UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

VINÍCIUS DE BONA FARINON

Avaliação experimental de métodos de desambiguação de autores em bibliotecas digitais

Trabalho de Graduação.

Prof. Dr. Carlos A. Heuser
Orientador

Porto Alegre, janeiro de 2009.

AGRADECIMENTOS

Agradeço aos meus pais Ademir Farinon e Dolores Regina De Bona Farinon pela oportunidade de realizar a graduação, não só pela ajuda financeira, mas também pela motivação e pelo exemplo que eles são para mim.

Agradeço também ao meu irmão Bruno De Bona Farinon por ser um grande amigo e companheiro.

Também gostaria de agradecer à minha namorada Morgana Gualdi Laux, que além de ser uma grande amiga e motivadora, é também um exemplo de dedicação e força de vontade.

Por fim, gostaria de agradecer a ajuda do professor Carlos Alberto Heuser durante o ano de 2009 para a realização desse trabalho.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

BDBComp	Biblioteca Digital Brasileira de Computação
DBLP	Digital Bibliography & Library Project
DLs	Digital Libraries (Bibliotecas Digitais)
TREC	Text REtrieval Conference
XML	eXtensible Markup Language

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Devido a grande diversidade de fontes de dados utilizadas pela maioria das bibliotecas digitais (DLs), podem existir problemas de ambigüidade em suas bases de dados. Pensando em melhorar esse quadro, esse trabalho propõe uma heurística que busca amenizar um problema de ambigüidade de nomes de autores bastante comum em DLs chamado *split citation*.

Esse problema ocorre quando um autor possui seu nome representado de maneiras distintas nas diferentes publicações de sua autoria. Dessa forma, cada uma dessas variações de nome, podem ser consideradas como pessoas diferentes, dividindo a produção de um determinado autor. O *split citation* é um problema bastante corriqueiro, pois é muito comum, por exemplo, a abreviação ou até mesmo a supressão de sobrenomes muito extensos, além de outras práticas que geram variações de um nome.

Para corrigir esse inconveniente, é feita uma análise em cima das diferentes informações contidas em uma publicação para decidir a sua autoria. Com as informações extraídas dos registros das bases de dados, são feitas medidas de similaridade que, ao final do processo, servem para ordenar uma lista de publicações onde as primeiras posições devem representar as publicações relevantes do autor em questão. Essas medidas de similaridade são calculadas utilizando evidências presentes nas publicações em forma de *metadados*, como por exemplo, os nomes dos autores, nomes dos co-autores, títulos, veículos de publicação, etc.

A heurística foi avaliada em termos de precisão e revocação com a ajuda do software chamado trec_eval, disponibilizado pela conferência TREC (Text REtrieval Conference) que apóia pesquisas na área de recuperação de informação. Esse software permite a análise da precisão do método para diferentes níveis de revocação, e com isso facilita também a comparação entre as variantes do método proposto.

Palavras-Chave: desambiguação, similaridade, precisão, revocação, bibliotecas digitais, split citation, trec_eval.

Experimental evaluation of methods for authors disambiguation in digital libraries

ABSTRACT

Due to the wide variety of data sources used by most digital libraries, there may be problems of ambiguity in their databases. Thinking of improving this situation, this paper proposes a heuristic method that seeks to alleviate a authors name ambiguity problem that is very common in digital libraries called split citation.

This problem occurs when an author has its name represented in different ways in his different publications. Thus, each of these name's variations can be considered as different authors, dividing the production of a particular author. The split citation is a fairly common problem. It is very common, for example, shortening or even elimination of long last names, and other practices that generate name variations.

To correct this drawback, an analysis is made on different information contained in a publication to decide on his own. With the information obtained from the records of the databases, are made similarity measures that, at the end of the process, are used to sort a list of publications where the top positions must represent the relevant publications of the author in question. These similarity measures are calculated using evidence from the publications in the form of metadata such as authors' names, co-authors' names, titles, publication venue, etc.

The heuristic method was evaluated in terms of recall/precision with the help of software called trec_eval, made available by the conference TREC (Text REtrieval Conference) which supports research in information retrieval. This software allows the analysis of the accuracy for different levels of recall, and it also facilitates the comparison between the proposed method variants.

Keywords: disambiguation, similarity, precision, recall, digital libraries, split citation, trec_eval.

1 INTRODUÇÃO

As DLs são importantes sistemas de informação e estão se tornando cada vez mais complexas (Gonçalves et al., 2004). Essa complexidade se torna maior, principalmente pela maneira como essas bibliotecas são alimentadas. As fontes de dados são diversas e cada uma dessas diferentes fontes pode adotar padrões diferentes de representação para um mesmo objeto. Repositórios de publicações científicas, por exemplo, disponibilizam o auto-arquivamento, ou seja, um serviço que permite a submissão de objetos digitais para as bibliotecas pelo próprio autor (Silva et al., 2007). Adicionalmente, esses repositórios, coletam dados em páginas da Internet e também de outros repositórios. Essa diversidade de fontes torna o sistema bastante dinâmico e completo, porém, traz complexidade e problemas de redundância e ambigüidade dos dados.

A comunidade acadêmica tem utilizado amplamente os serviços disponibilizados pelas DLs. Bibliotecas como DBLP (Digital Bibliography & Library Project) e BDBComp (Biblioteca Digital Brasileira de Computação), são cada vez mais populares e utilizadas para pesquisas bibliográficas. Os estudos baseados nas publicações dessas bibliotecas revelam resultados interessantes sobre o impacto das publicações, sobre os temas mais discutidos pela comunidade, sobre a qualidade das publicações, além de revelar tendências e padrões de colaboração em redes sociais. Esse tipo de estudo pressupõe que as bases de dados possuam informações de qualidade, caso contrário, os resultados serão distorcidos. De fato, como a coleta de dados é feita por meio de diferentes fontes que não utilizam necessariamente os mesmos padrões, é praticamente impossível fugir de problemas relacionados à redundância e ambigüidade de informações.

Dos diferentes tipos de ambigüidade que podem surgir em uma biblioteca digital, a ambigüidade de nomes de autores é uma das mais comuns e talvez um dos problemas mais difíceis de se lidar. A ambigüidade de nomes traz diversos problemas, sendo que os mais importantes são os relacionados à autoria de artigos. Basicamente, no contexto de citações bibliográficas, o problema de ambigüidade de nomes pode ser dividido em dois subproblemas (Lee et al., 2005) conhecidos como *split citation* (um autor possui variações do seu nome) e o *mixed citation* (autores com nomes iguais). O problema *split citation* acontece quando um autor possui seu nome representado de diversas maneiras nas diferentes publicações de sua autoria cadastradas na base de dados. Dessa forma ele pode ter sua produção dividida. Na Figura 1.1 podemos visualizar um exemplo retirado da biblioteca digital BDBComp onde a produção de um autor está dividida entre duas diferentes variações de grafia do seu nome.

Figura 1.1: Exemplo de *split citation* retirado da biblioteca digital BDBComp.

Já o problema conhecido como *mixed citation*, acontece quando dois autores distintos são identificados pelo mesmo nome, trazendo erros de autoria de publicações.

Nesse trabalho é proposta uma heurística que faz a análise das diferentes informações contidas em uma publicação para decidir se essa publicação é de autoria de um determinado autor, dessa forma, corrigindo o problema *split citation*. O método gera uma lista para cada autor contendo as publicações presentes na base de dados ordenada da maior similaridade para a menor, de maneira que as primeiras colocações da lista representam as publicações relevantes para o autor em questão. Essa similaridade é calculada utilizando as evidências presentes nas publicações em forma de *metadados* (dados sobre outros dados como por exemplo: co-autores, titulo, veículo de publicação, ano de publicação, etc) para chegar a um valor de similaridade entre dois autores.

Para avaliar os resultados da heurística, foi utilizado o software chamado trec_eval, disponibilizado pela conferência chamada TREC (Text REtrieval Conference) que apóia pesquisas na área de recuperação de informação.

Para fazer comparações entre os resultados obtidos com os experimentos e verificar se o método realmente contribui com a desambiguação dos autores, foi definido um *baseline*. Esse *baseline* representa o resultado da desambiguação das bases de dados utilizando apenas os nomes dos autores como evidência para o processo. Para o cálculo da similaridade entre dois nomes de autores foi utilizada a Distância de Levenshtein, também conhecida como Distância de Edição, que é um algoritmo já muito estudado e utilizado exaustivamente em diversas áreas como compressão de imagens, mineração de dados, reconhecimento de padrões, etc.

A grande maioria das DLs disponibiliza serviços de busca por publicações, e geralmente, essas ferramentas trabalham utilizando os nomes dos autores como indexador para as pesquisas. Dessa forma, o *baseline* definido está coerente com muitas ferramentas utilizadas no mundo real, permitindo verificar que o método proposto pode ser aplicado na prática em DLs. Os resultados dos experimentos mostram uma melhora significativa na desambiguação de nomes utilizando a abordagem proposta nesse trabalho em relação ao *baseline* que foi definido, mostrando que seria possível um

ganho na qualidade dos resultados desse tipo de ferramenta de busca e uma conseqüente melhora na qualidade do conteúdo dessas DLs.

Esse trabalho está organizado da seguinte maneira. O Capítulo 2 aborda o conceito de DLs utilizado no trabalho, descreve alguns conceitos básicos necessários para se entender o trabalho, e descreve o funcionamento de alguns trabalhos relacionados. O Capítulo 3 apresenta todos os detalhes e descreve o funcionamento do método proposto. São definidas aqui as bases de dados utilizadas nos experimentos, as medidas de similaridade feitas em cima das evidências encontradas nas publicações e são expostos os resultados dos experimentos. Por fim, o Capítulo 4 conclui o trabalho e apresenta algumas alternativas de melhorias para método para trabalhos futuros.

2 BIBLIOTECAS DIGITAIS, DESAMBIGUAÇÃO E CONCEITOS BÁSICOS

O problema da desambiguação de nomes, juntamente com o problema da deduplicação de registros são temas bastante estudados e possuem diversas propostas de solução. Essa é uma área que recebeu bastante atenção ultimamente, principalmente devido ao crescimento da popularidade e da importância das DLs. Muitas das propostas de solução dos problemas encontrados nas DLs possuem falhas, ou porque resolvem somente parte do problema, ou porque necessitam de informações que estão externas as DLs. Nesse capítulo serão descritas brevemente algumas propostas de solução existentes e suas diferenças em relação à heurística proposta por esse trabalho. Além disso, será explicado como funciona a ferramenta trec_eval, que é a ferramenta padrão utilizada pela conferência TREC (Text REtrieval Conference) para avaliar os resultados dos seus experimentos.

2.1 Bibliotecas Digitais

No contexto desse trabalho, podemos definir uma biblioteca digital como um repositório de informações sobre a produção cientifica de pesquisadores. Além das informações das publicações, também existem serviços associados a essas informações. Assim, é possível realizar pesquisas no conteúdo, filtrar os dados segundo alguns critérios estabelecidos, visualizar estatísticas sobre o conteúdo, e muitas vezes existe também a possibilidade de submissão de novas publicações, além de outros serviços possíveis. Para esse trabalho, vamos restringir bastante a definição de biblioteca digital. Será considerado o modelo de biblioteca digital utilizado pela biblioteca DBLP, que organiza as informações sobre as publicações por meio de *metadados* que registram informações sobre os nomes dos autores, os títulos de seus trabalhos, o ano e o veículo de publicação.

2.2 Desambiguação

A desambiguação de nomes de autores no caso de citações bibliográficas como acontece nas DLs, por exemplo, pode lidar basicamente com dois problemas: o chamado *mixed citation* e o *split citation*. O problema chamado de *mixed citation* acontece quando temos na biblioteca digital, vários autores sendo representados pelo mesmo nome. Já o problema conhecido como *split citation* é mais comum, e acontece quando um único autor possui diferentes variações de seu nome dentro da biblioteca digital. Existem trabalhos que tratam apenas do problema *mixed citation*, outros que

procuram resolver apenas o *split citation* e também, os que tentam lidar com os dois problemas simultaneamente.

Existem alguns trabalhos realizados na área de desambiguação de nomes de autores em DLs que trabalham com a abordagem de *clusters*. Um *cluster* é um grupo de registros compatíveis, ou seja, é um grupo que teoricamente pertenceria a um mesmo autor. Tipicamente, nesses trabalhos os *clusters* são formados aos poucos com base em medidas de similaridade calculadas em cima das evidências encontradas nos registros (nomes dos co-autores, título do trabalho, veículo de publicação, etc). Esses métodos fundem sucessivamente esses clusters considerados compatíveis. A informação dos clusters fundidos é então agregada fornecendo mais dados para a próxima rodada de fusões.

Um método que obteve resultados bastante satisfatórios foi proposto por Oliveira et al. (2005). Nesse trabalho ele propõe um método para a desambiguação de nomes que cria um índice unificado que registra todas as variações possíveis de nomes de autores existentes em uma biblioteca digital. Esse método usa um algoritmo de casamento de padrões chamado de *Fragment Comparsion* para agrupar nomes que apresentam algum grau de semelhança entre si. Esse algoritmo toma como entrada duas *strings*, correspondendo a dois nomes de autores, e compara cada um dos fragmentos individuais que compõe os nomes usando distância de edição, ignorando a ordem em que esses fragmentos aparecem na entrada. Se os dois nomes forem considerados similares para um determinado limiar, então outros dados como co-autor, títulos e veículos de publicação, são usadas como evidência adicional para determinar se os dois nomes correspondem a um mesmo autor e podem ser unidos em um mesmo *cluster*. Embora esta estratégia gere *clusters* de alta qualidade, ela tende a gerar muitos clusters quando não há evidencias suficiente para desambiguar um nome. Ou seja, quando não há evidências suficientes para a desambiguação, o problema *split citation* permanece presente na biblioteca digital, pois dois autores podem estar presentes em diferentes *clusters* mesmo que representem a mesma pessoa. Por isso, na abordagem que será proposta nesse trabalho, foi feita a escolha por um método que ao final do processo gera uma lista ordenada das publicações para cada autor. Dessa forma, mesmo que a base de dados não possua evidências suficientes para agrupar autores que representam uma mesma pessoa em um *cluster*, um autor terá uma lista de publicações onde as publicações de sua autoria terão grande chance de estarem bem posicionadas. Esse posicionamento se deve ao fato de que qualquer mínima evidência auxilia nas medidas de similaridade e consequentemente na ordenação da lista de publicações.

Pereira et al. (2009) realizou um trabalho onde a desambiguação dos nomes utiliza dados externos a base de dados da biblioteca digital. A idéia consiste em reunir informação das publicações e submeter consultas para um motor de busca na Internet a fim de encontrar os currículos e páginas que contenham publicações ou informações dos autores ambíguos. A partir do conteúdo dos documentos retornados pelo motor de busca, informações úteis que podem ajudar no processo de desambiguação são extraídas. Utilizando essas informações, a desambiguação dos nomes dos autores utiliza um método de agrupamento em *clusters* de forma parecida com a descrita anteriormente. Existem grandes desafios para que essa abordagem de fato funcione. É preciso formular muito bem as consultas para os motores de busca para que seja possível encontrar páginas que realmente sejam documentos pessoais de um autor, além de encontrar quais as páginas na Internet são mais importantes e de fato confiáveis como fonte de informação para o processo de desambiguação. Outro desafio é extrair do

resultado das consultas as informações relevantes para a desambiguação dos autores. Muitos casos de falhas foram encontrados devido a citações não encontradas na Internet. Uma possível causa desse problema são erros de ortografia em mais de uma palavra em títulos de publicações. Outro problema acontece quando citações somente são encontradas em páginas de conteúdo de outras DLs. Se uma citação aparece somente em uma biblioteca digital, o método gera *clusters* fragmentados. Além disso, DLs também possuem erros que podem ocasionar em agrupamentos incorretos de *clusters*. Para evitar esses problemas, a proposta desse trabalho utiliza exclusivamente os dados contidos na própria base de dados da biblioteca digital evitando assim, a inserção de conteúdo que eventualmente pode estar em discordância com o conteúdo presente nas bases de dados.

Algumas soluções propostas fazem uso de redes sociais (Malin, 2005). Uma rede social é gerada para nomes ambíguos. Cada nó nesta rede corresponde a um nome diferente e dois nós de nomes são ligados uns aos outros se eles co-ocorrem em pelo menos uma publicação. Para capturar o número de co-ocorrências, as arestas ligando dois nós são ponderadas de acordo com o número de publicações nas quais seus correspondentes nomes co-ocorrem. Então, caminhamentos aleatórios sobre o grafo gerado são executados. Cada caminhamento começa em um nó com um nome ambíguo e prossegue até que ou um nó de nome ambíguo é atingido, ou um número máximo de passos é executado. Depois de um certo número de passos, a probabilidade de se chegar a um nó B dado um caminhamento originado em um nó A, é estimada e utilizada para determinar a similaridade entre A e B, de modo que os nós não ambíguos podem ser removidos do grafo. A similaridade é então utilizada em um processo de agrupamento em *clusters* para remover arestas cujas similaridades estiverem abaixo de um valor limiar, de maneira que cada sub-grafo resultante corresponde a um autor em particular.

Técnicas de aprendizado de máquina também têm sido extensivamente utilizadas para tratar do problema da desambiguação de nomes em citações de autores. Han et al. (2004) propõem duas abordagens baseadas em técnicas de aprendizado supervisionado que usam nomes de co-autores, títulos e veículos de publicação como evidências para a desambiguação.

2.3 Similaridade entre strings

Para trabalhar com a similaridade dos nomes dos autores, existem diversas alternativas. Essa é uma área muito estudada e explorada por diferentes comunidades como, a comunidade de banco de dados, a de estatística e também a comunidade de inteligência artificial, por exemplo. Para se ter uma noção do tempo que se vem estudando esse tema, a distância de edição, por exemplo, já era utilizada em 1965 pelo russo Vladimir Levenshtein. Existe uma vasta bibliografia sobre a similaridade de *strings*, porém, esse tema foge do escopo do trabalho. Por isso, apesar de existirem métodos comprovadamente melhores (Cohen et al., 2003), a versão mais geral da distância de edição foi escolhida por ser poderosa o suficiente para uma quantidade muito grande de aplicações, além de não ser complexa (Fonseca; Reis, 2002) e será utilizada para realizar as medidas de similaridade entre nomes de autores.

O problema de verificar a similaridade entre dois nomes pode ser traduzido como o problema do casamento aproximado de caracteres. Na sua forma mais geral, o problema pode ser descrito como a tarefa de encontrar entre dois nomes, um determinado padrão,

permitindo um número de erros entre eles. Tipicamente, considera-se um modelo de erros onde é permitida a remoção, a inserção ou a substituição de caracteres de ambos os nomes que estão sendo comparados, e cada operação representa um custo. A distância de edição é o custo total das operações realizadas para transformar um nome em outro. Existem muitos métodos que propõe melhorias para a distância de edição e também métodos híbridos que mesclam a distância de edição com outros métodos de similaridade de *strings*, entretanto, essas variações são muito dependentes do tipo de erro considerado, e como foi dito anteriormente, esse não é o foco do trabalho.

Foram realizadas algumas alterações para que a Distância de Levenshtein retornasse um número sempre entre zero e um, ou seja, um valor normalizado. Também foi preciso alterações para que esse valor passasse a representar a similaridade entre dois nomes e não a distância. Sabe-se que a Distância de Levenshtein é no máximo igual ao tamanho do nome mais longo, por isso, para retornar um valor sempre entre zero e um, a Distância de Levenshtein é dividida pelo tamanho do maior nome. Dessa forma, o resultado para dois nomes idênticos seria zero, e para dois nomes completamente diferentes, o resultado seria um, representando a menor e a maior distância possíveis respectivamente. Para que esse resultado represente a similaridade entre dois nomes, ainda é preciso mais uma alteração. Um seria a maior similaridade possível entre dois nomes, e a Distância de Levenshtein um representa a menor similaridade possível. Então, basta subtrair de um, a distância normalizada.

2.4 Precisão / Revocação

As medidas mais comuns para avaliar a qualidade de um sistema de busca e recuperação de informação são conhecidas com precisão e revocação. Elas foram originalmente propostos por Kent et al. (1955) e são medidas utilizadas para avaliar a eficácia de um sistema de recuperação de informações, ou seja, elas medem a habilidade do sistema de recuperar os documentos relevantes e, ao mesmo tempo, de evitar os não relevantes (van Rijsbergen, 1979). Precisão é a fração de documentos recuperados que é relevante (Baeza-Yates; Ribeiro-Neto, 1999), ou seja, é uma medida da capacidade do sistema de recuperar somente documentos relevantes. Revocação é a fração de documentos relevantes recuperados (Baeza-Yates; Ribeiro-Neto, 1999), ou seja, é uma medida da capacidade do sistema de recuperar todos os documentos relevantes.

Os valores de precisão e revocação são calculados assumindo-se que todos os registros do conjunto resposta foram analisados pelo usuário, sem levar em conta a ordenação do resultado. No entanto, o usuário analisa os documentos a partir do topo da lista ordenada de registros do conjunto resposta, o que implica que os valores de precisão e revocação variam à medida que o usuário prossegue a sua análise da lista ordenada. No caso desse trabalho, o resultado do processo de desambiguação sempre retorna todos os registros de uma base de dados ordenados por similaridade. Dessa forma, a análise da precisão só tem sentido se for feita ao longo da lista ordenada para vários níveis de revocação. Uma forma comum de apresentar os resultados é por meio de gráficos de precisão versus revocação para vários níveis de revocação. Por exemplo, a precisão é calculada quando 10% dos registros relevantes são analisados, quando 20% dos registros relevantes são analisados, e assim por diante, até que 100% dos registros

relevantes sejam analisados. Esses gráficos são obtidos por meio do cálculo da precisão média em pontos padrão de revocação, tais como 10% (0.1), 20% (0.2), etc.

Os níveis de revocação para as várias consultas podem ser diferentes dos níveis de revocação padrão (0.1, 0.2, 0.3, etc). Utiliza-se então, a interpolação para calcular a precisão média nos níveis padrão de revocação.

Nesse trabalho, para um determinado autor, uma publicação é de sua autoria ou não é. Não existe meio termo. Essa abordagem de tratar a autoria de forma binária é exatamente a mesma idéia utilizada em avaliações de sistemas de recuperação de informações, onde um determinado documento recuperado é considerado relevante ou não, permitindo a confecção de gráficos de precisão versus revocação.

2.5 Trec_eval

A conferência Text REtrieval Conference foi iniciada em 1992 com o objetivo de apoiar a investigação na área de recuperação de informação, fornecendo a infra-estrutura necessária para a avaliação em larga escala de metodologias de recuperação de texto. Os objetivos da Text REtrieval Conference são os seguintes: incentivar a investigação na recuperação de informação em grandes bases de dados; criação de um fórum aberto para facilitar a troca de idéias e pesquisas entre indústria, academia e governo; acelerar a transferência da tecnologia desenvolvida em laboratórios para produtos comerciais e problemas do mundo real; aumentar a disponibilidade de técnicas de avaliação apropriadas para uso tanto na indústria quanto nas universidades, incluindo o desenvolvimento de novas técnicas de avaliação mais aplicáveis aos sistemas atuais.

A conferência é supervisionada por um comitê composto por representantes do governo, indústria e academia. Em cada edição da conferência, são fornecidos conjuntos de testes sobre os quais os participantes executam seus próprios sistemas de recuperação de dados e retornam seus resultados em forma de uma lista de documentos recuperados ordenados que são avaliados em termos de precisão versus revocação. Ao final, acontece um workshop onde os participantes podem compartilhar suas experiências.

O software utilizado para fazer a avaliação é chamado de trec_eval e está disponível para a comunidade em geral para que qualquer pessoa possa avaliar seus sistemas de recuperação a qualquer instante.

2.5.1 Formato dos arquivos de entrada do software trec_eval

Para avaliar o método de desambiguação proposto nesse trabalho, foi utilizado o software trec_eval. Para uma determinada base de dados, o método proposto retorna para cada autor, uma lista de publicações contendo todas as publicações da base ordenadas de maneira que as primeiras posições representam as publicações que tem mais chance de pertencer ao autor em questão e as últimas, por sua vez, tem menos chance de pertencer a esse autor. Um dos arquivos de entrada do programa trec_eval, possui a informação que representa, para cada autor da base, quais são as publicações relevantes. Ou seja, esse arquivo indica quais as publicações devem estar nas primeiras

posições na ordenação da lista para cada autor. O outro arquivo é a saída gerada pelo método após o processo de desambiguação, ou seja, esse arquivo possui uma lista de publicações, contendo todas as publicações, para cada autor, ordenadas pela similaridade calculada pelo método. Na Figura 2.1 podemos verificar o exemplo de um arquivo de entrada com a informação das publicações relevantes para cada autor presente em uma base de dados.

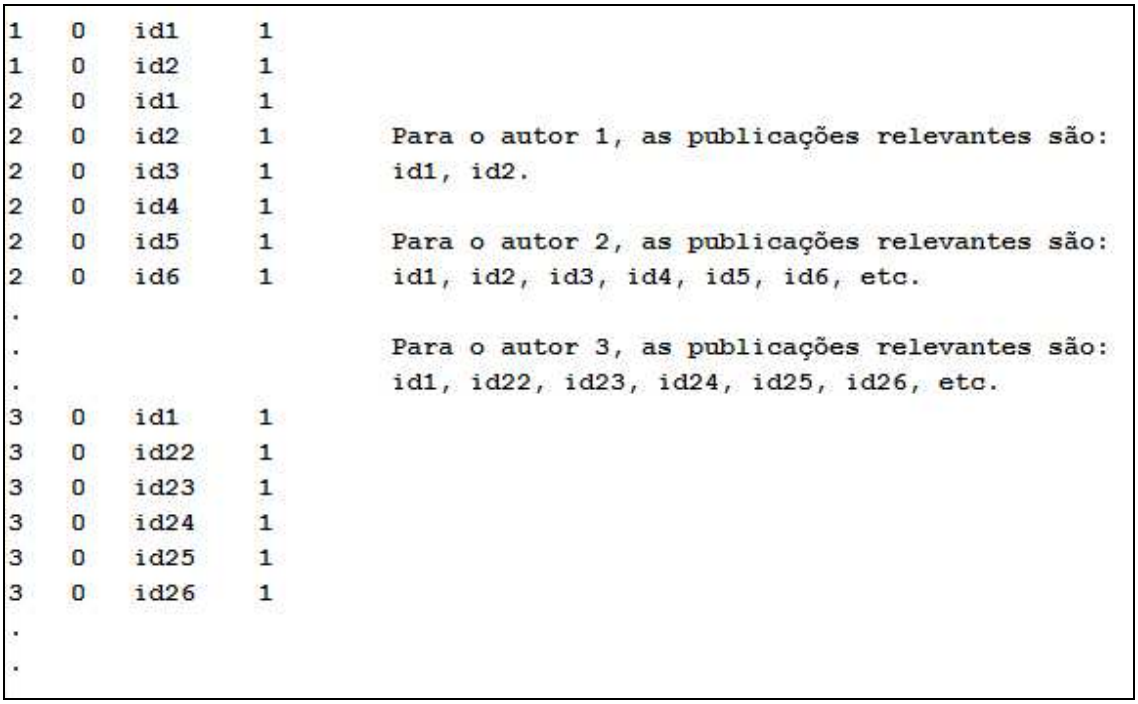

Figura 2.1: Arquivo de entrada para o software trec_eval contendo a informação de relevância de cada dos artigos para cada autor da base de dados.

Após o processamento do método de desambiguação, o segundo arquivo de entrada para o software trec_eval é gerado. Na Figura 2.2 podemos ver um exemplo de arquivo gerado pelo processamento em cima de uma base de dados. Esse arquivo contém a ordenação das publicações para cada autor por ordem de similaridade e deve servir de entrada para o programa. Baseado no arquivo que contém as publicações relevantes para cada autor e nesse arquivo de saída do processo, o software trec_eval vai avaliar o sistema gerando uma tabela com as medidas de precisão para os níveis padrão de revocação.

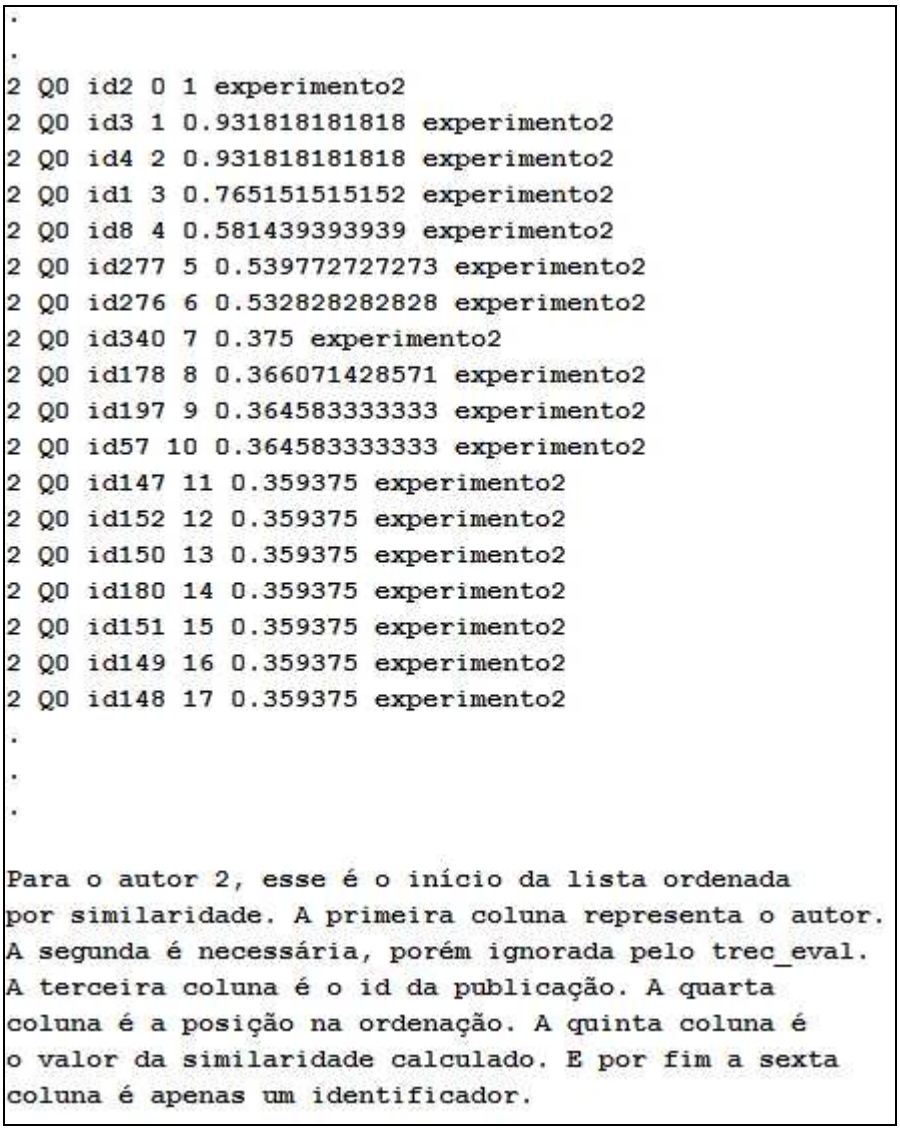

Figura 2.2: Exemplo de arquivo de entrada para o software trec_eval, contendo uma lista ordenada de publicações para cada autor, resultado do processo de desambiguação realizado em cima de uma base de dados.

2.5.2 Resultado do processamento do software trec_eval

O programa trec_eval tem como resultado de sua execução uma tabela mostrando a precisão para cada nível de revocação. Na Figura 2.3 é possível ver um exemplo da interface de saída do desse software. O resultado é uma tabela de precisão versus revocação interpolada a qual podemos transcrever facilmente para um gráfico, de maneira que fique mais simples a anáise dos resultados.

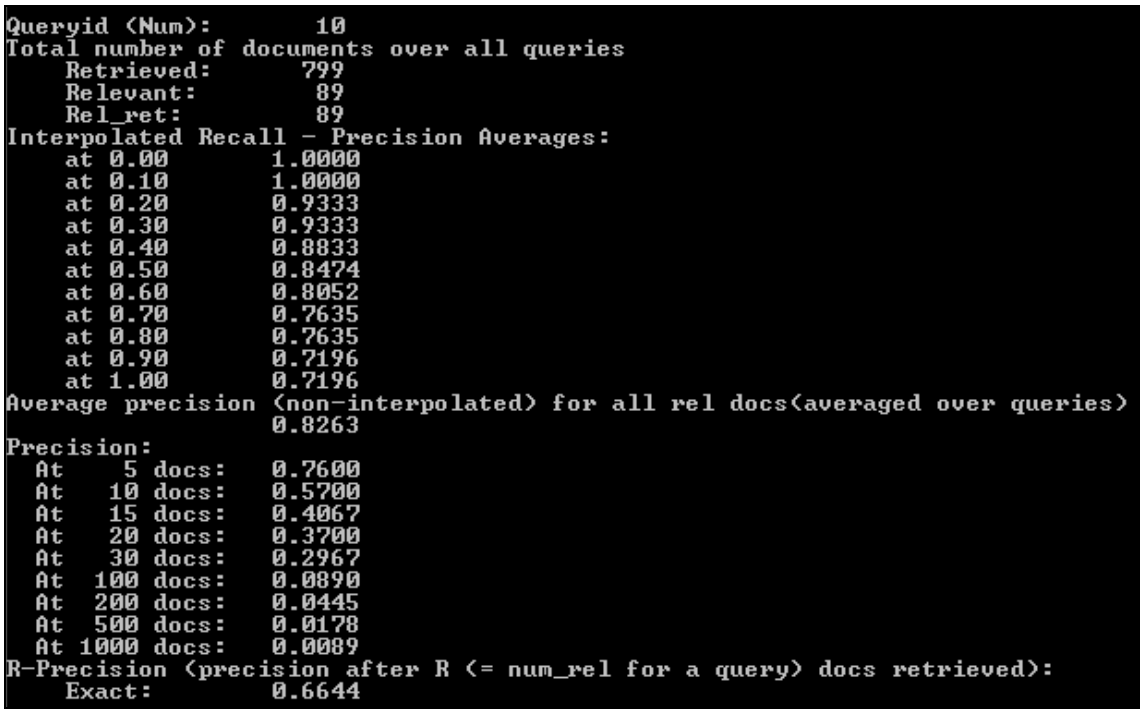

Figura 2.3: Interface do software trec_eval após a sua execução.

3 MÉTODO DE DESAMBIGUAÇÃO

Nesse capitulo será descrito com todos os detalhes a execução do método de desambiguação. Todas as medidas de similaridade realizadas com base nas evidências encontradas em cada publicação serão detalhadas e além disso, será explicado também todo o processo de confecção das bases de dados utilizadas nos experimentos. Ao final, serão mostrados os resultados obtidos e suas particularidades.

3.1 Setup

Aqui serão descritas exatamente como são as bases de testes utilizadas nos experimentos e também será esmiuçado o método explicando todas as suas particularidades.

3.1.1 Datasets

Para realizar a avaliação da heurística proposta nesse trabalho, é necessário possuir uma base de dados onde para cada autor presente, já se conheçam todas as publicações relevantes. O método deve posicionar essas publicações nas primeiras posições da lista. Além disso, para uma avaliação coerente, é necessário que a base de dados tenha um tamanho razoável.

A biblioteca BDBComp possui atualmente mais de 5 mil publicações cadastradas em sua base de dados, já a biblioteca digital DBLP possui mais de 1,2 milhões de publicações em seu acervo. A DBLP disponibiliza para *download* em sua página oficial na Internet, a sua base de dados completa no formato XML. Como para coleções tão grandes é impossível conhecer todas as publicações relevantes para um determinado autor, foi necessário trabalhar com um subconjunto de publicações dessas bases. Inicialmente, foi criada uma base de dados contendo 80 publicações e 160 nomes diferentes de autores, baseado no material das duas bibliotecas. Nessa base não existe redundância de informações, porém dentre os autores, existem muitos casos em que um autor possui variações do seu nome, representando o problema *split citation*. Esta primeira base de dados será chamada de **Dataset 1**.

Mais tarde, surgiu a necessidade de uma base maior e melhor elaborada, a fim de verificar a eficácia do método em uma base que fosse mais próxima de uma base de dados encontrada em sistemas reais. Essa segunda base de dados contém 361 publicações e 820 diferentes nomes de autores e é um subconjunto da biblioteca digital BDBComp, que disponibiliza informações dos trabalhos da comunidade brasileira de computação. Esse *dataset* foi fornecido pelo Laboratório de Banco de Dados da UFMG (Universidade Federal de Minas Gerais). Essa segunda base de dados será chamada de **Dataset 2**.

Essas bases de dados foram devidamente analisadas de maneira que para cada autor nelas presente, é possível saber quais são as publicações relevantes. Dessa forma é possível avaliar o desempenho do método ao final dos experimentos.

O formato das bases de dados utilizadas nos experimentos é o formato XML. A base de dados disponibilizada pelo DBLP encontra-se nesse formato e como inicialmente o Dataset 1 era um subconjunto do DBLP, esse padrão foi mantido ao longo dos experimentos. Além disso, o formato XML foi o escolhido, pois é um formato bastante utilizado e conhecido, possuindo ampla compatibilidade e facilidades de manipulação em muitas linguagens de programação. Na Figura 3.1 é possível visualizar um exemplo de um registro retirado de uma dessas bases de dados no formato XML.

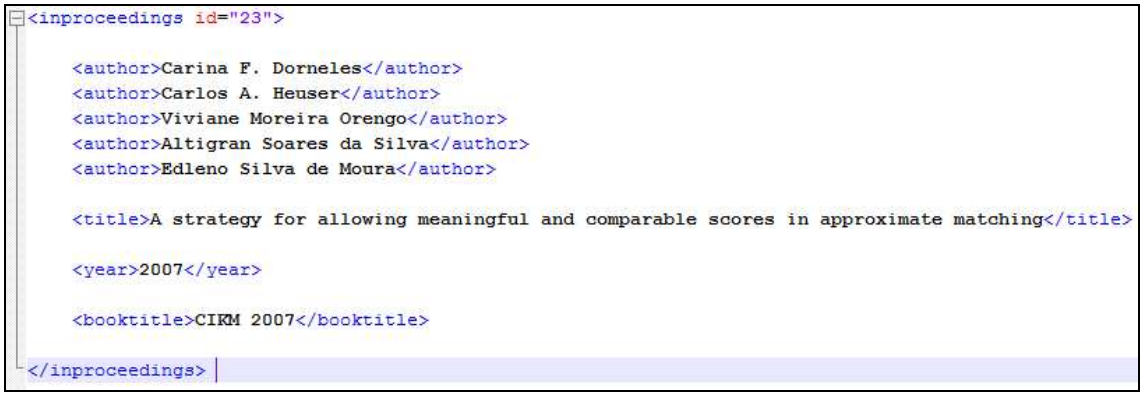

Figura 3.1: Formato do arquivo XML das bases de dados utilizadas nos experimentos (Dataset 1 e Dataset 2).

3.1.2 Procedimento

O procedimento funciona para qualquer base de dados que esteja no mesmo formato da base de dados do DBLP, ou seja, deve ser um documento XML e que possua a estrutura de *tags* idêntica à utilizada pelo DBLP.

Na estrutura XML em questão, cada publicação possui *metadados* representando os nomes dos seus autores, o seu título, o ano em que foi publicada, e o veículo responsável pela publicação. Dessa estrutura, podemos extrair algumas informações sobre cada um dos autores presentes. Além do seu nome, um autor presente em um registro nessa base de dados pode ter informações sobre seus co-autores, pode ter também informações importantes que ajudem na desambiguação no título da publicação e também no veículo de divulgação. As informações sobre os autores que podem ser extraídas dos registros das bases de dados foram divididas nesse trabalho em 4 componentes: nome, conjunto de co-autores, conjunto de títulos e conjunto de veículos. Para cada um dos componentes, foram criados métodos que funcionam de forma isolada e calculam a similaridade de um determinado autor com os autores das publicações da base de dados. Com esse cálculo eu consigo saber, para um determinado autor, qual nome presente em uma publicação é o mais similar ao seu. Ou seja, dentre os autores de um artigo, qual o que tem mais chance de ser o autor em questão. Isso permite que para esse autor eu ordene as publicações por ordem de similaridade, onde as primeiras posições representam os artigos relevantes, ou seja, os que tem mais chance de ser de sua autoria. Dessa forma, mesmo que um autor possua variações de seu nome com

similaridades baixas, existem outras medidas de similaridade que podem auxiliar no cálculo de similaridade geral e conseqüentemente ajudar a ordenação correta das publicações.

O procedimento em si começa com a base de dados sendo percorrida a fim de coletar todos os diferentes nomes de autores presentes. Aqui, independente de representar a mesma pessoa ou não, cada nome distinto é considerado um autor diferente. O nome dos autores é o primeiro componente extraído das publicações que é utilizado para fazer a desambiguação dos autores.

Após o algoritmo percorrer todas as publicações presentes, temos armazenados todos os nomes de autores da base de dados. Para ordenar a lista de publicações para um determinado autor utilizando somente o componente nome, percorremos todas as publicações presentes na base de dados novamente, comparando o autor da busca com cada autor presente em cada publicação. Essa comparação é feita utilizando a Distância de Levenshtein normalizada descrita no capítulo anterior, o que vai resultar em um valor entre zero e um que representa a similaridade do nome do autor da busca com cada um dos autores da publicação. Para exemplificar, vamos imaginar que estamos ordenando as publicações de maneira que queremos a produção do autor Carlos Alberto Heuser, nas primeiras posições. Percorrendo a base de dados encontramos a publicação descrita na Figura 3.2.

```
<inproceedings id="24">

    <author>Carlos A. Heuser</author>
    <author>Francisco N. A. Krieser</author>
    <author>Viviane Moreira Orengo</author>

    <title>SimEval - A Tool for Evaluating the Quality of Similarity Functions</title>

    <year>2007</year>

    <booktitle>ER (Tutorials, Posters, Panels e Industrial Contributions)</booktitle>

</inproceedings>
```

Figura 3.2: Exemplo de publicação presente nas bases de dados no formato XML.

A similaridade resultante da comparação entre Carlos Alberto Heuser e Carlos A. Heuser é aproximadamente 0.71, comparando com Francisco N. A. Krieser, a similaridade fica em 0.26, e por fim, comparando com Viviane Moreira Orengo a similaridade é de 0.14. Levando em consideração somente os nomes dos autores, podemos considerar que o nome Carlos Alberto Heuser é mais similar a Carlos A. Heuser nesse registro. Portanto, podemos concluir que dentre os três autores do artigo, o que tem mais chance de ser a mesma pessoa que Carlos Alberto Heuser é o primeiro autor, ou seja, Carlos A. Heuser. A pontuação dessa publicação (id= 24), utilizada para a posterior ordenação dos registros é 0.71. Sabemos que Carlos Alberto Heuser e Carlos A. Heuser representam a mesma pessoa, portanto, essa é uma das publicações relevantes para esse autor e por isso ela deve ficar a frente das publicações não relevantes na lista ordenada ao final do processo.

Supondo que o processo siga em frente e encontremos a publicação descrita na Figura 3.3.

```
<inproceedings id="61">

    <author>Sidnei Silveira</author>
    <author>Dante Augusto Couto Barone</author>

    <title>Formacao de Grupos Colaborativos utilizando Algoritmos Geneticos</title>

    <year>2004</year>

    <booktitle>XV Simposio Brasileiro de Informatica na Educacao</booktitle>

</inproceedings>
```

Figura 3.3: Exemplo de publicação presente nas bases de dados no formato XML.

Nesse caso, dentre os autores da publicação, o nome mais similar a Carlos Alberto Heuser é Dante Augusto Couto Barone, com similaridade 0.27. Então, a pontuação desse artigo para a ordenação é 0.27. Sabemos que esse não é um artigo relevante para a nossa pesquisa, e de fato, a pontuação dele foi inferior à publicação anterior. Considerando o resultado final do método, a ordenação teria a publicação com id 24 em uma posição superior à publicação com id 61. Esse era o resultado esperado, já que o primeiro artigo pertence ao autor Carlos Alberto Heuser e o segundo artigo não.

No exemplo anterior, utilizando apenas os nomes dos autores para a desambiguação, o resultado foi correto, porém em alguns casos, essa abordagem pode gerar erros. Existem casos em que um autor tem seu nome escrito de diversas maneiras bem diferentes, resultando em similaridades baixas para algumas comparações, e em contra partida, existem nomes de pessoas diferentes que podem gerar resultados de similaridade bastante altos. Também há casos em que o autor muda de nome (nome de solteiro pode ser diferente do nome de casado) e fica com sua produção dividida entre esses dois nomes.

Devido a essas situações os resultados podem não ser muito bons ao se utilizar apenas os nomes como evidência para a desambiguação. Por isso a idéia da heurística proposta é considerar os outros componentes que podem ser extraídos dos registros presentes nas bases de dados, aumentando assim a quantidade de infomação que pode ser utilizada para a desambiguação. O *baseline* considerado é exatamente a ordenação e a avaliação utilizando apenas os nomes dos autores para o processo e a seguir veremos que ele pode ser superado ao se utilizar as demais evidências. Após fazer a ordenação utilizando o restante dos componentes, será possível verificar a melhora nos resultados.

A próxima medida de similaridade calculada pelo método é a similaridade entre os co-autores do autor da busca e os co-autores dos autores das publicações. Um autor que publica trabalhos em uma determinada área, geralmente trabalha com outros autores dessa mesma área. Esse autor pode ter seu nome escrito de diversas maneiras nas várias publicações em que ele participou, porém, é bastante provável que essas várias publicações tenham alguns autores em comum.

A Figura 3.4 mostra um caso onde dois nomes de autores diferentes, porém que representam a mesma pessoa, tem um co-autor em comum (Edgar Jamhour).

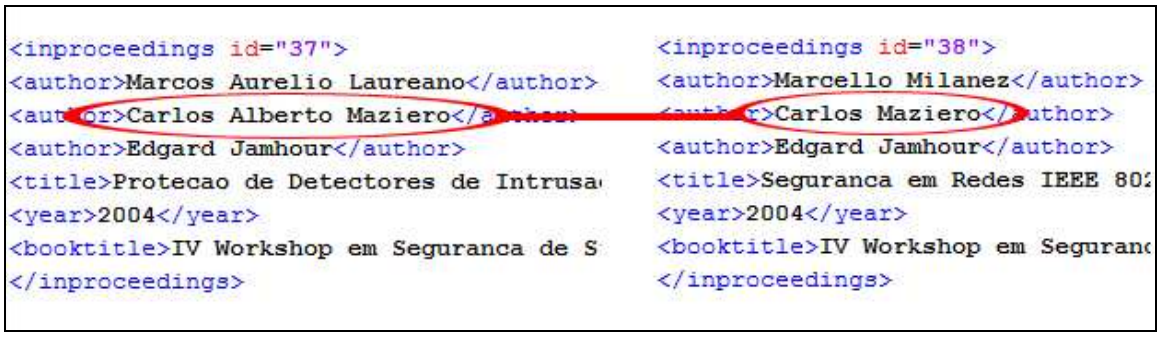

Figura 3.4: Duas publicações de uma mesma pessoa, porém com variações do seu nome. Podemos perceber que os autores destacados têm um co-autor em comum.

Cada autor da base tem um conjunto de co-autores. Para descobrir esses co-autores percorremos a base de dados buscando por publicações de um determinado autor e ao encontrar uma publicação desse autor, os demais nomes presentes são identificados como seus co-autores. Supondo que estamos coletando os co-autores do autor Carlos Alberto Heuser, vamos percorrer a base de dados à procura de publicações em que ele seja um dos autores. Identificando essas publicações, o restante dos autores que aparecem em suas publicações são seus co-autores. Esse processo de busca de co-autores é exemplificado na Figura 3.5.

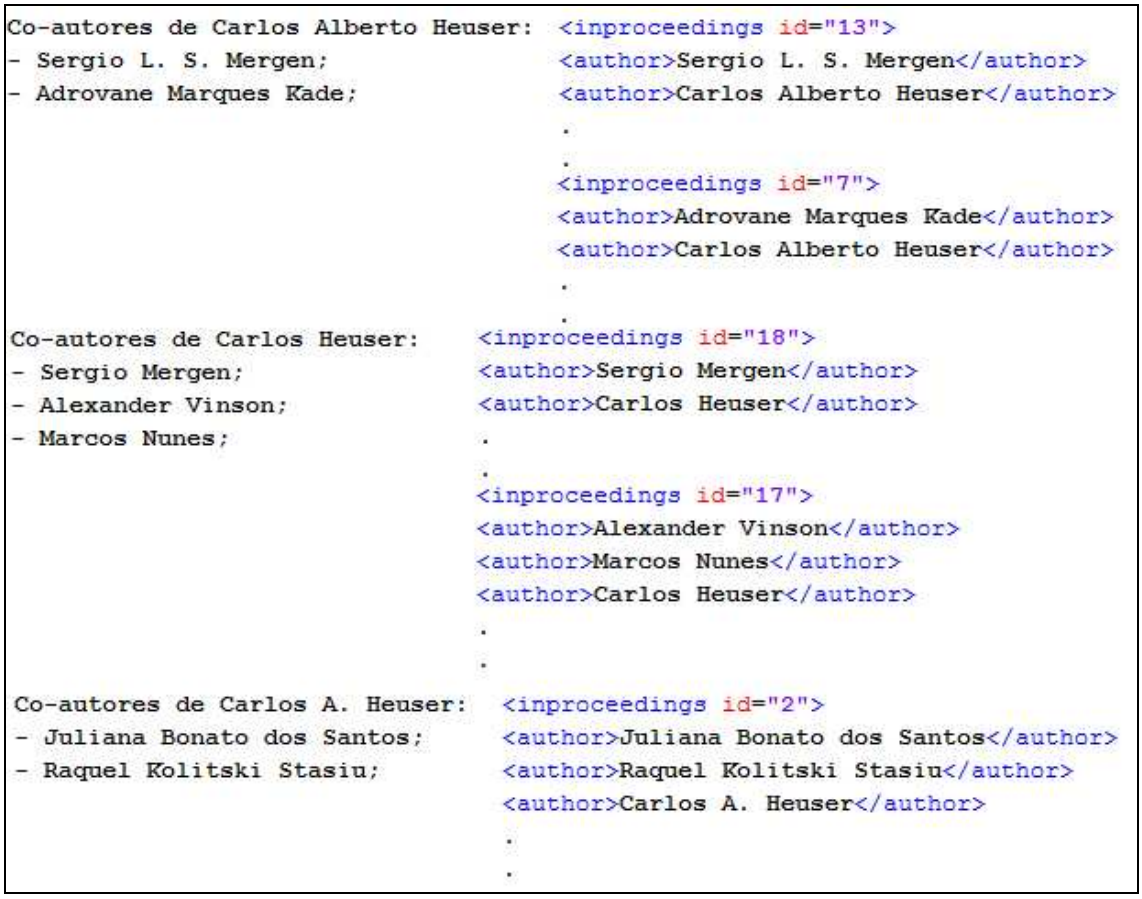

Figura 3.5: Identificação de co-autores nos registros de uma base de dados.

Novamente, vamos comparar o autor da busca com cada um dos autores em cada uma das publicações. A diferença é que dessa vez, a comparação dos autores é feita utilizando o conjunto de co-autores de cada um ao invés de utilizar os seus nomes. Primeiramente, é necessário verificar se os autores que estão em comparação não são co-autores entre si. Caso eles sejam co-autores, está comprovado que eles não podem representar a mesma pessoa e a similaridade resultante é zero. Caso eles não sejam co-autores, aí partimos para a comparação dos seus conjuntos de co-autores para verificar o quanto esses conjuntos tem em comum. A comparação dos conjuntos de co-autores é feita de maneira que todos os co-autores de um autor sejam comparados com todos os co-autores do outro. Para contabilizar pontos nessa medida de similaridade, não é necessário que dois co-autores tenham nomes idênticos. A exemplo da medida de similaridade com os nomes, aqui fazemos uso também da Distância de Levenshtein. Se o nome de dois co-autores possuírem similaridade superior a um limite pré-estabelecido (o limite utilizado nos experimentos é 0.7), esse valor será agregado e posteriormente dividido pelo número de comparações que ultrapassaram esse limite. Ou seja, a medida de similaridade dos co-autores é a média dos resultados de similaridade das comparações dos conjuntos de co-autores que ultrapassaram o valor limite. Esse valor limite foi definido por meio de testes e observação dos resultados.

O terceiro componente utilizado para auxiliar a desambiguação dos autores são os títulos das publicações. Cada publicação presente em uma base de dados possui um título. Os títulos das publicações geralmente revelam informações importantes sobre as áreas ou assuntos que um determinado autor estuda. Por exemplo, o título "**Semiautomatic Generation of Data-Extraction Ontologies from Relational Databases**" é de uma publicação do autor Carlos Alberto Heuser. Já o título "**UXQuery: building updatable XML views over relational databases**" é de outra publicação dessa vez de autoria de Carlos A. Heuser. Esses autores tem nomes diferentes, porém representam a mesma pessoa. Podemos verificar que a palavra **databases** aparece nos dois títulos. Essa palavra revela que os dois autores trabalham com **databases**, portanto, essa é mais uma evidência que pode auxiliar na correta ordenação das publicações de um autor. O fato de dois autores possuírem nomes parecidos e além disso ter mesma área de interesse, pode ser um indício de que esses autores representem a mesma pessoa. Pode acontecer de diferentes pessoas possuírem palavras em comum em seus títulos, porém é bem provável que nas outras medidas de similaridade (utilizando nomes, co-autores e veículos) o resultado não seja tão bom, impedindo que essa coincidência afete o resultado final.

A comparação entre os títulos dos autores é feita da seguinte maneira. Cada autor presente na base de dados possui um conjunto de títulos, que são todos os títulos de suas publicações. Quando o método está percorrendo a base de dados para elaborar a ordenação das publicações para um determinado autor, ele compara as palavras chaves extraídas do conjunto de títulos desse autor com as palavras chaves extraídas do conjunto de títulos dos autores das publicações.

Na Figura 3.6 podemos visualizar um exemplo simples dos conjuntos de palavras chave extraídos dos títulos dos artigos. É possível também perceber nesse exemplo, que os autores possuem palavras chave em comum.

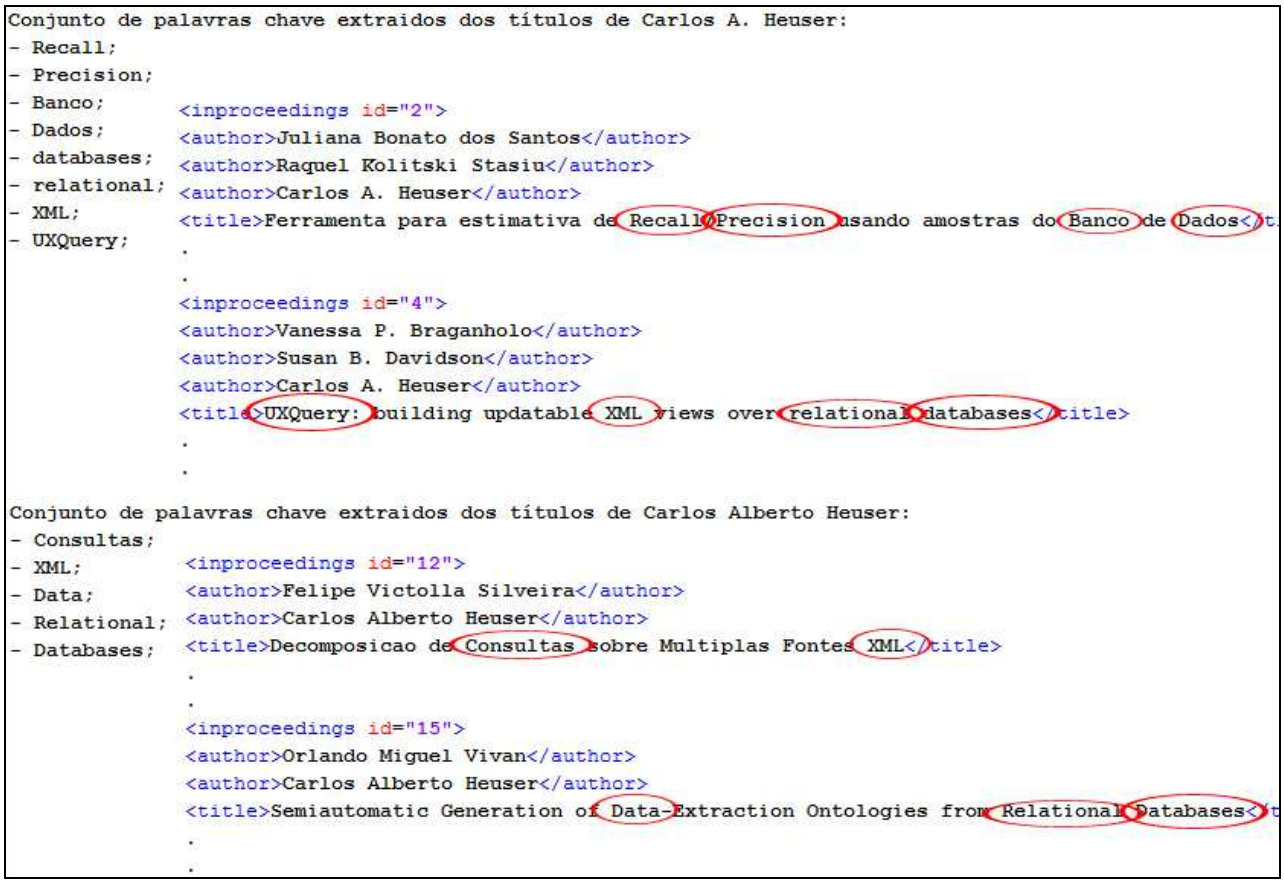

Figura 3.6: Conjuntos de palavras chave extraídas dos títulos de dois autores.

Para extrair essas palavras chave do conjunto de títulos de um autor, foi elaborado um algoritmo que inicialmente retira todos os símbolos e caracteres de pontuação presentes nos títulos e após isso, elimina as palavras mais comuns, tanto da língua portuguesa como da língua inglesa. São eliminadas preposições, pronomes, artigos, etc, além de palavras genéricas que podem estar presentes em títulos de qualquer área de estudo tais como sistema, ferramenta, gerenciar, automatizar, implementar, e suas derivações e equivalentes na língua inglesa. Dessa maneira, para cada autor, temos um conjunto de palavras chave que representam os termos centrais de seus trabalhos que identificam na maioria das vezes os interesses desses autores. Para comparar o conjunto de palavras chave de dois autores, fazemos a comparação de todas as palavras do conjunto de um com todas as palavras do conjunto do outro, ou seja, fazemos a intersecção dos dois conjuntos. Sabemos que o número máximo possível de palavras comuns entre os autores é igual ao tamanho do menor conjunto. Por isso, a medida de similaridade é o resultado da divisão do número de palavras comuns (intersecção) encontradas nessa comparação pelo tamanho do menor conjunto. Caso o número máximo de palavras em comum seja encontrado, o resultado da medida de similaridade será um, ou seja, a maior similaridade possível. Um exemplo desse cálculo está presente na Figura 3.7.

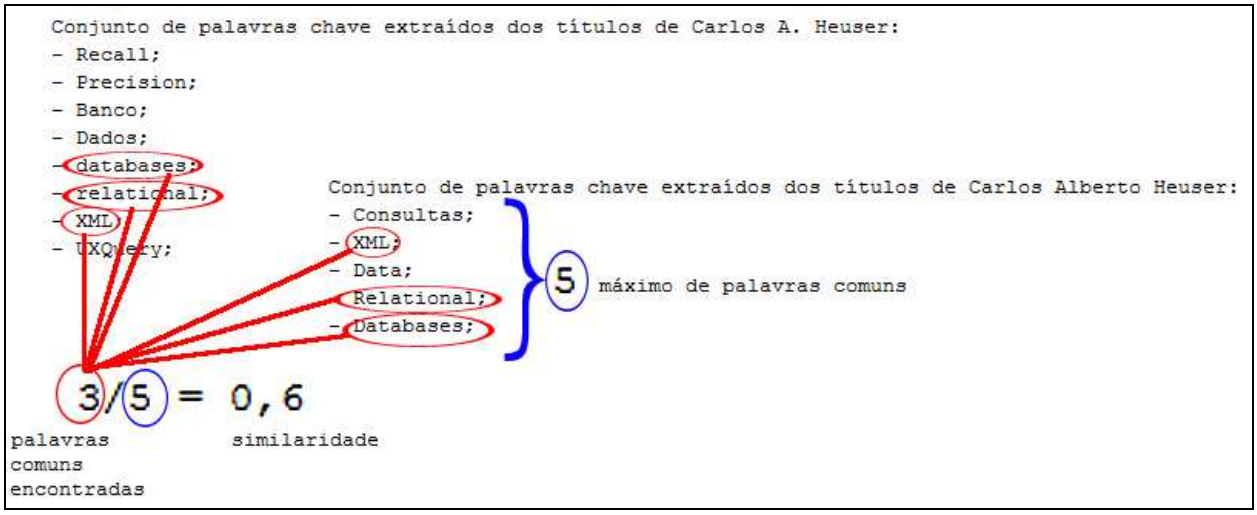

Figura 3.7: Cálculo de similaridade entre as palavras chave extraídas dos títulos de dois autores.

Os *metadados* presentes nas DLs, além de possuir os nomes dos autores e os títulos dos artigos, possuem também informação sobre os veículos de publicação. Os nomes desses veículos de publicação também podem conter pistas sobre a área de trabalho dos autores. Para trabalhar com o quarto componente extraído dos registros das bases de dados, foi utilizado o mesmo método utilizado sobre os títulos das publicações, ou seja, extrair palavras chave do conjunto de veículos dos autores. Dessa maneira, um determinado autor possui um conjunto de palavras chave, extraídas do conjunto de veículos presentes nas publicações as quais ele é autor. Para exemplificar, vamos analisar o exemplo para o autor Carlos Alberto Heuser. Alguns veículos presentes nas suas publicações são os seguintes: "**VII Workshop de Teses e Dissertações em Banco de Dados**", "**II Escola Regional de Banco de Dados**", "**XIX Simposio Brasileiro de Banco de Dados**". Já para o autor Carlos A. Heuser, que representa a mesma pessoa, encontramos veículos tais como: "**XXIII Simposio Brasileiro de Banco de Dados**", "**I Escola Regional de Banco de Dados**", etc. Ainda representando a mesma pessoa, temos o autor Carlos Heuser que dentre seu conjunto de veículos está presente o seguinte: "**III Escola Regional de Banco de Dados**". Nesse pequeno conjunto de veículos desses três autores, podemos perceber a repetição do termo **Banco de dados**. É claro que as publicações desses autores possuem outros veículos que não possuem esse termo e que talvez não tenham nenhuma palavra em comum, mas quando extrairmos as palavras chave desses autores, dentro do conjunto de termos resultante, teremos pelo menos as palavras **Banco** e **dados** para todos os três autores. Da mesma maneira que acontece com os títulos, quando existem termos em comum no conjunto de palavras chave, consideramos uma evidência que pode ajudar a definir se esses nomes representam a mesma pessoa.

Assim como acontece na extração de palavras chave dos títulos dos artigos, aqui também é preciso eliminar algumas palavras genéricas, palavras comuns e também símbolos e pontuação. Para os veículos, palavras como Simpósio, Workshop, Journal, Encontro, etc, não fazem sentido e por isso são eliminadas. Além disso, nos veículos aparecem outros problemas, como o número que representa a edição de uma convenção, por exemplo. Nos exemplos de veículos descritos anteriormente, podemos verificar a presença de algarismos romanos para representar as edições dos eventos. Essa

representação pode ocorrer também com números decimais e em alguns casos, também há a ocorrência do ano em que aquele encontro ocorreu. Para eliminar esse tipo de informação que aparece em alguns veículos e não são úteis nesse caso, foram criadas duas expressões regulares que, dada uma *string*, identificam esses números. Ao identificar um número, o algoritmo o elimina do conjunto de palavras chave. Ao final dessa fase de obtenção das palavras chave, o cálculo é feito da mesma maneira que é feito para os títulos, o número de palavras em comum é dividido pelo número de palavras do menor conjunto entre os dois autores.

Nesse ponto, temos como calcular uma medida de similaridade isolada para o nome dos autores, para os co-autores, para os títulos e por fim, para os veículos. Dessa maneira, é possível ordenar as publicações para cada autor da base, utilizando essas medidas isoladamente e também fazendo combinações entre elas. As publicações da base de dados ficam ordenadas de maneira que as publicações melhores posicionadas têm mais chance de pertencer a esse autor.

Após os cálculos de similaridade e posterior ordenação, um algoritmo transforma essa lista ordenada em entrada para o programa trec_eval, seguindo a sintaxe exigida pelo software que fará a avaliação desse resultado.

Com a saída do programa trec_eval, é possível analisar as medidas de similaridade isoladamente e também de maneira combinada, verificando assim, quais as evidências que mais ajudam a melhorar os resultados em relação ao resultado *baseline* definido. A partir dos resultados, foram elaborados gráficos de precisão versus revocação para cada medida de similaridade isoladamente de modo que fosse possível verificar se realmente faz sentido utilizar essas medidas para realizar a desambiguação. Após isso, foram realizados experimentos onde a similaridade dos nomes (baseline) foi combinada com cada uma das outras medidas de similaridade. O intuito desses experimentos era visualizar graficamente, quais as medidas de similaridade que mais contribuem para a melhoria do processo de desambiguação.

O resultado final da heurística de desambiguação é composto pela média de todas as medidas de similaridade. Comparando o resultado *baseline*, que é obtido utilizando apenas a similaridade de nomes, com o resultado final, é possível verificar uma melhora bastante considerável. Além disso, analisando os resultados isolados e também combinados, é possível definir pesos para cada medida de similaridade, resultando em uma média ponderada que melhora ainda mais a ordenação.

Para realizar os experimentos, primeiramente foi executado o método *baseline* para todos os autores presentes nas bases de dados. Com a ferramenta trec_eval, foi possível verificar os resultados de precisão versus revocação para cada autor isoladamente, com isso, foi possível identificar os autores que tinham problemas de nomes ambíguos. Com esse conjunto de autores que não tem um bom ordenamento ao se utilizar somente a similaridade de nomes, é possível visualizar se realmente o cálculo feito em cima das demais evidências está de fato auxiliando na ordenação. Esse conjunto de autores escolhido representa um pior caso em uma base de dados, pois para esse conjunto o resultado da desambiguação não é bom apenas utilizando a similaridade de nomes de autores.

3.2 Resultados

A seguir serão descritos os resultados obtidos utilizando as diferentes variações do método proposto nesse trabalho. Esses resultados estão expostos em forma de tabela (saída do software trec_eval) e também em forma de gráficos. Tanto as tabelas como os gráficos estão sempre mostrando os resultados do processo em forma de comparação com os resultados obtidos com o *baseline*. Além disso, os resultados da desambiguação para as duas bases de testes (Dataset 1 e Dataset 2) estão sempre dispostos lado a lado para que seja possível também essa comparação entre bases de tamanhos diferentes.

3.2.1 Baseline

Primeiramente, vamos analisar os resultados que são obtidos utilizando apenas os nomes dos autores para a desambiguação das bases de dados. Como descrito anteriormente, esse é o método *baseline* que faz a ordenação dos artigos para cada autor com base no cálculo de similaridade dos nomes dos autores que por sua vez é feito utilizando uma variação da Distância de Levenshtein. Esse processo é feito para que depois seja possível verificar se, ao agregarmos novas medidas de similaridade em cima das demais evidencias, ocorrerão melhorias nos resultados.

Para facilitar a análise, foi escolhido um grupo de autores em que o processo de desambiguação somente utilizando os nomes não era satisfatório. Isso se deve a problemas do tipo *split citation*. Para cada uma das bases de teste, foi escolhido esse grupo de autores e foram feitos os gráficos *baseline*, que ao longo dos experimentos serão utilizados para fins de comparação.

A Tabela 3.1 mostra os valores de precisão versus revocação obtidos como resultado do processo de desambiguação *baseline* nas duas bases de testes Esses mesmo dados foram transcrevidos para dois gráficos representados na Figura 3.8.

Tabela 3.1: Valores de precisão atingidos para os diversos níveis de revocação utilizando o método baseline para a desambiguação.

Revocação	Precisão Dataset 1	Precisão Dataset 2
0	1,0000	0,8666
0,1	1,0000	0,8597
0,2	0,9333	0,8377
0,3	0,9333	0,7819
0,4	0,8833	0,7802
0,5	0,8474	0,7627
0,6	0,8052	0,5141
0,7	0,7635	0,4806
0,8	0,7635	0,3978
0,9	0,7196	0,3708
1	0,7196	0,3406

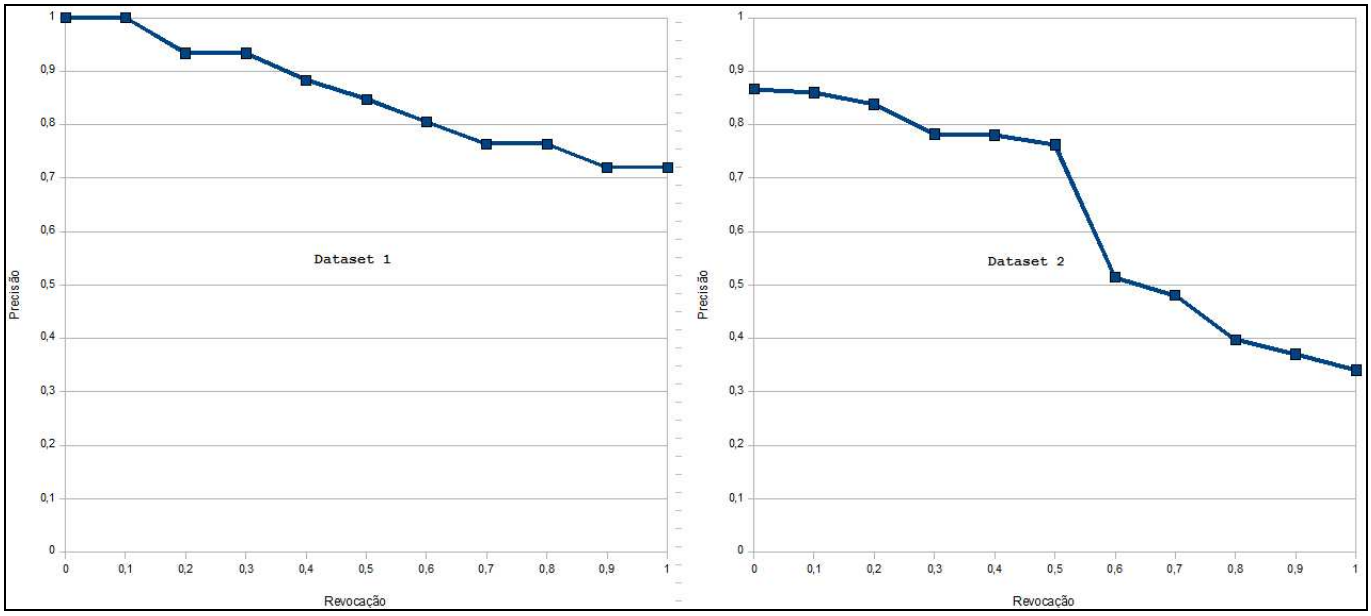

Figura 3.8: Resultado do método baseline transcrito em gráficos de precisão versus revocação.

3.2.2 Medidas de similaridade isoladas

Foram realizados experimentos utilizando cada uma das medidas de similaridade e avaliando-as separadamente. Após o cálculo isolado de cada medida, foi feita a

confecção de gráficos com os resultados para que fosse possível realizar uma análise do comportamento de cada medida, permitindo assim, a elaboração da média geral ponderada ao final do processo. A Tabela 3.2 contém os valores de precisão obtidos para as diferentes medidas de similaridade e a Figura 3.9 contém os gráficos para esses valores de precisão em relação à diferentes níveis de revocação.

Tabela 3.2: Valores de precisão obtidos após a execução do processo utilizando as evidências isoladamente.

Revocação	**Precisão Dataset 1**				**Precisão Dataset 2**			
	nomes	**co-autores**	**títulos**	**veículos**	**nomes**	**co-autores**	**títulos**	**veículos**
0	1,0000	0,9500	0,9500	0,8278	0,8666	0,6359	0,7335	0,2172
0,1	1,0000	0,9500	0,8647	0,7628	0,8597	0,6192	0,7189	0,2172
0,2	0,9333	0,8484	0,7631	0,6661	0,8377	0,6015	0,6873	0,2172
0,3	0,9333	0,6562	0,5961	0,6414	0,7819	0,5885	0,6545	0,2089
0,4	0,8833	0,6562	0,5461	0,5914	0,7802	0,5882	0,6239	0,2089
0,5	0,8474	0,5905	0,5461	0,5914	0,7627	0,5835	0,6172	0,2082
0,6	0,8052	0,5812	0,5046	0,5914	0,5141	0,4188	0,5203	0,2011
0,7	0,7635	0,5803	0,4929	0,5899	0,4806	0,4049	0,4755	0,2011
0,8	0,7635	0,4954	0,4049	0,5762	0,3978	0,3789	0,4390	0,1947
0,9	0,7196	0,4114	0,3561	0,4103	0,3708	0,3501	0,4182	0,1869
1	0,7196	0,4114	0,3552	0,4058	0,3406	0,3480	0,4027	0,1822

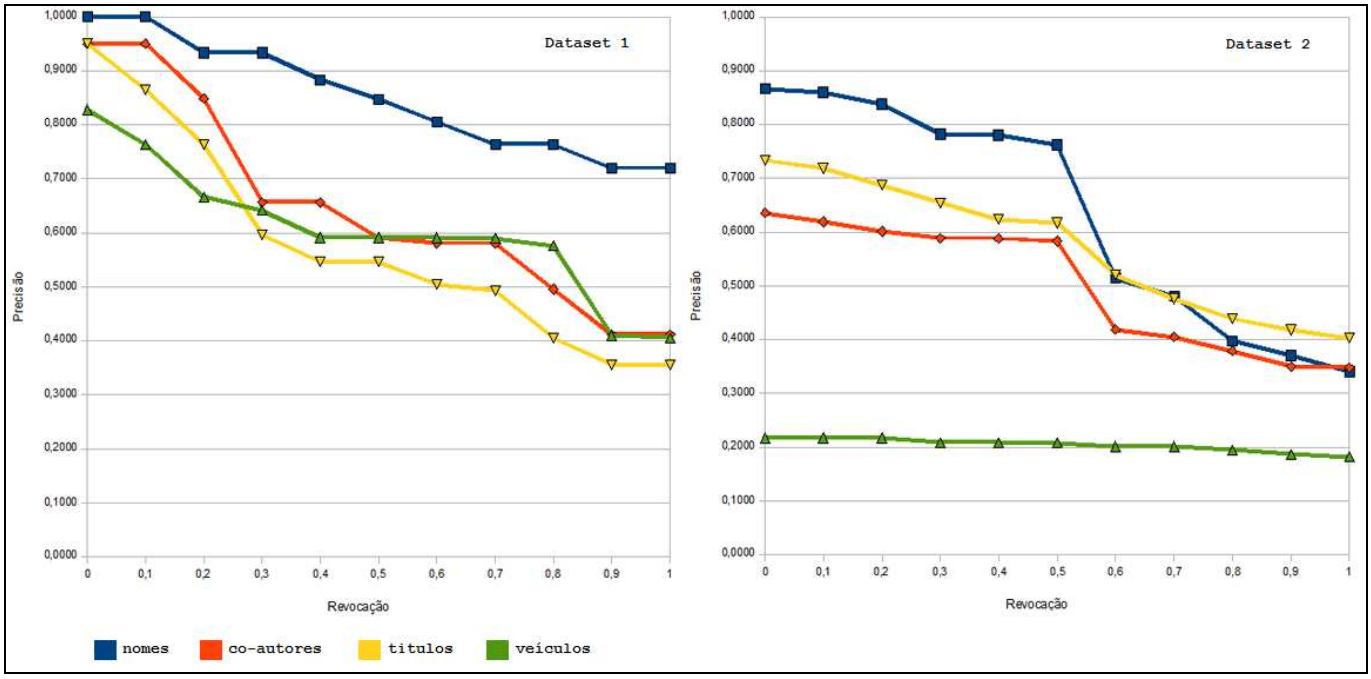

Figura 3.9: Gráficos de precisão versus revocação obtidos com a execução do processo de desambiguação utilizando as medidas de similaridade isoladamente.

3.2.3 Medidas de similaridade combinadas

Para uma melhor análise das medidas de similaridade, foram realizados testes com cada uma das medidas combinadas com a similaridade de nomes (*baseline*). Dessa forma foi possível verificar qual das evidencias realmente contribui mais para a melhoria dos resultados da desambiguação em relação ao *baseline*. Para fazer a ordenação da lista ao final do procedimento, foi calculada uma média simples entre o resultado de similaridade dos nomes e cada uma dos outros componentes (co-autores, títulos e veículos de publicação) extraídos das bases de dados. Os resultados obtidos com esse experimento podem ser vistos na Tabela 3.3 e na Figura 3.10.

Tabela 3.3: Valores de precisão calculados para a combinação da similaridade de nomes com cada um dos componentes utilizados como evidências pelo método.

Revocação	**Precisão Dataset 1**				**Precisão Dataset 2**			
	nomes	**nomes + co-autores**	**nomes + títulos**	**nomes + veículos**	**nomes**	**nomes + co-autores**	**nomes + títulos**	**nomes + veículos**
0	1,0000	1,0000	1,0000	1,0000	0,8666	0,8654	0,8660	0,8644
0,1	1,0000	1,0000	1,0000	1,0000	0,8597	0,8512	0,8511	0,8644
0,2	0,9333	0,9333	0,9333	0,9292	0,8377	0,8488	0,8376	0,8609
0,3	0,9333	0,8455	0,8905	0,9292	0,7819	0,8301	0,8137	0,8074
0,4	0,8833	0,8455	0,8905	0,9292	0,7802	0,8270	0,8082	0,8050
0,5	0,8474	0,8262	0,8905	0,9292	0,7627	0,7982	0,7935	0,7893
0,6	0,8052	0,8040	0,8683	0,9292	0,5141	0,6349	0,6890	0,6161
0,7	0,7635	0,7873	0,8040	0,9292	0,4806	0,6114	0,6566	0,6083
0,8	0,7635	0,7873	0,7740	0,9292	0,3978	0,5472	0,6159	0,5546
0,9	0,7196	0,7433	0,7222	0,8277	0,3708	0,5005	0,5883	0,4879
1	0,7196	0,7433	0,7222	0,8117	0,3406	0,4863	0,5761	0,4606

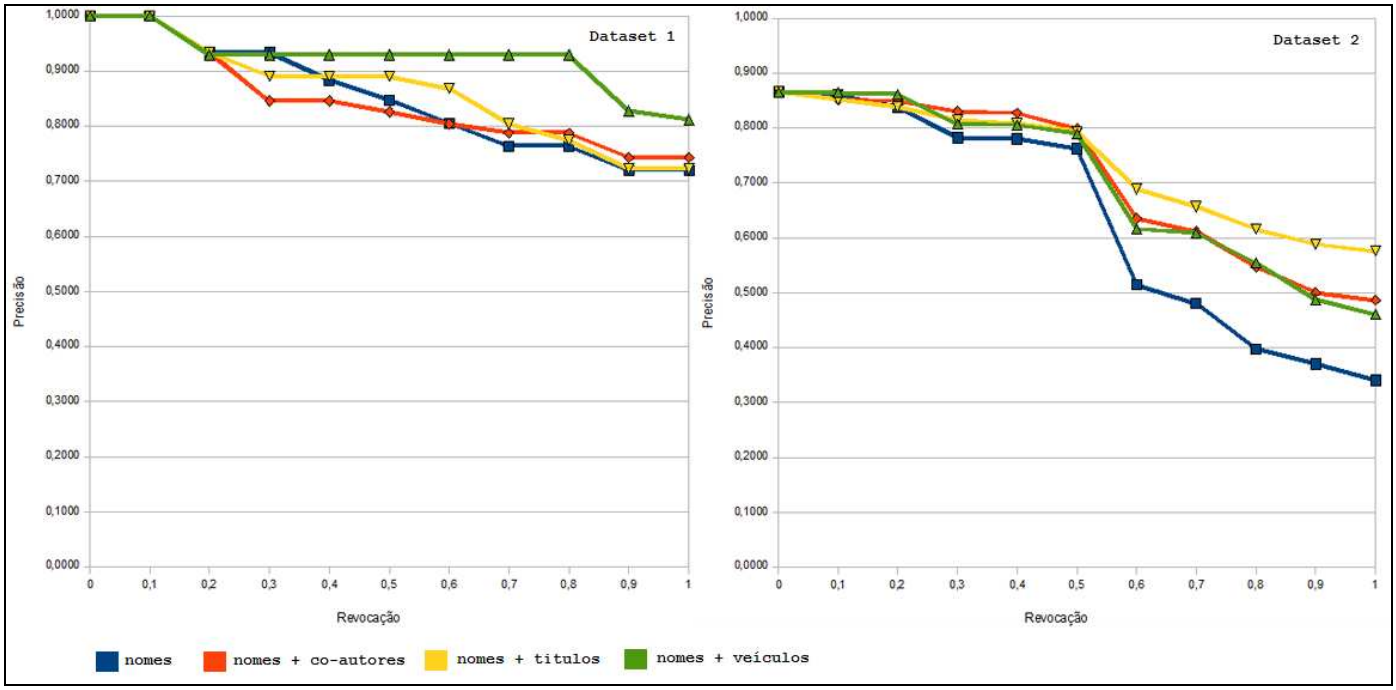

Figura 3.10: Gráficos mostrando as curvas de precisão versus revocação para as combinações da medida de similaridade dos nomes com cada uma das outras medidas.

3.2.4 Medida geral

A medida geral de similaridade geral é a média entre todas as medidas de similaridade. Inicialmente, foram realizados experimentos onde o cálculo era feito de maneira que todas as medidas tinham o mesmo peso, ou seja, uma média simples das medidas de similaridade. Com a análise feita em cima das curvas de precisão versus revocação geradas anteriormente e com base em observações, foi possível constatar que algumas das medidas tinham mais condições de influenciar positivamente no resultado da ordenação da lista. Por isso, foi elaborada uma média com pesos diferenciados para cada uma das medidas de similaridade. Os pesos utilizados forma os seguintes: peso 5 para a similaridade dos nomes, peso 1 para a similaridade dos co-autores, peso 2 para a similaridade dos títulos e por fim, peso 2 para a similaridade dos veículos de publicação. Dessa maneira, foi possível melhorar ainda mais os resultados do método, que já estava gerando resultados melhores do que o *baseline*.

3.2.4.1Média simples

Na Tabela 3.4 e na figura Figura 3.11 estão os resultados da heurística completa em relação ao método *baseline*. Aqui, após todas as medidas de similaridade, é feita a média simples dos valores resultantes para gerar a similaridade geral de um artigo em relação a um autor. Assim, definimos a ordenação das publicações para cada autor. Com o cálculo sendo efetuado dessa maneira, algumas distorções podem aparecer, pois algumas das evidências tem menos poder para a desambiguação como foi possível verificar nos gráficos que mostram as curvas de similaridades isoladas.

Tabela 3.4: Valores de precisão do método utilizando média simples em comparação com o baseline.

Revocação	Precisão Dataset 1		Precisão Dataset 2	
	baseline	média simples	baseline	média simples
0	1,0000	1,0000	0,8666	0,8669
0,1	1,0000	1,0000	0,8597	0,8457
0,2	0,9333	0,9292	0,8377	0,8319
0,3	0,9333	0,8417	0,7819	0,8002
0,4	0,8833	0,8417	0,7802	0,7956
0,5	0,8474	0,8417	0,7627	0,7851
0,6	0,8052	0,8417	0,5141	0,6750
0,7	0,7635	0,7833	0,4806	0,6485
0,8	0,7635	0,7611	0,3978	0,6282
0,9	0,7196	0,6983	0,3708	0,5828
1	0,7196	0,6638	0,3406	0,5765

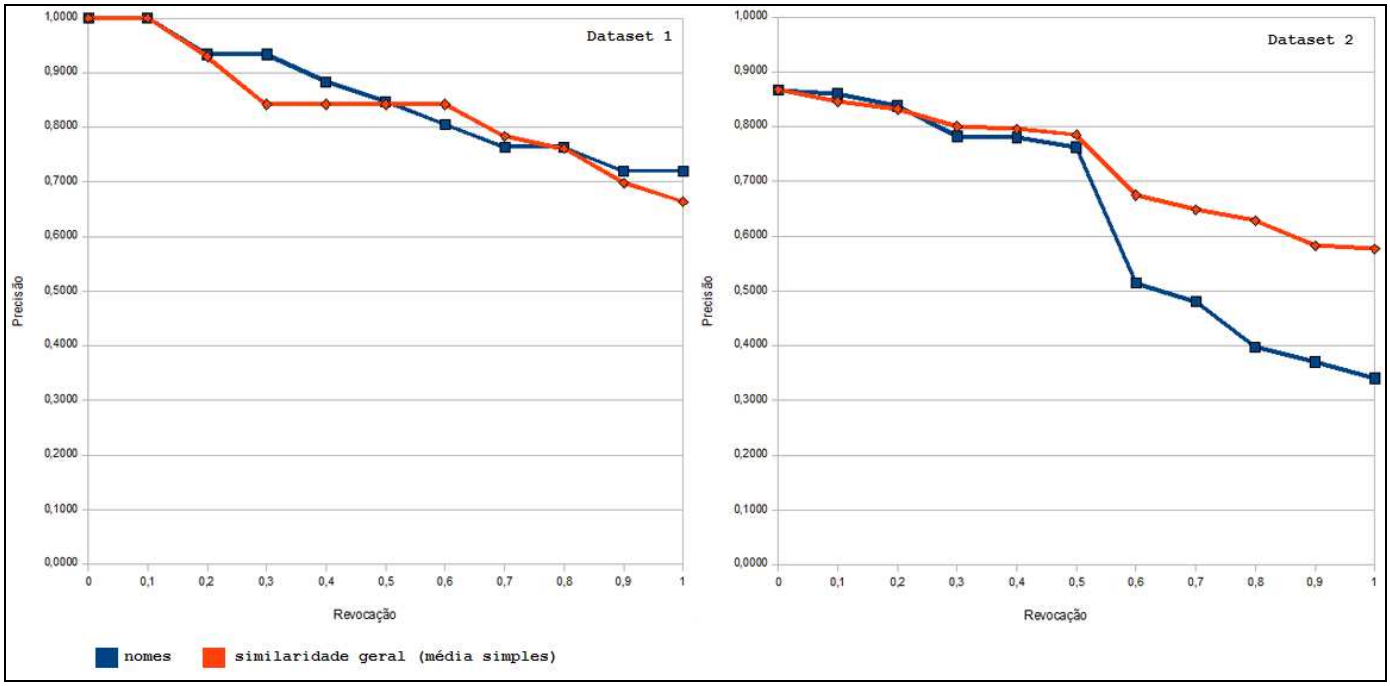

Figura 3.11: Comparação entre a precisão do baseline e o resultado gerado pelo método utilizando média simples.

3.2.4.2 Média ponderada

A fim de melhorar o resultado do método, foram examinados os resultados isolados e se chegou aos seguintes pesos para as medidas de similaridade: peso 5 para similaridade de nomes, peso 1 para a similaridade de co-autores, peso 2 para a similaridade de títulos e peso 2 para a similaridade de veículos de publicação. Os pesos devem ser maiores para aquelas medidas que de fato mais auxiliam na desambiguação dos autores. Como é possível verificar na Tabela 3.5 e na Figura 3.12, os resultados melhoraram sensivelmente apenas definindo pesos para as medidas. Esses pesos foram definidos de uma maneira arbitrária, somente com base em observação dos resultados e testes de execução do método.

Tabela 3.5: Comparação dos resultados de precisão das variantes (média simples e média ponderada) do método.

Revocação	**Precisão Dataset 1**			**Precisão Dataset 2**		
	baseline	**média simples**	**média ponderada**	**baseline**	**média simples**	**média ponderada**
0	1,0000	1,0000	1,0000	0,8666	0,8669	0,8661
0,1	1,0000	1,0000	1,0000	0,8597	0,8457	0,8587
0,2	0,9333	0,9292	0,9292	0,8377	0,8319	0,8474
0,3	0,9333	0,8417	0,8958	0,7819	0,8002	0,8224
0,4	0,8833	0,8417	0,8958	0,7802	0,7956	0,8224
0,5	0,8474	0,8417	0,8958	0,7627	0,7851	0,8093
0,6	0,8052	0,8417	0,8958	0,5141	0,6750	0,7028
0,7	0,7635	0,7833	0,8958	0,4806	0,6485	0,6814
0,8	0,7635	0,7611	0,8958	0,3978	0,6282	0,6614
0,9	0,7196	0,6983	0,8414	0,3708	0,5828	0,6285
1	0,7196	0,6638	0,8414	0,3406	0,5765	0,6185

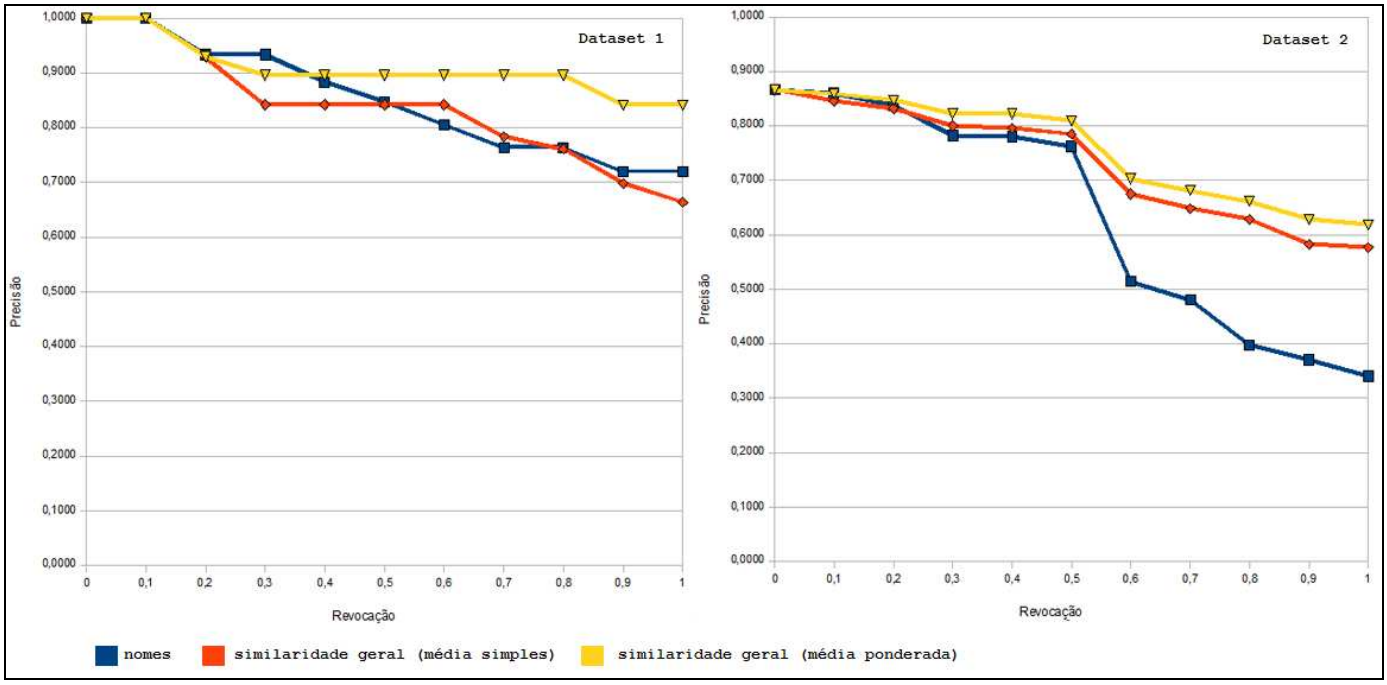

Figura 3.12: Comparação das curvas de precisão e revocação das duas variantes (média simples e média ponderada) do método.

4 CONCLUSÃO

Serviços comuns disponibilizados por DLs geralmente não tratam problemas como a ambigüidade de nomes de autores ou a redundância de dados em suas bases. Tipicamente, esses serviços são bem simples, o que pode gerar resultados distorcidos e até mesmo errados caso a base possua de fato esses problemas. Esses resultados apresentam distúrbios principalmente no que diz respeito à autoria de publicações devido a grande variedade de fontes que as DLs utilizam e também devido ao grande número de padrões e convenções utilizadas para representar uma publicação.

Nesse trabalho foi apresentado um método para a desambiguação de nomes de autores em citações de DLs. O método trabalha com as evidências encontradas nos próprios registros presentes nas bases de dados dessas bibliotecas. Para cada evidência extraída das citações, existe um algoritmo que calcula uma medida de similaridade. Cada uma dessas medidas foi analisada para se chegar a um melhor método de desambiguação. Esse método é o resultado de uma média ponderada das diferentes medidas de similaridade calculadas. Com o resultado das medidas de similaridade para todos os componentes presentes nos registros, é feita essa média ponderada gerando um valor de similaridade que ao final servirá para ordenar a lista de publicações para cada autor. Essa lista fica ordenada de modo que as primeiras posições representam as publicações relevantes ao autor em questão.

Os experimentos foram realizados em cima de duas bases de testes. A base Dataset 1 foi criada durante a implementação do método e é uma base que foi utilizada para os testes iniciais, portanto os resultados baseados nela não devem ser considerados seriamente. Devido a seu tamanho reduzido, ela foi utilizada para que se pudesse observar o comportamento do algoritmo para garantir a correção do processo de desambiguação. Já a base Dataset 2 é bem mais elaborada e complexa, imitando uma base real. Os resultados obtidos com os experimentos realizados em cima dessa base mostram que de fato o processo auxiliou na desambiguação dos nomes, amenizando o problema *split citation*. O método proposto nesse trabalho obteve ganhos bastante significativos de precisão nos resultados em relação ao *baseline* definido.

Para trabalhos futuros, seria possível explorar a área de *machine learning* para a definição dos pesos para a média geral ponderada, por exemplo. Dessa maneira, poderíamos melhorar ainda mais os resultados, pois teríamos pesos muito melhores ajustados do que os pesos definidos atualmente por meio de tentativa e observação. Outra possibilidade de melhoria seria extender o método para que ele trabalhasse não só com o problema *split citation*, mas também com o problema conhecido como *mixed citation*. Para isso, poderíamos utilizar a abordagem de agrupamento em *clusters* por exemplo, que é uma técnica que lida bem com esse problema e já foi amplamente utilizada e estudada. Ainda, outra extensão possível seria buscar informações sobre os

autores fora da base de dados. Dessa maneira, seria possível melhorar muito o processo de desambiguação, pois um dos grandes problemas nesse tipo de trabalho é a falta de informações suficientes para a comparação e cálculos de similaridade. Essa extensão traria grandes benefícios, porém ela requer um grande aprofundamento em outras áreas como a elaboração de *querys* para buscar essas informações, o tratamento dos documentos encontrados para extrair somente conteúdo relevante e benéfico para o processo, etc.

REFERÊNCIAS

Gonçalves, M.A., Fox, E.A., Watson, L.T. and Kipp, N.A. **Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries.** ACM Trans. Inf. Syst. 22, 2 (2004): 270-312.

Silva, L.V., Gonçalves, M.A. and Laender, A.H.F. **Evaluating a digital library selfarchiving service: The BDBComp user case study.** Information Processing and Management 43, 5 (2007): 1103-1120.

Lee, D., On, B.-W., Kang, J. and Park, S. **Effective and scalable solutions for mixed and split citation problems in digital libraries.** In Proceedings of the 2nd International Workshop on Information Quality in Information Systems, Baltimore, Maryland, 2005, pp. 69-76.

Oliveira, J.W.A, Laender, A.H.F and Gonçalves, M.A. **Remoção de ambigüidades na identificação de autoria de objetos bibliográficos.** In Proceedings of the 20th Brazilian Symposium on Databases, Uberlândia, MG, 2005, pp. 205-219.

Pereira, D. A., Ribeiro-Neto, B., Ziviani N., Laender, A. H. F., Gonçalves, M. A., Ferreira, A. A. **Using Web Information for Author Name Disambiguation.** 2009.

Malin, B. **Unsupervised name disambiguation via social network similarity.** In Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining. Newport Beach, CA. 2005, pp. 93-102.

Han, H., Giles, C.L., Zha, H., Li, C. and Tsioutsiouliklis, K. **Two supervised learning approaches for name disambiguation in author citations.** In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, Arizona, USA, June 2004, pp. 296-305.

Fonseca, B. M., Reis, D. C. **O fantástico mundo da distância de edição.** 2002.

A. Kent, M. M. Berry, L. V. Luehrs Jr, and J. W. Perry. **Machine literature searching III: Operational criteria for designing information retrieval systems.** American Documentation, 6(2):93-101, 1955.

C. J. van Rijsbergen. **Information Retrieval.** Butterworths, Lodon, 1979.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. **Modern Information Retrieval.** Addison Wesley, Harlow, 1999.

Cota, R. G., Gonçalves, M. A., Laender, A. H. F. **A Heuristic-based Hierarchical Clustering Method for Author Name Disambiguation in Digital Libraries.** SBBD, 2007.

Web site **BDBComp: Biblioteca Digital Brasileira de Computação**. Disponível em: < http://www.lbd.dcc.ufmg.br/bdbcomp/ >. Acesso em: set. 2009.

Web site **DBLP: Digital Bibliography & Library Project.** Disponível em: < http://www.informatik.uni-trier.de/~ley/db/ >. Acesso em: set. 2009.

Web site **Citeseer: Scientific Literature Digital Library**. Disponível em: < http://citeseer.ist.psu.edu/ >. Acesso em: set. 2009.

Web site **SimMetrics: Open Source Extensible Library of Similarity or Distance Metrics.** Disponível em: < http://www.dcs.shef.ac.uk/~sam/simmetrics.html >. Acesso em: set. 2009.