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

SOLANGE DE LURDES PERTILE

Combinando Métricas Baseadas em Conteúdo e em Referências para a Detecção de Plágio em Artigos Científicos

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Orientador: Profa. Dra. Viviane Pereira Moreira

Porto Alegre
2015

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Pertile, Solange de Lurdes

Combinando Métricas Baseadas em Conteúdo e em Referências para a Detecção de Plágio em Artigos Científicos / Solange de Lurdes Pertile. – Porto Alegre: PPGC da UFRGS, 2015.

79 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2015. Orientador: Viviane Pereira Moreira.

1. Detecção de plágio. 2. Similaridade de conteúdo. 3. Análise de citações. I. Moreira, Viviane Pereira. II. Título.

“O talento vence jogos, mas só o trabalho em equipe ganha campeonatos.”

— MICHAEL JORDAN

AGRADECIMENTOS

Agradeço primeiramente, a Deus por tudo e, em especial, pelas oportunidades que me tem concedido e pelas pessoas que tem posto em meu caminho. Sou muito grata aos meus pais e ao meu esposo por serem meu porto seguro. O amor, o companheirismo, a amizade e a dedicação incondicional que eles tem me oferecido foram essenciais para aguentar esta jornada.

A minha mãe, pelas palavras de conforto ao telefone.

Ao meu esposo, pela dedicação e paciência durante todo esse período.

Agradeço intensamente a Viviane pela paciência e pela ótima orientação, na qual aprendi muito.

O meu doutorado foi uma jornada. Por vezes, considerei-a longa demais. Nessas horas, a amizade, as brincadeiras, os chopps, a parceria e a presença dos colegas do laboratório ajudaram torná-la mais curta. Agradeço também pela ajuda nas anotações manuais que foram necessárias para realização dos exeprimentos.

Enfim, a todos que direta ou indiretamente contribuíram para a realização deste trabalho.

RESUMO

A grande quantidade de artigos científicos disponíveis on-line faz com que seja mais fácil para estudantes e pesquisadores reutilizarem texto de outros autores, e torna mais difícil a verificação da originalidade de um determinado texto. Reutilizar texto sem creditar a fonte é considerado plágio. Uma série de estudos relatam a alta prevalência de plágio no meio acadêmico e científico. Como consequência, inúmeras instituições e pesquisadores têm se dedicado à elaboração de sistemas para automatizar o processo de verificação de plágio. A maioria dos trabalhos existentes baseia-se na análise da similaridade do conteúdo textual dos documentos para avaliar a existência de plágio. Mais recentemente, foram propostas métricas de similaridade que desconsideram o texto e analisam apenas as citações e/ou referências bibliográficas compartilhadas entre documentos. Entretanto, casos em que o autor não referencia a fonte original pode passar despercebido pelas métricas baseadas apenas na análise de referências/citações. Neste contexto, a solução proposta é baseada na hipótese de que a combinação de métricas de similaridade de conteúdo e de citações/referências pode melhorar a qualidade da detecção de plágio. Duas formas de combinação são propostas: (i) os escores produzidos pelas métricas de similaridade são utilizados para ranqueamento dos pares de documentos e (ii) os escores das métricas são utilizados para construir vetores de características que serão usados por algoritmos de Aprendizagem de Máquina para classificar os documentos. Os experimentos foram realizados com conjuntos de dados reais de artigos científicos. A avaliação experimental mostra que a hipótese foi confirmada quando a combinação das métricas de similaridade usando Aprendizagem de Máquina é comparada com a combinação simples. Ainda, ambas as combinações apresentaram ganhos quando comparadas com as métricas aplicadas de forma individual.

Palavras-chave: Detecção de plágio. similaridade de conteúdo. análise de citações.

Combining Content- and Citation-Based Metrics for Plagiarism Detection in Scientific Papers

ABSTRACT

The large amount of scientific documents available online makes it easier for students and researchers reuse text from other authors, and makes it difficult to verify the originality of a given text. Reusing text without crediting the source is considered plagiarism. A number of studies have reported on the high prevalence of plagiarism in academia. As a result, many institutions and researchers have developed systems that automate the plagiarism detection process. Most of the existing work is based on the analysis of the similarity of the textual content of documents to assess the existence of plagiarism. More recently, similarity metrics that ignore the text and just analyze the citations and/or references shared between documents have been proposed. However, cases in which the author does not reference the original source may go unnoticed by metrics based only on the references/citations analysis. In this context, the proposed solution is based on the hypothesis that the combination of content similarity metrics and references/citations can improve the quality of plagiarism detection. Two forms of combination are proposed: (i) scores produced by the similarity metrics are used to ranking of pairs of documents and (ii) scores of metrics are used to construct feature vectors that are used by algorithms machine learning to classify documents. The experiments were performed with real data sets of papers. The experimental evaluation shows that the hypothesis was confirmed when the combination of the similarity metrics using machine learning is compared with the simple combining. Also, both compounds showed gains when compared with the metrics applied individually.

Keywords: plagiarism detection, content similatity, citation analysis.

LISTA DE ABREVIATURAS E SIGLAS

ACL	*Association for Computational Linguistics*
ACM	*Association for Computing Machinery*
API	*Application Programming Interface*
COPE	*Committee on Publication Ethics*
WEKA	*Waikato Environment for Knowledge Analysis*
IDF	*Inverse document frequency – Inverso da Frequência nos Documentos*
IEEE	*Institute of Electrical and Electronics Engineers*
GCT	*Greedy Citation Tiling*
HTML	*HyperText Markup Language*
LCS	*Longest Common Substring*
PlaMIR	*Plagiarism by Missing and Incorrect Reference*
SCnG	*Surrounding Context N-grams*
SVM	*Support Vector Machine*
TF	Frequência do termo
URL	*Uniform Resource Locator*
XML	*eXtensible Markup Language*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Esta tese visa contribuir para a melhoria de sistemas de detecção de plágio em artigos científicos. A partir da combinação de métricas baseadas em conteúdo, referências e citações, a solução proposta visa auxiliar na detecção de diferentes casos de plágio de texto.

A Seção 1.1 deste Capítulo apresenta a definição do problema de pesquisa e a motivação que levou ao desenvolvimento deste trabalho. A Seção 1.2 descreve o objetivo principal da pesquisa e os passos percorridos para alcançá-lo. Na Seção 1.3 são descritas as principais contribuições desta tese. A Seção 1.4 fornece um resumo sobre como esta tese está organizada.

1.1 Problema e Motivação

A grande disponibilidade de trabalhos científicos tem facilitado aos estudantes e pesquisadores o reuso de textos de outros autores. No entanto, muitos não estão fazendo uso deste fácil acesso de forma correta. Reusar texto sem creditar os autores originais é considerado plágio, mesmo que feito de forma não intencional.

Uma série de estudos relatam a alta prevalência de plágio no meio acadêmico. Mccabe (2005) realizou uma pesquisa com mais de 80.000 alunos nos EUA e Canadá e descobriu que 36% dos estudantes de graduação e 24% dos estudantes de pós-graduação admitem ter copiado ou parafraseado frases da Internet sem fazer a devida referência ao original. Walker (2010) conduziu um estudo em que foram analisados práticas reais de plágio em uma universidade da Nova Zelândia. O autor concluiu que apesar dos alunos terem sido alertados sobre o plágio e que os trabalhos seriam avaliados por um software antiplágio, mais de um quarto dos trabalhos apresentados continha plágio e cerca de 10% foram amplamente plagiados. Um estudo realizado por Youmans (2011) também apontou que alertar os alunos sobre os conceitos de plágio e o uso de ferramentas de detecção não reduziu o plágio.

Enquanto os referidos estudos tratam de plágio em trabalhos de disciplinas universitárias, a literatura também tem alguns relatos sobre casos de plágio em publicações científicas. Zhang (2010) usou a ferramenta CrossCheck para analisar 662 trabalhos submetidos à revista da Universidade de Zhejiang (China), destes, 22,8% apresentaram ocorrências de cópia ou de auto-plágio, e 25,8% apresentaram sérias suspeitas de plágio e infração de direitos autorais. A similaridade de conteúdo entre os documentos originais e suspeitos chegou a 83% em alguns casos.

Os autores Gupta e Rosso (2012) também analisaram o reuso de texto em artigos ci-

entíficos. Eles estudaram tendências de reuso de texto na coleção da ACL (*Association for Computational Linguistic*) e encontraram altos níveis de reuso. Esta investigação incidiu apenas sobre reuso de cópia exata. Os resultados mostraram que o auto-reuso é mais frequente do que o reuso de outros autores.

Recentemente, um banco de dados chamado Déjà vu[1](GARNER, 2014) foi gerado a partir de aproximadamente 80.000 pares de artigos científicos do Medline (um repositório de artigos científicos da área médica e biomédica) para os quais uma alta similaridade de conteúdo foi encontrada pelo motor de busca eTBLAST [2]. Nem todos os casos são de fato plágio, pois existem razões legítimas para que dois documentos tenham alta similaridade. Por exemplo, um artigo de conferência pode ser expandido e publicado em um periódico. García-Romero e Estrada-Lorenzo (2014) realizaram uma análise bibliométrica de uma amostra de artigos do Déjà vu que haviam sido examinados por revisores. Os autores concluíram que os casos de plágio são publicados em revistas com menor visibilidade, corroborando uma constatação feita por Fang *et al.* (2012), e recebem menos citações.

A enorme quantidade de documentos digitais disponíveis torna a análise manual de plágio inviável. Como consequência, são propostas técnicas de detecção automática para lidar com as diversas formas de plágio. A análise de plágio é geralmente baseada na comparação do conteúdo dos documentos. Esta comparação normalmente atribui um grau de similaridade entre os documentos analisados, que é quantificada por um escore de similaridade.

Sistemas de detecção de plágio podem ser classificados como *intrínsecos* ou *extrínsecos*. Sistemas de detecção intrínsecos visam identificar as partes de um documento que provavelmente foram escritas por um autor diferente, enquanto que a detecção extrínseca funciona comparando um texto suspeito com uma coleção de referência de documentos originais.

De acordo com os conceitos de plágio encontrados na literatura (ANDERSON; STENECK, 2011; STEIN; EISSEN, 2006) e os tipos considerados pela IEEE e ACM (Seção 2.1), os métodos baseados apenas em análise de conteúdo não são suficientes para identificar se um par de documentos apresenta plágio ou não. No entanto, enquanto a maioria dos métodos de detecção de plágio extrínsecos focam na comparação do conteúdo textual dos documentos (KASPRZAK; BRANDEJS, 2010; BARRÓN-CEDEÑO; ROSSO, 2009; BALAGUER, 2009; MALCOLM; LANE, 2009; GROZEA; GEHL; POPESCU, 2009), mais recentemente, têm surgido métodos que visam detectar plágio com base na análise de referências e citações (ALZAHRANI et al., 2012; GIPP; BEEL, 2010; GIPP; MEUSCHKE, 2011; MEUSCHKE; GIPP; BREITINGER, 2012).

[1]<http://dejavu.vbi.vt.edu/dejavu/>
[2]<http://etest.vbi.vt.edu/etblast3/>

Gipp *et al.* (2014) realizou uma comparação entre abordagens que analisam conteúdo (usando as ferramentas Sherlock e Encoplot) e abordagens que analisam apenas citações. A proposta mostrou que, para análise de plágio onde ocorre cópia exata, as abordagens de análise de conteúdo apresentaram melhores resultados. Na análise de paráfrases, a detecção baseada em citação superou as abordagens que analisam apenas conteúdo.

No caso da detecção extrínseca, é necessária a disponibilidade de textos completos para poder realizar a análise de plágio. Entretanto, nem sempre isso é possível. Bases de dados digitais como, IEEE[3], ACM[4] e Springer[5] não disponibilizam acesso livre ao texto completo dos artigos publicados, mas permitem acesso ao resumo e à lista de referências bibliográficas do artigo.

Além disso, o plágio não é apenas a cópia do texto, uma vez que também existem casos de plágio dos resultados, de tabelas e reuso de figuras sem creditar a fonte original. Ademais, detectar casos de plágio em que o plagiador traduz o texto para um outro idioma não é trivial.

A literatura sobre as diversas formas de detecção de plágio apresenta uma lacuna quanto à combinação da detecção baseada em conteúdo e baseada em referências/citações. Sendo assim, o foco desta tese é comparar e propor uma combinação destes dois tipos de métricas a fim de beneficiar-se das vantagens de cada uma.

1.2 Objetivos

O objetivo desta tese é investigar a hipótese de que a combinação de métricas de similaridade de conteúdo e de referências/citações pode melhorar a qualidade dos sistemas de detecção de plágio.

Para alcançar este objetivo, foi necessária a execução das seguintes tarefas:

- Levantamento das abordagens de detecção de plágio existentes no estado da arte, visando identificar suas limitações. As abordagens foram classificadas em: (i) baseada em conteúdo, (ii) baseada em conteúdo e estrutura, e (iii) baseada em citações/referências foi aplicado. Um conjunto de métricas de similaridade foram aplicadas.
- Proposta de uma métrica que contabiliza as coocorrências de citações.
- Comparação das abordagens com objetivo de avaliar se elas são complementares e se a sua combinação pode melhorar na qualidade da detecção. As métricas, então, foram

[3]<http://ieeexplore.ieee.org/Xplore/home.jsp>
[4]<http://dl.acm.org/>
[5]<http://link.springer.com/>

combinadas de duas maneiras: (i) baseada em aprendizagem de máquina e (b) uma combinação simples.

- Análise de casos de plágio reconhecidos com objetivo de identificar as diferenças entre a avaliação de plágio por sistemas de detecção e por avaliadores humanos.

1.3 Contribuições

A principal contribuição da tese está na proposta de combinação de métricas de conteúdo e métricas de citações/referências. Além dessa, são também contribuições deste trabalho:

- um levantamento bibliográfico sobre as técnicas de detecção de plágio, organizando e comparando as diferentes abordagens propostas na literatura;
- experimentos com artigos científicos reais;
- proposta de uma métrica de similaridade de artigos baseada na contagem de coocorrências de citações;
- proposta de uma estratégia de combinação de métricas utilizando aprendizagem de máquina;
- disponibilização de avaliações humanas sobre casos de reuso de texto em artigos científicos;
- análise de casos de plágio reais e reconhecidos nos quais os artigos foram retratados; e,
- criação de uma coleção de teste com casos de plágio criados artificialmente.

1.4 Organização da Tese

O restante do texto está organizado da seguinte forma:

- O Capítulo 2 apresenta uma revisão bibliográfica que aborda os assuntos permeados nesta tese. Este Capítulo inclui algumas das definições de plágio encontradas na literatura e uma visão geral de um processo de detecção de plágio;
- O Capítulo 3 discute os trabalhos relacionados e apresenta uma análise comparativa entre eles. As abordagens pesquisadas foram organizadas em três diferentes tipos: (i) baseada em conteúdo, (ii) baseada em conteúdo e estrutura, e (iii) baseada em citações/referências;
- No Capítulo 4, é apresentada a solução proposta nesta tese para identificar reuso de texto em artigos científicos a partir da combinação de métricas baseadas em conteúdo e em refe-

rências/citações. São detalhadas as etapas que compõem a solução, como: representação dos documentos, análise de similaridade, ranqueamento dos documentos e combinação das métricas.

- No Capítulo 5, é realatada uma série de experimentos que avaliam o método proposto. Esses experimentos envolvem duas coleções reais de artigos científicos, as quais foram coletadas a partir de bases de dados de publicações científicas. Duas estratégias de combinação foram avaliadas em função da melhoria na detecção de plágio. Os resultados são comparados em relação às métricas aplicadas de forma individual.
- O Capítulo 6 apresenta casos reconhecidos de retratações de artigos científicos. Uma análise foi realizada a partir das métricas de similaridade baseadas em conteúdo e referências/citações.
- Por fim, no Capítulo 7, são apresentadas as considerações finais. As principais contribuições desta tese e os resultados obtidos. Também são discutidos alguns pontos importantes que podem ser explorados como trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo apresenta as definições de plágio encontradas na literatura e uma visão geral do processo de um sistema de detecção. Também são sumarizados alguns conceitos relacionados à aprendizagem de máquina, mais especificamente sobre classificadores. Estes são necessários para o entendimento da solução proposta nesta tese.

A Seção 2.1 descreve algumas das definições de plágio encontradas na literatura. Na Seção 2.2 pode-se ter uma visão geral do problema de detecção de plágio extrínseco, seguida pela Seção 2.3 que apresenta as fases principais do processo. Uma série de métricas que avaliam a qualidade de sistemas de detecção de plágio é especificada na Seção 2.4. A Seção 2.5 descreve alguns termos que serão necessários para o entendimento desta pesquisa. Por fim, a Seção 2.6 apresenta a definição de aprendizagem de máquina, destacando algoritmos relacionados à tarefa de classificação supervisionada.

2.1 Definições de Plágio

Plágio é uma das formas mais graves de má conduta acadêmica. Ele é definido como o ato de apropriar-se de ideias, palavras ou obras de outra pessoa sem dar crédito à fonte original (ANDERSON; STENECK, 2011; STEIN; EISSEN, 2006). A literatura cita vários tipos de plágio que podem ser classificados em cinco categorias (COLLBERG; KOBOUROV, 2005; MAURER; KAPPE; ZAKA, 2006).

- **Cópia exata:** cópia literal do original.
- **Parafraseado:** mudar palavras do texto original usando sinônimos, reordenação, ou reafirmando o mesmo conteúdo em palavras diferentes.
- **Traduzido:** traduzir o conteúdo em diferentes idiomas.
- **Auto-Plágio:** reutilizar partes ou textos inteiros de um trabalho anterior de sua própria autoria.
- **Plágio de ideias:** usar ideias de outro autor como sendo sua própria.

IEEE[1] e ACM [2] apresentam orientações relativas ao reuso de texto e plágio considerando os tipos identificados acima. Além disso, a IEEE define os seguintes cinco níveis de plágio:

- **Nível Um:** cópia literal de um trabalho completo, ou a cópia exata de uma porção grande

[1] <http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html>
[2] <http://www.acm.org/publications/policies/plagiarism_policy>

(maior que 50%), ambas sem créditos.

- **Nível Dois:** cópia literal não creditada de uma grande porção (entre 20 e 50%) do trabalho original.
- **Nível Três:** cópia literal não creditada de elementos, parágrafos, sentenças, ilustrações, etc.
- **Nível Quatro:** paráfrases indevidas não creditadas de páginas ou parágrafos.
- **Nível Cinco:** cópia literal creditada de uma parte importante de um trabalho, sem definição clara (por exemplo, aspas ou endentação).

Pode-se observar que os níveis de um a três podem ser mais facilmente detectados. O nível quatro requer a identificação de paráfrases, enquanto que o nível cinco requer detecção de plágio baseada na análise de estrutura. Os cinco níveis de análise de plágio precisam avaliar se a fonte foi devidamente referenciada.

Um nível seis poderia ser incluído na lista acima para tratar dos casos em que o plagiador visa explicitamente enganar softwares de detecção de plágio. As ações maliciosas podem consistir, por exemplo, na substituição de alguns caracteres em Latim pelo caractere equivalente no alfabeto cirílico (BEALL, 2013). Como resultado, o texto plagiado e sua fonte teriam uma semelhança muito baixa.

O *Committee on Publication Ethics* (COPE) [3] é um fórum importante para editores de revistas e editoras. O comitê define diretrizes em matéria de reciclagem de texto e recomenda um conjunto de ações (representado como fluxogramas) sobre como proceder quando há uma suspeita de plágio em um documento submetido ou publicado. Quando uma sobreposição entre um artigo submetido e outras publicações for encontrada e considerada significativa, o artigo deve ser rejeitado. Se a sobreposição for considerada pequena, os autores poderão ser convidados a reescrevê-lo. Para trabalhos publicados, se a sobreposição for considerada significativa, uma retratação do artigo pode ser necessária. Ainda assim, a decisão quanto até que ponto a reutilização de texto é tolerada fica a critério do editor do periódico.

O conceito de auto-plágio é bastante controverso, já que, como apontado em um editoral de Cronin (2013), há muitas razões legítimas para que os autores possam reutilizar textos anteriores da sua própria autoria. Por exemplo, o artigo que está sendo publicado pode ser uma versão estendida de um artigo de um Workshop. García-Romero e Estrada-Lorenzo (2014) reportaram que um fórum organizado pelo COPE foi incapaz de chegar a um consenso sobre esta matéria. Isso mostra que até mesmo especialistas humanos encontram dificuldades em determinar se e

[3] <http://publicationethics.org/>

qual o nível de reuso de texto é, de fato fraudulento.

Outro aspecto importante que influencia o julgamento que avalia se o texto reutilizado pode ser considerado plágio é a localização (ou seja, a seção no artigo) em que foi encontrada a reutilização. Geralmente, o reuso de texto na revisão da literatura ou na seção da metodologia é mais tolerado do que na seção de resultados. Este fato é mencionado por COPE, Garner (2014) e Alzahrani *et al.* (2012).

A atribuição de autoria é uma área de pesquisa relacionada à detecção de plágio em que o objetivo é determinar o autor de um texto dado com base em características, tais como vocabulário, sintaxe e estrutura. Este tópico, no entanto, está fora do escopo desta tese, uma vez que o nosso foco é na detecção de plágio extrínseco, onde um documento suspeito é comparado com ukma coleção de documentos ditos originais. Para mais informações sobre a atribuição de autoria, consulte Juola (2006).

O plágio é uma questão ética muito importante, sendo cada vez mais comum dedicar esforços para combatê-lo, assim como debates e ferramentas de detecção. O grande número de documentos para comparar, a definição vaga do que consiste em plágio e os truques praticados pelos infratores para enganar softwares de detecção contribuem para a complexidade desta tarefa, tornando-a muito desafiadora. Além disso, uma das dificuldades acrescidas em lidar com publicações reais é que julgar artigos que compartilham conteúdos como sendo casos de plágio é problemático. Assim, alguns trabalhos preferem referir-se simplesmente a “reuso de texto".

2.2 Visão Geral do Problema

O problema de detecção de plágio é comumente definido como (POTTHAST et al., 2010b): dada uma coleção de documentos D e um conjunto de casos de plágio S, *a tarefa de um sistema de detecção de plágio* é identificar um conjunto de detecções R maximizando $S \bigcap R$. Dado um documento suspeito d_{plg}, $D_{src} \subseteq D$ é o conjunto de possíveis documentos fontes selecionados a partir de D. Um caso de plágio é representado pela tupla $s = \langle s_{plg}, d_{plg}, d_{src}, s_{src} \rangle$, onde s_{plg} é uma passagem plagiada a partir do documento d_{plg} e s_{src} corresponde à passagem original no documento fonte d_{src}.

Do mesmo modo, uma *detecção de plágio* para o documento d_{plg} associa uma passagem supostamente plagiada r_{plg} em d_{plg} para r_{src} em d'_{src}, e é denotada por $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$. Consideramos que r detecta s se $r_{plg} \bigcap s_{plg} \neq \emptyset \wedge r_{src} \bigcap s_{src} \neq \emptyset \wedge d_{src} = d'_{src}$.

Nas próximas seções, discutiremos detecção de plágio em função dessas definições.

2.3 Fases do Processo de Detecção de Plágio

O processo de detecção de plágio é comumente dividido em três fases, conforme mostra a Figura 2.1. A primeira etapa é conhecida como *recuperação dos candidatos*. Nesta etapa, dado um documento suspeito d_{plg}, o objetivo é identificar a partir de D os documentos D_{src} que são as fontes prováveis para os casos de plágio em d_{plg}. Essa etapa reduz significativamente o número de candidatos uma vez que geralmente $D_{src} \ll D$.

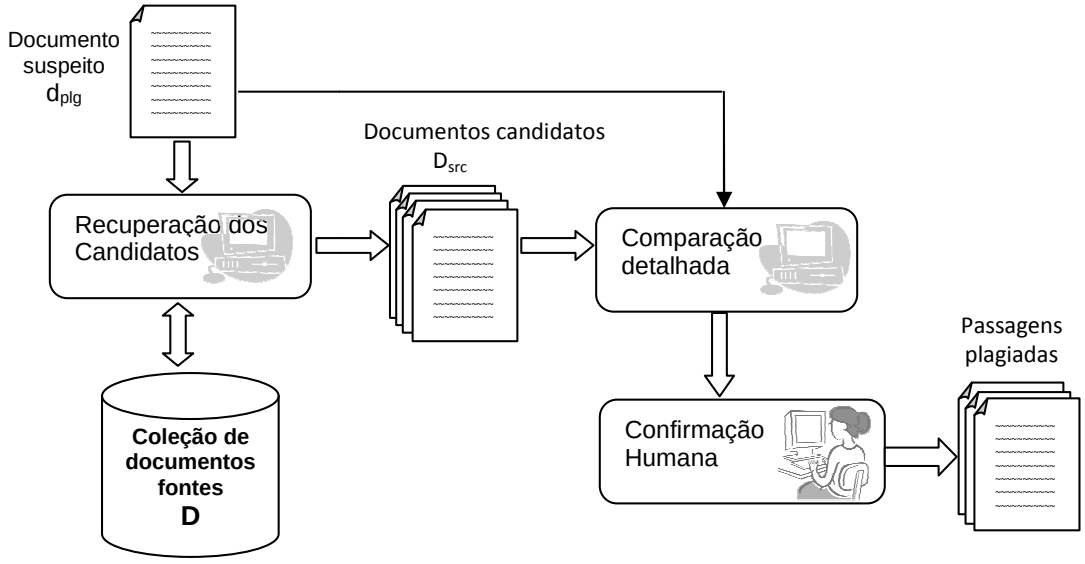


Figura 2.1 – Processo de Detecção de Plágio
Fonte: Adaptado de Stein; Eissen e Pottast (2007)

Na etapa *comparação detalhada*, d_{plg} é comparado com cada documento em D_{src} para permitir a identificação das passagens plagiadas. A terceira e última etapa exige o julgamento humano para decidir se a passagem suspeita de fato representa uma instância de plágio. No Capítulo 3, explicaremos como os vários métodos encontrados na literatura executam cada uma dessas fases.

2.4 Métricas de Avaliação

Há um crescente interesse no desenvolvimento de abordagens para auxiliar na tarefa de detecção de plágio, conforme será visto no Capítulo 3. Sendo assim, surgem juntamente a necessidade de métricas de avaliação e de padrões de referência, o que motivou a criação do *PAN Workshop Series*[4] em 2007. Desde 2009, o PAN organiza competições para ambos os tipos

[4] <http://pan.webis.de/>

de plágio, intrínseco e extrínseco. Desde a sua criação, o PAN avaliou mais de 56 sistemas de detecção de plágio extrínseco.

A fim de classificar tais sistemas, um *framework* de avaliação foi desenvolvido por Potthast (2010b). Este framework contém coleções anotadas[5] com casos de plágio artificiais. O framework define uma métrica de desempenho chamada *plagdet* para permitir a classificação de abordagens de detecção de plágio. O escore *plagdet* para um conjunto de casos de plágio, denotado por S, e um conjunto de detecções, denotado por R, é definido de acordo com a Equação 2.1. O *plagdet* é baseado em F_1 (Eq. 2.2), uma média harmônica entre a precisão (Eq. 2.3) e a revocação (Eq. 2.4). Uma vez que a precisão e revocação não levam em conta os casos em que o detector reporta múltiplas ou sobrepostas detecções para a mesma passagem, a métrica de *granularidade* (Eq. 2.5) é introduzida para punir tais casos, como se segue

$$plagdet(S, R) = \frac{F_1}{log_2(1 + granularidade(S, R))} \tag{2.1}$$

$$F_1 = \frac{2 \times precisao \times revocacao}{precisao + revocacao} \tag{2.2}$$

$$precisao(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{\left|\bigcup_{s \in S} s \bigcap r\right|}{|r|} \tag{2.3}$$

$$revocacao(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{\left|\bigcup_{r \in S} s \bigcap r\right|}{|s|} \tag{2.4}$$

onde $s \bigcap r$ é o número de caracteres sobrepostos entre um caso de plágio s e uma detecção r. $s \in S$ e $r \in R$ são conjuntos de referência aos caracteres no documento plagiado e sua fonte.

[5] <http://www.webis.de/research/corpora>

$$granularidade(S, R) = \frac{1}{S_r} \sum_{s \in S_r} |S_r| \tag{2.5}$$

2.5 Terminologia deste Trabalho

Esta subseção explica a terminologia que será utilizada ao longo desta tese:

- **Bloco de referências:** É a seção de um documento que contém a lista de referências bibliográficas citadas no texto.
- **Referência:** Uma referência aparece no bloco de referências contendo metadados bibliográficos das obras citadas no texto. A Tabela 2.1 enumera alguns exemplos comuns desses metadados, apresentando uma breve descrição para cada um deles.
- **Citação:** Encontra-se presente no corpo do texto do documento e contém informações suficientes, tais como, Autor e Ano para identificar uma referência a partir do bloco de referências do documento.

Campo	Descrição
Author	O(s) nome(s) do(s) Autor(es)
Title	Título do Trabalho
Date	Ano da Publicação
Booktitle	Título do Livro/Conferência
Institution	Instituição envolvida na publicação
Journal	Jornal ou revista onde o trabalho foi publicado
Note	Informação adicional
Pages	Números das páginas
Publisher	Nome do editor
Volume	O volume da revista

Tabela 2.1 – Metadados Bibliográficos

2.6 Aprendizagem de Máquina

Aprendizagem de Máquina (AM) é uma subárea da Inteligência Artificial concentrada no desenvolvimento de modelos que possam aprender através da experiência. Algoritmos de aprendizagem de máquina são fundamentalmente dependentes de uma fase de aprendizagem na qual, a partir de um conjunto de dados de entrada, tem a função de produzir um modelo para extração de novos conhecimentos (MONARD; BARANAUSKAS, 2003).

A AM utiliza o princípio da indução, uma forma de inferência lógica que permite obter conclusões genéricas a partir de um conjunto de exemplos. Um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. Os algoritmos de aprendizagem indutiva podem ser basicamente classificados em dois tipos: aprendizagem supervisionada e não supervisionada (WITTEN; FRANK; HALL, 2011).

- **Aprendizagem Supervisionada:** requer um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Cada exemplo é representado por um vetor de características e pelo rótulo da classe associada. Esses exemplos são então usados para aprender uma função de classificação que possa determinar corretamente a classe de novos exemplos ainda não rotulados.
- **Aprendizagem Não Supervisionada:** Distingue-se da aprendizagem supervisionada no sentido em que nenhum dado de treinamento é fornecido. Neste tipo de aprendizagem, os exemplos fornecidos são analisados para tentar determinar se alguns deles podem ser agrupados de alguma maneira.

2.6.1 Classificadores Supervisionados

O objetivo da classificação é o seguinte: a partir de um modelo aprendido rotular automaticamente novas instâncias com uma determinada classe. De acordo com a Figura 2.2, num processo de classificação, pode-se usar o conhecimento sobre o domínio para escolher os dados ou fornecer alguma informação previamente conhecida como entrada ao indutor. Após induzido, o classificador é geralmente avaliado e o processo de classificação pode ser repetido, se necessário, podendo-se adicionar outros atributos ou ajustar parâmetros durante o processo de indução (WITTEN; FRANK; HALL, 2011).

O fenômeno de interesse, ou seja, o conceito que se deseja aprender para fazer previsões a respeito. Este conjunto é composto por exemplos contendo valores de atributos bem como a classe associada.

Encontra-se na literatura um largo conjunto de algoritmos de classificação. Estes algoritmos podem ser organizados em diferentes tipos, de acordo com as características utilizadas no processo de aprendizagem.

Classificadores bayesianos são modelos estatísticos que classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe. Produz resultados rapidamente, de grande correção quando aplicados a grandes volumes de dados,

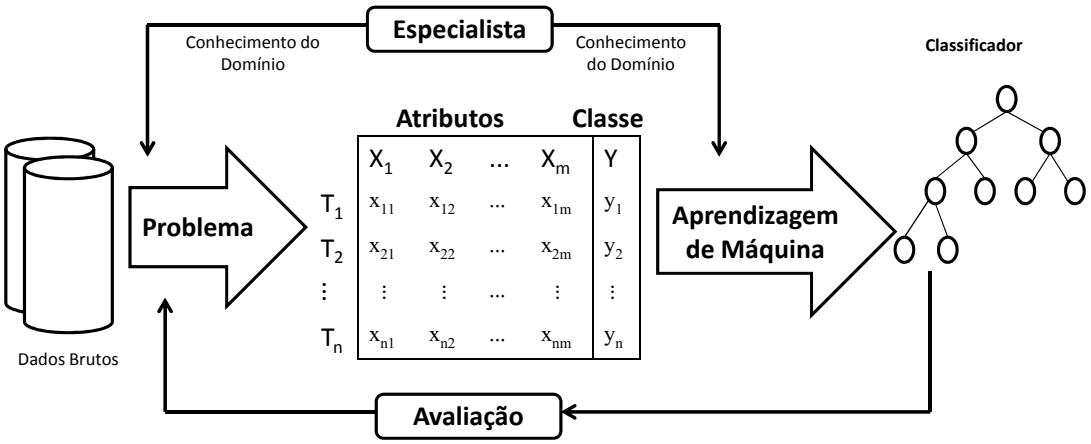


Figura 2.2 – Processo de Classificação
Fonte: Adaptado de Monard e Baranauskas (2003)

comparáveis aos resultados produzidos por árvores de decisão e redes neurais. Um classificador bayesiano simples (*Naive Bayes* (JOHN; LANGLEY, 1995)) é considerado ingênuo (*naïve*) por supor que os atributos de entrada sobre uma determinada classe são totalmente independentes dos valores dos outros atributos. Já um classificador de crença (*Bayes Net* (COOPER; HERSKOVITS, 1992)) assume em seu modelo que existem relações de dependência condicional entre os atributos de entrada. Uma rede bayesiana é representada por um grafo acíclico direcionado e uma Tabela de probabilidades condicionais para cada nó. Os nós da rede são os atributos de entrada e as arestas correspondem às dependências entre eles.

Entre outros classificadores baseados em redes neurais artificias, destacam-se o *Multilayer Perceptron (MLP)* (HAYKIN, 1998) e *RBF Network* (BROOMHEAD; LOWE, 1988). O MLP é uma rede constituída por um conjunto de nós de origem, os quais formam a camada de entrada (*input layer*), uma ou mais camadas de nós ocultos (*hidden layers*), e uma camada de saída (*output layer*). O algoritmo *backpropagation* (HECHT-NIELSEN, 1989) é usado para treinamento da rede. *RBF Network* é uma rede neural artificial que usa funções de base radial como funções de ativação.

Um outro tipo de classificador baseia-se em funções matemáticas. *Support Vector Machine - SVM* é um algoritmo de classificação binária que, usando um hiperplano busca encontrar a maior margem de separação possível que separe os vetores de duas classes. Quando os dados não são linearmente separáveis, o espaço de entrada é transformado aplicando uma função de núcleo que eleva o número de dimensões até que seja encontrado um espaço passível de separação linear. A *Sequential Minimal Optimization - SMO)* (PLATT, 1998) é uma forma analítica de resolver o problema de otimização da programação quadrática. O SMO permite trabalhar com problemas de otimização sem a necessidade de trabalhar com matriz das funções de núcleo, uma das partes mais custosas do SVM.

Algoritmos baseados em árvores e regras de decisão são utilizados para representar, através de expressões, o que é aprendido sobre os atributos dados. Em uma árvore de decisão o objeto é classificado seguindo o caminho da raiz da árvore até uma folha, enquanto as suas características satisfazem os nodos e suas ligações (WITTEN; FRANK; HALL, 2011). Entre os classificadores baseados em árvore de decisão temos o *J48* (QUINLAN, 1993) e *RandomTree*. Regras de decisão podem ser geradas a partir de um conjunto de dados, utilizando um algoritmo de indução de regras, ou com base nos conceitos aprendidos pela árvore. Entre os algoritmos de regras de decisão conhecidos estão o *ConjunctiveRule* e *DTNB* (HALL; FRANK, 2008).

Entre os tipos de algoritmos aqui apresentados temos os de meta-aprendizagem. Meta-aprendizagem consiste em combinar algoritmos para melhorar o desempenho de um determinado processo por meio do uso de informações coletadas sobre o próprio processo de aprendizagem. Classificadores como *AdaBoost* (FREUND; SCHAPIRE, 1995), *LogitBoost* (FRIEDMAN; HASTIE; TIBSHIRANI, 1998) e *Bagging* (BREIMAN, 1996) são considerados de meta-aprendizagem.

É sabido que certos tipos de classificadores funcionam melhor ou pior para um dado problema – não existe um algoritmo de classificação que obtenha bons resultados em todas as tarefas. Desta maneira, é preciso estabelecer através de testes quais os melhores classificadores para a tarefa de detecção de plágio.

2.7 Sumário do Capítulo

Neste capítulo, foram abordadas as definições de plágio encontradas na literatura. Além de uma visão geral do problema, foram descritas as três etapas que compõem um processo de detecção plágio: recuperação dos candidatos, comparação detalhada e a análise final que deve ser feita por um avaliador humano. As principais métricas utilizadas para avaliação de qualidade dos métodos de detecção de plágio foram então apresentadas. Ainda, foram descritos conceitos e algoritmos relacionados à Aprendizagem de Máquina, os quais foram divididos em 6 categorias: baseados no teorema de Bayes, funções matemáticas,M redes neurais, meta-aprendizagem, regras e árvore de decisão. No próximo capítulo, serão descritos conceitos e métodos de detecção de plágio, bem como uma análise comparativa entre eles.

3 TRABALHOS RELACIONADOS

Neste capítulo, discutiremos uma série de trabalhos encontrados na literatura desenvolvidos para auxiliar na detecção de plágio extrínseco. A Figura 3.1 mostra os métodos pesquisados organizados de acordo com o seu tipo: (*i*) baseados em conteúdo; (*ii*) baseados em conteúdo e estrutura; e (*iii*) baseados em citações e referências. A Seção 3.1 discute uma seleção de trabalhos que buscam detectar plágio utilizando apenas técnicas baseadas na similaridade do conteúdo dos documentos. Os trabalhos relatados nas seções 3.2 e 3.3 propõem métodos para analisar plágio em artigos científicos. Enquanto a Seção 3.2 descreve trabalhos que consideram que a combinação do conteúdo e a estrutura do documento podem ser um bom indicativo de plágio, a Seção 3.3 destaca alguns trabalhos que utilizam apenas as referências e citações do documento. Por fim, a Seção 3.4 sintetiza e compara as principais características dos trabalhos analisados entre os três grupos.

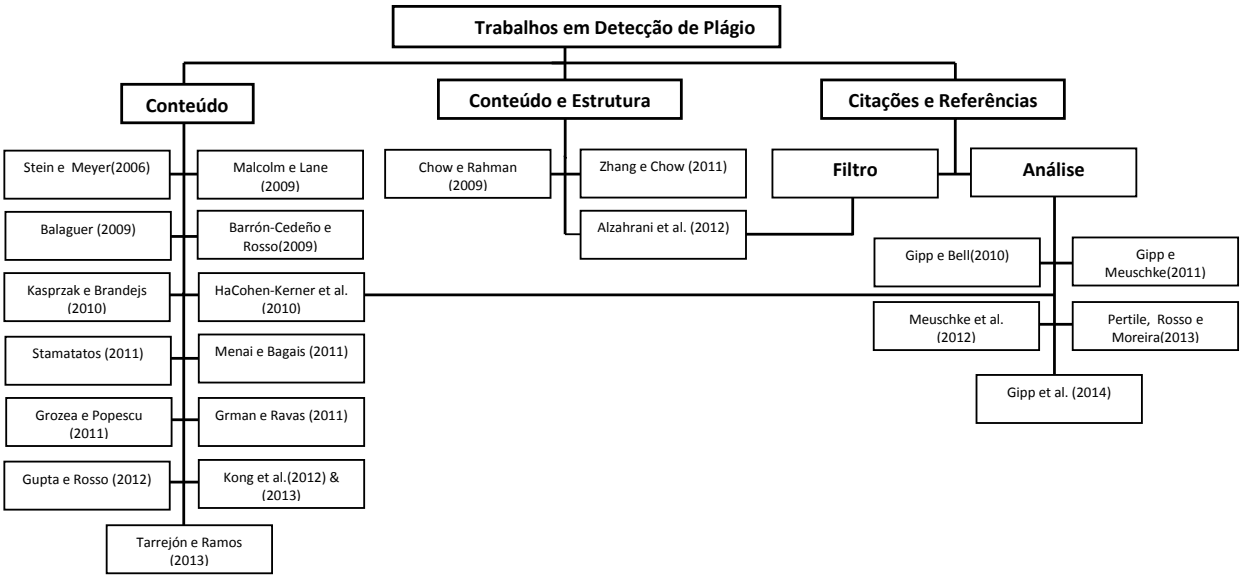


Figura 3.1 – Taxonomia de Métodos de Detecção de Plágio

3.1 Detecção de Plágio baseada em Conteúdo

A detecção baseada em conteúdo é a técnica mais utilizada para a identificação de plágio. Ela geralmente consiste em comparar fragmentos de texto (como, parágrafos, frases, palavras, n-gramas de palavras ou de caracteres) dos documentos suspeitos contra possíveis fontes. Geralmente são usadas três estratégias para a detecção com base em conteúdo, são elas: *bag-*

of-words, n-gramas, e *fingerprints*.

Bag-of-words são usadas principalmente para a etapa de recuperação de candidatos. Como o nome sugere, esta estratégia desconsidera a ordem das palavras. Consiste na aplicação de um Sistema de Recuperação de Informações para recuperar documentos que compartilham palavras com o documento suspeito. A recuperação pode ser baseada no modelo de espaço vetorial (VSM), por exemplo. Neste modelo, os documentos são representados como um vetor n-dimensional contendo a força da associação entre o documento e cada um dos n termos distintos na coleção. A similaridade entre um documento suspeito e as possíveis fontes pode ser calculada como o co-seno dos seus vetores. A recuperação dos candidatos utilizando uma abordagem *bag-of-words* foi usada em vários trabalhos (TORREJÓN; RAMOS, 2013; KONG et al., 2012; KONG et al., 2013; SANCHEZ-PEREZ; SIDOROV; GELBUKH, 2014).

A partir de 2012, a competição do PAN considerou um cenário mais realista, em que documentos originais são recuperados a partir da Web. A API ChatNoir (POTTHAST et al., 2012) foi disponibilizada para realizar a recuperação dos documentos candidatos para cada consulta. Esta API é um mecanismo de busca que indexa toda a parte em inglês do corpus ClueWeb09 [1] (isto é, uma parte estática rastreada da Web em 2009). Os melhores resultados no PAN-12 foram obtidos pelos autores Kong *et al.* (2012). Consultas foram compostas tomando os principais termos classificados por TF-IDF. Os top-n documentos recuperados são selecionados para a etapa de comparação detalhada, que é feita a nível de sentença. Pares de sentenças são computados por duas funções de similaridade: cosseno e uma variação do coeficiente de Dice. Pares com escore de similaridade superior a limiares pré-definidos para ambas as funções são submetidos a um algoritmo chamado *Bilateral Alternating Sorting*, baseado na formação de grupos alternados de sentenças adjacentes entre o documento suspeito e o documento original. Este algoritmo está descrito em detalhes em Kong *et al.* (2013).

N-gramas em comum é a abordagem predominante em detecção baseada em conteúdo. Um n-grama é uma sequência de n palavras consecutivas. A intuição é que quanto mais n-gramas em comum um par de documentos apresentar, mais eles são semelhantes. Uma série de estudos tem tentado estabelecer o valor para n que produza o melhor resultado. Os autores Barrón-Cedeño e Rosso(2009) consideram que textos independentes apresentam uma pequena quantidade de n-gramas em comum. Sendo assim, o principal objetivo da abordagem é identificar o melhor valor de n para os n-gramas a serem usados na tarefa de detecção de plágio. Uma comparação dos conjuntos de n-gramas das sentenças suspeitas e os documentos originais é realizada. A sentença é considerada candidata se o número de n-gramas for maior que um

[1] <lemurproject.org/clueweb09/>

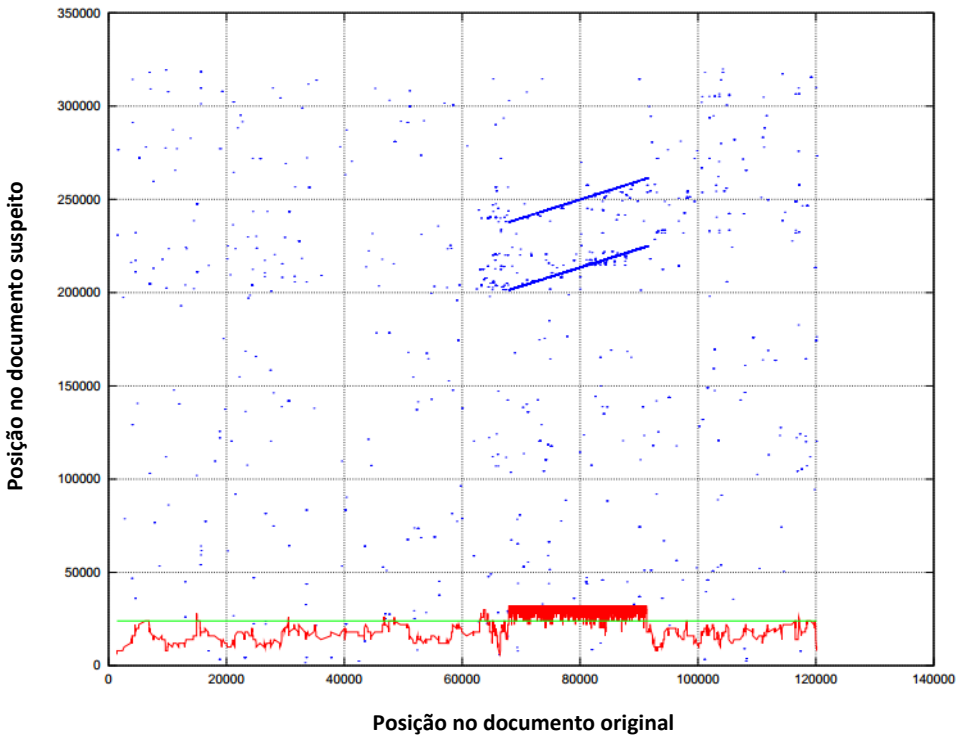


Figura 3.2 – Representação do par de documentos pelos seus n-gramas compartilhados
Fonte: Grozea and Popescu (2012)

determinado limiar. Experimentos com a coleção METER[2] mostraram que melhores resultados em termos de F1 são obtidos por bigramas e trigramas. Enquanto que bigramas alcançaram melhor revocação e trigramas a melhor precisão. Em outro estudo, (BALAGUER, 2009) foram testados valores de n entre 4 e 6 sobre o corpus PAN-09. Os autores defendem o uso de 6-gramas, uma vez que alcançou uma melhor precisão com uma pequena perda de revocação. 6-gramas também foi utilizado por Gupta e Rosso (2012).

Para o PAN-11, Grozea and Popescu (2011) introduziram uma nova medida de similaridade, na qual o conjunto de n-gramas correspondentes são computados para um par de documentos, seguido por uma projeção desse conjunto no eixo que representa o documento original (Figura 3.2). Então, é realizada uma contagem do número de vezes que uma janela móvel de tamanho fixo (256 caracteres) contém um certo número de n-gramas correspondentes acima de um determinado limiar (64). Esta contagem é considerada como sendo a similaridade entre dois documentos. Uma vez que a matriz de similaridade é obtida, os documentos são classificados com base nesses valores.

[2] <http://nlp.shef.ac.uk/meter/>

Para comparação detalhada dos documentos candidatos, num primeiro momento, o conjunto de n-gramas correspondentes entre dois documentos são obtidos. Em seguida, os n-gramas são agrupados em passagens correspondentes.

O sistema CoReMo participou em várias versões do PAN. Em 2010, Torrejón e Ramos (2010) aplicaram uma abordagem para gerar os n-gramas com base no contexto das sentenças. A principal ideia é representar o contexto com um número reduzido de caracteres (remoção de *stopwords*, *stemming*, remoção de palavras de um caractere, ordenação alfabética das palavras internas ao n-grama, utilização de unigramas contextuais consecutivos compartilhados).

No PAN-13, Torrejón e Ramos (2013) foram os vencedores da fase de comparação detalhada. Os autores estenderam o conceito de n-gramas de contexto para n-gramas de contexto próximo (*SCnG - Surrounding Context N-grams*). Este novo conceito visa identificar na sentença qual palavra foi removida, alterada ou incluída a partir da sentença original. Foi considerado $n = 3$, onde os SC3G são obtidos a partir de 4-gramas, dos quais se elimina a primeira, segunda, penúltima e a última palavra para formar até 4 n-gramas: por eliminar da primeira e última palavra (SC3G direto, ou simplesmente CTnG), ignorar a segunda palavra (SCnG direito), e por ignorar a a penúltima palavra (SCnG esquerdo). O tratamento dos n-gramas nesta abordagem conseguiu obter melhores resultados em relação a outras propostas.

O método desenvolvido por Stamatatos (2011) é baseado exclusivamente em n-gramas de *stopwords*. O objetivo é encontrar os n-gramas de *stopwords* comuns entre o documento suspeito e o original. Os documentos são representados unicamente pelo aparecimento de uma lista pré-definida de *stopwords* no texto. Essa lista contém as 50 palavras mais frequentes do idioma inglês fornecida pelo *British National Corpus*, que inclui cerca de 90 milhões de termos. Para evitar a atribuição de uma similaridade elevada em passagens independentes que contenham *stopwords* muito frequentes, uma restrição é adicionada à etapa de recuperação dos candidatos para descartar pares de documentos que apresentem esta característica. Os n-gramas de *stopwords* são ordenados de acordo com a sua primeira aparição no documento. O processo de transformar uma sentença em um conjunto de n-gramas de *stopwords* é demonstrado na Figura 3.3.

Os experimentos foram realizados com a coleção do PAN-10 e apresentaram melhores resultados em comparação com outras abordagens em casos de plágio simulado e plágio artificial com alta ofuscação (ou seja, casos de plágio em que há tentativa de disfarçar a ofensa com técnicas como paráfrase, por exemplo). Além disso, a abordagem mostrou que o comprimento da passagem plagiada afeta os resultados, casos de passagens longas (>10K caracteres) são mais fáceis de detectar.

These savage birds are very common in Maine, where they make great havoc among the flocks of wild-ducks and Canada grouse, and will even, when driven by hunger, venture an attack on the fowls of the farm-yard.

(a) Uma passagem de texto

are in they the of and and will by an on the of the

(b) O texto após a remoção dos termos que não constam na lista de *stopwords*

[are, in, they, the, of, and, and, will]
[in, they, the, of, and, and, will, by]
[they, the, of, and, and, will, by, an]
[the, of, and, and, will, by, an, on]
[of, and, and, will, by, an, on, the]
[and, and, will, by, an, on, the, of]
[and, will, by, an, on, the, of, the]

(c) 8-gramas de *stopwords* do texto

Figura 3.3 – Representação de uma sentença pelos seus n-gramas de *stopwords*
Fonte: Stamatatos (2011)

Fingerprints são representações compactas de documentos que visam capturar sua identidade. Para um algoritmo de *fingerprints*, a possibilidade de gerar a mesma representação de diferentes documentos deve ser insignificante. A fim de fazer isso, os documentos são separados em blocos e uma função *hashing* é aplicada a cada um dos blocos que produzem um valor inteiro.

O algoritmo *MD5 hash* (RIVEST, 1992) é uma das propostas mais populares para detecção de plágio de cópia exata (ou seja, fragmentos de texto idênticos). Stein e Meyer (2006) propuseram uma melhoria significativa para o algoritmo MD5, as *fuzzy fingerprints*. O objetivo é gerar o mesmo código *hash* para fragmentos similares. Esta proposta baseia-se em uma representação do modelo de espaço vetorial de n-gramas de palavras. Termos de índice são combinados em um pequeno número de classes equivalentes, de tal modo que todos os termos da classe comecem com o mesmo prefixo. Um vetor de frequências relativas (pf) de cada classe de prefixo é calculado para os índices de termos no bloco. Então, um vetor de desvios relativos de pf é computado. Este desvio é abstraído em um esquema de desvio *fuzzy*, e valores de *hash* são computados de acordo com a Eq. 3.1.

$$\gamma = \sum_{i=0}^{k-1} \delta_i . r^i, \;\; com \delta_i \in \{0, ..., r-1\} \tag{3.1}$$

onde k é o número de classes de prefixo, e δ_i é o desvio *fuzzy* da frequência da classe

de prefixo i. A *fuzzy fingerprint* é a união dos valores de *hash* para um n-grama. O objetivo do estudo foi investigar o desempenho em tempo de execução e a diferença entre os escores de *fuzzy fingerprints* e de cosseno sob o modelo de espaço vetorial. Os autores concluíram que *fuzzy-fingerprints* se assemelham mais à similaridade do cosseno do que a *MD5 fingerprint*. Esta comparação foi sobre documentos selecionados a partir da coleção RFC[3].

Hoad e Zobel (2003) fornecem uma análise detalhada de *fingerprints* e as comparam com uma medida de identidade (e suas variações) que eles propõem. Seus experimentos mostraram que as medidas de identidade propostas superaram métodos de *fingerprints*. Os autores também notaram que *métodos baseados em âncora* alcançam bons resultados. Uma âncora é uma cadeia (ou um n-grama) no texto do documento. Âncoras devem ser escolhidas de modo que haja pelo menos uma em cada documento, mas não tão comuns que a *fingerprint* torne-se demasiado grande.

Na abordagem de Kasprzak e Brandejs (2010), o texto é dividido em palavras. Um *MD5 hash* é calculado a partir de 5-gramas de palavras em comum e então um bloco é representado pelos 30 bits mais significativos do *hash*. A similaridade entre pares de documentos é calculada a partir de seus blocos comuns. Pares de documentos que contêm 20 ou mais blocos comuns são selecionados como candidatos. Para cada par, o método analisa se os blocos comuns formam um ou mais intervalos válidos, onde o intervalo entre dois blocos comuns vizinhos não é maior do que 50 blocos. Esses intervalos válidos são então relatados como passagens plagiadas. Uma modificação na fase de pós-processamento (em relação à apresentada em Kasprzak *et al.* (2009), que mantinha apenas a maior passagem do conjunto que se sobrepõem), foi a remoção de passagens sobrepostas que estivessem abaixo de um dado limiar (600 caracteres). Este método obteve o melhor desempenho no PAN-10.

Mais recentemente, pesquisadores têm proposto métodos que visam detectar plágio em artigos científicos. Hacohen-Kerner *et al.* (2010) descrevem uma variedade de métodos que foram desenvolvidos para identificar similaridade em artigos. Estes métodos foram aplicados a uma coleção de 10100 artigos acadêmicos selecionados a partir da *ACL Anthology*[4]. O artigo relata quantos pares de documentos foram considerados semelhantes por cada método. Eles concluíram que *full-fingerprints* foi considerado o melhor método.

Gupta e Rosso (2012) analisaram as tendências de reuso de texto em artigos científicos disponíveis na coleção ACL. A ferramenta WCopyFind foi usada para executar esta análise, onde os autores consideraram que documentos que excederem um limiar de 500 6-gramas de palavras contêm reuso de texto. Esta investigação centrou-se apenas no reuso de texto literal.

[3] <http://www.rfc-editor.org/>
[4] <http://aclweb.org/anthology/>

Os resultados apresentaram altos níveis de reuso de texto, e mostraram que o auto-plágio é mais frequente do que reuso de outros autores.

Lidar com paráfrases é uma característica importante para a identificação de casos de plágio em que o infrator tenta disfarçar a duplicação. Sistemas que lidam com paráfrases tiveram um bom desempenho no PAN.

O sistema desenvolvido por Grman e Ravas (2011) foi o vencedor da competição PAN-11. O método baseia-se em calcular o número de palavras correspondentes para um par de passagens entre os documentos suspeitos e originais. Em primeiro lugar, os pares de passagens em que o número de palavras correspondentes excede um determinado limiar, são selecionados. WordNet [5] então é usado como uma fonte de sinônimos. O uso de sinônimos e o desrespeito pela ordem das palavras auxiliam na detecção de plágio parafraseado e traduzido. Para uma análise mais profunda do impacto de análise de paráfrases em sistemas de detecção de plágio, pode-se consultar (BARRÓN-CEDEÑO et al., 2013).

3.2 Detecção de Plágio baseada em Conteúdo e Estrutura

Os métodos discutidos na Seção anterior são baseados apenas no texto do documento. No entanto, algumas abordagens usam informações sobre a estrutura do documento (CHOW; RAHMAN, 2009; ZHANG; CHOW, 2011; MENAI; BAGAIS, 2011; ALZAHRANI et al., 2012).

Em alguns tipos de documentos, como artigos científicos, a estrutura desempenha um papel importante. A informação estrutural é representada por cabeçalhos, seções, parágrafos, referências, etc. De acordo com Alzahrani *et al.* (2012), os métodos de particionamento de publicações científicas consideram que a estrutura é geralmente apresentada por elementos visuais, tais como localização, posição, pontuação, comprimento, tipo e tamanho da fonte. Alguns métodos empregam estratégias com base em palavras-chave para rotular conteúdos específicos, por exemplo, usando palavras como Capítulo, introdução e cabeçalhos de Seção (BURGET, 2007; STOFFEL et al., 2010).

Representar documentos como árvores é uma alternativa para lidar com a sua estrutura. Para a detecção de plágio, no trabalho de Chow e Rahman (2009), documentos no formato HTML são representados como uma estrutura de árvore em três camadas (documentos-páginas-parágrafos). Primeiramente, um documento é segmentado em um número de parágrafos. Os

[5] <http://wordnet.princeton.edu/>

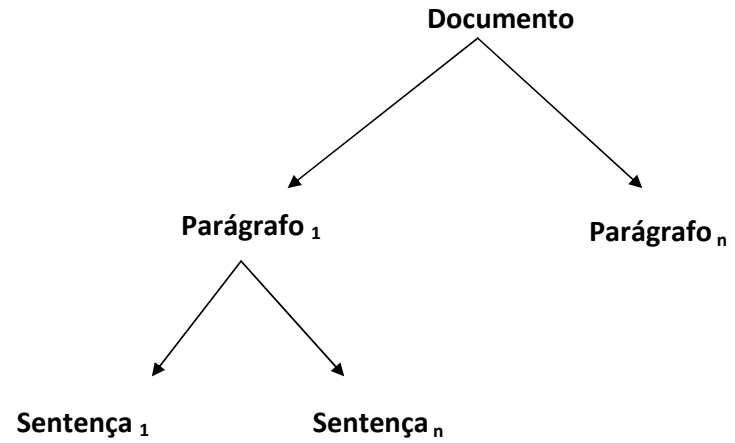


Figura 3.4 – Estrutura da representação do documento no APlag
Fonte: Adaptado de Menai e Bagais (2011)

parágrafos são então concatenados para criar uma página. Se o número de palavras em uma página ultrapassar um determinado limiar então uma nova página é criada. A camada superior da árvore é utilizada para a recuperação dos documentos candidatos, executando o agrupamento de documentos. A camada inferior é utilizada para identificar passagens plagiadas usando a medida de similaridade do cosseno. Mais tarde, os mesmos autores (ZHANG; CHOW, 2011) representaram os documentos em níveis de documento-parágrafo-sentença para a detecção de plágio. Experimentos em uma coleção de 10K documentos HTML[6] foram relatados em ambos os trabalhos, apresentando bons resultados. No entanto, não houve resultados reportados em coleções padrão para a análise de plágio.

Menai e Bagais (2011) descrevem uma novo sistema de detecção de plágio para textos escritos em árabe, onde cada documento é representado em estrutura de árvore (Figura 3.4). A comparação é realizada nível por nível da raiz às folhas. O sistema Aplag é baseado na segmentação de palavras. Primeiramente as palavras são segmentadas em cada sentença do documento, seu *hash* é calculado. Se dois documentos apresentam um número comum de *hashes* acima de um limiar fixo, então o par de documentos apresenta uma provável similaridade, e a análise então segue para nível de parágrafo, caso contrário, o processo é interrompido. Se for detectada uma similaridade no nível de parágrafo, o processo continua a nível da sentença. Se houver uma possível similaridade entre duas sentenças, o algoritmo *Longest Common Substring* (LCS) é usado. O algoritmo LCS consiste em encontrar a sequência comum mais longa de duas strings. Duas sentenças são consideradas correspondentes, se o comprimento do LCS for maior do que um dado limiar. Os experimentos relatam bons resultados, no entanto, eles foram realizados sobre um conjunto muito pequeno de textos em árabe (12 textos).

[6] <http://www.ee.cityu.edu.hk/~twschow/Html_CityU1.rar>

Considerar a estrutura de um artigo científico foi fundamental para a abordagem proposta por Alzahrani *et al.* (2012). Partindo dessa representação, os autores consideram que um caso de plágio nos componentes *introdução* e *definições* é menos importante comparado com um caso de plágio nos componentes *avaliação* ou *discussão*. A Figura 3.5 mostra um artigo científico com seus diferentes componentes na primeira e última página. Esses componentes são extraídos e classificados em classes genéricas. Por exemplo, a Seção Metodologia de um artigo científico, normalmente expande-se em vários componentes estruturais, e um classificador genérico deve ser capaz de agrupar estes componentes sob a mesma classe. Para as fases de extração, rotulação e classificação dos componentes em suas classes, os autores utilizaram a ferramenta *SectLabel* (KAN; LUONG; NGUYEN, 2010). A abordagem introduz diferentes funções para computar o fator de peso dos componentes para representação dos documentos. Para computar tais funções automaticamente, foram aplicadas três medidas estatísticas, as quais são descritas a seguir:

- **Inverse generic class frequency (IGF):** Dado que uma classe G genéria tem $NC_1, C_2, ..., N_C$ componentes, e um termo t que ocorre em N_{tC} componentes de G, o IGF de um termo t em C é definido como (Eq. 3.2):

$$IGF(t, G) = log\frac{N}{N_{t,C}} \tag{3.2}$$

- **Spread:** A *Spread* de um termo t em um artigo A é o número de componentes estruturais em A que contêm t. De acordo com a Eq. 3.3, esta função pode ser vista como a frequência estrutural de um termo.

$$Spread(t, A) = \sum^{C \in A} i \; onde \; i = \left\{ \begin{array}{c} 1 \; se \; t \; \in \; C \\ 0; \; caso \; contrário \end{array} \right\} \tag{3.3}$$

- **Depth:** O *Depth* de um termo t em uma classe genérica G refere-se à frequência de t em G ($TF_{t,G}$) normalizada pela frequência máxima em G ($MAX_{t',G}$) (Eq. 3.4).

$$Depth(t, G) = \frac{TF_{t,G}}{MAX_{t',G}} \tag{3.4}$$

Um caso de plágio é identificado quando um componente a partir de um documento

suspeito recebe um escore de similaridade alto em relação a um componente a partir de um documento original. Os experimentos foram realizados com uma coleção artificial composta por 15412 artigos científicos. Deste conjunto, 8657 são usados como documentos originais e 6755 como suspeitos. O conjunto de artigos foi obtido a partir de várias revistas listadas no *Directory of Open Access Journals (DOAJ)*[7] e de anais de conferências disponíveis na Web. A coleção está disponível a partir de: <http://www.c2learn.com/plagiarism/corpus/v1/>. Os resultados mostraram que a utilização de informação estrutural pode contribuir tanto para a etapa de recuperação dos candidatos como para a detecção plágio.

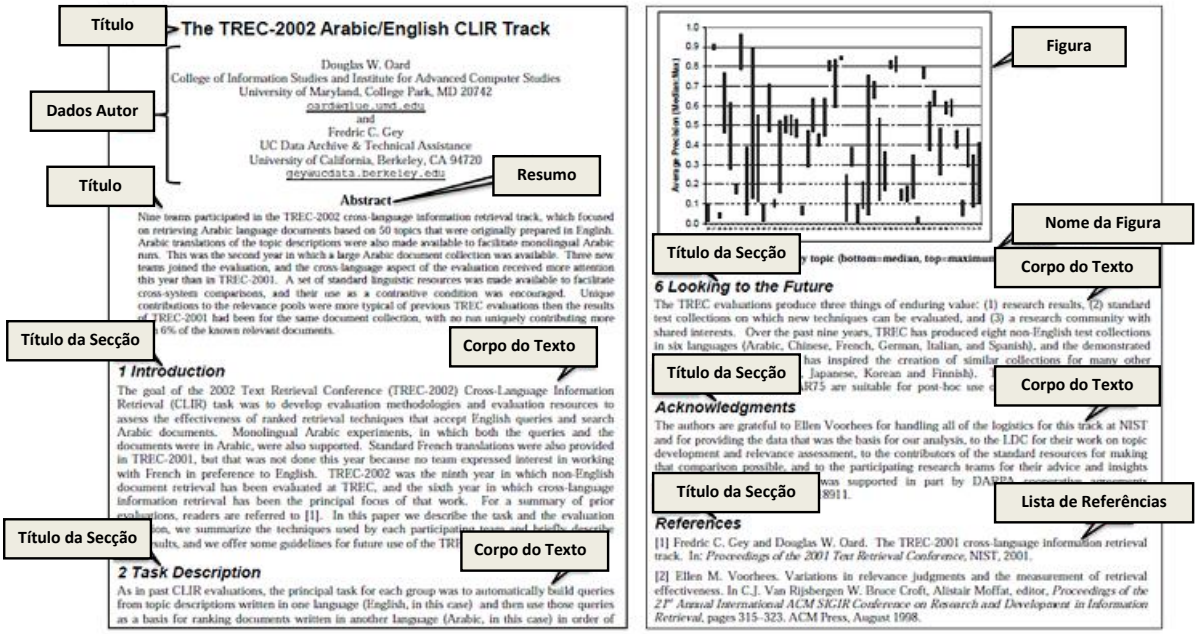


Figura 3.5 – Componentes da primeira e última página de um artigo científico
Fonte: Adaptado de Alzahrani *et al.* (2012)

3.3 Detecção de Plágio Baseada em Citações e Referências

Estratégias que analisam citações e referências podem ser usadas de duas formas diferentes por sistemas de detecção de plágio: como filtro (ALZAHRANI et al., 2012; SOROKINA et al., 2006) (isto é, verificar se a citação está dando crédito à fonte original); ou examinando citações/referências similares entre documentos (HACOHEN-KERNER; TAYEB; BEN-DROR, 2010; GIPP; BEEL, 2010; GIPP; MEUSCHKE, 2011; MEUSCHKE; GIPP; BREITINGER, 2012; GIPP; MEUSCHKE; BREITINGER, 2014; PERTILE; ROSSO; MOREIRA, 2013).

[7] <http://www.doaj.org/>

O uso de análise de citação como filtro para descartar falsos positivos foi a estratégia usada por Sorokina *et al.* (2006) e Alzahrani *et al.* (2012). Eles supõem que, se o autor de um artigo A aparece na lista de referências do artigo B, então isso poderia ser um caso de " plágio suave ", uma vez que os plagiadores normalmente não citam suas fontes. A desvantagem é que descartar grandes porções de passagens idênticas (mesmo com referências apropriadas) significa que os casos de plágio Nível 5 (ver Seção 2.1) iriam passar despercebidos. Os experimentos focaram principalmente na análise de plágio baseado nas informações estruturais dos documentos, sendo assim, os autores não analisam os ganhos/perdas trazidos por essa filtragem.

O uso de análise de citação como uma fonte de similaridade foi introduzido por Gipp e Beel (2010), Gipp e Meuschke (2011), Meuschke *et al.* (2012), Gipp (2013) e Gipp *et al.* (2014). A seguir, são descritos três algoritmos propostos para análise de similaridade entre padrões de citação (GIPP; MEUSCHKE, 2011; MEUSCHKE; GIPP; BREITINGER, 2012). Os autores definem padrões de citação como sequências de citações compartilhadas entre dois documentos.

- ***Longest Common Citation Sequence (LCCS):*** é uma adaptação do algoritmo tradicional para medir a similaridade de duas strings. O LCCS consiste do número máximo de citações que se pode tomar a partir uma sequência de citações sem alterar sua ordem.
- ***Greedy Citation Tiling (GCT):*** é uma adaptação da conhecida medida de similaridade para uma sequência de texto (*Greedy String Tiling*). O GCT foca na comparação exata na sequência de citações.
- ***Citation Chunking:*** é um conjunto de procedimentos heurísticos que identificam o local dos padrões de citação.

Os autores consideram que documentos mais citados são mais propensos a aparecer em um padrão de citação correspondente. Sendo assim, também propõem a função *Citing Frequency-Score*, derivada de um escore a partir da análise da frequência de citações. Para as abordagens citadas anteriormente, os documentos são representados apenas pelo seu padrão de citação, conforme ilustra a Figura 3.6.

Gipp e Meuschke (2011) também consideram que dois documentos que compartilham um número de referências em comum (Acoplamento Bibliográfico) podem apresentar um grau de similaridade. Quanto mais referências um par de documentos compartilha, mais similares eles são. Estas funções de similaridade serão apresentadas com mais detalhes no Capítulo 5.

Os autores relatam experimentos em três conjuntos de dados: (i) a tese de doutorado de um ex-Ministro da Defesa da Alemanha, que foi reconhecida como plagiada de diversas

fontes; (ii) uma investigação que utiliza *crowd-sourcing* contínua de acusações de plágio em teses alemãs; e (iii) os documentos extraídos do subconjunto de acesso aberto da PubMed[8]. Seus resultados indicam que a detecção baseada em citação supera a detecção baseada na análise de conteúdo em casos de plágio fortemente disfarçado.

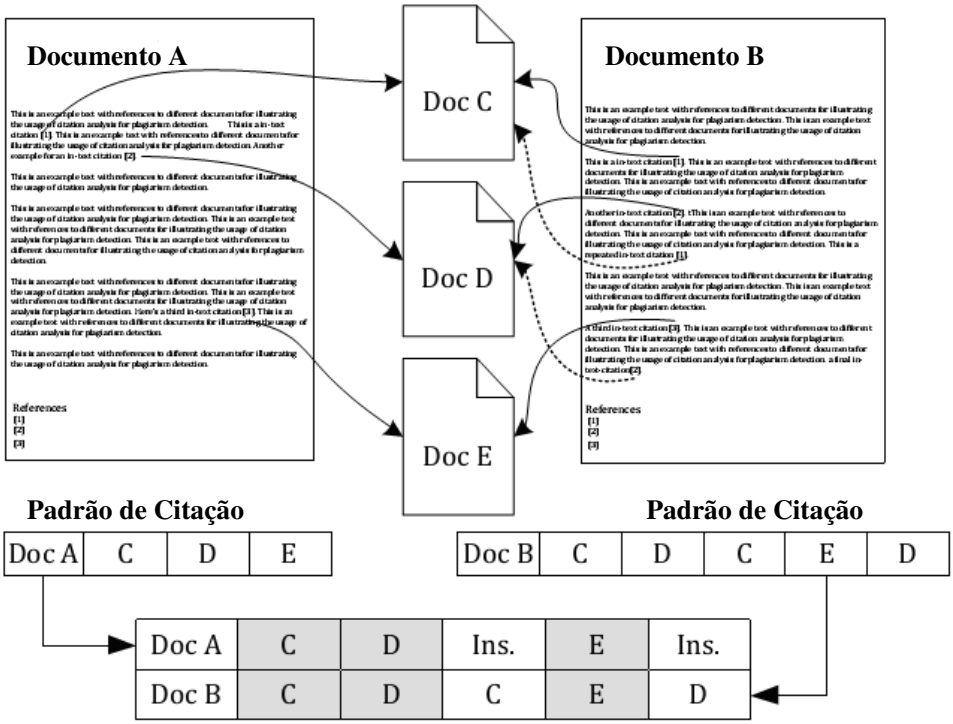


Figura 3.6 – Comparação de Padrões de Citações

Fonte: Adaptado de Meuschke *et al.* (2012)

Em nosso trabalho anterior (PERTILE; ROSSO; MOREIRA, 2013), o objetivo foi comparar a similaridade de trabalhos científicos com base na análise de coocorrências de citações. Nossa hipótese era de que uma alta taxa de coocorrências entre pares de citações pudesse ser uma indicação de plágio. Os resultados experimentais mostraram que a maioria dos casos com coocorrências em citações correspondem a plágio. No entanto, a coleção usada nos experimentos foi criada artificialmente usando casos simulados de plágio, portanto, não poderíamos tirar conclusões sobre o que acontece nas coleções reais de documentos científicos. O método e a coleção utilizada serão descritos em detalhes no Capítulo 4.3.1.

[8] <http://www.ncbi.nlm.nih.gov/pubmed>

3.4 Análise Comparativa

Nesta seção, apresentamos uma análise comparativa dos métodos de detecção de plágio descritos neste trabalho. Os critérios utilizados para esta comparação são os seguintes:

- **Proposta:** indica se a proposta baseia-se na análise de conteúdo (Co), em análise de citações (Ci), análise de conteúdo e estrutura (Co&S), ou uma combinação de todas as três.
- **Recuperação de candidatos:** técnica usada na etapa de recuperação dos candidatos.
- **Comparação Detalhada:** técnica usada na etapa de comparação detalhada.
- **Coleção:** a coleção usada nos experimentos.
- **Tipo de Documento:** tipo de documento analisado (artigo científico ou texto simples).
- **Avaliação (P, F1, Plagdet):** resultados da avaliação em termos de precisão, F1 e Plagdet.

De acordo com a Tabela 3.1, a maioria dos trabalhos são baseados apenas na análise de conteúdo e usam técnicas baseadas em palavras n-gramas e *fingerprints*. Apenas alguns autores (ALZAHRANI et al., 2012; GIPP, 2013; PERTILE; ROSSO; MOREIRA, 2013) têm desenvolvido propostas que exploram a análise de citações. As abordagens propostas se dizem capazes de detectar plágio, mas nenhuma delas é capaz de tratar todos os níveis de plágio relatados na Seção 2.1. Além disso, nenhuma abordagem propõe combinar a análise de conteúdo e referências/citações.

Os resultados da avaliação apresentados nas últimas três colunas da Tabela 3.1 não são comparáveis, uma vez que os experimentos foram realizados em diferentes coleções. Para permitir uma comparação justa, é necessário uma coleção de referência comum. Este é o objetivo do *PAN Workshops*, e também, dos experimentos que foram realizados neste trabalho.

Tabela 3.1 – Análise Comparativa das Propostas

Trabalho	Abordagem	Recuperação dos Candidatos	Comparação Detalhada	Coleção	Tipo de Doc	Avaliação		
						P	F1	Plagdet
Stein & Meyer (2006)	Co	fuzzy-fingerprints	N/A	RFC	plano	-	-	-
Barrón-Cedeño & Rosso (2009)	Co	n-gramas em comum (n=2 e 3)	classificação baseada em uma medida de contenção	METER	plano	0.74	0.66	-
Kasprzak & Br&ejs (2010)	Co	n-gramas em comum (n=5)	diferença entre dois n-gramas comuns vizinhos	PAN 2010	plano	0.94	0.80	0.80
Stamatatos (2011)	Co	n-gramas de stopwords em comum (n=11)	sequência de palavras n-gramas de stopwords em comum (n=8)	PAN 2010	plano	0.94	0.56	0.82
Grman & Ravas (2011)	Co	palavras correspondentes entre um par de passagens + WordNet para sinônimos	palavras correspondentes entre um par de passagens + WordNet para sinônimos	PAN 2011	plano	0.95	0.82	0.82
Grozea & Popescu (2011)	Co	matriz de similaridade com o número de n-gramas em comum dentro de uma janela	n-gramas em comum + agrupamento	PAN 2011	plano	0.81	0.48	0.42
Kong et al. (2012)	Co	ChatNoir API,TF-IDF, classificação VSM	similaridade do cosseno e Dice	PAN 2012	plano	0.82	0.73	0.73
Gupta & Rosso (2012)	Co	n-gramas em comum (n=6)	n-gramas em comum (n=6)	ACL Anthology	sci	-	-	-
Kong et al. (2013)	Co	ChatNoir API, TF-IDF, PatTree e TF-IDF ponderado	similaridade do cosseno, Bilateral Alternating Merging	PAN 2013	plano	0.83	0.82	0.82
Torrejón & Ramos (2013)	Co	sistema de RI & Reference Monotony Pruning	Surrounding Context N-grams & Odd-Even N-grams	PAN 2013	plano	0.89	0.81	0.82
Gipp (2013) Gipp (2014)	Ci	citações/referências em comum	citações/referências em comum	PubMed OAS GuttenPlag Wiki VroniPlag Wiki	sci	-	-	-
Pertile et al. (2013)	Ci	coocorrência em citação	coocorrências em citações	Azahrani	sci	0.47	0.03	-
Chow & Rahman (2009)	Co&S	Multilayer Self Organizing Map, histograma de palavras, PCA	Multilayer Self Organizing Map, histogramas de palavras, PCA	Html_CityU1	plano	-	-	-
Bachir Menai & Bagais (2011)	Co&S	fingerprints	Longest Common Substring	Texto Arábe	plano	0.93	0.96	0.96
Zhang & Chow (2011)	Co&S	Multilayer Self Organizing Map, histograma de palavras, PCA, Distância Earth Mover's	Multilayer Self Organizing Map, histograma de palavras, PCA, Distância Earth Mover's	Html_CityU1	plano	0.74	0.74	-
Alzahrani et al. (2012)	Co&S&Ci	TF-IDF ponderado para seções diferentes, similaridade do cosseno	Similaridade de Jaccard	Alzahrani	sci	-	-	-

3.5 Ferramentas de Detecção de Plágio

Ferramentas de detecção de plágio são programas de computador que comparam documentos com fontes prováveis, a fim de identificar similaridades e, assim, encontrar possíveis casos de plágio. Elas podem ser usadas para detectar e prevenir o plágio. A seguir, descrevemos várias dessas ferramentas. Muitas delas são soluções proprietárias, portanto, não há detalhes sobre os algoritmos que elas usam. A ferramenta Antiplag[9] baseia-se no vencedor da competição de detecção de plágio em PAN-2011 (GRMAN; RAVAS, 2011). Ela fornece um portal Web que consulta uma base de dados composta de trabalhos acadêmicos (dissertações de graduação, mestrado, doutorado, etc.) e publicações selecionadas a partir de outras fontes da Web. Universidades eslovacas são obrigadas a enviar todas as publicações para serem comparadas com o conteúdo de publicações da base de dados. A WCopyFind[10] é uma ferramenta desenvolvida em 2004 na Universidade da Virginia (USA). A entrada para esta ferramenta é um conjunto de documentos suspeitos e de origem. A saída apresenta os pares de documentos com pontuações

[9] <http://www.svop.sk/en/antiplag.aspx>
[10] <http://plagiarism.phys.virginia.edu/>

de similaridade maior do que um dado limiar. Esta análise baseia-se em n-gramas de palavras (n é um parâmetro que pode ser definido), e apenas as correspondências exatas são relatadas. Conforme relatado na Seção 3.1, esta foi a ferramenta utilizada por dois estudos, Gupta e Rosso (2012 e Vallés Balaguer (2009).

O Ferret[11] é também baseado em n-gramas de palavras. É um sistema de detecção de plágio gratuito desenvolvido na Universidade de Hertfordshire (Reino Unido). A pontuação de similaridade é calculada como o número de trigramas compartilhados entre dois documentos dividido pelo número de trigramas distintos no par. A versão desktop do sistema realiza comparações com todos os documentos de um conjunto de dados criado pelo usuário. A saída destaca as seções semelhantes entre cada par.

O Grammarly[12] verifica um determinado texto contra um grande banco de dados de documentos da Web. Ele destaca as seções que tenham sido previamente publicados em outros lugares. Como uma ajuda de escrita, ele sugere possíveis referências para os textos em que são encontrados similaridades.

O Turnitin[13] é um serviço baseado na Web que detecta material copiado a partir Web e também faz a verificação cruzada com documentos de texto em uma base de dados composta por documentos submetidos anteriormente para análise.

eTBLAST [14] é também um motor de busca baseado na Web que procura a literatura que corresponde exatamente ao parágrafo dado como entrada. Os bancos de dados pesquisados incluem Medline, PubMed, e arXiv.

O iThenticate[15] é um serviço privado desenvolvido pela Turnitin que visa verificar a originalidade das pesquisas para as instituições de ensino em todo o mundo. Os documentos suspeitos são confrontados com o CrossCheck[16], um grande banco de dados.

O Ephorus[17] pode ser integrado com ambientes digitais de aprendizagem. Ele também tem um banco de dados de documentos apresentados pelas instituições participantes. Nós não encontramos detalhes sobre como a similaridade entre os documentos é calculada.

EVE[18], Plagium[19], PlagScan[20], Plagaware[21], e Checkforplagiarism[22] comparam um de-

[11] <http://peterlane.info/software/ferret.html>
[12] <http://http://www.grammarly.com/>
[13] <http://turnitin.com/pt_br/>
[14] <http://etest.vbi.vt.edu/etblast3/>
[15] <http://www.ithenticate.com/>
[16] <http://www.crossref.org/crosscheck/>
[17] <https://www.ephorus.com/>
[18] <http://www.canexus.com/>
[19] <http://www.plagium.com/>
[20] <http://www.plagscan.com/>
[21] <http://www.plagaware.com/>
[22] <http://www.checkforplagiarism.net/>

terminado texto contra materiais disponíveis gratuitamente na Web.

Torrejón e Ramos 2013 desenvolveram o CoReMo[23] um sistema que permite comparar documentos suspeitos com fontes de uma coleção local ou da Web. Além disso, instituições privadas podem aderir a este sistema e indexar seus documentos como parte da coleção de documentos originais.

O sistema CitePlag[24] foi desenvolvido por Gipp (2013) e executa a detecção de plágio baseada exclusivamente em padrões de citação e referências. Este protótipo aceita o carregamento dos dois documentos que são comparados entre si. Outra opção é escolher duas publicações de documentos a partir do PubMed para serem comparadas.

A Tabela 3.2 mostra algumas características das ferramentas de detecção de plágio discutidas nesta seção. Note-se que quase todas as ferramentas focam na análise de correspondência exata e não na análise de citações e referências. Outra característica que está ausente na maioria das ferramentas é a análise de plágio parafraseado; apenas o CoReMo, CheckForPlagiarism.net e Ithenticate relatam que permitem este tipo de análise. Nenhuma das ferramentas apresentadas na Tabela 3.2 combina análise de conteúdo e citações/referências.

Tabela 3.2 – Ferramentas de Detecção de Plágio

Ferramentas	Características							
	Fonte Aberta	Gratuita	Privada	Plataforma	Análise de Citação	Análise de Conteúdo	Análise Estrutural	Plágio Parafraseado
Antiplag			✓	Web		✓		
WCopyfind	✓			Desktop		✓		
Ferret		✓		Desktop		✓		
Grammarly			✓	Web		✓		
Turnitin			✓	Web		✓		
eTBLAST		✓		Web		✓		
iThenticate			✓	Web		✓		✓
Ephorus			✓	Web		✓		
CheckForPlagiarism.net			✓	Web		✓		✓
Compilatio.net			✓	Web		✓		
DupliChecker		✓		Web		✓		
EVE2			✓	Desktop		✓		
Plagium		✓	✓	Web		✓		
PlagScan			✓	Web		✓		
PlagAware		✓	✓	Web		✓		
CoReMo			✓	Web		✓		✓
CitePlag		✓		Web	✓			

[23] <http://www.coremodetector.com/>
[24] <http://citeplag.org/>

3.6 Sumário do Capítulo

Neste capítulo, foram abordados diferentes métodos de detecção de plágio, os quais foram organizados de acordo com seu tipo: métodos baseados apenas na análise de conteúdo; baseados em conteúdo e estrutura; e baseados apenas em referências/citações. Ainda, uma análise comparativa entre os métodos foi descrita. Ferramentas de detecção de plágio encontradas na literatura também foram apresentadas neste capítulo. No próximo capítulo, serão abordadas as etapas que envolvem o método proposto nesta tese, assim como, a representação dos documentos, as métricas de similaridade computadas e as abordagens de combinação propostas.

4 COMBINANDO MÉTRICAS BASEADAS EM CONTEÚDO E CITAÇÃO

Este Capítulo apresenta em detalhes o método proposto para detectar reuso de texto em artigos científicos. A Seção 4.1 apresenta uma visão geral do método proposto e enfatiza o objetivo da tese. Nas seções seguintes, cada etapa do método é detalhada.

4.1 Visão Geral

O objetivo principal desta tese é propor uma combinação de métricas de detecção de plágio baseados em análise de conteúdo e análise de referências/citações. A Figura 4.1 apresenta uma visão geral das etapas que compõem o método proposto. A partir de um dado documento suspeito e uma coleção de documentos originais, (1) os documentos são representados pelo conteúdo e pelas referências e citações (Seção 4.2). (2) Para cada par de documentos são computadas métricas de similaridade baseadas em conteúdo, referências e citações (Seção 4.3). Os pares de documentos são ordenados em ordem decrescente de acordo com seu escore de similaridade. (3) A partir dos escores computados, as métricas então são combinadas de duas maneiras: combinação simples e utilizando aprendizagem de máquina (Seção 4.4). Por fim, uma avaliação humana é necessária para analisar se um par de documentos é de fato um caso de plágio ou não.

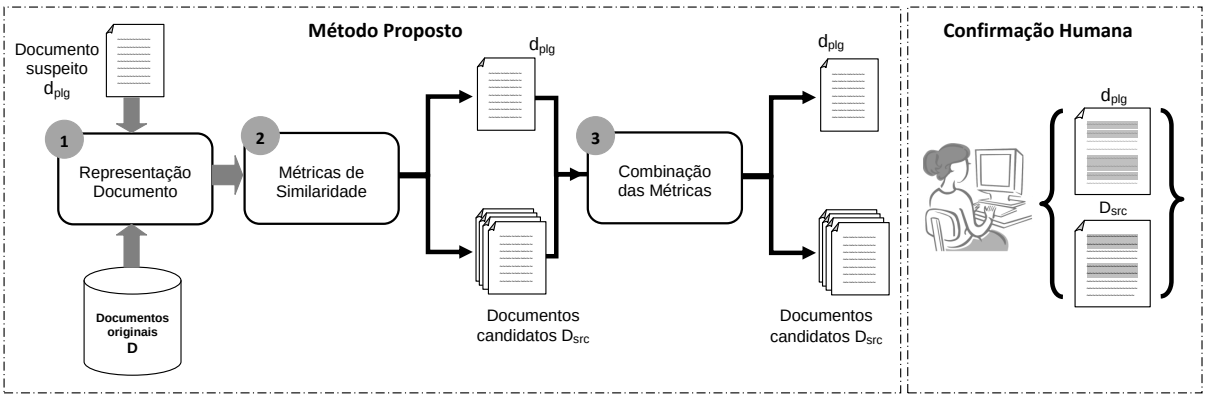


Figura 4.1 – Visão Geral do Método

A solução proposta é baseada na hipótese de que a combinação de métricas baseadas em conteúdo e referências pode aumentar a eficácia da detecção de plágio em artigos científicos. Esta tese também visa demonstrar a eficácia da combinação das métricas quando aplicadas sobre casos de reuso reais.

4.2 Representação dos documentos

Para computar os métodos baseados na análise de referências, os documentos precisam ser representados apenas pelas suas referências, conforme apresenta a Figura 4.2.

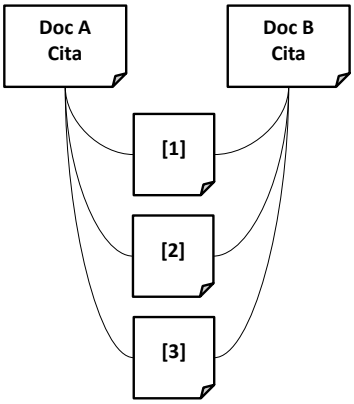


Figura 4.2 – Acoplamento Bibliográfico
Adaptado de Gipp e Beel (2010)

A ferramenta ParsCit [1] foi utilizada para pré-processar os artigos e extrair os metadados necessários. A ParsCit aplica um método de aprendizagem de máquina supervisionada, baseado em um algoritmo de *Conditional Random Fields* para analisar artigos científicos. Esta ferramenta identifica as citações no texto e as relaciona com suas referências correspondentes no bloco de referências. As referências também são segmentadas e seus metadados etiquetados como nome do autor, título, data, etc. De acordo com a Figura 4.3 (saída *XML* gerada pela ParsCit), pode-se perceber que além da ferramenta extrair os metadados mais comuns que constituem uma referência bibliográfica (Tabela 2.1), o campo *Marker* também é considerado neste trabalho. Este metadado corresponde a um identificador exclusivo para a fonte bibliográfica.

Os autores do ParsCit relatam resultados de avaliações da sua abordagem (COUNCILL; GILES; KAN, 2008). A ferramenta alcançou 0,9 de F1 na identificação dos campos nomes dos autores, títulos do artigo e data de publicação. A abordagem foi aplicada sobre as coleções Cora e CiteSeer. Nós examinamos uma amostra da saída produzida pela ferramenta ParsCit e concluímos que o mesmo nível de F1 foi alcançado em nossas coleções. Uma das dificuldades que encontramos foi que diferentes artigos podem citar a mesma referência de diferentes formas. Deste modo, a computação de dois artigos que citam a mesma referência não pode ser feita por meio de correspondência exata.

[1] <http://aye.comp.nus.edu.sg/ParsCit/>

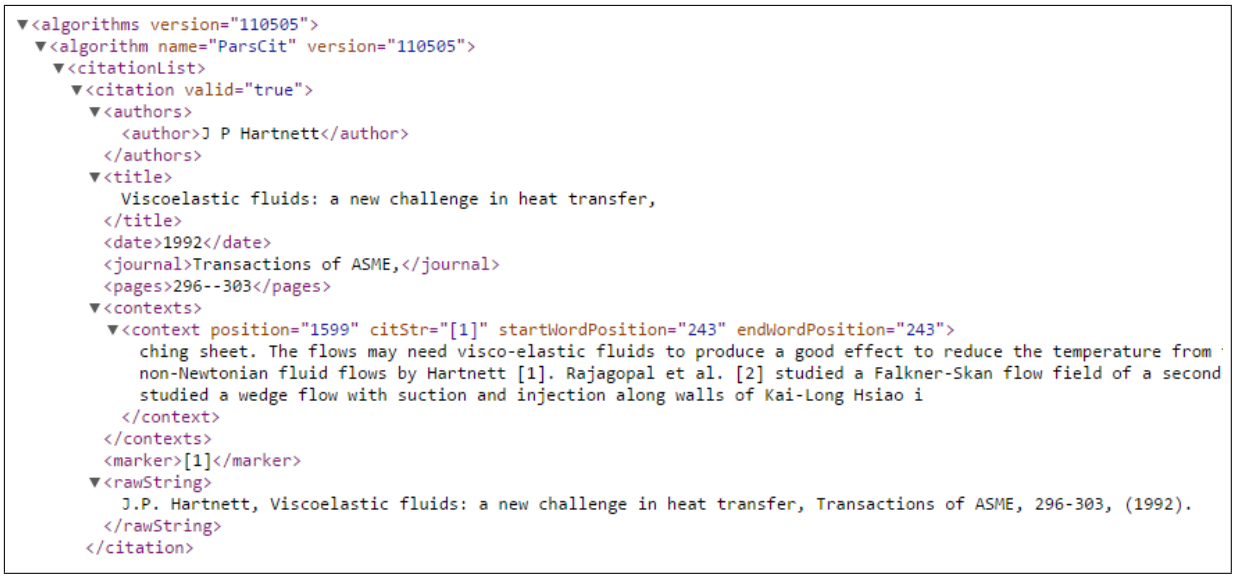


Figura 4.3 – Saída gerada pela Ferramenta ParsCit

Sendo assim, para cada par de referência de um par de documentos que está sendo analisado, usamos funções de similaridade para computá-las e definir se são correspondentes. A nossa análise incidiu sobre os metadados nome do autor, título do trabalho, título do livro/conferência, periódico, ano e página de publicação. A seguir apresentamos como computamos cada um dos metadados de uma referência:

- **Autor:** Num primeiro momento, tomamos as iniciais de todos os autores (por exemplo, se o nome é "Diane Litman and Kate Forbes" ou "D. Litman and K. Forbes" tomamos "DLKF"). E em seguida usamos a função NAMEMATCH proposta por Borges *et al.* (2011) para computar o campo autor. Deste modo, o escore de similaridade do algoritmo representa a porcentagem de casamentos encontrados entre os nomes dos autores de duas referências bibliográficas, e pode ser representado como:

$$sim\_autor(a_i, a_j) = \frac{n_{i,j}}{MAX(n_{i,j})}$$

  onde $n_{i,j}$ é o número de casamentos encontrados entre as duas cadeias de caracteres formadas pelas iniciais dos nomes dos autores, e MAX é o tamanho da maior cadeia. A Figura 4.4 apresenta um exemplo em que o nome do autor foi escrito de diferentes maneiras.

- **Título do Artigo/Periódico/Título do Livro:** Para comparar se estes metadados são correspondentes, computamos a distância de edição entre cada um deles.

- **Ano da publicação:** A compatibilidade dos anos de publicação é calculada pelo mesmo ano ou por apenas um ano de diferença. Por exemplo, para o ano 2013, também seriam

considerados os anos 2012 e 2014.

- **Páginas da Publicação:** O campo que representa as páginas de publicação é calculado pelo mesmo número das páginas de publicação.

D. Litman and K. Forbes. 2003. Recognizing emotion from student speech in tutoring dialogues. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).
D. J. Litman, C. P . Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In Proc. Intelligent Tutoring Systems.

Figura 4.4 – Exemplo de Referências em que ocorre variação no nome dos autores

Após obter o escore de similaridade entre os metadados bibliográficos, computamos se o par de referências em análise é correspondente ou não. Os metadados que não estiverem presentes nas duas referências comparadas não são computados.

Assim, primeiramente analisamos o campo autor. Se 50% das iniciais do nome dos autores corresponder, analisamos os demais campos. Num segundo momento, consideramos os campos título e ano da publicação. Se uma das referências que está sendo comparada não conter o campo ano, o limiar de similaridade utilizado para o campo título é de 80%. Caso contrário, o limiar é de 75%. No entanto, se o campo ano for correspondente e o título for maior que o limiar determinado, o par de referências é considerado correspondente. A Figura 4.5 mostra um par de referências que corresponde a mesma publicação. Mesmo com a variação no nome dos autores, o escore de similaridade entre o campo autor é 1 e o título é 0,9. Pode-se perceber que ocorre um erro de digitação na palavras “Emotions” no título da segunda referência.

Litman, Diane, and Kate Forbes (2003). Recognizing Emotions from Student Speech in Tutoring Dialogues. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), St. Thomas, Virgin Islands, November-December, 2003.
D. Litman and K. Forbes. 2003. Recognizing emotion from student speech in tutoring dialogues. In roc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Figura 4.5 – Exemplo de Referência com variação no nome dos autores

Notamos que em alguns casos os metadados título e ano da publicação não são suficientes para decidir se um par de referências corresponde a mesma publicação. Sendo assim, casos em que o título não está presente em uma das referências em análise, consideramos o nome do

Periódico/Título do Livro, e consideramos o número de páginas ao invés de considerar o ano. A Figura 4.6 mostra um par de referências em que a publicação é do mesmo autor e foi publicada no mesmo periódico no mesmo ano. Neste caso, apenas o número de páginas pode diferenciar as duas publicações.

Umezurike, G. M. (1971). Biochim. Biophys. Acta 227, 419-428.
Umezurike, G. M. (1971). Biochim. Biophys. Acta 250, 182-191.

Figura 4.6 – Exemplo de Referência sem Título da Publicação

Outro problema na variação do estilo das referências ocorre quando alguns autores são omitidos. Por exemplo, a Figura 4.7 mostra a variação no nome dos autores de uma mesma publicação. Neste caso, considerar os autores como sendo da mesma publicação não é trivial. Uma combinação do escores de todos os metadados seria uma possível solução, mas o par de referências apresentado na Figura 4.6 seria retornado como um falso positivo.

Fleck MPA, Lousada S, Xavier M, Chachamovich E, Vieira G, Santos L, et al. Aplicação da versão em português do instrumento abreviado de avaliação da qualidade de vida "WHOQOL-bref". Rev. Saúde Pública 2000 Abr; 34(2):178-83.
Fleck MPA et al. Aplicação da versão em português do instrumento abreviado de avaliação da qualidade de vida "WHOQOL-bref". Rev. Saúde Pública 2000 Abr; 34(2):178-83.

Figura 4.7 – Exemplo de referência em que ocorre omissão de autores

4.3 Métricas de Similaridade

A seguir, são descritas as medidas de similaridade baseadas em conteúdo, referências e citações que foram computadas para realizar as combinações do método proposto.

- **Acoplamento Bibliográfico:** calcula um escore de similaridade entre um par de documentos como o número absoluto de referências em comum entre eles (MEUSCHKE; GIPP; BREITINGER, 2012). Por exemplo, se considerarmos a Figura 4.2, percebemos que o documento A e B compartilham 3 referências em comum.
- ***Citing Frequency-Score (CF-Score):*** calcula um escore de similaridade entre um par de documentos com base na correspondência de padrões de citação (MEUSCHKE; GIPP; BREITINGER, 2012). Esta métrica considera que os documentos raramente citados são mais importantes e devem receber uma pontuação mais elevada (detalhes são descritos

na Seção 3.3). Nos nossos experimentos, esta função foi adaptada para calcular um escore entre dois documentos que compartilham referências correspondentes (ao invés de padrões de citação), como:

$$CF(d_i, d_j) = \sum^{n} \frac{N}{r_{i;j}}$$

onde n é o número de referências em comum entre o documento d_i e d_j, N é o número de documentos na coleção, e $r_{i;j}$ é o número de vezes que a referência r aparece na coleção.

- **Similaridade de Jaccard:** computa o escore de similaridade entre um par de documentos como a interseção entre suas referências dividida por sua união. Esta métrica é calculada como:

$$sim(d_i, d_j) = \frac{r_i \cap r_j}{r_i \cup r_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

onde r_i e r_j são as referências nos documentos d_i e d_j, respectivamente, $n_{i,j}$ é o número de referências compartilhadas entre os documentos d_i e d_j, n_i e n_j são o número de referências nos documentos d_i e d_j, respectivamente.

- **Coocorrências em Citações:** esta métrica (PERTILE; ROSSO; MOREIRA, 2013) calcula as Coocorrências em Citações entre dois documentos pelo deslizamento de duas janelas de tamanho s ao longo do texto dos documentos (uma para cada documento, conforme a Figura 4.8), e computa a similaridade de Jaccard com base naa coocorrências encontradas nos documentos. Detalhes dessa técnica encontram-se na próxima Seção. Nesta tese, $s = 15$ foi adotado por apresentar os melhores resultados em termos de F1 (Tabela A.1).

Para a análise baseada em conteúdo, foi utilizada a ferramenta WCopyFind (descrita na Seção 3.5). O escore de similaridade usado pela ferramenta é dado pelo número de n-gramas de palavras correspondentes entre um par de documentos. Nesta tese, adotamos o valor de n=6 (*Shortest Phrase to Match*).

4.3.1 Coocorrências em Citações

Esta seção apresenta um método desenvolvido para comparar o grau de similaridade entre artigos científicos com base na análise de coocorrências em citações. Se dois documentos compartilham pelo menos um par de citações dentro de um fragmento de texto, isto é considerado uma coocorrência. O método foi proposto em um artigo publicado na *Conference and Labs of the Evaluation Forum* (PERTILE; ROSSO; MOREIRA, 2013).

Nossa hipótese na solução proposta (PERTILE; ROSSO; MOREIRA, 2013) é que uma alta taxa de coocorrências pode ser um indicativo de plágio. Dado um par de documentos, estes são representados por meio das coocorrências de suas citações. Estas coocorrências dentro do documento são computadas por uma janela deslizante de tamanho s através do documento. As coocorrências entre dois documentos são, então, calculadas como o coeficiente de similaridade de Jaccard dessas coocorrências como na Eq. 4.1. A ideia é que os documentos com uma alta sobreposição são os candidatos a terem conteúdo similar.

$$sim(d_i, d_j) = \frac{d_i \cap d_j}{d_i \cup d_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \tag{4.1}$$

onde d_i e d_j são os documentos i e j, respectivamente, $n_{i,j}$ é o número de coocorrências compartilhadas entre os documentos i e j, n_i and n_j são o número de coocorrências nos documentos i e j, respectivamente.

Por exemplo, considere a Figura 4.8. Em primeiro lugar, todas as coocorrências em uma janela de $s = 7$ linhas dentro de um documento são computadas. Em seguida, essas coocorrências são usadas para comparar os documentos. A similaridade entre as duas janelas selecionadas é calculada de acordo com a Eq. 4.1 que gera $\frac{3}{4+6-3} = 0.43$.

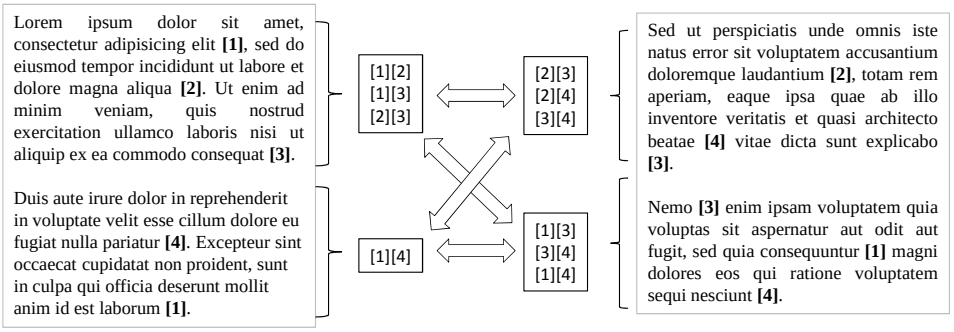


Figura 4.8 – Coocorrências em Citações

Mais especificamente, os passos envolvidos nessa abordagem são os seguintes:

- **Etapa 1: Pré-processamento**
  - Identificar as citações no texto do documento e ligá-las a sua entrada correspondente no bloco de referências. Uma vez que as citações e as referências podem ter diferentes estilos (por exemplo, numeradas, autor-data, etc.), dentro desta etapa elas são normalizadas para permitir uma comparação entre os artigos. Esta etapa exige que os campos nas referências sejam segmentados, a fim de identificar cada um deles.
  - Deslizar uma janela de tamanho s através dos documentos e computar as coocorrências dentro do documento.
- **Etapa 2: Computar as coocorrências entre pares de documentos** Para cada par de coocorrências entre o documento suspeito e o documento original, verificar se elas são correspondentes. Esta comparação não pode ser feita usando correspondência exata já que mesmo com a normalização, algumas diferenças ainda podem permanecer. Por exemplo, um artigo pode utilizar o nome completo dos autores em uma referência, enquanto outro artigo pode ter apenas o último nome e as iniciais do primeiro nome.

Na proposta de Pertile *et al.* (2013) cada citação foi representada pelo título da referência e pelo primeiro nome do autor. Uma extensão da distância de edição de *Levenshtein's*, chamada Carla (RITT et al., 2009), foi utilizada para comparar as referências entre dois documentos. Para o método proposto nesta tese, as citações foram representadas e comparadas conforme descrito na Seção 4.2. Os experimentos e resultados estão disponíveis no Anexo A.

4.4 Combinando as Métricas de Similaridade

O principal objetivo desta tese é propor uma estratégia combinando métricas baseadas em conteúdo e referências para identificar pares de documentos que contenham reuso de texto. Este objetivo visa auxiliar avaliadores humanos na detecção de plágio. Sendo assim, duas estratégias de combinação das métricas foram propostas: uma combinação simples e uma combinação utilizando aprendizagem de máquina.

A primeira estratégia proposta consiste simplesmente na união dos $top - |S|$ pares classificados pelas medidas baseadas em conteúdo, referências e citações. Ou seja, as métricas são computadas para cada par de documentos. Os documentos recuperados são ranqueados em ordem decrescente de acordo com seu escore de similaridade, e então combinados. Na

combinação, os ranques não são preservados.

A segunda combinação foi mais sofisticada, uma vez que fizemos uso de algoritmos de aprendizagem de máquina. Desta forma, cada par de documentos é representado como um vetor de características composto pelos escores atribuídos por cada uma das métrica de similaridade (Seção 4.3). O objetivo principal desta abordagem é construir um modelo baseado na combinação das métricas capaz de classificar novos pares de documentos. Para a construção do modelo, são necessários exemplos positivos (casos significativos de reuso de texto) e exemplos negativos (casos em que o reuso não é considerado significativo). Estes exemplos compõem o conjunto de treinamento. Idealmente, o número de exemplos positivos e negativos deve ser balanceado a fim de que o classificador não tenha nenhum viés quanto a uma classe ou outra.

A Figura 4.9 apresenta um exemplo do conjunto de treinamento, onde o rótulo @relation descreve o nome da coleção e os rótulos @attribute apresentam os escores calculados pelas métricas para cada par de documento. O valor `S` ou `N` no fim de cada instância indica se aquele par de documentos contém reuso de texto ou não (de acordo com as anotações atribuídas pelos avaliadores).

$$
\begin{array}{l}
@relation\ treinamento \\
\\
@attribute\ acoplamento\ bibliografico\ integer \\
@attribute\ jaccard\ real \\
@attribute\ cf\_score\ real \\
@attribute\ coocorrencia\ real \\
@attribute\ conteudo\ real \\
@attribute\ 'Class'\ \{S,\ N\} \\
\\
@data \\
41, 0.11, 35186, 0, 0.11, N \\
22, 1, 11327, 0, 0.16, N \\
2, 1, 1747, 0.00549, 0.99, S \\
66, 0.8, 56865, 0, 0.98, S \\
64, 0.19, 50749, 0, 0.52, N \\
29, 0.05, 23942, 0, 0.37, N \\
1, 0.11, 873, 0, 0.07, N \\
53, 0.79, 35510, 0, 0.41, S \\
116, 0.32, 92308, 0, 0.62, S \\
41, 0.37, 35558, 0, 0.21, S \\
\ldots
\end{array}
$$

Figura 4.9 – Exemplo de arquivo com instâncias de treinamento

Os seguintes algoritmos de aprendizagem de máquina foram utilizados para treinar o classificador. A lógica foi escolher algoritmos que usam diferentes estratégias de aprendizagem.

- Algoritmos baseados no teorema de Bayes - *Naive Bayes* e *BayesNet*;
- Algoritmos baseados em função - *SMO* uma variação do algoritmo SVM;
- Algoritmos baseados em Redes Neurais - *Multilayer Perceptron* (rede neural artificial) e *RBFNetwork* (rede de função de base radial);
- Algoritmos baseados em meta-aprendizagem - *AdaBoost*, *LogitBoost* e *Bagging*;
- Algoritmos baseados em regras de decisão - *ConjunctiveRule* e *DTNB* (combina uma Tabela de decisão (DT) com NaiveBayes (NB));
- Algoritmos baseados em árvore de decisão - *J48* e *RandomTree*;

4.5 Sumário do Capítulo

Este capítulo abordou uma visão do método proposto, bem como suas etapas. Ainda, foram descritas as abordagens desenvolvidas para combinação das métricas de similaridade. No próximo capítulo, serão abordados os passos e objetivos seguidos para realização de cada um dos experimentos.

5 AVALIAÇÃO EXPERIMENTAL

Este Capítulo descreve uma série de experimentos que avaliam a metodologia proposta para detecção de plágio em artigos científicos. Um dos objetivos desta tese é estudar a relação entre os métodos de detecção de plágio baseados em conteúdo e os métodos baseados em análise de referências/citações. Sendo assim, nossos experimentos são destinados a responder as seguintes perguntas: *Existe alguma correlação entre os métodos baseados em análise de conteúdo e baseados em análise de referências/citações?* Em outras palavras: eles detectam os mesmos casos ou eles complementam um ao outro? e *Uma combinação entre os dois tipos de métodos pode melhorar o processo de detecção de plágio?* Nossos experimentos calcularam quatro métricas baseadas na análise de referências/citações que são comparados com os resultados de uma ferramenta livremente disponível para a detecção de plágio baseada em conteúdo. Note que o foco está em comparar a eficácia e não a eficiência dos métodos.

Uma diferença importante entre os experimentos aqui relatados e os realizados por Gipp *et al.* (2014), é que, em seu estudo, abordagens de similaridade de conteúdo foram aplicadas apenas aos pares de documentos que compartilham pelo menos uma referência em comum. Acreditamos que isso faz com que o teste tenha um viés para a detecção baseada em citação, pois casos em que o infrator copia o conteúdo de um documento original, mas não a citação, iriam passar despercebidos. Nos experimentos aqui apresentados, uma vez que visamos avaliar a complementaridade entre os dois tipos de métodos, eles são executados de forma independente.

A seção 5.1 apresenta detalhes sobre as coleções utilizadas. O processo de geração do *ground truth* é apresentado na seção 5.2. A seção 5.3 define as métricas de avaliação da eficácia dos métodos baseados em conteúdo e referências. Foram executados três experimentos que são apresentados na seção 5.4. A seção 5.5 reporta os resultados dos experimentos conforme foram divididos.

5.1 Coleções de artigos científicos

A avaliação de detecção de plágio em documentos científicos tem uma limitação importante: a disponibilidade de um conjunto de teste. Idealmente, uma coleção de teste deve incluir casos reais de plágio, um *ground truth* e referências/citações para permitir que o método de detecção possa determinar se a fonte foi devidamente reconhecida.

Infelizmente, não há bases de teste que satisfaçam todas estas condições. As coleções que usamos aqui já foram utilizadas em outros estudos de detecção de plágio (GUPTA; ROSSO,

Tabela 5.1 – Detalhes das Coleções

Coleção	# Documentos	# Pares no Pool	# Pares Positivos
ACL	4686	93	41
PubMed	1513	85	64

2012; GARCÍA-ROMERO; ESTRADA-LORENZO, 2014). Porém, não há um *ground truth* disponível para elas, o que nos levou a criar um para cada uma das coleções (detalhes sobre a geração do *ground truth* podem ser encontrados na Seção 5.2).

A primeira coleção, `ACL`, foi utilizada para identificar as tendências de reuso de texto em Gupta e Rosso (2012). Para construir esta coleção, os autores coletaram artigos completos, resumidos, e de *workshop* publicados na conferência ACL em 1990-1997 e 2004-2011. Em nossos experimentos, foram utilizados 4686 documentos publicados entre 2004 e 2011, período em que Gupta e Rosso (2012) encontraram o maior número de casos de reuso. Cada documento foi comparado a todos os outros documentos na coleção.

A segunda coleção, `PubMed`, foi obtida a partir dos artigos PubMed disponibilizados pelo projeto Déjà vu[1]. Esta base de dados inclui cerca de 80000 pares de artigos para o qual o software eTBLAST encontrou altos índices de similaridade. Desses, selecionamos os 797 pares para os quais o texto completo estava disponível, totalizando 1513 documentos. Note que um mesmo artigo pode estar em mais de um par. Além disso, um mesmo artigo pode ora ser considerado fonte, ora suspeito. Como resultado o número de artigos é menor do que o dobro do número de pares. A Tabela 5.1 resume os detalhes das duas coleções de documentos usadas nos experimentos realizados nesta tese.

5.2 Geração do *Ground Truth*

Para gerar os *ground truths*, foi empregado o método *pooling* que é amplamente utilizado em Recuperação de Informação (VOORHEES; HARMAN, 1998). Assim, para cada coleção, computamos todas as métricas de similaridade entre todos os pares de documentos. Em seguida, os 30 pares com escores mais altos para cada métrica foram selecionados para compor o *pool*. Removendo as duplicatas (como o mesmo par pode ter sido identificado por mais de uma métrica), obteve-se 93 pares para a coleção `ACL` e 85 para `PubMed`. Estes pares foram então analisados por avaliadores humanos que receberam as mesmas instruções emitidas por Gipp *et al.* (2014), que pede aos avaliadores para relatarem como positivos os casos com

[1] <http://dejavu.vbi.vt.edu/dejavu/>

similaridade que um examinador que está fazendo uma checagem de plágio acharia importante conhecer. Com isso, todos os cinco níveis mencionados na Seção 2 seriam cobertos.

Os pares do *pool* foram analisados por um grupo de 10 avaliadores. Cada par foi visto por dois avaliadores e um terceiro foi solicitado sempre que havia um desacordo. As anotações podem ser obtidas em <http://www.inf.ufrgs.br/~slpertile/collections.html>.

Foi realizado o teste estatístico para computar a medida Kappa que reflete a concordância entre os avaliadores. De acordo com a Tabela 5.2, a concordância média para as avaliações (não considerando a terceira avaliação, que foi utilizada para os casos de discordância) foi de 84% para `ACL` e 80% para `Pubmed`, e o Fleiss' kappa (FLEISS, 1971) foi 0,675 (considerado *substancial*) e 0,524 (considerado *moderado*) para `ACL` e `PubMed` [2], respectivamente. O valor de Kappa foi menor para a coleção `Pubmed`, pois a maioria dos casos foram julgados como positivos por todos os avaliadores, o que aumentou a chance "da concordância ocorrer por acaso" que é levado em consideração nesta métrica.

Coleção	Porcentagem Média de Concordância	Fleiss' Kappa
ACL	84%	0.675
PubMed	80%	0.524

Tabela 5.2 – Concordância das Avaliações Manuais

5.3 Métricas de Avaliação

Uma vez que nosso *ground truth* foi gerado a nível de documento (e não de passagem), não podemos calcular as métricas como definidas na Seção 2.4. No entanto, calculamos a precisão, revocação e F1 para os pares de documentos.

Uma vez que cada métrica gera uma classificação de pares de documentos, é necessário selecionar um ponto de corte. Nós optamos por cortar o ranque dos documentos na posição $|S|$, onde $|S|$ é o número de detecções no *ground-truth*. Basicamente, esta é a métrica *Precisão-R*, uma métrica muito usada que avalia quantas detecções corretas são encontradas até os $|S|$-ésimos pares de documentos do ranque. Cortar o ranque na posição $|S|$, faz com que a precisão e revocação sejam iguais. Também calculamos a Precisão Média (AvP), que é a métrica padrão

[2] No conjunto de dados PubMed, 54 dos pares de suspeitos com textos completos tinham seu conteúdo examinado por revisores. No entanto, não foi possível utilizar essas avaliações, uma vez que, eles não indicam claramente se as semelhanças foram consideradas "significativa"

para avaliar os resultados classificados e é dada na Equação 5.1.

$$AvP = \frac{\sum_{k=1}^{|S|}(P(k) \times correto(k))}{|S|} \tag{5.1}$$

onde $|S|$ é o número de detecções no *ground truth*, $P(k)$ é a precisão até os k pares classificados, e *correto* é a função que retorna 1 se a detecção até os k classificados for correta (ou seja, se está no *ground-truth*) e 0, caso contrário.

5.4 Procedimento Experimental

Nosso primeiro experimento tem como objetivo analisar a correlação entre métodos de detecção baseados em conteúdo e baseados em citação. Sendo assim, uma vez que as medidas de similaridade foram computadas, analisamos a concordância entre elas. Um aspecto importante é que a saída das métricas é um escore de similaridade para cada par de documentos, que nos permite classificar os pares em ordem decrescente com base nos escores. Portanto, temos que escolher um limiar para separar os pares que têm reuso "significativo". Neste caso, escolhemos o número de detecções no *ground truth* como limiar para cada uma das coleções (Figura 5.1). Neste experimento, também analisamos a correlação entre as métricas baseadas em conteúdo e citação, a qual limitamos aos 100 pares de documentos no topo do ranque.

Em nosso segundo experimento, o objetivo foi comparar a precisão de tais métricas isoladamente e combiná-las para avaliar se essa combinação pode melhorar a qualidade de detecção. O fato de ambos os grupos de técnicas apresentarem similaridades identificadas em diferentes pares de documentos poderia indicar que elas são complementares uma da outra. Sendo assim, uma combinação destas abordagens poderia potencialmente melhorar a precisão de sistemas de detecção de plágio. A fim de avaliar se essa hipótese é verdadeira, nós realizamos um segundo experimento no qual os casos de plágio identificados por cada métrica foram comparados aos pares do *pooling* anotados pelos avaliadores. Dessa forma, fomos capazes de calcular a precisão, revocação e F1 para cada métrica.

Um terceiro experimento foi realizado com intuito de verificar se a combinação das métricas baseadas em conteúdo e citação usando aprendizagem de máquina melhora os resultados em relação a maneira como foram combinadas no experimento anterior. No entanto, este experimento tem como finalidade, com base nos escores gerados pelas métricas, construir um modelo baseado no conjunto de treinamento e classificar novos pares de documentos usando o

modelo gerado.

Para o conjunto de treinamento são necessários pares de documentos anotados por avaliadores humanos. Nesta etapa, visamos considerar o balanceamento em relação as classes. Neste sentido, para a coleção *ACL* o conjunto é representado por 82 instâncias (pares de documentos). Já para a coleção *PubMed* tivemos muitos casos positivos em relação aos casos negativos, sendo necessário obter mais casos negativos para compor o conjunto de treino. Para isso, foram selecionados aleatoriamente mais pares de documentos classificados pelas métricas, os quais foram anotados pelos avaliadores. O conjunto de dados da coleção *PubMed* é composto por 128 instâncias. Pares que não foram retornados por alguma das métricas recebe escore 0. Para a avaliação dos algoritmos de aprendizagem de máquina testados, foi utilizado o método *10-Fold Cross Validation* em todas as execuções. A análise de qualidade dos algoritmos foi feita com base na combinação de todas as métricas comparada à execução somente com a similaridade de conteúdo.

Por fim, usamos o modelo gerado para classificar os demais pares de documentos, os quais não possuem anotações. Esta classificação está disponível em: <http://www.inf.ufrgs.br/~slpertile/collections.html>.

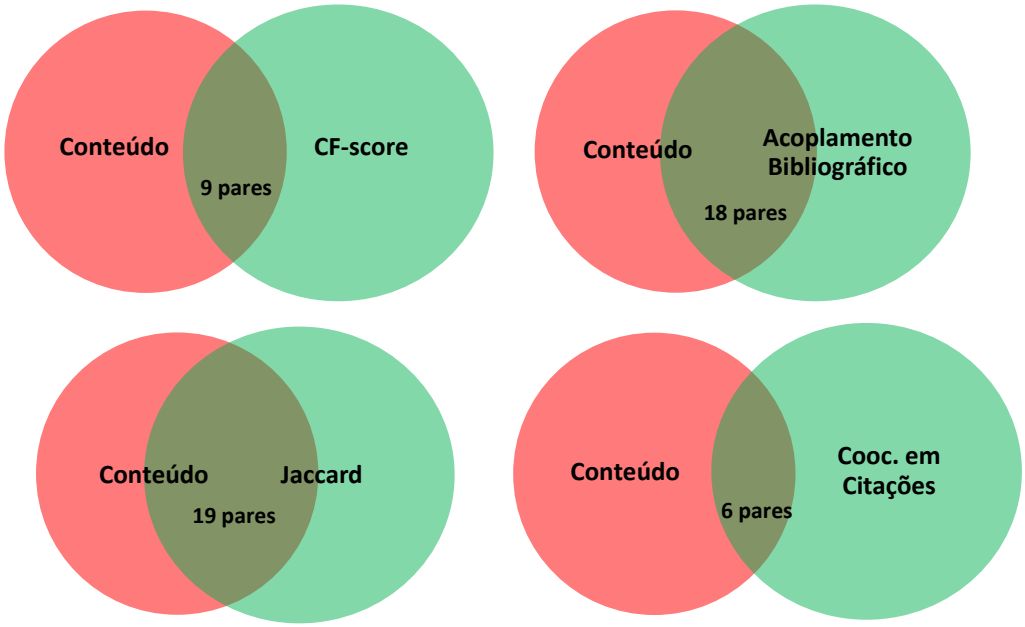
5.5 Resultados

Nesta seção são reportados os resultados dos três experimentos descritos na seção anterior.

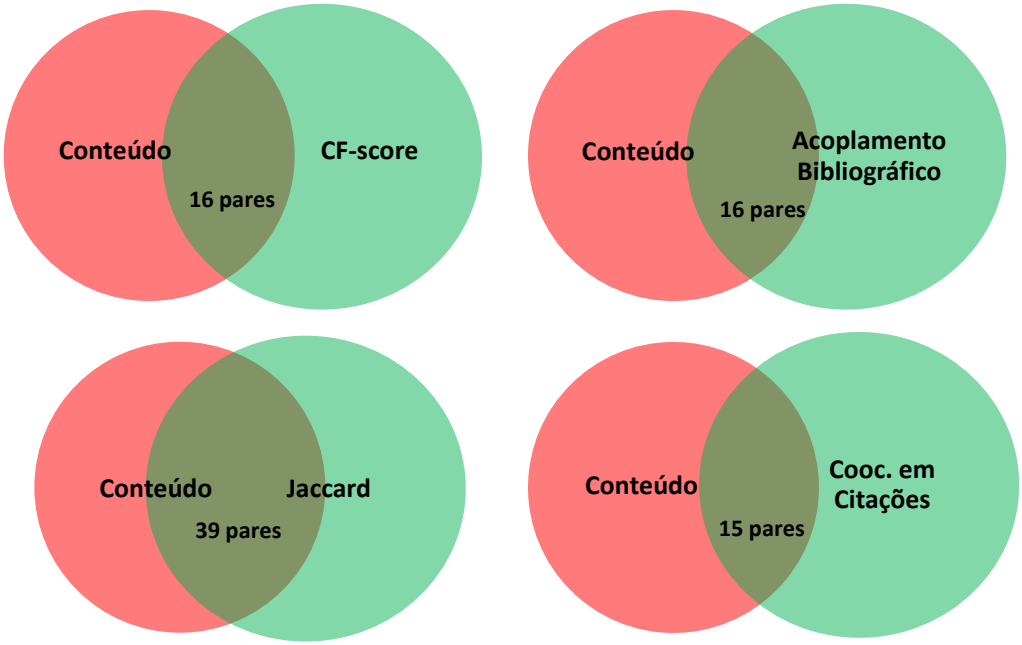
Correlação entre abordagens baseadas em Conteúdo e Referências/Citações.

Uma vez que as métricas de similaridade baseadas em conteúdo e referências/citações são calculadas, analisamos a concordância entre elas. Os ranques gerados pelas métricas foram bastante longos, sendo necessário limitá-los. Sendo assim, optamos por analisar os 100 primeiros pares. A Figura 5.1 mostra a interseção entre a métrica baseada em conteúdo e cada uma das métricas baseadas em referências e citações para ambas as coleções, `ACL` e `PubMed`.

Na coleção `ACL`, a interseção entre o conteúdo e as demais métricas foi entre 15-46%. Enquanto para a coleção `PubMed`, a intersecção tende a ser maior (23 a 61%) para todas as métricas. A menor sobreposição foi entre as métricas baseadas em conteúdo e coocorrências, para ambas as coleções. Apresentar uma interseção maior não significa necessariamente que os métodos são mais corretos, isso significa apenas que eles detectaram mais casos em comum.



(a) Coleção ACL



(b) Coleção PubMed

Figura 5.1 – Interseção entre pares de documentos identificados pelas métricas baseadas em conteúdo e referências/citações

Além de encontrar a interseção dos casos de reuso identificados pelos diferentes métodos, também calculamos a correlação de *Spearman's* (ρ), comparando a classificação dos pares de documentos gerado por cada métrica. A Tabela 5.3 mostra os coeficientes de correlação. Correlações significativas em α = 0,05 estão marcadas com um asterisco. As métricas que têm as correlações mais baixas com Conteúdo são Acoplamento Bibliográfico e Jaccard, para ambas as coleções. As correlações foram mais fortes na coleção `PubMed`.

Tabela 5.3 – Correlação entre Métricas Baseadas em Conteúdo e Referências/Citações

(a) Coleção ACL

		Baseados em Citação				Baseados em Conteúdo
		Acoplamento Bibliográfico	*CF-score*	*Jaccard*	*Coocorrências*	
Baseados em Citação	*Acoplamento Bibliográfico*	1.00*				
	CF-score	0.43*	1.00*			
	Jaccard	0.36*	0.16	1.00*		
	Coocorrências	0.10	0.38*	0.40*	1.00*	
Baseados em Conteúdo		0.10	0.35*	0.24*	0.14	1.00*

(b) Coleção PubMed

		Baseados em Citação				Baseados em Conteúdo
		Acoplamento Bibliográfico	*CF-score*	*Jaccard*	*Coocorrências*	
Baseados em Citação	*Acoplamento Bibliográfico*	1.00*				
	CF-score	0.98*	1.00*			
	Jaccard	0.63*	0.65*	1.00*		
	Coocorrências	-0.26*	-0.23*	0.10	1.00*	
Baseados em Conteúdo		0.33*	0.39*	0.38*	0.04	1.00*

Qualidade das métricas baseadas em Conteúdo e Referências/Citações. Para avaliar a precisão das métricas de similaridade, foram computadas a precisão, revocação, F1, AvP, e o número de verdadeiros positivos (VP), comparando os pares identificados pelas métricas e os pares no *ground truth*. Os resultados estão representados na Tabela 5.4. Os números mostram que a análise de conteúdo produz os melhores resultados em todas as métricas de avaliação. As melhores métricas baseada em citações foram Acoplamento Bibliográfico para a coleção `ACL` e Jaccard para `PubMed`. A métrica de Coocorrências em Citações apresentou os piores resultados.

Tabela 5.4 – Avaliação das Métricas baseadas em Conteúdo e Referências/Citações

Métricas de Similaridade	**ACL**			**PubMed**		
	P, R, F	**AvP**	**VP**	**P, R, F**	**AvP**	**VP**
Acoplamento Bibliográfico	0.61	0.57	25	0.44	0.41	28
CF-score	0.32	0.24	13	0.44	0.42	28
Jaccard	0.51	0.43	21	0.61	0.61	39
Coocorrências em Citações	0.20	0.09	8	0.36	0.28	23
Conteúdo	**0.76**	**0.78**	**31**	**0.67**	**0.76**	**43**

Combinando Métricas baseadas em Conteúdo e Citações. O fato de ambos os grupos de métricas terem identificado similaridade em diferentes pares de documentos poderia indicar que são complementares um do outro. Assim, uma combinação destas abordagens poderia potencialmente melhorar a precisão de sistemas de detecção de plágio. Para avaliar se a combinação de conteúdo e métricas baseadas em citações podem beneficiar tarefas de detecção de plágio, combinamos dois tipos de métricas de duas maneiras diferentes (descritas na seção 4.4).

A Tabela 5.5 mostra os resultados para a abordagem de combinação simples e utilizando aprendizagem de máquina. Os resultados correspondem ao valor mais alto de F1 alcançado pelos algoritmos testados. A primeira linha corresponde aos resultados obtidos quando se utilizou apenas o escore de similaridade baseada em conteúdo. A combinação simples traz melhorias na revocação (como pares mais positivos são encontrados), mas a precisão diminui à medida que o número de pares recuperados também cresce. O ganho na revocação não compensa a perda de precisão, refletindo em valores de F1 mais baixos.

Tabela 5.5 – Combinando Métricas de Similaridade

Métricas de Similaridade		ACL				PubMed			
		Revocação	Precisão	F1	VP	Revocação	Precisão	F1	VP
Simples	Apenas Conteúdo	0.76	**0.76**	**0.76**	31	0.67	**0.67**	**0.67**	43
	Acoplamento Bibliográfico + Conteúdo	0.95	0.56	0.70	39	0.88	0.50	0.64	56
	CF-score + Contenteúdo	0.85	0.44	0.58	35	0.88	0.50	0.64	56
	Jaccard + Conteúdo	0.85	0.51	0.64	35	0.75	0.54	0.63	48
	Coocorrências em Citações + Conteúdo	0.80	0.40	0.54	33	0.83	0.47	0.60	53
	Todas métricas de Citações/Referências	0.76	0.26	0.38	31	0.92	0.37	0.53	59
	Combinação de todas as métricas	**1.00**	0.29	0.45	**41**	**1.00**	0.36	0.53	**64**
Aprendizagem de Máquina	Apenas Conteúdo	0.94	0.95	0.94	36	0.92	0.91	0.91	56
	Acoplamento Bibliográfico + Conteúdo	0.94	0.95	0.94	36	0.91	0.91	0.91	58
	CF-score + Contenteúdo	0.94	0.95	0.94	36	**0.95**	**0.95**	**0.95**	**62**
	Jaccard + Conteúdo	0.94	0.95	0.94	36	0.90	0.90	0.90	59
	Coocorrências em Citações + Conteúdo	**0.95**	**0.96**	**0.95**	**37**	0.91	0.91	0.91	56
	Todas métricas de Citações/Referências	0.73	0.76	0.72	23	0.88	0.88	0.88	54
	Combinação de todas as métricas	**0.95**	**0.96**	**0.95**	**37**	0.93	0.93	0.93	60

O uso de aprendizagem de máquina produziu ligeiras melhorias para ambas as coleções. Os classificadores foram capazes de filtrar alguns dos falsos positivos, levando a um ganho em termos de precisão. Ao mesmo tempo, a combinação forneceu mais fontes de evidência trazendo uma revocação superior. Para a coleção `PubMed`, 5 dos 12 algoritmos testados produziram ganhos de F1, enquanto que, para a `ACL`, apenas 2 beneficiaram-se da combinação das métricas. De acordo com a Tabela 5.6, os melhores resultados em termos de F1 foram obtidos pelos algoritmos DTNB (classificador híbrido que combina uma tabela de decisão com o classificador Naive Bayes) e J48 (um classificador de árvore de decisão), para as coleções `ACL` e `Pubmed`, respectivamente. Em muitos casos, os resultados permaneceram os mesmos e, em alguns casos, os resultados até mesmo degradaram pois o número de falsos positivos aumentou.

Ao comparar as métricas baseadas em citações que resultam em um número absoluto (Acoplamento Bibliográfico e CF-Score), com as métricas que usam proporções (Jaccard e Coocorrências em Citações), argumentamos que ambos têm vantagens e desvantagens. A desvantagem do uso de números absolutos é que dois documentos podem ter um elevado número de referências em comum (e, portanto, um alto Acoplamento Bibliográfico), mas se eles têm um grande número de referências, isso pode não indicar reutilização maliciosa. Sendo assim, isso será refletido como uma baixa pontuação de Jaccard. Por outro lado, se dois artigos têm apenas

uma coocorrência cada, e esta é compartilhada entre eles, então eles vão ter uma pontuação de 1 para essa métrica. No entanto, se eles não compartilham muitas referências, eles vão ter um baixo Acoplamento Bibliográfico, o que pode ajudar a excluir o par do conjunto de candidatos. Portanto, usar uma variedade de métricas parece ser uma boa solução, podendo uma compensar as deficiências da outra.

Tabela 5.6 – Avaliação dos Algoritmos de Aprendizagem de Máquina

Algoritmo de Aprendizagem de Máquina	ACL		PubMed	
	Conteúdo	Combinação de todas as Métricas	Conteúdo	Combinação de todas as Métricas
NaiveBayes	0.93	0.92	0.91	0.90
BayesNet	0.94	0.90	0.91	0.86
SMO	0.94	0.93	0.91	0.90
MultilayerPerceptron	0.94	0.93	0.91	0.91
RBFNet	0.93	0.92	0.89	0.89
AdaBoost	0.94	0.89	0.91	0.92
LogitBoost	0.93	0.90	0.89	0.91
Bagging	0.94	0.93	0.91	0.91
ConjunctRule	0.94	0.94	0.88	0.88
DTNB	**0.94**	**0.95**	0.91	0.92
J48	0.93	0.90	**0.91**	**0.93**
RandomTree	0.87	0.89	0.84	0.93

Encontrando a fonte para um único documento suspeito. Nas avaliações anteriores, calculamos a similaridade entre pares de artigos, utilizando todas as métricas para todos os documentos suspeitos e utilizado os escores para criar uma única classificação dos pares. Agora vamos explorar um cenário diferente: a classificação que cada métrica produz para um único documento. Assim, cada documento suspeito é comparado com a coleção inteira e os ranques produzidos são avaliados em termos da Média das Precisões Médias (MAP).

A pontuação MAP para um conjunto de detecções é a média das pontuações de AvP (Eq.5.1) para cada detecção. É calculada como:

$$MAP = \frac{\sum_{k=1}^{|S|} AvP(k)}{|S|} \tag{5.2}$$

O valor ideal para o MAP é 1, e isso significa que a métrica ranqueou todos os artigos com reuso significativo (de acordo com o *ground truth*) com escore maior do que qualquer documento para o qual não foi encontrado reuso significativo. Os resultados da Tabela 5.7 mostram que a análise de conteúdo gerou as melhores classificações; sendo quase perfeita para ambas as coleções, `ACL` e `PubMed`. Métricas baseadas em citações apresentaram um desempenho pior,

mas ainda, para artigos mais suspeitos, elas atribuíram o maior escore para o documento fonte correspondente. Isto sugere que todas as métricas poderiam potencialmente ser utilizadas como indicativos de plágio, desde que os documentos originais estejam disponíveis para comparação.

Tabela 5.7 – MAPs para a classificação produzida por um único documento

Coleção	Conteúdo	CF-score	Acoplamento Bibliográfico	Jaccard	Coocorrências em Citações
ACL	1.00	0.80	0.81	0.81	0.62
PubMed	0.96	0.90	0.90	0.90	0.43

Limitações. Este estudo incidiu sobre os tipos de plágio definidos pela ACM e IEEE, que consideram cópias exatas ou grandes porções de texto parafraseado (páginas inteiras ou parágrafos). Em nossos experimentos, utilizamos o número de detecções do *ground truth* como ponto de corte. Entretanto, na prática não seria possível conhecer este número. Sendo assim, determinar um limiar para decidir quais pares manter pode ser considerada uma questão importante.

5.6 Sumário do Capítulo

Este capítulo apresentou três experimentos executados sobre duas coleções reais de artigos científicos. O objetivo principal dos experimentos foi estudar a relação entre os tipos de métodos de detecção de plágio e se a combinação das métricas baseadas em conteúdo e referências/citações é capaz de melhorar o processo. Sendo assim, as duas coleções utilizadas foram descritas, bem como o processo para geração dos pares analisados pelos avaliadores. Os melhores resultados foram obtidos pela combinação utilizando algoritmos de aprendizagem de máquina, que também são descritos neste capítulo. O próximo capítulo apresentará casos reconhecidos de artigos retratados por plágio, bem como experimentos realizados com três casos.

6 ANALISANDO CASOS DE PLÁGIO RECONHECIDOS

Neste capítulo serão apresentados casos de artigos que foram retratados por plágio em revistas científicas. Na Seção 6.1 serão descritos alguns destes casos. Experimentos foram realizados com três dos casos reconhecidos e serão descritos na Seção 6.2

6.1 Casos de Plágio Reconhecidos

A prática de plágio tem se apresentado como um problema tanto no âmbito acadêmico como cientifico, uma vez que os plagiadores têm praticado este ato mesmo com as consequências que isto pode vir a causar. Segundo um estudo realizado por Fang *et al.* (2012), no qual foram analisados 2047 artigos científicos indexados pelo banco de dados PubMed como retratados, revelou que apenas cerca de 21% dos casos foram atribuídos a erro. Já os demais casos foram atribuídos à má conduta, sendo cerca de 10% devido a plágio e 14% foram devido a publicações duplicadas. A causa mais frequente de retração foi fraude, com aproximadamente 43% (ou seja, fabricação e falsificação de dados).

Outra análise com os artigos indexados no PubMed foi realizada por Amos (2014). O autor analisou 821 artigos retratados (0,02% da literatura do PubMed) entre 2008-2012. Dentre os artigos analisados, cerca de 16,6% são retratações por plágio e 18,1% por publicações duplicadas. O estudo ainda relata que os Estados Unidos é o país que apresentou o maior número de artigos retratados (24,2%), seguido pela China (17,4%) e pelo Japão (6,9%). Entretanto, dos 20 países que apresentaram mais de 5 retrações, a maior taxa de retração por plágio foi encontrada na Itália, 66,7% das retratações (de 24 retratações, 16 delas foram por plágio). A Turquia vem em segundo lugar com 61,5% (8 de 13 retratações) seguida do Irã (6 de 14) e Tunísia (3 de 7) com 42,9% cada. O Brasil apareceu em 17º com 9 casos de retratação, sendo que um terço por plágio e um caso por duplicação.

Em 2010, a IEEE relatou um caso de plágio[1] praticado por um estudante de doutorado do *Institut Teknologi Bandung* na Indonésia. Após uma acusação, a IEEE investigou o artigo publicado nos anais da *IEEE Conference on Cybernetics and Intelligent Systems* de 2008 e determinou que era uma duplicação quase completa do trabalho de um pesquisador austríaco que já havia sido publicado nos anais do *International Workshop on Database and Expert System Applications* em 2000. Com este ato, o doutorado do estudante foi revogado e o mesmo inabilitado de publicar artigos na IEEE por três anos.

[1] <http://dx.doi.org/10.1109/ICCIS.2008.4670963>

No mesmo ano, a revista *Biochemical Pharmacology*[2] retratou um artigo publicado em 2008. O artigo reproduzia imagens de microscopia eletrônica idênticas às que haviam sido divulgadas originalmente em outra revista em 2003. Além disso, trechos semelhantes foram publicados sem dar os devidos créditos a fonte original. A contestação partiu da autora do artigo original. Um dos autores do trabalho suspeito era um professor e foi demitido após a confirmação do plágio.

Em 2011, um caso publicado na revista *Economic Modelling* em 2007[3] também foi investigado e retratado pelo comitê de pesquisas em economia *RePEc(Research Papers in Economics)*[4]. O comitê identificou que o artigo era uma reescrita de um artigo publicado em uma revista da área de mecânica estatística em 2003.

Um outro caso de plágio foi descoberto em 2013[5] e resultou na retratação de um artigo publicado em 2011 na revista Brasileira de Geriatria e Gerontologia. O artigo derivou da dissertação de mestrado de um estudante da Universidade Federal de Viçosa que teve seu título cassado. A revista considerou que o artigo apresentava diversos parágrafos que foram transcritos fielmente do artigo original, incluindo a metodologia e os resultados. A próxima seção descreve experimentos realizados com alguns destes casos.

6.2 Experimentos

Nesta seção apresentaremos experimentos realizados com alguns dos casos de plágio reais relatados na Seção anterior. Para execução dos experimentos, as métricas baseadas em conteúdo e referências foram aplicadas.

Primeiramente foi necessário obter o artigo suspeito e o original de cada um dos casos. Foram analisados três pares de documentos a partir das métricas baseadas em conteúdo e referências/citações. Uma vez que os pares de artigos científicos analisados neste experimento foram coletados de diferentes veículos de publicação, para computar o CF-score não temos os valores de N (número de documentos na coleção) e $r_{i;j}$ (número de vezes que a referência r aparece na coleção) (detalhes na Seção 4.3). Sendo assim, dado um par de artigos, utilizamos o Google Acadêmico para obter o valor de r para cada referência em comum encontrada. E para N atribuímos o maior valor de $r_{i;j}$.

[2]<doi:10.1016/j.bcp.2008.05.003>
[3]<http://dx.doi.org/10.1016/j.econmod.2006.06.007>
[4]<http://repec.org/>
[5]<http://dx.doi.org/10.1590/S1809-98232011000300004>

Tabela 6.1 – Resultados dos experimentos com casos de plágio reconhecidos

		Métricas de Similaridade				
Documento Suspeito	**Documento Original**	**Acoplamento Bibliográfico**	**CF-score**	**Jaccard**	**Coocorrências em Citações**	**Conteúdo**
J.X. Zhang, Q.L. Da, Y.H. Wang. Analysis of nonlinear duopoly game with heterogeneous players. Econ. Model., 24 (2007), pp. 138–148	H.N. Agiza, A.A. Elsadany. Nonlinear dynamics in the cournot duopoly game with heterogeneous players. Physica A, 320 (2003), pp. 512–524	12	338	0.37	0.06	0.19
Zuliansyah, M.; Supangkat, S. H.; Priyana, Y.; Machbub, M. 3D Topological Relations for 3D Spatial Analysis. Proceedings of the 2008 IEEE Conference on Cybernetics and Intelligent Systems. pp. 585-590	Zlatanova, S. (2000). "On 3D Topological Relationships." 11th Int. Workshop on Database and Expert Systems Applications, Greenwich, London, UK.	0	0	0	0	0.88
Moreira PHB, Mafra SCT, Pereira ET, Silva VE. Qualidade de vida de cuidadores de idosos vinculados ao Programa Saúde da Família – Teixeiras, MG. Ver. Bras. Geriatr. Gerontol. 2011;14(3):433-40.	Amendola F, Oliveira MA, Alvarenga MR. Qualidade de vida dos cuidadores de pacientes dependentes no programa de saúde da família. Texto & Contexto Enferm. 2008;17(2):266-72.	14	1464	0.87	0.24	0.59

A Tabela 6.1 apresenta os resultados da análise. O primeiro caso de plágio analisado foi retratado pela revista *Economic Modelling* em 2011. O artigo suspeito apresentou apenas 19% de similaridade quando comparado ao artigo original. Pode-se perceber que o baixo escore de similaridade se deu principalmente devido ao artigo suspeito ter parafraseado o original. Além disso, o plagiador utilizou equações e figuras do artigo original, dificultando a análise do conteúdo por ferramentas de detecção de plágio baseadas apenas no conteúdo. Pode-se considerar que, apesar dos escores retornados pelas métricas baseadas referências/citações terem sido baixos, todas elas podem ser utilizadas como alertas para o avaliador na detecção de plágio. Neste caso o autor do artigo suspeito referenciou o artigo original.

O segundo caso foi o artigo publicado na IEEE em 2008. O documento suspeito apresentou 88% de similaridade no conteúdo, mas não apresentou nenhuma referência em comum. De acordo com a Figura 6.1 pode-se observar que o plagiador fez cópias exatas de passagens do documento original mantendo as citações (formato numérico), mas alterou as referências bibliográficas. Além disso, notamos que algumas das referências citadas foram simplesmente substituídas por publicações mais recentes do mesmo autor referenciadas no artigo original.

O caso descoberto pela revista Brasileira de Geriatria e Gerontologia em 2013 apresentou 59% de similaridade do conteúdo. Mas percebemos que grande parte do documento suspeito foi parafraseado pelo plagiador, dificultando a análise pela ferramenta de similaridade de conteúdo. A Figura 6.2 apresenta um exemplo de paráfrase. Neste caso, podemos considerar que a análise de referências contribuí para a identificação do plágio. O plagiador usou apenas uma referência diferente das que foram citadas no artigo original. Além disso, percebeu-se que o plagiador reestruturou as referências e citações, o que pode ter atribuído um baixo escore para a métrica de Coocorrências em Citações.

Com o experimento realizado em artigos científicos reais retratados por plágio conseguimos perceber que além de parafrasear o conteúdo, os plagiadores também tentam despistar o plágio manipulando referências e citações, o que torna mais difícil a tarefa dos sistemas de

detecção. Sendo assim, podemos considerar que combinar os dois tipos de métricas pode ajudar a alertar os revisores.

***Passagem suspeita**: This paper presents a unified set of conditions for deriving the possible relationships between multidimensional simple spatial objects in 1,2 and 3D space. The conditions are systemised on the basis of dimension, co-dimension and connectivity of boundaries. Thus most of the conditions (15 of 23, see [6]) derived for 2D space are propagated in 3D space and the overal number of conditions is reduced.*
Referências
[6] R. Reis, M. Egenhofer, and J. Matos, Topological relations using two models of uncertainty for lines, 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences,286-295, 2006.

***Passagem original**: This paper presents a unified set of conditions for deriving the possible relationships between multidimensional simple spatial objects in 1,2 and 3D space. The conditions are systemised on the basis of dimension, co-dimension and connectivity of boundaries. Thus most of the conditions (15 of 23, see [6]) derived for 2D space are propagated in 3D space and the overal number of conditions is reduced.*
Referências
[6] Egenhofer, M. J. and J. R Herring, 1992, Categorising topological relations between regions, lines and points in geographic databases, in: The 9-intersections: formalism and its use for natural language spatial predicates, Technical report 94-1, NCGIA, University of California.

Figura 6.1 – Exemplo de passagem plagiada - Cópia Exata

***Passagem suspeita:** A amostra foi composta por cuidadores domiciliares de idosos, atendida por equipes de Saúde da Família distribuídas entre quatro unidades básicas de saúde.*

***Passagem original:** A amostra foi composta por cuidadores familiares de pacientes dependentes, atendida por equipes de Saúde da Família distribuídas entre sete Unidades Básicas de Saúde.*

Figura 6.2 – Exemplo de plágio parafraseado

7 CONCLUSÃO

Esta tese apresentou uma pesquisa sobre abordagens e ferramentas de detecção de plágio. No decorrer do texto, foram abordados os trabalhos relacionados ao tema e foi apresentada uma classificação dos trabalhos segundo três abordagens: (i) baseadas na análise de conteúdo; (ii) baseadas na estrutura e conteúdo; (iii) e baseadas em análise de citações e referências. Nosso foco está na combinação de abordagens baseadas em conteúdo e referências/citações.

Para avaliar se estas abordagens baseadas em conteúdo e em referências são complementares, realizamos experimentos com duas coleções reais de artigos científicos (ACL e PubMed). *Ground truths* com julgamentos humanos foram produzidos para ambas as coleções. Observou-se que, no máximo, metade dos pares identificados pelas métricas baseadas em conteúdo também foram identificadas pelas métricas baseadas em citação (e vice-versa). A correlação entre essas métricas é mais fraca na coleção da ACL do que no PubMed. Avaliamos a qualidade das métricas a partir do *Ground truth* e observamos que os resultados da similaridade baseada em conteúdo foi superior aos resultados das demais métricas. Em seguida, avaliamos se uma combinação das métricas poderia trazer melhores resultados. Quando as métricas foram combinadas através de algoritmos de aprendizagem de máquina, obtivemos ganhos pequenos em comparação ao uso das métricas por isoladamente. Em alguns casos, os algoritmos de aprendizagem de máquina foram capazes de reduzir falsos positivos e falsos negativos, levando a uma maior precisão e revocação.

Outra constatação importante é que a maioria dos casos em que foi encontrado reuso de texto significativo, as publicações compartilhavam pelo menos um autor. Assim, eles são candidatos a serem considerados auto-plágio, o que, como discutido no Capítulo 2, é bastante controverso. Sendo assim, tal análise nos leva a outra importante questão - “O que razoavelmente podemos esperar de um sistema de detecção de plágio?". Sistemas automáticos só podem ser utilizados para identificar o *reuso* de texto. O julgamento sobre se o reuso consiste em plágio não pode ser feito automaticamente. Pode-se considerar que é altamente improvável que um sistema automático descartaria documentos com um alto escore de similaridade, como é o caso de erratas ou versões estendidas de artigos de conferências publicados em revistas.

Apesar dos resultados superiores terem sido obtidos pelas métricas de similaridade de conteúdo, acreditamos que a análise de referências é promissora, especialmente em casos de plágio multilíngue. Neste cenário, um método baseado em n-gramas não seria capaz de identificar as potenciais fontes. Os experimentos realizados com casos reais de artigos retratados também mostraram que os plagiadores reescrevem/parafraseiam o texto, a fim de dificultar a

descoberta do ato. Além disso, em alguns casos, o texto completo do documento pode não estar disponível, mas as referências estão, portanto, métricas como Acoplamento Bibliográfico, Jaccard, e CF-score podem ser utilizadas. A desvantagem é que elas podem acabar identificando falsos positivos como encontramos em nossos experimentos, onde os autores tendem a reutilizar as mesmas referências, mesmo quando o corpo da obra difere de forma significativa.

A detecção de plágio multilíngue também tem recebido a atenção de pesquisadores (PEREIRA; MOREIRA; GALANTE, 2010; POTTHAST et al., 2010a), uma vez que a tarefa de detecção se torna mais complexa quando os artigos originais e suspeitos são escritos em diferentes idiomas. Nesse contexto, consideramos que combinar métricas de análise de plágio multilíngue com as métricas baseadas em referências/citações aqui propostas pode melhorar a qualidade da detecção. No entanto, uma análise de reuso de texto em exemplos de artigos que contenham plágio multilíngue se faz necessária.

Além das métricas baseadas em conteúdo e referências/citações, a combinação com métricas que analisam a estrutura do documento pode trazer ganhos. Ou seja, pares de documentos que compartilham reuso de texto na Seção de Metodologia seriam mais propensos a serem considerados plágio do que aqueles que compartilham reuso na Seção de Revisão da Literatura.

Outra possibilidade de trabalho futuro é a realização de uma análise detalhada de reuso de texto. Passagens detectadas como sendo reuso devem ser analisadas para verificar se estão creditando a fonte original.

Por fim, métodos mais elaborados podem ser utilizados para verificar se um par de referências corresponde a uma mesma publicação. Existe uma série de métodos que objetivam a deduplicação de dados bibliográficos que podem ser analisados. Esta tarefa é essencial para garantir a qualidade do cálculo das métricas baseadas em referências/citações.

Publicações:

O trabalho desenvolvido nesta tese rendeu 3 publicações. A primeira (PERTILE; MOREIRA, 2012) trata da construção de uma coleção de casos de plágio artificiais em artigos científicos foi publicada na *Conference and Labs of the Evaluation Forum (CLEF)* (Qualis B2 na Ciência da Computação). Esta coleção acabou não sendo usada nos nossos experimentos, por isso está relatada no Apêndice B. Em 2013, publicamos a proposta da métrica de similaridade baseada na contagem de coocorrências em citações (PERTILE; ROSSO; MOREIRA, 2013) também na *Conference and Labs of the Evaluation Forum (CLEF)*. Este ano, a comparação e combinação entre métricas baseadas em conteúdo e em referências/citações foi aceito pelo periódico *Journal of the Association for Information Science and Technology (JASIST)* – Qualis A1 na Ciência da Computação (PERTILE; MOREIRA; ROSSO, 2015).

REFERÊNCIAS

ALZAHRANI, S. et al. Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications. **JASIST**, v. 63, n. 2, p. 286–312, 2012.

AMOS, K. The ethics of scholarly publishing: exploring differences in plagiarism and duplicate publication across nations. In: **Journal of the Medical Library Association**. [S.l.]: Medical Library Association, 2014. v. 102, p. 87–91.

ANDERSON, M. S.; STENECK, N. H. The Problem of Plagiarism. **Urologic Oncology: Seminars and Original Investigations**, v. 29, n. 1, p. 90–94, 2011.

BALAGUER, E. V. Putting Ourselves in SME's shoes: Automatic Detection of Plagiarism by the Wcopyfind tool. In: **Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse**. [S.l.: s.n.], 2009. p. 34–35.

BARRÓN-CEDEÑO, A.; ROSSO, P. On Automatic Plagiarism Detection Based on n-Grams Comparison. In: **ECIR**. [S.l.: s.n.], 2009. p. 696–700.

BARRÓN-CEDEÑO, A. et al. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. **Computational Linguistics**, v. 39, n. 4, p. 917–947, 2013.

BEALL, J. **Five Ways to Defeat Automated Plagiarism Detection**. 2013. Accessed 10-Feb-2015. Available from Internet: <http://scholarlyoa.com/2013/02/07/five-ways-to-defeat-automated-plagiarism-detection/>.

BORGES, E. N. et al. An unsupervised heuristic-based approach for bibliographic metadata deduplication. **Inf. Process. Manage.**, v. 47, n. 5, p. 706–718, 2011. Available from Internet: <http://dx.doi.org/10.1016/j.ipm.2011.01.009>.

BREIMAN, L. Bagging predictors. **Mach. Learn.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 2, p. 123–140, 1996.

BROOMHEAD, D. S.; LOWE, D. Radial basis functions, multi-variable functional interpolation and adaptive networks. **Complex Systems**, v. 2, p. 321–355, 1988.

BURGET, R. Automatic Document Structure Detection for Data Integration. In: **Proceedings of the 10th International Conference on Business Information Systems**. [S.l.: s.n.], 2007. (BIS'07), p. 391–397.

CHOW, T.; RAHMAN, M. Multilayer SOM with Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection. **Neural Networks, IEEE Transactions on**, v. 20, n. 9, p. 1385–1402, 2009.

COLLBERG, C.; KOBOUROV, S. Self-plagiarism in Computer Science. **Commun. ACM**, v. 48, n. 4, p. 88–94, 2005.

COOPER, G. F.; HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. **Machine Learning**, Kluwer Academic Publishers, v. 9, n. 4, p. 309–347, 1992. ISSN 0885-6125.

CORTEZ, E. et al. Ondux: on-demand unsupervised learning for information extraction. In: **SIGMOD**. [S.l.: s.n.], 2010. p. 807–818. ISBN 978-1-4503-0032-2.

COUNCILL, I. G.; GILES, C. L.; KAN, M. yen. ParsCit: An open-source CRF reference string parsing package. In: **International Language Resources and Evaluation**. [S.l.]: European Language Resources Association, 2008.

CRONIN, B. Self-plagiarism: An odious oxymoron. **Journal of the American Society for Information Science and Technology**, v. 64, n. 5, p. 873–873, 2013. ISSN 1532-2890.

FANG, F. C.; STEEN, R. G.; CASADEVALL, A. Misconduct accounts for the majority of retracted scientific publications. **Proceedings of the National Academy of Sciences**, v. 109, n. 42, p. 17028–17033, 2012.

FLEISS, J. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, v. 76, n. 5, p. 378–382, 1971.

FREUND, Y.; SCHAPIRE, R. E. **A desicion-theoretic generalization of on-line learning and an application to boosting**. Springer Berlin Heidelberg, 1995. 23-37 p. (Lecture Notes in Computer Science, v. 904). Available from Internet: <http://dx.doi.org/10.1007/3-540-59119-2_166>.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting. **Annals of Statistics**, v. 28, p. 2000, 1998.

GARCÍA-ROMERO, A.; ESTRADA-LORENZO, J. M. A Bibliometric Analysis of Plagiarism and Self-plagiarism through Déju vu. **Scientometrics**, Springer Netherlands, v. 101, n. 1, p. 381–396, 2014. ISSN 0138-9130.

GARNER, H. The Case of the Stolen Words. **Scientific American**, v. 310, n. 3, p. 64– 67, 2014. ISSN 00368733.

GIPP, B. **Doctoral Thesis: Citation-based Plagiarism Detection: Applying Citation Pattern Analysis to Identify Currently Non-Machine-Detectable Disguised Plagiarism in Scientific Publications**. Thesis (PhD) — Department of Computer Science, Otto-von-Guericke University Magdeburg, Germany, 2013.

GIPP, B.; BEEL, J. Citation based Plagiarism Detection: a New Approach to Identify Plagiarized Work Language Independently. In: **Proceedings of the 21st ACM Conference on Hypertext and Hypermedia**. [S.l.: s.n.], 2010. (HT '10), p. 273–274.

GIPP, B.; MEUSCHKE, N. Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In: **Proceedings of the 11th ACM Symposium on Document Engineering (DocEng2011)**. [S.l.: s.n.], 2011.

GIPP, B.; MEUSCHKE, N.; BREITINGER, C. Citation-based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus. **Journal of the Association for Information Science and Technology**, v. 65, n. 8, p. 1527–1540, 2014. ISSN 2330-1643.

GIPP, B. et al. Demonstration of Citation Pattern Analysis for Plagiarism Detection. In: **Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2013. (SIGIR '13), p. 1119–1120.

GRMAN, J.; RAVAS, R. Improved Implementation for Finding Text Similarities in Large Sets of Data - Notebook for PAN at CLEF 2011. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2011.

GROZEA, C.; GEHL, C.; POPESCU, M. Encoplot: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2009.

GROZEA, C.; POPESCU, M. The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Notebook for PAN at CLEF 2011. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2011.

GROZEA, C.; POPESCU, M. Encoplot - Tuned for High Recall (also Proposing a New Plagiarism Detection Score). In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2012.

GUPTA, P.; ROSSO, P. Text Reuse with ACL: (upward) trends. In: **Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries**. [S.l.: s.n.], 2012. (ACL '12), p. 76–82.

HACOHEN-KERNER, Y.; TAYEB, A.; BEN-DROR, N. Detection of Simple Plagiarism in Computer Science Papers. In: **Proceedings of the 23rd International Conference on Computational Linguistics**. [S.l.: s.n.], 2010. (COLING '10), p. 421–429.

HALL, M.; FRANK, E. Combining naive bayes and decision tables. In: **Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)**. [S.l.]: AAAI press, 2008. p. 318–319.

HAYKIN, S. **Neural Networks**: A comprehensive foundation. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

HECHT-NIELSEN, R. Theory of the backpropagation neural network. In: **Neural Networks, 1989. IJCNN., International Joint Conference on**. [S.l.: s.n.], 1989. v. 1, p. 593–605.

HOAD, T. C.; ZOBEL, J. Methods for Identifying Versioned and Plagiarized Documents. **J. Am. Soc. Inf. Sci. Technol.**, v. 54, p. 203–215, 2003.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: **Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (UAI'95), p. 338–345. ISBN 1-55860-385-9.

JUOLA, P. Authorship attribution. **Foundations and Trends in Information Retrieval**, v. 1, n. 3, p. 233–334, 2006. ISSN 1554-0669.

KAN, M.-Y.; LUONG, M.-T.; NGUYEN, T. D. Logical structure recovery in scholarly articles with rich document features. **Int. J. Digit. Library Syst.**, IGI Global, Hershey, PA, USA, v. 1, n. 4, p. 1–23, oct. 2010. ISSN 1947-9077. Available from Internet: <http://dx.doi.org/10.4018/jdls.2010100101>.

KASPRZAK, J.; BRANDEJS, M. Improving the Reliability of the Plagiarism Detection System - Lab Report for PAN at CLEF 2010. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2010. p. 1–10.

KASPRZAK, J.; BRANDEJS, M.; KRIPAC, M. Finding Plagiarism by Evaluating Document Similarities. In: **Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misus**. [S.l.: s.n.], 2009. p. 24–28.

KONG, L. et al. Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection. In: **CLEF (Online Working Notes/Labs/Workshop)**. [S.l.: s.n.], 2012.

KONG, L. et al. Approaches for Source Retrieval and Text Alignment of Plagiarism Detection Notebook for PAN at CLEF 2013. In: **Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.** [S.l.: s.n.], 2013.

MALCOLM, J. A.; LANE, P. C. R. Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plagiarism Detector. In: **CEUR-WS.org**. [S.l.: s.n.], 2009. v. 502, p. 29–33.

MAURER, H.; KAPPE, F.; ZAKA, B. Plagiarism - A Survey. **J-JUCS**, v. 12, n. 8, p. 1050–1084, 2006.

MCCABE, D. L. Cheating Among College and University Students : A North American Perspective. **International Journal for Educational Integrity**, v. 1, n. 1996, 2005.

MENAI, M. E. B.; BAGAIS, M. APlag: A Plagiarism Checker for Arabic Texts. In: **Computer Science Education (ICCSE), 2011 6th International Conference on**. [S.l.: s.n.], 2011. p. 1379–1383.

MEUSCHKE, N.; GIPP, B.; BREITINGER, C. CitePlag: A Citation-based Plagiarism Detection System Prototype. In: **Proceedings of the 5th International Plagiarism Conference**. Newcastle upon Tyne, UK: [s.n.], 2012.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes**: Fundamentos e Aplicações. Barueri, SP, Brasil: Manole, 2003. Cap. 4.

PEREIRA, R. C.; MOREIRA, V. P.; GALANTE, R. Ufrgs@pan2010: Detecting external plagiarism - lab report for pan at clef 2010. In: **CLEF (Notebook Papers/LABs/Workshops)**. [S.l.: s.n.], 2010.

PERTILE, S.; MOREIRA, V. P. A test collection to evaluate plagiarism by missing or incorrect references. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2012. p. 141–143.

PERTILE, S.; MOREIRA, V. P.; ROSSO, P. Comparing and combining content and citation-based approaches for plagiarism detection. **Journal of the Association for Information Science and Technology (JASIST)**, 2015. A ser publicado.

PERTILE, S.; ROSSO, P.; MOREIRA, V. P. Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2013. p. 150–154.

PLATT, J. C. **Sequential Minimal Optimization**: A fast algorithm for training support vector machines. [S.l.], 1998.

POTTHAST, M. et al. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2012. (SIGIR '12), p. 1004–1004.

POTTHAST, M. et al. Cross-language plagiarism detection. **Language Resources And Evaluation**, Springer Netherlands, v. 45, n. 1, p. 45–62, 2010.

POTTHAST, M. et al. An Evaluation Framework for Plagiarism Detection. In: **Proceedings of the 23rd International Conference on Computational Linguistics: Posters**. [S.l.: s.n.], 2010. (COLING '10), p. 997–1005.

QUINLAN, R. **C4.5**: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

RITT, M. et al. An integer linear programming approach for approximate string comparison. **European Journal of Operational Research**, v. 198, n. 3, p. 706 – 714, 2009.

RIVEST, R. L. **The MD5 Message-Digest Algorithm**. 1992. Internet RFC 1321. Available from Internet: <http://tools.ietf.org/html/rfc1321>.

SANCHEZ-PEREZ, M. A.; SIDOROV, G.; GELBUKH, A. F. A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In: **Working Notes for CLEF 2014 Conference, Sheffield, UK, 2014.** [S.l.: s.n.], 2014. p. 1004–1011.

SOROKINA, D. et al. Plagiarism detection in arxiv. In: **Proceedings of the Sixth International Conference on Data Mining**. [S.l.]: IEEE Computer Society, 2006. p. 1070–1075.

STAMATATOS, E. Plagiarism Detection using Stopword n-Grams. **J. Am. Soc. Inf. Sci. Technol.**, v. 62, n. 12, p. 2512–2527, 2011.

STEIN, B.; EISSEN, S. M. zu. Intrinsic Plagiarism Detection. In: **ECIR**. [S.l.: s.n.], 2006. p. 565–569.

STEIN, B.; EISSEN, S. M. zu; POTTHAST, M. Strategies for retrieving plagiarized documents. In: **Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2007. (SIGIR '07), p. 825–826.

STEIN, B.; MEYER, S. Near Similarity Search and Plagiarism Analysis. In: **GFKL**. [S.l.: s.n.], 2006. p. 430–437.

STOFFEL, A. et al. Enhancing Document Sstructure Analysis using Visual Analytics. In: **Proceedings of the 2010 ACM Symposium on Applied Computing**. [S.l.: s.n.], 2010. (SAC '10), p. 8–12.

TORREJÓN, D. A. R.; RAMOS, J. M. M. CoReMo System (Contextual Reference Monotony) - Lab Report for PAN at CLEF 2010. In: **CLEF (Notebook Papers/Labs/Workshops)**. [S.l.: s.n.], 2010.

TORREJÓN, D. A. R.; RAMOS, J. M. M. Text Alignment Module in CoReMo 2.1 Plagiarism Detector - Notebook for PAN at CLEF 2013. In: **CLEF (Notebook Papers/Labs/Workshop)**. [S.l.: s.n.], 2013.

VOORHEES, E. M.; HARMAN, D. Overview of the Sixth Text Retrieval Conference (TREC-6). In: **The Fifth Text REtrieval Conference (TREC-5). NIST Special Publication 500-238, National Institute of Standards and Technology**. [S.l.: s.n.], 1998. p. 347–366.

WALKER, J. Measuring Plagiarism: Researching What Students Do, Not What They Say They Do. **Studies in Higher Education**, v. 35, n. 1, p. 41–59, 2010.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: Practical machine learning tools and techniques. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560.

YOUMANS, R. J. Does the Adoption of Plagiarism-detection Software in Higher Education Reduce Plagiarism?. **Studies in Higher Education**, v. 36, n. 7, p. 749–761, 2011. ISSN 03075079.

ZHANG, H.; CHOW, T. W. A Coarse-to-fine Framework to Eefficiently Thwart Plagiarism. **Pattern Recognition**, v. 44, n. 2, p. 471–487, 2011.

ZHANG, Y. CrossCheck: an Effective Tool for Detecting Plagiarism. **Learned Publishing**, v. 23, p. 9–14, 2010.

APÊNDICE A —

A.1 Experimentos e Resultados da Métrica de Análise de Coocorrências em Citações

Para execução dos experimentos seria ideal uma coleção real de artigos científicos em que alguns deles tivessem casos de plágio. No entanto, tal coleção não existe e não seria simples criar uma, pois fazer julgamentos de plágio em trabalhos reais é bastante problemático.

Sendo assim, recorremos a um conjunto de dados artificial (ALZAHRANI et al., 2012). O arquivo de anotações desta coleção nos permite verificar se a abordagem de analisar Coocorrências em Citações é um bom indicador de plágio.

Nestes experimentos, consideramos apenas citações no estilo numerado. Sendo assim, foi necessário selecionar a partir da coleção apenas documentos com este estilo de citação. Foram comparados 2218 documentos suspeitos com 6035 documentos originais.

A fim de segmentar as referências, contamos com a ferramenta Ondux (CORTEZ et al., 2010), que representa o estado da arte em extração de informações por segmentação de texto. Para comparar as referências entre os documentos, foi utilizada uma extensão da distância de edição, chamada Carla (RITT et al., 2009), o que trata inversões de cadeias de caracteres (o que ocorre com frequência nas referências). O limiar de similaridade usado foi $t = 0.86$, com base em observações empíricas.

As coocorrências dentro do documento foram computadas deslizando janelas pelos documentos. Essas janelas foram conFiguradas em tamanhos $s = 5$, $s = 15$ e $s = 30$. De acordo com a Tabela A.1, pode-se notar que a janela mais pequena (isto é, com 5 linhas) proporcionou uma precisão mais alta (56,67%), o que significa que, na maioria dos casos em que foram encontrados coocorrências, de fato correspondem a casos de plágio. Os demais casos com coocorrências que não foram considerados como sendo plágio, foram devido a três principais razões: (i) o documento suspeito havia citado e referenciado o documento original de onde o texto foi copiado; (ii) duas referências similares foram erroneamente tratadas pelo método como sendo a mesma. Por exemplo, para as referências "Jemain, A.A., A.I. Al-Omari and K. Ibrahim, Multistage median ranked set sampling for estimating the population median" e "Jemain, A. A. and Al-Omari, A. I., Multistage median ranked setsamples for estimating the population mean" foi atribuído um índice de similaridade superior ao limiar estabelecido, e assim, foram considerados como sendo a mesma referência; ou (iii) trabalhos do mesmo autor e sobre o mesmo tópico tinham um nível elevado de coocorrências de co-citação, mas não eram considerados como plágio.

Um outro fator que também pode ter afetado os resultados, é que parágrafos do documento suspeito e do documento original da coleção tinham conteúdos idênticos e não estavam anotados como

Tabela A.1 – Resultados da análise de coocorrência em citações

	S=5	S=15	S=30
Coocorrências em citações	90	154	161
Plágio com coocorrências	51	76	64
Precisão	0.5667	0.4935	0.3975
Revocação	0.0220	0.0328	0.0277
F1	0.0424	0.0616	0.0517

casos de plágio. Acreditamos que estes casos correspondem a casos reais de auto-plágio e que não estão incluídos no arquivo de anotações, uma vez que são registrados apenas os casos de plágio artificiais que foram inseridos automaticamente.

Em termos de revocação, cerca de 3% dos casos de plágio foram identificados (s=15). A principal razão para a baixa revocação é que, na maior parte dos casos de plágio nesta coleção, o fragmento de texto copiado do documento original não incluem quaisquer referências. Além disso, em alguns casos, o fragmento plagiado tinha sido extraído a partir de um artigo de uma área totalmente diferente (por exemplo, o documento original de um texto plagiado na área de economia era da área de veterinária). Em tais casos, é muito pouco provável que o documento original e o suspeito iriam partilhar quaisquer referências.

Além disso, é interessante notar que, quase todos os casos com Coocorrências em Citações eram casos com paráfrases. Isso mostra que a análise de coocorrência em citações pode ajudar a identificar passagens plagiadas em que o texto tenha sido ofuscado, que normalmente é problemático para a detecção baseada em conteúdo (BARRÓN-CEDEÑO et al., 2013).

APÊNDICE B —

B.1 Coleção de teste para avaliar Plágio por Referência ausente ou incorreta

Como apresentado no Capítulo 3, várias estratégias e ferramentas para auxiliar na tarefa de detecção de plágio foram desenvolvidas. Sendo assim, para avaliar tais estratégias, coleções de teste se tornam necessárias. Geralmente, essas coleções são compostas de um corpus contendo os documentos e um conjunto de anotações que irá permitir a avaliação da qualidade de cada método de detecção. Neste contexto, criamos a coleção PlaMIR (PERTILE; MOREIRA, 2012), uma coleção de teste com casos de plágio por referência ausente e incorreta.

A estratégia utilizada para criar a coleção foi gerar documentos artificiais (ou seja, artigos acadêmicos) e inserir neles passagens (plagiadas ou não) de outros artigos. A coleção de teste PlaMIR é composta por quatro partes: (i) um corpus com documentos originais; (ii) registros bibliográficos para os documentos originais; (iii) um corpus com documentos suspeitos; e (iv) um arquivo de anotação descrevendo onde foram inseridas as passagens e se elas são consideradas plágio. Note que não podemos distribuir os documentos originais uma vez que são trabalhos de pesquisa reais protegidos por direitos autorais. No entanto, para tornar a coleção usável, fornecemos links para esses artigos. A coleção PlaMIR está disponível a partir de: <http://www.inf.ufrgs.br/~slpertile/plamir.html>. A Tabela B.1 mostra o número de documentos em cada corpus da coleção.

Tabela B.1 – Características da Coleção de Teste PlaMIR

Documentos	***Número de Documentos***
Suspeitos	1000
Originais	963
Registros Bibliográficos	963

A partir dos 1000 documentos suspeitos, 818 contêm casos de plágio. Os 182 documentos restantes não contêm passagens plagiadas. Nestes casos, é adicionada uma citação no texto adjacente à passagem inserida, e uma referência para o documento original correspondente é incluída no bloco de referências do documento suspeito. A seguir são descritas em detalhes as etapas executadas na construção da coleção de teste PlaMIR:

- **Geração dos documentos suspeitos:** O primeiro passo foi gerar os 1000 artigos científicos na área de Ciência da Computação, utilizando a ferramenta SCIgen [1]. SCIgen gera artigos científicos na área de Ciência da Computação, incluindo, Figuras, citações e referências a trabalhos não-existentes. Os artigos gerados não fazem qualquer sentido, mas eles são gramaticalmente corretos

[1] <http://pdos.csail.mit.edu/scigen/>

e bem formatados. Estes documentos passaram por uma etapa de pós-processamento na qual gráficos e imagens foram removidos, preservando apenas o conteúdo textual do documento.

- **Obtenção dos documentos originais:** Para obter os documentos originais a partir dos quais as passagens seriam extraídas, 963 artigos científicos (PDF) foram coletados aleatoriamente a partir da base de dados DBLP[2], juntamente com seus registros bibliográficos (XML).
- **Inserção de passagens nos documentos suspeitos:** Uma vez que o conjunto de documentos suspeitos e originais são obtidos, casos de plágio artificiais são simulados nos documentos suspeitos. Para tal inserção, o processo que se segue foi aplicado:

  1. Seleção aleatória dos documentos originais que serão utilizados para extrair passagens artificiais plagiadas e não plagiadas. Note que cada documento suspeito pode ter passagens retiradas de até 10 documentos originais. De acordo com a Tabela B.2, a maioria dos documentos suspeitos receberam passagens selecionadas a partir de 4 a 7 documentos originais (42% dos documentos).

Tabela B.2 – Número de documentos originais por documento suspeito

Documentos originais por supeito	Estatísticas da coleção
1-3	294 (29%)
4-7	413 (42%)
8-10	293 (29%)

  2. De cada documento original, foram escolhidas aleatoriamente um número de passagens a serem inseridas nos documentos suspeitos. Note que até 5 passagens foram retiradas de cada documento original e cada documento suspeito pode ter até 15 passagens plagiadas e 15 passagens não plagiadas.

     O tamanho das passagens corresponde a uma citação, desde o início da sentença até seu ponto final. Um documento suspeito pode receber simultaneamente passagens plagiadas e não plagiadas. A Tabela B.3 mostra que 946 documentos suspeitos receberam passagens não plagiadas e 818 documentos suspeitos receberam passagens plagiadas.

     Para cada passagem, escolhemos aleatoriamente se seria um caso não plagiado, ou um caso de plágio por referência ausente ou referência incorreta. A Tabela B.4 mostra que 774 documentos suspeitos receberam passagens plagiadas por referência ausente e 710 receberam passagens plagiadas por referência incorreta.

[2] <http://www.informatik.uni-trier.de/~ley/db/>

Tabela B.3 – Número de passagens por documento suspeito

Passagens por documento supeito	Passagens não plagiadas	Passagens plagiadas
1-3	205 (22%)	176 (22%)
4-7	312 (32%)	254 (31%)
8-11	216 (23%)	193 (23%)
12-15	213 (23%)	195 (24%)

Tabela B.4 – Número de passagens plagiadas por referência ausente e incorreta por documento suspeito

Passagens por documento supeito	Passagens plagiadas	
	Referência ausente	Referência incorreta
1-3	319 (41%)	389 (55%)
4-7	322 (42%)	281 (40%)
8-10	125 (14%)	40 (5%)
11-15	8 (3%)	-

3. Após as passagens serem selecionadas, elas foram extraídas e inseridas numa posição aleatória no documento suspeito. O comprimento da passagem, o documento original que foi retirado e a posição onde foi inserida no documento suspeito são cruciais para avaliar os sistemas de detecção. Deste modo, essas características são registradas em um arquivo de anotação.

   Quando a passagem não é plagiada, a referência bibliográfica correspondente para o documento original é inserida no bloco de referências do documento suspeito. Em casos de plágio por referência incorreta, a passagem recebe informações que nos permitem identificar a referência para o documento original. Nestes casos, a referência incluída no bloco de referências é extraída de um outro documento original, não correspondente à citada na passagem.

4. Finalmente, quando todas as passagens são inseridas no documento suspeito, um arquivo de anotações é criado. Essas anotações são um componente-chave da coleção de teste, sem elas não se pode avaliar um sistema de detecção de plágio. Cada documento suspeito tem o seu correspondente arquivo de anotações com os seguintes campos:

   - **documento_suspeito:** indica o documento suspeito para o qual a anotação é relacionada. .
   - **deslocamento_documento_suspeito:** indica a posição de partida (em caracteres), onde a passagem foi inserida no documento suspeito.
   - **comprimento_documento_suspeito:** indica o comprimento (em caracteres) da passagem inserida no documento suspeito.

```
<?xml version=1.0= encoding=UTF_8?>
<documento suspeito='Documento102.pdf'>
<características_documento_original
comprimento_documento_suspeito="185"
deslocamento_documento_suspeito="9699"
documento_original="Documento425.pdf"
documento_original_referenciado="Documento680.pdf"
deslocamento_documento_original="5753"
comprimento_deslocamento_original="185"
tipo="referência_incorreta"
</document>
```

Figura B.1 – Exemplo de um arquivo de anotação da coleção PlaMIR

- **documento_original:** indica o documento original a partir do qual a passagem foi extraída.
- **documento_original_referenciado:** indica o documento original usado para referenciar uma passagem.
- **deslocamento_documento_original:** indica a posição de partida (em caracteres) da passagem no documento original.
- **comprimento_documento_original:** indica o comprimento (em caracteres) de passagem no documento original.
- **Tipo:** indica se uma passagem é um caso de plágio por referência ausente ou incorreta, ou se não é um caso de plágio.

Note que quando a passagem representa um caso de plágio por referência incorreta, o campo *documento_original_referenciado* é diferente do campo *documento_original*. Quando a passagem é um caso de plágio por referência ausente o campo *documento_original_referenciado* recebe valor *nulo*. As anotações são definidas no mesmo formato utilizado na competição do PAN e são geradas em XML. Um exemplo de um arquivo de anotação para um documento suspeito é mostrado na Figura B.1.

Para os documentos suspeitos que não tem casos de plágio, foi necessário simular apenas passagens não plagiadas. Neste caso, as passagens dos documentos originais foram inseridas em conjunto com suas respectivas referências.