# Combinação de Evidências para Identificação de Comunidades na *Web*

Álvaro R. Pereira Jr[1] Nivio Ziviani[1] Ricardo Baeza-Yates[2]

[1]*Departamento de Ciência da Computação*
*Universidade Federal de Minas Gerais*
*Belo Horizonte, Brasil*
{*alvaro, nivio*}*@dcc.ufmg.br*

[2]*Departamento de Ciencias de la Computación*
*Universidad de Chile*
*Santiago, Chile*
*rbaeza@dcc.uchile.cl*

## Resumo

*Este trabalho apresenta um mecanismo para a identificação de comunidades na Web. O mecanismo se baseia na construção de um grafo rotulado não direcionado onde os vértices representam documentos Web e as arestas representam relações de similaridades entre os documentos Web. Os rótulos das arestas representam pesos entre vértices. O cálculo do peso de cada aresta do grafo é feito com base em evidências diferentes, compreendidas entre evidências de ligações, evidências sintáticas e evidências de localização. Quanto maior o peso das arestas, maior a similaridade entre os documentos Web. Uma vez construído o grafo, documentos com maiores pesos são agrupados, buscando identificar comunidades, através de um algoritmo baseado na árvore geradora mínima (AGM) do grafo. Experimentos preliminares foram realizados utilizando uma coleção de* $1.300$ *páginas pessoais.*

## Abstract

*This work presents a mechanism for identifying communities on the Web. The mechanism is based on the construction of a non-directed labeled graph, whose vertices represent Web documents and edges represent similarities among Web documents. The labels of edges represent the weights among vertices. The weight of each edge is based on distinct types of evidences such as syntactics, links and location evidences. Bigger the weight of edges, greater the similarity among documents. Once the graph is constructed, documents with the biggest weights are clustered, aiming to identify communities, by an algorithm based on the minimum spanning tree of the graph. Preliminary experiments were performed using a collection of* $1,300$ *personal pages.*

## 1 Introdução

Este trabalho apresenta parte de um mecanismo para identificação de comunidades na *Web*, realizando agrupamento de documentos de assuntos relacionados. O mecanismo é composto de dois estágios: combinação de evidências para geração de grafo rotulado e agrupamento. Nesse trabalho será apresentado o estágio de combinação de evidências.

No estágio de combinação de evidências um grafo é construído a partir do grafo da *Web* (no caso desse trabalho, de uma coleção de documentos HTML). O grafo da *Web* considera cada documento como sendo um vértice e cada ligação como sendo uma aresta. Três tipos de evidências são consideradas: evidências de ligação, evidências sintáticas e evidências de localização. As evidências de ligações (*links*), quando identificadas, podem introduzir novas arestas no grafo, além das arestas já existentes que identificam uma ligação entre documentos. As evidências sintáticas e de localização são somente complementares, podendo contribuir com o crescimento do peso das arestas, fortalecendo indícios de similaridade entre documentos. Detalhes sobre o estágio de combinação de evidências são apresentados na Seção 2. No estágio de agrupamento é utilizada a árvore geradora mínima para particionamento do grafo. Foi utilizada a implementação apresentada em [1] do algoritmo de Prim para obter a árvore geradora mínima do grafo.

## 2 Combinação de Evidências

As evidências consideradas são classificadas em três grupos: evidências de ligações, evidências sintáticas e evidências de localização. As evidências de ligações se

baseiam na perspectiva da *Web* como um grafo direcionado [2], onde os documentos representam os vértices e as ligações (*links*) representam arestas direcionadas. As evidências sintáticas utilizam informações textuais dos documentos para inferir relações entre os mesmos. Evidências de localização estão relacionadas ao domínio no qual ele se encontra hospedado. A cada evidência verificada é atribuído um peso entre zero e um para a aresta, valor que é somado ao peso já atribuído para aquela aresta, caso exista.

O mecanismo proposto para a fase de combinação de evidências é apresentado na Figura 1 e funciona da seguinte forma: na etapa 1 da figura é feito o *parsing* da coleção de documentos HTML, de forma a retirar os textos, os títulos e as ligações de cada documento. Na etapa 2 a estrutura de ligações entre os documentos da coleção é usada para gerar um grafo direcionado. As evidências de ligações são processadas de acordo com as direções das arestas, retornando um grafo não direcionado rotulado, com novas arestas, o que ocorre na etapa 3. Na etapa 4, as evidências sintáticas e de localização são processadas comparando somente os documentos que estão conectados por arestas, ou seja, documentos que até este ponto possuem algum indício de tratarem de assuntos relacionados. Nesse ponto, pares de documentos que já possuírem ligação e que apresentarem similaridades baseadas em evidências sintáticas ou de localização, terão os pesos de suas arestas acrescidos, como pode ser verificado no grafo da Figura 1. Um grafo não direcionado rotulado é gerado. As próximas seções apresentam as evidências consideradas neste trabalho.

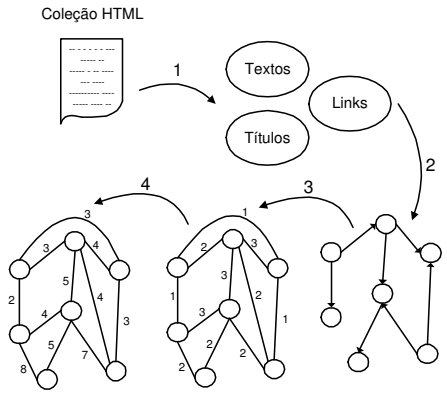


**Figura 1: Mecanismo proposto para a fase de combinação de evidências.**

## 2.1 Evidências de ligações

Foram consideradas quatro evidências diferentes de ligações, a saber: estrutura de ligações entre os documentos; relação de co-citação; relação de acoplamento bibliográfico e Amsler. Amsler combina as relações de co-citação e acoplamento bibliográfico. A estrutura de ligações é a primeira evidência de ligação a ser considerada. Portanto, se existe uma ligação de uma página $A$ para uma página $B$, existe também uma aresta entre essas páginas, com peso 1. Durante a fase de coleta de evidências de ligações novas arestas serão inseridas a medida em que as evidências são combinadas. As seções seguintes apresentam os demais tipos de evidências considerados. Detalhes sobre os tipos de evidências podem ser encontrados em [3].

### 2.1.1 Co-Citação

A co-citação foi proposta por Small [4] como uma medida de similaridade entre artigos científicos. Dois artigos são co-citados se um terceiro artigo tem citações para ambos. Esta afirmação supõe que um autor de artigos científicos citará apenas artigos relacionados com seu próprio trabalho. Apesar de ligações da *Web* serem bastante diferentes de citações, é possível assumir que muitos deles têm o mesmo significado, ou seja, um autor de uma página *Web* irá inserir ligações para páginas relacionadas à sua própria página. Assim, é possível aplicar a co-citação em documentos *Web* tratando ligações como citações.

### 2.1.2 Acoplamento Bibliográfico

Também com o objetivo de determinar a similaridade entre páginas, kessler [5] introduziu a medida de acoplamento bibliográfico. Dois documentos dividem uma unidade de acoplamento bibliográfico se ambos citam o mesmo artigo. Esta idéia é baseada na noção de que autores de artigos que trabalham no mesmo assunto tendem a citar os mesmos artigos. Assim como na co-citação pode-se aplicar este princípio para a *Web*. Assume-se que dois autores de páginas *Web* sobre o mesmo assunto tendem a inserir ligações para as mesmas páginas. Podemos dizer então que duas páginas tem uma unidade de acoplamento bibliográfico entre si se ambas possuem uma ligação para a mesma página.

### 2.1.3 Amsler

Em uma tentativa de tirar maior vantagem da informação disponível entre artigos, Amsler [6] propôs uma medida de similaridade que combina co-citação e

acoplamento bibliográfico. De acordo com Amsler, dois artigos A e B são relacionados se (1) A e B são citados pelo mesmo artigo, (2) A e B citam o mesmo artigo, ou (3) A cita um terceiro artigo C que cita B. Como para as medidas anteriores, pode-se aplicar a similaridade de Amsler para medir páginas *Web*, substituindo citações por ligações. A similaridade de Amsler também foi usada como uma evidência de similaridade entre as páginas *Web* nesse trabalho. As fórmulas de Amsler, co-citação e acoplamento bibliográfico podem ser obtidas em [3], bem como maiores detalhes sobre suas definições.

## 2.2 Evidências Sintáticas

Três evidências sintáticas diferentes foram consideradas, a saber: evidência de texto do documento; evidência de título e evidência de texto da URL, que serão detalhadas nas seções seguintes.

### 2.2.1 Evidência de Texto do Documento

Uma evidência de texto é identificada se o conteúdo texto de dois documentos ligados por arestas apresentam alguma similaridade. A similaridade é computada utilizando o conceito de *shingles*, proposto por Broder [7]. *Shingles* já foram usados para verificação de similaridade entre documentos [8]. O conjunto de *shingles* de tamanho $k$ de um documento é o conjunto de todas as sequências diferentes de $k$ termos do mesmo, considerando interseção de termos. Como exemplo: o conjunto de *shingles* do texto: "uma rosa é uma rosa é uma rosa" é: {(uma, rosa, é, uma), (rosa, é, uma, rosa), (é, uma, rosa, é)}, resultando em três *shingles* diferentes de tamanho $4$.

Para identificar uma evidência de texto, todos os *shingles* de um determinado documento do grafo são inseridos em uma tabela *hash* [1] e alguns *shingles* de cada documento que possui aresta com o documento em questão são pesquisados na tabela. Quanto mais *shingles* encontrados maior é a evidência de que os textos sejam similares, ou ao menos tratem do mesmo assunto. Se todos os *shingles* são encontrados, a evidência de texto fica em $1$.

### 2.2.2 Demais Evidências Sintáticas

Os títulos das páginas HTML foram identificados e agora são usados como um indício de similaridade entre documentos. A verificação é simples. Os termos do título de um documento são comparados com os termos dos títulos de todos os documentos que possuem aresta com o mesmo. Quanto maior o número de termos em comum, maior é o valor da evidência de título. Como exemplo, considere uma aresta $A$–$B$. Se o título de $A$ possui três termos, e dois termos do título de $B$ são encontrados entre os três termos de $A$, o peso da aresta $A$–$B$ fica acrescido de $0,66$.

Os termos contidos no endereço de URL das páginas também foram identificados e usados como indício de similaridade entre documentos. É considerado um termo todo o trecho de caracteres que estiverem entre barras. Por exemplo, para a URL:
http://www.dcc.ufmg.br/usuario/teste/index.html
temos três diferentes termos: usuario, teste e index.html. Quanto maior o número de termos na URL de um documento $B$ que são encontrados na URL de um documento $A$, maior a evidência de texto da URL dos documentos, contribuindo com o peso da aresta $A$–$B$.

## 2.3 Evidência de Localização

A única evidência de localização considerada foi o domínio da Internet. Páginas hospedadas sob o mesmo domínio podem abordar assuntos semelhantes. Se as páginas $A$ e $B$ estão sob o mesmo domínio da Internet, a evidência de localização contribui para o peso da aresta $A$–$B$ em $1$. Caso contrário, não existe contribuição. Como exemplo, considere as páginas:
$A$ – http://www.dcc.ufmg.br/usuario/teste/index.html
$B$ – http://www.dcc.ufmg.br/usuario/teste/index.html
$C$ – http://www.ufmg.br/novo/test/index2.html
Consideramos que as páginas $A$ e $B$ estão sob o mesmo domínio www.dcc.ufmg.br. No entanto, o domínio da página $C$ não é o mesmo de $A$ e $B$: www.ufmg.br.

# 3 Resultados Parciais

Foram realizados experimentos utilizando uma pequena coleção de $1.300$ páginas HTML sob o domínio www.dcc.ufmg.br. Para compor essa coleção foram utilizadas páginas pessoais de professores e alunos do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais. Infelizmente a coleção utilizada não permitiu a verificação de resultados fortes, devido ao seu tamanho limitado e ao baixo número de ligações entre as páginas. Foi realizada uma análise manual, e uma instância verificada será apresentada a seguir.

A Tabela 1 apresenta as URLs dos documentos de um dos agrupamentos retornados pelo algoritmo, que podem ser acessados via Web. Acreditamos que os documentos que compõem esse agrupamento realmente

tratam de assuntos correlacionados. Em uma primeira análise desse agrupamento, suspeitamos que ele pudesse ter sido formado devido aos seguintes fatores: $a$) as páginas estão sob o mesmo domínio, o que acontece com todos os documentos da coleção e $b$) o texto das URL ser similares, por serem do mesmo autor, com características similares (*home pages* de cursos anuais). Dessa forma, as evidências de localização e de texto da URL se destacaram, podendo não validar a boa performance do algoritmo.

Com as suspeitas identificadas, executamos novamente o algoritmo, porém atribuindo pesos bem maiores às evidências de texto. O resultado para esse agrupamento foi similar. Os documentos da execução anterior, apresentados na Tabela 1, também foram retornados no mesmo agrupamento. De fato, além da similaridade clara de localização e texto da URL, percebemos similaridade no título e no texto das páginas que compuseram esse agrupamento.

**Tabela 1: URLs dos documentos de um agrupamento retornado.**

| Documento |
|---|
| http://www.dcc.ufmg.br/~nvieira/cursos/tl/a00s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/tl/a02s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a01s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a02s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/tl/a03s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a04s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a02s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/md/a00s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/md/a02s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/md/a02s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/tl/a96s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/tl/a97s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/md/a97s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a01s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a00s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ftc/a97s1/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/lac/a03s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/ia/gr/a01s2/material.html |
| http://www.dcc.ufmg.br/~nvieira/cursos/md/a00s1/material.html |

## 4 Conclusões e Trabalhos Futuros

Este trabalho apresentou parte de um mecanismo para identificação de comunidades na *Web*, realizando agrupamento de documentos de assuntos relacionados. O mecanismo é composto de dois estágios: combinação de evidências para geração de grafo rotulado e agrupamento usando árvore geradora mínima para particionamento do grafo.

O mecanismo apresentado deve ser refinado e modelado. Como indicações de trabalhos futuros, sugerimos a identificação de novas evidências além das utilizadas e a avaliação do quanto cada evidência contribui para a formação coerente de um agrupamento. A coleção utilizada nos experimentos foi bastante limitada. Deve-se trabalhar com coleções maiores e com documentos mais heterogêneos, além de buscar formas de avaliar o modelo. Uma maneira de avaliar seria aplicá-lo ao problema da classificação. Ainda como indicação de trabalho futuro, seria interessante realizar agrupamento de agrupamentos. Certamente existem agrupamentos correlacionados, que devem ser identificados e agrupados. E ainda, buscar novos algoritmos para particionamento do grafo, além do algoritmo de Prim utilizado.

## Agradecimentos

Agradecemos ao colega Thierson Couto Rosa pela ajuda com relação à análise de ligações e disponibilização de seus programas, relacionados à parte de evidências de ligações, e ao colega José Augusto Nacif, por estudar, desenvolver e executar os algoritmos de particionamento do grafo.

## Referências

[1] N. Ziviani, *Projeto de Algoritmos com Implementações PASCAL e C*, 2nd ed. Pioneira Thomson Learning.

[2] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a graph: Measurements, models and methods," *Lecture Notes in Computer Science*, vol. 1627, pp. 1–18, 1999.

[3] M. Cristo, P. Calado, E. S. de Moura, N. Ziviani, and B. Ribeiro-Neto, "Link information as a similarity measure in web classification," in *Proceedings of the String Processing and Information Retrieval SPIRE, Lecture Notes in Computer Science 2857.* Springer Verlag, 2003, pp. 43–55.

[4] H. G. Small, "Co-citation in the scientific literature: A new measure of relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, pp. 265–269, 1973.

[5] M. M. Kessler, "Bibliographic coupling between scientific papers," American Documentation, 1963.

[6] R. Amsler, "Application of citation-based automatic classification," University of Texas at Austin, Linguistics Research Center, Austin, Texas, USA, Tech. Rep., 1972.

[7] A. Broder, "On the resemblance and containment of documents," in *Compression and Complexity of Sequences (SEQUENCES'97).* IEEE Computer Society, 1998, pp. 21–29.

[8] A. R. Pereira-Jr and N. Ziviani, "Syntactic similarity of web documents," in *First Latin American Web Congress*, Santiago, Chile, November 2003, pp. 194–200.