual ao número de exemplos negativos. Isso previne problemas em relação ao modelo gerado pelo classificador, pois se o conjunto de treinamento de entrada possuísse muito mais instâncias negativas, o modelo gerado seria tendencioso em favor da classe negativa.

Após a obtenção do modelo de classificação, o classificador está pronto para analisar um conjunto diferente de artigos para os quais o atributo de classe é desconhecido. A nossa abordagem não foi construída de forma a trabalhar com uma técnica de classificação específica. Em nossos experimentos, foi utilizado o modelo de árvores de decisão, no entanto, outras técnicas poderiam ter sido utilizadas.

3.5 Sumário

Este Capítulo apresentou a nossa abordagem (CLLFinder) para a descoberta de CLLs ausentes. Dada a importância desses *links*, o objetivo do nosso trabalho é encontrá-los de forma a contribuir para o aumento da cobertura de CLLs na Wikipédia.

A abordagem utilizada é composta da seleção do conjunto de candidatos, verificação de evidências de similaridade entre um determinado artigo e seus candidatos e submissão desse conjunto para um classificador. Para a primeira parte, foram descritos os algoritmos *Chain Link Hypothesis*, *CategoryLink* e *ArticleContentIndex*. Na segunda, são apresentados os atributos de similaridade `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`. Por fim, especifica-se a forma pela qual os escores gerados por tais atributos são submetidos ao classificador.

No Capítulo 4, estão descritos os experimentos realizados para a validação da abordagem proposta.

4 EXPERIMENTOS

Este capítulo apresenta os experimentos realizados para avaliar a abordagem CLLFinder proposta neste trabalho. Primeiramente, descreveremos a metodologia experimental utilizada (Seção 4.1) detalhando os elementos da Wikipédia envolvidos, os procedimentos para a extração do conteúdo dos artigos e a seleção do conjunto de artigos a ser verificado. Depois, são apresentados os resultados para a seleção de candidatos (Seção 4.2) e para a descoberta de CLLs ausentes (Seção 4.3). A comparação com o *baseline* proposto por SORG; CIMIANO (2008a) é realizada tanto na seleção de candidatos (Seção 4.2) quanto na etapa de descoberta de CLLs ausentes (Seção 4.3). Na Seção 4.4 realizamos a aplicação da nossa abordagem em um cenário real de descoberta de CLLs ausentes e na Seção 4.5 encontra-se o resumo do capítulo.

4.1 Metodologia

Nossos experimentos foram realizados de forma a encontrar CLLs ausentes entre artigos da Wikipédia em português e inglês, sendo que, o português será o idioma de origem e o inglês, o de destino.

4.1.1 Elementos da Wikipédia e suas Relações

Entenda-se como elementos da Wikipédia os seus artigos, categorias, CLLs, *pagelinks* e o conteúdo textual dos artigos. Para trabalhar com tais componentes, foi feito o *download* do *dump* da base de dados das Wikipédias nos idiomas inglês e português disponibilizados periodicamente pela própria Wikipédia no site *http://dumps.wikimedia.org/*. Para o idioma português, o *download* foi feito da data *10/04/2011* através do link *http://dumps.wikimedia.org/ptwiki/20110410/*, enquanto que para o idioma inglês, o *download* foi feito da data *05/04/2011* através do link *http://dumps.wikimedia.org/enwiki/20110405/*. Para ambos, foram importados os seguintes arquivos:

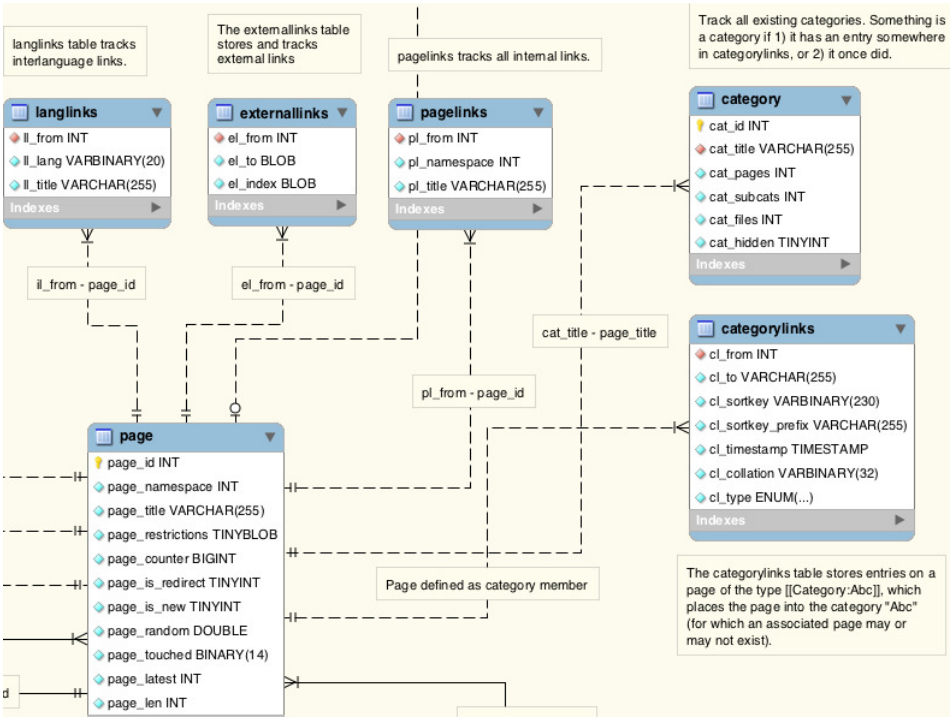
- *category.sql.gz*: tabela onde estão armazenadas as informações sobre as categorias da Wikipédia,
- *page.sql.gz*: tabela onde estão armazenadas as informações sobre os artigos da Wikipédia,
- *categorylinks.sql.gz*: tabela que faz o relacionamento entre artigo e as categorias ao qual ele pertence,

- *langlinks.sql.gz*: tabela onde estão armazenadas os CLLs dos artigos do idioma em questão para outros idiomas,
- *pagelinks.sql.gz*: tabela onde estão armazenadas os *pagelinks* para os artigos do mesmo idioma, e
- *pagesarticles.sql.gz*: documento *xml* contendo todos os artigos do idioma em questão.

O armazenamento desse conjunto de tabelas foi feito utilizando o banco de dados MySql 5.5 Server. Visto que ambos os idiomas português e inglês possuem o mesmo conjunto de tabelas disponibilizadas pela Wikipédia, para armazená-las em um mesmo banco foi alterado o nome das tabelas em português adicionando a letra "p" no início, de forma que pudessem coexistir em uma mesma instância do banco. Além disso, foram necessárias a realização de otimizações no banco de dados devido ao grande volume de informações a serem tratadas. A principais foram a criação de índice nas tabelas para as colunas mais acessadas durante as consultas e cruzamento de tabelas e a criação de subtabelas menores apenas com as informações necessárias para determinadas situações de consultas. Na Figura 4.1 é possível verificar o relacionamento entre as tabelas descritas acima.

Para a aplicação do atributo `CLLTransitivity`, foram escolhidos como idiomas intermediários o francês, o italiano e o espanhol. Ressaltamos que essas línguas não precisam ser morfologicamente similares. No entanto, caso tenham um contexto cultural semelhante, existe mais chance de possuírem um número maior de CLLs. Como exemplo, temos que os idiomas basco e espanhol seriam considerados "similares" pela definição acima, mesmo que sejam completamente diferentes sintaticamente. Diferentemente do português e inglês, que são, respectivamente, os idiomas de origem e destino, para

Figura 4.1: Relacionamento entre as tabelas da Wikipédia (MANUAL OF WIKIPEDIA DATABASE LAYOUT, 2011)



os intermediários, foi necessário realizar o *download* somente dos arquivos *page.sql.gz* e *langlinks.sql.gz*. Isso ocorre porque para a implementação do `CLLTransitivity` é necessário saber apenas os artigos e CLLs existentes dos idiomas intermediários. Elementos tais como categorias e links internos aos artigos não são necessários, conforme a descrição do método realizada na Seção 3.3.1.

4.1.2 Wikipedia Extractor e AWK para Extração do Conteúdo de Artigos

A análise do conteúdo do artigo, disponível no arquivo *pagesarticles.sql.gz*, faz parte da implementação proposta. Assim como existe o *dump* das tabelas da Wikipédia, esse arquivo é um *dump* do conteúdo do artigos. Quando descompactado, ele se transforma em um arquivo *XML*, de 4,1Gb para o idioma português e 38Gb para o inglês, contendo o texto de todos os artigos do idioma ao qual ele pertence. Visto que seu tamanho não permitia realizar operações rápidas de busca e leitura, a forma encontrada para poder manipulá-lo foi de quebrá-lo em outros arquivos menores de forma que cada um desses pequenos documentos contivesse o conteúdo de um único artigo e pudesse ser referenciado a partir da sua identificação única (*page_id*).

O processo de quebra do arquivo e separação em documentos menores envolve o processo de extração de conteúdo sem *tags* de marcação e a quebra e organização em uma hierarquia de diretórios. Sobre a extração de conteúdo, foi utilizado a ferramenta *Wikipedia Extractor* (WIKIPEDIA EXTRACTOR, 2013).

Tal ferramenta gera o texto plano a partir do arquivo *pagesarticles.sql.gz* descartando qualquer outra informação ou anotações presentes nos artigos da Wikipédia tais como imagens, tabelas, referências e listas. Abaixo, na Figura 4.2, segue um exemplo do conteúdo do artigo "Clima"antes e depois da aplicação do *Wikipedia Extractor*.

Figura 4.2: Aplicação do WikiExtractor

Antes do WikiExtractor

```
<page>
  <title>Clima</title>
  <id>599</id>
  <revision>
    <id>32270054</id>
    <parentid>32269221</parentid>
    <contributor>
      <username>HVL</username>
      <id>634769</id>
    </contributor>
    <minor />
    <sha1>5j72wmssl4vtp5z35ipzw4f5vsu9klk</sha1>
    <text xml:space="preserve">{{ver desambiguaÃ§Ã£o}}
O '''clima''' (do [[lÃ-ngua grega|grego]] para
&amp;quot;inclinaÃ§Ã£o&amp;quot;, referindo o Ã¢ngulo formado pelo eixo
de ...
...
== Cidades com cada clima ==
* [[Clima tropical|Tropical]] - [[Rio de Janeiro]] - {{BRA}}
...
== {{LigaÃ§Ãµes externas}} ==
* {{Link|pt|2=http://www.tempoagora.com.br|3=Tempoagora}}
</text>
  </revision>
</page>
```

Depois do WikiExtractor

```
<doc id="599>

O clima (do grego para "inclinação", referindo o ângulo
formado pelo eixo de rotação da terra e seu plano de
translação)
compreende um padrão dos diversos elementos
atmosféricos que ocorrem na atmosfera da Terra.
Fenômenos como frente frias,
tempestades, furacões e outros estão associados tanto às
variações meteorológicas preditas pelas leis físicas
determinísticas,
assim como a um conjunto de variações aleatórias dos
elementos meteorológicos (temperatura, precipitação,
vento, umidade, pressão do ar)
cuja principal ferramenta de investigação é a estatística.
...
...
...
Emmanuel de Martonne destacou-se no estudo da
geomorfologia climática. Seu estudo sobre problemas
morfológicos do Brasil tropical-atlântico
foi um dos primeiros trabalhos de geomorfologia
climática, sendo conhecida a "classificação do clima de
Martonne".
```

4.1.3 Seleção do Conjunto de Artigos

Para treinar o classificador e verificar o desempenho da nossa abordagem, foi necessário trabalhar com um domínio de artigos para o qual era conhecido o seu CLL, ou seja, o

seu artigo correspondente. Desta maneira, foi selecionado o conjunto que denominamos de $DS1000$, composto por 1000 artigos em português para os quais o artigo em inglês equivalente era conhecido. O $DS1000$ foi coletado utilizando uma função recursiva que, a partir de uma categoria da Wikipédia e do número desejado de artigos, verificava as subcategorias coletando os artigos pertencentes às mesmas até que o número de artigos desejado fosse atingido.

A coleta dos artigos foi realizada a partir das seguintes categorias em português: *animais*, *internet*, *automobilismo*, *moda*, *filmes*, *atores*, *biologia*, *matemática*, *física* e *computadores*. A estratégia utilizada foi escolher diferentes categorias evitando as que possuíssem artigos relacionados a uma determinada região ou cultura. Para cada uma das 10 categorias acima, foram escolhidos os 100 primeiros artigos retornados pela função recursiva que possuíssem CLL e não eram artigos de redirecionamento ou desambiguação. Essa coleta totalizou os 1000 artigos do conjunto $DS1000$.

4.2 Resultados para a Seleção de Candidatos

Conforme descrito na Seção 3.2, a geração do conjunto de candidatos combina os métodos *Chain Link Hypothesis*, proposto por SORG; CIMIANO (2008a), com os métodos *CategoryLink* e *ArticleContentIndex*, onde, para cada artigo $a_P \in DS1000$ em português, os métodos acima geram o conjunto de candidatos $C'(a_P)$ em inglês. Com o conjunto de candidatos gerado, verificamos se para cara artigo $a_P \in DS1000$ existe um artigo correspondente $a'_E \in C'(a_P)$.

Com o objetivo de analisar a qualidade dos métodos para a seleção de candidatos, os resultados são apresentados em duas etapas: utilizando o método *baseline Chain Link Hypothesis* do SORG; CIMIANO (2008a) combinado com o *CategoryLink* (Seção 4.2.1) e adicionando os candidatos selecionados pelo *ArticleContentIndex* aos candidatos da primeira etapa (Seção 4.2.2). Em ambas as etapas realizamos a comparação da nossa abordagem com o trabalho de *baseline*, que utiliza somente o *Chain Link Hypothesis*, proposto por SORG; CIMIANO (2008a).

Os experimentos são realizados para cada artigo do $DS1000$ e os resultados são apresentados pelo percentual de presença do artigo correspondente dentro de cada uma das 10 categorias utilizadas.

4.2.1 Chain Link Hypothesis e CategoryLink

Na primeira etapa, foram implementados os métodos *Chain Link Hypothesis* e o *CategoryLink*. O primeiro foi realizado conforme a descrição do algoritmo na Seção 3.2.1 que foi retirada do trabalho de SORG; CIMIANO (2008a). Nesse caso, o uso do método de SORG; CIMIANO (2008a) nos foi útil tanto na complementação dos nossos métodos através de uma terceira via para a seleção de candidatos quanto para a criação de um *baseline* para comparar com as soluções propostas nesse trabalho.

Através da Tabela 4.1, podemos verificar os resultados da primeira etapa para os métodos *Chain Link Hypothesis* e *CategoryLink*, bem como uma comparação dos nossos resultados e dos resultados do trabalho de *baseline*. A segunda coluna mostra os resultados para o *Chain Link Hypothesis* sozinho, da forma como foi aplicado no trabalho *baseline*, enquanto a terceira coluna mostra os resultados da combinação proposta para a primeira etapa do experimento. A coluna intitulada *Aumento* mostra, em termos de percentual, o aumento do número de casos em que o artigo correspondente está presente dentro do conjunto de candidatos utilizando o *CategoryLink*. Os resultados indicam que, na média,

Tabela 4.1: Número de artigos correspondentes dentro do conjunto de candidatos

	Chain Link Hypothesis	*Chain Link Hypothesis* + *CategoryLink*	**Aumento**
Animais	67%	84%	+25%
Internet	52%	64%	+23%
Automobilismo	18%	34%	+88%
Moda	54%	75%	+38%
Filmes	20%	95%	+375%
Atores	82%	98%	+19%
Biologia	73%	83%	+13%
Matemática	52%	55%	+5%
Física	55%	63%	+14%
Computadores	63%	85%	+34%
Média	53,6%	73,6%	+ 37,3%

o *CategoryLink* combinado com o *Chain Link Hypothesis* supera o *baseline*, que utiliza somente o *Chain Link Hypothesis*, em 37,3% a mais de presença do artigo candidato para os dados testados, número que valida o *CategoryLink* para a busca de candidatos.

Idealmente, tal melhoria deveria ocorrer sem aumento significativo do número de candidatos, ou seja, o incremento na qualidade não pode ser apenas consequência do aumento do número de candidatos. Além disso, ao analisar os conjuntos resultantes, verificou-se que possuíam, em média, cerca de 150 mil candidatos em cada um. Ainda que este seja um número bem menor do que todo o universo de artigos da Wikipédia em inglês, a utilização desse conjunto tornaria o processo de análise de similaridade custoso sob o ponto de vista computacional. De acordo com o método da Seção 3.2.4, foi considerada uma redução nesse conjunto a partir de um limiar N que limita o conjunto aos N primeiros candidatos ordenados conforme o número de ocorrências do mesmo dentro do conjunto. Dessa forma, foi construída a Tabela 4.2 para verificar a eficiência do *CategoryLink* dentro de cada um dos seguintes limiares N: 1000, 2000, 5000 e 10000. Portanto, para um dado artigo a_P e seu conjunto de candidatos $C'(a_P)$, verificamos se o seu artigo correspondente a'_E ocupa as N primeiras posições de $C'(a_P)$.

Através dos resultados, podemos verificar a evolução da presença dos artigos equivalentes conforme aumentamos o limiar N. Enquanto que para N = 1000 temos uma média de 51,1% da presença do artigo correspondente, para N = 10.000 esse número sobe para 68,1%, um aumento de 33%. Ainda que tal aumento não seja desprezível, esperava-se um crescimento maior visto que o conjunto de candidatos aumentou 10 vezes. Consideramos tal fato positivo, pois significa que existe uma concentração da distribuição dos artigos correspondentes dentro dos 1000 artigos mais bem classificados no conjunto de candidatos.

Além disso, percebemos que a eficiência do *CategoryLink* se repete ao trabalhar com todo o conjunto de candidatos e ao realizar a aplicação dos limiares propostos para selecionamos os N primeiros candidatos. Enquanto que para o conjunto geral de candidatos temos um aumento de 37,3% na média da presença do artigo correspondente, para os limiares de 1000, 2000, 5000 e 10000 obtivemos, respectivamente, 37%, 32%, 32% e 33%. Isso significa que o aumento das ocorrências de artigos gerado pelo *CategoryLink* ocorre em artigos candidatos com maior probabilidade de serem o artigo correspondente, o que melhora o seu *rank* frente a outros candidatos. Dessa forma, o *CategoryLink* mostra-se eficiente ao promover o artigo correspondente a um *rank* mais alto dentro do conjunto

Tabela 4.2: Presença dos artigos correspondentes entre os N primeiros candidatos

	N=1000		**N=2000**		**N=5000**		**N=10000**	
	CLH	CLH +CL	CLH	CLH +CL	CLH	CLH +CL	CLH	CLH +CL
Animais	54%	67%	60%	73%	64%	81%	66%	82%
Internet	40%	48%	44%	51%	48%	56%	49%	60%
Automobilismo	14%	15%	15%	16%	16%	18%	16%	18%
Moda	30%	45%	34%	50%	40%	63%	43%	66%
Filmes	35%	64%	41%	73%	43%	78%	45%	83%
Atores	36%	68%	47%	74%	61%	86%	70%	93%
Biologia	58%	62%	61%	67%	67%	74%	68%	81%
Matemática	30%	32%	34%	36%	39%	42%	47%	47%
Física	41%	48%	47%	54%	52%	60%	53%	63%
Computadores	39%	66%	48%	75%	55%	84%	58%	88%
Média	37,7%	51,5%	43,1%	56,9%	48,5%	64,2%	51,5%	68,1%
Aumento	**+36%**		**+32%**		**+33%**		**+32%**	

de candidatos. Isto aumenta a presença do artigo candidato no conjunto limitado por um limiar.

4.2.2 Adicionando os candidatos do ArticleContentIndex

Na segunda etapa dos experimentos, combinamos ao conjunto de candidatos $C'(a_P)$ gerados pelo *Chain Link Hypothesis* de SORG; CIMIANO (2008a) e pelo *CategoryLink* os candidatos resultantes da aplicação do *ArticleContentIndex*. Conforme relatado na Seção3.2.3, o método *ArticleContentIndex* possui dois parâmetros que orientam o seu funcionamento: o número de palavras NP do texto do artigo com o maior *TF-IDF* a serem usadas na formação da consulta e o limiar L para selecionar os L artigos mais relevantes do resultado da consulta e incluí-los no conjunto de candidatos $C'(a_P)$.

Para o parâmetro NP, quanto maior o número de palavras com o maior *TF-IDF* inseridas na consulta, maior será a chance de retorno de um artigo correspondente. No entanto, o tamanho da consulta é diretamente proporcional ao tempo de consulta ao índice. Visto que estamos interessados no melhor compromisso entre artigos correspondentes retornados e tempo para a execução do métodos, iniciamos os testes utilizando 20 palavras com maior *TF-IDF* Realizamos também, conforme a descrição do método, o acréscimo do título de a às consultas com as NP palavras com maior *TF-IDF*. Ao aumentar o número NP durante os experimentos, verificamos que ao ultrapassar as 50 palavras, não obtivemos melhorias significativas no método e, ao mesmo tempo, o uso de 50 palavras não onerou em demasia o tempo de execução do método. Sendo assim, para o primeiro parâmetro NP, escolhemos usar as 50 palavras de maior *TF-IDF* acrescidas do título do artigo de origem.

Para o parâmetro L, referenciando os L artigos mais relevantes retornados, fizemos experimentos que verificaram, inicialmente, a presença do artigo correspondente para um conjunto formado por $L = 200$. Tal limiar foi sendo ajustado até que chegássemos ao ponto ótimo de presença do artigo correspondente e tamanho do conjunto de candidatos que foi de $L = 1000$. Isso significa que para o *ArticleContentIndex* trabalhamos com os 1000 artigos mais relevantes retornados pela consulta ao índice.

Sendo assim, para determinado artigo a_P, o *ArticleContentIndex* acrescenta 1000 artigos candidatos a serem o correspondente de a_P ao conjunto de candidatos $C'(a_P)$ gerado pelo *Chain Link Hypothesis* de SORG; CIMIANO (2008a) e pelo *CategoryLink*. Este número foi escolhido porque, conforme será visto adiante, utilizaremos também o limiar $N = 1000$ para selecionar os N primeiros candidatos do conjunto $C'(a_P)$ gerados pelos dois primeiros métodos. Sendo assim, ao utilizar $L = 1000$ e $N = 1000$, trabalhamos com o mesmo número de candidatos tanto para propriedades de links entre artigos e links entre categorias quanto para a similaridade de conteúdo dos artigos.

Tabela 4.3: Presença dos artigos correspondentes entre os N primeiros candidatos

	N=1000		**N=2000**		**N=5000**		**N=10000**	
	CLH +CL	CLH +CL +ACI	CLH +CL	CLH +CL +ACI	CLH +CL	CLH +CL +ACI	CLH +CL	CLH +CL +ACI
Animais	67%	92%	73%	96%	81%	96%	82%	96%
Internet	48%	88%	51%	89%	56%	90%	60%	91%
Automobilismo	15%	90%	16%	92%	18%	92%	18%	92%
Moda	45%	85%	50%	85%	63%	85%	66%	85%
Filmes	64%	93%	73%	95%	78%	96%	83%	96%
Atores	68%	90%	74%	93%	86%	97%	93%	97%
Biologia	62%	86%	67%	89%	74%	90%	81%	91%
Matemática	32%	62%	36%	65%	42%	65%	47%	65%
Física	48%	66%	54%	70%	60%	71%	63%	71%
Computadores	66%	91%	75%	93%	84%	93%	88%	93%
Média	51,5%	84,3%	56,9%	86,7%	64,2%	87,5%	68,1%	87,7%
Aumento	**+63%**		**+52%**		**+36%**		**+28%**	

Através dos resultados da Tabela 4.3, percebemos que o uso do *ArticleContentIndex* aumenta muito a ocorrência do artigo correspondente dentro do conjunto de candidatos. Esse aumento é ainda mais perceptível quando trabalhamos com um conjunto pequeno ($N = 1000$) de candidatos, pois nessa situação, o *ArticleContentIndex* consegue cobrir um maior número de artigos correspondentes que não estavam presentes no conjunto de candidatos.

Além disso, ao analisar os resultados para cada um dos limiares N (1000, 2000, 5000, e 10000), verificamos que com o uso do *ArticleContentIndex*, a eficiência geral do método aumenta muito pouco à medida que vamos aumentando o limiar N. Para N = 1000 temos uma média de 84,3% de presença do artigo correspondente enquanto que para N = 10000 esse número sobe para apenas 87,7%, um aumento muito pequeno se considerarmos que estamos aumentando em 10 vezes o tamanho do conjunto de candidatos.

Sendo assim, ao verificar o resultado da combinação dos três métodos (*Chain Link Hypothesis*, *CategoryLink* e *ArticleContentIndex*), escolhemos trabalhar com o limiar $N = 1000$. Tal decisão foi tomada em virtude do seu excelente custo-benefício, pois a aplicação desse limiar, acrescidos dos 1000 artigos do método *ArticleContentIndex* ($L = 1000$), resulta em um conjunto final de candidatos composto por 2000 artigos. Se compararmos esse conjunto final de candidatos (2000 artigos) com o tamanho médio dos conjuntos de candidatos gerados (150 mil artigos), teremos uma redução, em média, do número de artigos de 75 vezes. Já a presença do artigo correspondente, apesar da re-

dução do conjunto de candidatos, é reduzida em cerca de dois pontos percentuais. Devido à redução expressiva no esforço computacional ao processar 2000 candidatos ($L = 1000$ + $N = 1000$) ao invés de 150 mil, tal perda nos parece aceitável em vista do benefício trazido pela diminuição do conjunto.

Ao trabalharmos com o limiar $N = 1000$, o *baseline*, que utiliza somente o *Chain Link Hypothesis*, apresenta, conforme a Tabela 4.2, um resultado de 37,7% da presença do artigo correspondente no conjunto de candidatos. Já a nossa abordagem, que combina o *Chain Link Hypothesis*, o *CategoryLink* e o *ArticleContentIndex*, apresenta, conforme a Tabela 4.3, um resultado de 84,3%. Sendo assim, a nossa abordagem supera o *baseline* em 123,3% a mais de presença do artigo correspondente dentro do conjunto de candidatos.

Tendo em vista os resultados da Tabela 4.3, concluímos que os métodos empregados apresentaram excelente desempenho, mesmo com a limitação imposta ao conjunto. Isso nos possibilitou ter um conjunto confiável da dados e viável em termos de tamanho para a aplicação dos atributos de descoberta de CLLs ausentes.

4.3 Resultados para a Descoberta de Cross-language Links Ausentes

Após a seleção do conjunto de candidatos, fazemos o cálculo dos indicadores de similaridade para cada artigo a_P do conjunto $DS1000$ com seus candidatos c_I em inglês pertencentes ao conjunto $C'(a_P)$. Conforme descrito na Seção 3.3, temos quatro atributos que visam quantificar a similaridade entre um artigo e seus candidatos: `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`. Os coeficientes de similaridade gerados serão utilizados para treinar o classificador e os resultados, bem como a comparação da nossa abordagem com o *baseline* do SORG; CIMIANO (2008a), serão apresentados na sequência.

4.3.1 Geração do Conjunto WPMAIN

Com o objetivo de organizar os artigos para serem analisados pelo classificador, foi gerado o conjunto chamado *WPMAIN*. Tal conjunto possui os códigos dos artigos envolvidos (*page_id*), o escore de cada um dos quatros atributos citados acima e o atributo de classe que indica se cada par em questão é ou não correspondente.

Sendo assim, os quatro atributos foram computados para cada par de artigos $\langle a_P, c_I \rangle$ | $a_P \in DS1000$ e $c_I \in C'(a_P)$ (restrito a $N = 1000$).Dado que a definição de *instância* para o conjunto *WPMAIN* consiste em entradas formadas por um par de artigos seguidos pelos seus escores de similaridade e atributo de classe, o conjunto *WPMAIN*, portanto, possui 2 milhões de instâncias. Tais entradas são resultantes da comparação de todos os 1000 artigos $a_P \in DS1000$ contra os seus 2000 candidatos $c_I \in C'(a_P)$.

Utilizando o limiar escolhido ($N = 1000$) para o número de candidatos de cada artigo, podemos observar na Tabela 4.3 que 84,3% dos artigos do $DS1000$ possuem o seu artigo correspondente dentro do seu conjunto de candidatos. Isso significa que 843 artigos $a_P \in DS1000$ fazem parte de instâncias dentro do *WPMAIN* formadas pelo par $\langle a_P, c_I \rangle$ | $c_I \xrightarrow{CLL} a_P$'. Já os outros 157 artigos remanescentes do $DS1000$, não possuem artigo candidato que seja o seu correspondente. Dessa forma, não existe registro dentro do *WPMAIN* onde os mesmos façam parte de pares $\langle a_P, c_I \rangle$ para os quais existe $c_I \xrightarrow{CLL} a_P$.

4.3.2 Medidas de Avaliação: Precisão, Revocação e Medida-F

Utilizando o conjunto *WPMAIN*, é possível saber quando uma instância se refere a um exemplo positivo (i.e. artigos correspondentes) ou negativo (i.e. artigos não correspondentes) devido ao atributo de classe associado a cada par de artigos. Dessa forma, é possível realizar o cálculo das medidas de *precisão* (Eq. 4.1), *revocação* (Eq. 4.2), e Medida-F (Eq. 4.3). Nesse ponto, destacamos uma das vantagem de trabalharmos com um conjunto cujo resultado da correspondência entre artigos é conhecida. Tal característica possibilitou a análise automática de 2 milhões de pares de artigos, número que seria impossível de verificar manualmente.

$$Precisao(P) = \frac{\#VerdadeiroPositivo}{\#VerdadeiroPositivo + \#FalsoPositivo} \tag{4.1}$$

$$Revocacao(R) = \frac{\#VerdadeiroPositivo}{\#VerdadeiroPositivo + \#FalsoNegativo} \tag{4.2}$$

$$Medida - F = \frac{2 \times P \times R}{P + R} \tag{4.3}$$

onde Verdadeiro Positivo refere-se a artigos correspondentes que foram identificados como tal; Falsos Positivos são artigos que não são correspondentes mas que foram considerados como tal; e Falso Negativo são artigos correspondentes que não foram identificados como tal.

4.3.3 Submissão de Conjuntos ao Classificador

A partir do conjunto *WPMAIN*, três conjuntos foram selecionados: *TrainingSet*, *TestSet* e *LargeTestSet*. O primeiro, denominado de *TrainingSet*, possui 842 instâncias sendo metade com exemplos positivos e a outra metade com exemplos negativos. Esse conjunto foi gerado a partir das primeiras 421 instâncias positivas do conjunto *WPMAIN* da seguinte forma: para cada instância positiva, a próxima instância negativa também era selecionada. A estratégia utilizada objetiva gerar um conjunto balanceado em termos de exemplos positivos e negativos. Dessa forma, o *TrainingSet* foi utilizado para treinar o classificador e gerar uma árvore de decisão usando o algoritmo J48 que será validada com os próximos conjuntos.

O segundo conjunto, denominado de *TestSet*, possui 844 instâncias (50% de exemplos positivos e 50% de exemplos negativos). Ele foi gerado a partir das últimas 422 instâncias positivas do conjunto WPMAIN e da mesma forma que o conjunto acima, para cada instância positiva que era coletada, era também selecionada a próxima instância negativa. O *TestSet* foi empregado para testar o modelo de classificação gerado a partir do conjunto de treinamento. Dessa forma, ele também é um conjunto balanceado, porém com instâncias diferentes justamente para testar a árvore de decisão. Destacamos aqui que a soma dos exemplos positivos do *TrainingSet* com os do *TestSet* resultam nas 843 instâncias positivas que temos no conjunto geral.

O último conjunto, denominado de *LargeTestSet*, possui um total de 25.987 instâncias: 422 positivas e 25.565 negativas. As instâncias positivas são as mesmas do *TestSet*, no entanto, as instâncias negativas foram selecionadas de maneira aleatória dentro do conjunto *WPMAIN* sem repetir as instâncias negativas do *TrainingSet*. Seu objetivo também é validar o modelo de classificação, porém, conta com um maior número de instâncias e não é balanceado. Essa quantidade significativamente maior tem o objetivo justamente de validar as várias combinações de coeficientes de atributos que podem ocorrer nas comparações entre os pares de artigos.

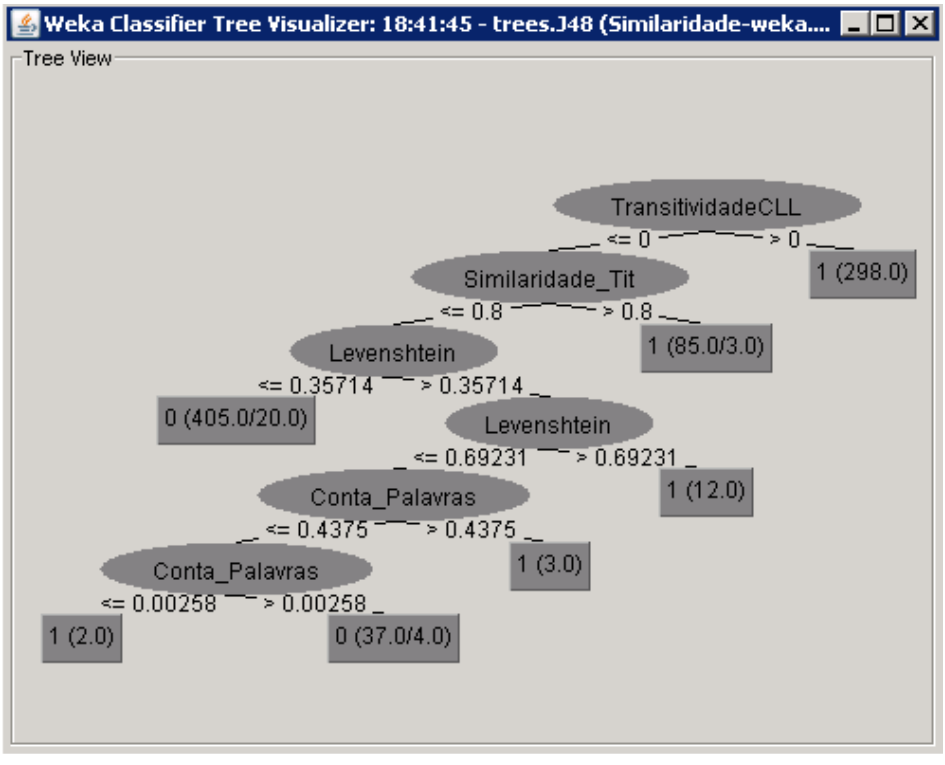
Para classificar as instâncias, utilizamos o algoritmo J48 provido pelo WEKA (HALL et al., 2009) que cria uma árvore de decisão. No entanto, para podermos identificar e fazer uma análise manual das instâncias que eram classificadas de forma errada, resolvemos usar a *API* do WEKA disponível para a linguagem Java. O pacote contendo as classes e métodos dessa *API* foram fornecidos por USE WEKA IN YOUR JAVA CODE (2014).

Escolhemos utilizar o modelo de árvore de decisão devido ao fato do conhecimento ser representado por meio de regras *se-então*. Essas regras expressam a linguagem natural facilitando o entendimento do porque um par de artigos será considerado equivalente ou não. Através da aplicação desse conjunto de regras, a classe atribuída assume somente valores discretos: par em questão é ou não é correspondente. O resultado da nossa abordagem é a sugestão categórica de pares de artigos que são ou não correspondentes, ou seja, a árvore de decisão se adapta a nossa proposta ao atribuir valores discretos para o resultado.

Em uma árvore de decisão a classificação de um caso se inicia pela raiz da árvore, e esta árvore é percorrida até que se chegue a uma folha. Em cada nó de decisão será feito um teste que irá direcionar o caso para uma sub-árvore. Este processo irá guiar-se até que uma folha seja alcançada. O atributo de classe para o par de artigos a ser analisado que indica se são ou não correspondentes está armazenada nesta folha.

Na Figura 4.3, podemos observar a árvore de decisão gerada pelo algoritmo J48 provido pelo WEKA. Esta árvore foi construída a partir da submissão do conjunto de candidatos *TrainingSet*. O modelo de decisão descrito na imagem será utilizado para avaliar a correspondências dos pares de artigos dos conjuntos *TestSet* e *TrainingSet*. Para cada par de artigo, os atributos implementados serão verificado nos nós de decisão da árvore. A escolha para qual sub-árvore o modelo deverá seguir será de acordo com decisão devido à pontuação atribuída pelos atributos de similaridade. Cada um desses atributos será analisado de forma individual na Seção 4.3.6.

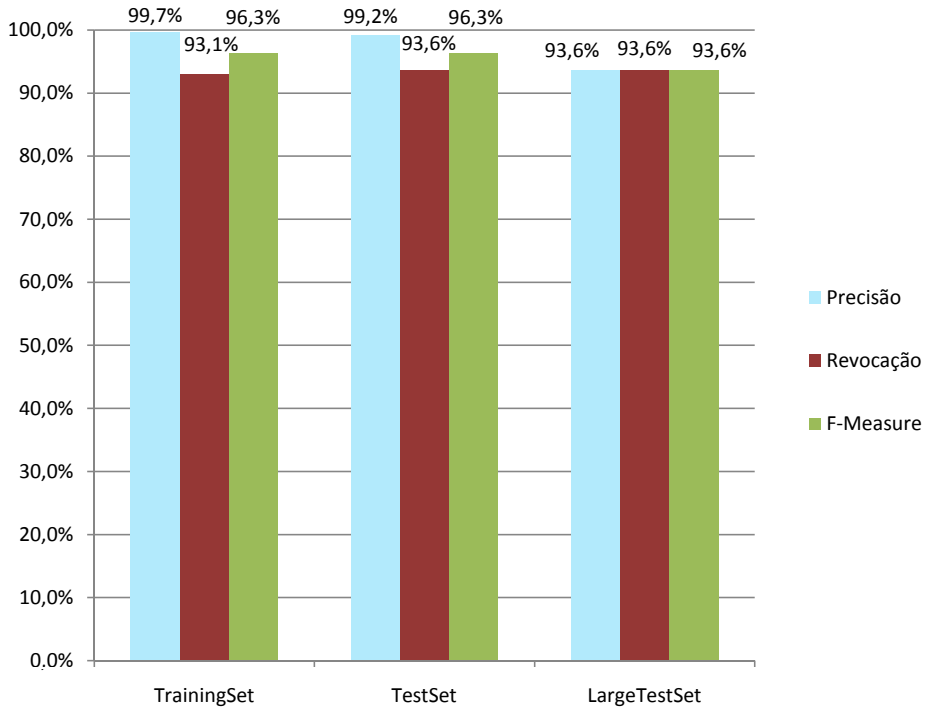
Figura 4.3: Árvore de Decisão gerada pelo J48 a partir do conjunto *TrainingSet*



4.3.4 Descoberta de Cross-langauge Links Ausentes

O resultado para a identificação de CLLs é apresentado na Figura 4.4. Conforme citado anteriormente, no nosso experimento utilizamos a ferramenta WEKA (HALL et al., 2009) para criar um modelo de decisão utilizando o algoritmo J48.

Figura 4.4: Valores para Precisão, Revocação e Medida-F do CLLFinder



Na Figura 4.4, o primeiro conjunto de gráfico de barras à esquerda mostra o desempenho do modelo de classificação gerado a partir do conjunto *TrainingSet*. Nessa situação, o modelo gerado é aplicado aos próprios dados utilizados para gerá-lo. Ainda que tais dados não validem o modelo, visto que seria um teste "viciado", percebe-se um ótimo desempenho da solução desenvolvida com 99,7% de precisão (apenas 1 falso positivo), 93,1% de revocação (29 falsos negativos) e 96,3% de Medida-F. Sendo assim, das 842 instâncias do *TrainingSet*, tivemos 1 falso positivo, 29 falsos negativos e 812 instâncias classificadas corretamente.

Já o conjunto de barras do meio mostra os resultados quando aplicamos o modelo gerado ao conjunto *TestSet*. Podemos verificar que o desempenho praticamente se manteve constante com escores de 99,2% de precisão (3 falsos positivos), 93,6% de revocação (27 falsos negativos) e 96,3% de Medida-F. Das 844 instâncias do *TestSet*, tivemos 3 falsos positivos, 27 falsos negativos e 814 instâncias classificadas corretamente.

O último conjunto do gráfico de barras refere-se à aplicação do *LargeTestSet* ao modelo gerado. Aqui, tivemos uma redução principalmente na precisão do CLLFinder com índice de 93,6% (27 falsos positivos). Já a revocação se manteve constante com valor de 93,6%. Ao analisar os falsos negativos, verificamos que foram as mesmas 27 instâncias do *TestSet* que não foram atribuídas como correspondente pelo CLLFinder. Das 21.522 instâncias do *LargeTestSet*, tivemos 27 falsos positivos, 27 falsos negativos e 21.468 instâncias classificadas corretamente.

Apesar da queda em termos de precisão para o último conjunto, estes números são bastante positivos. Adicionalmente, foram realizados experimentos com a opção *10-fold*

cross-validation que demonstraram resultados semelhantes.

4.3.5 Comparação com o trabalho *baseline* de SORG; CIMIANO (2008a)

Com o objetivo de validar a nossa abordagem, realizamos a comparação da mesma com a desenvolvida pelo SORG; CIMIANO (2008a). Para facilitar, o mesmo será referenciado a frente como *S&C Baseline*.

Da mesma forma que foi feita a seleção de candidatos pela *Chain Link Hypothesis*, foi necessário codificar os atributos de cálculo de similaridade utilizados por SORG; CIMIANO (2008a). Sendo assim, para um dado par de artigos $\langle a_P, c_I \rangle$ | $a_P \in DS1000$ e $c_I \in C'(a_P)$, os sete atributos abaixo do *S&C Baseline* exploram as propriedades de similaridade entre artigos da seguinte forma:

- `Chain Link Count`: dado que um *Chain Link* entre a_P e c_I é definido por $a_P \xrightarrow{pl} x_P \xrightarrow{CLL} x_I \xleftarrow{pl} c_I$, esse algoritmo realiza a contagem de *Chain Link* para o par $\langle a_P, c_I \rangle$.

- `Normalized Chain Link Count`: esse algoritmo normaliza os valores do `Chain Link Count` pelo limiar utilizado para a restrição do conjunto de candidatos realizada por SORG; CIMIANO (2008a). Visto que não tivemos acesso a tal número, porém, implementamos o *Chain Link Hypothesis*, utilizamos o nosso limiar médio de corte para $N = 1000$ que era de $2,52$ *Chain Links*.

- `Chain Link Inlink Intervals`: dado que para um *Chain Link* definido por $a_P \xrightarrow{pl} x_P \xrightarrow{CLL} x_I \xleftarrow{pl} c_I$ os artigos intermediários x_P e x_I são chamados de *Chain Link Intermediate Articles* (CLIA), e que INLINKS (x) = |{$y \in WP_\alpha$ | $y_\alpha \xrightarrow{pl} x_\alpha$}| (i.e. INLINKS($x$) = número de *pagelinks* para x), a motivação desse algoritmo é que quanto menor o número de INLINKS dos artigos que compõem o CLIA de um par de artigos $\langle a_P, c_I \rangle$, mais específico será o *Chain Link* entre a_P e c_I e maior a chance de a_P e c_I serem correspondentes.

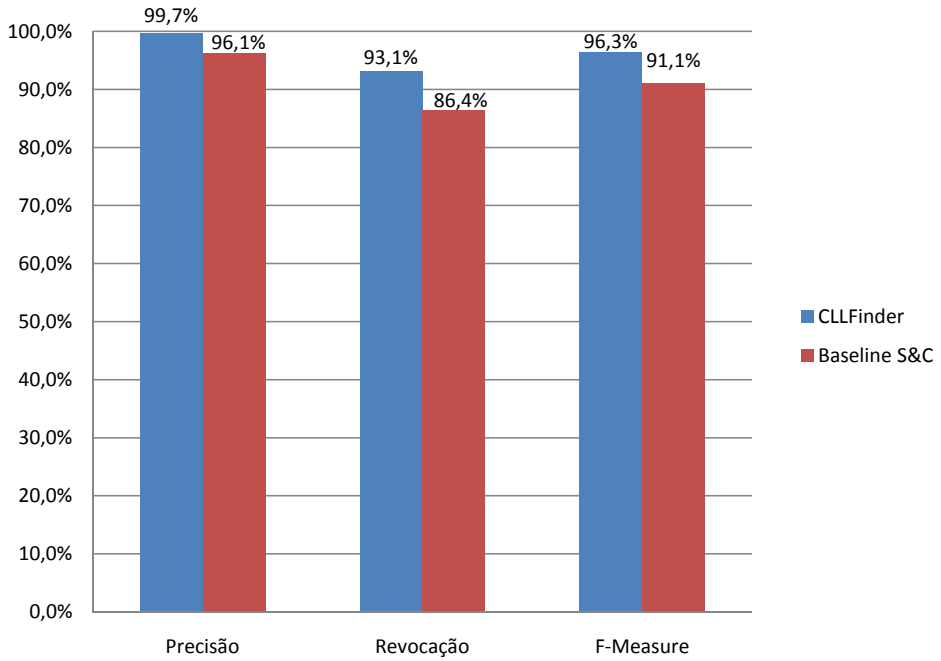
- `Common Categories`: para o par de artigos $\langle a_P, c_I \rangle$, esse algoritmo conta o número de categorias em comum entre a_P e c_I de forma que categorias em diferentes idiomas são ditas ser em comum quando estão relacionadas através de CLLs.

- `CLIA Graph`: dado um par de artigos $\langle a_P, c_I \rangle$ e todos os possíveis *Chain Links* $a_P \xrightarrow{pl} x_P \xrightarrow{CLL} x_I \xleftarrow{pl} c_I$ entre a_P e c_I, esse algoritmo avalia a similaridade entre dois grafos G_P e G_I. O primeiro é gerado de forma que seus vértices sejam todos os possíveis artigos x_P e suas arestas os *pagelinks* entre tais artigos. Da mesma forma, G_I é gerado de forma equivalente, porém, referente aos possíveis artigos x_I e seus *pagelinks*. A análise de similaridade se da a partir do número comum de vértices e arestas de G_P e G_I que é calculada baseada na existência ou não de CLLs para esses elementos.

- `Distância de Edição`: esse atributo realiza a comparação de títulos através do algoritmo *Distância de Levenshtein*.

- `Termos em Comum`: contagem do número de palavras em comum entre os textos de um par de artigos sem realizar a tradução das mesmas.

Figura 4.5: Comparando o CLLFinder com o *S&C Baseline*



Após a implementação dos algoritmos, a nossa preocupação foi em utilizar a mesma forma de classificação que SORG; CIMIANO (2008a) descreveu em seu trabalho. Verificamos que ao invés de utilizar o algoritmo J48 como no CLLFinder, o *baseline* utilizou o SVMlight (JOACHIMS, 1999).

O SVMlight é uma implementação na linguagem de programação C do *Support Vector Machines* (SVM). De acordo com CORTES; VAPNIK (1995), o SVM é definido como um conjunto de métodos de aprendizado supervisionado que analisam os dados e fazem o reconhecimento de padrões. Tais algoritmos têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros. Fundamentalmente, o SVM toma como entrada um conjunto de dados e prediz, para cada entrada dada, para quais das duas possíveis classes, positiva ou negativa, a entrada faz parte.

Para gerar o modelo, o conjunto do CLLFinder (*TrainingSet*) foi utilizando. Uma vez criado o modelo, a comparação entre o CLLFinder e a abordagem do SORG; CIMIANO (2008a) foi realizada com o aplicação do conjunto *TestSet* para ambos os modelos. Os resultados estão na Figura 4.5.

Podemos verificar que o CLLFinder alcançou resultados superiores ao *S&C Baseline* em termos de precisão, revocação e Medida-F. A precisão, ainda que superior no CLLFinder, foi bastante semelhante em ambas as abordagens. A maior diferença entre os trabalhos ficou por conta da revocação onde o CLLFinder obteve um resultado 7% pontos percentuais melhor que o *S&C Baseline*. Por consequência, a Medida-F também foi superior com valores de 96,3% contra 91,1%. Abaixo, listamos alguns motivos que fazem com que a nossa proposta tenha conseguido melhores resultados em relação ao *baseline*:

(i) O atributo `Common Categories`, utilizado pelo *S&C Baseline*, não é um bom discriminador pois para um dado artigo e seu candidato que não é o correspondente, mas é similar, podem existir um grande numero de categorias em comum.

(ii) O algoritmo *Chain Link Count* usado no *S&C Baseline* também não é uma evidência muito confiável visto que artigos não correspondentes também podem estar conecta-

dos várias vezes através de um *Chain Link*.

(iii) Com o objetivo de comparar a similaridade entre títulos, o *S&C Baseline* utiliza apenas o `Distância de Edição` sem realizar a tradução dos títulos. O CLLFinder também possui tal método bem como o `Similaridade de Títulos` que realiza a comparação após a tradução do título. O que percebemos é que muitos títulos de artigos correspondentes tornam-se praticamente iguais após a tradução. Conforme colocamos na descrição da distância de edição (3.3.3), o mesmo é util quando o artigo diz respeito a entidades ou objetos universais. Acreditamos, porém, ser necessário complementar a análise de títulos com técnicas que realizam a tradução automática do mesmo.

Além dos itens citados acima, o atributo `CLLTransitivity` foi um dos responsáveis pelos resultados superiores do CLLFinder. A existência de um caminho de CLLs do artigo de origem para o artigo de destino através de uma língua intermediária é uma forte evidência de que tratam-se de artigos correspondentes. Com isso, tal atributo apresentou altos índices de precisão conforme podemos observar na análise individual dos algoritmos de similaridade (Seção 4.3.6).

Percebemos que os resultados obtidos pela nossa implementação do *S&C Baseline* foram levemente superiores em termos de precisão (+3%) e razoavelmente superiores em termos de revocação (+16%) aos resultados reportados pelo SORG; CIMIANO (2008a). Acreditamos que isso ocorreu devido ao uso dos diferentes idiomas de origem e destino (utilizamos português e inglês enquanto SORG; CIMIANO (2008a) utilizou alemão e inglês) e, por consequência, diferentes conjuntos de artigos.

4.3.6 Contribuição de cada Atributo de Similaridade

Nessa etapa, foi verificado a contribuição de cada um dos atributos de similaridade entre artigos que compõem o CLLFinder. Essa análise foi realizada de duas formas:

(i) gerando o modelo de classificação quatro vezes, sendo que, em cada execução, um dos atributos era retirado, e

(ii) gerando o modelo de classificação novamente por quatro vezes, porém, em cada vez apenas um dos quatro atributos era utilizado.

As Tabelas 4.4 e 4.5 mostram os resultados em termos de precisão, revocação e Medida-F para, respectivamente, os testes *(i)* e *(ii)*. Em ambos os casos, o conjunto *TrainingSet* foi utilizado para treinar o classificador e gerar a árvore de decisão, enquanto o *TestSet* foi empregado para testar o modelo de classificação gerado pelo conjunto anterior. Abaixo, segue uma análise dos aspectos positivos, limitações e contribuições dos quatro atributos utilizados:

`CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`

- `CLLTransitivity`: Analisando a Tabela 4.4, percebemos que a execução em que retiramos o `CLLTransitivity` é a que ocorre a maior queda dos valores de precisão, revocação e Medida-F, quando comparamos à abordagem com os quatro atributos. Já na análise do desempenho individual (Tabela 4.5), verificamos que o `CLLTransitivity` alcançou uma precisão de 100%, número superior, inclusive, ao uso dos quatro atributos juntos. Tal resultado nos mostra que quando esse atributo prediz que um par de artigos é equivalente, temos a certeza de que realmente tratam-se de artigos correspondentes. No entanto, a revocação de 70,7% significa

que tal atributo não é capaz de predizer todos os pares de artigos correspondentes existentes no nosso conjunto com eficiência semelhante a da precisão. Por fim, ao combinarmos os quatro atributos e compararmos com a execução somente do `CLLTransitivity`, nota-se uma pequena perda de precisão (0,8 pontos percentuais) que é vantajosamente compensada com um ganho de 22,9 pontos percentuais de revocação. A principal limitação desse atributo é que logo que um artigo é publicado na Wikipédia, se ele não possuir CLLs que realizem a transitividade, o `CLLTransitivity` não será capaz de encontrar o artigo correspondente. Sendo assim, esse atributo exige a existência de outros CLLs para poder aplicar a transitividade e identificar o artigo equivalente.

- `Similaridade de Títulos`: Verificamos que a retirada da `Similaridade de Títulos` implica na execução com perdas significativas de precisão, revocação e Medida-F, ficando atrás somente do CLLTransitivity. Além disso, análise individual nos indica que o `Similaridade de Títulos` possui a segunda colocação tanto em precisão quanto em revocação, ficando atrás, respectivamente, do CLLTransitivity e do `Distância de Edição`. Por ter atingido índices equilibrados de precisão e revocação, o `Similaridade de Títulos` é o atributo que possui o maior valor de Medida-F, contribuindo efetivamente para a nossa abordagem. A boa precisão (88,4%) e revocação (81,7%) alcançadas devem-se ao êxito da aplicação desse atributo e ao fato de que muitos artigos correspondentes efetivamente possuem títulos muito similares ou idênticos quando comparados em um mesmo idioma. A principal limitação desse atributo é que ele confia no sucesso da tradução realizada. Caso o mecanismo de tradução falhe em traduzir para a palavra correta no idioma de destino, o escore de similaridade atribuído será penalizado.

- `Distância de Edição`: Na execução do CLLFinder sem o atributo `Distância de Edição` nota-se um ganho de precisão de 0,8% que faz com que ela alcance o valor de 100% nessa execução. No entanto, sobre a revocação, verificamos uma perda de 1,2 pontos percentuais, acarretando, também, em uma menor Medida-F. Na análise individual, verificamos número razoáveis de precisão (80,8%), revocação (77,1%) e Medida-F (78,9%). Ainda que a contribuição da `Distância de Edição` seja a menor entre os quatro atributos, a sua inclusão na nossa abordagem foi mantida visto que a mesma agrega valor ao CLLFinder. Provavelmente, em conjuntos de artigos onde tivesse uma maior representatividade de pessoas, lugares ou entidades, a retirada desse atributo significaria um impacto maior que o observado nos nossos experimentos. A principal limitação desse atributo é que como não é realizada a tradução, não será atribuído uma pontuação alta

Tabela 4.4: Precisão, Revocação e Medida-F removendo cada atributo

	Precisão	**Revocação**	**Medida-F**
CLLFinder (Todos algoritmos)	**99,2%**	**93,6%**	**96,3%**
Removendo o `CLLTransitivity`	92,4%	86,4%	89,3%
Removendo o `Similaridade de Títulos`	97,9%	89,8%	93,6%
Removendo o `Distância de Edição`	100%	92,4%	96%
Removendo o `Termos em Comum`	98,5%	93,3%	95,8%

Tabela 4.5: Precisão, Revocação e Medida-F para cada um dos atributos analisados de forma individual

	Precisão	**Revocação**	**Medida-F**
CLLFinder-`CLLTransitivity`	100%	70,7%	82,8%
CLLFinder-`Similaridade de Títulos`	88,4%	81,7%	84,9%
CLLFinder-`Distância de Edição`	80,8%	77,1%	78,9%
CLLFinder-`Termos em Comum`	65,4%	83,1%	73,2%

para títulos que são idênticos quando comparados em um mesmo idioma.

- `Termos em Comum`: A retirada do atributo `Termos em Comum` acarreta em uma pequena perda de precisão (0,7%), revocação (0,3%) e Medida-F (0,5%). Ao analisar o seu desempenho individualmente, verificamos que ele é o atributo que possui tanto a menor precisão (65,4%), quanto a maior revocação dentre os quatro, alcançando um índice de 83,1%. Dessa forma, ainda que possua uma baixa precisão, consideramos o `Termos em Comum` fundamental para garantir uma boa revocação do CLLFinder. Tal desempenho tende a aumentar principalmente para idiomas que são morfologicamente similares ou compartilham nomes próprios, número, datas e nomes de localizações, pois, em geral, trata-se de palavras iguais independente do idioma. A principal limitação desse atributo é que os textos podem variar muito de tamanho quando se compara artigos correspondentes em idiomas distintos. Isso pode resultar em algumas diferenças no compartilhamento de palavras penalizando o escore atribuído pelo `Termos em Comum`.

Através dos resultados apresentados nas Tabelas 4.4 e 4.5, concluímos que o atributo *CLLTransitivity* é o que mais contribui com a nossa abordagem. É possível fazer tal afirmação visto que a retirada do mesmo representa a maior perda não só de precisão, mas também da Medida-F. Esse último, representa a média harmônica que combina precisão e revocação em um único número, permitindo visualizar que a retirada do *CLLTransitivity* causa o maior impacto no CLLFinder. É importante destacar também que a medida de precisão é a qualidade mais desejada em abordagens que realizam a identificação de CLLs ausentes, pois criar links errados entre artigos que não são correspondentes é pior do que a não existência de CLLs. Tal fato reforça a importância do *CLLTransitivity* visto que o mesmo alcançou 100% de precisão.

Os bons resultados do CLLTransitivity confirmam a hipótese de que um autor, ao editar um artigo, pode ter colocado CLL somente para os idiomas que ele conseguiu identificar. Nesse caso, fica evidente que o *CLLTransitivity* conseguiu identificar os CLLs ausentes através da transitividade entre idiomas. Entretanto, esse atributo baseia-se na existência de um caminho de CLLs entre o artigo de origem e o de destino, e, em alguns casos, esse caminho não existe. Essa é a situação representada na Figura 1.1(b), onde não existe a transitividade de CLLs entre os artigos correspondentes. Nesse mesmo exemplo, é o atributo `Similaridade de Títulos` que permitiu encontrar o artigo correspondente, pois os títulos ficaram idênticos após a tradução do título em português para o inglês.

Quando um artigo não possui CLLs para serem explorados pelo *CLLTransitivity*, os outros atributos tornam-se fundamentais para identificar as evidências de similaridade. A combinação dos quatro atributos, portanto, é necessária para garantir que a nossa abordagem seja robusta para lidar com os casos em que não existam CLLs. Sendo assim,

ressaltamos que todos os atributos contribuem para o CLLFinder, ajudando a alcançar excelentes níveis de precisão, revocação e Medida-F.

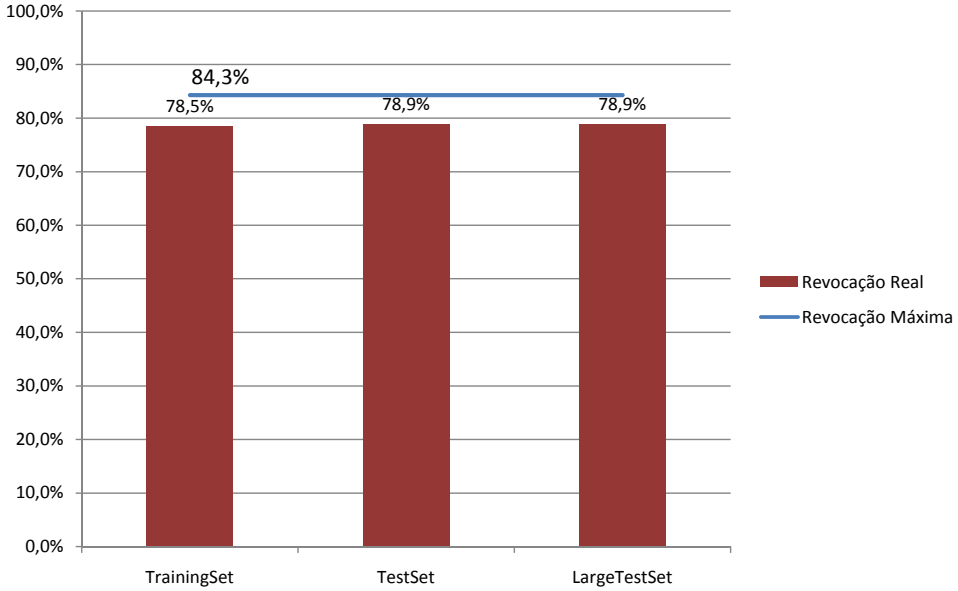
4.3.7 Revocação Geral

Até aqui, os resultados de revocação apresentados para a descoberta de CLLs ausentes foram sobre os conjuntos *TrainingSet*, *TestSet* e *LargeTestSet*. É importante destacar que tais conjuntos foram constituídos a partir dos 843 artigos $a_P \in DS1000$, do total dos 1000 artigos, que possuíam o artigo correspondente dentro do seu conjunto de candidatos. Isso quer dizer que dos 1000 artigos iniciais o módulo de seleção de candidatos conseguiu um percentual de 84,3% de presença do artigo correspondente. Já para 15,7% dos artigos, não se conseguiu selecionar o artigo correspondente.

Sendo assim, a revocação geral dos atributos de similaridade do CLLFinder implementados (`CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`) está limitada a 84,3%. Torna-se necessário, portanto, verificar qual foi a revocação geral obtida pelo CLLFinder levando-se em conta que para 15,7% dos artigos o nossa abordagem não foi capaz de trazer o correspondente.

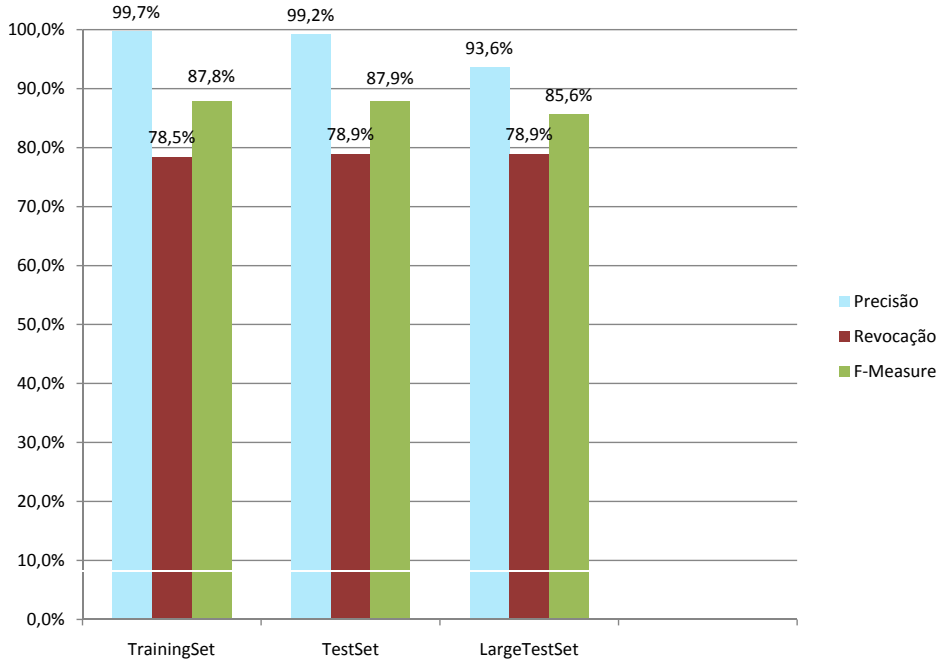
Conforme podemos observar na Figura 4.6, pegamos o mesmo percentual de 15,7% dos artigos sem candidato correspondente dentro do $DS1000$ e descontamos da revocação apresentada na Figura 4.4. O cálculo é o mesmo que supor que para os conjuntos *TrainingSet*, *TestSet* e *LargeTestSet* existiam 15,7% a mais de artigos sem seu correspondente. Dessa forma, respectivamente para os conjuntos acima, temos os índices de revocação geral de 78,5%, 78,9% e 78,9%.

Figura 4.6: Revocação Máxima limitada pelo Conjunto de Candidatos e Revocação Geral do CLLFinder



Com os valores gerais de revocação, podemos verificar na Figura 4.7 os valores finais de precisão, revocação e Medida-F do CLLFinder. A precisão não muda visto que o seu cálculo é o número de pares de artigos correspondentes recuperados sobre o número de pares de artigos recuperados, ou seja, não leva em consideração o número total de pares de artigos correspondentes. No entanto, como a Medida-F depende da revocação, seus valores diminuem para, respectivamente, de 87,8%, 87,9% e 85,6%.

Figura 4.7: Valores para Precisão, Revocação Geral (limitada pelo conjunto de candidatos) e Medida-F do CLLFinder



4.3.8 Análise de erros de predição do modelo de classificação

Foi feita uma análise sobre os casos em que o nosso classificador cometeu erros ao prever se um par de artigos é ou não correspondente. Abaixo, seguem análises para casos de falsos positivos e falsos negativos.

Casos de falsos positivos ocorreram quando os artigos tinham títulos muito similares e compartilhavam um bom número de palavras no seu texto. Um exemplo é o caso em que o CLLFinder classificou os artigos *MacBook* (em português) and *MacBook Pro* (em inglês) como sendo artigos correspondentes, quando, na verdade, eles não eram. Dos quatro atributos de similaridade, o único que resultou em um escore baixo para esse caso foi o `CLLTransitivity`, entretanto, os outros três atributos obtiveram um escore alto para esse caso visto que os artigos possuíam títulos semelhante e textos com muitas palavras em comum.

Abaixo, seguem exemplos de pares de artigos <português,inglês> erroneamente avaliados como correspondentes pelo CLLFinder:

- *Apple_I* e *Apple_II*: alta similaridade entre títulos e muitos termos em comum no texto;
- *Macintosh_XL* e *Macintosh_LC*: semelhante ao caso acima;
- *Amiga* e *AmigaOS*: aqui o CLLFinder confundiu o equipamento "Amiga"com o sistema operacional do mesmo devido à similaridade entre títulos e o compartilhamento de muitos termos em comum no texto. Em ambos os idiomas, existem na Wikipédia tanto os artigos sobre o "Amiga"quanto sobre o "AmigaOS";
- *Robert Downey Jr* e *Robert Downey, Sr*, devido a similaridade entre títulos, o CLLFinder errou ao dizer que, respectivamente, filho e pai eram correspondentes;

Falsos negativos são, na sua maioria, causados por erros de tradução somados à inexistência de pontuação do *CLLTransitivity*. Como exemplo, o CLLFinder falhou em detectar a correspondência entre *Moda Sustentável* e *Sustainable Fashion*. Os motivos foram que a palavra *moda* foi erroneamente traduzida, o que resultou em um baixo escore para o `Similaridade de Títulos`. Além disso, o `Distância de Edição` e o `Termos em Comum` atribuíram uma baixa pontuação visto que, respectivamente, os títulos não eram semelhantes antes da tradução e os textos não compartilhavam muitas palavras. Por fim, não existia transitividade de CLL para que o `CLLTransitivity` pudesse atribuir um escore capaz de influenciar na decisão do classificador.

Abaixo, seguem exemplos de pares de artigos <português,inglês> correspondentes que não foram avaliados como tal pelo CLLFinder:

- *Invariância* e *Invariant_(physics)*: erro devido à baixa pontuação do *CLLTransitivity* (não existem CLLs para francês, italiano ou espanhol) e a tradução de *Invariância* para *Invariance*;

- *Descritor_específico* e *Specific_name_(zoology)*: erro devido aos títulos distintos, baixo compartilhamento de termos do texto, visto que o texto do artigo *Descritor_específico* possui somente uma linha, e ausência de pontuação do *CLLTransitivity*;

- *Evidências_da_evolução* e *Evidence_of_common_descent*: semelhante ao caso acima, o exemplo não possui CLLs para o francês, italiano ou espanhol e os títulos não são parecidos.

4.4 Aplicação do CLLFinder para Descoberta de novos CLLs

Após o desenvolvimento da abordagem, resolvemos aplicá-la em um cenário real com o objetivo de encontrar novos *Cross-language Links* na Wikipédia. O cenário real implica em trabalhar com um conjunto de artigos do idioma de origem para os quais não existe CLL para o idioma de destino. Nessa situação, o fato de não existir CLL pode ocorrer por duas razões:

(i) o artigo correspondente não existe no idioma de destino, ou

(ii) o artigo correspondente existe, porém o mesmo não está conectado ao artigo origem.

Visto que na aplicação real do CLLFinder não existe atributo de classe, é necessária a avaliação manual de cada par de artigos. Esta avaliação foi feita para os pares que o CLLFinder considerou como correspondentes, para julgar se de fato o são. As seguintes etapas foram realizadas para iniciar o processo de descoberta de CLLs ausentes:

(i) coleta de 1000 artigos aleatórios da Wikipédia em português sem CLL para a Wikipédia em inglês;

(ii) aplicação dos procedimentos para a seleção do conjunto de candidatos (descritos na Seção 3.2);

(iii) avaliação dos pares de artigos através dos atributos de similaridade (descritos na Seção 3.3);

(iv) treino do classificador com o conjunto *TrainingSet* (Seção 4.3.3) para o qual se tem o atributo de classe;

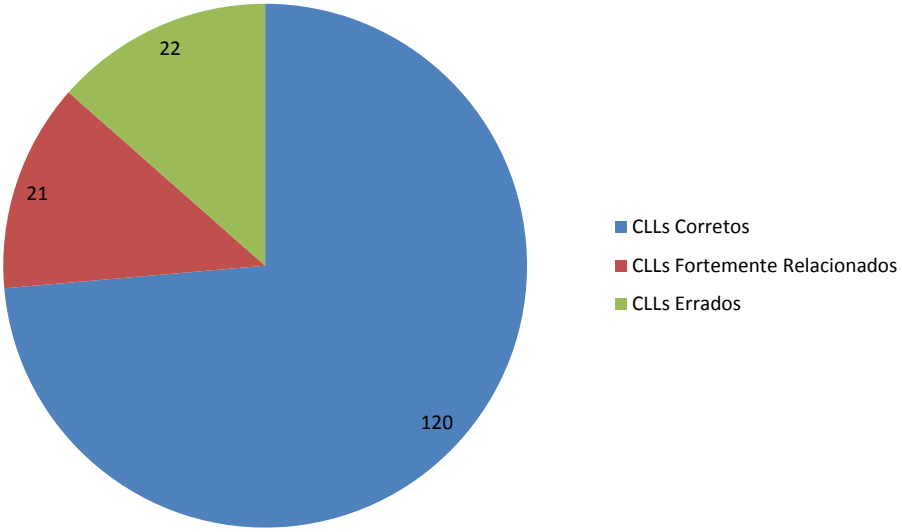
(v) submissão do conjunto de pares de artigos e seus escores de similaridade para o modelo de classificação gerado;

(vi) avaliação manual dos resultados.

Quando processamos os 1000 artigos, obtivemos 233 novos CLLs indicados por uma tupla de artigos <português,inglês>. Desses 233 CLLs, 32 pares continham redirecionamentos para outros artigos. Nesses casos, o artigo original não existia mais e a Wikipédia redirecionava para algum artigo, em geral, semelhante. Isso dificultou a nossa análise manual para verificar a correspondência entre artigos, portanto resolvemos excluir tais artigos da nossa análise.

Dos 201 pares restantes, 38 não existiam mais na Wikipédia e foram excluídos também da nossa análise. Sendo assim, calculamos a precisão para os 163 pares restantes. Para cada um deles, verificamos se eram equivalentes através da interpretação do conteúdo de ambos artigos.

Figura 4.8: Descoberta de novos CLLs a partir do CLLFinder



Através da Figura 4.8, podemos observar a distribuição dos 163 novos CLLs sugeridos pelo CLLFinder entre CLLs corretos, CLLs fortemente relacionados e CLLs errados. Dos 163 novos CLLs sugeridos pelo CLLFinder, 120 eram de fato corretos, 21 possuíam conteúdo fortemente relacionado e 22 estavam errados. Isso significa que o CLLFinder aplicado em um cenário real foi capaz de identificar um *Cross-language Link* correto para o artigo correspondente em 73,6% dos casos. Se levarmos em consideração a indicação de CLLs tanto para artigos correspondentes quanto para artigos fortemente relacionados, o nosso resultado aumenta para 86,5%. Acreditamos, portanto, que os resultados obtidos são plenamente satisfatórios e que o CLLFinder cumpre o seu papel conseguindo, de fato, aumentar a cobertura de links na Wikipédia. Os 163 pares de CLLs sugeridos pelo CLLFinder estão disponíveis no Anexo I desta dissertação.

4.5 Sumário

Neste Capítulo foram relatados os experimentos realizados para a validação da abordagem CLLFinder. Na Seção 4.1 apresentamos a metodologia utilizada bem como os elementos da Wikipédia envolvidos e o conjunto de artigos selecionados para a aplicação do CLLFinder. Já na Seção 4.2 foram apresentados os resultados para a seleção de candidatos, enquanto que na Seção 4.3 são apresentados os resultados para a descoberta de CLLs ausentes. Por fim, na Seção 4.4 o CLLFinder é utilizado em um cenário real para a descoberta de CLLs ausentes.

Os experimentos realizados demonstram que o CLLFinder é uma abordagem viável para a descoberta de CLLs ausentes. Na etapa de seleção de candidatos, o *CategoryLink* e *ArticleContentIndex* foram capazes de reduzir o universo de conjunto de candidatos formado por todos artigos da Wikipédia em inglês (mais de 3.600.000 artigos) para apenas 2000 candidatos. Em 84,3% das vezes o artigo correspondente estava presente no conjunto de candidatos. Dessa forma, a nossa abordagem, que combina o *Chain Link Hypothesis*, o *CategoryLink* e o *ArticleContentIndex*, supera o *baseline*, que utiliza somente o *Chain Link Hypothesis*, em 123,3% a mais de presença do artigo correspondente dentro do conjunto de candidatos.

Já na etapa de descoberta de CLLs ausentes, os atributos `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum` obtiveram 99,2% de precisão, 78,9% de revocação geral e 87,9% de Medida-F, superando o *baseline*. Ao testar o CLLFinder em um cenário real de descoberta de CLLs ausentes, para 86,5% dos casos ele foi capaz de encontrar um CLL entre pares de artigos que são correspondentes ou fortemente relacionados.

5 CONCLUSÃO

Esta dissertação apresenta o Cross-language Link Finder (CLLFinder), uma abordagem que possui um conjunto de técnicas que foram implementadas de forma a realizar a descoberta de CLLs ausentes entre artigos da Wikipédia. A abordagem utilizada para o desenvolvimento e implementação do CLLFinder é composta por três módulos.

O primeiro módulo tem por objetivo selecionar o conjunto de artigos candidatos a serem comparados na segunda etapa com o artigo para o qual se busca o CLL. Para isso, desenvolvemos o *CategoryLink* e o *ArticleContentIndex* que foram combinados com o *Chain Link Hypothesis* para efetuarem a seleção de candidatos. O uso desses métodos foi capaz de reduzir todo o universo de possíveis artigo correspondentes do idioma de destino para um conjunto de candidatos com apenas 2000 artigos. Os experimentos mostraram que para esse conjunto reduzido de candidatos, o artigo correspondente estava presente 84,3% das vezes. Visto que para esse mesmo conjunto de candidatos o trabalho de *baseline*, que utiliza somente o *Chain Link Hypothesis*, apresenta um resultado de 37,7%, a nossa abordagem supera o *baseline* em 123,3% a mais de presença do artigo correspondente dentro do conjunto de candidatos.

Sendo assim, o uso da abordagem proposta para a seleção de candidatos mostrou-se eficiente em selecionar os artigos que, efetivamente, possuíam maior chance de serem os correspondentes. Isso nos permitiu ter um conjunto confiável da dados e viável em termos de tamanho para a aplicação dos atributos de descoberta de CLLs ausentes do segundo módulo.

O segundo módulo é responsável pela análise de atributos que refletem a similaridade entre um par de artigos da Wikipédia. Esse par de artigos a ser avaliado é formado pelo artigo do idioma de origem, para o qual se busca o CLL, combinado com cada um dos seus candidatos selecionados no primeiro módulo. Neste módulo, foram apresentados o seguintes atributos de similaridade: `CLLTransitivity`, `Similaridade de Títulos`, `Distância de Edição` e `Termos em Comum`. Destacamos a alta precisão do atributo `CLLTransitivity` que contribuiu para a qualidade da abordagem proposta.

No último módulo, foi desenvolvido um modelo de classificação utilizando os coeficientes de similaridade calculados no segundo módulo. Os experimentos foram realizados a partir de 1000 artigos em português e 2000 candidatos em inglês selecionados para cada artigo. No total, foram avaliados mais de dois milhões de pares de artigos durante os experimentos. O CLLFinder alcançou valores de 99,2% de precisão e 78,9% de revocação geral, superando o trabalho utilizado como *baseline*.

Adicionalmente, aplicamos o CLLFinder em um cenário real com o objetivo de encontrar novos CLLs na Wikipédia. Para avaliarmos os resultados obtidos, tivemos que realizar a inspeção manual sobre a correspondência entre artigos para sabermos se o CLL-

Finder acertou ou não em indicar a existência de um CLL entre um par de artigos. Como resultado, verificamos que para 86,5% dos casos o CLLFinder foi capaz de indicar CLL entre um par de artigos que são correspondentes ou fortemente relacionados. Para 73,6% dos casos, o CLL descoberto era, de fato, entre artigos equivalentes.

Como produção científica, obtivemos a publicação do seguinte artigo completo no JIDM:

MOREIRA, C. E. M.; MOREIRA, V. P. Finding Missing Cross-Language Links in Wikipédia.**JIDM**, v.4, n.3, p.251 - 265, 2013.

Neste artigo, a etapa de seleção de candidatos utilizou apenas o *CategoryLink*. Ainda assim, apresentamos um ganho de 37% sobre o *baseline*.

Acreditamos que os resultados da abordagem CLLFinder foram muito positivos, pois alcançamos ótimos valores tanto na seleção do conjunto de candidatos quanto na busca de CLLs ausentes sem utilizar nenhuma propriedade específica dos idiomas escolhidos para os experimentos. Além disso, implementamos algoritmos que não necessitam de grande poder computacional.

Carlos: trabalhos futuros

Como trabalhos futuros, são sugeridas as seguintes atividades:

- utilização de outros idiomas intermediários para a aplicação do atributo `CLLTransitivity`: quanto maior o número de idiomas intermediários, maior a chance do autor do artigo de algum desses idiomas ter reconhecido o correspondente no idioma de destino. Tal mudança irá contribuir com uma maior revocação do atributo `CLLTransitivity`.

- inclusão de mais "saltos" entre os idiomas do `CLLTransitivity`: ao invés de utilizar somente um idioma de cada vez para realizar a transitividade, poderíamos utilizar mais idiomas ao mesmo tempo para verificar a existências de um caminho alternativo entre o artigo de origem e o seu correspondente. Dessa forma, vários idiomas iriam compor a transitividade aumentando a eficácia desse atributo de similaridade.

- aumentar o número de *pagelinks* antes de aplicar o *Chain Link Hypothesis*: o *Chain Link Hypothesis* baseia-se na presença de *pagelinks* para formar o caminho chamado *Chain Link*. Se o primeiro passo antes da aplicação desse método para a seleção de candidatos fosse realizar a descoberta de *pagelinks*, melhoraríamos a cobertura do métodos, resultando na presença de um maior número de artigos correspondentes dentro do conjunto de candidatos.

Além dessas atividades, sugere-se a realização de testes no CLLFinder tanto de idiomas morfologicamente similares, como português e espanhol, quanto distintos, como português e japonês. Assim, seria possível verificar o comportamento dos métodos para a seleção de candidatos e dos atributos de similaridade em ambas as situações.

REFERÊNCIAS

ADAFRE, S. F.; RIJKE, M. de. Finding Similar Sentences across Multiple Languages in Wikipedia. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** ACL, 2006. p.62–69.

ADAR, E.; SKINNER, M.; WELD, D. S. Information Arbitrage across Multi-lingual Wikipedia. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, New York, NY, USA. **Proceedings...** ACM, 2009. p.94–103. (WSDM).

BOUMA, G.; DUARTE, S.; ISLAM, Z. Cross-lingual Alignment and Completion of Wikipedia Templates. In: Proceedings of the International Workshop on Cross Lingual Information Access: addressing the information need of multilingual societies, Stroudsburg, PA, USA. **Anais...** ACL, 2009. p.21–29. (CLIAWS3).

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Mach. Learn.**, Hingham, MA, USA, v.20, n.3, p.273–297, Sept. 1995.

ERDMANN, M. et al. Improving the extraction of bilingual terminology from Wikipedia. **ACM Transactions on Multimedia Computing, Communications, and Applications**, New York, NY, USA, v.5, n.4, p.31:1–31:17, Nov. 2009.

GEY, F. C.; KANDO, N.; PETERS, C. Cross-Language Information Retrieval: the way ahead. **Inf. Process. Manage.**, [S.l.], v.41, n.3, p.415–431, 2005.

GREFENSTETTE; GREGORY. **Cross-Language Information Retrieval**. [S.l.]: Springer US, 1998.

HALL, M. et al. The WEKA Data Mining Software: an update. **SIGKDD Explorations Newsletter**, New York, NY, USA, v.11, n.1, p.10–18, Nov. 2009.

JOACHIMS, T. Making Large-scale Support Vector Machine Learning Practical. In: SCHöLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. (Ed.). **Advances in Kernel Methods**. Cambridge, MA, USA: MIT Press, 1999. p.169–184.

LEVENSHTEIN, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. **Soviet Physics Doklady**, [S.l.], v.10, p.707, 1966.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008.

MANUAL of Wikipedia Database Layout. 2011.

MILNE, D.; WITTEN, I. H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: AAAI WORKSHOP ON WIKIPEDIA AND ARTIFICIAL INTELLIGENCE: AN EVOLVING SYNERGY. **Proceedings...** AAAI Press, 2008. p.25–30.

MOREIRA, C. E. M.; MOREIRA, V. P. Finding Missing Cross-Language Links in Wikipedia. **JIDM**, [S.l.], v.4, n.3, p.251–265, 2013.

NASTASE, V.; STRUBE, M. Transforming Wikipedia into a large scale multilingual concept network. **Artif. Intell.**, [S.l.], v.194, p.62–85, 2013.

NGUYEN, T. et al. Multilingual Schema Matching for Wikipedia Infoboxes. **Proceedings of the Very Large Data Base Endowment**, [S.l.], v.5, n.2, p.133–144, Oct. 2011.

NIE, J.-Y. **Cross-Language Information Retrieval**. [S.l.]: Morgan and Claypool Publishers, 2010. (Synthesis Lectures on Human Language Technologies).

OH, J.-H. et al. Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECHNOLOGY - VOLUME 01, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2008. p.322–328.

PENTA, A. et al. Discovering Cross-language Links in Wikipedia through Semantic Relatedness. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Anais...** [S.l.: s.n.], 2012. p.642–647.

PORTER, M. F. An Algorithm for Suffix Stripping. **Program**, [S.l.], v.14, n.3, p.130–137, July 1980.

POTTHAST, M.; STEIN, B.; ANDERKA, M. A Wikipedia-Based Multilingual Retrieval Model. In: **Advances in Information Retrieval**. [S.l.]: Springer Berlin Heidelberg, 2008. p.522–530. (Lecture Notes in Computer Science, v.4956).

RINSER, D.; LANGE, D.; NAUMANN, F. Cross-lingual Entity Matching and Infobox Alignment in Wikipedia. **Information Systems**, [S.l.], v.38, n.6, p.887 – 907, 2013.

SORG, P.; CIMIANO, P. Enriching the Crosslingual Link Structure of Wikipedia - a classification-based approach. In: WORKSHOP ON WIKIPEDIA AND ARTIFICIAL INTELLIGENCE (WIKIAI). **Proceedings...** [S.l.: s.n.], 2008.

SORG, P.; CIMIANO, P. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Working Notes for the Cross-Language Evaluation Forum Workshop (CLEF). **Anais...** [S.l.: s.n.], 2008.

STOKOE, C.; OAKES, M. P.; TAIT, J. Word Sense Disambiguation in Information Retrieval Revisited. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMAION RETRIEVAL, 26., New York, NY, USA. **Proceedings...** ACM, 2003. p.159–166. (SIGIR '03).

USE WEKA in your Java code. 2014.

WIKIPEDIA Extractor. 2013.

WIKIPEDIA. 2014.

ANEXO I

- Branchiostoma Branchiostoma CORRESPONDENTES;
- Accipitriformes Accipitriformes CORRESPONDENTES;
- Cariamiformes Seriema CORRESPONDENTES;
- Autorreplicação Selfreplication CORRESPONDENTES;
- Phaethontiformes Tropicbird CORRESPONDENTES;
- Suliformes Galliformes FORTEMENTE RELACIONADOS;
- Suliformes Piciformes FORTEMENTE RELACIONADOS;
- Aeroplâncton Aeroplankton CORRESPONDENTES;
- Aminoácido Amino_acid CORRESPONDENTES;
- Bioclimatologia Bioclimatology CORRESPONDENTES;
- Bioinformática Bioinformatics CORRESPONDENTES;
- Biologia_da_conservação Conservation_biology CORRESPONDENTES;
- Leporinus_striatus Leporinus_fasciatus CORRESPONDENTES;
- Leporinus_striatus Leptoconchus_striatus FORTEMENTE RELACIONADOS;
- Olhodefogo Fireeye NÃO CORRESPONDENTES;
- Pimelodella_gracilis Farlowella_gracilis FORTEMENTE RELACIONADOS;
- Clitelo Clitellum CORRESPONDENTES;
- Pimelodus_ornatus Pimelodus_pictus FORTEMENTE RELACIONADOS;
- Selar US_Seal NÃO CORRESPONDENTES;
- Stenolicmus_ix Stenolicmus CORRESPONDENTES;
- Dente_de_ovo Egg_tooth CORRESPONDENTES;
- Hemitórax Hemothorax FORTEMENTE RELACIONADOS;

- Mesogleia Mesoglea CORRESPONDENTES;
- Complexo_de_espécies_crípticas Cryptic_species_complex CORRESPONDENTES;
- Código_Internacional_de_Nomenclatura_de_Plantas_Cultivadas International_Code_of_Nomenclature_for_Cultivated_Plants CORRESPONDENTES;
- Ovopositor Oviposition FORTEMENTE RELACIONADOS;
- Espongívoro Spongivore CORRESPONDENTES;
- Ovopositor Ovipositor CORRESPONDENTES;
- Palpo Palpi CORRESPONDENTES;
- Palpo Palp CORRESPONDENTES;
- Tórax Chest CORRESPONDENTES;
- Agregador_de_links Link_aggregation NÃO CORRESPONDENTES;
- Antena_direcional Directional_antenna CORRESPONDENTES;
- AMD_Live! AMD_Live! CORRESPONDENTES;
- Ciberspaço Cyberspace CORRESPONDENTES;
- COMUT Switch NÃO CORRESPONDENTES;
- Apple_IIc Apple_IIc CORRESPONDENTES;
- Conta_Google Google_Account CORRESPONDENTES;
- Steve_Wozniak Steve_Wozniak CORRESPONDENTES;
- Acer_Aspire Acer_Aspire CORRESPONDENTES;
- Enewsletters Newsletter CORRESPONDENTES;
- 48_bits 48bit CORRESPONDENTES;
- Fax_internet Internet_fax CORRESPONDENTES;
- Internet_a_rádio Internet_radio FORTEMENTE RELACIONADOS;
- Internet_Fax Internet_fax CORRESPONDENTES;
- Internet.br Internet NÃO CORRESPONDENTES;
- Podcast Podcast CORRESPONDENTES;
- Apple_Newton Newton_(platform) CORRESPONDENTES;
- Sistema_de_moderação Moderation_system CORRESPONDENTES;
- Upload Upload CORRESPONDENTES;

- Zire_Handheld Zire_Handheld CORRESPONDENTES;
- Chess_Engine_Communication_Protocol Chess_Engine_Communication_Protocol CORRESPONDENTES;
- Yahoo_Video Yahoo!_Video NÃO CORRESPONDENTES;
- ICMP_tunneling ICMP_tunnel CORRESPONDENTES;
- IPv9 IPv6 FORTEMENTE RELACIONADOS;
- IPv9 IPv4 FORTEMENTE RELACIONADOS;
- Palm_Treo Palm_Treo CORRESPONDENTES;
- Computador_analógico Analog_computer CORRESPONDENTES;
- Computador_de_DNA DNA_computing CORRESPONDENTES;
- Microsoft_Media_Services Microsoft_Media_Server CORRESPONDENTES;
- Netflow Netflow CORRESPONDENTES;
- Netflow Netflix NÃO CORRESPONDENTES;
- 1997_Marlboro_500 1997_Marlboro_500 CORRESPONDENTES;
- 1997_Marlboro_500 1999_Marlboro_500 FORTEMENTE RELACIONADOS;
- Indy_Japan_300_de_2007 2007_Indy_Japan_300 CORRESPONDENTES;
- XM_Satellite_Radio_Indy_300_de_2007 2007_XM_Satellite_Radio_Indy_300 CORRESPONDENTES;
- Peugeot_307_WRC Peugeot_307_WRC CORRESPONDENTES;
- Peugeot_307_WRC Peugeot_206_WRC FORTEMENTE RELACIONADOS;
- Atech_Grand_Prix Atech_Grand_Prix CORRESPONDENTES;
- Kingdom_Racing King_Racing FORTEMENTE RELACIONADOS;
- Campos_Racing Campos_Racing CORRESPONDENTES;
- Peugeot_307_WRC Peugeot_307_WRC CORRESPONDENTES;
- Toyota_Corolla_WRC Toyota_Corolla CORRESPONDENTES;
- Porsche_Supercup Porsche_Supercup CORRESPONDENTES;
- John_Casablancas Julian_Casablancas NÃO CORRESPONDENTES;
- Collins An_Collins NÃO CORRESPONDENTES;
- Olheiro Scout_(sport) CORRESPONDENTES;
- Women's_wear_daily Women's_Wear_Daily CORRESPONDENTES;

- Vionnet Madeleine_Vionnet CORRESPONDENTES;
- Lea_T Lea_T CORRESPONDENTES;
- Best_Player Best_Player CORRESPONDENTES;
- Lottery_Ticket The_Lottery_Ticket FORTEMENTE RELACIONADOS;
- Alan_Rickman Alan_Rickman CORRESPONDENTES;
- The_Mutiny_of_the_Bounty The_Mutineers_of_the_Bounty NÃO CORRESPONDENTES;
- As_Pupilas_do_Senhor_Reitor_(1924) As_Pupilas_do_Senhor_Reitor NÃO CORRESPONDENTES;
- Min_and_Bill Min_and_Bill CORRESPONDENTES;
- A_Severa_(filme) A_Severa_(film) CORRESPONDENTES;
- Peter_Sellers Peter_Sellers CORRESPONDENTES;
- Laurence_Olivier Laurence_Olivier CORRESPONDENTES;
- Berlin_Alexanderplatz_(filme) Berlin_Alexanderplatz CORRESPONDENTES;
- Scandal_Sheet Scandal_Sheet CORRESPONDENTES;
- O_Pecado_de_Madelon_Claudet The_Sin_of_Madelon_Claudet CORRESPONDENTES;
- Footlight_Parade Footlight_Parade CORRESPONDENTES;
- Just_Around_the_Corner Just_Around_the_Corner CORRESPONDENTES;
- Gentlemen's_Agreement Gentlemen's_agreement CORRESPONDENTES;
- King_of_Burlesque King_of_Burlesque CORRESPONDENTES;
- Satan_Met_a_Lady Satan_Met_a_Lady CORRESPONDENTES;
- Revolução_de_Maio May_Revolution CORRESPONDENTES;
- Adenograma Adenoma CORRESPONDENTES;
- Bursite_olecraniana Olecranon_bursitis CORRESPONDENTES;
- Cistocentese Cystocentesis CORRESPONDENTES;
- Clínica Clinic CORRESPONDENTES;
- Comorbidade Comorbidity CORRESPONDENTES;
- Enxerto_ósseo Bone_grafting CORRESPONDENTES;
- Espasticidade Spasticity CORRESPONDENTES;

- Estadiamento_do_câncer_de_próstata Prostate_cancer_staging CORRESPONDENTES;
- Fibroplasia Fibroblast CORRESPONDENTES;
- Fibrose_pulmonar Pulmonary_fibrosis CORRESPONDENTES;
- Higidez Health NÃO CORRESPONDENTES;
- Hipocinesia Hypokinesia CORRESPONDENTES;
- Influxo Influx NÃO CORRESPONDENTES;
- Inibidor_da_recaptação_de_dopamina Dopamine_reuptake_inhibitor CORRESPONDENTES;
- Internação Hospitality CORRESPONDENTES;
- Micologia_médica Medical_Mycology NÃO CORRESPONDENTES;
- Smartpill SmartPill CORRESPONDENTES;
- Suporte_avançado_à_vida Advanced_life_support CORRESPONDENTES;
- Suporte_básico_de_vida Basic_life_support CORRESPONDENTES;
- Transudato Transudate CORRESPONDENTES;
- Trombólise Thrombolysis CORRESPONDENTES;
- Bacteriostático Bacteriostatic CORRESPONDENTES;
- Bomba_de_infusão Infusion_pump CORRESPONDENTES;
- Espagíria Spagyric CORRESPONDENTES;
- Farmacodinâmica Pharmacodynamics CORRESPONDENTES;
- Farmacóforo Pharmacophore CORRESPONDENTES;
- Bacterioplâncton Bacterioplankton CORRESPONDENTES;
- Cariolinfa Nucleoplasm CORRESPONDENTES;
- Cascata_trófica Trophic_cascade CORRESPONDENTES;
- Competição_interespecífica Interspecific_competition CORRESPONDENTES;
- Condroplasto Chondroblast CORRESPONDENTES;
- Eletrócito Electricity NÃO CORRESPONDENTES;
- Extrusoma Extrusome CORRESPONDENTES;
- Filhote Puppy CORRESPONDENTES;
- Hematose Gas_exchange CORRESPONDENTES;

- Isolamento_Reprodutivo Reproductive_isolation CORRESPONDENTES;
- Macronutriente Micronutrient FORTEMENTE RELACIONADOS;
- Medula_das_adrenais Adrenal_medulla FORTEMENTE RELACIONADOS;
- Neurobiologia Neurobiology CORRESPONDENTES;
- Orgânico Organic CORRESPONDENTES;
- Origem_da_vida Abiogenesis CORRESPONDENTES;
- Osmorregulação Osmoregulation CORRESPONDENTES;
- Bacterioplâncton Bacterioplankton CORRESPONDENTES;
- Cariolinfa Nucleoplasm CORRESPONDENTES;
- Idioblasto Idioblast CORRESPONDENTES;
- Idiograma Ideograms NÃO CORRESPONDENTES;
- Idiograma Idiogram CORRESPONDENTES;
- As_bruxas_de_Eastwick The_Witches_of_Eastwick CORRESPONDENTES;
- Epígrafe Epigraph_(literature)CORRESPONDENTES ;
- Epígrafe Epigraphy NÃO CORRESPONDENTES;
- Leitor The_Reader NÃO CORRESPONDENTES;
- Literatura_sânscrita Sanskrit_literature CORRESPONDENTES;
- O_Empreendedor Entrepreneur NÃO CORRESPONDENTES;
- On_Dumpster_Diving Dumpster_diving FORTEMENTE RELACIONADOS;
- Provérbio Proverb CORRESPONDENTES;
- Revisão_sistemática Systematic_review CORRESPONDENTES;
- Microinformatica Music_informatics NÃO CORRESPONDENTES;
- Notetaker Notetaking CORRESPONDENTES;
- Advance_86 Advanced/36 FORTEMENTE RELACIONADOS;
- Spectrum_ED Spectrum NÃO CORRESPONDENTES;
- Aventura_na_Selva Jungle_Adventurer NÃO CORRESPONDENTES;
- Evil_Dead:_Hail_To_The_King Evil_Dead:_Hail_to_the_King CORRESPONDENTES;
- KChess Chess FORTEMENTE RELACIONADOS;

- Microsoft_Flight_Simulator_2000 Microsoft_Flight_Simulator FORTEMENTE RELACIONADOS;
- Microsoft_Flight_Simulator_2000 Microsoft_Flight_Simulator_X FORTEMENTE RELACIONADOS;