

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

ANÁLISE E DESENVOLVIMENTO DE METÓDOS DE
REMOÇÃO DE AMBIGUIDADE DE NOMES EM
CITAÇÕES BIBLIOGRÁFICAS DE BIBLIOTECAS
DIGITAIS.

Aluno: Herculano Gripp Neto
Matricula: 07.1.4192

Orientador: Anderson Almeida Ferreira

Ouro Preto
16 de abril de 2011

Universidade Federal de Ouro Preto - UFOP
Instituto de Ciências Exatas e Biológicas - ICEB
Departamento de Computação - DECOM

ANÁLISE E DESENVOLVIMENTO DE METÓDOS DE
REMOÇÃO DE AMBIGUIDADE DE NOMES EM
CITAÇÕES BIBLIOGRÁFICAS DE BIBLIOTECAS
DIGITAIS.

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia I (BCC390).

Aluno: Herculano Gripp Neto
Matricula: 07.1.4192

Orientador: Anderson Almeida Ferreira

Ouro Preto
16 de abril de 2011

Resumo

Devido a grande diversidade de fontes de dados utilizados pelas bibliotecas digitais, problemas de ambiguidade em citações bibliográficas são comumente encontrados. Em bibliotecas digitais, é frequente a ocorrência de casos em que o nome de um autor é representado de formas diferentes. Dessa forma, cada uma das diferentes formas de representação é tratada como publicações de pessoas diferentes. Deste modo é criado uma “divisão de citações”, ou *split citation*, onde publicações de uma mesma pessoa se encontram divididas de acordo com as variações do nome do autor.

O desafio é identificar quando duas referências pertencem ao mesmo autor. O objetivo deste trabalho é apresentar, analisar as técnicas presentes na literatura e ainda propor um método que suavize a ambiguidade de nomes em citações bibliográficas, para isso utilizando em conjunto técnicas supervisionadas e não-supervisionadas.

Palavras-chave: Ambiguidade de Nome, Bibliotecas Digitais, Citações Bibliográficas, Split Citation, Mixed Citation

Sumário

1	Introdução	1
2	Justificativa	2
3	Objetivos	3
3.1	Objetivo geral	3
3.2	Objetivos específicos	3
4	Metodologia	4
5	Cronograma de atividades	5

Lista de Figuras

1 Exemplo de *split citation* retirado da BDBComp 1

Lista de Tabelas

1 Cronograma de Atividades. 5

1 Introdução

As bibliotecas digitais(DLs) são importantes sistemas de gestão de informações na internet. As DLs são fontes massivas de informação para diversos segmentos. Atualmente, estas são de grande relevância em diversas áreas, principalmente a acadêmica e com sua ajuda vem atigindo um alto crescimento. Com sua ampliação estas passam a se tornar estruturas complexas. Com isso provocando uma serie de problemas, dentre eles o problema de ambiguidade em suas informações.

As DLs são alimentadas por diversas fontes de dados e cada uma destas fontes podem adotar padrões diferentes de representação para a gestão das informações. Um dos pontos mais críticos no contexto de padrões de representação é a ambiguidade de nomes de autores em citações bibliográficas.

Dos diferentes tipos de ambiguidades em uma biblioteca digital, a ambiguidade de nomes de autores é uma das mais comuns e possivelmente uma das mais difíceis de se resolver. Esse problema pode ser dividido em dois subproblemas o *split citation* (um autor possui variações do seu nome) ou *mixed citation* (dois autores diferentes com o mesmo nome). O primeiro ocorre quando o nome de um autor é representado de diferentes formas em suas publicações, assim suas publicações serão divididas. Isso pode ser observado na figura 1.

BDBComp
Biblioteca Digital Brasileira de Computação

Home Pesquisar: Autor Título Ano Evento Periódico L

Aleksandra do Socorro da Silva - Trabalhos Publicados

Veja também em: [ACM DL](#) - [CiteSeer](#) - [DBLP*](#) - [Google Scholar](#)

*Somente retorna uma resposta se o nome do autor na BDBComp e n

1 registro retornado

2001

• [Aleksandra do Socorro da Silva](#) [Arturo Hernández-Domínguez](#)

2004

• [Aleksandra Silva](#) [Arturo Hernandez-Dominguez](#) [Bruno Silva](#)
[Informática na Educação](#) [Buscar na Web](#)

2 registros retornados

2003

- [Maria da Penha de Andrade Abi Harb](#) [Silvana Rossy de Brito](#) [Aleksandra do Socorro Silva](#)
[aprendizagem multiparadigmático](#) [XIV Simpósio Brasileiro de Informática na Educação](#)
- [Aleksandra do Socorro Silva](#) [Silvana Rossy de Brito](#) [Eloi Luiz Favero](#) [Arturo Hernández-Domínguez](#)
[ambientes interativos de aprendizagem baseado em agentes, componentes e framework](#) [XIV Simpósio Brasileiro de Informática na Educação](#)

Figura 1: Exemplo de *split citation* retirado da BDBComp

O *mixed citation* acontece quando dois autores distintos possuem o mesmo nome em suas citações, o que pode provocar erro de autoria das publicações. O problema da ambiguidade de nomes surge quando o nome de um autor é representado de diversas formas, por exemplo abreviando um dos sobrenomes ou omitindo um ou vários deles. Neste trabalho iremos citar alguns métodos propostos para a remoção da ambiguidade de nomes. A partir dos métodos citados será elaborado um esboço para a solução do problema em questão, podendo ou não agregar heurísticas abordadas nestes.

2 Justificativa

O problema da ambiguidade de nomes juntamente com o problema de duplicidade de citações são temas intensivamente estudados. No entanto, ambos temas ainda não possuem uma solução suficientemente eficaz.

Com o crescimento das bibliotecas digitais e sua popularização esses problemas foram somente agravados. Como sistemas deste tipo utilizam o nome do autor como indexador para suas pesquisas, ambiguidade de nomes podem causar mal-funcionamento do mesmo. Logo, a proposta de um método que seja suficientemente eficaz se faz necessário.

3 Objetivos

3.1 Objetivo geral

- Desenvolvimento de um método eficiente de remoção de ambiguidade de nomes em citações bibliográficas.

O objetivo geral é propor um método de remoção de ambiguidade de nomes mais eficiente que os já encontrados na literatura. Com base nas análises feitas sobre os métodos já disponíveis e pelo estudo do problema.

3.2 Objetivos específicos

- 1º Fase - Exploração do tema em busca de métodos para resolução do problema.
- 2º Fase - Análise dos métodos existentes.
- 3º Fase - Propor um método alternativo para a resolução do problema

Os objetivos específicos serão a base para a construção de um novo método, a partir da exploração do problema. Nesta fase é necessário identificar como, quando e porque o problema ocorre.

Na segunda fase, serão analisados os métodos de resolução já existentes na intenção de identificar suas possíveis deficiências e comparar a eficiência de cada um motivo de um ser mais eficaz que os outros.

Na terceira fase, será proposto um algoritmo para resolução do problema com base no conhecimentos adquirido na primeira e agregando heurísticas dos métodos citados na segunda fase caso necessário.

4 Metodologia

No processo de desenvolvimento do novo método o ponto principal é identificar as limitações do problema e analisar a forma como os métodos existentes lidam com ela. Na fase inicial será feito um estudo detalhado do problema e delimitado o escopo a ser trabalhado. Ainda nesta fase serão determinadas informações relevantes para a resolução do problema, como por exemplo o nomes dos co-autores ou o veículo de publicação.

Na segunda fase iremos analisar o métodos já existentes e extrair metodologias que vem a ser úteis para elaboração do novo método. Na terceira etapa serão usadas todas as informações coletadas para a elaboração do novo método, ressaltando que nesta fase não haverá implementação do método. A implementação e testes do novo método serão realizadas na disciplina Monografia II.

A última fase consiste em redigir a monografia e preparar o material para a apresentação do trabalho fim do semestre letivo.

5 Cronograma de atividades

Na Tabela 1, deixa claro como será feita a divisão do tempo para a realização das fases para a conclusão da monografia.

Atividades	Mar	Abr	Mai	Jun	Jul
Exploração do tema	X				
Análise e comparação dos métodos	X	X			
Elaboração do novo método			X	X	
Redigir a Monografia				X	X
Apresentação do Trabalho					X

Tabela 1: Cronograma de Atividades.

Referências

- Anderson A. Ferreira, Adriano Veloso, Marcos André Gonçalves, Alberto H. F. Laender: **Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries**. JCDL 2010: 39-48, 2010
- Ricardo G. Cota, Anderson A. Ferreira, Cristiano Nascimento, Marcos André Gonçalves, Alberto H. F. Laender: **An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations**. JASIST 61(9): 1853-1870 2010
- N. R. Smalheiser and V. I. Torvik: **Author Name Disambiguation**, volume 43, pages 287-313. 2009.
- Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles: **Efficient topic-based unsupervised name disambiguation**. In Proc. of JCDL, pages 342-351, 2007.
- Web site BDBComp: Biblioteca Digital Brasileira Computação. Disponível em: <<http://www.lbd.dcc.ufmg.br/bdbcomp/>>. Acesso em: Abril 2011.