

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

UMA ABORADAGEM INCREMENTAL PARA REMOÇÃO
DE AMBIGUIDADE DE NOMES EM CITAÇÕES
BIBLIOGRÁFICAS

Aluno: Herculano Gripp Neto
Matricula: 07.1.4192

Orientador: Anderson Almeida Ferreira

Ouro Preto
29 de março de 2012

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

UMA ABORADAGEM INCREMENTAL PARA REMOÇÃO
DE AMBIGUIDADE DE NOMES EM CITAÇÕES
BIBLIOGRÁFICAS

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 II (BCC391).

Aluno: Herculano Gripp Neto
Matricula: 07.1.4192

Orientador: Anderson Almeida Ferreira

Ouro Preto
29 de março de 2012

Resumo

Problemas de ambiguidade de nomes de autores de artigos científicos são comumente encontrados, devido a grande diversidade de fontes de dados utilizadas pelas bibliotecas digitais para a coleta dos dados, a falta de padronização na forma de escrever os nomes de autores nos artigos e a ocorrência de vários autores com o mesmo nome. Assim, é comum encontrar, em bibliotecas digitais, artigos de um mesmo autor com diferentes nomes deste autor, o que pode levar a tratar cada nome como se fosse de um autor diferente. Outra situação, que também ocorre, é quando autores distintos possuem o mesmo nome em seus artigos. Esta situação pode levar a tratar todos os artigos como se fossem do mesmo autor. Vários métodos já foram propostos para tentar resolver este problema. No entanto, nenhum deles resolve completamente este problema e normalmente eles são aplicados sobre todo o conjunto de artigos representados pelos seus meta-dados. O objetivo deste trabalho é estudar técnicas de remoção de ambiguidade de nomes de autores de artigos científicos e propor um método capaz de remover a ambiguidade de nomes de autores de modo incremental, ou seja, remover a ambiguidade apenas dos novos artigos, cujos meta-dados estão sendo inseridos em uma biblioteca digital, bem como a sua avaliação por meio de análises comparativas com outros métodos.

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

Sumário

1	Introdução	1
2	Justificativa	4
3	Objetivos	5
3.1	Objetivo geral	5
3.2	Objetivos específicos	5
4	Método Proposto	6
4.1	Primeira Etapa	7
4.2	Segunda Etapa	7
5	Métricas de Similaridade de Strings	9
5.1	Similaridade de Distância de Levenshtein	9
5.2	Similaridade por Comparação de Fragmentos	9
5.3	Coeficiente de Jaccard	9
5.4	Medida do Cosseno	10
6	Cronograma de atividades	11

Lista de Figuras

1	Exemplo de <i>split citation</i> retirado da BDBComp.	2
2	Exemplo de <i>mixed citation</i> retirado do DBLP	3
3	Fluxograma Funcionamento do método	6

Lista de Tabelas

1	Cronograma de Atividades.	11
---	-----------------------------------	----

1 Introdução

O problema de ambiguidade de nomes pode ser observado em diversos contextos. Esse problema afeta principalmente os sistemas computacionais que na maioria das vezes não conseguem identificá-los e corrigi-los. Alguns exemplos de ambiguidade de nomes são encontrados em nomes de lugares como a cidade Ouro Preto e o bairro Ouro Preto localizado em Belo Horizonte, em nomes de pessoas como os ex-presidentes dos EUA George W. Bush pai e filho ou ainda em veículos de publicação onde, por exemplo, SBBD e Simpósio Brasileiro de Banco de Dados se remetem ao mesmo simpósio.

Este trabalho, trata do problema de desambiguação de nomes de pessoas, mais especificamente, do nome de pessoas em citações bibliográficas de artigos científicos, que comumente são armazenados em bibliotecas digitais.

Uma citação bibliográfica em uma biblioteca digital é representada pelos seus metadados que são constituídos por nomes dos autores, título do trabalho, título do veículo de publicação e o ano de publicação. Neste contexto, este problema pode afetar o desempenho de recuperação de documentos, de máquinas de pesquisa, da integração de banco de dados e pode causar a atribuição indevida de crédito a um autor.

Nas subseções a seguir serão introduzidos o conceito e a importância das bibliotecas digitais, os problemas relativos a ambiguidade de nomes neste contexto, bem como o escopo a ser abordado neste trabalho.

Bibliotecas Digitais

Bibliotecas digitais (DLs) são sistemas de informação complexos, que são projetados para um público específico, possuem um conjunto grande de objetos digitais e seus meta-dados, várias estruturas organizacionais e fornecem diversos serviços para manter e acessar esses objetos digitais [Gonçalves et al., 2004]. As bibliotecas digitais são importantes sistemas de gestão de informações na Internet. As DLs, ou *Digital Libraries*, são fontes massivas de informação para diversos segmentos. Atualmente, são de grande relevância em diversas áreas, principalmente a acadêmica e com a ajuda desta vem atingindo um alto crescimento.

As DLs além de um importante veículo de referência é uma grande rede de colaboração (co-autoria), e com sua ampliação está se tornando cada vez mais complexa. Entretanto, para qualidade destas características é necessário que as bases de dados possuam informações que condizem com a realidade, ou seja, que não haja ambiguidade nas informações da DL.

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

Ambiguidade de Nomes de Autores em Citações Bibliográficas

A ambiguidade de nomes de autores em citações bibliográficas é um problema que está presente nas mais diversas bibliotecas digitais. Segundo Lee et al. [2005] podemos dividi-lo em dois sub-problemas: *split citation*(SC) e *mixed citation*(MC).

O primeiro ocorre quando há uma variação no modo como o nome de um autor é representado. Nesse caso, publicações de um mesmo autor podem estar divididas como

se pertencessem a pessoas distintas. Isso acontece principalmente devido a um erro ortográfico, abreviatura de um nome ou sobrenome ou ainda a mudança na ordem de nome e sobrenomes.

Segue um exemplo na Figura 1 para ilustrar um caso de SC retirado de uma pesquisa na BDBComp¹ para a autora “Aleksandra Silva”. Neste exemplo, foram recuperadas três diferentes representações de uma mesma autora, sendo elas, “Aleksandra Silva”, “Aleksandra do Socorro Silva” e “Aleksandra do Socorro da Silva” em que todas as citações se referem a uma mesma autora.

The image shows a screenshot of the BDBComp website. The header includes the BDBComp logo and the text 'Biblioteca Digital Brasileira de Computação'. Below the header, there is a navigation menu with links: Home, Pesquisar, Autor, Título, Ano, Evento, Periódico, Li. The main content area shows search results for 'Aleksandra Silva'. There are two search results for the year 2004. The first result is for 'Aleksandra do Socorro da Silva' with a link to 'Trabalhos Publicados'. The second result is for 'Aleksandra Silva' with a link to 'Trat'. Both results are circled in red. The page also includes navigation links like 'Home', 'Pesquisar', 'Autor', 'Título', 'Ano', 'Evento', 'Periódico', and 'Li'.

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

O problema *mixed citation* ocorre quando diferentes autores compartilham o mesmo nome ou a mesma variação de nome e suas publicações aparecem como se pertencessem a um mesmo autor. Na Figura 2 é apresentado um exemplo de MC retirado de uma pesquisa na DBLP para o autor “Mohammed Zaki”. Analisando somente o nome exato utilizado para pesquisa “Mohammed Zaki”, notamos que essa classe referencia dois autores diferentes.

¹ <http://www.lbd.dcc.ufmg.br/bdbcomp/>

Um deles indica um “Mohammed Zaki” que pertence ao corpo docente da Universidade de Al-Zhar, na cidade de Nasr, Cairo, Egito. Enquanto o outro aponta para o “Mohammed Zaki” que pertence ao Departamento de Ciência da Computação do Instituto Politécnico Rensselaer dos Estados Unidos.

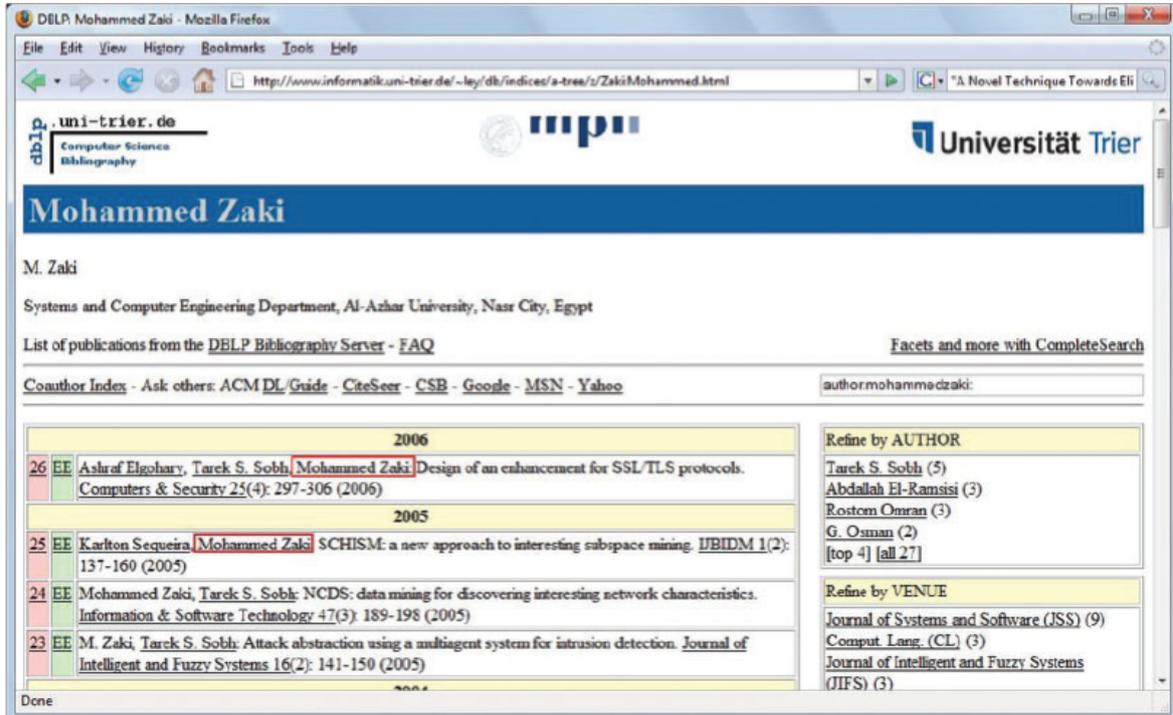


Figura 2: Exemplo de *mixed citation* retirado do DBLP

Escopo

Agora que já foi apresentado o conceito de bibliotecas digitais e os sub-problemas relacionados a ambiguidade de nomes em suas citações bibliográficas, será delimitado o escopo do trabalho. Na literatura existem diversos trabalhos que tratam do problema de remoção de ambiguidade de nomes em uma coleção de citações bibliográficas contidas em uma DL, alguns deles serão expostos e explicados na seção Trabalhos Relacionados. Entretanto, pouco se fala em uma abordagem incremental deste problema. Essa abordagem incremental consistiria em desambiguar apenas os registros das novas citações que estão sendo inseridas em uma biblioteca digital.

A proposta desta abordagem é simples de se compreender. Por exemplo, suponha que a partir de uma coleção desambiguada de registros de citações de uma DL seja necessário incluir uma publicação de um determinado autor. O método a ser proposto deverá ser capaz de identificar se uma publicação a ser inserida na DL já pertence a um autor que já possui artigos na biblioteca digital ou não. Caso pertença a biblioteca, a publicação a ser inserida deverá ser atribuída a esse autor. Caso contrário, deve-se identificar o trabalho como sendo de um novo autor.

2 Justificativa

O problema de ambiguidade de nomes afeta diretamente diversas áreas como os sistemas de recuperação de informação, o estabelecimento de rede de colaboração (co-autoria) e agências de fomento.

Em um sistema de recuperação de informação, se torna mais simples notar o problema de ambiguidade de nomes. Nesses sistemas, a importância de se manter os nomes desambiguados é imensa. Ao se fazer uma pesquisa sobre as publicações de um determinado autor $a1$, caso haja alguma ambiguidade neste nome os resultados da pesquisa poderiam conter publicações referentes a outros autores ou mesmo não incluir todos os trabalhos deste autor.

No caso dos outros dois exemplos a percepção deste problema não é tão simples. No estabelecimento de redes de colaboração havendo uma atribuição errada de um autor a uma publicação irá gerar uma ligação que não deveria existir na rede de co-autoria. Já com relação as agências de fomento é levado em consideração o mesmo problema citado para as redes de co-autoria. Se uma publicação for atribuída erroneamente a outro autor haverá uma redução no número de publicações do autor que realmente publicou. Como as agências de fomento são responsáveis pelo incentivo e patrocínio de pesquisas e publicações (exemplo CNPq) elas levam em consideração a quantidade de trabalhos relacionados aquele autor para aprovar ou não o investimento no projeto.

3 Objetivos

3.1 Objetivo geral

- Manter uma coleção de citações bibliográficas livre de ambiguidade.

3.2 Objetivos específicos

- Fazer uma revisão bibliográfica sobre métodos de remoção de ambiguidade.
- Análise de métodos existentes, visando descobrir seus pontos fracos e fortes.
- Propor um método incremental de remoção de ambiguidade.
- Avaliar o método proposto comparando-o a métodos representativos existentes na literatura.

4 Método Proposto

Na seção I, foi apresentado o problema e o escopo do trabalho. Nesta será apresentado o método proposto detalhadamente.

Como o objetivo é manter uma coleção de registros de citações bibliográficas livre de ambiguidade o desafio é , dado um conjunto de citações bibliográficas a serem inseridos nesta coleção, identificar de forma única o autor a que se refere cada nome de cada citação bibliográfica. Para isso, uma das grandes questões é identificar se os autores a que se referem os nomes nas citações bibliográficas já possuem trabalhos cadastrados na biblioteca digital. Pois, se já possuem, essas citações devem ser atribuídas a esses autores já existentes e se não possuírem essas citações devem ser atribuídas a um novo autor.

Indicar se uma citação a ser inserida em uma DL se refere a autores com trabalhos já cadastrados nessa DL, requer, inicialmente, uma comparação dos nomes desta citação com os autores que já possuem trabalhos cadastrados na DL e o modo como será feita essa comparação influencia na eficiência e na eficácia do método de desambiguação.

Imagine por exemplo, que para reconhecer um determinado autor *a1*, tivéssemos que comparar *a1* com cada autor presente na DL para garantir que é um novo autor. O custo computacional seria muito alto. Com o intuito de minimizar comparações excessivas iremos separar o método em duas etapas: a primeira será a seleção de citações da DL de possíveis autores e a segunda comparação entre os registros dessas citações com os das citações a serem inseridas. Abaixo a Figura 3 ilustra a idéia de funcionamento do método.

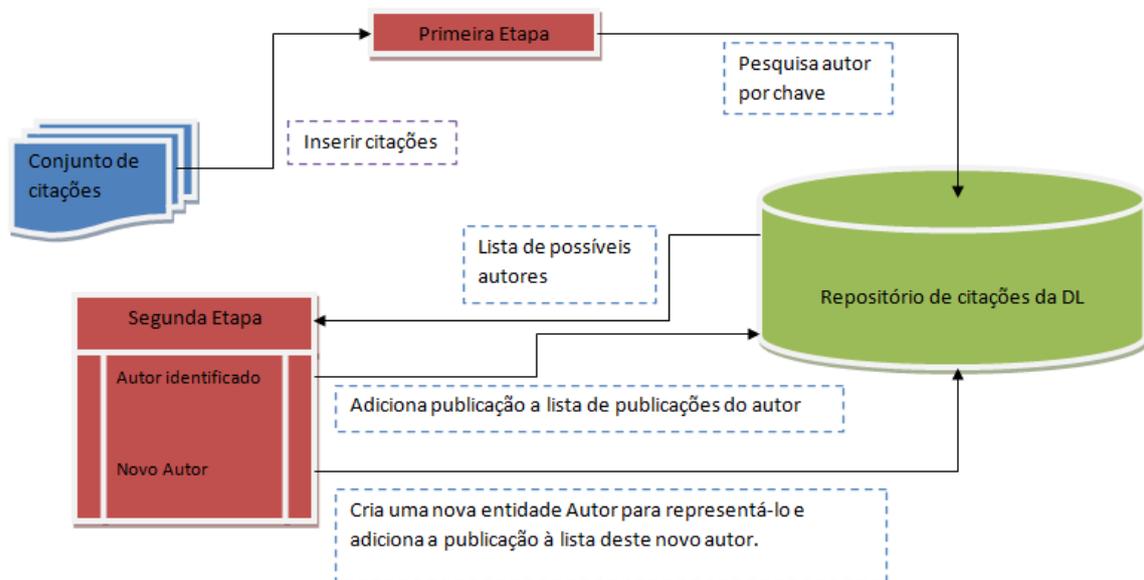


Figura 3: Fluxograma Funcionamento do método

Na etapa inicial serão selecionados registros de citações cujos nomes de autores são similares aos nomes de autores dos registros de citações a serem inseridos. Na próxima etapa serão avaliadas técnicas supervisionadas e não supervisionadas para identificar se os autores da citação a ser inserida já possuem publicações cadastradas na DL.

4.1 Primeira Etapa

Na etapa inicial dado uma citação c a ser inserida com n autores sendo eles a_1, a_2, \dots, a_n , para cada nome de c deve-se recuperar uma lista de possíveis autores contidos na DL. Como este primeiro refinamento deve ser computacionalmente barato será usado uma tabela hash para armazenar todos os autores contidos na DL que usará como chave a inicial do nome do autor e o último sobrenome. Cada registro nesta tabela será uma lista da entidade *Autor*. Essa entidade é uma tupla contendo o nome representativo do autor, uma lista de publicações deste autor e o valor da similaridade entre a_i e o nome representativo da entidade, esse valor será definido somente na segunda etapa. Um exemplo de registro da tabela é a chave *A. Ferreira* que irá armazenar uma lista de entidades *Autor* que possuem o nome representativo iniciado com a letra *A* e tenham *Ferreira* como seu último sobrenome, ou seja, a lista de autores por exemplo pode conter os seguintes nomes *Anderson Almeida Ferreira, André Luiz Ferreira ou Aida Silva Ferreira* desde que estes autores possuam alguma publicação na DL.

4.2 Segunda Etapa

Na segunda etapa cada nome (a_1, \dots, a_n) da citação c terá uma lista (l_1, \dots, l_n) de possíveis autores associados a ele. Como nesta lista podem haver nomes de pessoas totalmente diferentes ao nome do autor da citação c será necessário utilizar um método de similaridade entre strings para fazer um refinamento. Na seção 5 serão apresentados diferentes métricas de similaridade strings. As funções de similaridade de strings retornam um valor entre 0 e 1 de acordo com o grau de similaridade entre duas strings, neste caso o nome de dois autores. Baseado neste valor será realizada uma filtragem na lista de possíveis autores de cada nome pertencente a c , onde um valor α será tomado como base. Isso quer dizer que só irá permanecer na lista os nomes de autores o qual a similaridade entre strings retornar um valor maior ou igual ao valor base α , sendo ela aplicada entre o nome da lista e um autor de c . Este valor da similaridade será armazenado no atributo valor da similaridade (SM), como dito na primeira etapa.

Com a lista de possíveis autores refinada, serão utilizados outros atributos para ajudar a identificar se alguns dos autores pertencentes a essa lista é realmente um autor da citação c . Os atributos que serão utilizados são o nome do autor, nome dos co-autores, título das publicações e veículo de publicação. Antes de utilizar estes dados é necessário realizar um pré-processamento dos mesmos com o intuito de extrair as informações não relevantes a identificação dos autores. Este pré-processamento ocorrerá exclusivamente sobre o título da publicação. O mesmo consiste em retirar as palavras de parada, do inglês *stop-words*, e utilizar um método de radicalização de palavras, tornando assim o processo de similaridade mais eficiente.

Após a fase pré-processamento é necessário selecionar uma citação de cada possível autor em sua lista de publicações. Para isso será usado um critério simples, primeiro a posição do autor nesta publicação deve ser 1, ou seja, ele deve ser o autor da publicação e segundo a citação a ser escolhida deve o maior número de co-autores em comum com a citação c entre as citações da lista.

Como foi escolhida uma citação de cada possível autor, podemos utilizar os outros atributos para definir se algum destes autores é o autor da citação c . Agora serão usados os co-autores para definir uma similaridade entre os autores. Serão comparados a lista

de de co-autores de c com a lista de co-autores da citação escolhida. Essa similaridade de co-autores (SCA) será definida assim:

$$SCA = \frac{N_c}{N_t}$$

onde, N_c é número de co-autores de c que são iguais aos co-autores da citação escolhida e N_t é o número total de co-autores de c .

Para ajudar a identificar o autor serão utilizados os atributos título(ST) e veículo de publicação(SV). Quanto ao título será aplicada uma métrica de similaridade que irá comparar a lista de palavras das citações. Essa comparação poderá ser casamento exato ou inexato entre as palavras, lembrando que do título foram removidas as *stop-words* e as demais foram reduzidas ao seu radical. No veículo de publicação será utilizado o mesmo método, casamento de padrão na tentativa de identificar foram publicados no mesmo veículo ou em áreas a fins.

Para identificar se o autor da citação c é um autor presente na DL utilizaremos uma média aritmética ponderada utilizando a similaridade de todos os atributos, ou seja, a similaridade de nomes de autores (SM), co-autores (SCA), título (ST) e de veículo de publicação (SVP). Cada um destes atributos terão um peso P decrescente começando por P_{SM} , P_{SCA} , P_{ST} até P_{SVP} . Sendo assim os atributos mais relevantes terão um peso maior. A função de identificação de autores é mostrada abaixo :

$$F = \frac{SM * P_{SM} + SCA * P_{SCA} + ST * P_{ST} + SVP * P_{SVP}}{P_{SM} + P_{SCA} + P_{ST} + P_{SVP}}$$

Para que a identificação do autor seja confirmada será necessário que o valor de F seja maior que um valor δ pré-definido. O possível autor que tiver o maior resultado de F será identificado como um autor da citação c , desde que $F \geq \delta$. Uma vez que nome da citação c tenha sido identificado como um autor da DL, a publicação é inserida na lista deste publicações do autor.

Caso os possíveis autores da citação c obtenham um valor de $F \leq \delta$ será considerado como um novo autor. Considerado um novo autor será criada uma entidade *Autor* para representá-lo. O nome representativo da entidade será o mesmo nome contido em c e será criada uma lista de publicações contendo apenas uma publicação, ou seja, a publicação a qual a citação c se refere.

Os valores base citados acima α e δ serão definidos na implementação visto que eles influenciam diretamente no resultado do método, assim também como os pesos P_{SM} , P_{SCA} , P_{ST} e P_{SVP} . Testes serão executados afim de identificar valores aos quais o método produz um melhor resultado.

5 Métricas de Similaridade de Strings

Nesta seção serão apresentadas as principais métricas de similaridade entre strings utilizadas. No entanto, somente algumas serão utilizadas na construção do método de remoção ambiguidade.

5.1 Similaridade de Distância de Levenshtein

Distância de Levenshtein [1966] foi nomeada em homenagem ao cientista russo Vladimir Levenshtein, que desenvolveu o algoritmo em 1965. Também sendo conhecida pelo nome de *edit distance*. O cálculo de similaridade utilizando distância de edição baseia-se no número de mínimo de transformações (inserção, exclusão e substituição) necessário para transformar uma string S em uma string T . Quanto menor a distância de edição mais similares são as strings. Existem diversas variações deste método: alguns atribuem pesos diferentes para cada operação e outros utilizam métricas diferentes da distância de Levenshtein como a distância de Hamming.

5.2 Similaridade por Comparação de Fragmentos

Segundo French et al. [2000] a similaridade por comparação de fragmentos é uma função de casamento de padrão que, através do algoritmo de distância de edição, avalia um a um cada fragmento de duas cadeias de caracteres que representam nomes.

Os parâmetros necessários são duas cadeias de caracteres normalizadas, S e T , e o limite L utilizado para a distância de edição. O resultado retornado é verdadeiro se S e T são compatíveis e eventualmente podem representar a mesma entidade do mundo real, e falso caso contrário. O limite L pode ser um número inteiro e, neste caso é utilizado diretamente pelo algoritmo, ou um número entre 0 e 1, em cujo caso o número máximo de erros permitidos será igual ao tamanho do menor fragmento comparado multiplicado por L .

5.3 Coeficiente de Jaccard

Segundo Cohen et al. [2003] a métrica de similaridade de Jaccard é baseada em *tokens*. Jaccard considera as strings a serem comparadas como palavras separadas por espaços para definir a similaridade. Sua equação é dada a seguir:

$$jaccard = \frac{S \cap T}{S \cup T}$$

onde, S e T são conjuntos de *tokens* derivados da quebra em palavras das duas strings de entrada s e t . Esta métrica retorna o quociente do número de *tokens* que representam a intersecção dos conjuntos S e T pelo número de *tokens* que representam a união desses conjuntos. Observe o exemplo a seguir:

Jaccard ("Universidade Federal Ouro Preto", "Universidade Ouro Preto") = 0.750

Neste exemplo há três *tokens* em comum (Universidade, Ouro e Preto), e há 4 *tokens* no total desta forma basta realizar a divisão $\frac{3}{4}$ para obter o resultado final como no exemplo.

5.4 Medida do Cosseno

Segundo Salton et al. [1975] uma string é nada mais que um conjunto de palavras. Existe um conjunto finito de palavras que determinam o vocabulário. Logo em cada elemento da string podem aparecer quaisquer palavras do vocabulário. Sendo assim é possível criar uma representação vetorial em um espaço Euclidiano multidimensional para cada elemento do conjunto. Cada eixo deste espaço corresponde a uma palavra. A coordenada de um elemento e na direção correspondente a uma palavra p é determinada por duas medidas:

- TF (*term frequency*), que corresponde ao número de vezes que uma palavra p aparece em um elemento e .
- IDF (*inverse document frequency*), que corresponde ao peso da palavra de acordo com o inverso da frequência no elemento

A medida IDF deve ser utilizada, pois os eixos no espaço vetorial não são igualmente importantes. Assim, palavras de maior frequência no conjunto de elementos devem ser penalizadas. Sendo E o conjunto de todos os elementos de mesmo rótulo pertencentes a um repositório R e E_p o conjunto dos elementos de E que contêm determinada palavra p , uma forma comum para o cálculo do IDF de p é:

$$w_p = \log \left(1 + \frac{|E|}{|E_p|} \right)$$

Elementos longos tendem a ser favorecidos por conterem um maior número de palavras diferentes. Com isso, é necessário realizar uma normalização em função do tamanho do elemento, que é determinada pela fórmula:

$$w_e = \sqrt{\sum w_{e,p}^2}$$

onde $w_{e,p}$ corresponde ao peso das palavras em relação ao elemento e é calculado através da regra $TF * IDF$

$$w_{e,p} = (1 + \log(f_{e,p})) * w_p$$

sendo $f_{e,p}$ o número de ocorrências da palavra p no elemento e .

A similaridade entre dois elementos é calculada através da medida do cosseno entre suas representações vetoriais. Quanto maior o cosseno, maior a similaridade. Para isto, utilizamos a fórmula:

$$\cos(e_1, e_2) = \frac{1}{w_{e_1} * w_{e_2}} * \sum_{p \in (e_1 \cap e_2)} (w_{e_1,p} * w_{e_2,p})$$

6 Cronograma de atividades

Na Tabela 1, estão representadas todas as etapas de desenvolvimento da monografia contemplando as disciplinas monografia I e II .

Atividades	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Revisão Bibliográfica	X								
Estudo dos métodos	X	X							
Projetar um novo método		X	X						
Implementar o método				X	X	X	X		
Testar o método							X	X	
Análise Comparativa							X	X	
Redigir a Monografia							X	X	X
Apresentação do Trabalho									X

Tabela 1: Cronograma de Atividades.

Referências

- W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. pages 73–78, 2003.
- J. C. French, A. L. Powell, and E. Schulman. Using clustering strategies for creating authority files. volume 51, pages 774–786, 2000.
- M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22(2):270–312, 2004.
- D. Lee, B.-W. On, J. Kang, and S. Park. Effective and scalable solutions for mixed and split citation problems in digital libraries. *Proceedings of the 2nd international workshop on Information quality in information systems*, pages 69–76, 2005. doi: <http://doi.acm.org/10.1145/1077501.1077514>. URL <http://doi.acm.org/10.1145/1077501.1077514>.
- I. V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. pages 707–710, 1966.
- G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, 1975.