Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Aprendizado Automático de Relações Semânticas entre Tags de Folksonomias

Alex Sandro da Cunha Rêgo

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Banco de Dados e Sistemas da Informação

Dr. Leandro Balby Marinho / Dr. Carlos Eduardo Santos Pires
(Orientadores)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

R343a Rêgo, Alex Sandro da Cunha.

Aprendizado automático de relações semânticas entre tags de folksonomias / Alex Sandro da Cunha Rêgo. – Campina Grande, 2016.

167 f. : il. color.

Tese (Doutorado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2016.

"Orientação: Prof. Dr. Leandro Balby Marinho, Prof. Dr. Carlos Eduardo Santos Pires".

Referências.

1. Folksonomia. 2. Relações Semânticas. 3. Aprendizado de Máquina. 4. Sinonímia. 5. Hiponímia. I. Marinho, Leandro Balby. II. Pires, Carlos Eduardo Santos. III. Título.

CDU 004.738.5:025.4(043)

„PRENDIZADO AUTOMÁTICO DE RELAÇÕES SEMÂNTICAS ENTRE TAGS DE FOLKSONOMIAS"

ALEX SANDRO DA CUNHA RÊGO

TESE APROVADA EM 15/03/2016

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Orientador(a)

LEANDRO BALBY MARINHO, Dr., UFCG
Orientador(a)

CLAUDIO ELIZIO CALAZANS CAMPELO, PhD., UFCG
Examinador(a)

EVANDRO DE BARROS COSTA, D.Sc, UFAL
Examinador(a)

JUSSARA MARQUES DE ALMEIDA, Dra., UFMG
Examinador(a)

ANA CAROLINA BRANDAO SALGADO, Drª, UFPE
Examinador(a)

CAMPINA GRANDE - PB

Resumo

As folksonomias têm despontado como ferramentas úteis de gerenciamento *online* de conteúdo digital. A exemplo dos populares *websites* Delicious, Flickr e BibSonomy, diariamente os usuários utilizam esses sistemas para efetuar *upload* de recursos web (e.g., url, fotos, vídeos e referências bibliográficas) e categorizá-los por meio de *tags*. A ausência de relações semânticas do tipo sinonímia e hiperonímia/hiponímia no espaço de *tags* das folksonomias reduz a capacidade do usuário de encontrar recursos relevantes. Para mitigar esse problema, muitos trabalhos de pesquisa se apoiam na aplicação de medidas de similaridade para detecção de sinonímia e construção automática de hierarquias de *tags* por meio de algoritmos heurísticos. Nesta tese de doutorado, o problema de detecção de sinonímia e hiperonímia/hiponímia entre pares de *tags* é modelado como um problema de classificação em Aprendizado de Máquina. A partir da literatura, várias medidas de similaridade consideradas boas indicadoras de sinonímia e hiperonímia/hiponímia foram identificadas e empregadas como atributos de aprendizagem. A incidência de um severo desbalanceamento e sobreposição de classes motivou a investigação de técnicas de balanceamento para superar ambos os problemas. Resultados experimentais usando dados reais das folksonomias BibSonomy e Delicious mostraram que a abordagem proposta denominada CPDST supera em termos de acurácia o *baseline* de melhor desempenho nas tarefas de detecção de sinonímia e hiperonímia/hiponímia. Também, aplicou-se a abordagem CPDST no contexto de geração de listas de *tags* semanticamente relacionadas, com o intuito de prover acesso a recursos adicionais anotados com outros conceitos pertencentes ao domínio da busca. Além da abordagem CPDST, foram propostos dois algoritmos fundamentados no acesso ao WordNet e ConceptNet para sugestão de listas especializadas com *tags* sinônimas e hipônimas. O resultado de uma avaliação quantitativa demonstrou que a abordagem CPDST provê listas de *tags* relevantes em relação às listas providas pelos métodos comparados.

Palavras-chave: Folksonomia; Relações semânticas; Aprendizado de Máquina; Similaridade Semântica; Sinonímia; Hiperonímia; Hiponímia.

Abstract

Folksonomies have emerged as useful tools for online management of digital content. Popular websites as Delicious, Flickr and BibSonomy are now widespread with thousands of users using them daily to upload digital content (e.g., webpages, photos, videos and bibliographic information) and tagging for later retrieval. The lack of semantic relations such as synonym and hypernym/hyponym in the tag space may diminish the ability of users in finding relevant resources. Many research works in the literature employ similarity measures to detect synonymy and to build hierarchies of tags automatically by means of heuristic algorithms. In this thesis, the problems of synonym and subsumption detection between pairs of tags are cast as a pairwise classification problem. From the literature, several similarity measures that are good indicators of synonymy and subsumption were identified, which are used as learning features. Under this setting, there is a severe class imbalance and class overlapping which motivated us to investigate and employ class imbalance techniques to overcome these problems. A comprehensive set of experiments were conducted on two large real-world datasets of BibSonomy and Delicious systems, showing that the proposed approach named CPDST outperforms the best performing heuristic-based baseline in the tasks of synonym and subsumption detection. CPDST is also applied in the context of tag list generation for providing access to additional resources annotated with other semantically related tags. Besides CPDST approach, two algorithms based on WordNet and ConceptNet accesses are proposed for capturing specifically synonyms and hyponyms. The outcome of an evaluative quantitative analysis showed that CPDST approach yields relevant tag lists in relation to the produced ones by the compared methods.

Keywords: Folksonomy; Semantic Relations; Machine Learning, Semantic Similarity; Synonymy; Hypernymy; Hyponymy.

Agradecimentos

No tocante desta importante fase da minha vida acadêmica, quero agradecer primeiramente a Deus, força maior para minha perseverança e empenho na conquista dos objetivos que me são oferecidos.

Aos meus orientadores, Dr. Leandro Balby Marinho e Dr. Carlos Eduardo Santos Pires, os mais sinceros agradecimentos pelo constante acompanhamento do trabalho realizado nesta tese de doutorado, respeito, sensibilidade no trato de questões que em alguns momentos atrapalharam meu ritmo de trabalho e condução da transformação científica vivenciada ao longo dos anos de pesquisa; vocês são agentes transformadores na vida de muitos alunos que passam por uma pós-graduação e devem sim ser sempre exaltados pelo fundamental papel que exercem à educação e pesquisa no nosso país. Aprendi bastante e terei-os sempre como referência.

À minha mãe, Maria Alice dos Santos, meu eterno amor, admiração e gratidão pelos ensinamentos de vida, construção de valores e total dedicação à criação sozinha dos seus filhos, mostrando sempre que a educação é o melhor caminho para superar as adversidades da vida.

Agradeço também aos pesquisadores autores de trabalhos relacionados ao tema da minha pesquisa que gentilmente partilharam experiência com sugestões e/ou pontos de vista sobre questões que em determinado momento implicaram em maiores dificuldades de resolução. Muito obrigado Dr. Christoph Trattner (pesquisador do *Know-Center Research Center for Big Data Analytics at Graz University of Technology*, Graz, Áustria), Dr. Vladimir Soares Catão (Professor da Universidade Federal de Campina Grande campus de Cuité, Paraíba, Brasil), Dr. Gustavo Enrique A.P.A. Batista (Professor do Instituto de Ciências Matemáticas e de Computação - USP, São Carlos, São Paulo, Brasil), Dr. Anderson de Rezende Rocha (Professor do Instituto de Computação -UNICAMP, São Paulo, Brasil) e Dr. Nitesh Chawla (Professor da Universidade de Notre Dame, Indiana, USA).

E por fim, aos amigos do programa de doutorado da UFCG que tive o prazer de conhecer, compartilhar conhecimentos e criar laços de amizade: José Gildo de Araújo Júnior, Demétrio Gomes Mestre e Dimas Cassimiro do Nascimento Filho.

Dedicatória

Dedico este trabalho à minha esposa **Danielle de Carvalho Pereira**, minha fortaleza, o amor da minha vida, mulher que não me deixou fraquejar nos momentos de maior dificuldade e que aceitou a renúncia de me ter por completo como esposo para que eu pudesse me dedicar inteiramente aos estudos. Essa vitória é nossa!

Também dedico este trabalho ao meu primogênito **Nathan Pereira Cunha**, que nasceu no início do meu doutorado e me presenteia até hoje com amor, sorrisos e dias mais felizes. Ao seu lado, você fez passar despercebido o peso do cansaço mental de dias inteiros de trabalho. Amo vocês!

Conteúdo

Lista de Símbolos

AM - Aprendizado de Máquina

ANOVA - *ANalisys Of VAriance*

API - *Application Programming Interface*

ARIPPER - *Alternate Repeated Incremental Pruning to Produce Error Reduction*

BWRF - *Balanced and Weighted Random Forests*

ARFF - *Attribute-Relation File Format*

ASC - Aprendizado Sensível ao Custo

BREL - *BibSonomy RELated*

BSIM - *BibSonomy SIMilar*

CE - Configuração Experimental

CPDST - Classificação Para Detecção de relações Semânticas entre pares de Tags

CSR - *Classification-based Learning of Subsumption Relations*

CNET - *ConceptNet*

CNN - *Condensed Nearest Neighbor Rule*

CNNT -*Complementary Neural Network*

DBLP - *DataBase systems and Logic Programming*

DE - Distância de Edição

DEN - Distância de Edição Normalizada

FN - Falso Negativo

FP - Falso Positivo

FDA - Função de Distribuição Acumulada

HTTP - *HyperText Transfer Protocol*

IA - Inteligência Artificial

IADIS - *International Association for Development of the Information Society*

IF - *Information Gain*

JAWS - *Java API for WordNet Searching*

JRip - *Java implementation of the rule learner Repeated Incremental Pruning to Produce Error Reduction - RIPPER*

JSON - *JavaScript Object Notation*

KDD - *Knowledge Discovery in Databases*

kNN - *k Nearest Neighbors*

LSA - *Latent Semantic Analysis*

OSS - *One-Sided Selection*

OAA - *One-Against-All*

OAO - *One-Against-One*

PLN - Processamento de Linguagem Natural

QP - Questão de Pesquisa

RBF - *Radial Basis Function*

REST - *REpresentational State Transfer*

RI - Recuperação da Informação

RIPPER - *Repeated Incremental Pruning to Produce Error Reduction*

RU - *Random Undersampling*

SAC - *Symposium of Applied Computing*

SATLG - *Semantic-Aware Tag List Generation*

SMOTE - *Synthetic Minority Over-sampling TEchnique*

SQL - *Structured Query Language*

SVM - *Support Vector Machines*

TL - *Tomek Link*

URL - *Uniform Resource Locator*

VN - Verdadeiro Negativo

VP - Verdadeiro Positivo

WNET - *WordNet*

WS4J - *WordNet Similarity for Java*

WWW - *World Wide Web*

Lista de Figuras

Lista de Tabelas

Lista de Códigos Fonte

Capítulo 1

Introdução

A popularidade da Web 2.0 fez surgir uma série de novos serviços e funcionalidades que modificaram a forma como os usuários acessam informações na web, introduzindo um hábito em que os usuários passam a interagir e colaborar uns com os outros. Neste cenário, os usuários participam de forma ativa na criação, organização e interação com o conteúdo, de maneira dinâmica e colaborativa. Mesmo quando o conteúdo não é gerado pelos usuários, este pode ser enriquecido com comentários ou avaliações.

As folksonomias[1] são um exemplo clássico de sistema que incorpora o paradigma Web 2.0. Folksonomias são sistemas de classificação em que o próprio usuário fica responsável por criar e gerenciar *tags* de forma colaborativa para anotar ou categorizar recursos, definindo desta forma sua própria estratégia de organização. Um recurso pode ser entendido como qualquer objeto de conteúdo digital como, por exemplo, foto, vídeo, documento ou URL (*Uniform Resource Locator*). *Tags* são palavras-chave escolhidas livremente as quais refletem o vocabulário empregado pelos usuários de uma folksonomia para identificar os recursos (QUINTARELLI, 2005).

Descrever recursos por meio de *tags* tornou-se uma prática popular aos usuários da WWW (*World Wide Web*) desde o lançamento dos precursores sites Delicious[2] e Flickr[3] (GUPTA et al., 2010). Ambos são, respectivamente, sistemas de gerenciamento e compartilhamento de *bookmarks* e fotos. Desde então, outros sistemas sociais foram

[1] Outras denominações para o termo folksonomia: *Social Bookmarking*, *Social Web*, *Collaborative Tagging Systems* e *Social Tagging*.

[2] <http://www.delicious.com>

[3] <http://www.flickr.com>

desenvolvidos para prover suporte à anotação de uma variedade de recursos. Por exemplo, existem folksonomias para classificação e compartilhamento de referências bibliográficas[4], músicas[5], livros[6], vídeos[7] e até mesmo organização de metas pessoais[8]. Aplicações de outro domínio, como o gerenciador de emails Gmail[9], também oferecem a opção de utilizar *tags* ao invés de pastas para identificar e organizar *emails*. Entretanto, este não é considerado um exemplo típico de folksonomia porque não permite que as *tags* sejam compartilhadas com outros usuários do sistema.

De um modo geral, as principais características que tornam as folksonomias atrativas à comunidade de usuários web são:

- O esforço exigido para aprender a manusear uma folksonomia é mínimo, pois não requer que o usuário possua habilidades especializadas. O processo de anotação é intuitivo e adapta-se facilmente ao perfil cognitivo de seus usuários colaboradores;
- O usuário emprega a *tag* que melhor expressa seu entendimento ou necessidade acerca de um recurso, aplicando sua visão pessoal ao processo de anotação;
- Por meio de suas *tags*, as folksonomias indexam qualquer tipo de mídia, não sendo limitadas apenas a documentos textuais;
- A folksonomia permite que usuários de diversos países utilizem palavras do seu próprio idioma como *tags*;
- Com o passar do tempo, a própria folksonomia se adapta automaticamente ao surgimento de novas *tags*. *Tags* consideradas ultrapassadas serão cada vez menos utilizadas e isto fará com que caiam em desuso naturalmente. Em contrapartida, *tags* que alcançarem alta frequência de utilização estarão sempre em evidência.

Apesar de todas as suas vantagens, a democracia promovida pela livre criação de *tags* acaba desenvolvendo um vocabulário descontrolado, repleto de termos redundantes e sem conectividade semântica explícita. Neste contexto, existe um grande interesse no

[4] <http://www.bibsonomy.org>, <http://www.citeulike.org>
[5] <http://www.last.fm>
[6] <http://www.librarything.com>
[7] <http://www.youtube.com>
[8] <http://www.43things.com>
[9] <https://mail.google.com/>

desenvolvimento de estratégias para tentar inferir semântica no espaço de *tags* e então identificar relações úteis que possam incrementar a descoberta de recursos relevantes, o que indiretamente impacta na melhoria da qualidade dos serviços de Recuperação da Informação (RI) que utilizam *tags* como elementos de acesso à informação. Esta tese focaliza justamente o problema de detecção automática de relações semânticas entre *tags* de folksonomias e tem por objetivo propor uma nova abordagem que utiliza técnicas de aprendizado supervisionado para detectar com maior nível de acerto relações de sinonímia e hiperoníma/hiponímia entre *tags* de folksonomias.

1.1 Motivação

A principal característica que torna uma folksonomia tão popular e atrativa acarreta problemas que limitam o acesso ao conteúdo armazenado. Somados à diversidade cultural, linguística e comportamental dos seus usuários, sem uma padronização inevitavelmente será desenvolvido um espaço de *tags* inconsistente (MAGABLEH et al., 2010). Os principais problemas disseminados na literatura estão relacionados à ambiguidade, redundância e inconsistência no uso de *tags* (ANGELETOU; SABOU; MOTTA, 2008; MAGABLEH et al., 2010; WU; ZHANG; YU, 2006; YEUNG; GIBBINS; SHADBOLT, 2007), decorrentes dos seguintes fatores:

1. **Semântico**: inexiste um controle para tratamento de sinonímia, polissemia e homonímia. Além disso, conexões hierárquicas do tipo hiponímia e hiperonímia (tipicamente uma relação *is-a*) não são estabelecidas entre *tags* (ANGELETOU; SABOU; MOTTA, 2008; GOLDER; HUBERMAN, 2006);

2. **Linguístico**: erros ortográficos, ausência de convenção de termos (e.g., nomes próprios começando com letra maiúscula), variações léxicas da mesma palavra, uso de *tags* em diferentes idiomas, acrônimos e *tags* compostas por múltiplas palavras;

3. **Cognitivo**: heterogeneidade de vocabulário decorrente da variação entre o nível de conhecimento básico e especializado dos usuários durante o processo de anotação. Por exemplo, as *tags* `C` e `java` podem ser específicas demais para alguns usuários, enquanto `programação` pode ser muito geral para outros.

Desconsiderar relações semânticas entre *tags* reduz as chances do usuário descobrir recursos relevantes. Cada usuário enxerga um recurso de forma diferente, portanto, usuários distintos podem aplicar diferentes *tags* para expressar entendimentos pessoais acerca do recurso anotado. Entretanto, o conjunto heterogêneo de *tags* atribuído pelos usuários aos recursos de mesma contextualização tende a refletir conceitos estreitamente associados. Assim, embora tais recursos sejam fortemente relacionados, acabam não sendo descobertos por causa do problema de incompatibilidade de palavras.

A natureza colaborativa das folksonomias contribui naturalmente para o surgimento de muitas *tags* sinônimas, provocando, eventualmente, baixo desempenho na recuperação de recursos relevantes. A detecção de sinonímia é motivada principalmente pela presença de termos redundantes. Por exemplo, um recurso que trata sobre veículos de passeio pode ser anotado com as *tags* `car`, `automobile` ou `motorcar`. Se um usuário deseja recuperar tal recurso utilizando a *tag* `car` mas o recurso em questão está anotado com a *tag* `automobile`, o sistema não será capaz de encontrá-lo, embora `automobile` seja um sinônimo de `car`.

A detecção de relações hierárquicas é particularmente importante porque certos recursos podem ser anotados com um vocabulário que sintetiza os diferentes níveis de conhecimento de seus usuários acerca de um conteúdo, envolvendo interpretações desde sua forma mais abrangente até a mais específica. Tomando como referência o exemplo anterior (recurso sobre veículos de passeio), é coerente o uso das *tags* `honda`, `sedam`, `flexcar` e `conceptcar`. Ao observar todo o conjunto de *tags* utilizado para descrever o referido recurso (`car`, `automobile`, `motorcar`, `honda`, `sedam`, `flexcar` e `conceptcar`), pode-se inferir que a *tag* `car` é mais genérica do que `sedam`, o que presumivelmente fornece um indicativo de hierarquia implícita. Na linguística, `sedam` é compreendido como hipônimo de `car` e `car` por sua vez é um hiperônimo de `sedam`[10]. Mais uma vez, em uma tarefa de busca, o uso da *tag* inapropriada impede que tal recurso seja recuperado.

Uma maneira de reduzir não só o problema de sinonímia, como também amenizar os de ambiguidade em folksonomias, é adotar o conceito de folksonomia controlada com o suporte de um *thesaurus* (NORUZI, 2007), no qual o próprio sistema interage com os usuários apresentando sugestões de *tags* de acordo com o contexto, identifica casos de polissemia para

[10]Um relacionamento hipônimo-hiperônimo (*is-a*) também é conhecido como relação de subordinação.

desambiguação ou aplica corretor ortográfico em tempo real. Por outro lado, um *thesaurus* evolui lentamente em relação ao vocabulário praticado em uma folksonomia e por isso não é capaz de acompanhar e/ou reconhecer *tags* que, por exemplo, refletem termos típicos de uma região, gírias ou informalidades (PETERS, 2009). O sistema Faviki[11] é um exemplo de folksonomia controlada que permite aos usuários utilizar conceitos da Wikipedia como *tags*. Entretanto, alguns pesquisadores não concordam com este tipo de abordagem porque a folksonomia controlada limita os usuários a um restrito conjunto de *tags*, contrapondo-se à principal característica de uma folksonomia: permitir que os usuários criem livremente suas *tags*, mesmo sabendo que a consistência pode ser comprometida.

Na RI, a técnica tradicional de busca textual com base em palavra-chave define um tipo de pesquisa na qual os termos de uma consulta são comparados com os termos existentes em documentos individuais, armazenados em um banco de dados (indexados) e ranqueados algoritmicamente (BEALL, 2008). Entretanto, a técnica em si é ineficiente na recuperação textual quando se leva em conta questões de ordem semântica, pois não consegue, por exemplo, reconhecer sinônimos ou identificar homônimos indesejados.

A literatura aborda diferentes estratégias para lidar com esta limitação, dentre as quais destacam-se: (i) aplicação de técnicas de Processamento de Linguagem Natural (PLN) para combinar o conteúdo sintático/semântico dos termos da consulta com o conteúdo semântico dos documentos (reconhecimento de padrões léxico-sintático (HEARST, 1992), suporte de um *thesaurus*, utilização de algoritmos de *stemming* para extração da raiz morfológica da palavra, entre outros); e (ii) aplicação de métodos estatísticos para classificar numericamente documentos que estimam maior utilidade em relação aos termos da consulta, utilizando medidas de similaridade (SINGHAL, 2001).

Embora os problemas linguísticos tenham sido amplamente abordados pela comunidade científica da área de RI, a maioria das soluções necessita de adaptação quando aplicada às folksonomias. As técnicas tradicionais de PLN não são adequadas para detectar sinônimos, hiperônimos ou hipônimos em folksonomias, visto que *tags* são palavras isoladas, na maioria dos casos, e os recursos podem assumir vários formatos além de texto, tais como vídeo, áudio e fotografia.

[11] <http://www.faviki.com>

A justificativa para desenvolvimento de uma nova abordagem que vise a descoberta de relações semânticas do tipo sinonímia e hiperonímia/hiponímia se deve à importância com que o tema vem sendo explorado pela comunidade científica, pois tais relacionamentos semânticos estão nitidamente relacionadas à tríade de problemas típicos de uma folksonomia: semântico, linguístico e cognitivo. Uma estratégia amplamente empregada na literatura para lidar com os referidos problemas é aplicar medidas de similaridade/distância semântica para mensurar o grau de parentesco entre *tags*, mas sem distinguir o tipo de relação semântica capturado (CATTUTO et al., 2008; CLEMENTS; VRIES; REINDERS, 2008; MARKINES et al., 2009; MOUSSELLY-SERGIEH et al., 2014; RADELAAR et al., 2011; SOLSKINNSBAKK; GULLA, 2011; WU; ZHOU, 2009). Após revisão minuciosa da literatura relacionada, não foi encontrado outro trabalho que aplique a técnica de Aprendizado de Máquina (AM) para identificar, automaticamente, relações de sinonímia e hiperonímia/hiponímia entre *tags* de folksonomias.

1.2 Objetivos

O principal objetivo desta tese é propor uma abordagem de aprendizado indutiva para a detecção de semântica entre *tags* de folksonomias, direcionada especificamente para a identificação de dois tipos de relação semântica: sinonímia e hiperonímia/hiponímia. Na abordagem proposta, deste ponto em diante denominada **Classificação Para Detecção de relações Semânticas entre pares de Tags - CPDST**, as relações semânticas são aprendidas diretamente dos dados da folksonomia. Para esta finalidade, o problema de detecção de relações semânticas entre *tags* é modelado como um problema de classificação, no qual os atributos de aprendizado são medidas de similaridade/distância *tag-tag* empregadas como heurísticas por diversos trabalhos relacionados (cf. Capítulo 3) para capturar relações de parentesco entre *tags*.

Formular o problema desta forma implica em maiores dificuldades de treinamento pois o número de pares de *tags* que constitui uma relação específica (exemplos positivos) é largamente inferior ao número de pares de *tags* que não apresenta qualquer conexão semântica (exemplos negativos). Tal contexto é conhecido como problema de desbalanceamento de classes e, como consequência, o classificador é fortemente enviesado

a predizer sempre a classe majoritária. A isso, alia-se o fato de haver a necessidade de combater o fenômeno de sobreposição de classes, ou seja, instâncias de diferentes classes com características similares, o que torna o problema ainda mais complexo de se resolver.

Este objetivo geral pode ser delineado em três objetivos específicos, conduzidos pelas seguintes questões de pesquisa:

- *Questão de Pesquisa 1 (QP1): Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de sinonímia provê melhor acurácia na identificação de sinonímia entre tags de folksonomia?*

  Aperfeiçoar os resultados em torno da descoberta de relações de sinonímia tem sido alvo de pesquisas. Percebe-se que, de um modo geral, heurísticas são aplicadas isoladamente para conduzir essa tarefa e as avaliações são feitas de forma qualitativa. Para responder essa questão, propõe-se uma nova estratégia fundamentada na técnica de aprendizado supervisionado para predizer automaticamente uma semântica específica entre *tags*. Para fins de detecção de sinonímia, realizou-se uma avaliação não só qualitativa, como também quantitativa em comparação com 5 (cinco) heurísticas comumente empregadas para este propósito.

- *Questão de Pesquisa 2 (QP2): Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de relações hierárquicas provê melhor acurácia na identificação de relações hierárquicas (hiperonímia e hiponímia) entre tags de folksonomias?*

  Para responder essa questão, estende-se o modelo geral de predição semântica mencionado na QP1 e seleciona-se novos atributos ou funções objetivas que capturem prováveis relacionamentos do tipo hiperonímia/hiponímia entre *tags*. Normalmente, técnicas probabilísticas embasadas na medição de sobreposição mútua de instâncias de *tags* são usadas para estimar o quanto o significado de uma *tag* está incorporado ao significado de outra e vice-versa.

- *Questão de Pesquisa 3 (QP3): É possível melhorar a relevância de uma lista de tags relacionadas utilizando tags geradas a partir da abordagem CPDST?*

  Considere o cenário em que um usuário deseja acessar os recursos que foram

anotados por uma *tag* em particular. Como resultado, além de poder visualizar a lista de recursos anotados explicitamente pela *tag* de busca, também é sugerida uma lista de *tags* relacionadas que o usuário poderia utilizar para acessar um conjunto secundário de recursos relacionados ao tema abordado pela *tag* de busca. O sistema BibSonomy provê essa funcionalidade por meio dos métodos "*tags* relacionadas" e "*tags* similares". A fim de responder a QP3, propõe-se a aplicação da abordagem CPDST como estratégia para criação de listas de *tags* relacionadas. Além da abordagem CPDST, são propostos algoritmos de seleção de *tags* fundamentadas no acesso aos dicionários WordNet e ConceptNet. Uma avaliação quantitativa usando as métricas relevância e sobreposição (*overlap*) (LEGINUS; DOLOG; LAGE, 2013a) foi efetuada, comparando-se o desempenho de diferentes métodos em relação a um conjunto de *tags* usadas como chave de busca.

1.3 Contribuições

No tocante ao cumprimento dos objetivos propostos, este trabalho incorpora à literatura as seguintes contribuições:

- Formulação do problema de detecção de relações semânticas, especificamente sinonímia e hiperonímia/hiponímia, como um problema de classificação em pares de *tags*. Relações semânticas não são explícitas no espaço de *tags* de folksonomias e os métodos existentes para mensurar similaridade semântica entre *tags* não são capazes de determinar o tipo do relacionamento semântico. A partir dos trabalhos relacionados, foram identificadas várias medidas de similaridade consideradas boas indicadoras de sinonímia e hiperonímia/hiponímia sob diferentes aspectos, as quais são empregadas como atributos de aprendizado;

- Enfrentamento dos problemas de desbalanceamento e *overlapping* de classes utilizando diferentes algoritmos de aprendizado e técnicas de balanceamento que extraem os exemplos mais discriminativos. Além disso, para efeito comparativo, foram realizados experimentos com amostras de dados submetidas inicialmente a um critério de filtragem com o intuito de reduzir o nível de desbalanceamento. Os resultados

mostram que alguns classificadores conseguem obter um melhor desempenho em termos de acurácia quando treinados a partir de bases de dados provenientes de um filtro de pré-processamento, ao invés de amostras geradas por técnicas de balanceamento;

- Avaliação quantitativa da abordagem CPDST em duas bases de dados de sistemas reais: BibSonomy e Delicious. Para a detecção de sinonímia, os resultados mostraram que a abordagem CPDST é estatisticamente significante superior ao *baseline* de melhor desempenho apoiado em heurística, conseguindo alcançar índices de acurácia superiores a 90% nas duas bases de dados. Em se tratando de hiperonímia/hiponímia, por ser uma relação de difícil detecção, a acurácia medida foi menor (entre 10% e 30%), porém estatisticamente significante superior ao *baseline* de melhor desempenho em cada base de dados;

- Proposta da funcionalidade de geração de lista de *tags* relacionadas com semântica bem-definida em relação a uma chave de busca inicial, utilizando a abordagem CPDST como fonte geradora de *tags* sinônimas, hipônimas ou ambas. Além da abordagem CPDST, foram concebidos algoritmos para prover as mesmas funcionalidades mediante acesso às bases de conhecimento WordNet e ConceptNet. Algumas folksonomias (e.g., BibSonomy) disponibilizam listas de *tags* relacionadas empregando medidas de similaridade que inferem a existência de parentesco semântico entre *tags*, mas sem se preocupar em determinar a especificidade do relacionamento semântico envolvido. Logo, os usuários não conseguem discernir qual o relacionamento semântico que as *tags* da lista estabelecem em relação à chave de busca;

- Realização de uma avaliação quantitativa das listas de *tags* providas pelos métodos comparados, utilizando métricas que mensuram a qualidade das *tags* relacionadas sob a perspectiva dos usuários. Na literatura consultada, não foi encontrado outro trabalho que realizasse esse tipo de avaliação considerando diferentes métodos e especificidade semântica empregada na geração de listas de *tags*.

Até o presente momento, os resultados dessa pesquisa culminaram com as seguintes publicações científicas: (a) *Learning Synonym Relations from Folksonomies*, nos

Proceedings da IADIS (*International Association for Development of the Information Society*) Internet Conference 2012 (WWW/Internet 2012), Madri, Espanha (RÊGO; MARINHO; PIRES, 2012), e (b) *A Supervised Learning Approach to Detect Subsumption Relations Between Tags in Folksonomies*, nos *Proceedings* da SAC 2015 (*Symposium of Applied Computing*), Salamanca, Espanha (RÊGO; MARINHO; PIRES, 2015). Um novo artigo se encontra em produção a ser submetido para um periódico internacional, o qual aborda a proposta de concepção e avaliação de listas de *tags* semanticamente relacionadas a uma chave de busca.

1.4 Estrutura do Documento

O restante deste documento de tese está organizado da seguinte forma:

- **Capítulo 2 - Fundamentação Teórica**: oferece uma explanação básica acerca dos principais assuntos que abrangem o presente trabalho. O texto discorre sobre a importância da semântica na RI, os tipos de relações semânticas exploradas no âmbito desta tese, introduz os dicionários linguísticos utilizados como ferramenta auxiliar ao mapeamento semântico, apresenta uma definição formal de folksonomias e medidas de similaridade/distância semântica e, por fim, provê um entendimento sobre Aprendizado de Máquina e as técnicas de Mineração de Dados utilizadas no trabalho;

- **Capítulo 3 - Trabalhos Relacionados**: neste capítulo é apresentada a revisão bibliográfica acerca dos principais trabalhos relacionados encontrados na literatura e suas respectivas propostas. Essencialmente, os trabalhos estão agrupados por tema: (i) pesquisas que abordam o problema de detecção de sinonímia em folksonomias, (ii) propostas focalizadas na construção de relações hierárquicas entre *tags*, e (iii) interfaces de navegação com base na exploração de conceitos relacionados semanticamente;

- **Capítulo 4 - A Abordagem CPDST**: introduz todos os detalhes acerca da formulação do problema de classificação proposto, arquitetura, extração de atributos e outros detalhes inerentes à sua concepção;

- **Capítulo 5 - Metodologia Experimental**: apresenta detalhes acerca das atividades de pré-processamento dos dados, criação das instâncias de treinamento e teste e protocolo de avaliação que norteia a avaliação dos experimentos realizados no trabalho;

- **Capítulo 6 - Resultados e Discussão**: neste capítulo, os resultados dos experimentos para detecção de sinonímia e hiperonímia/hiponímia são exibidos e discutidos. Uma análise comparativa entre a abordagem CPDST e os *baselines* em termos de acurácia é apresentada;

- **Capítulo 7 - Listas de *Tags* Relacionadas**: apresenta uma motivação para o uso da abordagem CPDST direcionado à criação de listas de *tags* relacionadas com semântica mais específica. Além da motivação, são definidas as métricas para avaliação quantitativa, protocolo de avaliação, resultados obtidos e considerações sobre os experimentos;

- **Capítulo 8 - Conclusões e Trabalhos Futuros**: por fim, neste capítulo são expostas as principais conclusões extraídas com a realização da pesquisa, limitações e direcionamentos para possíveis trabalhos futuros;

- **Apêndices**: fornece uma discussão mais detalhada sobre os casos típicos de sinonímia comumente encontrados em folksonomias, um estudo sobre a variação de limiar para detecção de sinônimos utilizando a medida de distância de edição e uma explanação mais completa sobre o funcionamento do algoritmo *Taxonomy Learning* proposto por Marinho, Buza e Schmidt-Thieme (2008).

Capítulo 2

Fundamentação Teórica

Neste capítulo é fornecido um embasamento teórico acerca dos principais temas que caracterizam o trabalho proposto. A Seção 2.1 discorre brevemente sobre a importância do tratamento de semântica nos sistemas de RI. Na Seção 2.2, é apresentado um entendimento sobre os tipos de relações semânticas entre palavras que normalmente se manifestam em folksonomias, enquanto que a Seção 2.3 descreve os dicionários linguísticos utilizados neste trabalho. Uma definição formal de folksonomia é introduzida na Seção 2.4. A Seção 2.5 discute sobre as terminologias empregadas na literatura para definir o grau de similaridade semântica entre objetos. Por fim, na Seção 2.6 é provida uma visão geral a respeito da técnica de Aprendizado de Máquina e outros tópicos pertencentes ao domínio que estão conectados diretamente ao trabalho.

2.1 O Papel da Semântica na Recuperação da Informação

Os sistemas de RI normalmente recebem como entrada palavras expressas em linguagem natural e o grande desafio é fazer com que estes sistemas saibam exatamente quais as reais pretensões do usuário após formular uma consulta.

A indexação convencional, levando-se em conta apenas a ocorrência dos termos em um documento sem saber seu real significado, provavelmente não terá a eficiência desejada visto que muitos recursos relevantes poderão ficar de fora, ou serão retornados recursos impróprios ao sentido do termo. Isto porque, desconsiderar fenômenos linguísticos relacionados à semântica das palavras, tais como polissemia, homonímia, sinonímia, hiperonímia e

hiponímia (cf. Seção 2.2), eventualmente diminui a relevância dos resultados retornados por um sistema de RI. Por exemplo, se o usuário utiliza o termo `prato` em uma consulta, a que está se referindo? Comida, instrumento musical ou peça de louça? Se o termo utilizado for `fruta`, por que não considerar recursos que foram anotados com os termos `laranja`, `uva` ou `acerola`? Se o usuário digita o termo `Beijing`, mas só existem documentos contendo o termo `Pequim`, porque não incluir estes documentos no resultado da consulta?

No campo da linguística, o dicionário Houaiss (HOUAISS, 2009) define o significado da palavra **semântica** como "*o componente do sentido das palavras e da interpretação das sentenças e enunciados*". A semântica determina a conotação do uso das palavras, símbolos e sinais em uma frase, e da relação entre elas para compreender o significado da expressão humana por meio da linguagem.

A descoberta de relações semânticas entre *tags* no ambiente de uma folksonomia é um requisito importante para potencializar a busca por recursos relevantes. Ainda mais porque, diante da diversidade de formatos digitais que podem ser anotados por *tags*, muitos deles não oferecem a possibilidade da indexação convencional que é realizada com texto. A seguir, são apresentados conceitos básicos acerca de alguns aspectos semânticos explorados neste trabalho. Uma explanação completa sobre outros tipos de relação semântica não cobertos neste trabalho pode ser encontrado em Bean e Green (2001).

2.2 Relações entre Palavras

As palavras estabelecem entre si relações de significado, as quais podem ser de ordem semântica ou fonéticas/gráficas (AZEREDO, 2008). Um entendimento sobre os tipos de relações entre palavras mais comuns em folksonomias é descrito nesta seção.

2.2.1 Relação Semântica

As relações de ordem semântica a seguir discriminadas conceitualmente são: sinonímia, hiperonímia e hiponímia. No que diz respeito aos significados de ordem fonéticas/gráfica, são apresentadas as relações de homonímia e polissemia.

Sinonímia

Sinonímia é o nome dado para a relação de sentido entre dois vocábulos que possuem significados idênticos ou muito próximos (HOUAISS, 2009), os quais podem ser usados no mesmo contexto, sem que exista uma alteração de significado do enunciado em que ocorrem. Por exemplo, as palavras `condutor` e `motorista` expressam uma relação de sinonímia no seguinte contexto: *O motorista/condutor do veículo não prestou socorro à vítima.*

A existência de sinonímia é muito comum em folksonomias, pois não existe uma convenção para definição de palavras e uma mesma *tag* pode ser escrita de várias formas. Por exemplo, é comum o uso de letras capitalizadas, utilização de caracteres especiais (_,-,{,}) e ocorrência de erros ortográficos. Todos esses fatores acabam provocando redundância no espaço de *tags*.

Hiperonímia e Hiponímia

Hiperonímia é uma relação semântica hierárquica estabelecida entre um vocábulo de sentido mais genérico e outro de sentido mais específico, ambos com traços semânticos comuns. O primeiro vocábulo sempre impõe suas propriedades ao segundo. Por exemplo, `animal` estabelece uma relação de hiperonímia com `gato` ou `cachorro`, portanto, pode-se afirmar que `animal` é hiperônimo de `gato`.

Hiponímia expressa uma ideia inversa à definição de hiperonímia. Hiponímia é uma relação semântica estabelecida entre um termo de sentido mais específico e outro de sentido mais genérico. No exemplo dos vocábulos `animal` e `gato`, pode-se afirmar que `gato` é hipônimo de `animal`.

Um hiperônimo pode substituir, em todos os contextos, qualquer um de seus hipônimos. A recíproca não é verdadeira porque um hipônimo possui propriedades especializadas. Em folksonomias, as *tags* pertencem a um espaço plano, portanto, inexiste uma definição quanto ao relacionamento hierárquico do tipo hiperonímia e hiponímia. Neste caso, necessita-se de mecanismos ou abordagens específicas para inferir esse tipo de relação.

2.2.2 Relação Fonética/Gráfica

Homonímia e polissemia são dois conceitos que estão relacionados com o emprego de uma mesma palavra em expressões que assumem significados diferentes. A diferença é sutil, mas simples de perceber.

Polissemia (do Grego poli, "muitos", e sema, "significado") é a característica que um vocábulo ou expressão apresenta em adquirir um novo significado, além do original, de acordo com o contexto em que foi empregado. É uma consequência do uso figurado das palavras, analogia, metáfora ou extensão de sentido (HOUAISS, 2009). Por exemplo, nas frases abaixo,

1. O carro estava estacionado com a chave na ignição.

2. A chave geral foi desligada e ficamos sem luz.

a palavra *chave* exerce significados polissêmicos: *chave de carro* e *chave de luz*.

A homonímia é uma propriedade semântica expressa por dois vocábulos de grafia idêntica ou fônica, mas com significados distintos. Os tipos de homonímia podem ser classificados em:

- **Homônimos homógrafos**: palavras de mesma grafia e pronúncia diferente. Por exemplo, "gosto" (substantivo) e "gosto" (verbo gostar, presente do indicativo);
- **Homônimos homófonos**: palavras com mesma pronúncia, mas com grafia diferente. Por exemplo, "seção" (setor) e "sessão" (atividade);
- **Homônimos homógrafos e homófonos**: palavras com mesma grafia e pronúncia. Por exemplo, "grama" pode significar uma unidade de massa ou planta rasteira.

Homonímia e polissemia são um problema maior em folksonomias porque geram ambiguidade, pois o uso de uma *tag* pode apresentar duplicidade de sentido levando-se em conta fatores polissêmicos ou homônimos. Embora a homonímia e polissemia não sejam contemplados como foco deste trabalho, optou-se por apresentar estes conceitos em virtude de serem comumente mencionados na literatura como problemas de ordem semântica que afetam a qualidade da recuperação.

2.3 Dicionários Linguísticos

Principalmente no campo das folksonomias, nas quais *tags* são criadas sem levar em conta qualquer questão de ordem semântica, fontes de conhecimento externas podem ser utilizadas para atender diferentes propósitos como, por exemplo, inferir a relação semântica correspondente entre *tags*. Em particular, são destacadas brevemente duas ferramentas de mapeamento semântico utilizadas neste trabalho:

- **WordNet**: é um dicionário léxico eletrônico que cobre a maioria dos substantivos, adjetivos, verbos e advérbios da língua inglesa (MILLER, 1995). A informação no WordNet é organizada em agrupamentos lógicos denominados *synsets*, no qual cada *synset* reúne um conjunto de palavras sinônimas relacionadas a um conceito. Além disso, os *synsets* são interligados entre si por diferentes *links* semânticos que expressam diferentes tipos de relações, as quais variam de acordo com a categoria sintática. A relação *is-a* conecta um hipônimo (*synset* mais específico) a um hiperônimo (*synset* mais geral), definindo desta forma uma relação de hierarquia. Outros tipos de relações semânticas que podem ser estabelecidas entre palavras no WordNet incluem: sinonímia (relação básica, visto que os *synsets* são formados por sinônimos), antonímia (relação de oposição), meronímia (*Part-of*) e seu inverso holonímia (*Has-part*, ou seja, o todo em que algo é parte). Relações semânticas podem ser consultadas no WordNet utilizando a API (*Application Programming Interface*) JAWS[12] (*Java API for WordNet Searching*) em um arquivo de programa na linguagem Java. Neste trabalho, o WordNet foi utilizado para auxiliar no processo de rotulação automática de instâncias para sinonímia e hiperonímia/hiponímia;

- **Projeto ConceptNet5**[13]: é uma base de conhecimento de acesso livre em que os conceitos são estruturados como uma rede semântica, para que o conhecimento seja utilizado por computadores. A rede semântica do ConceptNet5 é representada por um hipergrafo, no qual suas assertivas são representadas por arestas que conectam seus nós, ou simplesmente conceitos (e.g., uma palavra ou frase curta da linguagem natural). A conexão entre nós é rotulada por relações interlinguísticas aplicáveis a

[12] <http://lyle.smu.edu/~tspell/jaws/index.html>
[13] <http://conceptnet5.media.mit.edu/>

conceitos de diferentes idiomas. Exemplos de relações semânticas suportadas pelo ConceptNet incluem: *IsA* (sentido mais específico para um mais geral, ou seja, de um termo hipônimo para um hiperônimo), *Instance-Of* (um subtipo de *IsA* em que um $conceito_1$ é tido como uma única instância do $conceito_2$), *Part-Of* (meronímia), *Translation-Of* (tradução de uma palavra), *Synonym* (sinônimo), entre outros. O acesso à base de conhecimento do ConceptNet pode ser feito via *website* ou usando a API ConceptNet Web 5.4[14]. A API segue o padrão REST (*REpresentational State Transfer*), o qual usa solicitações HTTP (*HyperText Transfer Protocol*) simples para interagir com o servidor. Utilizou-se a linguagem Java para construir um programa que verifica a existência de relações específicas entre dois conceitos no ConceptNet por intermédio dos métodos *lookup* e *search* da API, utilizando URL parametrizadas. Alternativamente, é possível fazer o *download* de uma cópia dos dados do ConceptNet e acessá-lo *offline* utilizando o SQLite[15], uma biblioteca que implementa um banco de dados SQL (*Structured Query Language*) embutido.

2.4 Folksonomias

Nesta seção, é apresentada uma definição formal de folksonomia introduzida por Hotho et al. (2006), a qual será utilizada ao longo do documento.

Definição: Uma folksonomia $\mathbb{F}$ é definida como uma tupla $\mathbb{F} := (U, T, R, Y)$ com U, T, R, e Y representando, respectivamente, os conjuntos de usuários, *tags*, recursos e relação ternária entre eles. Cada tripla $(u, t, r) \in Y$ designa uma atribuição da *tag* $t \in T$ ao recurso $r \in R$ pelo usuário $u \in U$. Um *post* corresponde ao conjunto de atribuições de *tags* de um usuário para um determinado recurso, ou seja, uma tripla $(u, T_{u,r}, r)$ com $u \in U$, $r \in R$, e um conjunto não-vazio $T_{u,r} := \{t \in T \mid (u, t, r) \in Y\}$. O conjunto de associação $A_R(t)$ de uma *tag* $t \in T$ é o conjunto de recursos em R associado a *tag* t. Logo, por definição, $A_R(t) \subseteq R$. Por fim, o conjunto

$$T_r := \{t \in T \mid \exists u \in U \colon (u, t, r) \in Y\}$$

[14]Disponível em <https://github.com/commonsense/conceptnet5/wiki/API>.
[15]<https://www.sqlite.org/>

corresponde à união de todas as *tags* atribuídas pelos usuários que marcaram um determinado recurso $r \in R$.

A Figura 2.1 ilustra uma folksonomia com dois usuários, três *tags* e três recursos. O conjunto U é composto pelos usuários u_1 e u_2. O conjunto R contém os recursos r_1, r_2 e r_3. O conjunto T é constituído pelas *tags* t_1, t_2 e t_3. O conjunto Y de atribuição de *tags* é dado por $Y := \{(u_1, t_1, r_1), (u_1, t_2, r_1), (u_2, t_2, r_1), (u_2, t_2, r_2), (u_2, t_3, r_3)\}$. Como exemplo de *post* introduzido pelo usuário u_1 tem-se $(u_1, \{t_1, t_2\}, r_1)$. O conjunto de associação $A_R(t_2)$ compreende os recursos r_1 e r_2.

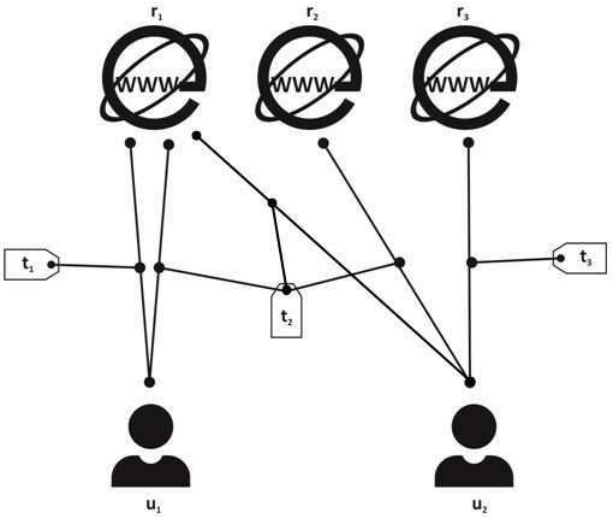

Figura 2.1: Representação visual de uma folksonomia.

Em bases de dados do mundo real, os usuários são normalmente representados por um ID exclusivo. As *tags* podem ser representadas por caracteres arbitrários enquanto que os recursos são dependentes da finalidade do sistema. Por exemplo, no BibSonomy os recursos são URL ou publicações armazenadas no formato BibTex[16], enquanto que no Flickr, os recursos são fotografias. Os recursos também são regularmente associados a um ID único.

2.5 Medidas de Similaridade/Distância Semântica

O uso de *tags* em folksonomias estabelece não só um processo de categorização individual, mas também um processo social de indexação. Wu e Zhou (2009) destacam que "*tags*

[16]Uma ferramenta de formatação de referências usada no sistema de preparação de documentos LaTex (<http://www.latex-project.org/>).

não estão associadas apenas ao recurso, mas relacionadas à pessoas". Isto significa que a *tag* nem sempre expressa apenas um tópico ou termo que indexa um assunto, mas também uma estrutura subjetiva de organização do conhecimento empregada pelo usuário. Descobrir conexões semânticas entre *tags* com base no comportamento de marcação dos usuários da folksonomia é uma prática amplamente adotada na literatura para diversas finalidades.

A literatura utiliza diferentes terminologias quando se trata do tema similaridade semântica. De acordo com Budanitsky e Hirst (2006), três principais tipos de medidas estão relacionadas ao referido tópico: parentesco semântico, similaridade semântica e distância semântica. É perceptível que a interpretação sobre a definição destas medidas não é tratada com clareza ou homogeneidade conceitual por diferentes autores. Optou-se por adotar a distinção tratada em Budanitsky e Hirst (2006) porque fornece esclarecimentos plausíveis e também é citada em outros trabalhos na área (CATTUTO et al., 2008; GRACIA; MENA, 2008; PATWARDHAN; BANERJEE; PEDERSEN, 2003).

- **Parentesco semântico** é descrito como uma conexão entre entidades em razão de uma relação semântica estabelecida ou visível (KADLEC, 2010). Uma entidade pode ser um conceito, palavra, *tag* ou outro componente textual. Alguns relacionamentos entre entidades cobertos pela noção de parentesco semântico incluem:

  - Holonímia: o todo pelo qual uma determinada palavra é parte. Por exemplo, *chapéu* é holonímia para *brim*;
  - Meronímia: sentido inverso de holonímia, palavra que designa uma parte do todo. Por exemplo, *tinta* é meronímia para *caneta*;
  - Antonímia: palavras com significados opostos como, por exemplo, *quente* e *frio*;
  - Qualquer tipo de relacionamento funcional ou associação frequente como, por exemplo, *lapis-papel*, *pinguim-Antártica* e *chuva-inundação*.

- O dicionário Houaiss (HOUAISS, 2009) define "similar" como algo semelhante, que possui características em comum. **Similaridade semântica** é um caso especial de parentesco semântico que diferencia-se por considerar relações léxicas mais específicas entre palavras. São considerados exemplos de similaridade semântica as relações léxicas de sinonímia e de hiperonímia/hiponímia (GRACIA; MENA,

2008). Em Resnik (1995) o autor afirma que, quando duas entidades são similares, elas também estão relacionadas. Para esclarecer o entendimento entre parentesco semântico e similaridade semântica, considere o seguinte exemplo: os termos *sanfona* e *forró* não são similares por completo, todavia, estão fortemente relacionados. Portanto, *sanfona* e *forró* constituem um caso de parentesco semântico. Por outro lado, *sanfona* e *bandoneon* são similares e bastante relacionados. Similares porque compartilham características em comum como, por exemplo, são instrumentos que possuem fole (recipiente de ar), podem ser diatônicos (teclas equivalentes ao do piano, tocadas com a mão direita) e possuem botões de baixo (botões acionados com a mão esquerda que emitem um som mais grave). Neste caso, tem-se um exemplo de similaridade semântica entre termos, visto que *bandoneon* pode ser visto como uma espécie de *sanfona*.

- **Distância semântica** é uma medida que quantifica o grau de proximidade ou distância entre dois conceitos, seja em termos de significado (similaridade semântica ou parentesco semântico) ou forma lexical (MOHAMMAD; HIRST, 2012). Dois conceitos são ditos distantes um do outro se sua similaridade ou parentesco é pequena, caso contrário, são considerados próximos.

As noções de similaridade e dissimilaridade entre termos são bastante difundidas principalmente no campo da RI, PLN e AM (LIU; ZHOU; ZHENG, 2007). Exemplos de sua aplicação incluem tarefas que envolvem mineração de texto (HOTHO; NÜRNBERGER; PAAß, 2005), recuperação de documentos (MANNING; RAGHAVAN; SCHÜTZE, 2008) e desambiguação de sentido de palavras (RESNIK, 1999; SANDERSON, 1994).

A similaridade entre dois objetos de dados quantifica numericamente o grau de semelhança entre esses objetos. Em oposição ao significado de similaridade, a dissimilaridade (distância) entre dois objetos determina o grau de diferença entre os objetos envolvidos. Seguindo a formalização definida por Gan, Ma e Wu (2007), considerando $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$ dois itens de dado no espaço n-dimensional pertencentes ao conjunto $\mathcal{X}$, com x_i e y_i denotando o i-ésimo atributo escalar, a medida de dissimilaridade d é uma função da forma

$$d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_0^+,$$

que, quando satisfaz as seguintes propriedades:

- **Simetria**: $d(x, y) = d(y, x)$;
- **Positiva-definida**: $d(x, y) \geq 0$ e $d(x, y) = 0 \Rightarrow x = y$;
- **Desigualdade triangular**: $d(x, z) \leq d(x, y) + d(y, z)$;

para quaisquer x, y e $z \in \mathcal{X}$, é chamada de métrica ou função de distância.

A medida de similaridade (ou parentesco) s é uma função da forma

$$s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_0^+,$$

que satisfaz as seguintes propriedades:

- **Simetria**: $s(x, y) = s(y, x)$;
- **Positiva-definida**: $s(x, y) \geq 0$ e $s(x, y) = 1 \Rightarrow x = y$;

para quaisquer $x \in \mathcal{X}$ e $y \in \mathcal{X}$.

As medidas de similaridade mais prestigiadas normalmente operam com valores no intervalo [0,1], em que 1 e 0 significam máxima e mínima similaridade (o oposto deve ser considerado para dissimilaridade), respectivamente. Se os valores de similaridade s e dissimilaridade d estão inseridos no intervalo [0,1], então a similaridade pode ser definida em função da dissimilaridade por meio da fórmula $s = 1 - d$, e vice-versa para dissimilaridade. Algumas vezes, as medidas de similaridade são definidas no intervalo [-1; 1] como, por exemplo, o coeficiente de *Pearson*.

Nas folksonomias, a noção de similaridade entre *tags* desempenha um papel essencial devido à sua aplicabilidade no suporte à navegação (LI et al., 2007), análise de ambiguidade e redundância de *tags* (GEMMELL et al., 2009; YEUNG; GIBBINS; SHADBOLT, 2007), expansão de consultas (ABBASI, 2011), recomendação de *tags* (WARTENA; BRUSSEE; WIBBELS, 2009) e aprendizado de ontologias (BENZ, 2007; TANG et al., 2009). Ontologia, na Ciência da Computação, é um termo técnico que designa a especificação formal de um conjunto de conceitos (vocabulário) inerentes a um domínio e do relacionamento (e.g., hiperonímia/hiponímia) existente entre eles, utilizada para definir a estrutura organizacional básica do conhecimento (GRUBER, 1993).

Medidas de similaridade/parentesco vêm sendo exploradas em diversas pesquisas na literatura com vistas à extração de semântica diretamente da estrutura da folksonomia, formada pela tríade *tag*, usuário e recurso. Especificamente para as tarefas de detecção de sinonímia e construção de hierarquia entre *tags*, muitos pesquisadores introduziram medidas para computar a similaridade entre *tags*. A Seção 4.2 apresenta, define e formula todas as medidas de similaridade/parentesco/distância utilizadas como atributos no processo de aprendizado supervisionado proposto neste trabalho.

2.6 Aprendizado de Máquina

O Aprendizado de Máquina é uma subárea da Inteligência Artificial (IA) que lida com métodos computacionais e estatísticos para adquirir conhecimento, habilidades e organizar o conhecimento existente de forma automática (MITCHELL, 1997b). O aprendizado é extraído diretamente dos dados, baseado em exemplos passados, sem que haja uma programação explícita. Isto é concretizado por meio de algoritmos e técnicas que permitem ao computador "aprender" e melhorar seu desempenho em uma determinada tarefa por meio da observação, análise e generalização de dados. A generalização é uma habilidade dos algoritmos de AM em executar com precisão a aprendizagem de novas tarefas logo após ser submetido ao reconhecimento de um conjunto de dados de aprendizado (MONARD; BARANAUSKAS, 2003).

O uso de AM está presente hoje em diversos tipos de aplicações comerciais. Exemplos da aplicação de AM estão presentes em: reconhecimento de face, reconhecimento de escrita manual, recomendação de produtos na Amazon.com (HARRINGTON, 2012), reconhecimento de fala e aprendizagem de jogadas de xadrez (MITCHELL, 1997b; WITTEN; FRANK, 2011).

A decisão pela utilização de AM em problemas computacionais requer conhecimento avançado por parte do profissional em computação. Sua concepção necessita de uma formulação bem especificada da tarefa, familiaridade com as técnicas disponíveis, conhecimento acerca dos algoritmos de aprendizado que melhor se adequam ao problema, entendimento sobre como preparar um conjunto de dados para treinamento e teste, escolha das métricas de desempenho a serem aplicadas, interpretação de dados e realização de

ajustes.

A literatura é extensa e não trivial para tratar com detalhes os requisitos citados no parágrafo anterior. Harrington (2012), Mitchell (1997b), Witten e Frank (2011) oferecem uma cobertura teórica completa acerca do tema para maior aprofundamento. Nesta seção, o principal objetivo é apresentar um entendimento geral sobre AM e outros componentes pertencentes ao domínio que estão diretamente conectados ao trabalho. As decisões metodológicas inerentes à abordagem de AM objeto deste trabalho, incluindo a preparação e seleção dos dados para treinamento, algoritmo de aprendizado e avaliação dos resultados de saída, estão definidas nos Capítulos 4 e 5, que tratam sobre o delineamento e avaliação da solução proposta.

2.6.1 O Processo de Extração de Conhecimento

A Extração de Conhecimento, também conhecida como processo KDD, um acrônimo para *Knowledge Discovery in Databases*, é um processo de análise de dados de baixo nível para extração de informações potencialmente úteis, implícitas e desconhecidas a partir de um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Neste processo, o conhecimento obtido é compactado e transformado em uma estrutura compreensível para as pessoas.

De forma sumarizada, o processo KDD (Figura 2.2) é constituído pelas seguintes etapas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; GOEBEL; GRUENWALD, 1999; HAN; KAMBER, 2006):

1. **Compreensão do domínio da aplicação**: planejamento do objetivo que se deseja atingir com a mineração de dados;
2. **Seleção de dados**: recuperação de dados relevantes para a análise;
3. **Pré-processamento**:
    - **Limpeza de dados**: remoção de ruídos e dados inconsistentes;
    - **Integração de dados**: combinação de múltiplas fontes de dados;
4. **Transformação de dados**: conversão dos dados para um formato apropriado para a mineração;

5. **Mineração de dados**: processo essencial, no qual algoritmos específicos são aplicados com o intuito de extrair padrões nos dados;

6. **Interpretação e avaliação dos resultados**: visualização, análise e consolidação do conhecimento.

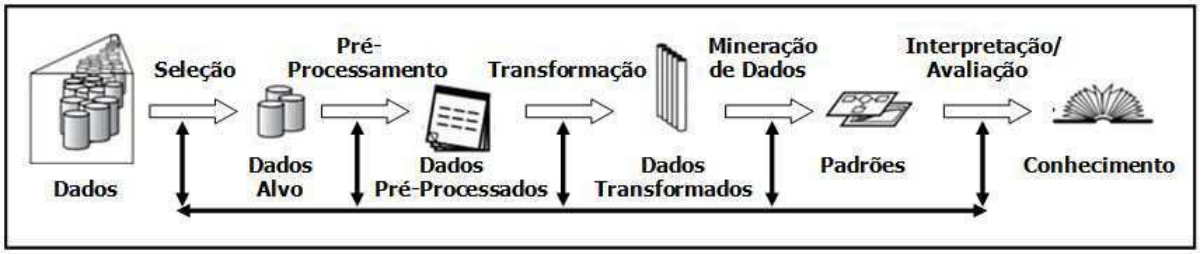

Figura 2.2: Visão geral dos passos constituintes do processo de KDD.
Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

O processo KDD é interativo e iterativo, podendo conter retroalimentação entre as etapas do processo, com muitas decisões tomadas pelo usuário (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Dentre as etapas da KDD, o foco deste trabalho está centrado na etapa de Mineração de Dados. Entretanto, as demais etapas não são descartadas e exercem a sua importância dentro do processo.

É importante ressaltar que a AM fornece a base técnica para a Mineração de Dados porque, além de ser um campo multidisciplinar, seus algoritmos de aprendizado são utilizados rotineiramente para revelar conhecimento valioso. Segundo Mitchell (1997a), *"a Mineração de Dados envolve a aplicação de métodos de AM em um histórico de dados para aprimorar futuras decisões"*. Alpaydin (2010) complementa que a aplicação de métodos de AM em grandes bases de dados é chamada de Mineração de Dados.

Existem diferentes técnicas para a realização da tarefa de Mineração de Dados. As mais populares são: classificação, regressão, agrupamento, associação e sumarização (HAN; KAMBER, 2006). Diversos tipos de algoritmos estão disponíveis de acordo com a técnica de mineração de dados escolhida. Uma leitura mais aprofundada sobre os tipos de algoritmos existentes para a tarefa de Mineração de Dados pode ser feita em Alpaydin (2010), Han e Kamber (2006), Mitchell (1997b) e Witten e Frank (2011). Dentre as técnicas citadas, duas foram utilizadas neste trabalho. A técnica de classificação (cf. Seção 2.6.2) é a base para realização das predições semânticas do tipo sinonímia e hiperonímia/hiponímia entre duas

tags. A técnica de associação (cf. Seção 2.6.3) é utilizada para identificar potenciais regras candidatas a uma relação de hiperonímia/hiponímia.

2.6.2 Classificação

No AM, o aprendizado do tipo **indutivo** permite extrair conclusões genéricas (regras) sobre um conjunto particular de exemplos; uma espécie de raciocínio que parte do mais específico para o mais geral. Os métodos pertencentes a esta categoria são mais populares e muito utilizados para derivar conhecimento novo, mesmo sabendo que se trata de um método desafiador tendo em vista que não existe garantia de que esse conhecimento seja verdadeiro, o que dificulta a análise dos resultados obtidos (BATISTA, 2003; MONARD; BARANAUSKAS, 2003).

O aprendizado **indutivo** é categorizado em **supervisionado** e **não-supervisionado** (MONARD; BARANAUSKAS, 2003). No aprendizado **supervisionado**, as classes dos exemplos[17] de treinamento são conhecidas. Uma classe é representada por um rótulo que nomeia um determinado grupo de dados. No aprendizado **não-supervisionado**, como o próprio nome sugere, um conjunto de exemplos é fornecido como entrada, mas sem informação explícita sobre a classe dos exemplos.

Uma formulação padrão da abordagem de aprendizado supervisionado é o problema de classificação, foco deste trabalho. A tarefa de classificação consiste em produzir um modelo (ou função de inferência) a partir de um conjunto de exemplos de treinamento, que mapeia novos exemplos para uma determinado conjunto de classes. Este tipo de abordagem é utilizado para problemas preditivos, ou seja, tarefas em que são realizadas induções nos dados a fim de gerar predições (HAN; KAMBER, 2006). Por exemplo, uma empresa de cosméticos gostaria de prever quais de seus clientes estão propensos a realizar mais compras se for concedido um cartão de afinidade. Neste caso, o conjunto de treinamento teria informações relevantes sobre clientes que usaram um cartão de afinidade no passado e a tarefa de previsão consistiria em determinar se um novo cliente é candidato ou não a receber um cartão de afinidade.

A tarefa de classificação pode ser definida como binária (existência de apenas duas classes) ou multiclasse (mais de duas classes são consideradas). Na tarefa de classificação

[17] Alguns autores utilizam os termos *observação* ou *instância* em lugar de *exemplo*.

binária, adotada neste trabalho, geralmente se está interessado em saber se uma instância pertence a uma determinada classe ou não. Por exemplo, em um cenário bancário, um sistema de aprendizado necessitaria de subsídios, baseado no histórico de movimentação do cliente, para atender a solicitação de um empréstimo ou negá-lo. Os casos de clientes que tiveram empréstimos atendidos são denominados de *exemplos positivos*, e os demais casos de *exemplos negativos*, ou seja, contra-exemplos de empréstimos aprovados. Por questão de simplicidade, será utilizado o símbolo (+) para identificar os exemplos positivos e o símbolo (-) para identificar os exemplos negativos. Neste caso, o aprendizado de classe consiste idealmente em encontrar descrições (propriedades) compartilhadas por todos os exemplos positivos e nenhum dos exemplos negativos.

A fronteira de decisão ou separação entre os dois exemplos é formada pelas propriedades em comum observadas entre todos os exemplos negativos e todos os exemplos positivos, conforme ilustrado na Figura 2.3. A fronteira de decisão é determinada por uma função de inferência, também chamada de *classificador*, que separa as instâncias de uma classe das demais (ALPAYDIN, 2010).

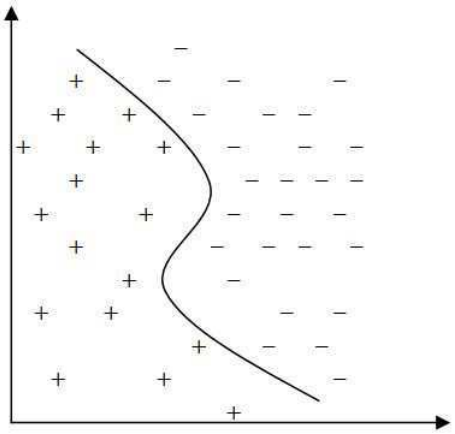

Figura 2.3: Fronteira de decisão entre classes.

Para gerar um classificador, é necessário treinar um algoritmo de classificação[18] de modo que o mesmo aprenda com os dados. O treinamento é feito a partir de um conjunto de dados denominado conjunto de treinamento, o qual contém exemplos qualificados denominados de instâncias cujas classes são conhecidas. Além da classe, cada instância é constituída por um conjunto de atributos representados por medidas mensuráveis que

[18] A escolha do algoritmo é uma decisão de projeto.

caracterizam um fenômeno observado. A escolha dos atributos em um problema de classificação é fundamental para que a predição seja bem sucedida. O conjunto de teste é um conjunto independente que contém exemplos exclusivos, os quais não foram utilizados para treinamento, para fins de avaliação do desempenho do classificador. Neste conjunto, o rótulo de classe de cada exemplo não é fornecido ao classificador, pois cabe ao mesmo decidir a qual classe cada exemplo pertence. Uma definição formal do problema de classificação é feita concomitantemente com a formulação do problema de detecção de relações semânticas em pares de *tags* na Seção 4.1.

O desempenho do classificador é determinado pelos diferentes tipos de erros e acertos encontrados na tarefa de classificação. Estes podem ser sintetizados em uma tabela informativa denominada matriz de confusão. A Tabela 2.1 exibe um modelo de matriz de confusão para um problema de classificação binária em que:

- a é o número de predições **corretas** de exemplos **positivos**;
- b é o número de predições **incorretas** de exemplos **positivos**;
- c é o número de predições **incorretas** de exemplos **negativos**;
- d é o número de predições **corretas** de exemplos **negativos**.

Tabela 2.1: Componentes elementares de uma matriz de confusão.

	Predição (+)	Predição (-)
Classe (+)	Verdadeiro positivo (a)	Falso negativo (b)
Classe (-)	Falso positivo (c)	Verdadeiro negativo (d)

A partir da matriz de confusão, é possível determinar a taxa de erro e acurácia do classificador por intermédio das Equações 2.1 e 2.2. A taxa de erro é a fração de predições incorretas do classificador para um conjunto de teste. Consequentemente, a acurácia é a fração de predições corretas realizada pelo classificador.

$$\textit{Erro} = \frac{c+b}{a+b+c+d}. \tag{2.1}$$

$$\textit{Acurácia} = \frac{a+d}{a+b+c+d}. \tag{2.2}$$

Existem diversos algoritmos de AM na literatura apropriados para classificação, dentre os quais destacam-se: árvore de decisão, Naive Bayes, *Suport Vector Machines* (SVM), AdaBoost, regressão logística, redes neurais, Random Forest, entre outros (ALPAYDIN, 2010; DOMINGOS, 2012; HAN; KAMBER, 2006; MITCHELL, 1997b). O desempenho de um algoritmo de AM pode variar de acordo com as características do problema.

2.6.3 Regras de Associação

Mineração de Regras de Associação é uma técnica popular da Mineração de Dados que visa encontrar padrões de relacionamentos significantes entre itens em grandes conjuntos de dados, os quais geralmente não são aparentes sob o ponto de vista superficial (AGRAWAL; IMIELINSKI; SWAMI, 1993; HARRINGTON, 2012; KOTSIANTIS; KANELLOPOULOS, 2006). Por exemplo, a partir de uma base de dados que armazena registros de itens comprados por clientes em uma padaria, a mineração de regras de associação poderia gerar a regra $\{$*pão,ovo*$\} \rightarrow \{manteiga\}$, a qual indica que "quem compra pão e ovo com um determinado grau de certeza também compra manteiga". Dá-se o nome de *conjunto de itens frequentes* ao conjunto de itens que normalmente ocorrem juntos em uma base de dados.

Quando aplicadas a uma base de dados constituída por inúmeras transações[19], as regras de associação possibilitam encontrar regras do tipo $X \rightarrow Y$, ou seja, conjuntos de itens frequentes no banco de dados que contém X (antecedente) e que tendem a conter Y (consequente). Tanto o antecedente quanto o consequente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens.

Parâmetros de suporte e confiança são medidas estimadas para selecionar regras candidatas a mais importantes. Suporte é uma indicação de frequência com que um conjunto de itens aparece na base de dados. Para um conjunto de itens X, o suporte $sup(X)$ é definido como a fração de transações da base de dados que contém X (Equação 2.3) ou simplesmente a quantidade de transações na qual se observa X (Equação 2.4). O suporte de uma regra de associação $X \rightarrow Y$ é dado por $Sup(X \cup Y)$.

[19]Relação de itens registrados em uma operação específica.

$$Sup(X) = \frac{\#X}{\#\textit{transações}} \tag{2.3}$$

$$Sup(X) = \#X \tag{2.4}$$

A confiança é uma medida que quantifica a confiabilidade de uma regra. Expressa em termos de porcentagem, a confiança estima a probabilidade de X ocorrer quando Y ocorre. Sob o ponto de vista estatístico, confiança é a probabilidade condicional de X dada a condição Y (Equação 2.5).

$$\textit{Confiança}(X \rightarrow Y) = P(X|Y) = \frac{Sup(X \cup Y)}{Sup(X)} \tag{2.5}$$

O problema da mineração de regras de associação consiste, então, em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo e uma confiança mínima especificados pelo usuário. Agrawal, Imielinski e Swami (1993) ressaltam que suporte não deve ser confundido com confiança. Enquanto a confiança é uma medida de força da associação de itens, o suporte corresponde à uma significância estatística.

Neste trabalho, projeta-se uma folksonomia para uma base de transações com o intuito de aplicar a técnica de mineração de regras de associação, baseando-se na mesma ideia empregada por Marinho, Buza e Schmidt-Thieme (2008) e Solskinnsbakk e Gulla (2010) de minerar conjuntos de itens frequentes de tamanho 2 no espaço de *tags*, ou seja, associações entre duas *tags* do tipo $t \rightarrow t'$, para um suporte e confiança mínimos pré-estimados. As transações são compostas por *tags* usadas em um *post* para anotar um determinado recurso. O suporte da regra é determinado pela contagem de coocorrência entre duas tags t e t'.

2.6.4 Desbalanceamento de Classes

À medida em que as técnicas de AM são aplicadas em problemas reais, novos temas, até então não previstos pela comunidade de AM, naturalmente surgem e abrem rumo para discussões. Os dados do mundo real são normalmente desbalanceados e estão presentes em diferentes domínios. O desbalanceamento de classes acontece quando uma das classes

(classe majoritária) pode conter muito mais exemplos do que outra (classe minoritária) na base de dados (GU et al., 2008). Esta situação é conhecida na literatura como **problema de desbalanceamento de classes** e muitas vezes é vista como um obstáculo para a indução de bons classificadores pelos algoritmos de AM, além de tratar-se de uma tarefa de difícil resolução (BATISTA; PRATI; MONARD, 2004; DRUMMOND; HOLTE, 2005).

Exemplos típicos de aplicações cujo conjunto de dados apresenta a característica do desbalanceamento de dados são: detecção fraudulenta de chamadas telefônicas (FAWCETT; PROVOST, 1996) e em transações de cartões de crédito (STOLFO et al., 1997). Tomando como referência esses dois exemplos, o desbalanceamento é natural porque o número de transações legítimas é muito maior do que o número de transações fraudulentas. Outros exemplos de cenários que apresentam dados desbalanceados são: detecção de derramamento de óleo na superfície do mar a partir de imagens de satélite (KUBAT; HOLTE; MATWIN, 1998), análise de risco em seguradoras (PEDNAULT; ROSEN; APTE, 2000) e diagnóstico médico de doenças cardiovasculares (MENA; GONZALEZ, 2006).

Além dos exemplos citados, o presente trabalho também é afetado pelo problema de desbalanceamento de classes, pois o número de pares de *tags* que de fato expressam uma relação semântica específica do tipo sinonímia ou hiperonímia/hiponímia, é muito menor do que a quantidade de pares de *tags* que não expressam as relações semânticas citadas.

Os algoritmos tradicionais de AM não apresentam bom desempenho quando aprendem a partir de dados desbalanceados. Isto porque a função de classificação gerada tende a classificar todos os dados como sendo da classe majoritária, o que normalmente é a classe de menor interesse (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006). Tal fenômeno é explicado pelo fato de que os algoritmos de AM direcionam seus esforços visando a redução da taxa de erro geral, desconsiderando as diferenças entre os tipos de erro de classificação por julgá-las igualmente importantes (GU et al., 2008). Para alcançar maior eficiência na classificação, o ideal é que o conjunto de treinamento para ambas as classes esteja balanceado em número, ou seja, as instâncias de treinamento de todas as classes devem ser numericamente iguais (SPILIOPOULOS; VOUROS; KARKALETSIS, 2010)

O problema de desbalanceamento de classes se torna ainda mais crítico quando o custo da classificação incorreta da classe minoritária é muito maior do que o custo da classificação incorreta da classe majoritária. Isto pode ser bem exemplificado em bases de dados de

aplicações do domínio da medicina. Por exemplo, o risco de não diagnosticar a existência de câncer no pulmão e brônquios em um paciente (falso negativo da classe minoritária) é mais grave do que classificá-lo incorretamente em pacientes sadios (falso positivo da classe majoritária).

Atualmente, existem soluções na literatura para lidar especialmente com esse tipo de problema em nível de dados, cujo objetivo é modificar a distribuição das classes para uma distribuição mais equilibrada. Basicamente, as técnicas se enquadram em duas abordagens de amostragem: *oversampling*, na qual o balanceamento no conjunto de treinamento é realizado por meio da replicação randômica de instâncias pertencentes à classe minoritária, e *undersampling*, na qual são eliminadas aleatoriamente instâncias pertencentes à classe majoritária. Ambas as abordagens atuam na fase de pré-processamento de dados, ou seja, são aplicadas para fins de treinamento em uma etapa anterior à de extração de conhecimento (BATISTA, 2003). As abordagens *undersampling* e *oversampling* possuem suas limitações:

- *Oversampling* pode aumentar a probabilidade de ocorrer *overfitting*, visto que a maioria dos métodos *oversampling* cria cópias exatas dos exemplos pertencentes à classe minoritária (BATISTA, 2003; CHAWLA, 2005; GU et al., 2008);

- *Undersampling* pode eliminar instâncias potencialmente úteis que poderiam ser importantes para o classificador.

O termo *overfitting*, citado como limitação da abordagem *oversampling*, é um termo utilizado no AM para sinalizar que o modelo de aprendizagem se ajusta ao conjunto limitado de pontos de dados do conjunto de treinamento (baixa taxa de erro para treino), ou seja, o modelo é construído para representar perfeitamente apenas o conjunto de treinamento e não generaliza bem para exemplos futuros. Logo, se o modelo for aplicado a um conjunto de teste com dados até então desconhecidos pelo classificador, o desempenho será insatisfatório (MITCHELL, 1997b). Por exemplo, considere que uma quantidade de exemplos é dividida em conjuntos de treino e teste. O modelo é criado usando o conjunto de treino como entrada. Quanto mais informações o modelo contiver acerca do conjunto de treino, mais haverá uma degradação da sua habilidade de generalização quando submetido

ao arquivo de teste. Dessa forma, um classificador simbólico pode construir regras que são aparentemente precisas, mas que na verdade cobrem um único exemplo replicado.

Os métodos Tomek Link (TOMEK, 1976), *Condensed Nearest Neighbor Rule* - CNN (HART, 1968) e *One-Sided Selection* (OSS) (KUBAT; MATWIN, 1967) são exemplos de métodos precursores mais complexos que realizam a re-amostragem usando o conceito *undersampling*. Tomek Link é um método aprimorado que remove apenas exemplos da classe majoritária identificados como ruído, ou seja, exemplos encontrados no lado errado da borda de decisão, ou linha de borda, isto é, exemplos próximos à borda de decisão pouco confiáveis visto que uma pequena quantidade de ruído em um dos atributos pode mover esses exemplos para o lado errado da borda de decisão. CNN elimina exemplos da classe majoritária que estão distantes da borda de decisão, uma vez que esses exemplos podem ser considerados irrelevantes para o aprendizado. O método OSS efetua uma remoção cuidadosa dos exemplos da classe majoritária, preservando os exemplos da classe minoritária. Basicamente, o método OSS consiste em detectar e eliminar casos não confiáveis considerados como ruído, próximos à borda e redundantes. Exemplos redundantes são aqueles que podem ser representados por outros presentes no conjunto de treinamento.

SMOTE (*Synthetic Minority Over-sampling TEchnique*) é um método *oversampling* aprimorado no qual a classe minoritária é incrementada ao se criar novos exemplos sintéticos desta classe por meio da interpolação entre vários exemplos da classe minoritária que residem próximos uns dos outros (CHAWLA et al., 2002) . Outros métodos são derivados da combinação das técnicas *undersampling* e *oversampling*, originando dessa forma uma abordagem híbrida como efetuado por Batista, Prati e Monard (2004) ao propor o SMOTE + Tomek Link.

Neste trabalho, foram avaliados os métodos *random undersampling*, *random oversampling*, CNN, Tomek Link e SMOTE para obter distribuições de dados balanceadas artificialmente nos experimentos e, consequentemente, perseguir um melhor desempenho na taxa de aprendizado da classe minoritária.

2.6.5 Aprendizado Sensível ao Custo

Assim como ressaltado na Seção 2.6.4, a maioria dos algoritmos de classificação normalmente tentam minimizar a porcentagem de predições incorretas para as classes

envolvidas. Em particular, os algoritmos assumem que todos os erros de classificação possuem o mesmo custo. Entretanto, para muitos domínios de aplicações, a classificação incorreta da classe minoritária é mais custosa do que classificar incorretamente exemplos da classe majoritária. Para esses domínios é possível utilizar o sistema de Aprendizado Sensível ao Custo (ASC), o qual tem o objetivo de minimizar o custo total da classificação ao invés da taxa de erro geral da classificação (LING; SHENG, 2008).

O ASC associa um custo para classificação incorreta nas diferentes classes. Neste caso, uma matriz de custo codifica as penalidades ao classificar um exemplo de uma classe em outra. Assumindo um problema de classificação binária, a Tabela 2.2 adaptada de Elkan (2001) ilustra um exemplo de matriz de custo para os casos Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN), enquanto que C(i,j) representa o custo de classificar incorretamente uma instância de sua classe real j para a classe predita i (considerar 0 para positivo e 1 para negativo).

Tabela 2.2: Exemplo de matriz de custo para um problema de classificação binária.

	Classe (+)	Classe (-)
Predição (+)	C(0,0) ou TP	C(1,0) ou FN
Predição (-)	C(0,1) ou FP	C(1,1) ou VN

A ASC também é uma solução indicada para lidar com o problema de distribuições de dados severamente desbalanceadas. Neste caso, normalmente se está interessado em reconhecer os casos positivos (raros) ao invés dos negativos, portanto o custo para a classificação incorreta de uma instância positiva em negativa é maior do que o custo de classificar incorretamente uma instância negativa em positiva. Isso significa que o valor de C(1,0) para FN é normalmente maior do que C(0,1), ou seja, FP. Desta forma, a classe minoritária se torna mais custosa e espera-se que ela seja melhor classificada. Por outro lado, quando as instâncias positivas ou negativas são preditas corretamente para os elementos da diagonal principal, não existe custo envolvido, ou seja, C(0,0)=C(1,1)=0.

Para aprender usando conjuntos de classes desbalanceadas, é necessário treinar um classificador sensível ao custo, ou seja, um classificador otimizado para levar em conta a matriz de custo ao desempenhar a tarefa de classificação. A maior dificuldade na implementação do ASC é que os custos de classificação são frequentemente desconhecidos

e deve assumir um valor constante (WEISS, 2004). Parte desta dificuldade se deve ao fato de que a mensuração dos custos muitas vezes depende de múltiplas considerações acerca do problema que não são fáceis de serem determinadas. Neste trabalho, a distribuição de custo adequada para cada classe foi determinada empiricamente, a fim de identificar o ajuste que melhor beneficia o desempenho de classificação aplicado ao problema de detecção de relações de subordinação.

2.7 Considerações Finais

Neste capítulo, apresentou-se um entendimento geral acerca dos principais tópicos pertencentes ao domínio desta pesquisa. No cenário das folksonomias, a descoberta de relações semânticas entre *tags* tanto em um nível geral (indicação de parentesco) ou específico (e.g., sinonímia e hiperonímia/hiponímia) é um requisito importante para aprimorar o desempenho de diversos tipos de aplicações, dentre as quais a expansão de consultas, suporte à navegação e recomendação de *tags*. Normalmente, a literatura foca o uso de medidas de similaridade semântica para identificar possíveis conexões semânticas entre *tags* com base no comportamento de marcações dos usuários de uma folksonomia.

O campo da AM possui um conjunto de métodos e procedimentos para extrair conhecimento diretamente dos dados de forma automática. A partir do conhecimento obtido, uma função de inferência pode ser usada para realizar predições e determinar a classe de novas observações, problema este tipicamente conhecido como classificação. Para alcançar este objetivo, a etapa de Mineração de Dados do processo de KDD é então executada para que sejam aplicados algoritmos de AM específicos com o intuito de aprimorar futuras decisões.

Os dicionários linguísticos eletrônicos são comumente utilizados para prover suporte semântico aos sistemas que processam texto(s) em linguagem natural. Diversos tipos de aplicações fazem uso desse recurso, o que o torna ideal para tratar problemas relacionados à desambiguação de sentido das palavras, anotação semântica e recuperação da informação.

Com base no suporte provido pela técnica de AM, esta tese de doutorado utiliza medidas de similaridade/parentesco/distância semântica como atributos para capturar diferentes aspectos semânticos inerentes à especificidade de sinonímia e hiperonímia/hiponímia. Estas medidas são apoiadas por trabalhos relacionados os quais são discutidos no próximo capítulo.

Capítulo 3

Trabalhos Relacionados

Este capítulo aborda os trabalhos relacionados encontrados na literatura que exploram de forma direta ou indireta a estimação de similaridade semântica entre *tags* de folksonomias. Especialmente, são destacadas as medidas de similaridade/parentesco/distância utilizadas como heurísticas para identificar as relações semânticas de sinonímia e hiperonímia/hiponímia, bem como as estratégias utilizadas para a construção de hierarquias entre *tags*. Uma relação semântica do tipo hiperonímia/hiponímia é abordada na literatura com diferentes denominações, tais como relações hierárquicas, relações de subordinação e relações super/sub conceito. Neste trabalho, adota-se a convenção "relação de subordinação" para se referir às relações no sentido hiperonímia/hiponímia entre dois conceitos.

3.1 Detecção de Sinonímia

Várias pesquisas na área têm concentrado seus esforços na tentativa de superar ou amenizar os problemas causados pelo livre uso de *tags* em folksonomias, o qual afeta a recuperação de recursos relevantes. Em geral, observa-se que a aplicação de medidas de similaridade/parentesco/distância no espaço de *tags* vêm sendo frequentemente empregadas como heurísticas para identificar automaticamente relações semânticas entre *tags*. Mousselly-Sergieh et al. (2014) propõem uma abordagem para medir parentesco entre *tags* com base em estatísticas de coocorrência, na qual *tags* relacionadas são determinadas de acordo com a distância entre suas respectivas distribuições.

Wartena (2010) computa a similaridade entre *tags* usando a noção de que *"tags são*

semanticamente relacionadas quando ocorrem em contextos semelhantes". Para o autor, um contexto é definido por intermédio do perfil de distribuição de *tags* por todos os recursos. Esta teoria, conhecida como hipótese de distribuição, foi derivada da RI para definir uma medida de similaridade semântica que introduz a distribuição de coocorrência como formalização para o contexto de uma *tag*. Portanto, a similaridade entre os contextos de duas *tags* é um indicador de similaridade semântica.

Os autores Clements, Vries e Reinders (2008) postulam que *tags* normalmente aplicadas ao mesmo recurso tendem a ser sinônimas, ainda que as *tags* sejam fornecidas por diferentes usuários. Esta hipótese é exemplificada usando o par de *tags* `color` (inglês americano) e `colour` (inglês britânico). A heurística é embasada na hipótese de que grupos de usuários de diferentes idiomas empregam os termos que têm o costume de usar e, naturalmente, ambas as *tags* vão despontar quando aumentarem a frequência de coocorrência. A similaridade entre *tags* é medida pelo cálculo da correlação de Pearson em vetores de perfil de distribuição de usuário-*tag* e recurso-*tag*. Visto que o artigo foi publicado como *short paper*, não foram disponibilizados detalhes acerca da experimentação e do método de avaliação.

Solskinnsbakk e Gulla (2011) combinaram distância de edição, para identificar similaridade sintática (variações morfológicas), com a medida do cosseno para reconhecer sinônimos (ou quase sinônimos) em folksonomias. Os resultados apresentados mostraram que o cosseno da similaridade é capaz de reconhecer tipos mais complexos de relações semânticas (e.g., associações e abstrações, que são generalização e especialização de *tags*), mas não necessariamente sinônimos.

No sistema TagPlus (LEE; YONG, 2007), a busca por imagens no Flickr é enriquecida ao permitir que uma *tag* de entrada tenha seus respectivos sinônimos e homônimos recuperados por meio de acesso ao dicionário WordNet. Antes de exibir o resultado da consulta, a interface do TagPlus recupera e exibe todos os sentidos da *tag* e seus respectivos conjuntos de sinônimos, para que o usuário colabore com o processo de desambiguação. Entretanto, imagens indesejadas podem ser recuperadas devido à ausência de controle de homonímia. Isto acontece porque, no momento de inserção da imagem no Flickr, os usuários não distinguem a semântica das *tags*. Deste modo, torna-se difícil saber a qual domínio a *tag* pertence. Embora o WordNet exerça papel fundamental na construção do processo de consulta do TagPlus, obviamente serão descartadas *tags* relevantes no processo de

enriquecimento da busca em virtude da existência de oscilações na grafia de uma mesma palavra.

Várias medidas de similaridade/parentesco foram analisadas por Cattuto et al. (2008) utilizando uma base de dados do Delicious com o intuito de expor, sob o ponto de vista semântico, o tipo de relação que elas estabelecem. Os resultados experimentais indicaram que a medida do cosseno fundamentada na contagem de coocorrência *tag*-*tag* (Similaridade do Contexto de *Tags*) e *tag*-recurso (Similaridade do Contexto de Recursos) pode ser usada para descobrir relações de sinonímia. Benz (2007) adotou a distância de edição e um modelo estatístico de sinonímia como métricas para detectar sinônimos, em uma das etapas de um algoritmo de aprendizado de ontologias proposto pelo autor.

Para avaliar o impacto que a ambiguidade e redundância de *tags* causam na eficácia de algoritmos de recomendação de *tags*, Gemmell et al. (2009) empregaram uma abordagem focalizada no uso de *clusters* para agrupar recursos e *tags* em partições coesas, com base no princípio de coocorrência. Os *clusters* de recursos e *tags* são criados por intermédio da aplicação do algoritmo K-Means (HARRINGTON, 2012). Os recursos são modelados sob a forma de um vetor recurso-*tag*, enquanto que as *tags* têm seu vetor modelado na forma *tag*-recurso. Ambos utilizam como peso para os vetores a frequência do termo. Para um par recurso-*tag*, o peso corresponde à quantidade de vezes que um recurso foi anotado com a *tag*. O raciocínio é análogo para *tag*-recurso. A similaridade entre vetores de *tags* ou recursos é calculada por meio da aplicação da medida do cosseno. Sinonímia é a fonte causadora de redundância, portanto, para saber se duas *tags* são consideradas redundantes, a abordagem verifica se ambas pertencem ao mesmo *cluster*. A avaliação experimental não contempla aspectos relacionados à validação semântica das *tags* existentes em cada *cluster*, ou seja, se as *tags* consideradas sinônimas são realmente sinônimas.

Uma abordagem holística para descobrir sinonímia, homonímia e relações de subordinação em folksonomias é introduzida por Dattolo, Eynard e Mazzola (2011), na qual foram aplicadas diferentes medidas de parentesco em uma base de dados do Delicious. Para detecção de sinonímia, os autores usaram ferramentas de *stemming* (extração da raiz morfológica da palavra), distância de edição normalizada e buscas por sinônimos no WordNet. Experimentos mostraram que a medida do cosseno é capaz de identificar relações

de sinonímia. Os padrões *Hearst on the web*[20] (HEARST, 1992) foram adotados como estratégia para detectar relações de subordinação do tipo *is-a* e, posteriormente, construir uma árvore de hierarquia entre *tags*. Sua principal desvantagem se deve à necessidade de submeter várias consultas a um motor de busca à procura dos padrões *Hearst*, com a finalidade de obter a quantidade de páginas retornadas como um indicador de força para a relação submetida. Isso torna a abordagem não escalável para um elevado número de *tags*.

Eynard, Mazzola e Dattolo (2013) estendem o trabalho de Dattolo, Eynard e Mazzola (2011) e aplicam uma versão personalizada do conceito de *Similaridade do Contexto de Tags* abordado por Cattuto et al. (2008) para tratar o problema de detecção de sinonímia. A descoberta de relações *is-a* foi aprimorada com a substituição dos padrões *Hearst on the web* por um algoritmo de *clustering* que tem a finalidade de identificar o contexto de uso das *tags*.

A utilização de medidas de similaridade para detectar relações de sinonímia no espaço de *tags* é um artifício comumente empregado na literatura. Basicamente, as medidas de similaridade estimam numericamente um nível de semântica estabelecido entre duas *tags*. Porém, não conseguem qualificar o relacionamento semântico estabelecido entre os pares de *tags* diante dos vários tipos que podem estar presentes. A abordagem CPDST, por outro lado, é concebida para limitar essa deficiência, oferecendo a capacidade de tratar e detectar relações semânticas mais específicas, neste caso, sinonímia e hiperonímia/hiponímia.

3.2 Detecção de Relações Hierárquicas

Em se tratando de detecção de relações de subordinação em folksonomias, as abordagens existentes na literatura convergem para a construção de hierarquias entre *tags*, utilizando processos lógicos de maior complexidade. Algoritmos são normalmente propostos para identificar prováveis relações entre *tags* por meio de medidas de similaridade predefinidas.

Heymann e Garcia-Molina (2006) descrevem um algoritmo para converter um *corpus* de *tags* do Delicious em uma taxonomia de *tags* hierárquica, com base na ordem ascendente da centralidade *closeness* em um grafo de similaridade de *tags*. As *tags* são representadas como vetores *tag*-recurso e a similaridade entre elas é computada por meio do cálculo do cosseno

[20]Conjunto de padrões usado para encontrar relações de sub/super conceitos entre termos em uma frase como, por exemplo, “Conceito1 *such as* Conceito2, Conceito3”. Neste exemplo, *such as* é o elemento que identifica Conceito1 como hiperônimo de Conceito2 e Conceito3 (hipônimos).

da similaridade entre duas *tags*. Um grafo não-ponderado é então construído de modo que duas *tags* (vértices) são conectadas por uma aresta se a similaridade entre os vetores das respectivas *tags* for superior a um limiar estimado.

Tibély et al. (2013) apresentam um *framework* para extração automática de hierarquias com base na análise de estatísticas de coocorrência de *tags*. Fundamentado na teoria das redes complexas (*Complex Network Theory*) (ALBERT; BARABÁSI, 2002), a abordagem constrói um grafo acíclico direcionado entre *tags* a partir de uma base de dados, na qual as arestas orientam a relação no sentido da *tag* de maior nível (genérico) na hierarquia em direção aos seus descendentes de menor nível (específico), criando desta forma uma rede de conexões. O peso de uma aresta corresponde ao número de recursos compartilhados por duas *tags* em particular. Duas versões de algoritmos são implementadas para que sejam aplicadas em bases de dados distintas, visando obter melhor desempenho na extração de hierarquias.

Alguns trabalhos adotam a técnica de mineração de itens frequentes como suporte à construção de hierarquias ou aprendizado de ontologias. Neste contexto, o item é um termo genérico que define o objeto alvo do problema. Schmitz et al. (2006) focaram na técnica de regras de associação (AGRAWAL; IMIELINSKI; SWAMI, 1993) para minerar pares de *tags* que coocorrem frequentemente. As regras geradas podem ser visualizadas como candidatas a uma relação de subordinação, de modo que possam ser usadas para aprender uma estrutura taxonômica. Em parte do trabalho de Marinho, Buza e Schmidt-Thieme (2008), os autores desenvolveram um algoritmo fundamentado na mineração de itens frequentes para aprender uma ontologia a partir de *tags* de uma folksonomia. As relações de maior frequência de coocorrência são adicionadas iterativamente a um grafo, o qual modela a construção de uma hierarquia a partir da análise de hipóteses intuitivas predefinidas, até que sejam convergidas para a ontologia final.

Solskinnsbakk e Gulla (2010) apresentam uma abordagem não-supervisionada para gerar uma estrutura hierárquica entre *tags*. A construção da estrutura é guiada principalmente pela mineração de regras de associação entre pares de *tags*. Os autores propõem uma representação semântica de *tags* denominada *Tag Vector*, resultante do pré-processamento de recursos com conteúdo textual. Este componente vetorial é usado juntamente com a similaridade do cosseno para assegurar a qualidade das relações hierárquicas construídas a partir das regras de associação refinadas.

O algoritmo proposto por Benz e Hotho (2007) estende o trabalho de Heymann e Garcia-Molina (2006) para induzir relações de hierarquia entre *tags*. De acordo com a metodologia do algoritmo, uma combinação de métricas protagonizam a decisão sobre o local em que uma nova *tag* é inserida na hierarquia em construção. O processo é executado de forma iterativa e em cada passo são aplicados, de forma sequencial, uma medida inicial de ocorrência de *tags* (seleção de *tags* mais populares), uma medida de generalidade[21] (número de relações de coocorrência com outras *tags*), uma medida de similaridade (contagem de coocorrência *tag-tag* normalizada com base nas associações feitas pelo maior número de usuários) e duas medidas que identificam possíveis sinônimos (distância de edição e modelo estatístico de sinonímia, esta última computada por meio do cálculo da probabilidade condicional de coocorrência de *tags* embasado em recursos).

Almoqhim, Millard e Shadbolt (2013) defendem que anotar recursos com pares de *tags* na forma de um relacionamento *is-a* ao invés de *tags* individuais pode melhorar a qualidade da hierarquia de *tags* resultante. Ao solicitar que os usuários forneçam pares de *tags* relacionadas, com t' sendo a *tag* para um dado recurso e t uma generalização de t', os autores deduzem que a qualidade da estrutura hierárquica é aprimorada com pouco impacto de esforço na usabilidade. O algoritmo proposto emprega medidas para estimar similaridade (coocorrência entre duas *tags*) e generalidade (número de coocorrências de uma *tag* com as demais).

Meo, Quattrone e Ursino (2009) organizam um grupo de *tags* semanticamente relacionadas em um modelo hierárquico utilizando técnicas probabilísticas para mensurar o grau de similaridade e generalização entre duas *tags*. A abordagem permite ao usuário tanto visualizar *tags* de interesse de acordo com o nível de granularidade semântica desejado, quanto sugerir *tags* que auxiliem a anotação. Primeiramente, para um conjunto de *tags* de entrada, deriva-se um grupo de *tags* semanticamente relacionadas por meio da aplicação do coeficiente de Jaccard. Em seguida, as *tags* derivadas são organizadas hierarquicamente de acordo com a interpretação dos relacionamentos mútuos de generalização e inclusão. Cai et al. (2011) também adotam medidas de sobreposição mútua para medir o grau de inclusão e generalização entre *tags*.

[21] Os autores seguem a ideia intuitiva de que uma *tag* que ocorre junto com muitas *tags* diferentes é mais genérica.

Si, Liu e Sun (2010) analisam a descoberta de relações hierárquicas ao estimar a probabilidade condicional $p(t|t')$ entre *tags* sob três aspectos de coocorrência: *tag-tag*, *tag-word* e *tag-reason*, sendo os dois últimos métodos dependentes de conteúdo textual. A motivação é embasada pelo seguinte raciocínio: t é um super-conceito de t' ($t \succ t'$) se os recursos em que t' ocorre são um subconjunto dos recursos em que t ocorre. Uma vez que t é mais frequente na hierarquia, t é visto como "pai" de t'. Os *top-k* pares com maior $p(t|t')$ são selecionados para compor o conjunto final de relações hierárquicas descobertas.

É possível encontrar trabalhos na literatura que exploram técnicas de *clustering* como estratégia para alocar *tags* em agrupamentos hierárquicos (ANGELETOU et al., 2007; ZHOU et al., 2007). Embora seja possível capturar similaridade entre *tags*, a técnica de *cluster* não consegue distinguir internamente relações mais específicas tais como sinonímia ou hiperonímia/hiponímia. Além disso, *clusters* ignoram o sentido do relacionamento.

Spiliopoulos, Vouros e Karkaletsis (2010) utilizam a técnica de classificação na concepção de uma abordagem denominada CSR (*Classification-Based Learning of Subsumption Relations*), com o objetivo de aprender relações de subordinação entre conceitos de duas ontologias distintas. Dado um par de ontologias (origem e destino), o objetivo do CSR é aprender padrões de atributos que fornecem evidências para a relação de subordinação entre conceitos. Assim, por meio da tarefa de classificação, CSR define se dois conceitos estão associados por uma relação de subordinação.

As estratégias apresentadas nesta seção para construção de hierarquias são consolidadas em algoritmos que combinam métodos e métricas para determinar o melhor local de conexão para as *tags*. Uma das vantagens da abordagem CPDST para detecção de relações de subordinação é que o modelo de predição não depende de uma aplicação específica, podendo ser utilizado para diferentes finalidades como, por exemplo, auxiliar na expansão de consultas, construir estruturas hierárquicas, prover listas de *tags* semanticamente relacionadas e recomendar *tags*.

3.3 Navegabilidade e Busca em Folksonomias

Muitas pesquisas têm sido propostas com o intuito de aprimorar a experiência do usuário em encontrar recursos armazenados nas folksonomias (BINDELLI et al., 2008;

KNAUTZ; SOUBUSTA; WOLFGANG, 2010; LANIADO; EYNARD; COLOMBETTI, 2007; LOHMANN; ZIEGLER; TETZLAFF, 2009). Além da opção de busca direta por palavra-chave, normalmente a interface dos sistemas de anotação são enriquecidas com funcionalidades alternativas que visam dinamizar a forma como os usuários podem navegar pelos recursos disponíveis. O objetivo é evitar que os usuários tenham que navegar através de listas enormes de potenciais resultados antes de encontrar o recurso desejado.

O conceito de nuvem de *tags* é uma consequência dessa tendência. A nuvem de *tags* é uma representação visual sumarizada das *tags* mais utilizados pelos usuários nas anotações de recursos em um sistema. Neste modelo, as *top-k tags* mais frequentes são selecionadas e normalmente dispostas em ordem alfabética, sendo proporcionalmente expandidas em termos de tamanho de fonte e ênfase. Além do clássico arranjo sequencial de *tags* (Figura 3.1), outros *layouts* alternativos de distribuição de *tags* estão disponíveis, tais como circular, *clustered* (agrupamento de *tags* semanticamente relacionadas) e randômico (LOHMANN; ZIEGLER; TETZLAFF, 2009).

Figura 3.1: Nuvem de *Tags*.

A ação de clicar em uma das *tags* da nuvem conduz ao conjunto de recursos anotado pela referida *tag*. Na prática, a nuvem de *tags* tipicamente transforma o vocabulário emergente de uma folksonomia em uma ferramenta de navegação de baixo esforço cognitivo (SINCLAIR; CARDEW-HALL, 2008). A nuvem de *tags* é oportuna para tarefas exploratórias e situações na qual o usuário não sabe exatamente o que está procurando, além de também servir de ponto de partida para buscas mais específicas condicionadas a uma palavra-chave (LEGINUS; DOLOG; LAGE, 2013b).

Uma vez que os algoritmos de seleção de nuvens de *tags* não provêem informações com respeito ao relacionamento semântico estabelecido entre *tags*, trabalhos de pesquisa vêm sendo propostos para explorar essa lacuna. Laniado, Eynard e Colombetti (2007) introduzem uma hierarquia de conceitos semanticamente relacionados como suporte à navegação na interface de um sistema inspirado no Delicious. A partir dos n recursos associados a uma dada *tag*, são coletadas as m *tags* mais frequentes em cada recurso. Então, para cada *tag* distinta presente na coleção, uma cadeia de hiperonímia/hiponímia assegurada pelo WordNet é criada com um caminho até a raiz original da hierarquia, ou seja, a chave de busca. A árvore resultante é então comprimida para eliminar *tags* sem utilidade, de modo a favorecer a navegação. Por fim, uma barra lateral é exibida ao usuário contendo a hierarquia resultante. No entanto, o trabalho não cobre uma avaliação sobre a utilidade da proposta sob o ponto de vista do usuário e a capacidade da abordagem de produzir hierarquia para qualquer chave de busca reconhecida pelo sistema.

Bindelli et al. (2008) apresentam um engenho de busca denominado *TagOnto* que oferece uma interface de navegação com base em conceitos de ontologias para a recuperação de recursos adicionais. O sistema recebe uma *tag* de busca como entrada, examina os conceitos em uma ontologia de referência para identificar a associação apropriada do par *tag*/conceito e procura termos relacionados que possam refinar a intenção de busca. A interface é dividida em duas porções horizontais: a parte superior reporta o resultado da busca, enquanto que a parte inferior é dedicada à apresentação dos conceitos da ontologia associados à chave de busca e navegação.

Knautz, Soubusta e Wolfgang (2010) desenvolveram um sistema de RI denominado *Tag Cluster* como proposta para uma nova ferramenta de visualização e expansão de consulta dirigida à visualização. Para oferecer suporte à detecção de semântica entre *tags*, o sistema incorpora o cálculo de similaridade por meio de diferentes medidas e coeficientes (dice, cosseno e Jaccard) e subsequente *clustering*. Através de sua interface gráfica, *Tag Cluster* proporciona um ambiente de maior interatividade com as funcionalidades do sistema, permite adicionar novos argumentos de busca à consulta inicial e provê acesso a *tags* semanticamente similares as quais o usuário não tinha conhecimento antes do início de sua busca.

No sistema BibSonomy (BENZ et al., 2010), após a submissão de uma busca fornecendo

uma *tag* de entrada, os resultados (*bookmarks* e artigos científicos) são ordenados por um algoritmo de ranqueamento e apresentados ao usuário. Adicionalmente, uma lista de *tags relacionadas* e outra de *tags similares* também são exibidas, propiciando opções de navegação semântica ao permitir acesso aos recursos que são (por tendência) anotados por *tags* semanticamente correlacionadas. As *tags relacionadas* são extraídas por meio da aplicação da medida de coocorrência entre *tags*, enquanto que as *tags similares* são obtidas por meio do cálculo da similaridade do cosseno em vetores *tag-tag*.

Os trabalhos mencionados oferecem diferentes visões sobre como a navegação pode ser aprimorada. Os sistemas expostos propõem funcionalidades úteis, entretanto carecem de uma avaliação experimental que possa mensurar a relevância das propostas para o usuário final, seja através da condução de um estudo focado na satisfação do usuário ou aplicação de métricas sintéticas que qualifiquem similar resultado. Focado no cenário de geração de listas de *tags*, este trabalho avalia a aplicação da abordagem CPDST para prover listas de *tags* com semântica específica, ou seja, sinonímia e relação de subordinação. Os métodos oferecidos pelo BibSonomy utilizam medidas de similaridade que estimam um grau de relacionamento semântico entre *tags*, sem se preocupar em determinar quais tipos de relacionamentos semânticos estão presentes no resultado. A realização de uma série de experimentos distintos sintetiza o comportamento dos métodos comparados sob diversos aspectos, inclusive quantitativamente sob a perspectiva de métricas que quantificam a qualidade das listas de *tags* geradas.

3.4 Caracterização dos Trabalhos Relacionados

Nesta seção, os trabalhos relacionados mais representativos mencionados nas Seções 3.1 e 3.2 são classificados de acordo com uma lista de particularidades. Para a tarefa de detecção de sinonímia, os seguintes itens são considerados:

- (C1) Procedimento empírico: medidas de similaridade/parentesco/distância são utilizadas como heurísticas para identificar relações semânticas entre *tags*;
- (C2) Detalhamento experimental: inclui detalhes acerca da preparação, apresentação e avaliação dos resultados experimentais;

- (C3) Validação semântica: realiza validação semântica para os resultados apresentados, seja mediante o uso de uma fonte de dados externa confiável, julgamento humano ou outro artifício aceito pela comunidade científica;

- (C4) Capacidade de adaptação para explorar outros tipos de relação semântica: se a proposta do trabalho não estiver focada em um escopo muito específico, significa que sua base metodológica pode ser estendida para incrementar a descoberta de outros tipos de relação semântica. O julgamento segue um ponto de vista subjetivo;

- (C5) Detecção de sinonímia: indica se a proposta do trabalho inclui métodos para a descoberta de sinonímia;

- (C6) Detecção de relações de subordinação: indica se a proposta do trabalho também contempla a detecção de relações de subordinação;

- (C7) Acesso a fontes de dados externas: sinaliza se a abordagem incorpora acesso a recursos linguísticos externos;

- (C8) Dependência de conteúdo: informa se a abordagem é dependente de um tipo específico de recurso como, por exemplo, texto, foto, vídeo, entre outros.

A Tabela 3.1 sumariza o resultado da categorização dos trabalhos relacionados para os itens de C1 a C8. No que diz respeito ao item (C7), verificou-se que Cattuto et al. (2008), Dattolo, Eynard e Mazzola (2011), Lee e Yong (2007) utilizam o WordNet como ferramenta de recurso linguístico externo. Além do WordNet, o trabalho de Dattolo, Eynard e Mazzola (2011) incorpora serviços *online* para tradução de uma *tag* em diferentes idiomas. O trabalho de Solskinnsbakk e Gulla (2011) adota uma metodologia que depende de conteúdo textual (C8).

Vale ressaltar que alguns trabalhos possuem um escopo bem restrito, seja para testar se uma medida é apropriada para identificar um tipo de relação semântica específico ou avaliar os impactos causados pelo vocabulário descontrolado no uso de *tags*.

Tabela 3.1: Caracterização dos trabalhos relacionados: detecção de sinonímia.

Autor(es)	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
Wartena (2010)	√	√			√			
Clements, Vries e Reinders (2008)	√				√			
Solskinnsbakk e Gulla (2011)	√	√			√			√
Lee e Yong (2007)		√	√		√		√	
Cattuto et al. (2008)	√	√	√		√		√	
Gemmell et al. (2009)	√	√			√			
Dattolo, Eynard e Mazzola (2011)	√			√	√	√	√	
Benz (2007)	√	√	√		√	√		
Eynard, Mazzola e Dattolo (2013)	√			√	√	√	√	
Mousselly-Sergieh et al. (2014)	√	√	√					

Em se tratando de detecção de relações de subordinação, a Tabela 3.2 sumariza as principais características dos trabalhos relacionados de acordo com as seguintes particularidades:

- (C1) Construção de hierarquias: sinaliza se a proposta do trabalho contempla a definição de um algoritmo destinado à construção de hierarquias ou aprendizado de ontologias;
- (C2) Aplicação da técnica de mineração de itens frequentes: define se a seleção de pares de *tags* candidatas a uma relação hierárquica é determinada pela análise de regras de associação;
- (C3) Detecção de hierarquias usando *clustering*: aponta se o trabalho implementa a alocação de *tags* similares em *clusters* hierárquicos;
- (C4) Validação semântica: aponta se é feita a realização de validação semântica para os resultados apresentados, seja mediante o uso de uma fonte de dados externa confiável, julgamento humano ou outro artifício aceito pela comunidade científica;
- (C5) Dependência de conteúdo: sinaliza se a abordagem é dependente de um tipo específico de recurso como, por exemplo, texto, foto, vídeo, entre outros.
- (C6) Detalhamento experimental: assinala se o trabalho inclui detalhes acerca da preparação, apresentação e avaliação dos resultados experimentais;

Tabela 3.2: Caracterização dos trabalhos relacionados: detecção de relações hierárquicas.

Autor(es)	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Heymann e Garcia-Molina (2006)	√					√
Tibély et al. (2013)	√			√		√
Schmitz et al. (2006)	√	√				√
Marinho, Buza e Schmidt-Thieme (2008)	√	√		√		√
Solskinnsbakk e Gulla (2010)	√	√			√	√
Benz (2007)	√					√
Almoqhim, Millard e Shadbolt (2013)	√			√		√
Meo, Quattrone e Ursino (2009)	√					√
Cai et al. (2011)				√		√
Si, Liu e Sun (2010)				√		√
Angeletou et al. (2007)			√			√
Zhou et al. (2007)			√		√	

Os trabalhos de Angeletou et al. (2007), Cai et al. (2011), Si, Liu e Sun (2010), Zhou et al. (2007) estão focados simplesmente na exploração da estrutura das folksonomias para descobrir relações de subordinação automaticamente. Cai et al. (2011) e Si, Liu e Sun (2010) se apoiam na aplicação de técnicas probabilísticas, enquanto que Angeletou et al. (2007), Zhou et al. (2007) derivam hierarquia semântica utilizando técnicas de *clustering* (C3). Dentre os trabalhos assinalados com o item (C1), apenas no de Solskinnsbakk e Gulla (2010) a proposta de geração de hierarquias é dependente de conteúdo textual.

Poucos trabalhos realizam avaliação qualitativa e quantitativa das estruturas hierárquicas resultantes de forma automática, com base em um *gold standard* (C4). Benz (2007) utilizam uma hierarquia de estilos musicais como referência para comparar as estruturas hierárquicas produzidas pelo algoritmo de Heymann e Garcia-Molina (2006) e pelo trabalho proposto. No trabalho de Almoqhim, Millard e Shadbolt (2013), a hierarquia construída é comparada com uma taxonomia de domínio criada manualmente por pesquisadores especialistas em Web Semântica. Em ambos trabalhos, foram empregadas as medidas *precision*, *recall* e *f-measure* taxonômicas para quantificar a qualidade das hierarquias em relação ao respectivo *gold-standard*.

As abordagens propostas por Angeletou et al. (2007), Meo, Quattrone e Ursino (2009), Tibély et al. (2013), Zhou et al. (2007) avaliam a qualidade das relações de subordinação

com base na intuição e senso comum. Heymann e Garcia-Molina (2006) e Schmitz et al. (2006) focalizam a discussão do modelo de construção de hierarquia proposto, sem realizar um exame detalhado sobre a qualidade das relações descobertas.

Experimentos com a participação de julgamento humano são encontrados nos trabalhos de Almoqhim, Millard e Shadbolt (2013), Cai et al. (2011), Si, Liu e Sun (2010), Solskinnsbakk e Gulla (2010). Si, Liu e Sun (2010) avaliam as relações de subordinação quantitativamente usando as medidas *precision* e *coverage*. O julgamento das relações corretas é feito por avaliadores humanos utilizando *pooling*, uma técnica amplamente aceita na RI para rotulação de relevância de documentos (VOORHEES; HARMAN, 2005). Cai et al. (2011) comparam as relações de subordinação detectadas pelo algoritmo proposto com uma hierarquia construída manualmente e utilizam as medidas *precision*, *recall* e *f-measure* para quantificar a qualidade dos conceitos de hierarquia produzidos.

3.5 Considerações Finais

Embora os trabalhos reportados neste capítulo sejam relevantes para identificar sinonímia ou construir estruturas hierárquicas entre *tags* de folksonomias, as medidas de similaridade sozinhas não fornecem garantias empíricas ou teóricas que são de fato adequadas para identificar as relações semânticas desejadas. Esta tese, por outro lado, apresenta uma abordagem com fundamentos plausíveis, pois as relações de sinonímia e subordinação entre *tags* são aprendidas diretamente dos dados da folksonomia. Os atributos são extraídos de pares de *tags* utilizando medidas de similaridade/parentesco/distância semântica e estimativas de generalização/especialização mútuas que refletem diferentes aspectos das relações de sinonímia e de subordinação entre *tags*. A sistemática associada à tarefa de classificação proposta é genérica, ou seja, não é dependente de um domínio específico. Além disso, destaca-se sua flexibilidade de adaptar-se facilmente à temporalidade dos dados, pois qualquer mudança no padrão dos atributos será percebida pelo modelo após novo treinamento.

A abordagem de AM para detectar relações semânticas entre *tags* apresentada neste trabalho é similar à apresentada por Spiliopoulos, Vouros e Karkaletsis (2010) para o alinhamento de ontologias, mas com direcionamento focado em folksonomias. De fato,

ideias do campo da identificação de objetos foram tomadas como referência para a concepção do trabalho. Na tarefa de identificação de objetos (BRIZAN; TANSEL, 2006), também conhecida como *record linkage*, resolução de entidades e detecção duplicada, o objetivo consiste em identificar grupos de objetos que são idênticos mesmo que esses objetos contenham atributos com ruídos[22] ou ausentes, ocasionados pela mistura de informações oriundas de múltiplas fontes de dados (KÖPCKE; THOR; RAHM, 2010; RENDLE; SCHMIDT-THIEME, 2006). De acordo com Rendle e Schmidt-Thieme (2006), os trabalhos neste campo do conhecimento visualizam a tarefa de identificação de objetos como um problema de classificação para aprender classes de equivalência entre objetos.

A navegabilidade em folksonomias também é relatada sob o ponto de vista dos sistemas de busca/recuperação proporcionarem aos seus usuários alternativas de aprimorarem a experiência de busca, oferecendo interfaces auxiliares que ampliam o acesso a recursos adicionais por meio de *tags* semanticamente relacionadas. O capítulo seguinte aborda os tópicos essenciais inerentes à modelagem do problema de classificação idealizado neste trabalho.

[22]Por exemplo, um determinado produto pode conter descrições diferenciadas quando discriminado em *websites* distintos.

Capítulo 4

Abordagem CPDST - Classificação Para Detecção de relações Semânticas entre pares de *Tags*

Neste capítulo, contextualiza-se o problema genérico de predição semântica no espaço de *tags* de uma folksonomia. Primeiramente, na Seção 4.1, é realizada a formalização do problema de detecção de semântica como um problema de classificação binária entre pares de *tags*. Em seguida, na Seção 4.2 são especificadas as medidas de similaridade empregadas como atributos para treinamento dos modelos preditores de sinonímia e relações de subordinação. A Seção 4.3 especifica a sistemática desenvolvida para rotular automaticamente as instâncias de treinamento. Por fim, o problema de classificação multiclasse em bases de dados desbalanceadas é discutido na Seção 4.4.

4.1 Classificação para Detecção Semântica

Para enfrentar o problema de detecção automática de relações semânticas entre *tags* de folksonomias, propõe-se a elaboração de um modelo genérico de classificação binária para capturar uma relação semântica em particular, a partir dos dados de uma folksonomia. Tomando-se como parâmetro a necessidade de detectar relações de sinonímia entre *tags*, um sistema de aprendizado necessita de subsídios com base em um conjunto de características que expressam tal fenômeno, para inferir se duas *tags* são sinônimas ou não. Os casos

de pares de *tags* que comprovam a existência de sinonímia correspondem aos *exemplos positivos*, e os contra-exemplos aos *exemplos negativos*. Deste modo, o aprendizado de classe dedica-se a encontrar propriedades compartilhadas por todos os exemplos positivos e ausentes nos exemplos negativos (ALPAYDIN, 2010).

Considere Ω uma denominação abstrata de relação semântica a ser utilizada como exemplo. Esta denominação é usada para representar o tipo de relação semântica a qual se deseja detectar como, por exemplo, sinonímia ou relação de subordinação. Seja $\mathcal{X}$ um conjunto denominado *espaço de atributos*, ou seja, o espaço abstrato que abriga o conjunto de atributos usados para caracterizar um fenômeno observado, e $\mathcal{Y} = \{+, -\}$ um conjunto de rótulos para as classes positiva e negativa, respectivamente. O conjunto de treinamento $\mathcal{D}^{\text{train}} := \{(\vec{x}_k, y_k), \ldots, (\vec{x}_n, y_n)\}$ contém exemplos qualificados cujas classes são conhecidas, em que $\vec{x}_k \in \mathbb{R}^m$ são os vetores de atributos que representam as instâncias observadas e $y_i \in \mathcal{Y}$ a classe a qual $\vec{x_k}$ pertence.

Dada uma função de similaridade *tag-tag* $f : T^2 \to \mathbb{R}$ a qual designa um grau de semântica Ω entre $(t, t') \in T^2$, o vetor de atributos é então definido como $\vec{x}_i = (f_1(t_i, t_j), \ldots, f_m(t_i, t_j))$, em que $t_i \neq t_j$. O conjunto de teste $\mathcal{D}^{\text{test}}$ é um conjunto independente caracterizado pelas seguintes particularidades: (i) os exemplos $\mathcal{D}^{\text{test}}$ não podem ter sido utilizados para treinamento, e (ii) o rótulo de classe y de cada exemplo não é fornecido ao classificador, para que o mesmo determine a classe a qual o exemplo pertence. A ideia é aprender uma função de classificação da forma

$$\hat{y} : \mathcal{X} \to \mathcal{Y},$$

na qual $\hat{y}$ é uma função de predição capaz de reconhecer relacionamentos entre os atributos e a variável alvo do conjunto de treinamento e minimize o erro no conjunto de teste $\mathcal{D}^{\text{test}} \subseteq \mathcal{X} \times \mathcal{Y}$ (indisponível durante treinamento), ou seja,

$$\text{err}(\hat{y}; \mathcal{D}^{\text{test}}) := \frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{(x,y) \in \mathcal{D}^{\text{test}}} \ell(y, \hat{y}(x)).$$

A expressão $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ é uma função de perda que mede, para qualquer instância de teste $(x, y) \in \mathcal{D}^{\text{test}}$, o desajuste entre o y verdadeiro e seu valor previsto $\hat{y}(x)$.

A Figura 4.1 sintetiza o funcionamento da abordagem de classificação semântica

proposta neste trabalho. Para o classificador trabalhar corretamente, primeiramente é necessário extrair vetores de atributos entre pares de *tags*, ou seja, $v : T^2 \rightarrow \mathbb{R}^n$, utilizando várias medidas de similaridade *tag-tag* (cf. Seção 4.2) apropriadas para a indicação da semântica alvo desejada[23]. Note que agora os dados de treinamento são constituídos por um conjunto de vetores de atributos $\mathcal{X}$, em que cada vetor de atributo é rotulado como positivo ou negativo, indicando se o par de *tags* representa ou não uma semântica Ω de acordo com o resultado de um *gold standard* (passo 1). A tarefa prossegue com a construção do modelo de classificação, com a diferença de que, agora, para qualquer par distinto de *tags* $(t, t') \in T^2$, deseja-se aprender uma função de classificação (passo 2) entre *tags* da forma $\hat{y} : f(t, t') \rightarrow \mathcal{Y}$, que prediz a classe (positiva ou negativa) de novas instâncias no conjunto de teste (passo 3) como positiva ou negativa (passo 4).

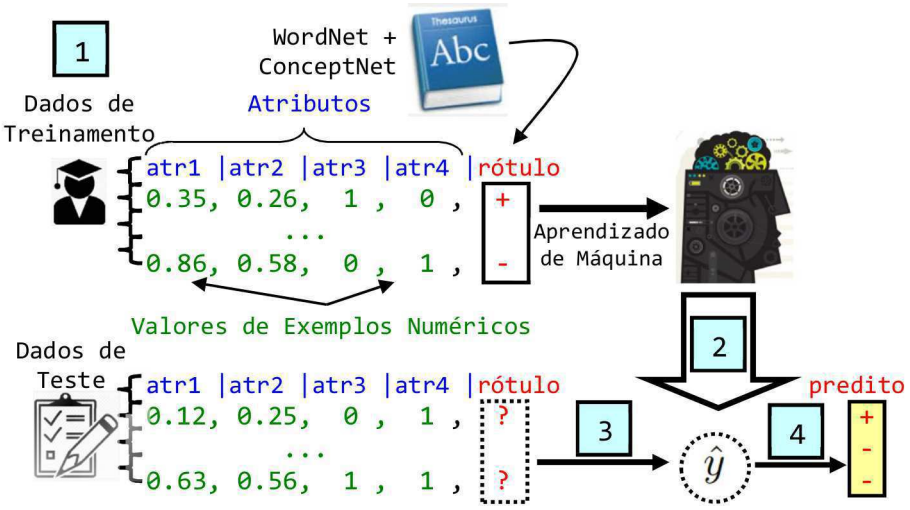

Figura 4.1: Classificação para detecção de relação semântica em folksonomias.

Por se tratar de um problema de classificação binária, a abordagem proposta é instanciada para tratar apenas um tipo de relação semântica por vez. Entretanto, unidades distintas de predição semântica (e.g., sinonímia e relação de subordinação) podem ser integradas para originar uma solução híbrida, permitindo que um único sistema possa se comunicar com as partes e obter predições em termos de sinonímia, relações de subordinação ou ambas.

Cada tipo de relação semântica apresenta suas singularidades. Algumas propriedades importantes acerca das relações de sinonímia e subordinação abordadas nesta tese, tomando como exemplo um par de *tags* de entrada (t, t'), são definidas como segue:

[23]Cada alvo de classificação semântica (sinonímia ou subordinação) normalmente utiliza atributos específicos.

- Uma relação (t, t') candidata a sinonímia é denotada pela simbologia $t \equiv t'$, ou seja, *t é equivalente a t'*. Para relação de subordinação, $t \succ t'$ expressa que *t é um super-conceito de t'*, ou seja, *t subordina t'*;

- A relação de sinonímia é simétrica, ou seja, se $t \equiv t'$ então $t' \equiv t$;

- A relação de subordinação é assimétrica, ou seja, $t \succ t'$ não implica em $t' \succ t$. Por definição, se t é hiperônimo de t', deduz-se que t' é hipônimo de t. Logo, a inversão da ordem de entrada das *tags* para (t', t) não preserva a indicação de uma provável relação de subordinação;

- A relação de subordinação é transitiva, ou seja, se $t_i \succ t_j$ e $t_j \succ t_k$, então $t_i \succ t_k$.

Uma importante consideração a ser assumida deste ponto em diante é que, para a detecção de relações de subordinação, o modelo de predição concebido nesta tese de doutorado está interessado em identificar apenas entradas (t, t') na qual t subordina t', ou seja, no sentido do conceito mais geral para o mais específico.

4.2 Extração de Atributos

Para fins de treinamento, propõe-se a extração de atributos entre pares de *tags* disponíveis em uma base de dados por meio da aplicação de medidas de similaridade/parentesco/distância. As medidas apresentadas a seguir são apoiadas pelos trabalhos relacionados e foram usadas para capturar relações semânticas entre *tags*, em particular, relações de sinonímia e de subordinação.

4.2.1 Atributos para Sinonímia

Tags sinônimas podem se manifestar em diferentes grafias como, por exemplo, mediante uso de acrônimos, variações de número gramatical, inflexões do substantivo ou verbo e preferência de vocabulário. A seguir são apresentadas as heurísticas mais discriminativas para a detecção de sinônimos no espaço de *tags*.

Distância de Edição - DE

A distância de edição é uma métrica conhecida para tratar o problema de correspondência de sequência de caracteres entre *strings*, embasada na noção de operações de edição primitivas: inclusão, exclusão e substituição de caracteres. A distância entre duas *strings* é calculada em termos do número mínimo de operações necessárias para transformar uma *string* em outra (DAMERAU, 1964). A distância de edição possui medidas variantes que diferem entre si em relação ao conjunto de operações aplicado a *string*. Uma de suas variantes mais usuais é a distância de Levenshtein (LEVENSHTEIN, 1966).

A DE é muito útil para detectar similaridade sintática entre cadeias de caracteres distintas. Particularmente, sua aplicação proporciona a identificação de termos sinônimos decorrentes da variação em número gramatical (e.g., `tablet` e `tablets`), adição de símbolos especiais (e.g., `case_study` e `casestudy`) e erros ortográficos (e.g., `computer` e `conputer`).

Uma vez que as *strings* comparadas podem apresentar diferentes comprimentos, a computação da métrica DE por meio da quantidade total de operações de edição pode ser inadequada para alguns tipos de aplicações se não for conduzida uma normalização. Por exemplo, 2 operações de edição observadas na comparação entre *strings* de comprimento dois são mais críticas do que 3 operações de edição em *strings* de comprimento igual a nove. Com o intuito de suprimir as eventuais diferenças de comprimento entre as *strings* comparadas, uma versão da DE normalizada é adotada neste trabalho. Assim, para qualquer par de *tags* $(t, t') \in T^2$, sua Distância de Edição Normalizada *DEN* é calculada como segue:

$$\text{DEN} = \frac{\text{DE}(t, t')}{\max(L(t), L(t'))}, \tag{4.1}$$

em que $L(t)$ e $L(t')$ são os comprimentos das *tags* t e t' respectivamente. Observe que, quando *DEN=0* as *tags* são idênticas em grafia, enquanto que valores próximos a 1 indicam alta dissimilaridade.

Para facilitar o entendimento das medidas adotadas como atributo para a detecção de sinonímia, a Figura 4.2 ilustra um cenário de folksonomia que servirá de base para exemplificar como são calculadas as medidas de similaridade/parentesco/distância contidas nesta seção. A figura é constituída por 2 usuários distintos (u_1 e u_2), 3 tags (t_1, t_2 e t_3) e 3 recursos (r_1, r_2 e r_3). Esta simbologia pode ser recordada na Seção 2.4. Cada *tag* possui suas

respectivas representações de vetores de coocorrência *tag-tag* e *tag*-recurso. Considerando que $t_1 = car$ e $t_2 = cars$, o cômputo da DE(t_1, t_2) é igual a 1, visto que apenas uma operação de edição (inserção do caractere 's' em car) é necessária para transformar car em $cars$. Em sua versão normalizada, DEN$(t_1, t_2) = 0.25$ conforme aplicação da Equação 4.1.

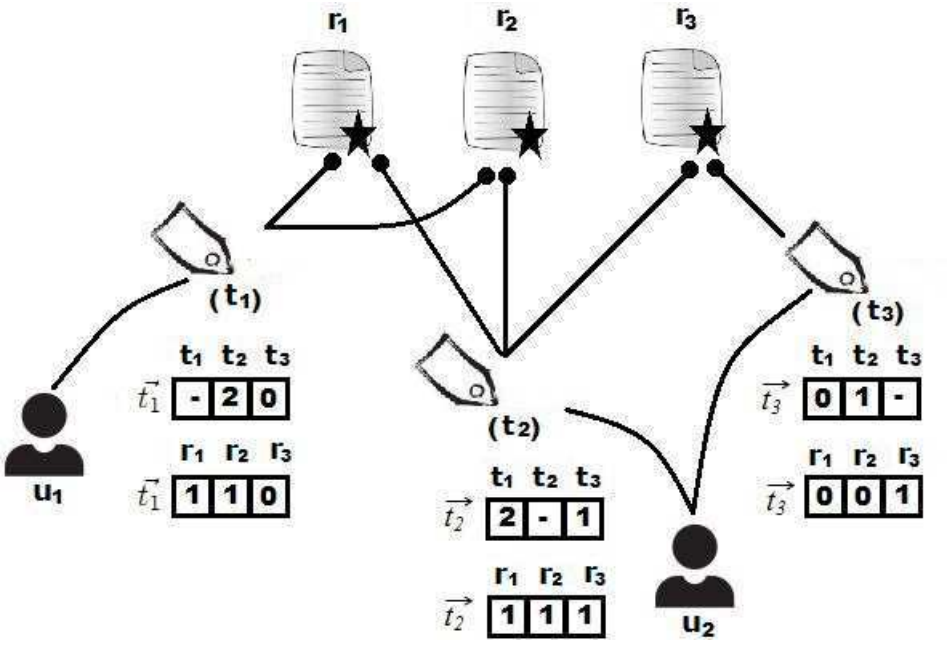

Figura 4.2: Exemplo ilustrativo de folksonomia para entendimento do cálculo de similaridade/distância entre *tags*.

Coocorrência *tag-tag*

A contagem de cooccorrência entre *tags* tem sido usada por vários autores para mensurar similaridade semântica entre *tags* (CATTUTO et al., 2008; CLEMENTS; VRIES; REINDERS, 2008; DATTOLO; EYNARD; MAZZOLA, 2011; GEMMELL et al., 2009). A heurística apoia-se no pressuposto que *tags frequentemente aplicadas ao mesmo recurso oferecem um sinal de que estão relacionadas semanticamente*. Esta conjuntura é largamente aceita na literatura e argumentada em experimentos efetuados por Begelman, Keller e Smadja (2006) e Li et al. (2007). Neste trabalho, adota-se a mesma definição introduzida por Cattuto et al. (2008) na qual a contagem de coocorrência entre duas *tags* distintas t e t' é definida como o número de *posts* em que elas coocorrem, ou seja,

$$\text{tag-tag}^{\text{post}}(t, t') := |\{(u, r) \in U \times R | t, t' \in T_{ur}\}|. \tag{4.2}$$

Alternativamente, a coocorrência por recurso pode ser contada como segue:

$$\text{tag-tag}^{\text{res}}(t, t') := |\{r \in R | t, t' \in T_r\}|. \tag{4.3}$$

Para manter todos os valores deste atributo no mesmo intervalo de $[0, 1]$, utiliza-se uma versão normalizada da Equação 4.2 representada pela Equação 4.4, no qual w e w' formam pares de *tags* coocorrentes.

$$\text{tag-tag}^{\text{norm}}(t, t') = \frac{\text{tag-tag}^{\text{post}}(t, t')}{\max_{(w,w' \in T^2 \,:\, w \neq w')} \text{tag-tag}^{\text{post}}(w, w')}. \tag{4.4}$$

A versão normalizada para a Equação 4.3 é definida analogamente, substituindo a função tag-tag$^{\text{post}}$ por tag-tag$^{\text{res}}$. É importante ressaltar que, na normalização da coocorrência, a distribuição dos valores no intervalo [0,1] pode apresentar bastante variação dependendo da base de dados utilizada (BENZ, 2007). Por exemplo, é comum encontrar poucas *tags* com alto valor de frequência de coocorrência, caracterizando *outliers* na série de dados, fazendo com que a maioria dos valores de coocorrência normalizada fique concentrada em um intervalo inferior como, por exemplo, [0, 0,2].

No exemplo da Figura 4.2, as *tags* t_2 e t_3 coocorrem em um *post* atribuído pelo usuário u_2 ao recurso r_3. Deste modo, tag-tag$^{\text{post}}(t_2, t_3) = 1$. Para tag-tag$^{\text{res}}(t_2, t_3)$ considerando o recurso r_3, o valor de coocorrência também é igual a 1, pois (t_2, t_3) marcam apenas uma vez o recurso r_3. Observe que, se t_3 for substituído por t_1, tag-tag$^{\text{res}}(t_1, t_2)$ para o recurso r_2 será igual a 2.

Similaridade do Cosseno

Cosseno é uma medida matemática amplamente utilizada na RI para quantificar a similaridade entre documentos modelados no espaço vetorial. O espaço vetorial é um modelo que representa documentos textuais (e quaisquer objetos em geral) como vetores de termos, para capturar a importância relativa dos termos em cada documento. Considerando o mesmo princípio, o cosseno pode ser empregado para computar a similaridade entre *tags*. Para calcular a similaridade do cosseno entre duas *tags* distintas t and t', aplica-se a seguinte equação:

$$\cos(\vec{t}, \vec{t'}) = \frac{\vec{t} \cdot \vec{t'}}{\|\vec{t}\| \cdot \|\vec{t'}\|} = \frac{\sum_{i=1}^{n} t_i \times t'_i}{\sqrt{\sum_{i=1}^{n} (t_i)^2} \times \sqrt{\sum_{i=1}^{n} (t'_i)^2}}, \tag{4.5}$$

em que $\vec{t}$ e $\vec{t'}$ são vetores de perfis para as *tags* t e t', respectivamente, $\cdot$ denota o produto escalar, e $\|\|$ a magnitude, geralmente definida como a norma Euclidiana. Os valores do cosseno variam no intervalo de 0 (máxima dissimilaridade) a 1 (máxima similaridade).

Dois tipos diferentes de componentes vetoriais são adotados: (i) a contagem de coocorrência por *post* e (ii) a contagem de coocorrência *tag*-recurso. Cattuto et al. (2008) mostraram que valores elevados do cosseno entre vetores de perfis de *tags* cujos componentes são (i) ou (ii) tendem a indicar sinonímia.

(i) Coocorrência tag-tag por Post. Cada componente vetorial corresponde à contagem de coocorrência *tag-tag* de acordo com a equação 4.2. Logo, para uma determinada *tag* $t \in T$, seu vetor de perfil é composto por

$$\vec{t} := (\text{tag-tag}^{\text{post}}(t, t_1), \text{tag-tag}^{\text{post}}(t, t_2), \ldots, \text{tag-tag}^{\text{post}}(t, t_{|T|})). \tag{4.6}$$

(ii) Coocorrência tag-recurso. Neste tipo de coocorrência, cada componente vetorial corresponde à contagem de quantas vezes uma *tag* $t \in T$ é usada para anotar um certo recurso $r \in R$, ou seja,

$$\text{tag-res}(t, r) := |\{u \in U \mid (u, t, r) \in Y\}|. \tag{4.7}$$

Portanto, o vetor de perfil de uma determinada *tag* $t \in T$ é composto por

$$\vec{t} := (\text{tag-res}(t, r_1), \text{tag-res}(t, r_2), \ldots, \text{tag-res}(t, r_{|R|})). \tag{4.8}$$

De acordo com a Figura 4.2, a aplicação da Equação 4.5 para calcular a similaridade do cosseno usando os vetores *tag-tag* de $\vec{t_1}$ e $\vec{t_2}$ resulta no valor 0.73, pois

$$\cos(\vec{t_1}, \vec{t_2}) = \frac{1 \times 2 + 2 \times 1 + 0 \times 1}{\sqrt{1^2 + 2^2 + 0^2} + \sqrt{2^2 + 1^2 + 1^2}} = \frac{2 + 2 + 0}{\sqrt{5} + \sqrt{6}} = 0.73.$$

Finalmente, o cosseno *tag*-res entre os vetores $\vec{t_1}$ e $\vec{t_2}$ é igual a 0.82.

Sobreposição Mútua

A noção de sobreposição mútua em perfis de anotações de *tags* pode ser aplicada para estimar relações de sinonímia. Originalmente introduzido por Cai et al. (2011) para prover suporte ao desenvolvimento de um algoritmo de aprendizado de ontologias, esta técnica probabilística proporciona diferentes percepções acerca do nível de informação compartilhada entre duas *tags*, por meio da análise do conjunto de recursos anotados por ambas.

Considere R_t e $R_{t'}$ o conjunto de recursos anotados por t e t', respectivamente. As funções $Min(|R_t|)$ e $Max(|R_t|)$ representam o número mínimo e máximo de recursos associados a t. Por fim, a sobreposição mútua para um par de *tags* t e t' é definida como uma tupla (a, b), em que a e b são estimativas de probabilidade de sobreposição (*overlap*) formuladas como:

$$a(t, t') = \frac{|R_t \cap R_{t'}|}{Min(|R_t|, |R_{t'}|)}, \tag{4.9}$$

$$b(t, t') = \frac{|R_t \cap R_{t'}|}{Max(|R_t|, |R_{t'}|)}, \tag{4.10}$$

em que a e b assumem valores no intervalo [0,1] (com $a \geq b$) e, intuitivamente, provêem como significado a porcentagem de interseção entre os conjuntos R_t e $R_{t'}$. No exemplo da Figura 4.2, $a(t_2, t_3) = 1$ e $b(t_2, t_3) = 0.33$.

A interpretação para os valores de sobreposição mútua a e b entre duas *tags* (t, t') é entendida da seguinte forma:

- Se o valor de $a_{t,t'}$ e $b_{t,t'}$ são suficientemente pequenos, a interseção entre os conjuntos de recursos das duas *tags* também é pequeno. Isso significa que é difícil encontrar recursos que tenham sido anotados com ambas as *tags* t e t'. **Conclusão:** é provável que não exista uma relação de equivalência (sinonímia) ou subordinação entre t e t';

- Se os valores de $a_{t,t'}$ é grande e $b_{t,t'}$ pequeno, grande parte dos recursos anotados por t' também são anotados por t, mas não vice-versa. Isso significa que o conjunto interseção $R_t \cap R_{t'}$ tem pequena ocorrência em t, mas grande abrangência no conjunto de t'. **Conclusão:** t pode ser considerado hiperônimo de t';

- Se os valores de $a_{t,t'}$ e $b_{t,t'}$ são suficientemente elevados, o conjunto interseção $R_t \cap R_{t'}$ tem grande representação em R_t e $R_{t'}$, portanto, recursos anotados por t provavelmente também são anotados por t' e vice-versa. **Conclusão:** t e t' podem ser considerados como equivalentes, ou seja, sinônimos.

Limiares são estimados empiricamente para definir o que é considerado "grande" (α) ou "pequeno" (β) na tentativa de examinar o melhor efeito na detecção da semântica pretendida. Um código computacional foi então desenvolvido para originar o atributo overlap^{syn}, com a finalidade de efetuar o cálculo dos valores da tupla (a, b) inerentes a cada par de *tags* e compará-los em relação aos limiares α e β pré-estimados, seguindo a interpretação descrita previamente. De acordo com a análise de um encadeamento de estruturas condicionais, o atributo overlap^{syn} pode assumir os seguintes valores: 1 (irrelevante), 2 (sinônimos), 3 (relevante) e 4 (t é hiperônimo de t'). O valor 1 sugere que o par de *tags* representado na instância não expressa um relacionamento semântico mútuo. O valor 3 insinua a existência de um parentesco semântico próximo, mas não uma relação de sinonímia ou subordinação. Os valores 2 e 4 são autoexplicativos.

4.2.2 Atributos para Hiperonímia/Hiponímia

A semântica por trás de uma relação hierárquica supõe a existência, primeiramente, de parentesco entre *tags* e graus de inclusão/generalização que devem satisfazer critérios empíricos. Assim como foi idealizado para sinonímia, foram empregados como atributos informativos as métricas introduzidas nos trabalhos relacionados para a detecção de relações de subordinação.

Suporte e Confiança

Alguns algoritmos propostos para a construção de hierarquia de *tags* têm como base a aplicação da técnica de mineração de *conjuntos de itens frequentes* para selecionar prováveis pares de *tags* candidatas a uma conexão hierárquica (MARINHO; BUZA; SCHMIDT-THIEME, 2008; HEYMANN; RAMAGE; GARCIA-MOLINA, 2008; SCHMITZ et al., 2006; SOLSKINNSBAKK; GULLA, 2010). Neste contexto, um *conjunto de itens*

frequentes é constituído por uma coleção de *tags* que coocorrem frequentemente, fornecendo evidências de correlação entre elas.

Para exemplificar a lógica dessa ideia, considere o exemplo ilustrado na Figura 4.2 para as *tags* t_1 e t_2. Sendo $A_R(t_1) = \{r_1, r_2\}$ e $A_R(t_2) = \{r_1, r_2, r_3\}$, se um subconjunto de recursos anotado por t_2 também foi marcado na totalidade por t_1, considera-se t_1 uma especialização de t_2. Isso remete à interpretação de que "*sempre que a tag t_2 é observada, pode-se com certa probabilidade observar t_1 no mesmo recurso*". Deste modo, regras de associação da forma $t_i \rightarrow t_j$ podem ser extraídas.

A fim de selecionar as regras de associação baseadas em *tags* mais relevantes em uma base de dados, as medidas de significância mais conhecidas são *suporte* e *confiança*. Neste trabalho, considera-se regras de tamanho 2. O *suporte* de uma regra de associação $t \rightarrow t'$ é definido pelo número de transações que contém ambas as *tags* t e t', ou seja, a frequência com que t e t' coocorrem (Equação 4.11). A *confiança* estima a confiabilidade de uma regra de associação, computada como a porcentagem de transações que contém as *tags* t e t' (probabilidade condicional $P(t'|t)$), conforme Equação 4.12.

$$\textit{sup}(t \rightarrow t') = sup(t' \cup t') \tag{4.11}$$

$$\textit{confiança}(t \rightarrow t') = P(t'|t) = \frac{sup(t \cup t')}{sup(t)} \tag{4.12}$$

Limiares para *suporte* e *confiança* devem ser estimados empiricamente para selecionar o conjunto de regras que melhor sugere relacionamentos de subordinação. O melhor valor de *suporte* e *confiança* toma como base as seguintes observações:

- regras com baixo valor de *suporte* não traduzem relações de subordinação porque são regras fracas, ou seja, a frequência com que t e t' coocorrem não é expressiva na base de dados;
- regras com alto valor de *suporte*, mas que foram marcados por uma minoria absoluta de usuários, não expressam um consenso coletivo de que t e t' são de fato potenciais candidatas a uma relação de subordinação;
- quanto maior o valor da *confiança*, em teoria, maior a probabilidade de que a regra

$t \rightarrow t'$ represente uma relação de subordinação;

- regras com alto valor de *confiança* nem sempre são unanimidade como potencial indicativo de subordinação entre *tags* caso o valor do *suporte* seja baixo. Como exemplo, um único usuário pode marcar 1 ou 2 recursos com duas *tags* que só ele tenha usado, tendenciando para que a regra tenha um alto valor de *confiança*.

Embora *suporte* e *confiança* não sejam medidas diretas para detectar relações de subordinação, elas são importantes para selecionar regras candidatas, sob um limiar mínimo, para uma decisão subsequente de conexão de subordinação entre *tags*.

Coeficiente de Jaccard vs. Cosseno da Similaridade vs. Coocorrência

A descoberta de relações hierárquicas entre *tags* demanda maior complexidade se comparada às práticas atuais empregadas para detectar *tags* sinônimas. Isso porque normalmente são concebidos algoritmos bem elaborados capazes de determinar se um par de *tags* pode ou não ser conectado hierarquicamente (MARINHO; BUZA; SCHMIDT-THIEME, 2008; BENZ, 2007; CAI et al., 2011; MEO; QUATTRONE; URSINO, 2009). Mas, antes que esta decisão seja tomada, é necessário que primeiramente um par de *tags* esteja semanticamente relacionado para que depois possa ser mensurado o grau de especialização e generalização entre ambas.

Abbasi (2011) expõe em seu trabalho várias medidas de similaridade/parentesco fundamentadas no princípio de coocorrência, a saber: contagem simples de coocorrência, cosseno da similaridade e os coeficientes *Dice* e *Jaccard*. Entre as medidas citadas, observou-se a aplicação do coeficiente *Jaccard* por Marinho, Buza e Schmidt-Thieme (2008), Meo, Quattrone e Ursino (2009) como medida de distância semântica em fase anterior a de construção de hierarquia.

O coeficiente *Jaccard* foi introduzido como atributo de aprendizado na abordagem CPDST a partir dos experimentos de detecção de relações de subordinação, motivado pela necessidade de investigar sua possível contribuição para a melhoria da acurácia do modelo preditor. O coeficiente *Jaccard* entre duas *tags* é calculado em função do conjunto de recursos marcados por t e t', respectivamente, R_t e $R_{t'}$. A Equação 4.13 então define a função de similaridade $Jaccard$ como a razão da interseção dos conjuntos R_t e $R_{t'}$ pela

união dos conjuntos R_t e $R_{t'}$.

$$sim(t, t') = Jaccard(t, t') = \frac{|R_t \cap R_{t'}|}{|R_t \cup R_{t'}|} \tag{4.13}$$

Os valores do coeficiente *Jaccard* variam no intervalo de 0 (máxima dissimilaridade) a 1 (máxima similaridade). No exemplo da Figura 4.2, a similaridade $Jaccard(t_1, t_2) = 0.67$ e $Jaccard(t_2, t_3) = 0.33$.

O coeficiente de *Jaccard* juntamente com as medidas de coocorrência *tag-tag* e similaridade do cosseno são pretendidas para quantificar similaridade semântica entre *tags*. Logo, utilizar todas essas medidas como atributos seria redundante. Embora o coeficiente de *Jaccard* tenha sido usado por Marinho, Buza e Schmidt-Thieme (2008) e Meo, Quattrone e Ursino (2009) em seus respectivos algoritmos de construção de hierarquias, optou-se neste trabalho por utilizar como atributo a função da similaridade do cosseno (cf. Equação 4.5), tendo em vista que a função do cosseno apresentou melhores resultados nos experimentos de seleção de atributos (cf. Seção 6.5). Coocorrência *tag-tag*, outra métrica concorrente, está sendo contemplada indiretamente devido a sua correspondência com a equação do atributo *suporte* (Equação 4.11). Mesmo ciente de que o cosseno, *Jaccard* ou coocorrência não são medidas específicas para detecção de relações hierárquicas, acredita-se que em combinação com outras medidas mais específicas, os atributos escolhidos podem ajudar a melhorar a acurácia de classificação geral, especialmente nos casos em que os atributos hierárquicos mais específicos falham na detecção de relações de subordinação.

Especialização e Generalização

Uma relação hierárquica é estabelecida entre dois conceitos quando um deles tem sentido mais genérico e o outro mais específico, com semântica em comum. Em uma folksonomia, para que uma *tag* t seja candidata a hiperônimo de t' é necessário verificar o quanto o significado de t incorpora o significado de t', e vice-versa.

Li et al. (2007 apud MEO; QUATTRONE; URSINO, 2009) defendem a ideia de que a subordinação hierárquica de uma *tag* t em relação a t' depende da fração de recursos de t que também foi anotada com t', e vice-versa. Deste modo, se a fração é numerosa, é possível concluir que a maioria dos recursos anotados com t também estão anotados com t'. Por outro

lado, se a fração é pequena, então a maioria dos recursos anotados com t não estão anotados com t'. E então, ou t e t' tem diferentes significados ou t' é mais específico do que t.

Meo, Quattrone e Ursino (2009) formalizam este raciocínio propondo uma medida chamada de *grau de inclusão*, para mensurar um nível de especificidade, e *grau de generalização*, para mensurar um nível de generalidade. Em concordância com estas definições, neste trabalho são adotados os termos *especialização* e *generalização* como atributos mais específicos para detecção de relações hierárquicas, os quais são definidos matematicamente por

$$\textit{especialização}(t, t') = \frac{|R_t \cap R_{t'}|}{|R_t|}, \tag{4.14}$$

$$\textit{generalização}(t, t') = \frac{|R_t \cap R_{t'}|}{|R_{t'}|} = \textit{especialização}(t', t), \tag{4.15}$$

em que R_t e $R_{t'}$ correspondem ao conjunto de recursos anotados por t e t', respectivamente. Um valor alto como resultado da Equação 4.14, pode implicar que t tem alta inclusão em t'. Note que a ordem de entrada das *tags* nas funções *especialização* e *generalização* influencia o valor de saída, de modo que $\textit{especialização}(t, t') \neq \textit{especialização}(t', t)$. No exemplo da Figura 4.2, $\textit{especialização}(t_1, t_2) = 1$ e $\textit{generalização}(t_1, t_2) = 0.67$

Sobreposição Mútua

Na Seção 4.2.1 é apresentada uma explanação sobre a noção de sobreposição mútua para a presunção de semântica. Cai et al. (2011) aplicam essa técnica probabilística para construir uma hierarquia de conceitos usando *tags* de folksonomias. Similarmente ao que pode ser observado na contextualização dos atributos *especialização* e *generalização* descritos anteriormente, a sobreposição mútua também mede a fração de recursos anotados por duas *tags* distintas. No entanto, as relações hierárquicas são estabelecidas de acordo com o resultado dos valores da tupla (a, b) em conformidade com seus respectivos limiares α e β.

Observe que, enquanto para a detecção de sinonímia o conjunto interseção $R_t \cap R_{t'}$ deve apresentar grande representação em R_t e $R_{t'}$, sinalizando que recursos anotados por t também são anotados por t', para presumir relações hierárquicas é necessário que $R_t \cap R_{t'}$

tenha pequena ocorrência em t mas grande cobertura no conjunto de t'. Então, após computar as probabilidades de sobreposição mútua a e b para $t \succ t'$, estas podem ser interpretadas como segue: se a é "grande" e b é "pequeno", recursos anotados com t' são prováveis de serem anotados com t, mas não o contrário. Deste modo, supõe-se que t é hiperônimo de t'.

Além das estimativas de probabilidade a e b, também é adicionado o mesmo atributo $overlap$ descrito na Seção 4.2.1, tendo em vista que um dos valores produzido pelo atributo sugere a indicação de prováveis relações de subordinação. Acrescenta-se, ainda, o valor de saída igual a 0 para indicar $|R_t| < |R_{t'}|$, ou seja, se t subordina t', teoricamente espera-se que t anote uma quantidade de recursos maior do que t'. Logo, se essa suposição não é atestada, o atributo designa um valor particular. Para a detecção de relações de subordinação, este atributo é referenciado por overlap^{hyp}.

Busca Hierárquica: *tsearch*

Visto que não existem muitos atributos para a detecção de relações hierárquicas na literatura, às vezes é necessário conceber atributos com base em abordagens existentes. O algoritmo *Taxonomy Learning based on Frequent Itemsets* desenvolvido por Marinho, Buza e Schmidt-Thieme (2008) (cf. Apêndice C) realiza um processo iterativo para a construção de uma taxonomia. Basicamente, os conjuntos de itens mais frequentes são extraídos iterativamente tal que t é considerado uma super-*tag* (super-conceito) de t' se, em cada iteração, uma condição de limiar de aresta entre ambas for satisfeito. O grafo resultante contém um conjunto de vértices (*tags*) conectados por arestas que representam uma cadeia de conceitos hierárquicos. O grafo não contém múltiplas relações de herança, ou seja, é modelado como uma árvore.

O algoritmo *Taxonomy Learning* foi reproduzido para conceber um novo atributo denominado *busca hierárquica*, ou simplesmente *tsearch*, e então tirar vantagem da rede de conexões hierárquicas produzida. A ideia do atributo *tsearch* é certificar a existência de conexão que une t a t', de tal forma que o tamanho do caminho entre dois vértices é medido pelo número de arestas existentes entre origem e destino.

Em relação à implementação original, adicionou-se um algoritmo de busca em largura em grafos para encontrar o número de arestas entre dois vértices t e t'. Seja $G = \{V, E\}$ um grafo composto por um conjunto de vértices V e $E \subseteq V \times V$ um conjunto de arestas. O peso

w, para qualquer aresta $(v, v') \in E$, é definido como $w : E \rightarrow 1$, ou seja, qualquer aresta no grafo possui peso de uma unidade. O caminho partindo de t_i para t_j é a sequência de vértices em $P = (v_1, v_2, \ldots, v_n)$ em que $v_1 = t_i$ e $v_n = t_j$. Deste modo, $tsearch$ é definido como segue.

$$tsearch(t_i \rightsquigarrow t_j) = \sum_{i=1}^{n-1} w(v_i, v_{i+1}). \tag{4.16}$$

.

Assim, se existe um caminho a partir da *tag* t_i até t_j no grafo, a suposição intuitiva é que t_i é hiperônimo de t_j. Caso contrário, $tsearch(t_i \rightsquigarrow t_j) = 0$. Quanto maior o valor de *tsearch*, mais fraca é relação de hierarquia entre as *tags*. Um valor ideal para *tsearch* é 1 (relação direta). No exemplo da Figura 4.3, $tsearch(t_1, t_7) = 2$ e $tsearch(t_3, t_5) = 0$. A ordem de entrada das *tags* é importante, visto que a função $tsearch$ é assimétrica. Logo, $tsearch(t_7, t_1) = 0$.

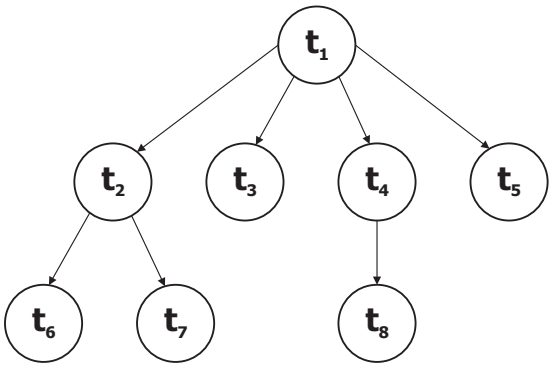

Figura 4.3: Taxonomia em árvore produzida pelo algoritmo *Taxonomy Learning*.

4.3 Rotulação de Instâncias

Após a extração dos atributos (cf. Seção 4.2), instâncias são rotuladas como positivas ou negativas conferindo se o par de *tags* que representa cada instância representa uma relação de sinonímia ou subordinação por meio de um módulo de rotulação estensível. Atualmente, este módulo incorpora acesso ao WordNet e ConceptNet para definir automaticamente o rótulo das instâncias de entrada.

Esta tarefa é realizada da seguinte maneira: dadas duas *tags* $t, t' \in T$ com $t \neq t'$, submete-se uma consulta (*query*) ao módulo de rotulação usando t como entrada. No caso

do WordNet, o módulo de rotulação retorna todos os *synsets* que representam conceitos sinônimos associados à *tag* t. Para recuperar os hipônimos, uma iteração a mais é realizada: a partir dos *synsets* retornados, são recuperados os *synsets* subordinados que são hipônimos de t. Assim, o par (t, t') é rotulado como positivo se t' está contido no conjunto de sinônimos ou hipônimos diretos de t. Caso contrário, o par (t, t') é rotulado como negativo. Considerando o ConceptNet, requisições HTTP via ConceptNet5 API verificam se as relações IsA(t', t) ou Synonym(t, t') são verdadeiras.

Na Seção 2.3 os dicionários/*thesaurus* WordNet e ConceptNet são apresentados em maiores detalhes, inclusive no que diz respeito às APIs utilizadas para realizar acesso às referidas bases de conhecimento. Particularmente, o resultado de uma busca ao ConceptNet pode ser um conjunto vazio (não foi constatada a relação de subordinação) ou uma lista de estruturas de dados JSON (*JavaScript Object Notation*), organizada em ordem decrescente de *score*, contendo todos os campos de uma aresta ConceptNet. O *score* é uma medida que quantifica o quão consistente o conceito t' recuperado é para a *query* t, de acordo com a relação estabelecida. As arestas são fragmentos de conhecimento de senso comum que interconectam conceitos com uma relação particular. As relações ConceptNet exploradas neste trabalho são *is-a* e *Synonym*.

4.4 Classificação Multiclasse em Dados Desbalanceados

Uma vez que a abordagem CPDST estava sendo empregada para detectar separadamente relações de sinonímia e de subordinação, decidiu-se então tentar convergir os dois problemas de classificação binária para um solução que fosse capaz de caracterizar 3 classes[24] Na literatura de AM, a aplicação da técnica de classificação multiclasse apresenta-se como um artifício natural para lidar com essa questão. Objetivamente, a técnica de classificação multiclasse consiste em classificar uma instância em uma das N classes diferentes (com $N > 2$). Certamente, o aprendizado a partir de múltiplas classes implica em maior dificuldade de aprendizado para os algoritmos de classificação, uma vez que a borda de decisão entre as classes pode ser sobreposta, causando uma diminuição no desempenho do

[24] 1 classe para indicar sinônimos, 1 classe para indicar relação de subordinação e 1 classe para indicar exemplos que não expressam as duas relações semânticas abordadas. e, ao mesmo tempo, investigar a possibilidade de obter algum ganho na acurácia da classificação das instâncias minoritárias.

classificador (FERNÁNDEZ; JESUS; HERRERA, 2010). Além disso, alia-se o fato de que muitos algoritmos de classificação foram concebidos basicamente para resolver problemas de classificação binária (ALY, 2005) (e.g., k-Nearest Neighbor (kNN) e SVM), enquanto outros podem ser estendidos para problemas multiclasse (e.g., árvore de decisão e redes neurais) (ROCHA; GOLDENSTEIN, 2014).

Uma forma de reduzir as dificuldades inerentes ao problema de classificação multiclasse é decompô-lo em vários problemas de classificação binária, os quais são mais fáceis de discriminar por meio da aplicação de técnicas de binarização de classe (EICHELBERGER; SHENG, 2013). As estratégias mais populares são: (i) um contra todos (*One-Against-All* - OAA) e um contra um (*One-Against-One* - OAO). A estratégia OAA consiste em construir um classificador para cada classe do problema, considerando os exemplos da classe atual como positivos e das demais classes como negativos (RIFKIN; KLAUTAU, 2004). Na estratégia OAO, um classificador binário é treinado para cada par de classes possível, ignorando os exemplos que não pertencem às classes relacionadas (FERNÁNDEZ; JESUS; HERRERA, 2010).

Muitos problemas de classificação do mundo real são em sua natureza multiclasse, tais como categorização de imagens (DENG et al., 2010) e classificação de proteína (ZHAO et al., 2008). Apesar da literatura prover soluções para lidar com as peculiaridades do problema de classificação multiclasse, muitas pesquisas (FERNÁNDEZ; JESUS; HERRERA, 2010; FERNÁNDEZ et al., 2013; JEATRAKUL; WONG, 2012; PHOUNGPHOL; ZHANG; ZHAO, 2012; WANG; YAO, 2012) destacam que manusear o problema de classificação multiclasse em dados desbalanceados é um desafio a ser superado. Isso ocorre porque: (i) soluções para tratamento de dados desbalanceados em problemas binários não são diretamente aplicáveis a problemas multiclasse, ou então podem alcançar um desempenho abaixo do que normalmente se espera (ZHOU; LIU, 2006); (ii) o alto nível de relacionamento entre *overlapping* e desbalanceamento pode impossibilitar a distinção entre as classes envolvidas; e (iii) um alto grau de desbalanceamento de classe pode aumentar ainda mais a dificuldade de produzir bons modelos preditores (ZHOU; LIU, 2006).

Uma revisão na literatura permitiu deduzir que o problema de classificação multiclasse em dados desbalanceados e com presença de *overlapping* é um campo de estudo em aberto e sua resolução dependente das características de cada aplicação. Trabalhos de pesquisa

focam em diferentes linhas de atuação, combinando e/ou adaptando uma série de técnicas a fim de obter resultados de classificação satisfatórios. Jeatrakul e Wong (2012) propõem um algoritmo que emprega as técnicas de balanceamento SMOTE e *Complementary Neural Network* (CNNT) e um classificador de rede neural artificial incorporado à técnica OAA. Fernández et al. (2013) desenvolveram um estudo experimental com o intuito de determinar as melhores abordagens a serem aplicadas em cenários de classificação multiclasse em bases de dados desbalanceadas. O objetivo se concentra em determinar a melhor combinação entre técnicas de binarização (OAA e OAO), utilizando os algoritmos de árvore de decisão, SVM e kNN, em conjunto com técnicas de balanceamento (*undersampling* e *oversampling*) ou com o uso da técnica ASC.

Phoungphol, Zhang e Zhao (2012) formulam um problema SVM multiclasse aplicando a técnica ASC para melhorar o desempenho de classificação em uma base de dados biomédica desbalanceada. Wang e Yao (2012) estudam os desafios impostos pelo problema de desbalanceamento multiclasse, analisando o impacto da utilização de técnicas de balanceamento (*undersampling* e *oversamling*) no cenário multiclasse. Com base nos resultados, os autores então defendem o uso do proposto *ensemble* algoritmo Adaboost.NC (*base learner* C4.5), sem aplicação de técnica de decomposição de classe, combinado com o método *random oversampling* para manusear o problema de desbalanceamento multiclasse. Zhang et al. (2013) abordam o problema de desbalanceamento multiclasse e *overlap* com a proposta de um sistema de suporte à decisão que adota técnicas de AM. O sistema incorpora uma modificação do algoritmo embasado em regras RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) para prover múltiplas predições combinado com uma técnica híbrida de balanceamento (SMOTE + Tomek Link), e o método *ensemble* BWRF (*Balanced and Weighted Random Forests*) para tratar os problemas de ruídos na base de dados, *overlap* e desbalanceamento de classes.

Zhao et al. (2008) apresentam uma nova abordagem para a tarefa de classificação de proteína em dados desbalanceados. O algoritmo proposto implementa uma técnica de amostragem híbrida que integra SMOTE e *undersampling* para tratar o desbalanceamento juntamente com um classificador *ensemble* denominado *EnClassifier*. O método OAA é empregado para reduzir o problema multiclasse em um conjunto de problemas binários. Por fim, Fernández, Jesus e Herrera (2010) propõem uma metodologia que emprega OAO para

decompor o problema em subproblemas binários desbalanceados e SMOTE é aplicado para balancear os dados antes do processo de aprendizado. O estudo empírico é desenvolvido com a aplicação do algoritmo linguístico *Fuzzy Rule Based Classification System*.

Ciente de todos os desafios que normalmente caracterizam a implementação de um problema de classificação multiclasse em dados desbalanceados, esforços foram dedicados no decorrer do trabalho para converter a presente abordagem binária em um problema de classificação multiclasse. Entretanto, os resultados preliminares utilizando as orientações mais diretas providas nos trabalhos relacionados (combinações de técnicas de balanceamento, métodos de decomposição de problemas multiclasse para binário, entre outras) não surtiram efeito para que ao menos fosse mantida a acurácia conseguida com a aplicação do modelo de classificação binária. Algumas soluções que podem ser examinadas com maior ênfase são complexas, além disso não há garantia de que a estratégia a ser adotada seja ideal para o cenário do problema em questão, pois o forte desbalanceamento aliado à incidência de *overlap* entre classes podem sugerir a criação de uma solução diferente.

Diante dos resultados desencorajadores, preferiu-se não seguir adiante com o propósito de migração do modelo de classificação binária para multiclasse e adiá-lo para trabalhos futuros, visto que não está contemplado no escopo desse trabalho resolver o problema de classificação multiclasse em dados desbalanceados com a presença de *overlapping*.

Uma providência tomada para simular o cenário do problema de classificação multiclasse foi combinar múltiplas execuções de classificadores binários em uma mesma unidade de software, reproduzindo o funcionamento de uma árvore de decisão para determinar a classe das instâncias. A predição é realizada da seguinte forma: a princípio, submete-se uma instância de entrada (t, t') para o modelo de detecção de sinonímia. Na ocasião de (t, t') representar um caso positivo, um rótulo de classe que designa sinonímia é atribuído e a tarefa de classificação é concluída. Caso contrário, a mesma instância é submetida ao modelo de detecção de relação de subordinação. Se for prevista como um caso positivo, a instância é qualificada com um rótulo de classe que designa a suposição de relação de subordinação, concluindo desta forma a tarefa de classificação. Na eventualidade da instância representada pelo par de *tags* (t, t') não ser considerada como um caso de sinonímia ou subordinação, um rótulo de classe que denota a classe majoritária (inexistência de ambas as relações semânticas) define o resultado da predição.

4.5 Considerações Finais

Neste capítulo, foi apresentada uma abordagem para detecção de relações semânticas entre *tags* de folksonomias, especificamente sinonímia e subordinação. Com base na técnica de aprendizado supervisionado, descreveu-se a formulação geral do problema como uma tarefa de classificação binária. Os atributos utilizados para treinamento foram extraídos por meio da aplicação de medidas de similaridade/parentesco/distância entre pares de *tags*, respaldadas nos trabalhos relacionados.

Embora as relações semânticas do tipo sinonímia e subordinação estejam sendo priorizadas neste trabalho, a abordagem proposta é genérica e pode ser estendida para explorar outros tipos de relações semânticas, desde que sejam extraídos os atributos apropriados. No campo prático, a proposta pode ser facilmente empregada para dar suporte a diversos tipos de aplicações ou novos serviços como, por exemplo, expansão de consultas, construção de hierarquias, recomendação de *tags*, sugestão de *tags* semanticamente relacionadas, entre outros.

A simplicidade da concepção do modelo binário de predição semântica não implica em igual facilidade de obter um bom desempenho na qualidade da predição. Especificamente para a captura de relações de subordinação, os atributos são escassos e a própria relação em si é mais complexa de se mensurar. Portanto, dependendo do tipo de semântica a ser explorado, muitos experimentos podem ser efetuados para determinar o melhor conjunto de atributos, examinar classificadores, ajustar parâmetros, amenizar problemas de *overlapping* e desbalanceamento, entre outros. Por fim, justifica-se o motivo de combinar os dois modelos binários de predição semântica (sinonímia e relação de subordinação) para desempenhar a mesma função de um problema de classificação multiclasse. O próximo capítulo discorre sobre a metodologia utilizada para avaliar a abordagem CPDST e os *baselines*.

Capítulo 5

Metodologia Experimental

O processo KDD, descrito na Seção 2.6.1 e sumarizado na Figura 2.2, foi usado como metodologia para descoberta de pares de *tags* convenientes a uma relação semântica de sinonímia e de subordinação. Neste capítulo são apresentados os detalhes acerca da configuração experimental e do protocolo de avaliação, para que os experimentos possam ser reproduzidos em condições controladas e atender aos objetivos específicos determinados neste trabalho. Além disso, também é descrito como são abordados os problemas de desbalanceamento e *overlapping* de classes que são intrínsecos à formulação do problema de detecção de relações semânticas.

5.1 Preparação dos Dados

Para avaliação da abordagem de aprendizado de relações semânticas, conduzimos os experimentos utilizando *snapshots* de duas folksonomias populares: BibSonomy e Delicious. O BibSonomy (BENZ et al., 2010) é um sistema de gerenciamento de publicações científicas e *bookmark* social que fornece seus dados publicamente na web para fins de pesquisa[25]. Deste modo, outros pesquisadores podem utilizar os mesmos dados para reproduzir os experimentos. Utilizou-se um *snapshot* da base de dados do BibSonomy de 1º de Julho de 2011. O Delicious é um sistema de gerenciamento e compartilhamento de

[25] *Snapshots* do BibSonomy estão disponíveis para download em <http://www.kde.cs.uni-kassel.de/bibsonomy/dumps>.

bookmark. Empregou-se a base de dados[26] utilizada no contexto do projeto TAGora[27], visto que o Delicious não disponibiliza bases de dados oficiais para *download*. A base de dados foi coletada automaticamente por um *web crawler* diretamente no portal do Delicious entre os anos de 2006 e 2007.

Embora as bases de dados normalmente apresentem centenas de milhares de *tags*, a maioria absoluta tem baixa frequência de utilização na folksonomia, ou seja, são usadas por poucos usuários. A Figura 5.1, originalmente publicada por Abbasi (2011), mostra um gráfico que caracteriza a frequência de usuários que utilizam uma *tag* particular em uma folksonomia. Nesta figura, observa-se que poucas *tags* são de fato usadas por muitos usuários e que a maioria das *tags* são usadas apenas por alguns usuários. Por isso, optou-se primeiramente por selecionar as *top-k tags* mais frequentes que atingem uma cobertura acima de 80% das anotações.

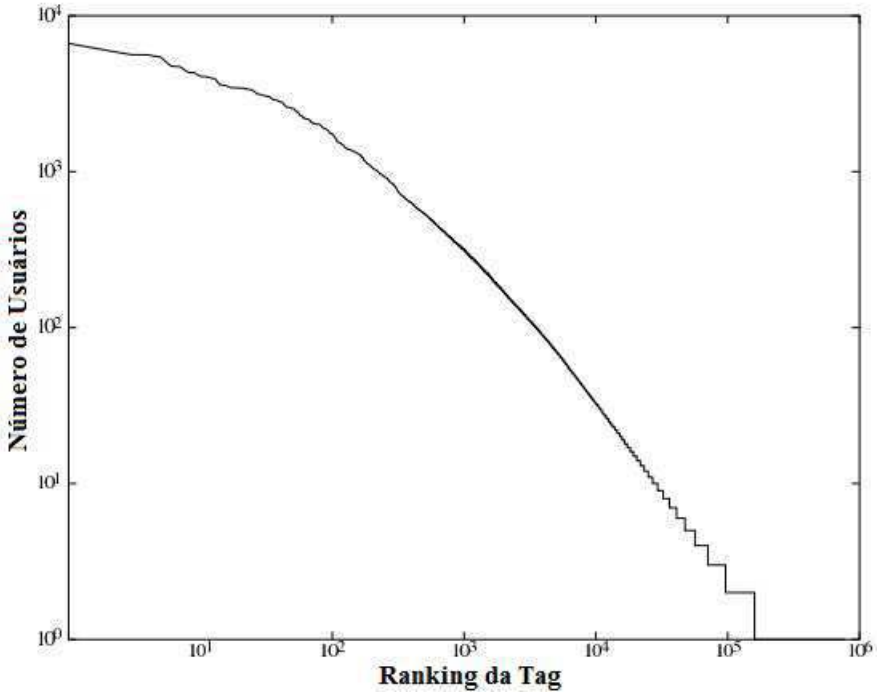

Figura 5.1: Número de usuários por *tag*.
Fonte: Abbasi (2011)

No BibSonomy, as 10.000 *tags* mais populares correspondem a 5% do número total de *tags* e estão associadas a mais de 90% dos recursos existentes. No Delicious, as 6.000 *tags*

[26] <https://www.uni-koblenz-landau.de/de/koblenz/fb4/AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_html>

[27] <http://www.tagora-project.eu/>

mais populares estão associadas a 80.1% das anotações e representam 0,25% do total de *tags*. Com base nestas observações, considera-se a quantidade de *tags* especificada em cada base de dados significativa para fins experimentais.

A partir das *tags* selecionadas, iniciou-se um processo de limpeza para remoção de símbolos especiais tais como vírgulas, ponto e vírgula, colchetes e chaves. Todas as *tags* foram normalizadas para caracteres minúsculos. *Tags* atribuídas automaticamente[28] pelo sistema foram desconsideradas. Finalmente, apenas as *tags* reconhecidas pelo WordNet e ConceptNet foram mantidas, visto que tais dicionários foram usados para rotular as instâncias de treinamento em termos de sinonímia e relação de subordinação. Deste ponto em diante, essas *tags* resultantes do processo de mapeamento são referenciadas por *tags do experimento*. As estatísticas da base de dados original e pré-processada estão resumidas na Tabela 5.1.

Tabela 5.1: Características dos dados.

	$\|R\|$	$\|U\|$	$\|T\|$	$\|Y\|$
Dados originais				
BibSonomy	852.292	7.243	216.094	2.740.834
Delicious	17.262.480	532.924	2.481.698	140.126.586
Dados pré-processados				
BibSonomy	338.492	4.800	3.177	722.999
Delicious	13.487.483	501.683	4.790	102.323.655

5.2 Construção das Instâncias de Treinamento/Teste

Para determinar a população de instâncias usada para fins experimentais, efetuou-se um arranjo simples das *tags* do experimento duas a duas. Entretanto, apenas os pares de *tags* que coocorreram ao menos uma vez em um dado recurso foram selecionados. Em seguida, para cada par de *tags* que constitui uma instância, foram extraídos os atributos que capturam diferentes particularidades relacionadas à semântica alvo. Excepcionalmente, para os atributos *overlap* e *tsearch*, foram ajustados empiricamente os limiares α, β e e

[28]Por exemplo, a *tag* $imported$ (a mais popular) foi eliminada da base de dados do BibSonomy, uma vez que é automaticamente atribuída pelo sistema quando importa recursos de outros sistemas como, por exemplo, da bibliografia em ciências da computação DBLP (*DataBase Systems and Logic Programming*).

(cf. Apêndice C). Todos os atributos foram normalizados para valores compreendidos no intervalo [0,1]. Por fim, um rótulo de classe foi adicionado a cada instância, conforme a metodologia descrita na Seção 4.3. A realização dessa sequência de tarefas resultou em um total de 503.676 instâncias no BibSonomy e 13.847.556 instâncias na base do Delicious disponíveis para fins de treinamento e teste.

A fim de assegurar um grau mínimo de parentesco para estimativa de semântica, um filtro de coocorrência foi aplicado para reter os pares de *tags* mais convenientes à tarefa de treinamento do modelo de predição. O valor ideal do filtro de coocorrência foi obtido mediante análise dos seguintes critérios: (i) menor razão de instâncias positivas/negativas, ou seja, menor nível de desbalanceamento, e (ii) maior quantidade de instâncias positivas remanescentes. Excepcionalmente, para detecção de relações de subordinação, além do filtro de coocorrência (também definido como o $suporte$ de uma regra de associação) foi aplicado o filtro de confiança para prover maior confiabilidade na seleção de pares de *tags* candidatas. A Tabela 5.2 exibe os valores estimados para os parâmetros de atributo e filtragem de instâncias via coocorrência/confiança, nas bases de dados do BibSonomy e Delicious.

Tabela 5.2: Ajuste Experimental.

	Sinonímia		Subordinação	
	BibSonomy	Delicious	BibSonomy	Delicious
filtro de coocorrência	≥ 10	≥ 10	≥ 5	≥ 20
filtro de confiança	-	-	≥ 0.1	≥ 0.01
sobreposição mútua	$\alpha = 0.35$ $\beta = 0.25$	$\alpha = 0.01$ $\beta = 0.15$	$\alpha = 0.4$ $\beta = 0.3$	$\alpha = 0.15$ $\beta = 0.14$
tsearch (limiar de aresta)	-	-	$e = 1.5$	$e = 1.4$

Ao final da etapa de coleta de dados, a base de dados resultante, também chamada de base de dados de treinamento, apresenta os dados na forma atributo-valor esperada pelos algoritmos de classificação. Cada experimento de detecção de semântica (sinonímia e relação de subordinação) possui sua base de dados de treinamento particular, a qual pode variar em quantidade de instâncias e rotulação.

5.3 Tratamento do Desbalanceamento e *Overlap* de Classes

Após o processo de geração e rotulação das instâncias de treinamento (Seção 5.2), observou-se a presença de um desbalanceamento de classes severo. Este fato não chega a ser uma surpresa, tendo em vista que poucos pares de *tags* expressam uma relação de sinonímia ou de subordinação no universo de instâncias. Para exemplificar, após a fase de geração de instâncias (sem filtro) utilizando a base de dados do BibSonomy, foram identificadas 1.634 instâncias como positivas para sinonímia e 502.042 negativas, o que implica em uma razão de desbalanceamento de 1:307 (1 positiva para 307 negativas). O mesmo problema de desbalanceamento é observado na base de dados para detecção de relações de subordinação, em uma razão menor, cujo nível de desbalanceamento é de 1:187 (1 positiva para 187 negativas).

O desbalanceamento é ainda mais crítico com a base de dados do Delicious, cuja quantidade de instâncias é 27 vezes maior se comparada à quantidade resultante na base de dados do BibSonomy. Após a tarefa de rotulação automática para sinonímia, observou-se uma razão de desbalanceamento de 1:1.340 (1 positiva para 1.340 negativas). Diante da perspectiva exposta, como ressaltado na Seção 2.6.4, o classificador é fortemente tendenciado a predizer sempre a classe majoritária, justamente a que denota menos interesse no domínio do problema investigado neste trabalho. Logo, o cenário exige a aplicação de uma solução que reduza o conjunto total de instâncias em sua parcela mais significativa, o que diretamente contribui para a redução do nível de desbalanceamento e provê melhores condições de efetuar a tarefa de classificação.

O problema de desbalanceamento de classes é mais um tema inserido no universo do AM. Algumas abordagens para tratar essa questão são apresentadas na Seção 2.6.4. Durante os experimentos, a maioria dessas abordagens foi aplicada e analisada como segue: (a) detecção de sinonímia: *random undersampling* e Tomek Link, e (b) detecção de relações de subordinação: Tomek Link, CNN, SMOTE e *random undersampling*. Foram empregadas mais técnicas de re-amostragem nos experimentos para detecção de relações de subordinação porque esta foi a semântica mais difícil de predizer na tarefa de classificação.

A base de dados experimental também é infligida pelo problema de sobreposição de classes (*class overlapping*). Esta se manifesta com maior intensidade na base de

dados utilizada para detecção de relações de subordinação. Dentre as técnicas de re-amostragem analisadas, Tomek Link é a mais efetiva para lidar com o problema porque consegue identificar e eliminar exemplos negativos ruidosos que confundem o classificador na diferenciação entre exemplos positivos e negativos.

Outra solução adotada para lidar com a questão do desbalanceamento diz respeito à aplicação de um filtro de pré-processamento de instâncias com base na frequência de coocorrência (cf. Tabela 5.2). Este procedimento é pertinente porque contribui para tratar ao mesmo tempo duas questões relevantes: (i) impor requisito mínimo de suposição de parentesco, e (ii) reduzir a desproporção entre o número de exemplos de cada classe. Visto que a base de dados experimental é muito esparsa, utilizou-se a parte mais densa dos dados para facilitar o aprendizado dos algoritmos. A Tabela 5.3 resume estatísticas acerca da redução do desbalanceamento de classes nas bases de dados do BibSonomy e Delicious. Mesmo após a aplicação do filtro de coocorrência, o nível de desbalanceamento ainda é elevado, principalmente na base de dados do Delicious, porém com uma proporção atenuada em relação ao desbalanceamento apresentado nos dados brutos.

Tabela 5.3: Estatística de Desbalanceamento.

Quantitativo de instâncias	BibSonomy		Delicious	
(dados brutos)	503.676		13.847.556	
Alvo semântico	Sinonímia		Subordinação	
	BibSonomy	Delicious	BibSonomy	Delicious
# Instâncias positivas	1.634	10.328	2.681	11.485
Desbalanceamento natural	1(+):307(-)	1(+):1.340	1(+):187(-)	1(+):1.285
# Instâncias após filtro	37.296	3.753.046	14.287	869.257
# Instâncias positivas	452	3.756	221	4.170
Desbalanceamento final	1(+):82(-)	1(+):999(-)	1(+):64(-)	1(+):492(-)

5.4 Protocolo de Avaliação

Para avaliar a capacidade de generalização do modelo de predição construído pela abordagem CPDST, os experimentos são realizados utilizando duas técnicas de avaliação distintas: *holdout* estratificada e *k-fold cross-validation*. A técnica *holdout* divide o conjunto

total de dados em conjuntos mutuamente exclusivos, em uma razão de 2/3 para treinamento e 1/3 para teste, mantendo a mesma proporção entre as classes nos dois conjuntos (WITTEN; FRANK, 2011). Uma porção dos dados de treinamento (1/3) é reservada para o conjunto de validação, conjunto este usado para realizar ajustes no modelo ou estimar o classificador mais adequado para o problema. A técnica *k-fold cross-validation* consiste em particionar o conjunto de dados em k subconjuntos mutuamente exclusivos de mesmo tamanho (*folds*), de maneira que cada iteração utilize $k-1$ subconjuntos para treino e um conjunto para teste em um processo de rotação. Este método é mais efetivo porque reduz a variância na avaliação do modelo preditor (BENGIO; GRANDVALET, 2004).

Para contornar os problemas de classificação normalmente causados por distribuições de classes desbalanceadas, as seguintes estratégias foram consideradas:

1. **Busca gulosa para identificação do nível de balanceamento que proporciona o melhor desempenho de classificação**: o nível de desbalanceamento é ajustado iterativamente no conjunto de validação partindo da razão perfeitamente balanceada (1+/1-) até o nível natural de desbalanceamento. Ao identificar o nível de desbalanceamento que produz a melhor acurácia na classificação das instâncias positivas, este é considerado para realização de treinamento e teste do modelo de predição. Diferentes técnicas de balanceamento são observadas sob essa estratégia;

2. **Aplicação direta dos algoritmos de classificação nas bases de dados filtradas**: ao invés de medir o desempenho dos classificadores a partir de uma distribuição de dados modificada por um método de balanceamento, o modelo de predição é treinado e avaliado utilizando todas as instâncias da base de dados experimental alvo.

Na base de dados do Delicious, mesmo após a aplicação do filtro de coocorrência, constata-se que o número remanescente de instâncias de treinamento ainda é muito elevado (cf. Tabela 5.2), o que impede a execução da tarefa de Mineração de Dados pelos *softwares* acadêmicos populares devido a problemas internos de alocação de memória. Para viabilizar o treinamento e teste, foram realizados experimentos de calibração com o *software* Weka. Observou-se que uma quantidade média de 40.000 instâncias possibilita a execução de sucessivas operações de treinamento e teste com diferentes algoritmos de classificação. Deste modo, amostras de dados foram selecionadas aplicando a técnica

random undersampling, na qual todos os exemplos positivos são mantidos fixos e as instâncias negativas são acrescentadas aleatoriamente até atingir o limiar da amostra.

Visando mitigar qualquer viés causado por uma amostra em particular escolhida pelo método *holdout*, o processo de geração de amostras randômicas de treino/validação/teste é executado 10 vezes. Deste modo, a taxa de acerto geral é determinada pela média da taxa de acerto apresentada como resultado por cada amostra. Em relação ao método *10-fold cross-validation*, a variabilidade dos resultados é medida ao longo de 10 execuções.

Apresentadas as técnicas de balanceamento e avaliação abordadas neste trabalho, a metodologia de avaliação é definida como segue: dados dois conjuntos de treinamento e teste, a aplicação de uma técnica de balanceamento é conduzida no conjunto de treinamento, logo o classificador é criado a partir de um conjunto de dados modificado. Entretanto, o desempenho do classificador é avaliado em um conjunto de teste que não sofre qualquer alteração em sua distribuição de dados original. Este protocolo de avaliação é mais próximo da realidade, pois os dados do mundo real são normalmente desbalanceados e o conjunto de teste preserva essa propriedade.

5.5 Avaliação dos *Baselines*

Nesta seção, descreve-se a metodologia utilizada para mensurar a eficiência dos *baselines* na tarefa de detecção de relações semânticas. Os *baselines* utilizados para fins comparativos variam de acordo com a predição semântica de referência.

Como *baselines*, adotou-se o conjunto de medidas de similaridade/parentesco/distância usadas como atributos pela abordagem CPDST, tendo em vista que estas representam as heurísticas usadas nos trabalhos relacionados para detectar similaridade entre *tags*. Logo, os atributos selecionados em cada problema foram (cf. Seção 4.2):

1. **Sinonímia**: Distância de Edição Normalizada (*DEN*), coocorrência *tag-tag* normalizada ($\text{cooc}^{\text{tag-tag}}$), cosseno fundamentado em vetores de contagem de coocorrência *tag-tag* ($\cos^{\text{tag-tag}}$), cosseno fundamentado em vetores de contagem de coocorrência *tag*-recurso ($\cos^{\text{tag-res}}$) e *overlap* normalizado para identificação de sinônimos ($\text{overlap}^{\text{syn}}$);

2. **Relação de subordinação**: suporte normalizado (*sup*), o qual é equivalente a $\text{cooc}^{\text{tag-tag}}$, generalização (gen), cosseno fundamentado em vetores de contagem de coocorrência *tag-tag* ($\cos^{\text{tag-tag}}$), *overlap* normalizado para identificação de relações hierárquicas ($\text{overlap}^{\text{hyp}}$) e busca hierárquica normalizada (*tsearch*).

Embora a maioria dos trabalhos relacionados adotem algumas destas medidas para detectar relações semânticas entre *tags* em folksonomias, observou-se na literatura consultada que este é o primeiro trabalho que efetua uma avaliação quantitativa destas medidas para o problema de detecção de sinonímia e relações de subordinação entre *tags*.

A metodologia de avaliação para os *baselines* é definida da seguinte forma: para um determinado par de *tags* $t, t' \in T$ e uma dada medida de similaridade $s : T^2 \rightarrow \mathbb{R}$, uma classe positiva é atribuída a t, t' se e somente se $s(t, t') \geq$ *thr*, em que *thr* é um limiar pré-estimado para cada *baseline*; e negativo, caso contrário. Note que, no caso da função de distância d, um rótulo de classe positiva é atribuído a t, t' se e somente se $s(t, t') \leq$ *thr*; e negativa, caso contrário.

O melhor limiar *thr* definido para cada medida de similaridade/distância foi estimado por meio de uma busca gulosa[29] ao longo do conjunto de validação, contendo este 1/3 das instâncias de treinamento. O *thr* assume várias frações no intervalo]0,1[.

5.6 Métricas para Avaliação

As medidas de avaliação tradicionais *precision* (precisão), *recall* (revocação) e *f-measure* (média harmônica da precisão e revocação) (MANNING; RAGHAVAN; SCHÜTZE, 2008) são empregadas para avaliar o desempenho dos *baselines* e da abordagem CPDST, as quais são computadas como segue:

$$\text{precision} = \frac{VP}{VP + FP}, \tag{5.1}$$

$$\text{recall} = \frac{VP}{VP + FN}, \tag{5.2}$$

[29] O algoritmo varia o limiar, coleta o resultado e atualiza a melhor escolha em cada iteração.

$$\text{f-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{5.3}$$

em que VP, FP e FN denotam *Verdadeiro Positivos*, *Falso Positivos* e *Falso Negativos*, respectivamente. De acordo com Manning, Raghavan e Schütze (2008), o valor 2 existente na fórmula do *f-measure* indica que o *precision* e *recall* são ponderados equitativamente. *Precision* determina, para uma classe em particular, a fração de verdadeiros positivos existente em todos os resultados positivos. *Recall* define a fração de verdadeiros positivos de uma determinada classe previstos corretamente. Optou-se por discutir o desempenho dos métodos comparados no Capítulo 6 majoritariamente em função do *f-measure* porque esta medida, isoladamente, pondera o *trade-off precision* vs. *recall* em um único valor.

No caso de uma medida de similaridade, um verdadeiro positivo entre duas *tags* $t, t' \in T$ para sinonímia ocorre quando $s(t, t') \geq$ *thr* e t, t' são sinônimos de acordo com os dicionários WordNet ou ConceptNet. Quando $s(t, t') <$ *thr* e t não é confirmado como sinônimo de t', tem-se um típico caso de verdadeiro negativo. O procedimento é inverso para medidas de distância: um VP acontece quando $d(t, t') <$ *thr* e t, t' são reconhecidos pelo Wordnet ou ConceptNet. No caso de detecção de relações de subordinação, um VP ocorre quando $s(t, t') \geq$ *thr* e t' é um hipônimo comprovado de t.

5.7 Algoritmos de Classificação

A abordagem CPDST é configurada para executar com os seguintes algoritmos de classificação: C4.5, Random Forest, SVM, Naive Bayes, Regressão Logística, JRip (*Java implementation of the rule learner Repeated Incremental Pruning to Produce Error Reduction* - RIPPER), kNN, Redes Neurais e o metaclassificador AdaBoost (com C4.5). A justificativa para a escolha desses algoritmos levou em conta duas questões: (i) reunir algoritmos que empreguem diferentes estratégias de aprendizado e predição das classes de um problema, a fim de descobrir qual deles é mais apropriado para detectar relações semânticas com melhor acurácia; e (ii) explorar os algoritmos de classificação mais influentes na literatura, enfatizados em livros-textos (HAN; KAMBER, 2006;

HARRINGTON, 2012; WITTEN; FRANK, 2011) e evidenciados em publicações científicas afins (FERNÁNDEZ-DELGADO et al., 2014; COHEN, 1995; KUMAR; VERMA, 2012; WU et al., 2007). Diante da diversidade de algoritmos examinados, para desviar-se de uma discussão longa decidiu-se apresentar e interpretar os resultados no Capítulo 6 apenas em relação ao classificador que produziu melhor *f-measure* nos experimentos para detecção de relações de sinonímia e subordinação.

Não faz parte do escopo deste trabalho realizar um estudo exaustivo e completo para comparar o desempenho dos classificadores utilizados pela abordagem CPDST, ou até mesmo encontrar os parâmetros ideais de ajuste para cada classificador base. O intuito é de apenas identificar, dentre uma lista de classificadores disponíveis, qual deles consegue apresentar melhor desempenho natural na classificação da classe positiva para o cenário do problema abordado, a fim de certificar que o problema de detecção de relações semânticas utilizando a técnica de aprendizado supervisionado compensa.

5.8 Considerações Finais

Neste capítulo, relatou-se a metodologia adotada para conduzir os experimentos e avaliar a abordagem de detecção de relações semânticas proposta. Descreveu-se a base de dados das aplicações Web 2.0 BibSonomy e Delicious, assim como a sistemática adotada para pré-processamento de dados e criação/seleção de instâncias de treinamento. Também foram apresentadas as técnicas empregadas para lidar com o problema de desbalanceamento e *overlap* de classes, as medidas de avaliação e os parâmetros de configuração para extração de atributos.

Particularmente para a base de dados do sistema Delicious, as tarefas de seleção e pré-processamento de dados foram as que concentraram o maior volume de esforço. O grande número de instâncias tornou mais demorada a rotulação usando o ConceptNet e inviabilizou o tratamento em memória de várias rotinas de processamento, o que teve de ser contornado com acessos sucessivos ao Sistema de Gerenciamento de Banco de Dados. Uma vez que a quantidade de recursos marcados por uma *tag* é consideravelmente elevada, a extração dos atributos *overlap* e $\cos^{\text{tag-res}}$ mostrou-se computacionalmente dispendiosa de processar em computadores convencionais. Embora não esteja no foco deste trabalho tratar questões de

desempenho relacionados à tarefa de extração de atributos, soluções para reduzir esse esforço podem ser futuramente analisadas. No próximo capítulo, são apresentados os resultados experimentais alcançados pelos *baselines* e pela abordagem CPDST.

Capítulo 6

Resultados e Discussão

Neste capítulo, são apresentados os resultados experimentais em termos de acurácia inerentes à comparação entre a abordagem CPDST e os *baselines*, de acordo com a metodologia relatada no Capítulo 5. Outra parte do conteúdo deste capítulo pode ser examinada nas publicações provenientes desta tese em Rêgo, Marinho e Pires (2012) e Rêgo, Marinho e Pires (2015). Recorde os dois primeiros problemas de pesquisa definidos no Capítulo 1:

- *QP1: Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de sinonímia provê melhor acurácia na identificação de sinonímia entre tags de folksonomia?*
- *QP2: Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de relações hierárquicas provê melhor acurácia na identificação de relações hierárquicas (hiperonímia e hiponímia) entre tags de folksonomias?*

Inicialmente, os resultados em termos de *f-measure* para detecção de sinonímia e relação de subordinação são avaliados (QP1 e QP2), confrontando a efetividade alcançada pela abordagem CPDST em relação aos *baselines*. A QP3 é abordada separadamente no Capítulo 7. Todos os resultados apresentados aqui foram obtidos no conjunto de teste utilizando os melhores parâmetros estimados no conjunto de validação, conforme relatado na Seção 5.4.

O software de Mineração de Dados Weka[30] foi utilizado para processar os algoritmos de classificação, considerando as configurações originalmente oferecidas (em particular, foi

[30]<http://www.cs.waikato.ac.nz/ml/weka>

usado o algoritmo SVM de classificação não-linear com kernel *Radial Basis Function* - RBF e parâmetro de custo C=1.0). Excepcionalmente, para o metaclassificador AdaBoost, substitui-se o classificador *DecisionStump* por C4.5.

6.1 Detecção de Sinonímia

Nesta seção, são apresentados e discutidos os resultados experimentais inerentes à captura de relações de sinonímia. A princípio, na Seção 6.1.1, é realizada uma análise qualitativa das indicações de sinonímia sugeridas pelos *baselines* referentes à base de dados do BibSonomy considerando as top-5 *tags* relacionadas para cada *tag* de referência. Por fim, a Seção 6.1.2 apresenta o comparativo de desempenho entre os *baselines* e a abordagem CPDST.

6.1.1 Análise Qualitativa: Visão dos *Baselines*

Preliminarmente, efetuou-se uma investigação qualitativa com respeito às indicações de sinonímia providas pelos *baselines* utilizando a base de dados do BibSonomy, tendo em vista que esta base prevaleceu unicamente durante grande parte do desenvolvimento desta tese. Para cada uma das medidas introduzidas na Seção 4.2.1, determinou-se para cada uma das 3.177 *tags* remanescentes do processo de Preparação dos Dados (cf. Seção 5.1) suas top-5 *tags* mais relacionadas. A Tabela 6.1 ilustra exemplos de *tags* retornadas pelos *baselines* para cinco *tags* selecionadas aleatoriamente, a fim de prover percepções qualitativas acerca de suas indicações. As células destacadas na cor cinza denotam os sinônimos reconhecidos pelo WordNet/ConceptNet.

Naturalmente, a redução do conjunto $|T|$ levando-se em conta a condição de existência da *tag* nos dicionários eletrônicos limita a detecção de sinônimos. Além disso, a decisão sobre quem é ou não sinônimo é tomada sob o ponto de vista dos dicionários eletrônicos, o que se torna mais restritivo pelo fato de se tratar de uma semântica mais específica. Entretanto, durante análise dos dados brutos na etapa de preparação dos dados, era visível o grande número de *tags* sinônimas que poderiam ser capturadas facilmente utilizando distância de edição. O Apêndice A exemplifica a ocorrência de diversos sinônimos da palavra `learning` no universo das 10.000 *tags* mais frequentes no BibSonomy. Por ser conceito popular, `learning` apresenta diferentes formações da mesma palavra, que vão desde a falta

Tabela 6.1: Top-5 *tags* mais relacionadas por cada *baseline*.

Tag	Rank	DE	$\cos^{\text{tag-tag}}$	$\cos^{\text{tag-res}}$	$\text{cooc}^{tag-tag}$	overlap^{syn}
study	1	studio	reading	pt	books	-
	2	stuff	publishing	geometry	math	-
	3	story	science	radiography	research	-
	4	studies	research	provenance	science	-
	5	student	reference	unread	reference	-
color	1	colour	colors	colour	design	colour
	2	colors	graphic	colors	tools	colors
	3	cohort	typography	scheme	photos	-
	4	cocoa	graphics	abstracts	art	-
	5	solar	inspiration	design	graphic	-
video	1	videos	free	videos	software	streaming
	2	Wide	music	lecture	tutorial	movie
	3	view	download	streaming	audio	webcam
	4	idea	cool	lectures	tools	capture
	5	kinder	tool	audio	music	documentary
web	1	weak	annotation	environment	semantic	deep
	2	debt	project	semantic	tools	crawler
	3	webcam	ontology	ajax	software	service
	4	Welt	ontologies	tools	search	http
	5	weka	desktop	pt	social	page
cite	1	site	citations	citations	citation	-
	2	city	citation	citation	citations	-
	3	cute	bibliographic	bibliographic	bibliographic	-
	4	cities	academics	archiving	reference	-
	5	crime	archiving	centrality	research	-

de padronização de letras em maiúsculo/minúsculo até o acréscimo de símbolos especiais.

Conforme pode-se observar na Tabela 6.1, a maioria dos sinônimos atestados pelo WordNet/ConceptNet correspondem a sutis variações léxicas de sua respectiva *tag* de referência. Como esperado, a DE é capaz de capturar muitos sinônimos relacionados a pequenas variações léxicas como, por exemplo, `study` e `studies`. Entretanto, para *tags* de tamanho curto (e.g., `web` e `cite`), a medida perde sua eficiência porque captura muitos termos de grafia semelhante mas sem parentesco semântico. Uma análise sobre a variação de limiar para caracterização de sinônimos usando distância de edição está disponível no

Apêndice B.

As medidas $\cos^{\text{tag-tag}}$, $\cos^{\text{tag-res}}$ e $\text{cooc}^{\text{tag-tag}}$, fundamentadas em distribuições estatísticas, também foram capazes de revelar sinônimos, seguindo a teoria de que o princípio de coocorrência provê indicativos para esta finalidade. O *baseline* overlap^{syn} não retorna candidatos a sinônimo para algumas *tags*. Como overlap^{syn} produz valores numéricos que denotam diferentes interpretações sobre a semântica compartilhada por duas *tags*, o *baseline* assume comportamento restritivo quando apenas um valor em particular caracteriza a informação de interesse. Então, após a aplicação do limiar de corte (cf. Tabela 6.2), é possível que não haja a indicação de *tags* para uma *tag* de entrada aleatória.

Assim como ressaltado por Cattuto et al. (2008) e Clements, Vries e Reinders (2008), mesmo que as medidas probabilísticas sejam capazes de capturar alguns casos diferentes de sinonímia (e.g., `cite` e `reference`), na maioria das vezes elas capturam outros tipos de relações semânticas, mas não necessariamente relações de sinonímia. Um exemplo que expressa esse ponto de vista é a *tag* `web` e suas relacionadas: `semantic`, `search`, `annotations` e `ajax`. Da mesma forma, embora `reading`, `publishing`, `science`, `research` e `reference` sejam certamente conceitos relacionadas à *tag* `study`, nenhuma delas é sinônima. Esta observação é confirmada pela análise quantitativa apresentada na seção subsequente. Neste trabalho, os esforços estão voltados principalmente para os casos mais complexos, que não podem ser resolvidos por uma simples uniformização de *tags* (e.g., remoção de caracteres especiais ou padronização dos caracteres para letras minúsculas).

Algumas *tags* denotam um sentido mais amplo e adquirem maior popularidade porque podem ser facilmente combinadas com outras *tags*. Neste caso, o que se observa é uma diversidade de *tags* distintas que despontam como semanticamente relacionadas em relação à *tag* mais genérica. Isto se observa, por exemplo, com as *tags* `video` e `web` diante dos resultados apresentados: `video`={`free`, `lecture`, `tutorial`, `software`} e `web`={`ontology`, `tools`, `social`, `http`}. Portanto, para uma *tag* de sentido geral, a lista das *top-n tags* semanticamente relacionadas tende a apresentar menos casos de sinonímia.

6.1.2 Análise Quantitativa

O tópico de pesquisa representado pela QP1 é abordado nesta seção. Para responder a QP1, foram comparadas diferentes heurísticas comumente empregadas na literatura para detecção de sinonímia com a abordagem CPDST.

Os primeiros resultados obtidos com os experimentos de detecção de sinonímia originaram uma publicação científica (RÊGO; MARINHO; PIRES, 2012). Utilizando uma base de dados do BibSonomy, foi executada uma análise comparativa de desempenho em termos de *f-measure* entre os *baselines* e a abordagem CPDST. O método *random undersampling* foi aplicado para gerar amostras de dados constituídas por diferentes proporções de instâncias positivas/negativas, ou seja, 10+/90-, 20+/80-, 30+/70-, 40+/60- e 50+/50-. A técnica de avaliação utilizada em cada partição foi a *holdout set* em 10 amostras distintas. A abordagem CPDST foi avaliada apenas com o classificador C4.5. Na ocasião, distância de edição se destacou como o *baseline* de melhor desempenho individual, além de ter sido o atributo mais relevante para a tarefa de classificação. Os resultados mostraram que a abordagem CPDST supera a melhor heurística em todas as partições com superioridade de até 8.1%. Particularmente nas amostras de dados mais balanceadas, distância de edição apresentou um *f-measure* mais próximo em relação ao que foi obtido pela abordagem CPDST.

Seguindo o protocolo de avaliação definido na Seção 5.4, os experimentos para detecção de sinonímia foram refeitos acrescentando-se o atributo overlap^{syn} e atualizando-se os valores de todos os atributos fundamentados na contagem de coocorrência, tendo em vista que a base de dados foi aperfeiçoada para evitar que um recurso em particular fosse referenciado com identificador numérico distinto em alguns *posts*[31]. As configurações experimentais são discriminadas a seguir. Relembre na Seção 5.6 que, para a abordagem CPDST, as configurações foram analisadas com todos os classificadores listados, entretanto apenas o de melhor desempenho em termos de *f-measure* é retratado na configuração final.

- CE1: Técnica *10-fold cross-validation*, sem aplicação de método de balanceamento de classes. Abordagem CPDST configurada com o metaclassificador AdaBoost(C4.5);
- CE2: Técnica *holdout set* e aplicação do método Tomek Link. Abordagem CPDST

[31] Este comportamento foi observado na base de dados do BibSonomy.

configurada com o metaclassificador AdaBoost(C4.5).

A Tabela 6.2 exibe o melhor valor de limiar estimado para os *baselines* ao longo do conjunto de validação, conforme procedimento descrito na Seção 5.5. Particularmente, o baixo valor revelado para $\text{cooc}^{tag-tag}$ é explicado pelo fato de que a normalização é computada por meio da máxima contagem de coocorrência constatada, logo a distribuição dos seus valores no intervalo [0;1] pode variar bruscamente dependendo da base de dados utilizada. Por exemplo, se existir um *outlier* com alto valor de contagem de coocorrência, a maioria dos valores de similaridade podem ficar comprimidos no intervalo [0; 0,15]. Para exemplificar, a maior contagem de coocorrência encontrada na base do BibSonomy foi de 3.168 relativa à associação do par de *tags* `programming` e `genetic`, enquanto que a contagem de coocorrência para a maioria dos pares de *tags* está contida na faixa de $[1, 350]$. Portanto, a escolha do limiar é altamente dependente da contagem de coocorrência máxima, a qual pode variar significativamente entre diferentes folksonomias.

Tabela 6.2: Estimativa de limiar dos *baselines* por técnica de avaliação.

Téc. Avaliação	*Baselines*				
	$\text{cooc}^{tag-tag}$	DEN	$\text{coS}^{tag-tag}$	overlap^{syn}	$\text{cos}^{tag-res}$
BibSonomy					
10-fold cross-validation	0,03	0,3	0,60	1,00	0,10
holdout set	0,03	0,3	0,60	0,67	0,10
Delicious					
10-fold cross-validation	0,01	0,5	0,90	0,33	0,10
holdout set	0,01	0,5	0,90	0,33	0,10

A Tabela 6.3 exibe os resultados de *f-measure* para a classe positiva na forma *média* $\pm$ *desvio padrão* ao longo de 10 repetições/amostras para todos os métodos comparados, agrupados por técnica de avaliação (*holdout set* e *10-fold cross-validation*). O *f-measure* apresentado pelos *baselines* é resultante da aplicação dos limiares definidos na Tabela 6.2. O método com melhor desempenho em cada técnica de avaliação está destacado em negrito.

Observa-se que os resultados foram expressivos neste novo cenário, no qual a abordagem CPDST obteve um desempenho notável na classificação correta das instâncias positivas tanto na base de dados do BibSonomy quanto na do Delicious (*f-measure* = 0,940 e *f-measure* = 0,925, respectivamente). O sucesso da abordagem CPDST é creditado aos atributos

Tabela 6.3: Medições de *f-measure* das instâncias positivas por técnica de avaliação.

Téc. Avaliação	*Baselines*					CPDST
	$\text{cooc}^{tag-tag}$	DEN	$\cos^{tag-tag}$	overlap^{syn}	$\cos^{tag-res}$	
BibSonomy						
10-fold CV	0,052±0,001	0,505±0,001	0,035±0,001	0,033±0,001	0,044±0,000	**0,940** ±0,04
holdout set	0,052±0,009	0,510±0,033	0,036±0,003	0,030±0,006	0,044±0,003	**0,774**± 0,03
Delicious						
10-fold CV	0,256±0,001	0,555 ±0,001	0,375±0,002	0,271±0,001	0,567 ± 0,001	**0,925** ± 0,01
holdout set	0,254±0,015	0,553±0,010	0,376±0,006	0,271±0,002	0,569 ± 0,011	**0,740** ± 0,02

utilizados, o estado da arte dos classificadores e à determinação do ponto ideal de corte (filtro) para seleção de instâncias de treinamento e teste. Note que o baixo desvio padrão (0,01 a 0,04) sugere que a abordagem CPDST é consistente levando-se em conta as diferentes distribuições de treinamento e teste avaliadas. Logo, o *f-measure* resultante é normalmente homogêneo e tende a estar próximo da média.

Analisando o resultado individual dos *baselines* na base de dados do BibSonomy, a melhor heurística é a distância de edição. Entretanto, todos os *baselines* seguiram a tendência de ter o desempenho deteriorado com o agravamento do nível de desbalanceamento de classes, uma vez que todas as instâncias foram utilizadas para treinamento e teste. Com isso, a margem de desempenho que separa a distância de edição da abordagem CDPST tornou-se maior.

Na base de dados do Delicious, os *baselines* apresentaram melhor desempenho para detecção de sinônimos. A melhor heurística é $\cos^{tag-res}$, seguida de DEN, $\cos^{tag-tag}$, overlap^{syn} e $\text{cooc}^{tag-tag}$. Visto que o valor médio de $\cos^{tag-res}$ está abaixo de 0,01 para as instâncias negativas e em torno de 0,18 para as instâncias positivas, o limiar estimado (0,1) consegue discriminar com maior sucesso as instâncias positivas das negativas. Por ser uma base de dados de propósito geral, em que os usuários aplicam *tags* para registrar *bookmarks*, o perfil de distribuição de *tags* se torna mais consistente tendo em vista a superioridade em termos de quantidade de anotações e usuários participantes.

A Tabela 6.4 exibe a lista de *tags* sinônimas previstas corretamente pela abordagem CPDST usando a técnica *10-fold cross-validation* na base de dados do BibSonomy, considerando as *tags* alvo constantes na Tabela 6.1 e outras suplementares. As indicações são asseguradas pelos dicionários eletrônicos WordNet e ConceptNet.

Tabela 6.4: Detecção de sinonímia - Aprendizado Supervisionado.

Tag	Sinônimos detectados
study	**learning, work**
color	**colour, colors**
video	**television, videos, TV**
web	**net, networks, network**
cite	**citations**
Outros Exemplos	
citation	**reference, references, citations, cite**
hack	**crack, hacker, hacks**
image	**icon, icons, imaging, pictures**
europe	**european-community, european-economic-community, european-union**
movie	**cinema, film, movies**

Nota-se na Tabela 6.4, que a abordagem de aprendizado supervisionado é capaz de capturar os seguintes casos típicos de sinonímia em folksonomias:

1. **Número gramatical**: *tags* podem ser encontradas na forma singular ou plural, tais como `hack` e `hacks`, por exemplo;

2. **Inflexões do substantivo/verbo**: as variações devido a inflexões do substantivo e verbo produzem *tags* similares tais como `image` e `imaging`;

3. **Vocabulário do usuário**: atribuições de *tags* às vezes dependem da preferência de vocabulário do usuário. Por exemplo, alguns usuários podem marcar um recurso que trata sobre o bloco econômico europeu com a *tag* `Europe`, enquanto outros podem preferir utilizar a *tag* `European_Union`;

4. **Multilinguismo de *tags***: algumas palavras podem ser escritas de forma diferente de acordo com o idioma utilizado. Algumas vezes as variações ocorrem dentro da mesma linguagem como, por exemplo, `color` (inglês americano) e `colour` (inglês britânico);

5. **Casos não triviais de sinonímia**: palavras distintas podem ter o mesmo significado como, por exemplo as *tags* `study` e `learning`, ou `movie` e `cinema`;

6. **Acrônimos**: a existência de acrônimos é normalmente vista como um problema de ambiguidade em folksonomias, porém, também traz evidências de sinonímia. Por exemplo, ambas as *tags* `TV` e `television` normalmente expressam o mesmo significado.

6.1.3 Teste de Significância

Para determinar se a diferença na média de acurácia da abordagem CPDST comparada com o melhor dos *baselines* após múltiplas execuções é estatisticamente significante, executou-se um teste estatístico de comparação de alternativas para confirmar ou refutar o seguinte teste de hipótese:

$\mathbf{H}_0$: Os métodos de detecção de sinonímia possuem a mesma acurácia em termos de média ($\mu_1 = \mu_2$).

$\mathbf{H}_1$: Há diferença em *f-measure* entre os métodos comparados ($\mu_1 \neq \mu_2$), o que implica em dizer que a abordagem CPDST apresenta melhor acurácia.

Ao confirmar a suposição de normalidade nas medições de *f-measure* usando os dados do BibSonomy (*holdout* e *10-fold cross-validation*), executou-se o teste paramétrico teste-t de *student* pareado de duas caldas (*paired t-test 2-sided*) com 95% de nível de confiança. Os resultados *p-valor= 4.856e-16 < 0.05* (*10-fold cross-validation*) e *p-valor= 1.596e-07 < 0.05* (*holdout*) rejeitam a hipótese nula, indicando que CPDST é estatisticamente significante superior ao melhor dos *baselines* DEN com 95% de confiança nas duas configurações experimentais CE1 e CE2. Com a base de dados do Delicious, o teste não-paramétrico *Wilcox Signed Rank* indicou um *p-valor = 0.0019 < 0.05* (*holdout*) e *p-valor= 0.00585 < 0.05* (*10-fold cross-validation*), confirmando que CPDST é estatisticamente significante superior ao melhor dos *baselines* $\cos^{tag-res}$. Deste modo, o resultado do teste de significância responde a QP1, confirmando que o uso da técnica de classificação para a detecção de relações de sinonímia provê melhor acurácia do que as heurísticas empregadas para o mesmo propósito.

De um modo geral, os resultados sugerem que os *baselines* isoladamente são insatisfatórios para a detecção específica de sinonímia em contextos nos quais a distribuição das classes é fortemente desbalanceada. Com sua configuração experimental aprimorada, a abordagem CPDST evidenciou-se como uma estratégia eficaz para essa tarefa, alcançando índices de acurácia acima de 90,0% nas duas bases de dados utilizadas. Vale ressaltar que as

normativas definidas no protocolo de avaliação reproduzem um ambiente mais próximo do mundo real, visto que a distribuição de teste não é submetida à ação de qualquer técnica de balanceamento de classes, mesmo com o alto grau de desbalanceamento de classes observado.

6.2 Detecção de Relações de Subordinação

Nesta seção, são apresentados os resultados da abordagem CPDST para a tarefa de detecção de relações de subordinação. Para responder a QP2, um extenso conjunto de experimentos foi conduzido com o intuito de determinar a combinação de técnicas que melhor contribui para o aprendizado de relações de subordinação. Assim, pode-se comparar o desempenho da abordagem CPDST com diferentes heurísticas empregadas para esta finalidade.

A identificação de relações de subordinação verídicas demonstrou ser um problema de difícil reconhecimento. Isso porque a natureza da semântica em questão pressupõe primeiramente a existência de parentesco entre *tags* e a presunção de nível de inclusão e generalização mútuos, os quais são definidos empiricamente. Aliadas à natural complexidade do problema, a construção de um modelo preditor razoável a partir de um conjunto de dados com acentuado desbalanceamento de classes e, principalmente, com a intensificação do problema de *overlapping* de classes, torna-se um obstáculo ainda maior para o aprendizado do classificador.

Diante desta conjuntura, a abordagem CPDST foi ajustada para a tarefa de detecção de relações de subordinação. A realização de um comparativo de *f-measure* entre a abordagem CPDST (configurada com o classificador de melhor desempenho *Random Forest*) e um conjunto de *baselines* empregados para estimação de relações de subordinação resultaram em uma publicação científica (RÊGO; MARINHO; PIRES, 2015). Os experimentos foram conduzidos em uma base de dados do BibSonomy sob o seguinte protocolo de avaliação: (i) geração de partições de dados com diferentes proporções de instâncias positivas e negativas (10+/90-, 20+/80-, 30+/70-, 40+/60- e 50+/50-), da mesma forma como foi feito em Rêgo, Marinho e Pires (2012), (ii) os métodos SMOTE, CNN, *random undersampling* e Tomek Link foram empregados com o intuito de originar as partições de treinamento e teste, para que fosse possível avaliar o efeito de suas propostas para a resolução do problema

tratado, (iii) utilização da técnica de avaliação *10-fold cross-validation*, e (iv) execução da abordagem CPDST com os classificadores C4.5, Random Forest, SVM, Naive Bayes, Regressão Logística e AdaBoost.

Tendo em vista que o desempenho da abordagem CPDST é discutido nesta seção sob uma nova metodologia de avaliação, os resultados apresentados em nossa publicação (RÊGO; MARINHO; PIRES, 2015) não estão reproduzidos neste documento. Entretanto, um breve resumo é apresentado para prestar conhecimento sobre as descobertas. Na ocasião, Tomek Link foi estimado como a melhor técnica para geração das partições. Overlap^{hyp} foi o *baseline* de melhor desempenho, embora tenha sido observado que todos os *baselines* não são fortemente afetados pela modificação da distribuição de classes provida pela técnica Tomek Link. A análise quantitativa mostrou que a abordagem CPDST supera todos os *baselines* em todas as partições, independentemente do classificador utilizado. A técnica Tomek Link teve relevante contribuição para enaltecer o desempenho da abordagem CPDST, pois possibilitou a seleção de conjuntos de treinamento e teste otimizados para que o classificador pudesse identificar com clareza os padrões que diferenciam as instâncias positivas das negativas. A média de acurácia da abordagem CPDST ao longo das partições foi de 98%.

6.2.1 Análise Quantitativa

Deste ponto em diante, a discussão está focada na apresentação dos resultados de acordo com o protocolo de avaliação relatado na Seção 5.4. Após a análise de várias combinações de técnicas de balanceamento de classes e algoritmos de classificação, em particular são analisados os resultados em torno da melhor configuração experimental identificada em cada técnica de avaliação. A metaclassificação ASC combinada com o classificador base Naive Bayes obteve a melhor acurácia na classificação. As configurações experimentais (CE) abordadas são:

- CE1 - Técnica *10-fold cross-validation*, sem aplicação da técnica de balanceamento de classes no conjunto de treinamento. Abordagem CPDST configurada com metaclassificador ASC e algoritmo de classificação Naive Bayes. Reforçando o que foi exposto na Seção 2.6.5, o custo para cada classe no ASC foi estimado empiricamente,

variando os valores do custo a fim de identificar o ajuste que proporciona o melhor desempenho na classificação. Matriz de custo estimada:

BibSonomy		Delicious	
0.0	20.0	0.0	1.0
9.0	0.0	6.0	0.0

- CE2 - Técnica *holdout set* e aplicação do método Tomek Link. Abordagem CPDST configurada com o metaclassificador ASC e algoritmo de classificação Naive Bayes para a base de dados do BibSonomy ou simplesmente com o classificador Naive Bayes para a base de dados do Delicious. Matriz de Custo estimada para a metaclassificação ASC:

0.0	1.0
3.0	0.0

Observe que Naive Bayes foi o classificador de melhor desempenho nas configurações experimentais. Assim como nos experimentos para detecção de sinonímia (cf. Seção 6.1.2), as medidas baseadas no princípio de coocorrência tiveram seus valores atualizados para se adequarem aos ajustes efetuados na nova base de dados experimental.

A Tabela 6.5 exibe os resultados de *f-measure* para a classe positiva na forma *média* $\pm$ *desvio padrão* ao longo de 10 amostras para todos os métodos comparados, agrupados por técnica de avaliação (*holdout set* e *10-fold cross-validation*). O método com melhor desempenho em cada técnica de avaliação está destacado em negrito.

Os limiares utilizados para obtenção dos resultados dos *baselines* foram:

- BibSonomy: 0,1 (suporte), 0,1 (gen e $\cos^{tag-tag}$), 0,75 (overlap^{hyp}) e 0,33 ($tsearch$);
- Delicious: 0,01 (suporte), 0,1 (gen), 0,9 ($\cos^{tag-tag}$), 0,25 (overlap^{hyp}) e 0,00 ($tsearch$).

$Tsearch$ é o *baseline* de melhor desempenho usando a base de dados do BibSonomy e assume maior valor de *f-measure* sob o ponto de vista da técnica de avaliação *10-fold cross-validation*. Na base de dados do Delicious, o atributo overlap^{hyp} destacou-se como

Tabela 6.5: Medições de *f-measure* das instâncias positivas por técnica de avaliação.

Téc. Avaliação	*Baselines*					CPDST
	$suporte$	gen	$\cos^{tag-tag}$	overlap^{hyp}	$tsearch$	
BibSonomy						
CE1: 10-fold CV	0,052±0,003	0,031±0,000	0,030±0,000	0,066±0,002	0,145±0,002	**0,160±0,006**
CE2: holdout set	0,041±0,005	0,031±0,000	0,030±0,000	0,063±0,006	0,107±0,018	**0,159± 0,033**
Delicious						
CE1: 10-fold CV	0,130±0,001	0,104±0,001	0,114±0,003	0,160±0,002	0,085±0,000	**0,185±0,003**
CE2: holdout set	0,132±0,010	0,107±0,014	0,118±0,008	0,159±0,008	0,085±0,000	**0,207±0,019**

melhor *baseline*, enquanto que $tsearch$ apresentou o pior desempenho. Esse revés é uma consequência da observação de 95% das instâncias positivas possuírem valor $tsearch = 0$, ocasionando *recall* máximo devido ao limiar estimado para $tsearch$ no Delicious.

Para a maioria dos *baselines*, o baixo desempenho individual em termos de *f-measure* é decorrente do fato de que o melhor limiar estimado se encontra em uma faixa de valores estreita. A fim de esclarecer essa argumentação, observe a Figura 6.1 que mostra a distribuição de valores para o atributo *generalização* na base de dados do BibSonomy em formato de *box/whisker plots*. O eixo das ordenadas denota o intervalo de valores alcançado. A parte inferior da caixa denota o 1º quartil (25º percentil) e o topo da caixa o 3º quartil (75º percentil). A linha no meio da caixa denota a mediana (50º percentil), enquanto que o topo da barra e a parte inferior da barra denotam, respectivamente, os valores máximo e mínimo observados. Há também alguns pontos denominados *outliers* representados em formato de círculo. Na Figura, os exemplos positivos estão concentrados em torno do valor *0,2*, aumentando desta forma o *recall* visto que a maioria das instâncias estão acima deste limiar, mas diminuindo o *precision* em virtude de não só os exemplos positivos estarem acima deste limiar.

A abordagem CPDST supera o melhor dos *baselines* tanto na base de dados do BibSonomy quanto na do Delicious, em qualquer uma das configurações experimentais avaliadas. Entretanto, a acurácia na detecção das classes positivas não foi largamente superior se comparada aos resultados para detecção de sinonímia diante do mesmo protocolo de avaliação. O agravamento dos problemas de desbalanceamento e *overlapping* de classes exigiram desde o início um esforço maior para se conseguir avanços na tarefa de predição. Neste cenário, foram testadas as técnicas CNN e SMOTE, mas com resultados

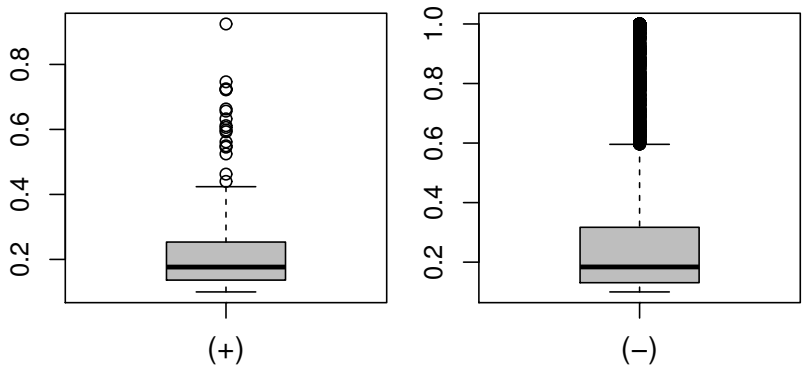

Figura 6.1: Distribuição de dados: atributo *generalização*.

insatisfatórios. *Random undersampling* é capaz de atenuar o problema de desbalanceamento de classes mas não de tratar o problema de *overlapping* de classes. Por sua vez, Tomek Link demonstra que é capaz de aliviar o problema de *overlapping* de classes no conjunto de treinamento ao detectar e remover exemplos de *overlapping* pertencentes à classe majoritária. Entretanto, o modelo perde efetividade na predição da classe positiva quando submetido à avaliação no conjunto de teste.

A aplicação da técnica Tomek Link, apontada como melhor solução para tratamento do desbalanceamento e *overlapping* de classes nas condições experimentais estabelecidas em Rêgo, Marinho e Pires (2015), não demonstrou a mesma potencialidade quando a avaliação da abordagem CPDST é realizada em um conjunto de teste sem modificação de sua distribuição de dados. Este comportamento é compreensível pelo seguinte raciocínio: ao aplicar uma técnica de tratamento de balanceamento/*overlapping* de classes no conjunto de treinamento, certamente será construído um modelo preditor que tende a apresentar boa acurácia na detecção da classe de interesse. No entanto, quando o modelo é submetido à tarefa de predição a partir de um conjunto de teste cujos dados são naturalmente desbalanceados, a qualidade da predição é afetada tendo em vista que os exemplos negativos ruidosos serão classificados incorretamente pelo modelo. O melhor desempenho obtido na CE2 pela abordagem CPDST usando a base de dados do Delicious pode ser atribuído à maior quantidade de instâncias positivas existente nas amostras de treinamento e teste, o que resulta em um menor nível de desbalanceamento de classes ajustado (cf. Seção 5.4).

A existência de um extenso *overlap* na Figura 6.1 é perceptível entre as instâncias positivas e negativas em torno do valor *0,2* para o atributo *generalização*. Nesta circunstância, a habilidade de qualquer classificador seria drasticamente prejudicada para discriminar com precisão as instâncias positivas das negativas. O ideal é que os *boxplots* estejam situados em regiões distintas dentro da faixa de valores assumida por um atributo, assim como pode ser observado na Figura 6.2 extraída de nossa publicação (RÊGO; MARINHO; PIRES, 2015) em relação ao padrão das instâncias positivas e negativas selecionadas pelos métodos Tomek Link (TL) e *Random Undersampling* (RU). Tomek Link seleciona os exemplos negativos que não se encontram na região de *overlapping*, ajudando então o classificador a determinar uma melhor borda de decisão.

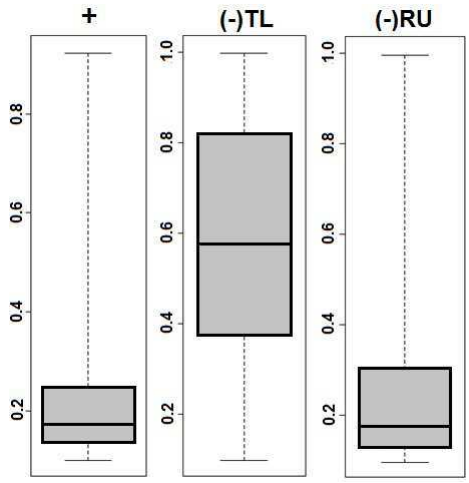

Figura 6.2: Atributo generalização: instâncias positivas vs. instâncias negativas selecionadas por Tomek Link e *Random Undersampling*.
Fonte: Rêgo, Marinho e Pires (2015).

Como ressaltado anteriormente, o aprendizado de relações de subordinação implicou em maiores dificuldades para a abordagem CPDST melhorar o baixo desempenho de *f-measure* apresentado nos experimentos, mesmo que ainda seja superior aos *baselines*. No entanto, após uma análise criteriosa, várias questões puderam ser levantadas para explicar o comportamento dos resultados:

- Percebe-se uma maior necessidade de quantitativo de instâncias positivas para aprender relações de subordinação se comparado com sinonímia. Isso porque, após a aplicação dos filtros de suporte e confiança, muitas instâncias positivas são descartadas

por não atenderem a condição do filtro. Dessa forma, o número de instâncias positivas remanescente perde representatividade para fins de treinamento e teste;

- Mesmo utilizando duas fontes linguísticas no suporte à rotulação (WordNet e ConceptNet), ainda é perceptível a limitação dos dicionários quanto à capacidade de reconhecer relações entre conceitos populares em um domínio específico ou até mesmo de conhecimento geral. Por exemplo, na base de dados do Delicious, as relações *scripting* $\succ$ *js*, *standards* $\succ$ *w3c* e *toys* $\succ$ *lego* podem ser intuitivamente interpretadas como relações de subordinação. Isso denota uma condição de restrição maior do rotulador do que da própria abordagem CPDST;

- Os principais atributos, overlap^{hyp} e $tsearch$, denotam interpretações particulares sobre prováveis pares de *tags* candidatas a uma relação de subordinação. Por exemplo, em overlap^{hyp} espera-se que as instâncias positivas tenham valor igual a 4 para o referido atributo, valor esse que sugere a indicação de uma relação de subordinação (cf. Seção 4.2.2). Para o atributo $tsearch$, o ideal é que as instâncias positivas apresentem os valores 1 ou 2. A Tabela 6.6 mostra a distribuição de frequência dos valores de overlap^{hyp} e $tsearch$ na base de dados do Delicious em uma das amostras utilizadas nos experimentos.

Tabela 6.6: Distribuição de frequência dos valores dos atributos overlap^{hyp} e $tsearch$.

Valor	overlap^{hyp}		$tsearch$	
	(+)	(-)	(+)	(-)
0	1.155	32.841	1.697	37.891
1	475	4.773	53	81
2	9	25	12	24
3	1	4	1	4
4	124	357	1	0
Total	1.764	38.000	1.764	38.000

Observa-se na Tabela 6.6 que os valores desejados para os atributos overlap^{hyp} e $tsearch$ predominam nas instâncias negativas. Ainda, para overlap$^{hyp} = 0$, a heurística argumenta que as instâncias deveriam ser descartadas pela intuição de que se $t \succ t'$, $|R_t|$ não pode ser inferior a $|R_{t'}|$, ou seja, t como hiperônimo deve anotar maior

quantidade de recursos do que seu hipônimo t' (princípio da generalidade). Entretanto, isso reduziria radicalmente a quantidade de instâncias para treinamento.

As heurísticas apoiam-se em probabilidades estatísticas para definir prováveis relações de subordinação e a constatação é de que suas indicações são tênues se comparado com o rigor de um dicionário linguístico. Logo, como se percebe, seus valores de saída tendem a divergir com a interpretação mais rigorosa do rotulador semântico, embora expressem um significado relacionado. Neste caso, entende-se que a rotulação de instâncias também deve levar em conta o que as heurísticas indicam, a fim de converter eventuais instâncias negativas em positivas. Deste modo, consegue-se introduzir melhorias na representatividade das instâncias para que possa ser refletida no treinamento do modelo de predição.

- Diante das taxas de acurácia exibidas na Tabela 6.5 e embasado nos questionamentos supracitados, resolveu-se analisar subjetivamente a qualidade das predições de saída da abordagem CPDST expondo um ponto de vista pessoal. O exame denota uma impressão sobre o significado das relações, originada a partir da análise de uma amostra de treinamento/teste com aproximadamente 40.000 instâncias, seguindo o senso comum de um modo geral. A Tabela 6.7 exibe fragmentos de relações de subordinação extraídas da base de dados do Delicious agrupados por categoria.

  - **(CAT1) Classificado corretamente**: duas *tags* constantes na relação de subordinação predita são asseguradas pelos dicionários linguísticos (VP). Também são consideradas algumas relações caraterizadas como falso positivos, porém com o entendimento de que poderia ser uma relação de subordinação correta por um julgamento humano;
  - **(CAT2) Classificação relevante**: duas *tags* preditas como uma relação de subordinação são próximas em termos de significado, mas não possuem uma relação de subordinação;
  - **(CAT3) Classificação invertida**: para um par de *tags* predito como uma relação de subordinação, observa-se uma inversão na ordem das *tags*, ou seja, da mais específica para a mais geral;

- **(CAT4) Classificação sem significado**: uma *tag* sem sentido aparece na relação de subordinação predita;
- **(CAT5) Classificado incorretamente**: duas *tags* constantes na relação de subordinação não expressam de fato uma relação de subordinação. Nesta categoria, são enquadradas *tags* funcionais (e.g., todo), de amplo significado (e.g., design) e que denotam um sentimento específico do usuário (e.g., cool).

Tabela 6.7: Predições efetuadas pela abordagem CPDST para detecção de relações de subordinação.

CAT1	(VP) *site* $\succ$ *portal*, *education* $\succ$ *learning*, *computer* $\succ$ *macintosh*, *science* $\succ$ *maths*, *sport* $\succ$ *rugby*, *interface* $\succ$ *gui*, *hotel* $\succ$ *hostel*, *humor* $\succ$ *satira*, *browsers* $\succ$ *opera* (FP considerados) *os* $\succ$ *xp*, *realty* $\succ$ *home*, *security* $\succ$ *firewall*, *gaming* $\succ$ *wii*, *decoration* $\succ$ *furniture*, *sourcecode* $\succ$ *java*, *relationships* $\succ$ *marriage*, *learning* $\succ$ *research*, *standards* $\succ$ *xml*
CAT2	*awk* $\succ$ *sripting*, *javascript* $\succ$ *firefox*, *vm* $\succ$ *virtualization*, *airline* $\succ$ *plane*, *hotels* $\succ$ *accommodation*, *sea* $\succ$ *marine*, *nuclear* $\succ$ *atomic*, *browse* $\succ$ *plugin*, *proxy* $\succ$ *http*;
CAT3	*ebay* $\succ$ *auction*, *sudoku* $\succ$ *games*, *cat* $\succ$ *pet*, *regex* $\succ$ *expressions*, *candy* $\succ$ *food*, *amd* $\succ$ *hardware*, *phone* $\succ$ *telecommunications*, *english* $\succ$ *language*, *bmw* $\succ$ *car*
CAT4	*what* $\succ$ *best*, *after* $\succ$ *and*, *companies* $\succ$ *trojan*, *frequent* $\succ$ *interesting*
CAT5	*windows* $\succ$ *linux*, *web* $\succ$ *photography*, *colours* $\succ$ *resource*, *chicken* $\succ$ *recipes*, *todo* $\succ$ *social*, *cool* $\succ$ *app*, *tools* $\succ$ *python*, *design* $\succ$ *service*, *photos* $\succ$ *beautiful*

De um modo geral, observa-se que a maioria das predições efetuadas pela abordagem CPDST estão concentradas nas categorias CAT1 e CAT2. Além das categorias citadas, dentre as predições corretas (VP) é possível encontrar algumas relações isoladas que estão mais próximas de um significado de sinonímia como, por exemplo, *football* $\succ$ *soccer*, *work* $\succ$ *job*, *image* $\succ$ *picture* e *holland* $\succ$ *netherlands*. Outras ocorrências em maior quantidade também são percebidas nos falsos positivos (FP) como, por exemplo, *maganizes* $\succ$ *magazine*, *children* $\succ$ *kids* e *helpdesk* $\succ$ *support*. É previsível que casos de sinonímia sejam compreendidos como relações de subordinação pela abordagem

CPDST, uma vez que emprega atributos que estimam similaridade semântica e que também são indicados para detecção de sinonímia (e.g., similaridade do cosseno).

As relações referentes à categoria CAT3 são vistas com pouca frequência. Essas ocorrências acontecem devido à decisão de considerar para treinamento as instâncias (t, t') na qual $|R_t| < |R'_t|$. Na categoria CAT4, também é reduzido o número de relações que se enquadram nessa classificação. Porém, a categoria CAT5 acumula uma série de exemplos de relações de subordinação inconsistentes, compostas pela combinação de *tags* populares[32] no sistema Delicious com *tags* de inferior frequência de uso. A alta frequência de uso das *tags* mais populares do Delicious faz com que elas coocorram com milhares de outras *tags* ao longo dos recursos, influenciando na maior incidência de exemplos da categoria CAT5.

6.2.2 Teste de Significância

O desempenho da abordagem CPDST para detecção de relações de subordinação na base de dados do BibSonomy apresentou praticamente a mesma acurácia em ambas configurações experimentais C1 e C2. Verifica-se que o desempenho do melhor dos *baselines* ($tsearch$) é mais acentuado quando avaliado sob à técnica de avaliação *10-fold cross validation*. Com isso, a abordagem CPDST é 10,3% superior a $tsearch$ usando a técnica *10-fold cross validation* e 48,6% superior com a técnica *holdout*. Presumivelmente, a avaliação utilizando a técnica *holdout* produz maior variância em relação à técnica *10-fold cross-validation*, pois a avaliação depende de como os pontos de dados foram distribuídos nos conjuntos de treinamento e teste, logo a avaliação pode ser significativamente diferente dependente de como foi efetuada a divisão dos dados. Com a base de dados do Delicious, os resultados mostraram que a abordagem CPDST é 15,6% superior ao melhor dos *baselines* (overlap^{hyp}) na CE1 e 30,2% superior na CE2.

Um teste estatístico foi conduzido para determinar se a diferença na média de acurácia da abordagem CPDST comparada com $tsearch$ (BibSonomy) e overlap^{hyp} (Delicious) é estatisticamente significante, após múltiplas execuções. Para isso, o seguinte teste de

[32] As *top-2 tags* mais populares no Delicious foram utilizadas, respectivamente, 1.802.552 (design) e 1.666.643 (software) vezes. No sistema BibSonomy, as *top-2 tags* mais populares foram aplicadas, respectivamente, 12.058 (Deutschland) e 11.839 (ZZZ_TO_SORT) vezes.

hipótese foi definido:

$\mathbf{H}_0$: Os métodos de detecção de relações de subordinação possuem a mesma acurácia em termos de média ($\mu_1 = \mu_2$).

$\mathbf{H}_1$: Há diferença em *f-measure* entre os métodos comparados ($\mu_1 \neq \mu_2$), o que implica em dizer que a abordagem CPDST apresenta acurácia estatisticamente superior.

BibSonomy: Ao confirmar a suposição de normalidade dos dados ao longo de 10 execuções em CE1 e CE2, executou-se o teste paramétrico teste-t de *student* pareado de duas caldas (*paired t-test 2-sided*) com 95% de nível de confiança. Os resultados *p-valor= 2.844e-07 < 0.05* (CE1) e *p-valor= 2.879e-05 < 0.05* (CE2) rejeitam a hipótese nula, indicando que a abordagem CPDST para a detecção de relações de subordinação é estatisticamente significante superior ao melhor dos *baselines* $tsearch$ com 95% de confiança.

Delicious: Uma vez que as medições de *f-measure* ao longo de 10 amostras segue uma distribuição normal, aplicou-se o teste estatístico teste-t de *student* pareado de duas caldas com 95% de confiança. Com um *p-valor= 9.59e-11 < 0.05* (CE1) e *p-valor= 4.408e-06 < 0.05* (CE2), a hipótese nula é rejeitada, logo deduz-se que a abordagem CPDST é estatisticamente significante superior ao *baseline* overlap^{hyp} na tarefa de detecção de relações de subordinação.

Em suma, o resultado do teste de significância responde a QP2, atestando que o uso de uma técnica de classificação para a detecção de relações de subordinação provê melhor acurácia em relação às heurísticas empregadas para o mesmo propósito.

6.3 Influência dos Dicionários Eletrônicos na Avaliação da Abordagem CPDST

Recapitulando o que foi reportado na Seção 4.3, as instâncias de treinamento (sinonímia e subordinação) foram rotuladas automaticamente por meio dos dicionários eletrônicos WordNet e ConceptNet. Embora o uso desses dicionários tenha legitimado como a melhor escolha para a realização da tarefa, constatou-se que a atuação dos mesmos impõe limitações na avaliação da abordagem CPDST.

Mesmo com o suporte de dois dicionários para efetuar a rotulação automática de instâncias de pares de *tags*, o número de casos positivos reconhecidos ainda é largamente

inferior ao volume de casos negativos, representando menos de 1% da cobertura das instâncias processadas dependendo da base de dados. Esse comportamento se justifica basicamente pelos seguintes motivos: (i) o próprio cenário é espontaneamente tendencioso ao desbalanceamento; (ii) os dicionários são rigorosos por natureza, pois desconsideram relações semânticas entre termos que intuitivamente são admitidos no cotidiano; (iii) relações entre conceitos populares em um domínio específico também carecem de reconhecimento por parte dos dicionários eletrônicos; e (iv) termos emergentes demandam maior tempo para que sejam estabelecidos os relacionamentos semânticos com outros conceitos relacionados.

Considerando os argumentos expostos, por ser um recurso independente e desconectado do domínio das folksonomias, os dicionários eletrônicos não assumem o compromisso de maximizar a cobertura da base de dados. Consequentemente, essa limitação contribui para que o problema de desbalanceamento seja intensificado, pois muitas instâncias positivas deixam de ser descobertas. Além disso, um menor quantitativo de instâncias positivas pode afetar a construção de um bom modelo de predição, principalmente quando é aplicado o filtro de coocorrência porque vários exemplos positivos são perdidos por não estarem em conformidade com o limiar (cf. Tabela 5.3).

Alguns relacionamentos semânticos podem ser mais complexos do que outros para aprender. A abordagem CPDST não enfrentou maiores dificuldades para conseguir uma relevante acurácia na detecção de relações de sinonímia diante das limitações do rotulador semântico (cf. Tabela 6.3). Entretanto, a detecção de relações de subordinação demonstrou maior dependência da acurácia dos dicionários para obter uma melhor desempenho na predição da classe positiva (cf. Tabela 6.5). A própria abordagem CPDST, utilizando-se de sua habilidade de generalização, inferiu relações semânticas que não estão nos dicionários eletrônicos, mas que foram avaliadas como corretas por alguns estudantes de pós-graduação pelo julgamento de bom senso.

A rotulação de relações de subordinação utilizando julgamento humano pode ser uma alternativa que venha contribuir para a melhoria da avaliação da abordagem CPDST. No entanto, conduzir experimentos de avaliação com o usuário implicam em maiores desafios visto que a tarefa de julgamento pode apontar diferenças ocasionadas pela subjetividade do avaliador. Por exemplo, uma relação de subordinação pode ser clara para um avaliador A_1 e ao mesmo incorreta sob o ponto de vista de outro avaliador A_2. Se a variação de

julgamento humano subjetivo for elevada, a avaliação se torna mais difícil. Mesmo diante dessa perspectiva, a necessidade de avaliação das relações semânticas com o usuário é evidente, embora não seja uma tarefa trivial para o problema em particular.

6.4 Custo vs. Benefício do Aprendizado de Máquina para Detecção de Relações Semânticas

Os testes de significância reportados nas Seções 6.1.3 e 6.2.2 indicam que, sob o ponto de vista estatístico, a abordagem CPDST provê um ganho de acurácia significante em relação ao *baseline* de melhor desempenho em cada base de dados. Esta seção amplia a discussão sobre o desempenho da abordagem CPDST com o intuito de examinar a relação custo vs. benefício do uso da técnica de classificação para o problema específico de detecção de relações semânticas de sinonímia e subordinação entre *tags* de folksonomias. Além do ganho estatístico, as questões de pesquisa QP1 e QP2 buscam responder *"Até que ponto usar a técnica de classificação em lugar das heurísticas melhora a acurácia na detecção de relações semânticas"*, de maneira que possa ser enfatizada as circunstâncias em que o uso da técnica de AM apresenta sinais de inviabilidade. Os resultados de *f-measure* obtidos com o uso da técnica *10-fold cross-validation* são utilizados como suporte para justificar o ganho de cada método/abordagem.

De fato, o uso de heurísticas para detectar similaridade semântica entre *tags* representa um método simples se comparado ao uso de AM. Neste caso, o princípio básico consiste em extrair as informações necessárias diretamente da fonte de dados, aplicar as transformações necessárias para derivar uma representação conveniente e armazená-las em um banco de dados ou arquivo texto. A partir dos dados pré-processados, o cálculo da heurística pode ser então codificado em uma linguagem de programação. Portanto, em pouco tempo é possível colocar em prática a heurística e ter acesso aos resultados. Além do ganho proporcionado pelo baixo custo de implementação, a utilização de heurísticas se mostrou viável para detectar sinônimos em amostras de dados balanceadas, como observado nos experimentos realizados com a base de dados do BibSonomy (RÊGO; MARINHO; PIRES, 2012), na qual distância de edição teve uma acurácia muito próxima à apresentada pela abordagem CPDST. Em oposição aos seus benefícios, as heurísticas se mostram insuficientes nos seguintes

quesitos:

- Individualmente, cada heurística focaliza um aspecto particular de uma relação semântica. Por exemplo, para descoberta de sinonímia, distância de edição é especializada em detectar similaridade sintática entre cadeia de caracteres distintas. No entanto, a heurística falha na detecção de outros casos de sinonímia decorrentes do uso de acrônimos, preferências de vocabulário ou termos distintos. Esta limitação afeta a cobertura da heurística;

- O desempenho das heurísticas tende a ser prejudicado com o aumento do desbalanceamento de classes e, de acordo com as características da base de dados, pode ser afetado com maior intensidade;

- As medidas de similaridade empregadas como heurística não conseguem discernir a especificidade do relacionamento semântico. Na prática, elas atuam como um indicador de nível semântico entre duas *tags*, podendo capturar diferentes tipos de relações semânticas;

- Para detecção de relações de sinonímia e subordinação, as melhores heurísticas apresentaram, respectivamente, médias de taxa de acurácia para a classe positiva em torno de 53,6% e 15,2%.

O uso da técnica de AM demanda um custo de implementação maior quando comparado com as heurísticas. Para aprender um modelo de classificação, é necessário extrair vetores de características e efetuar o procedimento de rotulação das instâncias de treinamento. Mesmo que a rotulação seja feita de forma automática, a quantidade de instâncias a ser rotulada e o tempo de resposta do dicionário eletrônico para determinar o rótulo pode acarretar em uma tarefa de longa duração. Por exemplo, a rotulação automática realizada pelo Wordnet é muito mais rápida em relação ao acesso via ConceptNet, uma vez que neste último são realizados acessos online sucessivos. Além disso, o desempenho do modelo preditor está condicionado à acurácia dos dicionários eletrônicos em identificar os casos positivos.

A maior dificuldade em utilizar a técnica de AM para detecção de relações semânticas reside em seu ciclo de concepção, principalmente em bases de dados numerosas, uma vez que os esforços estão concentrados nas fases de seleção, pré-processamento, transformação

e escolha do algoritmo de classificação ideal para o problema abordado. Na eventualidade de desbalanceamento de classes, técnicas devem ser empregadas para viabilizar a tarefa de classificação, o que pode requerer mais tempo para definir os ajustes necessários. Apesar da complexidade requerida para se chegar ao melhor modelo de predição, os resultados mostram que os custos de se utilizar AM compensam os benefícios pelas seguintes razões:

- Para detecção de sinonímia, a abordagem CPDST apresenta um ganho médio na acurácia de 73,9% em relação à melhor heurística, enquanto que para detecção de relação de subordinação o ganho médio na acurácia obtido com a técnica de classificação é de 13,2%. Embora o ganho revelado pelos experimentos de detecção de subordinação seja relativamente pequeno em relação ao conseguido nos experimentos para detecção de sinonímia, ainda assim se trata de um resultado significativo diante da dificuldade de determinar com precisão esse tipo de relação semântica;

- Ao implementar a técnica de AM, a abordagem CPDST pode ser estendida facilmente por meio da inclusão de novos atributos, permitindo desta forma que o modelo de predição seja aprimorado com a descoberta de novos padrões de dados. Eventualmente, esta versatilidade tende a influenciar a taxa de acurácia;

- Com a técnica de AM, a abordagem CPDST consegue agregar diferentes propriedades de uma relação semântica em um único modelo, ampliando desta forma o reconhecimento de instâncias associadas à relação semântica de interesse.

Em síntese, os compromissos relativos ao custo vs. benefício relatados nesta seção se mostram favoráveis para utilizar a técnica de AM orientada à detecção de relações semânticas do tipo sinonímia e subordinação, uma vez que o ganho na taxa de acurácia compensa o esforço empreendido. As heurísticas, isoladamente, perdem em precisão para desempenhar este tipo de tarefa. A aplicabilidade da técnica de AM neste cenário tende a ser comprometida quando: (i) os dicionários eletrônicos não conseguem prover uma cobertura de rotulação viável, o que afeta o treinamento do classificador, e (ii) a ocorrência de *overlap* de classes é intensa ao ponto de impossibilitar a discriminação das classes por qualquer algoritmo de classificação.

6.5 Seleção de Atributos

O conjunto final de medidas de similaridade adotadas como atributos de aprendizado contempla o resultado da aplicação da medida *Gain Ratio*, priorizando os atributos que maximizam o *gain ratio* e resultam em medições maiores do que 0. *Gain ratio*, em português "razão de ganho", é uma medida que avalia a utilidade de um atributo ao medir a taxa de ganho que o mesmo proporciona à discriminação das classes a serem aprendidas. Trata-se de uma modificação da medida *Information Gain* - IF (WITTEN; FRANK, 2011) que visa reduzir o viés provocado por IF de preferir atributos que apresentam a maior faixa de valores possível.

A técnica de Seleção de Atributos é comumente utilizada em AM para selecionar o melhor subconjunto de atributos disponível em uma base de dados para aplicação de um algoritmo de aprendizado, descartando desta forma aqueles que por ventura não favoreçam o processo de aprendizado por serem redundantes ou irrelevantes (LADHA; DEEPA, 2011). A fim de determinar os melhores atributos, experimentos de seleção de atributos foram então conduzidos, norteados pelas métodos existentes na literatura. O método *Wrapper* foi utilizado para realizar essa tarefa porque, segundo Talavera (2005), se trata de uma opção que tende a ser superior ao método *Filter* em problemas de aprendizado supervisionado. No método *Wrapper*, o algoritmo indutivo estima o valor de um subconjunto de atributos usado para treinar o próprio modelo, levando em conta o viés particular do algoritmo. O método *Filter* seleciona atributos com base em critérios discriminativos, tornando-o independente de qualquer algoritmo particular. Os melhores atributos foram avaliados no conjunto de validação, sendo selecionados aqueles que proveram o melhor resultado em termos de *f-measure*.

Nesta seção, uma discussão mais detalhada acerca do experimento de seleção de atributos é descrito. A determinação de um conjunto de atributos conveniente para detectar relações semânticas de sinonímia e subordinação denota uma contribuição adicional ao trabalho desenvolvido nesta tese, uma vez que os algoritmos de classificação utilizados para treinamento não são originais. O cenário específico de detecção de sinonímia usando a base de dados do BibSonomy é tomado como referência para enfatizar a escolha dos atributos.

O método *Wrapper* foi implementado levando-se em conta o metaclassificador AdaBoost

com o classificador base C4.5. Uma variação dos atributos apresentados na Seção 4.2.1 foi efetuada de maneira controlada para determinar a influência dos mesmos na acurácia do classificador. No total, 31 combinações foram examinadas com a técnica *1 run 10-fold cross-validation*, considerando desde subconjuntos com um único atributo até a situação em que todos os atributos são utilizados.

A Tabela 6.8 mostra os subconjuntos de atributos que apresentaram as melhores e piores medições de *f-measure* (cada subconjunto contendo pelo menos dois atributos). Por questão de concisão, optou-se apenas pela exibição dos *top*-5 melhores e piores desempenhos e seus respectivos subconjuntos de atributos para fins de análise.

Tabela 6.8: Seleção de atributos.

Subconjunto de Atributos	***f-measure***
Melhor Desempenho	
DE, $\cos^{\text{tag-tag}}$, $\cos^{\text{tag-res}}$	0,946
DE, $\cos^{\text{tag-tag}}$, $\text{cooc}^{\text{tag-tag}}$	0,941
DE, $\cos^{\text{tag-tag}}$, , $\cos^{\text{tag-res}}$, $\text{cooc}^{\text{tag-tag}}$	0,940
$\cos^{\text{tag-tag}}$, $\cos^{\text{tag-res}}$, $\text{cooc}^{\text{tag-tag}}$	0,939
$\cos^{\text{tag-tag}}$, $\cos^{\text{tag-res}}$	0,935
Pior Desempenho	
$\text{overlap}^{\text{syn}}$, $\text{cooc}^{\text{tag-tag}}$	0,000
$\text{overlap}^{\text{syn}}$, $\cos^{\text{tag-tag}}$	0,059
$\text{overlap}^{\text{syn}}$, $\cos^{\text{tag-res}}$	0,062
$\text{overlap}^{\text{syn}}$, $\cos^{\text{tag-res}}$,$\text{cooc}^{\text{tag-tag}}$	0,362
$\text{overlap}^{\text{syn}}$, DE	0,452

Com os dados do BibSonomy, a distância de edição foi o atributo individual que agregou maior ganho na tarefa de classificação. Ao expandir o resultado da Tabela 6.8 para os *top*-10 melhores subconjuntos de atributos, constatou-se que distância de edição está presente em 7 subconjuntos. Na situação em que os subconjuntos foram formados apenas por um único atributo, distância de edição apresentou *f-measure = 0,478*, seguido pelos atributos $\cos^{tag-tag}$ (*f-measure = 0,213*) e $\cos^{tag-res}$ (*f-measure = 0,144*). Combinado com outros atributos, $\cos^{tag-tag}$ e $\cos^{tag-res}$ melhoram a acurácia na classificação.

Os atributos overlap^{syn} e $\text{cooc}^{tag-tag}$ isoladamente não ofereceram ganhos na classificação. Combinado com outros atributos, overlap^{syn} evidencia um desempenho

inferior se comparado a cooc$^{tag-tag}$ principalmente quando é arranjado em pares. Isto é percebido na Tabela 6.8, visto que os subconjuntos que denotaram a menor acurácia incluem o atributo overlap^{syn}. Portanto, de acordo com o experimento de seleção de atributos para detecção de sinonímia usando a base de dados do BibSonomy, foi identificada a seguinte ordem de contribuição dos atributos (da maior para a menor contribuição): distância de edição, cos$^{tag-tag}$, cos$^{tag-res}$, cooc$^{tag-tag}$ e overlap^{syn}.

6.6 Considerações Finais

Neste capítulo, foram apresentados os resultados experimentais que serviram de apoio para responder as questões de pesquisas QP1 e QP2 propostas nesta tese de doutorado, utilizando duas bases de dados distintas. Os dois principais interesses inerentes à tarefa de detecção de semântica avaliados foram: (1) descoberta de sinonímia, e (2) descoberta de relações de subordinação.

Os resultados indicaram que CPDST apresenta maior desempenho para a detecção de sinonímia se comparado à tarefa de detecção de relações de subordinação. O desempenho dos *baselines*, isoladamente, é deteriorado com o agravamento do nível de desbalanceamento.

Embora as heurísticas empregadas para detecção de relações de subordinação revelem *tags* que subjetivamente expressam a ideia de hierarquia entre conceitos, observa-se que tais relações na maioria das vezes não são reconhecidas pelo WordNet ou ConceptNet como uma relação hiperonímia $\succ$ hiponímia. Esse comportamento também se percebe na abordagem CPDST e, consequentemente, afeta a acurácia na predição. Uma discussão detalhada sobre as técnicas empregadas para seleção dos atributos mais significativos foi conduzida, especificamente para a tarefa de deteção de sinonímia usando os dados do BibSonomy. Os resultados apontaram que o atributo distância de edição proporciona o maior ganho na tarefa de classificação da abordagem CPDST. No próximo capítulo, discute-se a aplicabilidade da abordagem CPDST para sugerir *tags* relacionadas semanticamente.

Capítulo 7

Geração de Listas de *Tags* Relacionadas Semanticamente

Neste capítulo, discute-se a aplicabilidade da abordagem CPDST para a tarefa de geração de listas de *tags* semanticamente relacionadas a uma *tag* de busca particular. Inicialmente, é apresentada uma contextualização do problema seguida das contribuições que a abordagem CPDST acrescenta à literatura. Um extenso conjunto de experimentos foi realizado a fim de avaliar a relevância das *tags* providas pela abordagem CPDST em comparação com outros métodos de captura de semântica entre *tags*, visando responder a QP3. A avaliação é realizada sob a perspectiva das métricas relevância e *overlap*.

7.1 Motivação

O usuário que interage com uma folksonomia está normalmente interessado em recuperar seus próprios recursos, quando possui credenciais para se identificar no sistema, ou explorar o conteúdo compartilhado pela comunidade de forma colaborativa. As folksonomias oferecem diferentes mecanismos para atender a essa necessidade. Quando a necessidade de informação ainda não está bem definida, uma interface conveniente de recuperação denominada *nuvem de tags* oferece ao usuário web um ponto de partida para ajudá-lo a definir seu objetivo de busca. A nuvem de *tags* provê uma visão sumarizada acerca dos principais tópicos praticados em um sistema, na qual as *tags* são visualizadas em ordem alfabética e visualmente ponderadas em tamanho de fonte de acordo com sua

popularidade (HASSAN-MONTERO; HERRERO-SOLANA, 2006; LEGINUS; DOLOG; LAGE, 2013b; RIVADENEIRA et al., 2007).

Em tarefas de RI, existe a necessidade de exploração de recursos relacionados ao contexto de uma chave de busca específica fornecida pelo usuário (simples ou múltipla), como uma forma de otimizar o acesso ao conteúdo que corresponde às pretensões de busca. O método mais utilizado é a busca direta condicionada a uma chave de busca, na qual são recuperados os recursos que foram marcados explicitamente por uma *tag* especificada.

No sistema BibSonomy, uma busca por palavra-chave resulta em uma página web contendo as publicações científicas e *bookmarks* anotados pela *tag* especificada. Além disso, é oferecido aos usuários uma opção de busca indireta por meio da sugestão de uma lista de *tags* relacionadas (*Related Tags*) e similares (*Similar Tags*) em relação à chave de busca, para que os usuários possam navegar por outros conceitos semanticamente relacionados. As opções de busca citadas estão realçadas na lateral direita da Figura 7.1. Neste caso, ao clicar em uma das *tags* da lista, serão exibidos os recursos anotados pela *tag* escolhida por qualquer usuário do sistema. A lista de *tags* relacionadas é construída por meio da aplicação da heurística de frequência de coocorrência entre *tags*. A lista de *tags* similares é embasada na aplicação da medida do cosseno em perfis de vetores *tag-tag*.

Esta funcionalidade diversifica a tarefa de busca para os usuários e se torna útil em alguns aspectos. Por exemplo, usuários com interesse exploratório podem usar as *tags* da lista para acessar e examinar um grupo de recursos, podendo retroceder para selecionar outras *tags* se achar necessário. Outros usuários com necessidade de busca mais específica podem utilizar a lista de *tags* para estreitar a conclusão da tarefa, uma vez que a lista de *tags* pode apresentar termos que de imediato combinem com o real interesse do usuário.

Como pode-se constatar, o BibSonomy disponibiliza as funcionalidades *tags relacionadas* e *tags similares* para aprimorar a navegação utilizando *tags* semanticamente relacionadas. Entretanto, as heurísticas empregadas pelos referidos métodos não são capazes de especificar qual o tipo de relacionamento semântico que as *tags* da lista compartilham com a chave de busca. A concepção de novas técnicas de geração de listas de *tags* pode aperfeiçoar a experiência do usuário na navegação à procura de recursos de interesse.

Nesta tese, propõe-se a aplicação da abordagem CPDST como uma nova estratégia para geração de listas de *tags* semanticamente relacionadas a partir de uma chave de busca inicial.

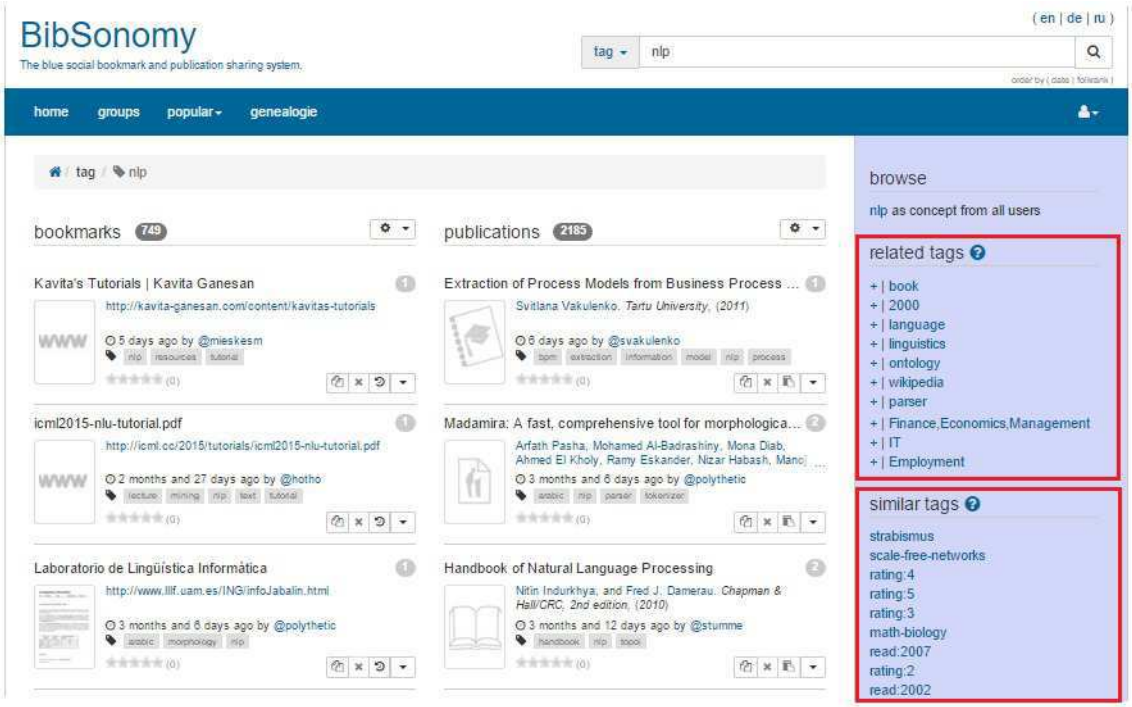

Figura 7.1: Interface de recuperação do BibSonomy com as opções de busca *related tags* e *similar tags* relacionadas à *tag* de busca *nlp*.

Como retratado no Capítulo 4, a abordagem CPDST pode ser ajustada para detectar dois relacionamentos semânticos mais específicos entre *tags*: sinonímia e subordinação, além de uma configuração que incorpora ambos os tipos de relações. Além da abordagem CPDST, propõem-se algoritmos fundamentados no acesso ao WordNet e ConceptNet para prover listas de *tags* com a mesma especificidade semântica. Com os novos métodos, amplia-se a perspectiva de geração de listas de *tags* relacionadas para atender aos propósitos dos usuários que desejam tanto efetuar uma busca exploratória quanto procurar algo mais específico.

A inclusão da abordagem CPDST como estratégia para geração de *tags* relacionadas adiciona as seguintes contribuições à literatura:

- Concepção de novos métodos para geração de listas de *tags* relacionadas em que as *tags* selecionadas denotam um relacionamento semântico mais específico em relação à chave de busca inicial. Os métodos propostos são especializados na recuperação de *tags* sinônimas e hipônimas, enquanto que os métodos existentes fundamentam-se na mensuração de grau de parentesco. Logo, não há a preocupação em discernir qual o relacionamento semântico que as *tags* recuperadas estabelecem com a chave de busca;

- Avaliação quantitativa de diferentes métodos de geração de *tags* relacionadas utilizando métricas que quantificam artificialmente a satisfação dos usuários em termos de relevância e capacidade das *tags* da lista de conduzir a um conjunto complementar de recursos (*overlap*).

7.2 Modelagem do Problema

Existem na literatura diferentes métodos empregados para a detecção de semântica, cada um fundamentado em uma estratégia distinta como, por exemplo, aplicação de medidas de distância semântica, análise de distribuição de *tags*, classificação e suporte de *thesaurus* eletrônico. Para unificar estratégias tão heterogêneas, implementou-se um *framework* extensível denominado *Semantic-Aware Tag List Generation* (SATLG), o qual permite agregar métricas adicionais de avaliação e outros métodos de detecção de semântica, além de modificar parâmetros para analisar diferentes configurações de busca.

Suponha a existência de uma função total de similaridade entre *tags* $sim(t, t')$ que recebe duas *tags* de entrada e produz um valor de saída no intervalo [0,1]. Considere os conjuntos R e T definidos na Seção 2.4. Cada recurso $r \in R$ está associado a um subconjunto de *tags* em T, simbolizado por T_r. Similarmente, R_t denota o conjunto de recursos em R associado a $t \in T$. Seja $q \in T$ uma *tag* que representa a chave de busca fornecida pelo usuário. Denota-se T_q^s o conjunto de *tags* relacionadas à q pela relação semântica s (e.g., sinonímia), extraídas a partir da aplicação de uma função de similaridade $sim : T^2 \rightarrow [0, 1]$ no espaço de *tags*. Quanto maior o valor de $sim(q, t)$, maior a similaridade semântica entre as duas *tags*. Obviamente, $T_q^s \subseteq T$.

O conjunto $R_{T_q^s} \subseteq R$ denota o conjunto de recursos associados com as *tags* existentes em T_q^s. Logo, para uma *tag* $t \in T_q^s$, $A_{R_{T_q^s}}(t)$ corresponde ao subconjunto de recursos em $R_{T_q^s}$ associados à *tag* t. Por definição, $A_{R_{T_q^s}}(t) \subseteq R_{T_q^s}$. O objetivo é selecionar uma lista de *tags* semanticamente relacionadas T_q^s de tamanho máximo k, gerada a partir de uma chave de busca particular $q \in T$. De agora em diante, assuma que $q \in T$ é a chave de busca usada pelo usuário.

A Figura 7.2 ilustra uma visão geral do funcionamento do *framework* SATLG para geração de listas de *tags*.

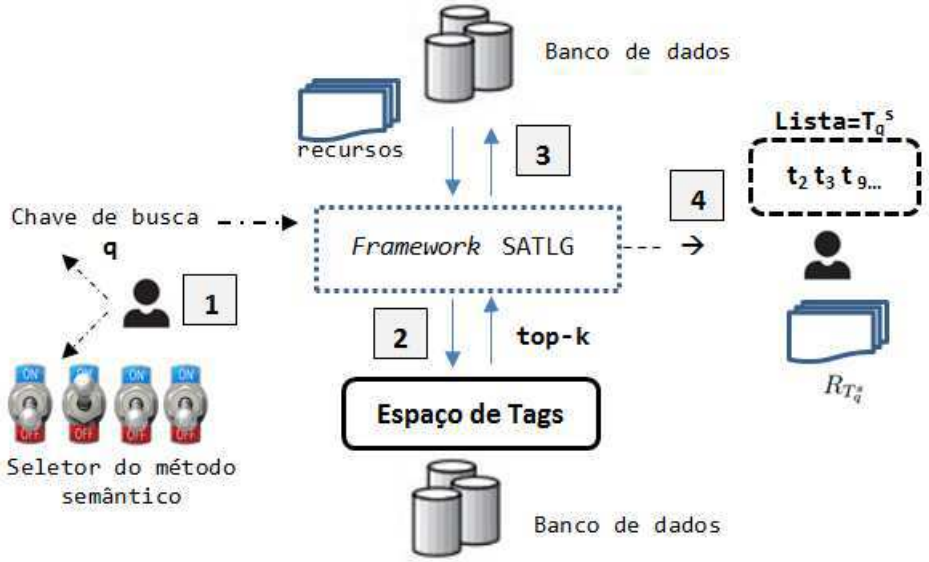

Figura 7.2: SATLG - Visão Geral do *Framework* para Geração de Listas de *Tags*.

A princípio, o usuário seleciona o método de detecção de semântica desejado por meio de uma interface de usuário e expressa sua necessidade de informação fornecendo uma chave de busca q (passo 1). Em seguida, a chave de busca é interpretada pelo SATLG que se encarrega de examinar o espaço de *tags* e extrair um subconjunto T_q^s com as *top-k tags* semanticamente mais significativas com respeito a q e à relação semântica s (passo 2). As *top-k tags* mais significativas são determinadas pelo maior valor de *score* semântico entre q e $t \in T_q^s$, computado de acordo com o método empregado para gerar a lista de *tags* (cf. Seção 7.4.1). No passo 3, são recuperados os recursos que foram marcados diretamente por q. Finalmente, o SATLG apresenta ao usuário os recursos marcados por q e a lista de *tags* ordenada por *score* decrescente de similaridade (passo 4).

7.3 Métodos para Seleção de *Tags*

Apresenta-se nesta seção os métodos propostos para a geração de listas de *tags* relacionadas. A aplicação geral dos métodos consiste basicamente em: (i) executar o algoritmo de seleção de *tags* considerando uma chave de busca fornecida pelo usuário, e (ii) recuperar as *top-k tags* semanticamente mais relevantes para compor a lista de *tags* a ser apresentada ao usuário.

A princípio, ressalta-se que a partir de uma *tag* de busca q, várias *tags* semanticamente relacionadas podem ser retornadas por um método, inclusive em quantidade superior ao

parâmetro k que define o ponto de corte relativo às *top-k tags* de interesse. Dentre as *tags* retornadas, algumas podem não pertencer ao conjunto T, as quais são automaticamente descartadas. Na ocasião de $|T_q^s| > k$, o conjunto resultante T_q^s é então reduzido para as *top-k tags*. No caso de $|T_q^s| < k$, preserva-se todas as *tags* existentes no conjunto. Os métodos são especificados como segue, levando-se em conta as considerações expostas.

- *CPDST*: a abordagem CPDST percorre o espaço de *tags* T para predizer relações semânticas mais específicas em relação a q. O seu princípio de funcionamento pode ser conferido na Seção 4.1. Denota-se CPDST-S a referência para a abordagem direcionada à detecção de sinonímia, CPDST-H a referência para detecção de relações de subordinação e simplesmente CPDST para prover simultaneamente as funcionalidades delimitadas em CPDST-S e CPDST-H;

- *WordNet Search*: o dicionário linguístico WordNet é utilizado para fornecer termos que expressem uma relação semântica específica em relação a um conceito de busca. Deste modo, a partir de uma chave de busca q e um parâmetro k, *WordNet Search* consulta a base conceitual do WordNet e seleciona todos os termos que são sinônimos ou hipônimos em relação a q. Uma vez que o WordNet pode retornar termos $t \notin T$, justifica-se o descarte automático desses termos. Denota-se por *WSyn* a referência para uso do WordNet direcionado à detecção de sinônimos, *WHyp* a referência para detecção de hipônimos e simplesmente WNET quando as duas são incorporadas;

- *ConceptNet Search*: o ConceptNet disponibiliza uma base de conhecimento na web que pode ser consultada por intermédio de sua API. A mesma sistemática definida no método *WordNet Search* para a recuperação de termos sinônimos e hipônimos em relação a q também é empregada no método *ConceptNet Search*, diferenciando-se apenas na base de conhecimento utilizada. Denota-se por CSyn a referência ao método ConceptNet para detecção de sinonímia, *CHyp* a referência para detecção de relações de subordinação e simplesmente CNET quando ambas são incorporadas.

Vale ressaltar que a abordagem CPDST utiliza os próprios dados da folksonomia para induzir relações semânticas. Os métodos com base no WordNet e ConceptNet acessam uma base de conhecimento externa para produzir suas respectivas listas de *tags*. Para quantificar

diferentes propriedades das listas de *tags* geradas pelos métodos comparados, empregam-se as métricas definidas na Seção 7.4.2.

7.4 Metodologia Experimental

Nesta seção, descreve-se a metodologia utilizada para conduzir a análise comparativa dos métodos de seleção de *tags* propostos na Seção 7.3 em conjunto com os métodos *Related Tags* e *Similar Tags* oferecidos pelo sistema BibSonomy. A seguir, são delineadas as definições para cálculo do *score* semântico (Seção 7.4.1), métricas (Seção 7.4.2) e protocolo de avaliação (Seção 7.4.3).

7.4.1 *Score* Semântico

Recapitulando o que foi retratado na Seção 7.3, os métodos para seleção de *tags* recebem como parâmetro de entrada uma *tag* de busca q e um valor k que define a quantidade máxima de *tags* relacionadas a ser retornada. Uma vez que a quantidade de *tags* retornadas pode exceder o valor de k, as *tags* precisam ser arranjadas em ordem decrescente de *score* semântico para priorizar as mais significativas e então estabelecer o ponto de corte que origina o conjunto final T_q^s, *score* este resultante da aplicação da função de similaridade $sim(q,t)$ relatada na Seção 7.2. A seguir, são descritas as estratégias que fundamentam o cálculo da similaridade semântica em cada método de seleção de *tags* integrado ao *framework* SATLG.

- CPDST: por se tratar de uma tarefa de classificação, CPDST realiza predição para uma classe específica com base no reconhecimento de padrões assimilado na etapa de treinamento. Evidentemente, o modelo de predição não fornece explicitamente uma informação sobre o quão relacionadas semanticamente estão duas *tags* (q,t). Para contornar esta limitação, identificou-se um valor mensurável na saída de cada predição que corresponde à *distribuição de probabilidade das classes*, ou seja, um valor real no intervalo [0,1] que sinaliza a probabilidade de uma dada instância pertencer à classe (+) ou (-). Assim, utilizou-se esta informação como critério para especificar o *score* de cada predição positiva;

- ConceptNet: a submissão de uma busca ao ConceptNet para uma relação específica resulta em uma lista de estruturas de dados na notação JSON (cf. Seção 4.3) que contém todos os campos de uma aresta ConceptNet. Um destes campos é o peso que quantifica a consistência semântica entre o conceito de entrada q e o conceito recuperado t. Deste modo, o valor do peso é utilizado para mensurar o nível de semântica em relação a um dado par de *tags* (q, t);

- WordNet: diferentes medidas podem ser empregadas para mensurar o grau de similaridade semântica entre dois *synsets* do WordNet. Budanitsky e Hirst (2001) apresentam 5 dessas medidas em um experimento de avaliação de desempenho de um sistema real de correção ortográfica. A API WS4J[33] (*WordNet Similarity for Java*) fornece a implementação de todos os algoritmos de similaridade abordados por Budanitsky e Hirst (2001). Então, optou-se pela escolha de uma medida bem avaliada nos experimentos e que provesse valor de similaridade no intervalo real [0,1]. Como resultado, definiu-se a *Lin's similarity* (LIN, 1998) como medida de *score* semântico entre *tags* com base no WordNet. Porém, outra medida pode ser facilmente empregada para essa função.

Os métodos $cos^{tag-tag}$ e $cooc^{tag-tag}$ empregados pelo BibSonomy para exibir respectivamente as listas de *tags similares* e *tags relacionadas*, produzem naturalmente um valor entre [0,1] para quantificar o nível de parentesco existente entre (q, t). Por isso, a seleção das *top-k tags* que definem T_q^s é direta.

7.4.2 Métricas

A qualidade de uma lista de *tags* é normalmente determinada por avaliadores humanos que subjetivamente julgam sua composição em relação à sua utilidade para uma determinada tarefa. Por exemplo, dada a *tag* de entrada $south\_america$ na qual se está interessado em obter uma relação de *tags* hipônimas relacionadas, um avaliador humano poderia qualificar como pertinente as sugestões $tango$, $brazil$, $amazon$, e $conmebol$. No entanto, julgamentos realizados por avaliadores humanos são dispendiosos e difíceis de implementar devido a

[33] <https://code.google.com/p/ws4j/>

vários fatores, tais como motivação, disponibilidade, nível de interesse, experiência, entre outros (LEGINUS; DOLOG; LAGE, 2013b).

Alternativamente, métricas *offline* ou indiretas são aplicadas para avaliar quantitativamente diferentes propriedades de um conjunto de *tags*. Porém, a maior dificuldade consiste exatamente em encontrar as métricas propícias para avaliar a qualidade das listas de *tags* sob o ponto de vista semântico. No segmento de pesquisas voltadas à geração de nuvens de *tags*, diferentes métricas foram introduzidas para medir aspectos qualitativos de uma nuvem de *tags* tais como cobertura, relevância, *overlap*, coesividade, popularidade, independência e equilíbrio (VENETIS; KOUTRIKA; GARCIA-MOLINA, 2011).

Devido à ausência de métricas específicas para avaliar a qualidade de listas de *tags* relacionadas semanticamente, enxergou-se a oportunidade de adotar métricas procedentes do contexto de avaliação de nuvens de *tags* para assumir essa atribuição. As métricas **relevância** e ***overlap*** foram identificadas como as que provêem informações mais expressivas sob o ponto de vista do usuário. Entretanto, a escolha dessas métricas tem as suas implicações, pois ambas são tendenciosas a favorecer métodos que capturam coocorência. Uma vez que os métodos comparados são heterogêneos quanto à estratégia que adotam para selecionar suas *tags*, supostamente as métricas **relevância** e ***overlap*** não serão capazes de mensurar essencialmente os aspectos qualitativos de maneira imparcial. As métricas relevância e *overlap* são apresentadas formalmente nos parágrafos seguintes.

Relevância

Considerou-se **relevância** como a métrica mais importante para medir a qualidade das listas de *tags* providas pelos métodos sob a perspectiva do usuário. A noção de relevância institui que uma *tag* $t \in T_q^s$ é mais relevante para a chave de busca q quando a interseção dos conjuntos R_q e R_t é alta. Se a interseção for alta, isto sugere que pelo princípio de coocorrência o par (q, t) está relacionado semanticamente. Neste trabalho, adotou-se a definição de relevância introduzida por Leginus, Dolog e Lage (2013b).

Seja T_q^s o conjunto de *tags* originado a partir de q pela relação semântica s. A relevância de T_q^s é definida da seguinte forma:

$$rel(T_q^s) = avg_{t \in T_q^s} \frac{|R_t \cap R_q|}{|R_t|} \tag{7.1}$$

em que avg é uma função que determina a média aritmética de um conjunto de valores, R_t constitui o conjunto de recursos associados com a *tag* t e R_q o conjunto de recursos anotados com a *tag* de busca q. A métrica assume valores no intervalo [0,1]. Um valor próximo de 1 significa que a maioria dos recursos anotados por uma *tag* particular t é coberta pelos recursos anotados pela *tag* de busca q.

Overlap

O conjunto T_q^s é constituído por uma relação de *tags* distintas. Uma vez selecionadas, essas *tags* podem conduzir ao mesmo subconjunto de recursos em R_q. Para quantificar esse aspecto, empregou-se a métrica *overlap* para medir o nível de redundância entre pares de *tags*. Assim, dado $t_i \in T_q^s$ e $t_j \in T_q^s$, o *overlap* é computado medindo-se a porção de recursos anotados por t_i que também foram anotados por t_j. Seguindo a definição introduzida por Leginus, Dolog e Lage (2013a), o *overlap* em T_q^s é definido como segue:

$$overlap(T_q^s) = avg_{ti \neq tj} \frac{|R_{t_i} \cap R_{t_j}|}{min\{|R_{t_i}|, |R_{t_j}|}. \tag{7.2}$$

O $overlap(T_q^s)$ assume valores no intervalo [0,1]. Se $overlap(T_q^s)$ é próximo de 0, significa que a interseção dos recursos anotados por ambas as tags t_i e t_j é pequena. Neste caso, cada *tag* provê acesso a uma parcela diferenciada dos recursos anotados pelas *tags* da lista. Por outro lado, se $overlap(T_q^s)$ é próximo de 1, as *tags* tendem a ser indiscrimináveis, uma vez que a escolha de t_i ou t_j resulta no acesso ao mesmo conjunto de recursos.

7.4.3 Avaliação

A fim de realizar uma análise comparativa entre os diferentes métodos de geração de *tags* (métodos propostos na Seção 7.3 e os métodos providos pelo BibSonomy), são empregadas as métricas descritas na Seção 7.4.2. As bases de dados do BibSonomy e Delicious utilizadas para experimentação são as mesmas descritas na Seção 5.1).

As métricas só podem ser aplicadas quando, para uma dada chave de busca q, todos os algoritmos envolvidos conseguem gerar uma lista de *tags* com no mínimo k elementos.

Assim, o tamanho da lista de *tags* é adequado para o tamanho k de acordo com a ordenação do *score* de ranqueamento. Deste modo, impõe-se condições de igualdade em todos os métodos para aplicar a métrica de avaliação.

Na eventualidade de insuficiência de chaves de busca para consultas que produzam listas de *tags* para todos os métodos comparados, o experimento é segmentado em grupos de métodos. Uma vez que deseja-se observar o desempenho da abordagem CPDST sob o ponto de vista da métrica relevância e *overlap*, a abordagem CPDST deve estar presente em cada grupo para efeito comparativo.

7.5 Resultados

Nesta seção, apresentam-se os resultados individualizados por tipo de experimento. Cada experimento aborda uma tarefa que contribui para entender melhor o domínio do problema. Os algoritmos apresentados na Seção 7.3 foram implementados no *framework* SATLG e comparados quantitativamente utilizando a metodologia de avaliação apresentada na Seção 7.4.3 de acordo com a métricas definidas na Seção 7.4.2. A abordagem CPDST emprega os mesmos classificadores ressaltados na Seção 6, tendo em vista que apresentaram o melhor desempenho na acurácia da classe positiva dentre os classificadores examinados. Recapitulando, para seleção de sinônimos a abordagem CPDST utiliza o AdaBoost com algoritmo C4.5 e para seleção de hipônimos a metaclassificação ASC com algoritmo base Naive Bayes.

Na condução dos experimentos, utilizou-se a seguinte configuração: o método $cos^{tag-tag}$ (*BibSonomy Similar* ou simplesmente BSIM) assume um limiar mínimo de valor 0,4 para recuperar *tags* relacionadas semanticamente. O método $cooc^{tag-tag}$ (*BibSonomy Related* ou simplesmente BREL) utiliza a própria condição de filtro usada para selecionar os pares de *tags* que definem as instâncias (cf. Seção 5.1). O tamanho k da lista de *tags* varia no intervalo de 3 a 10 para aplicação das métricas. Os resultados estão delineados por tipo de experimento que captam diferentes informações acerca da natureza do problema.

Na análise dos gráficos apresentados a seguir, os métodos introduzidos na Seção 7.3 são referidos pelas abreviações CPDST, CPDST-S, CPDST-H, WNET, WSyn, WHyp, CNET, CSyn e CHyp, enquanto que os métodos do BibSonomy são referidos por BSIM e BREL.

Utiliza-se o termo *query* como referência para chave de busca fornecida pelo usuário.

7.5.1 Experimento 1: Frequência de Geração de Lista de *Tags*

Neste experimento, investiga-se a quantidade de listas de *tags* de tamanho n que cada abordagem consegue gerar, com $1 \leq n \leq 31$. Tendo em vista que alguns métodos conseguem gerar listas de *tags* com mais de 30 elementos, assume-se que o valor 31 representa essa faixa, ou seja, 31 ou mais elementos. Para obter essa resposta, todas as *tags* do conjunto T foram consideradas como *query*, as quais foram submetidas a cada método. De acordo com o resultado de cada método, elaborou-se a distribuição de frequência em relação ao tamanho da lista gerada. As Figuras abaixo exibem os gráficos da Função de Distribuição Acumulada (FDA) para os métodos BSIM e BREL (Figura 7.3) e CPDST, WNET e CNET (Figuras 7.4, 7.5 e 7.6, respectivamente), bem como suas variantes. A FDA descreve a probabilidade de que a variável aleatória X (conjunto de *tags* do experimento) assuma um valor inferior ou igual a determinado x, ou seja, $F(x) = P(X \leq x)$. Para cada x (*tag* usada como *query* para geração de listas de *tags*), a função $F(x)$ assume um valor diferente.

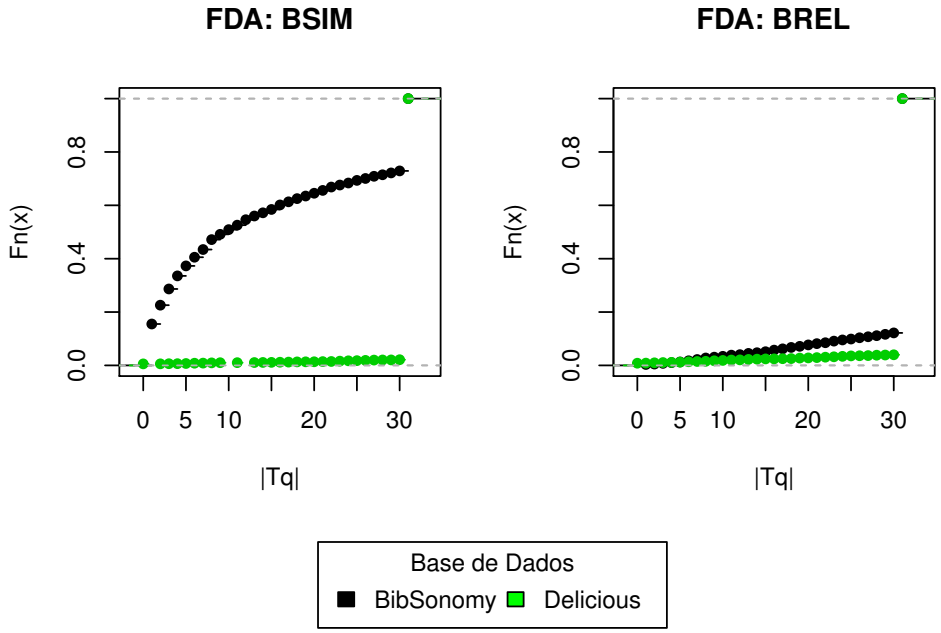

Figura 7.3: FDA dos métodos *Similar* e *Related* do BibSonomy.

Na base de dados do Delicious, os métodos BSIM e BREL produzem listas com 31 ou mais *tags* em mais de 96% das *queries* (Figura 7.3). Uma vez que no sistema Delicious o quantitativo de *tags* é numericamente superior ao existente no BibSonomy, aliado à elevada frequência de utilização das *tags*, observa-se que a quantidade de *tags* coocorrentes também aumenta na mesma proporção. Logo, a tendência é de que BSIM e BREL produzam listas de *tags* com cobertura potencializada.

Usando a base de dados do BibSonomy, o método BREL apresenta maior cobertura dentre os métodos comparados, produzindo listas de *tags* com 31 ou mais elementos em mais de 87% das *queries* analisadas. O método BSIM tem seus picos de geração em listas de tamanho [1-5] (25%) e [26-31] (30%). Vale ressaltar que BSIM e BREL são passíveis a não gerarem listas de *tags* para algumas *queries*, entretanto com uma baixa probabilidade especialmente para BREL (0,34%) se comparado com BSIM (15%). Os gráficos de FDA na Figura 7.3 reforçam a intuição de que os métodos embasados na estimativa de parentesco (em nível geral) são favoráveis ao fornecimento de listas de *tags*.

Como pode-se observar nas Figuras 7.4, 7.5 e 7.6, os métodos WNET, CNET e CPDST não proveem a mesma facilidade que BSIM e BREL de gerarem listas de *tags* para qualquer *query* nas duas bases de dados. Isto se deve à especificidade semântica abordada pelos métodos pois, embora uma *query* possa se relacionar com diversas *tags* em T, a mensuração de sinonímia ou hiponímia só é percebida em alguns pares de *tags*, diante das diferentes relações semânticas que podem estar presentes. Os referidos métodos se mostram mais restritivos quando são configurados para detectar apenas sinonímia (CPDST-S, WSyn e CSyn) ou hiponímia (CPDST-H, WHyp e CHyp).

Diante do cenário em que $|T_{bibsonomy}| = 3.177$ e $|T_{delicious}| = 4.790$, os métodos WHyp, CSyn e CPDST-H foram os mais limitados nas duas bases de dados analisadas. No BibSonomy, os respectivos métodos só conseguem gerar listas de *tags* para 36%, 25% e 8% das *queries* disponíveis. Na base de dados do Delicious, a porcentagem observada foi de 29% (CSyn), 23% (WHyp), e 22% (CPDST-H). Essa limitação é atenuada quando as listas são geradas com *tags* sinônimas e hipônimas (WNET, CNET e CPDST).

Os métodos fundamentados no WordNet (WNET, WSyn e WHyp) atuam melhor na base de dados do BibSonomy. Por outro lado, os métodos CPDST e CNET (e suas variantes) incrementam sua capacidade de gerar listas de *tags* com o aumento do volume da base de

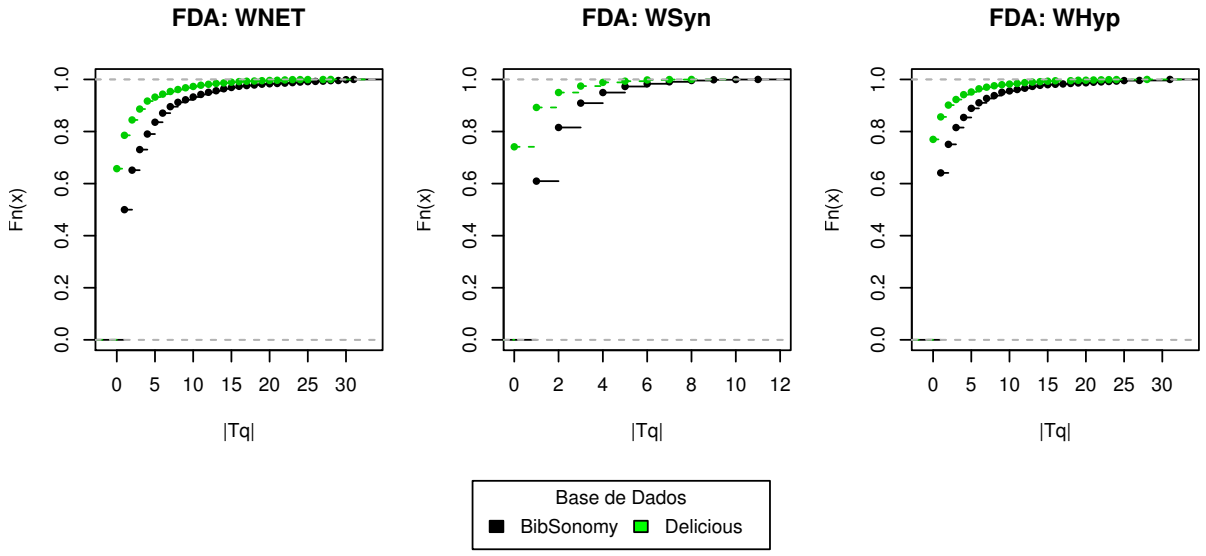

Figura 7.4: FDA do método Wordnet e suas variantes.

dados. Na base de dados do BibSonomy, WNET consegue gerar listas com 31 ou mais *tags* para apenas 3 *queries* em particular. Porém, os tamanhos de lista gerados pela maioria das *queries* se encontram no intervalo [1,10]. Na base de dados do Delicious, o maior tamanho de lista de *tags* observado foi 28, alcançado por uma única *query*.

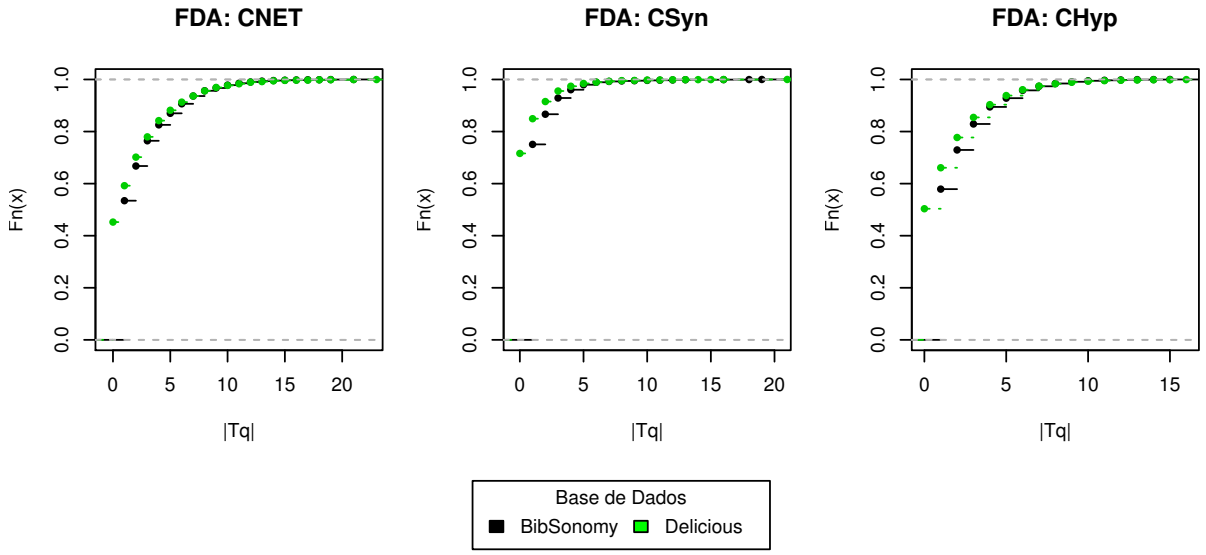

Figura 7.5: FDA do método ConceptNet e suas variantes.

Em relação ao método CNET, observou-se no BibSonomy o tamanho máximo de lista igual a 20 (1 *query*) e tendência das *queries* gerarem listas com tamanho no intervalo de [1,10]. No Delicious, o maior tamanho de lista observado foi 24 (1 *query*) e a maioria das *queries* produz tamanho de listas no intervalo de [1,12].

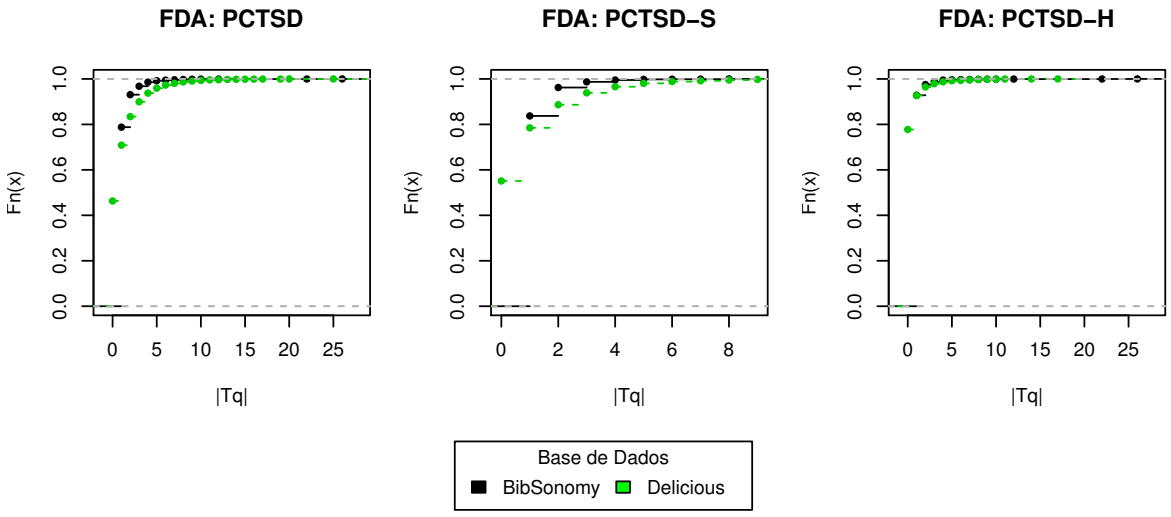

Figura 7.6: FDA do método CPDST e suas variantes.

Na abordagem CPDST, os tamanhos de listas produzidos pela maioria das *queries* se encontram nos seguintes intervalos: [1,5] na base de dados do BibSonomy e [1,7] na base de dados do Delicious. Observa-se um tamanho máximo de lista com 25 *tags* (1 *query*) nas duas bases de dados utilizadas. Por fim, os experimentos mostraram que, dentre os métodos concebidos para detecção de semântica específica, WNET provê melhor desempenho quanto à capacidade de gerar listas de *tags* na base de dados do BibSonomy. Na base de dados do Delicious, CNET sobressaiu-se como melhor método.

7.5.2 Experimento 2: Efetividade de Geração de Listas de *Tags*

No Experimento 1, observou-se que os métodos detectores de semântica mais específica (WNET, CNET e CPDST) são restritivos quanto à capacidade de gerar listas de *tags* para diferentes *queries*. O intuito deste experimento é avaliar o desempenho dos métodos quanto à sua aptidão de gerar listas de *tags* de qualquer tamanho considerando 100 *queries* escolhidas aleatoriamente. Em outras palavras, deseja-se levantar estatísticas úteis relacionadas às

seguintes questões: (i) para as 100 *queries* de entrada, qual método consegue gerar mais listas de *tags*?, (ii) qual a porcentagem de geração de listas de *tags* dos métodos para a amostra de *queries*?, e (iii) qual a média de tamanho de listas de *tags* gerada pelos métodos?

A Tabela 7.1 sumariza o resultado do experimento após 40 execuções (cada execução corresponde a uma amostra de 100 *queries* aleatórias). As questões são representadas por mnemônicos: MGQ = média geral de *queries* que resulta na criação de listas de *tags* de qualquer tamanho (das 100 escolhidas aleatoriamente), PQA = porcentagem de *queries* da amostra que geram listas de qualquer tamanho, e TML = tamanho médio das listas de *tags* geradas pelos métodos. Neste experimento, o tamanho das listas de *tags* varia no intervalo de 1 a 31.

Tabela 7.1: Desempenho de Geração de Listas de *Tags*.

Questão	Métodos de Geração de Lista de Tags										
	WNET	WSyn	WNET-H	CPDST	CPDST-S	CPDST-H	CNET	CSyn	CHyp	BSIM	BREL
BibSonomy											
MGQ	50	39	35	20	15	7	46	25	41	84	99
PQA	50,2%	39,4%	35,5%	20,8%	15,8%	7,2%	46,4%	25,0%	41,7%	84,5%	99,7%
TML	1	0	0	0	0	0	1	0	1	12	28
Delicious											
MGQ	34	25	22	53	45	22	55	29	50	99	99
PQA	34,5%	25,6 %	22,7%	53,7%	45,2%	22,6%	55,3%	29,3%	50,1%	99,3%	99,1%
TML	1	0	0	1	0	0	1	0	1	30	29

Algumas considerações podem ser extraídas da Tabela 7.1. Os métodos especializados em semântica explícita (CPDST, WordNet e ConceptNet) possuem menor probabilidade de prover listas de *tags* para uma *query* aleatória (MGQ). Para esses métodos, observa-se na base de dados do Delicious a ocorrência de listas de *tags* vazias para uma faixa de 45% a 65% das *queries*. Esta limitação é salientada quando se observa as variantes para detecção exclusivamente de sinônimos ou hipônimos.

A abordagem CPDST e suas variantes apresentaram o menor índice de geração de listas de *tags* usando os dados do BibSonomy (PQA). Porém, na base de dados do Delicious, a abordagem CPDST consegue equiparar-se ao melhor dos métodos correlatos (CNET). O baixo desempenho da abordagem CPDST se deve à pequena quantidade de instâncias positivas, uma vez que o classificador concentra seus esforços em acertar justamente esses casos. Logo, não há garantia que uma *query* aleatória seja contemplada com uma lista de *tags* pela abordagem CPDST, o que de certa forma é compreensível. Esta constatação afeta diretamente o item TML pois, no cálculo da média, as ocorrências de *queries* que resultam

em lista de *tags* nula degradam o desempenho da abordagem CPDST. Visto que WNET e CNET também apresentam limitações semelhantes, os resultados acabam sendo similares à abordagem CPDST.

Os melhores desempenhos foram observados para os métodos BSIM e BREL. Este resultado era esperado, visto que tais métodos não se preocupam em detectar uma relação semântica específica entre pares de *tags*, portanto, a probabilidade de entregar uma lista com mais de 12 *tags* a partir de uma *query* é elevada, de um modo geral (de 84,5% a 99,7%). Por outro lado, os métodos não são especializados para identificar sinônimos ou hipônimos com precisão.

7.5.3 Experimento 3: Seleção de *queries* para Aplicação das Métricas

No Experimento 1, pode-se observar por meio do gráfico FDA a probabilidade de cada método não prover uma lista de *tags* a partir de uma *query* aleatória. Vale ressaltar que uma *query* que gera listas com 31 ou mais *tags* no método BREL não proporciona garantia de fornecer listas de tamanho igual nos demais métodos. Além disso, é possível que alguns métodos sejam capazes de prover até mesmo listas de *tags* nulas. Para a aplicação das métricas descritas na Seção 7.4.2, é imprescindível que uma *query* aleatória seja capaz de gerar listas de *tags* de tamanho k em todos os métodos envolvidos. A fim de entender como esse fenômeno se manifesta de maneira geral, idealizou-se um experimento para analisar quais *tags* conseguem gerar lista para todos os métodos e qual o tamanho mínimo das listas observado.

Basicamente, o experimento em questão consiste em variar iterativamente o tamanho da lista k, com $1 \leq k \leq 31$, e identificar a quantidade (e denominação) de *queries* que consegue fornecer listas de *tags* com tamanho maior ou igual a k. A Tabela 7.2 resume o resultado do experimento. À medida que k aumenta, diminui a quantidade de *tags* que satisfazem a condição de k mínimo. Como é possível observar, apenas com um $|T_q|$ mínimo igual a 1 atinge-se 25 *queries* na base de dados do BibSonomy e 93 *queries* na base de dados do Delicious aptas à aplicação das métricas. Entretanto, um $|T_q| = 1$ é insuficiente para analisar o desempenho dos métodos em cada métrica.

Com o resultado da Tabela 7.2, identificou-se que não é possível selecionar um conjunto de *queries* aleatoriamente para conduzir o experimento de avaliação de desempenho dos

Tabela 7.2: Simulação de *queries* que geram listas de *tags* para todos os métodos.

$\|T_q\|$ mínimo	Número de *tags* selecionadas	
	BibSonomy	Delicious
1	25	93
2	1	16
3	-	4
4	-	-

métodos para as métricas selecionadas (Experimento 4). Quanto mais métodos forem comparados simultaneamente, menor a possibilidade de se alcançar tamanhos de lista maior. Para viabilizar a execução do Experimento 4 relatado na Seção 7.5.4, considerou-se prudente dividir os métodos em 2 grupos e assim incrementar o $|T_q|$ mínimo: um grupo contendo CPDST, WNET e CNET (grupo de detecção de semântica específica) e outro grupo contendo CPDST e os métodos que não lidam com semântica específica BSIM e BREL. A abordagem CPDST aparece nos dois grupos porque pretende-se observar seu desempenho em comparação aos demais métodos envolvidos.

7.5.4 Experimento 4: Análise de Desempenho para as Métricas

As Figuras 7.7 e 7.8 exibem os resultados dos diferentes algoritmos de seleção de *tags* em relação às métricas relevância e *overlap* e auxiliam na obtenção da resposta para a QP3: *É possível melhorar a relevância de uma lista de tags relacionadas utilizando tags fornecidas pela abordagem CPDST?* Os resultados apresentados correspondem à média de valores observados ao longo de diferentes *queries*. Os métodos CPDST, WordNet e ConceptNet são exibidos em sua versão completa de deteção de relações semânticas, ou seja, sinonímia e relações de subordinação.

Como observa-se, a abordagem CPDST apresenta desempenho superior para a métrica relevância na base de dados do BibSonomy. Por exemplo, na abordagem CPDST as medições iniciam em 0,224 para *top-k* = 7 e terminam em 0,635 para *top-k* = 9, enquanto que o segundo colocado (BSIM) apresenta relevância mínima de 0,158 para *top-k* = 9 e relevância máxima de 0,469 para *top-k* = 10. BSIM e BREL apresentam desempenho mais próximos da abordagem CPDST. Em contrapartida, as listas de *tags* providas por WNET e

CNET exibem baixa relevância. Isso se explica pelo fato de que WNET e CNET recuperam sinônimos e hipônimos diretamente de suas respectivas bases de conhecimento, sem levar em consideração a observação de coexistirem com a *query* ao longo dos *posts* introduzidos pelos diferentes usuários do sistema. Na base de dados do Delicious, BREL predomina como melhor método segundo a métrica relevância ($0,082 \leq$ relevância $\leq 0,291$), seguido pela abordagem CPDST ($0,087 \leq$ relevância $\leq 0,150$). Novamente, WNET e CNET apresentaram listas de *tags* com menor média de relevância.

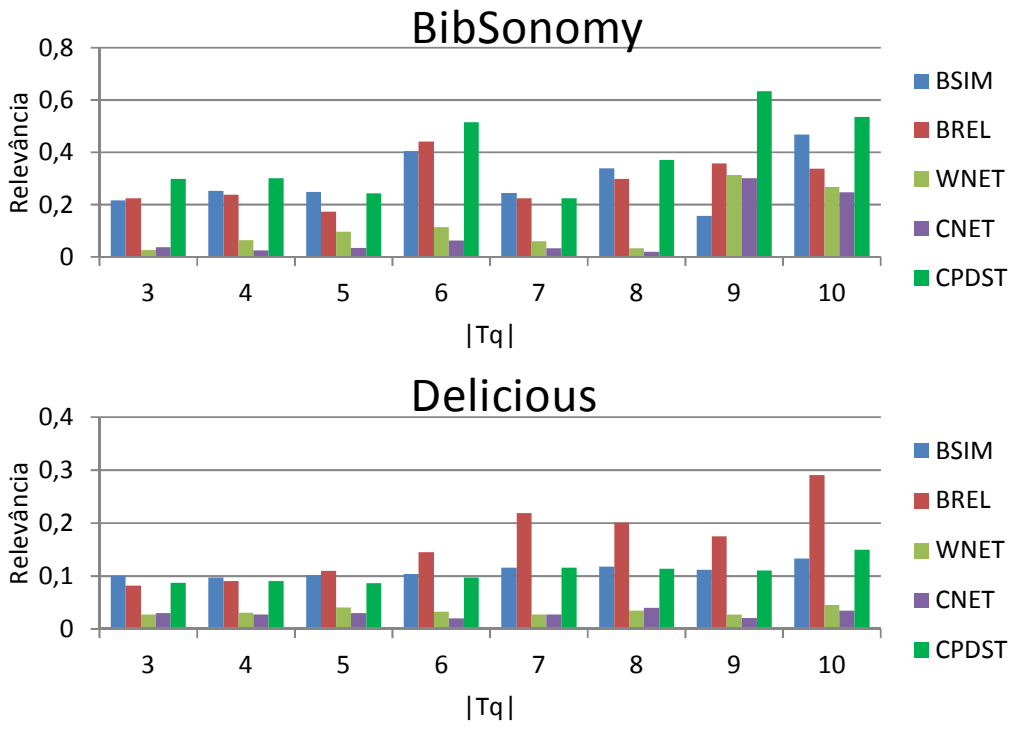

Figura 7.7: Média de relevância.

Em termos de *overlap* (Figura 7.8), os maiores valores observados na base de dados do BibSonomy foram para a abordagem CPDST e o método BREL. Entretanto, a média ao longo do eixo *top-k* demonstra que o nível de *overlap* se encontra em torno de 10%, sugerindo que a escolha de uma *tag* aleatória na lista acrescenta acesso a um conjunto de recursos que é 90% distintivo em relação aos recursos anotados pela *query*. Com o incremento do tamanho da lista, o gráfico sugere que o *overlap* aumenta para os métodos CPDST, BREL e BSIM, enquanto que não exerce influência para os métodos WNET e CNET.

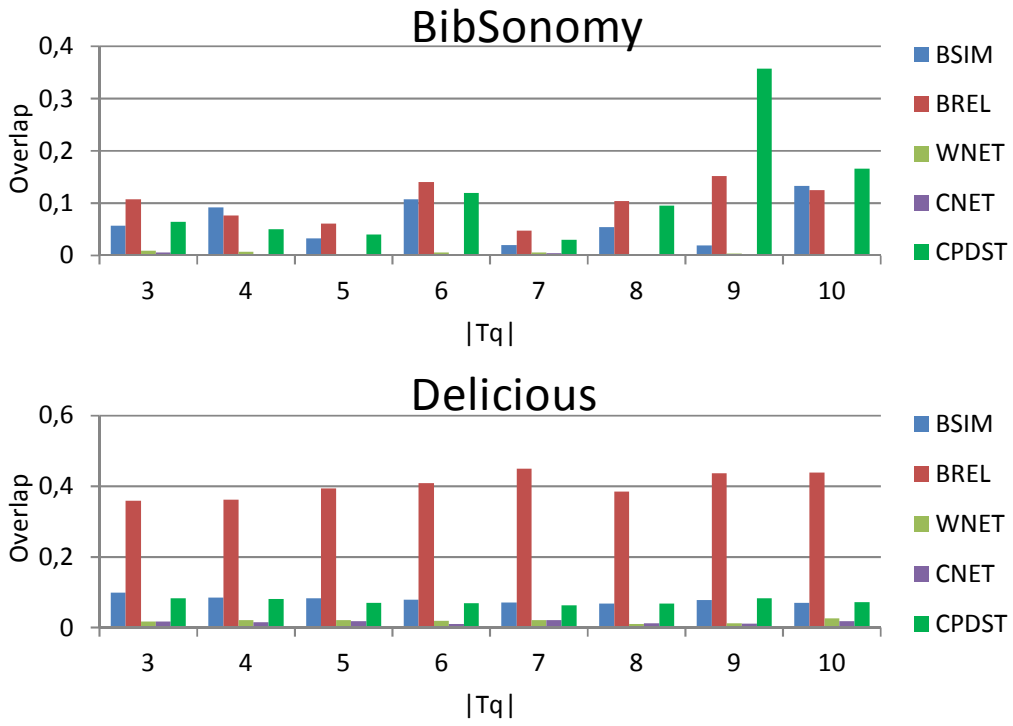

Figura 7.8: Média de *overlap*.

Em contraste, WNET e CNET apresentam os mais baixos níveis de *overlap* tanto na base de dados do BibSonomy quanto na do Delicious. Ao selecionar suas *tags*, o ranqueamento que define as *top-k tags* no WNET e CNET utiliza critérios particulares que independem da proeminência das *tags* na folksonomia e de sua ligação direta com a *query*. Como resultado, os sinônimos e hipônimos sugeridos pela lista não são empregados frequentemente entre si pelos usuários no processo de anotação. Para exemplificar, interprete como exemplo a *query* `photo`. A lista provida pela abordagem CPDST (*pics, photographs, photos*) na base de dados do BibSonomy obteve *overlap=0,0377*, enquanto que a lista provida por WNET (*print, exposure, frame*) apresentou *overlap=0,0000*.

Quanto aos resultados na base de dados do Delicious, BREL é absoluto como método de maior *overlap*. Com uma média de *overlap* em torno de 40%, BREL apresenta uma larga vantagem em relação ao segundo melhor método BSIM, que obteve uma média de 8%. Isso sugere que os métodos fundamentados em coocorrência são propensos a sobressairem no Delicious, de acordo com as métricas relevância e *overlap*.

Como ressaltado na Seção 7.4.2, as métricas relevância e *overlap* são tendenciosas a salientar coocorrência. Os resultados mostram que, indiretamente, tais métricas beneficiam

os métodos do BibSonomy (BREL e BSIM) e desfavorecem os métodos WNET e CNET, pois ambas dão ênfase à exploração de recursos em comum entre *tags* ao invés de estabelecer critérios mais independentes para a mensuração qualitativa das *tags* da lista no que diz respeito ao aspecto semântico. A abordagem CPDST aparenta não ser fortemente desfavorecida pelas métricas pois, em algumas situações, é possível observar medições mais próximas àquelas alcançadas pelos métodos BSIM e BREL. Na ausência de métricas que possibilitem extrair exatamente a informação desejada, entende-se que o julgamento qualitativo das listas de *tags* relacionadas providas pelos métodos comparados, sob o ponto de vista do usuário, é uma providência pertinente para poder obter conclusões complementares.

7.5.5 Análise de *Boxplots*

As figuras apresentadas neste experimento ilustram os resultados das métricas em formato de *box/whisker plots*. O eixo das ordenadas denota o intervalo de valores alcançado. Uma vez que a análise de *boxplots* é efetuada por $|T_q|$, este variando de 3 a 10, por questão de espaço decidiu-se apresentar apenas os 3 melhores *boxplots* de $|T_q|$ que reúnem o maior número de *queries* para os métodos comparados.

A Figura 7.9 mostra os *boxplots* para a métrica relevância usando os dados do BibSonomy, na comparação dois a dois da abordagem CPDST com os demais métodos. Pode-se observar que a abordagem CPDST predomina como método que seleciona *tags* mais relevantes em relação à *query* inicial. A Figura 7.10 mostra os *boxplots* para a métrica relevância usando os dados do Delicious, na qual a abordagem CPDST é comparada primeiramente com BSIM e BREL, e depois com WNET e CNET. Visualmente, os gráficos evidenciam uma similaridade entre CPDST, BSIM e BREL, enquanto que em relação aos métodos WNET e CNET a abordagem CPDST é discretamente superior.

Outras observações interessantes incluem:

- A abordagem CPDST demonstra superioridade em relevância quando comparada principalmente com WNET e CNET. Embora as *tags* providas por WNET e CNET tenham a devida relação semântica assegurada por um *thesaurus*, não se pode garantir que estas *tags* estejam alinhadas com o contexto de recursos que caracteriza a *query*,

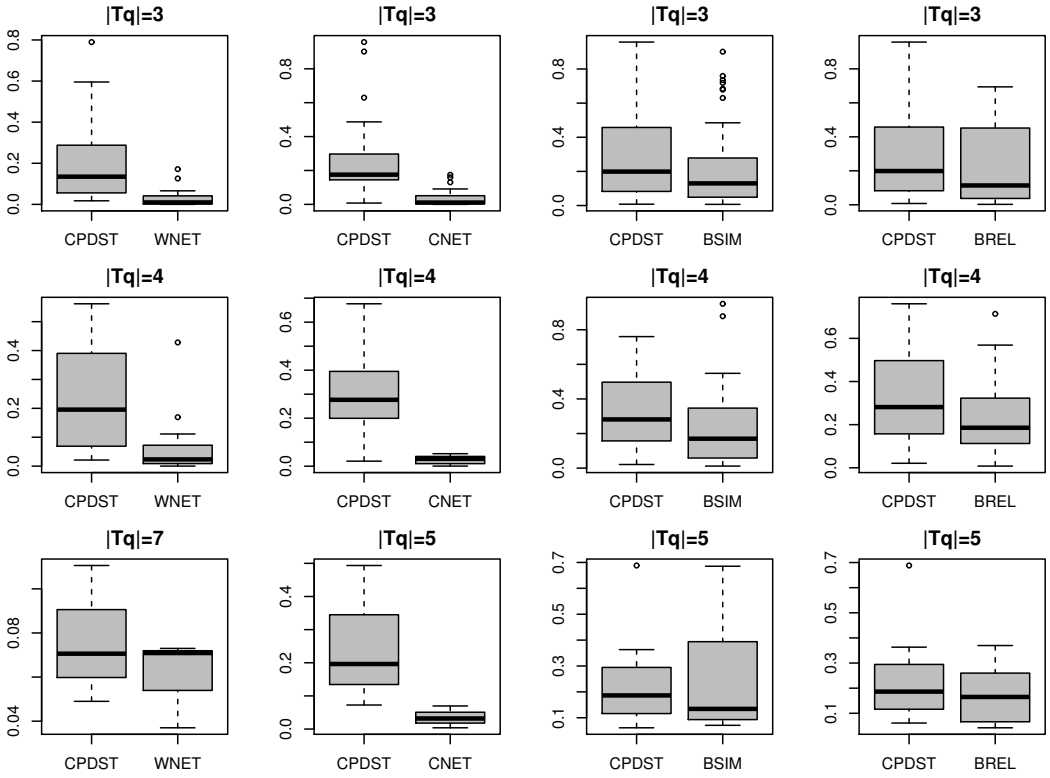

Figura 7.9: *Boxplots* para a métrica relevância na base de dados do BibSonomy.

pois as *tags* são selecionadas de acordo com a aplicação de medidas de distância semântica entre termos, as quais são dependentes do método;

- Os métodos CNET e WNET apresentam resultados semelhantes, com uma estreita vantagem para WNET. Uma possível justificativa para essa diferença pode ser atribuída ao peso da relação semântica indicada pelo CNET entre a *query* e o termo recuperado, os quais normalmente se repetem e estão próximos do limite inferior (1.0). Deste modo, termos menos úteis à consulta são mais prováveis de serem recuperados devido à tendência de haver sucessivos empates no momento da ordenação, e isto pode priorizar *tags* menos expressivas para o problema;

- Devido à maior probabilidade de BSIM e BREL em produzirem listas de *tags*, um maior número de *queries* em comum com CPDST foi utilizado para produzir os *boxplots* e permitir uma maior variedade de $|T_q|$. De um total de 100 *queries* resultantes na base de dados do BibSonomy, verificou-se que a abordagem CPDST apresentou melhor relevância em 60, BSIM em 28 e BREL em 12 *queries*. Entretanto,

BSIM apresenta intervalo de valores mais próximo da abordagem CPDST. Na base de dados do Delicious, para um total de 793 *queries*, a abordagem CPDST apresentou melhor relevância em 164, BSIM em 310 e BREL em 312 *queries*;

- BREL se mostrou menos efetivo em relação a CPDST e BSIM quanto à métrica relevância ao observar os *boxplots* apresentados para as duas bases de dados. A própria natureza do método BREL seleciona *tags* com base na ordem decrescente de frequência de coocorrência, assim, termos mais genéricos (de maior amplitude) adquirem natural precedência. Deste modo, as *tags* selecionadas tendem a ser menos específicas e, sucessivamente, reduz-se a interseção entre os recursos anotados pela *query* e a *tag* selecionada.

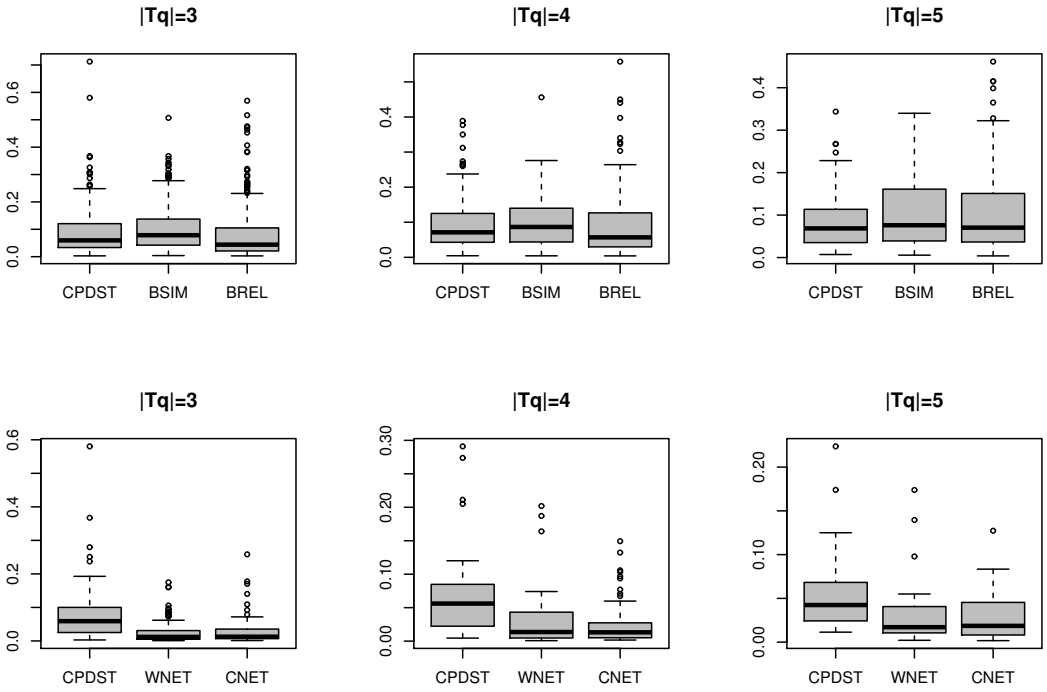

Figura 7.10: *Boxplots* para a métrica relevância na base de dados do Delicious.

As Figuras 7.11 e 7.12 exibem os *boxplots* para a métrica *overlap*. Os métodos WNET e CNET produzem o menor *overlap* dentre os métodos comparados. Essa constatação pode ser explicada da seguinte maneira: uma vez que tais métodos selecionam seus termos externamente à folksonomia, pode acontecer que as *tags* da lista não anotem recursos em comum, minimizando dessa forma o nível de *overlap*. Os métodos CPDST e BSIM

produzem similar *overlap*, com maior superioridade observada para o método BREL. O método BSIM produz menor variabilidade de *overlap* em relação ao grupo CPDST, BSIM e BREL.

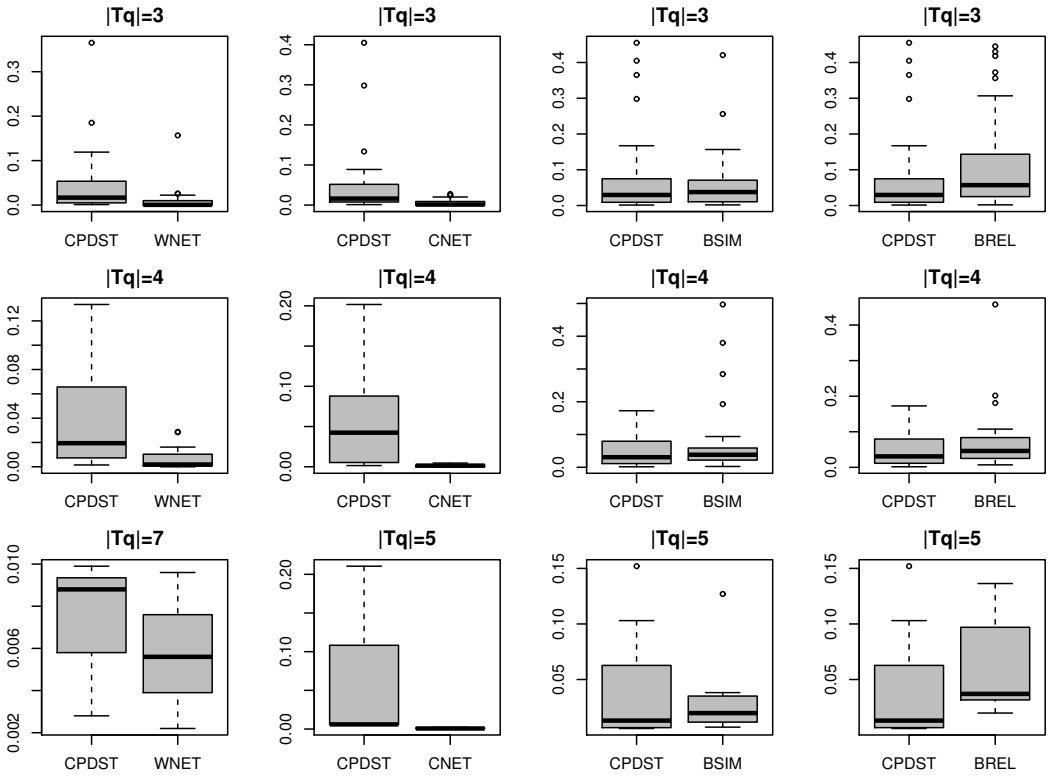

Figura 7.11: *Boxplots* para a métrica *overlap* na base de dados do BibSonomy.

Devido à natureza restritiva dos métodos específicos para detecção de sinonímia ou hiponímia, observa-se que para os métodos CPDST-S, WSyn, CSyn, CPDST-H, WHyp e CHyp predomina um $|T_q| = 3$. Isso se deve à pequena quantidade de *queries* que conseguem produzir lista de *tags* simultaneamente nos métodos comparados, o que restringe a faixa de $|T_q|$ disponível.

De acordo com a Figura 7.13, observa-se a mesma tendência de superioridade em favor das variantes da abordagem CPDST segundo a métrica relevância. Em particular, CPDST-H demonstra ser mais efetivo do que os métodos WHyp e CHyp nas duas bases de dados usadas nos experimentos. Em termos de detecção de sinonímia, CPDST-S e WSyn aparentam um desempenho similar quando os *boxplots* são observados em relação à base de dados do Delicious.

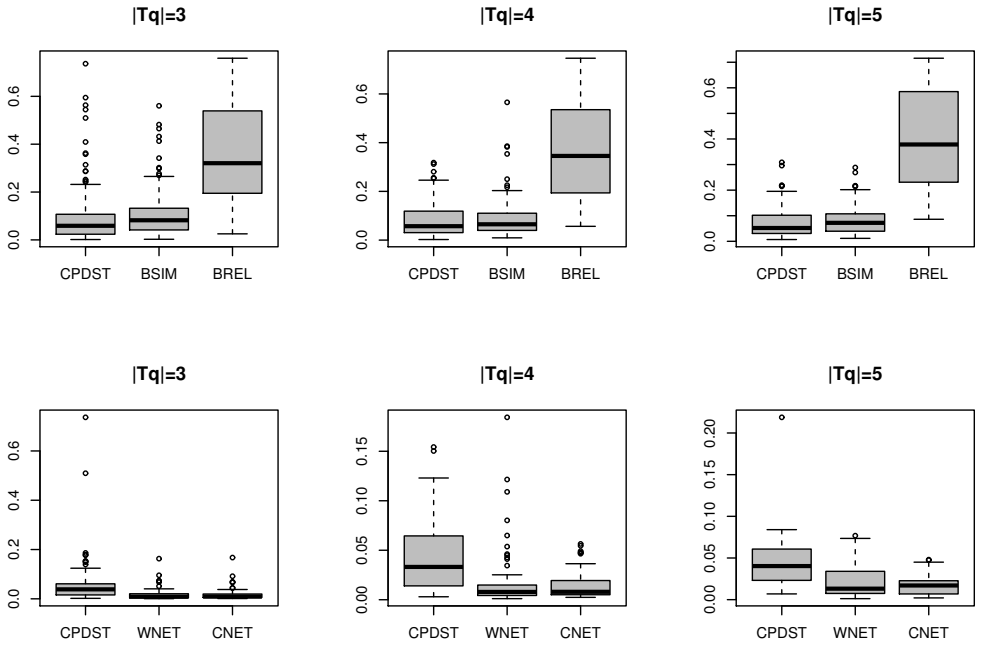

Figura 7.12: *Boxplots* para a métrica *overlap* na base de dados do Delicious.

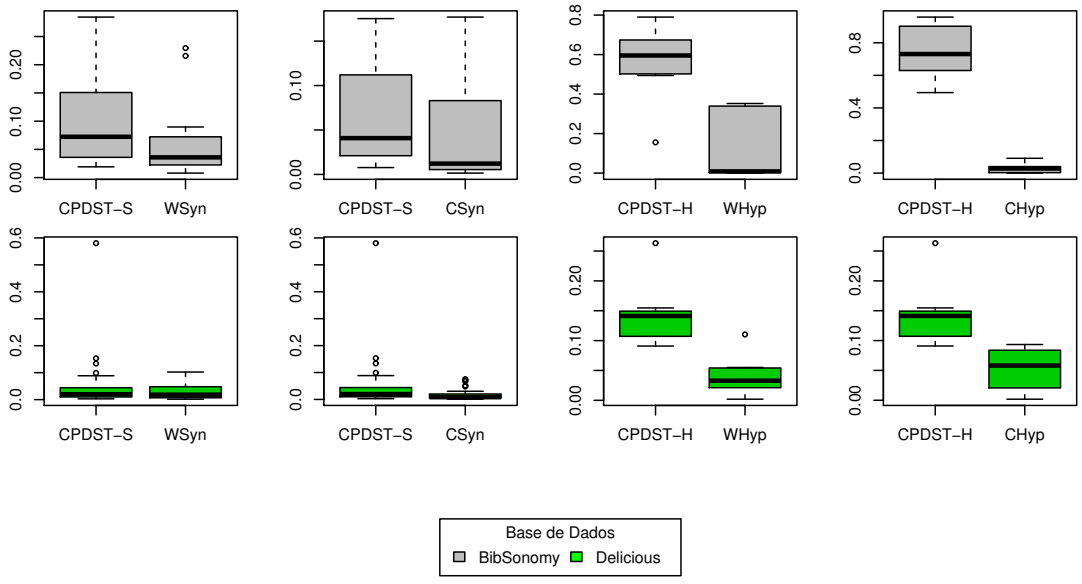

Figura 7.13: *Boxplots*: métrica relevância para sinonímia e hiponímia.

7.5.6 Teste de Significância

A fim de determinar se a diferença na média de relevância entre os métodos comparados é estatisticamente significante, executou-se um teste estatístico de comparação de alternativas. Para comparar a abordagem CPDST com WNET e CNET usando os dados do BibSonomy, foi necessário isolar CPDST com WNET e depois CPDST com CNET. Reforçando o que foi dito previamente, isso foi feito porque as *queries* em comum entre CPDST e WNET não são aproveitadas em sua maioria para executar a métrica relevância com os métodos CPDST e CNET. Por isso, justifica-se a necessidade de efetuar a comparação separadamente. Entretanto, usando os dados do Delicious foi possível comparar simultaneamente a abordagem CPDST com os métodos WNET e CNET.

Foram executados o teste paramétrico teste-t de *student* pareado de duas caldas, o teste não-paramétrico *Wilcoxon Signed-Rank Test* e o método *One-Way* ANOVA (*ANalisys Of VAriance*) para comparação múltipla de médias, com 95% de confiança. Logo, um *p-valor < 0,05* rejeita a hipótese nula de que os métodos comparados possuem a mesma relevância em termos de média.

A Tabela 7.3 resume o resultado dos testes para o $|T_q| = 3$, tamanho de lista que reúne o maior número de *queries* entre a abordagem CPDST e os demais métodos. No teste de comparação múltipla utilizando ANOVA, o *p-valor < 0,05* indica que a média dos três métodos comparados não são iguais estatisticamente. Logo, quando esta condição é satisfeita, a coluna *Resultado* prontamente exibe o método que é estatisticamente superior[34].

Pode-se observar que a abordagem CPDST (e suas variantes) é estatisticamente significante superior aos métodos correspondentes providos por WNET e CNET, exceto na base de dados do Delicious para as variantes focadas na descoberta de sinonímia. Na comparação entre os métodos CPDST, BSIM e BREL (BibSonomy), o teste de Análise de Variância aceita a hipótese nula de que as médias dos métodos podem ser consideradas estatisticamente iguais. Usando os dados do Delicious, o teste estatístico apontou que BSIM é estatisticamente significante superior.

[34]O desfecho da análise estatística identifica quais pares de médias apresentam diferença estatística no nível de significância adotado e, entre os pares de média, é aplicado o teste-t ou *wilcoxon* para identificar qual método possui média de relevância estatisticamente significante superior.

Tabela 7.3: Teste de Significância.

Comparativo	# queries	Teste Utilizado	p-valor	Resultado
BibSonomy				
CPDST vs. WNET	32	Wilcox	p-valor = 1.178e-07 < 0,05	CPDST é superior
CPDST-S vs. WSyn	13	Wilcox	p-valor = 0,02493 < 0,05	CPDST-S é superior
CPDST-H vs. WNET-H	7	Wilcox	p-valor = 0,03125 < 0,05	CPDST-H é superior
CPDST vs. CNET	27	Wilcox	p-valor = 1,49e-08 < 0,05	CPDST é superior
CPDST-S vs. CNET-S	8	Wilcox	p-valor = 0,02344 < 0,05	CPDST-S é superior
CPDST-H vs. CNET-H	5	teste t	p-valor = 0,0006898 < 0,05	CPDST-H é superior
CPDST vs. BSIM vs BREL	54	ANOVA	p-valor = 0,148 > 0,05	Métodos semelhantes
Delicious				
CPDST vs. WNET vs CNET	90	ANOVA	p-valor = 5,902e-29 < 0,05	CPDST é superior
CPDST-S vs. WSyn vs CSyn	34	ANOVA	p-valor = 0,0709 > 0,05	Métodos semelhantes
CPDST-H vs. WHyp vs CHyp	8	ANOVA	p-valor = 0,0001 < 0,05	CPDST-H é superior
CPDST vs. BSIM vs BREL	312	ANOVA	p-valor = 0,0256 < 0,05	BSIM é superior

7.6 Considerações Finais

Neste capítulo, foi introduzido um cenário de aplicação da abordagem CPDST para a tarefa de seleção de listas de *tags* semanticamente relacionadas. Foram propostos dois métodos fundamentados no acesso aos dicionários WordNet e ConceptNet e delineadas as contribuições que o presente estudo adiciona à literatura. Conduziu-se um extenso conjunto de experimentos utilizando as bases de dados do BibSonomy e Delicious com o intuito de extrair diferentes percepções acerca do problema abordado. Também apresentou-se uma análise comparativa entre os métodos propostos (CPDST, CPDST-S, CPDST-H, WNET, WSyn, WHyp, CNET, CSyn e CHyp) e os métodos providos pelo BibSonomy (BSIM e BREL) utilizando as métricas relevância e *overlap* provenientes do contexto de avaliação de nuvens de *tags*. Porém, ressaltou-se que a aplicação de tais métricas são tendenciosas a favorecer os métodos que capturam coocorrência.

Os métodos configurados para tratamento de semântica específica (sinonímia ou hiponímia) mostraram ser mais limitados quanto à capacidade de gerar listas de *tags*, uma vez que a probabilidade de gerar uma lista é bem inferior se comparado aos métodos BSIM e BREL. Por outro lado, BSIM e BREL não são claros quanto ao que está sendo recuperado, podendo sugerir *tags* que expressam qualquer tipo de relação semântica com a *query*. Trata-

se de um *trade-off* que deve ser analisado em questões de projeto para decidir qual dos métodos é mais interessante para um problema em particular.

Uma vez que os métodos CPDST-S e CPDST-H são mais restritivos, assume-se que a união da captura de ambas as semânticas melhora a capacidade da abordagem CPDST de fornecer listas de *tags*. Além disso, alia-se o fato de que a união tanto de *tags* sinônimas quanto de hipônimas podem coexistir na mesma lista de *tags* e manterem-se compatibilizadas com o sentido denotado pela *query*. No próximo capítulo, apresenta-se as conclusões finais sobre o trabalho realizado nesta tese de doutorado assim como as atividades idealizadas para trabalhos futuros.

Capítulo 8

Conclusões e Trabalhos Futuros

Apresentou-se neste trabalho uma nova abordagem para a detecção de relações semânticas entre *tags* em folksonomias, denotada por CPDST. A abordagem empenha-se em aprender relações semânticas do tipo sinonímia e de subordinação entre *tags* diretamente dos dados de uma folksonomia, usando um modelo genérico de classificação em pares e medidas de similaridade *tag-tag* introduzidas nos trabalhos relacionados como atributos de aprendizado. O modelo de classificação idealizado pela abordagem CPDST independe de fonte de dados e pode ser instanciado para detecção de outros tipos de relações semânticas.

A concepção da proposta é multidisciplinar e envolve a necessidade de conhecimentos básicos sobre diferentes temas, tais como PLN, medidas de similaridade/distância, Mineração de Dados e balanceamento de classes. Sob a perspectiva de um problema de AM, a formulação do problema de detecção de relações semânticas apresenta complexidade intrínseca, uma vez que o severo desbalanceamento de classes é evidente (menor número de instâncias em que se observa a existência de uma relação semântica específica) e a sobreposição de classes acentua a dificuldade em discriminar os exemplos positivos dos negativos. Para lidar com estes problemas, foram empregadas diferentes técnicas de re-amostragem as quais mostraram ser adequadas para tratar ambas as questões, permitindo aos classificadores utilizados alcançarem bons resultados na acurácia da predição da classe positiva.

Recorda-se no Capítulo 1 as principais questões de pesquisa dirigidas neste estudo:

- *QP1: Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de sinonímia provê melhor acurácia na identificação de sinonímia*

entre tags de folksonomia?

- *QP2: Até que ponto usar a técnica de classificação ao invés de heurísticas indicadas para a detecção de relações hierárquicas provê melhor acurácia na identificação de relações hierárquicas (hiperonímia e hiponímia) entre tags de folksonomias?*
- *QP3: É possível melhorar a relevância de uma lista de tags relacionadas utilizando tags geradas a partir da abordagem CPDST?*

As respostas para essas questões foram obtidas por meio de um extenso conjunto de experimentos. Os resultados mostram que a abordagem CPDST é estatisticamente significante superior em relação ao melhor dos *baselines* denotado nas bases de dados do BibSonomy e Delicious. Na tarefa de predição de sinonímia (QP1), a abordagem CPDST apresentou notável desempenho atingindo acurácia de classificação em 93% das instâncias positivas, resultado este que representa uma melhora de 86,13% em relação à acurácia determinada por distância de edição (BibSonomy) e 63,14% em relação a $\cos^{tag-res}$ (Delicious). Na tarefa de detecção de relações de subordinação (QP2), a abordagem CPDST teve um desempenho mais modesto, atingindo melhoria na ordem de 10,3% na base de dados do Bibsonomy e 30,19% na base de dados do Delicious.

No decorrer do trabalho, percebeu-se que identificar relações de subordinação com a abordagem CPDST se tornou um grande desafio, tendo em vista a complexidade natural de convencionar a referida relação semântica e o agravamento dos problemas de desbalanceamento de classes aliado ao *overlapping*, o que dificultou o aprendizado da classe positiva. Várias combinações de técnicas de balanceamento e algoritmos de aprendizagem foram testadas para encontrar a configuração que proporciona o melhor ganho no aprendizado das classes. Os trabalhos relacionados usualmente focam na construção de estruturas hierárquicas entre *tags*, empregando processos lógicos complexos que conduzem à identificação de prováveis relações de subordinação. Embora as conexões hierárquicas estabelecidas entre pares de *tags* tenham sentido, uma análise exploratória em hierarquias providas por diferentes trabalhos evidenciou que a maioria não é assegurada por um dicionário linguístico. Uma vez que a abordagem CPDST reúne medidas de similaridade introduzidas nestes trabalhos, eventualmente esta observância tende a ser refletida na tarefa de predição e ocasione influência no incremento da taxa de falsos positivos e negativos.

Foram propostos novos métodos voltados à tarefa de criação de listas de *tags* relacionadas semanticamente, condicionadas a uma chave de busca. Introduziu-se a abordagem CPDST e dois métodos fundamentados no acesso aos dicionários WordNet e ConceptNet para fornecer listas de *tags* constituídas especialmente por indicações de *tags* sinônimas e hipônimas. Diante da especificidade do relacionamento semântico, observou-se que os métodos propostos são mais limitados quanto à capacidade de prover lista de *tags* para uma *query* aleatória, o que é compreensível tendo em vista que as opções BSIM e BREL providas pelo BibSonomy estimam um grau de semântica entre *tags* que pode agregar qualquer tipo de relacionamento semântico.

A avaliação comparativa dos diferentes algoritmos de geração de listas de *tags* mostrou que, na base de dados do BibSonomy, a abordagem CPDST é mais relevante do que os métodos BSIM e BREL no enfrentamento por *query*, ou seja, CPDST proporciona listas de *tags* mais relevantes para mais de 50% das *queries* analisadas individualmente. Em contrapartida, o teste de significância apontou que a diferença não é significativa para alguns tamanhos de lista como, por exemplo, $|T_q| = 3$ (QP3). Em termos de *overlap*, os resultados revelaram que a escolha de uma *tag* aleatória na lista provida pela abordagem CPDST concede acesso a recursos que em média são 90% distintivos em relação às demais *tags* que constituem a lista.

Na base de dados do Delicious, BREL e SIM sobressaem-se como os métodos que apresentam listas de *tags* mais relevantes em, respectivamente, 40% e 39% das *queries*. Entretanto, considerando $|T_q| = 3$, um teste de comparação múltipla entre BSIM, BREL e CPDST mostrou que BSIM é estatisticamente significante superior. Os resultados para a métrica *overlap* indicaram que o método BREL apresenta *overlap* médio de 40% quando todos os $|T_q|$ são considerados, sugerindo que a escolha aleatória de uma *tag* da lista concede acesso a um grupo de recursos que é 60% distintivo em relação às demais *tags* presentes na lista.

Independentemente da base de dados utilizada nos experimentos, é visível a disparidade da métrica relevância (e estatisticamente significante) em favor da abordagem CPDST com relação aos métodos WNET e CNET para todos os $|T_q|$ analisados (QP3), tendo em vista que tais métodos sugerem *tags* sem atentar para sua coexistência com a *query* nos *posts* introduzidos pelos usuários. Deste modo,

Em linhas gerais, esta tese acrescenta à literatura contribuições direcionadas ao problema de detecção de relações semânticas entre *tags* (cf. Seção 1.3) e novos métodos para geração de listas de *tags*, os quais foram avaliados quantitativamente utilizando métricas existentes na literatura (cf. Seção 7.1).

Apesar de comprovar-se as virtudes da abordagem CPDST nos diversos experimentos realizados, aponta-se alguns aspectos que vão contribuir com a evolução da pesquisa em trabalhos futuros, os quais são discutidos a seguir:

- Visando aprimorar o processo de aprendizagem, principalmente para a detecção de relações de subordinação, pode-se investigar atributos adicionais mais discriminativos. A seleção dos atributos de treinamento é um dos fatores que influenciam no desempenho do modelo de classificação em dados desbalanceados. Embora constate-se a dificuldade em encontrar atributos que capturem as propriedades semânticas instituídas pela relação de subordinação, a literatura deve ser continuamente revisada a fim de extrair conhecimento que conduza à idealização de novos atributos;

- Muitos estudos relatam que, para certos conjuntos de dados desbalanceados, o grau de desbalanceamento na distribuição de classes não é o único princípio que deteriora o desempenho da tarefa de classificação. Outros elementos tais como *overlapping*, presença de instâncias ruidosas e vulnerabilidade na representatividade dos dados intensificam a complexidade da base de dados de treinamento. Os problemas de *overlapping* e ruídos residem nas bases de dados usadas neste trabalho, portanto não basta simplesmente tratar a questão do desbalanceamento. Deste modo, técnicas adicionais em nível de dados ou em nível de algoritmo podem ser investigadas para abordar ambos os problemas de *overlapping* e desbalanceamento, visto que esses fenômenos estão correlacionados. Um encaminhamento inicial a considerar seria: (a) aplicar a técnica OSS (cf. Seção 2.6.4) e avaliar experimentalmente o impacto de sua proposta no cenário do problema; e (b) investigar esquemas particulares de modelagem de base de dados com incidência de *overlapping* introduzidos em (XIONG; WU; LIU, 2010);

- Na Seção 4.4, foram relatados os esforços empregados para estender o modelo geral de classificação binária em um problema de classificação multiclasse para identificar

sinônimos e relações de subordinação entre *tags*. Diante das dificuldades que esta reformulação demanda para alinhar-se ao cenário de experimentação, principalmente no que diz respeito à presença de fenômenos que afetam o aprendizado das classes, entendeu-se que alocar tempo de estudo para explorar a complexidade inerente ao problema inviabilizaria a conclusão das atividades de pesquisa em andamento. Deste modo, um estudo focalizado pode ser conduzido para explorar esse desafio (classificação multiclasse em dados desbalanceados e com presença de *overlapping*), uma vez que se trata de um campo de estudo em aberto;

- O modelo de predição CPDST para detecção de sinonímia e relação de subordinação demonstrou que não é liberal para classificar instâncias como positivas. Por isso, justifica-se a pequena quantidade de *tags* relacionadas em relação a uma determinada *query*, independentemente dos falsos positivos revelados na classificação. Em muitos casos, nem sequer é possível recuperar *tags* relacionadas. Embora esta limitação seja consequência da especificidade da relação semântica tratada pela abordagem CPDST, esta limitação pode ser amenizada com a conversão de exemplos negativos em positivos, os quais não foram percebidos no período de execução da rotulação automática. Duas maneiras de incrementar o número de exemplos positivos consistem em: (a) submeter os pares de *tags* negativos para confirmação no ConceptNet (ação temporal), uma vez que o ConceptNet incorpora definições vindas de várias fontes de dados e esta atualização é feita com maior frequência; e (b) utilizar a plataforma *crowdsourcing* (YIN et al., 2014) para conduzir um experimento de julgamento humano direcionado à aprovação (ou não) de pares de *tags* candidatas a uma relação de sinonímia ou subordinação. Desta forma, a abordagem CPDST pode ser capaz de ampliar sua capacidade de detecção;

- Configurar o modelo de predição CPDST para detecção de outros tipos de relações semânticas como, por exemplo, meronímia (parte-de) e holonímia;

- Uma avaliação quantitativa dos métodos geradores de listas de *tags* foi realizada usando as métricas relevância e *overlap* em duas bases de dados de folksonomias reais: BibSonomy e Delicious. Os experimentos revelaram que a probabilidade de geração de lista de *tags* na abordagem CPDST é inferior aos métodos comparados, porém

esta probabilidade diminui em bases com maior volume de dados, como constatou-se com o Delicious em relação ao BibSonomy. Acredita-se que essa limitação pode ser reduzida se os experimentos forem reproduzidos novamente com um maior número de instâncias positivas utilizadas para treinamento;

- A relevância das listas de *tags* providas pelos métodos comparados foi mensurada por meio da aplicação de métricas *offline* em um experimento de análise quantitativa. Entretanto, considera-se que um experimento de avaliação subjetiva de satisfação do usuário pode ser conduzido futuramente para descobrir como os usuários percebem a relevância e utilidade das listas de *tags* providas pelos métodos propostos.

Como relatado previamente, existe a necessidade de conduzir experimentos de avaliação com o usuário para fins de julgamentos de relevância. Esta decisão visa obter uma percepção da qualidade das listas de *tags* sob o ponto de vista do usuário e admitir relações semânticas desconhecidas no processo de rotulação para aumentar o número de casos positivos. O uso dos dicionários eletrônicos exerceram influência na acurácia da predição da abordagem CPDST. Além disso, as métricas empregadas para avaliar quantitativamente as listas de *tags* foram tendenciosas aos métodos do BibSonomy. Por isso, considera-se importante realizar experimentos de avaliação com o usuário para enriquecer o trabalho realizado. Entretanto, a decisão de conduzir esse processo envolve a participação de avaliadores humanos, o que torna a tarefa onerosa e demorada. Esta atividade envolve desafios que merecem uma discussão mais detalhada.

Primeiramente, não se tem garantias de que o processo de julgamento humano seja sempre um sucesso devido à subjetividade do avaliador. Dependendo do esforço da tarefa, torna-se conveniente estabelecer o perfil do avaliador ideal para que a qualidade da avaliação não seja afetada. Caso a análise do resultado da avaliação aponte imperfeições, é necessário realizar ajustes na metodologia de avaliação (limiares, modificação dos algoritmos envolvidos, atualização das unidades de tarefa[35], entre outros) e iniciar outra sequência de julgamentos. Logo, é possível que haja múltiplas iterações de avaliação, o que denota uma tarefa não trivial para o problema particular tratado nesta tese.

O artifício da redundância das respostas pode ser aplicado para tratar incertezas nos

[35]Uma expressão utilizada para designar a realização de uma tarefa de julgamento pelo avaliador humano.

julgamentos, com o intuito de determinar a resposta correta (PONCIANO et al., 2014). Nesta estratégia, a mesma tarefa é replicada para vários avaliadores distintos. Assim, a partir de um conjunto de respostas redundantes aplica-se o conceito de voto majoritário, o qual de maneira simples considera como correta a resposta provida pela maioria dos avaliadores. Quanto maior o número de usuários, maior será a chance de definir a resposta correta, embora implique em maiores custos (financeiro, tempo, avaliação, entre outros) para realização do experimento.

Para examinar a qualidade dos julgamentos de relevância, aplica-se uma medida de concordância entre os avaliadores. Este procedimento é crucial para verificar se existe um nível mínimo de consistência entre os avaliadores. A literatura especializada aponta diversas formas de medição do nível de concordância entre avaliadores. Uma delas, o coeficiente *Kappa* de Cohen (VIERA; GARRETT, 2005), é um procedimento estatístico de concordância para itens qualitativos, usada para comparar a habilidade de diferentes avaliadores para classificar uma tarefa em uma das opções (rótulo) disponíveis. O valor de *kappa* varia de 0 a 1 e, dependendo da faixa de valores que esteja associado, pode ser interpretado como pobre, razoável, satisfatório e excelente.

Mesmo diante dos desafios citados, os experimentos com a participação do usuário especificados neste capítulo são encorajados pois os resultados podem acrescentar maiores contribuições ao trabalho realizado. A plataforma *crowdsourcing* tem emergido como uma viável solução para realizar diferentes tipos de avaliação humana. A principal razão por trás dessa tendência é que o *crowdsourcing* permite conduzir experimentos de forma rápida, com resultados normalmente satisfatórios a um baixo custo. Alonso (2013) destaca que, como em qualquer experimento, importantes características com respeito a detalhes de implementação[36] devem ser levadas em consideração para que se tenha um experimento *crowdsourcing* bem sucedido.

[36]Instruções claras sobre a tarefa, definição de diretrizes de interface com o usuário, utilização de métricas de concordância entre avaliadores, *feedback* dos avaliadores, controle de qualidade das respostas, entre outros.

Referências Bibliográficas

[Abbasi 2011]ABBASI, R. Query expansion in folksonomies. In: *Proceedings of the 5th international conference on Semantic and digital media technologies*. Berlin, Heidelberg: Springer-Verlag, 2011. (SAMT'10), p. 1–16. ISBN 978-3-642-23016-5. Disponível em: <http://dl.acm.org/citation.cfm?id=2032129.2032131>.

[Agrawal, Imielinski e Swami 1993]AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 22, n. 2, p. 207–216, jun. 1993. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/170036.170072>.

[Albert e Barabási 2002]ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, American Physical Society, v. 74, n. 1, p. 47–97, jan. 2002. Disponível em: <http://link.aps.org/doi/10.1103/RevModPhys.74.47>.

[Almoqhim, Millard e Shadbolt 2013]ALMOQHIM, F.; MILLARD, D.; SHADBOLT, N. An approach to building high-quality tag hierarchies from crowdsourced taxonomic tag pairs. In: JATOWT, A. et al. (Ed.). *Social Informatics*. Springer International Publishing, 2013, (Lecture Notes in Computer Science, v. 8238). p. 129–138. ISBN 978-3-319-03259-7. Disponível em: <http://dx.doi.org/10.1007/978-3-319-03260-3_12>.

[Alonso 2013]ALONSO, O. Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, Springer Netherlands, v. 16, n. 2, p. 101–120, 2013. ISSN 1386-4564. Disponível em: <http://dx.doi.org/10.1007/s10791-012-9204-1>.

[Alpaydin 2010]ALPAYDIN, E. *Introduction to Machine Learning 2nd Edition*. 2nd. ed. Massachusetts, USA: MIT Press, 2010. ISBN 978-0-262-01243-0.

[Aly 2005]ALY, M. Survey on multiclass classification methods. *Neural Networks*, p. 1–9, 2005.

[Angeletou, Sabou e Motta 2008]ANGELETOU, S.; SABOU, M.; MOTTA, E. Semantically enriching folksonomies with flor. In: *In Proceedings of 5th ESWC. Workshop: Collective Intelligence the Semantic Web*. [S.l.: s.n.], 2008. (ESWC'08).

[Angeletou et al. 2007]ANGELETOU, S. et al. Bridging the gap between folksonomies and the semantic web: An experience report. In: *Proceedings of 4th European Semantic Web Conference*. [S.l.: s.n.], 2007. (ESWC'07), p. 93.

[Azeredo 2008]AZEREDO, J. C. *Gramática Houaiss da Língua Portuguesa*. São Paulo, SP, Brasil: Publifolha, 2008.

[Batista 2003]BATISTA, G. E. *Pré-processamento de dados em Aprendizado de Máquina Supervisionado*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - ICMC-USP, Março 2003.

[Batista, Prati e Monard 2004]BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 6, p. 20–29, June 2004. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1007730.1007735>.

[Beall 2008]BEALL, J. The weaknesses of full-text searching. *Journal of Academic Librarianship*, Elsevier, Oxford, UK, UK, v. 34, n. 5, p. 95–145, 2008. ISSN 00991333.

[Bean e Green 2001]BEAN, A.; GREEN, R. *Relationships in the Organization of Knowledge*. Dordrecht, Nethersands: Kluwer Academic Publishers, 2001. ISBN 978-94-015-9696-1.

[Begelman, Keller e Smadja 2006]BEGELMAN, G.; KELLER, P.; SMADJA, F. Automated tag clustering: Improving search and exploration in the tag space. In: *Proceedings of the WWW Collaborative Web Tagging Workshop*. Edinburgh, Scotland: Citeseer, 2006.

[Bengio e Grandvalet 2004]BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. *J. Mach. Learn. Res.*, JMLR.org, v. 5, p. 1089–1105, dez. 2004. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1005332.1044695>.

[Benz 2007]BENZ, D. *Collaborative Ontology Learning*. Tese (Doutorado) — University of Freiburg, Freiburg, Germany, June 2007.

[Benz e Hotho 2007]BENZ, D.; HOTHO, A. Position paper: Ontology learning from folksonomies. In: HINNEBURG, A. (Ed.). *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*. Halle, Germany: Martin-Luther-Universität Halle-Wittenberg, 2007. p. 109–112. ISBN 978-3-86010-907-6. Disponível em: <http://www.kde.cs.uni-kassel.de/pub/pdf/benz2007position.pdf>.

[Benz et al. 2010]BENZ, D. et al. The social bookmark and publication management system bibsonomy. *The VLDB Journal*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 19, p. 849–875, December 2010. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/s00778-010-0208-4>.

[Bindelli et al. 2008]BINDELLI, S. et al. Improving search and navigation by combining ontologies and social tags. In: *Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems*. Berlin, Heidelberg: Springer-Verlag, 2008. (OTM '08), p. 76–85. ISBN 978-3-540-88874-1.

[Brizan e Tansel 2006]BRIZAN, D. G.; TANSEL, A. U. A survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, v. 6, n. 3, p. 41–50, 2006. Disponível em: <http://www.iima.org/CIIMA/8\%20CIIMA\%206-3\%2041-50\%20\%20Brizan.pdf>.

[Budanitsky e Hirst 2001]BUDANITSKY, A.; HIRST, G. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Reources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2001.

[Budanitsky e Hirst 2006]BUDANITSKY, A.; HIRST, G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, v. 32, p. 13–47, 2006.

[Cai et al. 2011]CAI, S. et al. Learning concept hierarchy from folksonomy. In: *Web Information Systems and Applications Conference (WISA), 2011 Eighth*. Chongqing, China: [s.n.], 2011. p. 47–51.

[Cattuto et al. 2008]CATTUTO, C. et al. Semantic grounding of tag relatedness in social bookmarking systems. In: *Proceedings of the 7th International Conference on The Semantic Web*. Berlin, Heidelberg: Springer-Verlag, 2008. (ISWC '08), p. 615–631. ISBN 978-3-540-88563-4. Disponível em: <http://dx.doi.org/10.1007/978-3-540-88564-1_39>.

[Chawla et al. 2002]CHAWLA, N. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Disponível em: <http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a.pdf>.

[Chawla 2005]CHAWLA, N. V. Data mining for imbalanced datasets: An overview. In: MAIMON, O.; LOKACH, L. (Ed.). *The Data Mining and Knowledge Discovery Handbook*. Springer US, 2005. cap. 40, p. 853–867. ISBN 978-0-387-24435-8. Disponível em: <http://www.springerlink.com/index/R824814907175608.pdf>.

[Clements, Vries e Reinders 2008]CLEMENTS, M.; VRIES, A. P. de; REINDERS, M. J. Detecting synonyms in social tagging systems to improve content retrieval. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008. (SIGIR '08), p. 739–740. ISBN 978-1-60558-164-4. Disponível em: <http://doi.acm.org/10.1145/1390334.1390479>.

[Cohen 1995]COHEN, W. W. Fast effective rule induction. In: *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*. [S.l.: s.n.], 1995. p. 115–123.

[Damerau 1964]DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Commun. ACM*, ACM, New York, NY, USA, v. 7, p. 171–176, March 1964. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/363958.363994>.

[Dattolo, Eynard e Mazzola 2011]DATTOLO, A.; EYNARD, D.; MAZZOLA, L. An integrated approach to discover tag semantics. In: *Proceedings of the 2011 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2011. (SAC '11), p. 814–820. ISBN 978-1-4503-0113-8. Disponível em: <http://doi.acm.org/10.1145/1982185.1982359>.

[Deng et al. 2010]DENG, J. et al. What does classifying more than 10,000 image categories tell us? In: DANIILIDIS, K.; MARAGOS, P.; PARAGIOS, N. (Ed.). *Computer Vision - ECCV 2010*. [S.l.]: Springer Berlin Heidelberg, 2010. p. 71–84. ISBN 978-3-642-15554-3.

[Domingos 2012]DOMINGOS, P. A few useful things to know about machine learning. *Commun. ACM*, ACM, New York, NY, USA, v. 55, n. 10, p. 78–87, out. 2012. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2347736.2347755>.

[Drummond e Holte 2005]DRUMMOND, C.; HOLTE, R. C. Severe Class Imbalance: Why Better Algorithms aren't the Answer. In: *Proceedings of the 16th European Conference of Machine Learning*. Porto, Portugal: Springer, 2005. p. 539–546.

[Eichelberger e Sheng 2013]EICHELBERGER, R. K.; SHENG, V. S. Does One-Against-All or One-Against-One Improve the Performance of Multiclass Classifications? In: *Association for the Advancement of Artificial Inteligence*. Bellevue, Washington, USA: AAAI Press, 2013. p. 1609–1610.

[Elkan 2001]ELKAN, C. The foundations of cost-sensitive learning. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (IJCAI'01), p. 973–978. ISBN 1-55860-812-5, 978-1-558-60812-2. Disponível em: <http://dl.acm.org/citation.cfm?id=1642194.1642224>.

[Eynard, Mazzola e Dattolo 2013]EYNARD, D.; MAZZOLA, L.; DATTOLO, A. Exploiting tag similarities to discover synonyms and homonyms in folksonomies. *Software: Practice and Experience*, v. 43, n. 12, p. 1437–1457, 2013. ISSN 1097-024X. Disponível em: <http://dx.doi.org/10.1002/spe.2150>.

[Fawcett e Provost 1996]FAWCETT, T.; PROVOST, F. Combining data mining and machine learning for effective user profiling. In: *The 2nd International Conference on Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. p. 8–13.

[Fayyad, Piatetsky-Shapiro e Smyth 1996]FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. In: FAYYAD, U. M. et al. (Ed.). *Advances in knowledge discovery and data mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. p. 1–34. ISBN 0-262-56097-6. Disponível em: <http://dl.acm.org/citation.cfm?id=257938.257942>.

[Fernández, Jesus e Herrera 2010]FERNÁNDEZ, A.; JESUS, M. del; HERRERA, F. Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning. In: HüLLERMEIER, E.; KRUSE, R.; HOFFMANN, F. (Ed.). *Computational Intelligence for Knowledge-Based Systems Design*. Dortmund, Germany: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6178). p. 89–98. ISBN 978-3-642-14048-8. Disponível em: <http://dx.doi.org/10.1007/978-3-642-14049-5_10>.

[Fernández et al. 2013]FERNÁNDEZ, A. et al. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, v. 42, n. 0, p. 97 – 110, 2013. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705113000300>.

[Fernández-Delgado et al. 2014]FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn.*

Res., JMLR.org, v. 15, n. 1, p. 3133–3181, jan. 2014. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=2627435.2697065>.

[Gan, Ma e Wu 2007]GAN, G.; MA, C.; WU, J. *Data clustering - theory, algorithms, and applications.* [S.l.]: SIAM - Society for Industrial and Applied Mathematics, 2007. I-XXII, 1-466 p.

[Gemmell et al. 2009]GEMMELL, J. et al. The impact of ambiguity and redundancy on tag recommendation in folksonomies. In: *Proceedings of the third ACM conference on Recommender systems*. New York, NY, USA: ACM, 2009. (RecSys '09), p. 45–52. ISBN 978-1-60558-435-5. Disponível em: <http://doi.acm.org/10.1145/1639714.1639724>.

[Goebel e Gruenwald 1999]GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 1, n. 1, p. 20–33, jun. 1999. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/846170.846172>.

[Golder e Huberman 2006]GOLDER, S. A.; HUBERMAN, B. A. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, Sage Publications, Inc., Thousand Oaks, CA, USA, v. 32, p. 198–208, April 2006. ISSN 0165-5515. Disponível em: <http://dl.acm.org/citation.cfm?id=1119738.1119747>.

[Gracia e Mena 2008]GRACIA, J.; MENA, E. Web-based measure of semantic relatedness. In: *Proceedings of the 9th international conference on Web Information Systems Engineering*. Berlin, Heidelberg: Springer-Verlag, 2008. (WISE '08), p. 136–150. ISBN 978-3-540-85480-7. Disponível em: <http://dx.doi.org/10.1007/978-3-540-85481-4_12>.

[Gruber 1993]GRUBER, T. R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, Academic Press Ltd., London, UK, UK, v. 5, n. 2, p. 199–220, jun. 1993. ISSN 1042-8143. Disponível em: <http://dx.doi.org/10.1006/knac.1993.1008>.

[Gu et al. 2008]GU, Q. et al. Data mining on imbalanced data sets. In: *Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on*. [S.l.: s.n.], 2008. p. 1020–1024.

[Gupta et al. 2010]GUPTA, M. et al. Survey on social tagging techniques. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 12, p. 58–72, November 2010. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1882471.1882480>.

[Han e Kamber 2006]HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques, 2nd Edition*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2006. ISBN 13: 978-1-55860-901-3.

[Harrington 2012]HARRINGTON, P. *Machine Learning in Action*. Shelter Island, NY, USA: Manning, 2012. ISBN 9781617290183.

[Hart 1968]HART, P. The condensed nearest neighbor rule. *Information Theory, IEEE Transactions on*, v. 14, n. 3, p. 515–516, 1968. ISSN 0018-9448.

[Hassan-Montero e Herrero-Solana 2006]HASSAN-MONTERO, Y.; HERRERO-SOLANA, V. Improving tag-clouds as visual information retrieval interfaces. In: *Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies*. [S.l.: s.n.], 2006. (InSciT2006), p. 1–6.

[Hearst 1992]HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th conference on Computational linguistics - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. (COLING '92), p. 539–545. Disponível em: <http://dx.doi.org/10.3115/992133.992154>.

[Heymann e Garcia-Molina 2006]HEYMANN, P.; GARCIA-MOLINA, H. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. [S.l.], 2006. 1–5 p.

[Heymann, Ramage e Garcia-Molina 2008]HEYMANN, P.; RAMAGE, D.; GARCIA-MOLINA, H. Social tag prediction. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008. (SIGIR '08), p. 531–538. ISBN 978-1-60558-164-4. Disponível em: <http://doi.acm.org/10.1145/1390334.1390425>.

[Hotho et al. 2006]HOTHO, A. et al. Information retrieval in folksonomies: Search and ranking. In: SURE, Y.; DOMINGUE, J. (Ed.). *The Semantic Web: Research and Applications*. Heidelberg: Springer, 2006. (LNAI, v. 4011), p. 411–426.

[Hotho, Nürnberger e Paaß 2005]HOTHO, A.; NÜRNBERGER, A.; PAAß, G. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, v. 20, n. 1, p. 19–62, maio 2005. ISSN 0175-1336. Disponível em: <http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf>.

[Houaiss 2009]HOUAISS, I. A. *Dicionário Houaiss Eletrônico*. [S.l.]: Editora Objetiva Ltda., 2009. CD-ROM. Versão 3.0.

[Jeatrakul e Wong 2012]JEATRAKUL, P.; WONG, K. W. Enhancing classification performance of multi-class imbalanced data using the oaa-db algorithm. In: *Neural Networks (IJCNN), The 2012 Annual International Joint Conference on*. Brisbane, Australia: [s.n.], 2012. p. 1–8. ISSN 2161-4393.

[Kadlec 2010]KADLEC, J. *Measures of semantic similarity in folksonomies*. Dissertação (Diploma Thesis) — Aalborg Universitet, Aalborg, Denmark, June 2010. Disponível em: <http://projekter.aau.dk/projekter/en/studentthesis/measures-of-semantic-similarity-in-folksonomies>.

[Knautz, Soubusta e Wolfgang 2010]KNAUTZ, K.; SOUBUSTA, S.; WOLFGANG, S. Tag clusters as information retrieval interfaces. In: *Proceedings of the 43rd Hawaii International Conference on System Sciences*. Washington, DC, USA: IEEE Computer Society, 2010. (HICSS '10), p. 1–10. ISBN 978-0-7695-3869-3. Disponível em: <http://doi.acm.org/10.1145/1935826.1935855>.

[Köpcke, Thor e Rahm 2010]KÖPCKE, H.; THOR, A.; RAHM, E. Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endow.*, VLDB Endowment, v. 3, n. 1-2, p. 484–493, set. 2010. ISSN 2150-8097. Disponível em: <http://dl.acm.org/citation.cfm?id=1920841.1920904>.

[Kotsiantis e Kanellopoulos 2006]KOTSIANTIS, S.; KANELLOPOULOS, D. Association rules mining: A recent overview. *International Transactions on Computer Science and Engineering*, Global Engineering, Science, and Technology Society (GESTS), v. 32, n. 1, p. 71–82, Jan 2006.

[Kotsiantis, Kanellopoulos e Pintelas 2006]KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Handling imbalanced datasets: A review. *GESTS - International Transactions on computer Science and Engineering*, v. 30, p. 25–36, 2006.

[Kubat, Holte e Matwin 1998]KUBAT, M.; HOLTE, R. C.; MATWIN, S. Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 30, n. 2-3, p. 195–215, fev. 1998. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1007452223027>.

[Kubat e Matwin 1967]KUBAT, M.; MATWIN, S. Addressing the curse of imbalanced training sets: one-sided selection. In: FISHER, D. H. (Ed.). *14th International Conference on Machine Learning - ICML*. Oregon, USA: [s.n.], 1967. p. 179–186. ISBN 1-55860-486-3. ISSN 0018-9448.

[Kumar e Verma 2012]KUMAR, R.; VERMA, R. Classification algorithms for data mining: A survey. *International Journal of Innovations in Engineering and Technology (IJIET)*, v. 1, n. 2, p. 7–14, August 2012. ISSN 2319-1058.

[Ladha e Deepa 2011]LADHA, L.; DEEPA, T. Feature selection methods and algorithms. *International Journal on Computer Science and Engineering (IJCSE)*, v. 3, n. 5, p. 1787–1797, may 2011. ISSN 0975-3397. Disponível em: <http://doi.acm.org/10.1145/2347736.2347755>.

[Laniado, Eynard e Colombetti 2007]LANIADO, D.; EYNARD, D.; COLOMBETTI, M. Using wordnet to turn a folksonomy into a hierarchy of concepts. In: *Semantic Web Application and Perspectives - Fourth Italian Semantic Web Workshop*. [s.n.], 2007. p. 192–201. Disponível em: <http://home.dei.polimi.it/eynard/papers/swap2007.pdf>.

[Lee e Yong 2007]LEE, S.-S.; YONG, H.-S. Tagplus: A retrieval system using synonym tag in folksonomy. In: *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering*. Washington, DC, USA: IEEE Computer Society, 2007. (MUE '07), p. 294–298. ISBN 0-7695-2777-9. Disponível em: <http://dx.doi.org/10.1109/MUE.2007.201>.

[Leginus, Dolog e Lage 2013]LEGINUS, M.; DOLOG, P.; LAGE, R. Graph based techniques for tag cloud generation. In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, 2013. (HT '13), p. 148–157. ISBN 978-1-4503-1967-6. Disponível em: <http://doi.acm.org/10.1145/2481492.2481508>.

[Leginus, Dolog e Lage 2013]LEGINUS, M.; DOLOG, P.; LAGE, R. Tag cloud generation for results of multiple keywords queries. In: DANIEL, F.; DOLOG, P.; LI, Q. (Ed.). *Web Engineering*. Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 7977). p. 233–248. ISBN 978-3-642-39199-6. Disponível em: <http://dx.doi.org/10.1007/978-3-642-39200-9_21>.

[Levenshtein 1966]LEVENSHTEIN, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, v. 10, p. 707–710, 1966.

[Li et al. 2007]LI, R. et al. Towards effective browsing of large scale social annotations. In: *Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007. (WWW '07), p. 943–952. ISBN 978-1-59593-654-7. Disponível em: <http://doi.acm.org/10.1145/1242572.1242700>.

[Lin 1998]LIN, D. An information-theoretic definition of similarity. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. (ICML '98), p. 296–304. ISBN 1-55860-556-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645527.657297>.

[Ling e Sheng 2008]LING, C. X.; SHENG, V. S. Cost-sensitive Learning and the Class Imbalanced Problem. In: ____. *Encyclopedia of Machine Learning*. [S.l.]: Springer, 2008. p. 1–8.

[Liu, Zhou e Zheng 2007]LIU, X.-Y.; ZHOU, Y.-M.; ZHENG, R.-S. Measuring semantic similarity in wordnet. In: *Machine Learning and Cybernetics, 2007 International Conference on*. [S.l.: s.n.], 2007. v. 6, p. 3431–3435.

[Lohmann, Ziegler e Tetzlaff 2009]LOHMANN, S.; ZIEGLER, J.; TETZLAFF, L. Comparison of tag cloud layouts: Task-related performance and visual exploration. In: *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I*. Berlin, Heidelberg: Springer-Verlag, 2009. (INTERACT '09), p. 392–404. ISBN 978-3-642-03654-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03655-2_43>.

[Magableh et al. 2010]MAGABLEH, M. et al. Towards a multilingual semantic folksonomy. In: *Proceedings of the IADIS International Conferences Collaborative Technologies 2010 and Web Based Communities 2010*. [S.l.: s.n.], 2010. p. 178–182.

[Manning, Raghavan e Schütze 2008]MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

[Marinho, Buza e Schmidt-Thieme 2008]MARINHO, L. B.; BUZA, K.; SCHMIDT-THIEME, L. Folksonomy-based collabulary learning. In: *Proceedings of the 7th International Conference on The Semantic Web*. Berlin, Heidelberg: Springer-Verlag, 2008. (ISWC '08), p. 261–276. ISBN 978-3-540-88563-4. Disponível em: <http://dx.doi.org/10.1007/978-3-540-88564-1_17>.

[Markines et al. 2009]MARKINES, B. et al. Evaluating similarity measures for emergent semantics of social tagging. In: *Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, 2009. (WWW '09), p. 641–650. ISBN 978-1-60558-487-4. Disponível em: <http://doi.acm.org/10.1145/1526709.1526796>.

[Mena e Gonzalez 2006]MENA, L.; GONZALEZ, J. A. Machine learning for imbalanced datasets: Application in medical diagnostic. In: SUTCLIFFE, G.; GOEBEL, R. (Ed.). *FLAIRS Conference*. [S.l.]: AAAI Press, 2006. p. 574–579.

[Meo, Quattrone e Ursino 2009]MEO, P. D.; QUATTRONE, G.; URSINO, D. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Inf. Syst.*, Elsevier Science Ltd., Oxford, UK, UK, v. 34, n. 6, p. 511–535, set. 2009. ISSN 0306-4379. Disponível em: <http://dx.doi.org/10.1016/j.is.2009.02.004>.

[Miller 1995]MILLER, G. A. Wordnet: A lexical database for english. *Commun. ACM*, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, nov. 1995. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/219717.219748>.

[Mitchell 1997]MITCHELL, T. Does machine learning really work? *Artificial Inteligence Magazine*, v. 18, n. 3, p. 11–20, 1997.

[Mitchell 1997]MITCHELL, T. M. *Machine Learning*. 1st. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

[Mohammad e Hirst 2012]MOHAMMAD, S.; HIRST, G. Distributional measures of semantic distance: A survey. *CoRR*, abs/1203.1858, 2012.

[Monard e Baranauskas 2003]MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: MANOLE. *Sistemas Inteligentes - Fundamentos e Aplicações*. [S.l.]: Manole, 2003. cap. 4, p. 89–114.

[Mousselly-Sergieh et al. 2014]MOUSSELLY-SERGIEH, H. et al. Tag relatedness using laplacian score feature selection and adapted jensen-shannon divergence. In: GURRIN, C. et al. (Ed.). *MultiMedia Modeling*. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8325). p. 159–171. ISBN 978-3-319-04113-1. Disponível em: <http://dx.doi.org/10.1007/978-3-319-04114-8_14>.

[Noruzi 2007]NORUZI, A. Folksonomies: Why do we need controlled vocabulary? *Webology*, v. 4, n. 2, 2007. Disponível em: <http://www.webology.ir/2007/v4n2/editorial12.html>.

[Patwardhan, Banerjee e Pedersen 2003]PATWARDHAN, S.; BANERJEE, S.; PEDERSEN, T. Using measures of semantic relatedness for word sense disambiguation. In: *Proceedings of the 4th international conference on Computational linguistics and intelligent text processing*. Berlin, Heidelberg: Springer-Verlag, 2003. (CICLing'03), p. 241–257. ISBN 3-540-00532-3. Disponível em: <http://dl.acm.org/citation.cfm?id=1791562.1791592>.

[Pednault, Rosen e Apte 2000]PEDNAULT, E.; ROSEN, B.; APTE, C. Handling imbalanced data sets in insurance risk modeling. In: *AAAI Workshop on Learning from Imbalanced Data Sets, Technical Report WS-00-05*. [S.l.]: AAAI Press, 2000. p. 01–06.

[Peters 2009]PETERS, I. *Folksonomies. Indexing and Retrieval in Web 2.0*. 1st. ed. Hawthorne, NJ, USA: Walter de Gruyter & Co., 2009. ISBN 3598251793, 9783598251795.

[Phoungphol, Zhang e Zhao 2012]PHOUNGPHOL, P.; ZHANG, Y.; ZHAO, Y. Robust multiclass classification for learning from imbalanced biomedical data. *Tsinghua Science and Technology*, v. 17, n. 6, p. 619–628, Dec 2012.

[Ponciano et al. 2014]PONCIANO, L. et al. Estratégia de replicação adaptativa para tarefas de computação por humanos. In: *XXXII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Florianópolis, Brasil: [s.n.], 2014. p. 1–14.

[Quintarelli 2005]QUINTARELLI, E. *Folksonomies: Power to the People*. 2005. Disponível em: <http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm>.

[Radelaar et al. 2011]RADELAAR, J. et al. Improving the exploration of tag spaces using automated tag clustering. In: *Web Engineering*. Berlin, Heidelberg: Springer-Verlag, 2011. (6757), p. 274–288. ISBN 978-3-642-22233-7.

[Rêgo, Marinho e Pires 2012]RÊGO, A. S.; MARINHO, L. B.; PIRES, C. E. Learning synonym relations from folksonomies. In: *International Conference WWW/Internet (IADIS)*. Madrid, Spain: [s.n.], 2012. p. 294–301.

[Rêgo, Marinho e Pires 2015]RÊGO, A. S. C.; MARINHO, L. B.; PIRES, C. E. S. A supervised learning approach to detect subsumption relations between tags in folksonomies. In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2015. (SAC '15), p. 409–415. ISBN 978-1-4503-3196-8. Disponível em: <http://doi.acm.org/10.1145/2695664.2695904>.

[Rendle e Schmidt-Thieme 2006]RENDLE, S.; SCHMIDT-THIEME, L. Object identification with constraints. In: *Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2006. (ICDM '06), p. 1026–1031. ISBN 0-7695-2701-9. Disponível em: <http://dx.doi.org/10.1109/ICDM.2006.117>.

[Resnik 1995]RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 448–453. ISBN 1-55860-363-8, 978-1-558-60363-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1625855.1625914>.

[Resnik 1999]RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, v. 11, n. 1, p. 95–130, 1999.

[Rifkin e Klautau 2004]RIFKIN, R.; KLAUTAU, A. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, JMLR.org, v. 5, p. 101–141, dez. 2004. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1005332.1005336>.

[Rivadeneira et al. 2007]RIVADENEIRA, A. W. et al. Getting our head in the clouds: Toward evaluation studies of tagclouds. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2007. (CHI '07), p. 995–998. ISBN 978-1-59593-593-9. Disponível em: <http://doi.acm.org/10.1145/1240624.1240775>.

[Rocha e Goldenstein 2014]ROCHA, A.; GOLDENSTEIN, S. K. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *Neural Networks and Learning Systems, IEEE Transactions on*, v. 25, n. 2, p. 289–302, Feb 2014. ISSN 2162-237X.

[Sanderson 1994]SANDERSON, M. Word sense disambiguation and information retrieval. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Springer-Verlag New York, Inc., 1994. (SIGIR '94), p. 142–151. ISBN 0-387-19889-X. Disponível em: <http://dl.acm.org/citation.cfm?id=188490.188548>.

[Schmitz et al. 2006]SCHMITZ, C. et al. Mining association rules in folksonomies. In: BATAGELJ, V. et al. (Ed.). *Data Science and Classification. Proceedings of the 10th IFCS Conf.* Heidelberg: Springer, 2006. (Studies in Classification, Data Analysis, and Knowledge Organization), p. 261–270.

[Si, Liu e Sun 2010]SI, X.; LIU, Z.; SUN, M. Explore the structure of social tags by subsumption relations. In: HUANG, C.-R.; JURAFSKY, D. (Ed.). *COLING*. [S.l.]: Tsinghua University Press, 2010. p. 1011–1019.

[Sinclair e Cardew-Hall 2008]SINCLAIR, J.; CARDEW-HALL, M. The folksonomy tag cloud: When is it useful? *J. Inf. Sci.*, Sage Publications, Inc., Thousand Oaks, CA, USA, v. 34, n. 1, p. 15–29, fev. 2008. ISSN 0165-5515. Disponível em: <http://dx.doi.org/10.1177/0165551506078083>.

[Singhal 2001]SINGHAL, A. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, v. 24, n. 4, p. 35–43, 2001.

[Solskinnsbakk e Gulla 2010]SOLSKINNSBAKK, G.; GULLA, J. A. A hybrid approach to constructing tag hierarchies. In: *Proceedings of the 2010 international conference on On the move to meaningful internet systems: Part II*. Berlin, Heidelberg: Springer-Verlag, 2010. (OTM'10), p. 975–982. ISBN 3-642-16948-1, 978-3-642-16948-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1926129.1926160>.

[Solskinnsbakk e Gulla 2011]SOLSKINNSBAKK, G.; GULLA, J. A. Mining tag similarity in folksonomies. In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. New York, NY, USA: ACM, 2011. (SMUC'11), p. 53–60. ISBN 978-1-4503-0949-3. Disponível em: <http://doi.acm.org/10.1145/2065023.2065037>.

[Spiliopoulos, Vouros e Karkaletsis 2010]SPILIOPOULOS, V.; VOUROS, G. A.; KARKALETSIS, V. On the discovery of subsumption relations for the alignment of ontologies. *Web Semant.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 8, n. 1, p. 69–88, mar. 2010. ISSN 1570-8268. Disponível em: <http://dx.doi.org/10.1016/j.websem.2010.01.001>.

[Stolfo et al. 1997]STOLFO, S. J. et al. Credit card fraud detection using meta-learning: Issues and initial results. In: *In AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management*. [S.l.: s.n.], 1997. p. 83–90.

[Talavera 2005]TALAVERA, L. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In: FAMILI, A. et al. (Ed.). *Advances in Intelligent Data Analysis VI*. Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, v. 3646). p. 440–451. ISBN 978-3-540-28795-7. Disponível em: <http://dx.doi.org/10.1007/11552253_40>.

[Tang et al. 2009]TANG, J. et al. Towards ontology learning from folksonomies. In: *Proceedings of the 21st international jont conference on Artifical intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009. (IJCAI'09), p. 2089–2094. Disponível em: <http://dl.acm.org/citation.cfm?id=1661445.1661779>.

[Tibély et al. 2013]TIBéLY, G. et al. Extracting tag hierarchies. *PLoS ONE*, Public Library of Science, v. 8, n. 12, p. 1–12, 12 2013. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0084133>.

[Tomek 1976]TOMEK, I. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 7(2), p. 679–772, 1976.

[Venetis, Koutrika e Garcia-Molina 2011]VENETIS, P.; KOUTRIKA, G.; GARCIA-MOLINA, H. On the selection of tags for tag clouds. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2011. (WSDM '11), p. 835–844. ISBN 978-1-4503-0493-1. Disponível em: <http://doi.acm.org/10.1145/1935826.1935855>.

[Viera e Garrett 2005]VIERA, A.; GARRETT, J. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, v. 37, n. 5, p. 360–363, 2005.

[Voorhees e Harman 2005]VOORHEES, E. M.; HARMAN, D. K. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. [S.l.]: The MIT Press, 2005. ISBN 0262220733.

[Wang e Yao 2012]WANG, S.; YAO, X. Multiclass imbalance problems: Analysis and potential solutions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, v. 42, n. 4, p. 1119–1130, Aug 2012. ISSN 1083-4419.

[Wartena 2010]WARTENA, C. Testing the distributional hypothesis for collaborative tagging systems. In: HANNEFORTH, T.; FANSELOW, G. (Ed.). *Language and Logos*. Berlin: Akademie Verlag, 2010, (Studia Grammatica, v. 72). p. 407–415.

[Wartena, Brussee e Wibbels 2009]WARTENA, C.; BRUSSEE, R.; WIBBELS, M. Using tag co-occurrence for recommendation. In: *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*. Washington, DC, USA: IEEE Computer Society, 2009. (ISDA'09), p. 273–278. ISBN 978-0-7695-3872-3. Disponível em: <http://dx.doi.org/10.1109/ISDA.2009.130>.

[Weiss 2004]WEISS, G. M. Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 6, n. 1, p. 7–19, jun. 2004. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1007730.1007734>.

[Witten e Frank 2011]WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0120884070.

[Wu e Zhou 2009]WU, C.; ZHOU, B. Semantic relatedness in folksonomy. In: *Proc. of the 2009 International Conference on New Trends in Information and Service Science*. Washington, DC, USA: IEEE Computer Society, 2009. p. 760–765. ISBN 978-0-7695-3687-3. Disponível em: <http://dl.acm.org/citation.cfm?id=1636710.1637435>.

[Wu et al. 2007]WU, X. et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, Springer-Verlag New York, Inc., New York, NY, USA, v. 14, n. 1, p. 1–37, dez. 2007. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/s10115-007-0114-2>.

[Wu, Zhang e Yu 2006]WU, X.; ZHANG, L.; YU, Y. Exploring social annotations for the semantic web. In: *Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006. (WWW '06), p. 417–426. ISBN 1-59593-323-9. Disponível em: <http://doi.acm.org/10.1145/1135777.1135839>.

[Xiong, Wu e Liu 2010]XIONG, H.; WU, J.; LIU, L. Classification with class overlapping - a systematic study. In: *Proceedings of International Conference on E-Business Intelligence (ICEBI-2010)*. Atlantis Press, 2010, (ICEBI-2010, v. 2972). p. 491–497. ISBN 978-90-78677-40-6. Disponível em: <http://www.atlantis-press.com/php/download_paper.php?id=2053>.

[Yeung, Gibbins e Shadbolt 2007]YEUNG, C.-m.; GIBBINS, N.; SHADBOLT, N. Understanding the semantics of ambiguous tags in folksonomies. In: HAASE, P. et al. (Ed.). *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC2007, Busan, South Korea*. Busan, South Korea: [s.n.], 2007. p. 108–121.

[Yin et al. 2014]YIN, X. et al. What? how? where? a survey of crowdsourcing. In: LI, S. et al. (Ed.). *Frontier and Future Development of Information Technology in Medicine and Education*. Springer Netherlands, 2014, (Lecture Notes in Electrical Engineering, v. 269). p. 221–232. ISBN 978-94-007-7617-3. Disponível em: <http://dx.doi.org/10.1007/978-94-007-7618-0_22>.

[Zhang et al. 2013]ZHANG, J. et al. On the application of multi-class classification in physical therapy recommendation. *Health Information Science and Systems*, p. 1–14, 2013.

[Zhao et al. 2008]ZHAO, X.-M. et al. Protein classification with imbalanced data. *Proteins: Structure, Function, and Bioinformatics*, Wiley Subscription Services, Inc., A Wiley Company, v. 70, n. 4, p. 1125–1132, 2008. ISSN 1097-0134. Disponível em: <http://dx.doi.org/10.1002/prot.21870>.

[Zhou et al. 2007]ZHOU, M. et al. An unsupervised model for exploring hierarchical semantics from social annotations. In: *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*. Berlin, Heidelberg: Springer-Verlag, 2007. (ISWC'07/ASWC'07), p. 680–693. ISBN 3-540-76297-3, 978-3-540-76297-3. Disponível em: <http://dl.acm.org/citation.cfm?id=1785162.1785212>.

[Zhou e Liu 2006]ZHOU, Z.-H.; LIU, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on*, v. 18, n. 1, p. 63–77, Jan 2006. ISSN 1041-4347.

Apêndices

Apêndice A

Casos Típicos de Sinonímia em Folksonomias

Devido ao vocabulário livre praticado nas folksonomias, os recursos são muitas vezes marcados por inúmeras *tags* sinônimas. Como consequência, o problema de sinonímia impede que os usuários consigam recuperar recursos que foram anotados com *tags* sinônimas que ele/ela normalmente não estão acostumados a usar. Este apêndice apresenta o resultado de uma análise particular efetuada no espaço de *tags* de um extrato da folksonomia BibSonomy, sobre as ocorrências típicas de sinonímia.

1. **Número gramatical**: *tags* podem ser encontradas na forma singular ou no plural, tais como `Animal` e `Animals`, por exemplo;

2. **Sensibilidade ao tamanho** (*case sensitivity*): *tags* são diferenciadas por letras maiúsculas e minúsculas, por isso, é possível encontrar uma mesma *tag* escrita em diferentes combinações de caracteres maiúsculos ou minúsculos. Por exemplo, as *tags* `Brazil` e `brazil`;

3. **Erro ortográfico**: *tags* escritas com erros ortográficos podem deixar recursos "órfãos", em virtude de terem sido marcados com apenas uma (ou poucas) *tag(s)*. Por exemplo, as *tags* `laptop` e `latpop`;

4. **Caracteres especiais e símbolos**: usuários normalmente adicionam símbolos e caracteres especiais tais como vírgula, ponto e vírgula, colchetes, chaves, parênteses,

entre outros, para customizar a identificação de uma *tag*. Como exemplos desse caso se enquadram as *tags* `(Psychology);` e `{DNA}`;

5. **Inflexões do substantivo/verbo**: as variações devido a inflexões do substantivo e verbo produzem *tags* tais como `account` e `accountancy`, que são termos similares;

6. **Vocabulário do usuário**: a atribuição de *tags* depende, às vezes, da preferência de vocabulário do usuário. Por exemplo, alguns usuários podem anotar um documento sobre computadores Macintosh com a *tag* `mac`, enquanto outros usuários podem decidir usar a *tag* `macintosh`;

7. **Multilinguismo**: algumas palavras podem ser escritas de forma diferente de acordo com o idioma do usuário. As *tags* `casa` (português), `house` (inglês), e `maison` (francês), por exemplo, têm o mesmo significado. Às vezes, as variações ocorrem dentro da mesma linguagem como, por exemplo, `behavior` (inglês americano) e `behaviour` (inglês britânico);

8. **Sinonímia típica**: palavras distintas podem ter o mesmo significado, como no exemplo das *tags* `picture`, `photo` e `photograph`;

9. **Acrônimos**: são abreviações formadas a partir das letras iniciais de uma frase ou palavra, geralmente em letras maiúsculas. A existência de acrônimos é normalmente vista como um problema de ambiguidade em folksonomias, mas também traz evidências de sinonímia. Por exemplo, as *tags* `IR` e `Information Retrieval` normalmente se referem ao mesmo conceito.

Nota-se que alguns casos de sinonímia mencionados, como por exemplo os casos 2 e 4, podem ser facilmente manipulados por algum tipo simples de normalização. Por exemplo, sensitividade ao tamanho pode ser resolvido ao normalizar todas as *tags* para caracteres minúsculos. Os símbolos especiais podem ser removidos por um analisador de expressão regular.

Um pouco mais da metade dos pares de *tags* que expressam relação de sinonímia no BibSonomy são decorrentes das situações especificadas pelos casos de 1 a 5. Isto pode

ser percebido durante a análise das 10.000 *tags* mais populares do BibSonomy quando as seguintes *tags* foram encontradas para o termo *learning*: `LEARNING`, `Learning`, `learning`, `Learning,`, `Learning;` e `Learning2.0`.

Apêndice B

Uma Análise sobre a Variação de Limiar vs. Distância de Edição

A medida distância de edição (DE) é perfeitamente apropriada para identificar os casos mais típicos de sinonímia observados na base de dados do BibSonomy: plural vs. singular, inflexões de verbo ou substantivo e modificações da palavra causadas pela adição de caracteres especiais. Estes casos tem como principal característica uma suave variação da palavra original em relação às suas candidatas a sinônimo. Certamente, este comportamento tende a ser observado em outras bases de dados, visto que a ausência de convenção na criação de *tags* tem como efeito o acúmulo de termos redundantes.

Escolher a melhor faixa de limiar que consegue reunir o maior número de sinônimos possível é uma tarefa subjetiva e varia de acordo com o cenário experimental. Entretanto, apenas encontrar a melhor faixa de limiar não oferece um entendimento mais detalhado sobre onde se concentram os casos de sinonímia referenciados no Apêndice A. Para responder a esta questão, uma análise manual da medida *DEN* foi realizada ao longo do intervalo $]0, 1]$, com incrementos de 0.1, em todas as instâncias positivas rotuladas como sinônimas pelo WordNet. A Tabela B.1 sumariza o resultado da análise.

Observa-se que o limiar 0.3 é adequado para capturar a grande ocorrência de variações de número gramatical de uma palavra. Entretanto, vale salientar que um valor pequeno de distância de edição não é garantia de que se tenha a afirmação de sinonímia, como pode-se perceber nos pares de *tags* $oncology$ e $ontology$ (0.125), $article$ e $particle$ (0.125) e $best$ e $test$ (0.25). Esses casos são minoria no universo analisado e a quantidade de ruídos

Tabela B.1: Estimativa de limiar vs. caracterização de sinonímia para a medida DE.

Tipo de Sinonímia	Melhor Limiar	Exemplos
Plural/Singular	$0.0 < thr \leq 0.3$	(art,arts): 0.25 (story,stories):0.4286
Inflexões (substantivo/verbo)	$0.3 < thr \leq 0.55$	(bank, banking):0.4286 (site,website):0.500
Sinônimos	$thr > 0.55$	(life, living):0.6667 (earth,world):0.8000 (cite, reference):1.000

produzida é discreta.

As inflexões começam a aparecer quando o limiar é estendido para 0.55, valor este que provê uma melhor margem qualitativa sem que os ruídos interfiram negativamente. Nesta faixa, é possível detectar os pares $telephone$ e $phone$ (0.4286), $measurement$ e $measure$ (0.3636) e $querying$ e $query$ (0.3750). Evidentemente, a quantidade de ruídos tende a aumentar significativamente à medida que o limiar vai sendo incrementado. Por fim, os pares de *tags* sinônimas que possuem grafias distintas são percebidos quando o limiar é superior a 0.55.

De acordo com a análise efetuada, um limiar máximo aceitável para capturar casos de sinonímia é **0.5**. Acima deste valor, não compensa considerar DE para identificar casos de sinonímia porque, neste caso, os casos de sinonímia se perdem diante da elevada quantidade de ruídos. Embora não esteja destacado na Tabela B.1, acrônimos também foram encontrados no intervalo $[0,57; 1,0]$, representados pelos exemplos www e web (0.6667), USA e $United\_States$ (0.8462) e CV e $resume$ (1.0000). Seguindo o mesmo raciocínio anterior, é evidente que DE não é apropriada para capturar estes casos.

Apêndice C

O Algoritmo *Taxonomy Learning*

O algoritmo *Taxonomy Learning* proposto por Marinho, Buza e Schmidt-Thieme (2008), baseia-se na técnica de mineração de conjuntos de itens frequentes para aprender uma ontologia a partir de uma folksonomia, com o objetivo de criar uma taxonomia em árvore. Este algoritmo foi explorado no presente trabalho para revelar relações taxonômicas entre *tags* com a finalidade de servir de base para computação de valores do atributo *busca hierárquic*a ($tsearch$), descrito na Seção 4.2.

A folksonomia (*tags* e recursos) é projetada para um banco de dados transacional, na qual cada transação é composta pelo agrupamento das *tags* usadas por um usuário em particular para anotar um referido recurso, da forma *idTransaction*:$\{t_1, t_2, ..., t_n\}$, com $t \in T$.

Em linhas gerais, o método de criação da estrutura taxonômica é descrito no Pseudocódigo C.1. Seja $m[]$ um array contendo os valores mínimos de $suporte$ que serão considerados em cada iteração, $e[]$ um array com os valores de limiar que vão determinar os pesos para cada tamanho de conjunto de itens e F o conjunto de itens frequentes.

Pseudo-código C.1: Funcionamento do algoritmo *Taxonomy Learning*

```
REQUISITOS:
a) Dados da folksonomia,
b) Array de Limiar Minimo de Suporte m[]
c) Array de Limiar de aresta e[]

1. Projetar os dados da folksonomia para o banco de dados transacional D
2. Inicializar o grafo S de representação da folksonomia
3. Para i=0 até i < m.length
```

```
4.      fqSet = mineFrequentItemset(D,m[i])
5.      S1 = buildTaxonomyPieces(fqSet,e)
6.      S2 = prunePieces(S1)
7.      S = addPrunePiecesToTaxonomy(S,S2)
8. Fim Para
```

O princípio de funcionamento do algoritmo é descrito da seguinte forma:

- A partir de um processo iterativo (passo 1), os conjuntos de itens mais frequentes são extraídos do banco de dados transacional (passo 4). Se um conjunto de itens frequentes f' possui um suporte mensurado como significativo, pode-se dizer que os itens que ocorrem em f' são estreitamente relacionados. O índice do vetor $m[]$ define o suporte mínimo da iteração, o qual é usado para selecionar os conjuntos de itens frequentes $\{f \in F | suporte(f) \geq m[i]\}$. O suporte é reduzido nas iterações subsequentes;

- Após o processo de mineração, é iniciada (ou incrementada) a construção da taxonomia (passo 5). Neste caso, as relações taxonômicas são aprendidas na iteração corrente, de modo que dois nós t_x e t_y são conectados (com t_x sendo um superconceito de t_y) de tal modo que $suporte(t_x) \geq e[j].suporte(t_y)$, em que $j = |f|$ e $e[j]$ é um limiar pré-estimado de aresta para conjuntos de itens de tamanho j. Caso t_x e t_y satisfaçam às condições estabelecidas, estes são adicionados como vértices do grafo S e conectados por uma aresta;

- Para evitar múltiplas relações de herança, ou seja, que um vértice possua mais de uma aresta de chegada, o grafo é podado para assumir a forma de uma árvore (passo 6). Neste passo, a preferência é para os caminhos mais longos (maior altura), uma vez que, segundo os autores, são geralmente mais informativos. Por exemplo, supondo a existência das arestas $t_1 \succ t_2$, $t_2 \succ t_3$ e $t_1 \succ t_3$, com t_3 tendo dois vértices pai, a aresta $t_1 \succ t_3$ é considerada redundante, portanto, é removida da árvore. Note na Figura C.1 que o caminho de t_1 passando por t_2 para chegar em t_3 é maior. Esta etapa é apoiada pela travessia de busca em profundidade diante dos fragmentos da taxonomia, para garantir que cada conceito tenha apenas um antecessor;

- No passo 7, as relações aprendidas na iteração corrente são mescladas com as relações aprendidas na iteração anterior, convergindo então para a taxonomia em árvore.

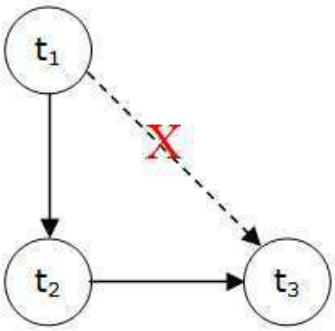

Figura C.1: Poda do grafo.

Nas iterações subsequentes (em um total de 7), o limiar de suporte mínimo é relaxado a fim de garimpar os conjuntos de itens menos frequentes que levam ao aprendizado de novos fragmentos. Após experimentos de calibração, o valor de limiar de suporte mínimo na sequência i das iterações foi determinado pela fórmula $\frac{0{,}025.|D|}{2^i}$ em que $|D|$ é o número total de transações. Os limiares de arestas foram definidos experimentalmente com os valores 1.5, 1.4, 1.3, 1.2 e 1.1 para os conjuntos de itens de tamanho 2,3,4,5 e 6, respectivamente.

Após implementação do algoritmo *Taxonomy Learning* para construir a taxonomia em árvore, os experimentos sugeriram ajustes que resultaram em um número total de iterações igual a 6, limiar de suporte mínimo de acordo com a fórmula $\frac{0{,}016.|D|}{2^i}$ e valor de limiar de aresta igual a 1.5 ou 1.4, visto que só foram considerados conjuntos de itens frequentes de tamanho 2.