Construção Automática de Diretórios Web usando Agrupamento Incremental de Termos

Ricardo M. Marcacini[1], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos - SP - Brasil

{rmm,solange}@icmc.usp.br

Abstract. *Hierarchical clustering methods are useful to support the construction of web directories in a unsupervised way. However, the traditional methods are ineffective in dynamic scenarios, with constant updating of knowledge. Moreover, these methods obtain a hierarchical structure that is difficult to be interpreted by users. In this paper, we propose an incremental term clustering approach that allows (1) the organization of document collections in dynamic scenarios and (2) obtain cluster descriptors to support the interpretation of results. An experimental evaluation was carried out through real data from a web directory, which presented good results.*

Resumo. *Métodos baseados em agrupamento hierárquico de documentos são úteis para apoiar a construção de diretórios web de maneira não supervisionada. No entanto, os métodos tradicionais são ineficientes em cenários dinâmicos, com constante atualização do conhecimento. Além disso, estes métodos obtêm uma estrutura hierárquica que é de difícil interpretação para os usuários. Neste trabalho, é proposta uma abordagem de agrupamento incremental de termos que permite (1) organizar coleções de documentos em cenários dinâmicos e (2) obter descritores ao agrupamento para apoiar a interpretação dos resultados. Uma avaliação experimental foi realizada em dados reais de um diretório web, apresentando bons resultados.*

1. Introdução

As plataformas online para publicação e armazenamento de conteúdo digital, como a Internet, têm contribuído de forma significativa para o surgimento de diversos repositórios textuais. Devido à necessidade de extrair conhecimento útil desses repositórios, abordagens para organização automática de documentos têm recebido grande atenção na literatura [Feldman and Sanger 2006, Premalatha and Natarajan 2010]. A organização do conteúdo da Internet nos chamados "diretórios web" é uma das abordagens mais populares, em que os documentos da web são organizados em diretórios e subdiretórios, e cada diretório contém documentos relacionados a um mesmo tema [Yang and Lee 2004, Kim 2006]. Desta forma, o usuário pode visualizar a informação de interesse em diversos níveis de granularidade e explorar interativamente grandes coleções de documentos.

Os diretórios web desempenham um papel importante na recuperação de informação da Internet, principalmente em tarefas de busca exploratória. Neste tipo de tarefa, o

usuário geralmente tem pouco domínio sobre o tema de interesse, dificultando expressar o objetivo diretamente por meio de palavras-chave [Marchionini 2006]. Assim, é necessário disponibilizar previamente algumas opções para guiar o processo de busca da informação. Para tal, cada diretório possui um conjunto de descritores que contextualizam e indicam o significado dos documentos ali agrupados. Esta organização está relacionada com a hipótese de que se um usuário está interessado em um documento específico pertencente a um determinado diretório deve também estar interessado em outros documentos desse diretório e de seus subdiretórios [Manning et al. 2008].

A maioria dos diretórios web é construída de maneira supervisionada, a exemplo do *Dmoz - Open Directory Project*[1] e *Yahoo! Directory*[2], editadas manualmente por especialistas humanos. Apesar da boa qualidade dos resultados obtidos, a construção manual desses diretórios é limitada pela grande quantidade de documentos disponíveis e pela alta frequência de atualização. Deste modo, métodos para automatizar a construção de diretórios web têm sido explorados, em especial com uso de algoritmos de agrupamento hierárquico [Yang and Lee 2004, Kim 2006].

Os algoritmos de agrupamento hierárquico organizam, de forma não supervisionada, um conjunto de objetos em grupos, no qual objetos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos objetos de outros grupos [Xu and Wunsch 2008]. Ainda, a organização é disposta em grupos e subgrupos, formando-se uma estrutura hierárquica similar à dos diretórios web. Por outro lado, os algoritmos clássicos de agrupamento hierárquico possuem algumas limitações que desafiam a construção automática desses diretórios. Uma das limitações está relacionada com a interpretação dos resultados do agrupamento, pois apenas a organização em grupos e subgrupos geralmente não é suficiente para análise dos resultados [Moura and Rezende 2010]. Assim, é necessário também extrair características do agrupamento que descrevem o significado de cada grupo e apresentá-las ao usuário durante a busca exploratória. Outra limitação se deve ao ambiente dinâmico dos diretórios web, uma vez que novos documentos são publicados a todo momento na Internet. O agrupamento hierárquico geralmente obtém uma estrutura estática para organização dos objetos, o que inviabiliza incorporar diretamente novo conhecimento disponível. Repetir todo o processo de agrupamento sempre que ocorrer atualizações significativas é computacionalmente custoso em grandes bases de dados. Neste cenário, é desejável que o agrupamento seja realizado de maneira incremental [Fung et al. 2008].

Em vista desses desafios, neste trabalho é proposta e avaliada uma abordagem para construção automática de diretórios web. Para tal, foi adaptado um algoritmo de agrupamento incremental de termos que permite adicionar novo conhecimento sem reprocessamento redundante. Além disso, durante o processo de agrupamento são extraídos automaticamente descritores para cada grupo, visando auxiliar a interpretação e exploração da organização. Uma avaliação experimental realizada sobre um conjunto de dados reais, obtido do Diretório Web *Dmoz (Open Directory Project)*, indica que a abordagem proposta obtém resultados competitivos em comparação com uma abordagem similar da literatura. Uma visão geral das contribuições alcançadas durante o desenvolvido deste trabalho é descrita a seguir:

[1]Dmoz - Open Directory Project: http://www.dmoz.org/

[2]Yahoo Directory!: http://dir.yahoo.com/

- É apresentada a ideia de agrupamento de termos para a construção de diretórios web. Para isto, foi realizada uma adaptação do algoritmo IHTC (*Incremental Hierarchical Term Clustering*) [Marcacini and Rezende 2010a], que permite extrair grupos de termos (palavras e expressões) para identificação dos tópicos existentes nos textos. Os grupos de termos são associados a grupos de documentos, auxiliando a interpretação dos resultados e atividades de busca exploratória.
- Para lidar com o cenário dinâmico dos diretórios web, é explorada uma estratégia de agrupamento incremental visando obter uma representação condensada dos dados. Um algoritmo de agrupamento hierárquico é aplicado sobre esta representação condensada e, assim, é possível extrair uma organização hierárquica com custo computacional reduzido.
- Por fim, um módulo computacional para construção e análise de diretórios web a partir de textos foi desenvolvido e incorporado à ferramenta *Torch - Topic Hierarchies* [Marcacini and Rezende 2010b], que está disponível *online*[3] juntamente com o conjunto de dados utilizado na avaliação experimental.

O restante deste artigo está organizado da seguinte maneira. Na Seção 2 é apresentada uma breve revisão da literatura, com os conceitos principais sobre agrupamento de documentos e seu uso na organização não supervisionada de coleções textuais. A abordagem proposta neste trabalho é detalhada na Seção 3, em que é descrito o processo de agrupamento incremental de termos, a estratégia para obter uma representação condensada dos textos e a construção dos diretórios web. Na Seção 4 é apresentada a avaliação experimental da abordagem proposta bem como a discussão dos resultados. Por fim, as conclusões e as direções para trabalhos futuros são descritas na Seção 5.

2. Conceitos Básicos

A tarefa de construção automática de diretórios web pode ser vista como um problema de agrupamento de documentos [Yang and Lee 2004]. Para tal, é necessário definir um modelo para representação estruturada dos textos, uma medida de similaridade entre os documentos e uma estratégia para formação dos grupos [Feldman and Sanger 2006].

O modelo espaço-vetorial é uma das formas mais comuns para representação dos textos [Manning et al. 2008]. Neste modelo, cada documento é representado como um vetor $x = \{t_1, t_2, ..., t_m\}$ no espaço formado pelos termos da coleção textual. Os termos são palavras selecionadas para representar os documentos de forma concisa e estruturada. Cada termo t_i possui um valor associado que indica sua importância (peso) para determinado documento. Em particular, neste trabalho utiliza-se a frequência de ocorrência de t_i no documento d, pois é um critério computacionalmente eficiente e, ainda, é facilmente aplicável no processamento de documentos em cenários dinâmicos. Um grupo de documentos $G = \{x_1, x_2, ..., x_n\}$ também possui representação no modelo espaço-vetorial por meio de um centroide $C_G = (\sum_{i=0}^{n} x_i)/|G|$, o vetor médio de todos os documentos de G.

A similaridade entre dois documentos (ou grupos de documentos) representados no modelo espaço-vetorial é usualmente computada pela medida cosseno [Zhao et al. 2005]. Nesta medida, sejam x_i e x_j dois documentos, o ângulo cosseno $cos(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}$ tem valor 1 quando os dois documentos são idênticos e valor 0

[3]Torch - Topic Hierarchies: http://sites.labic.icmc.usp.br/marcacini/ihtc/

quando não possuem nenhum termo em comum (são ortogonais entre si). Em alguns casos, é útil adaptar a medida de similaridade cosseno para uma medida de dissimilaridade por meio da equação $dis(x_i, x_j) = 1 - cos(x_i, x_j)$.

As estratégias para formação dos grupos, com uso de algoritmos de agrupamento hierárquico, podem ser divididas em aglomerativas e divisivas [Xu and Wunsch 2008]. No agrupamento hierárquico aglomerativo, inicialmente cada documento é um grupo unitário e, em cada iteração, os pares de grupos mais próximos são unidos até formar um único grupo. Já no agrupamento hierárquico divisivo, inicia-se com um grupo contendo todos os documentos que é, então, dividido em grupos menores até restarem grupos unitários. Avaliações experimentais indicam que o algoritmo UPGMA (aglomerativo) e o Bisecting-kmeans (divisivo) são os que obtêm melhores resultados em dados textuais [Zhao and Karypis 2002]. No entanto, tanto as estratégias aglomerativas quanto as divisivas possuem, em geral, complexidade quadrática de tempo e espaço, o que limita sua aplicação em grandes bases de dados e em cenários dinâmicos que requerem frequentes atualizações no agrupamento.

Para superar esta limitação, pode-se utilizar um método de agrupamento baseado em condensação [Xu and Wunsch 2008]. Nesses métodos, o agrupamento hierárquico é aplicado sobre uma representação condensada dos dados ao invés do conjunto total dos dados. A representação condensada mantém um sumário com as características principais dos dados, possibilitando processar grandes bases de dados com redução significativa do custo computacional [Nassar et al. 2004, Xu and Wunsch 2008]. Para exemplificar, considere a base de dados ilustrada na Figura 1. Em um primeiro momento, o agrupamento hierárquico é obtido a partir do conjunto total dos dados (Figura 1a). Alternativamente, na Figura 1b, a base de dados é sumarizada em uma representação condensada em 14 grupos. O agrupamento hierárquico aplicado sobre esta representação obtém uma estrutura hierárquica com propriedades similares à anterior. No entanto, ao usar 14 centroides como representação condensada, há uma significativa redução do custo computacional.

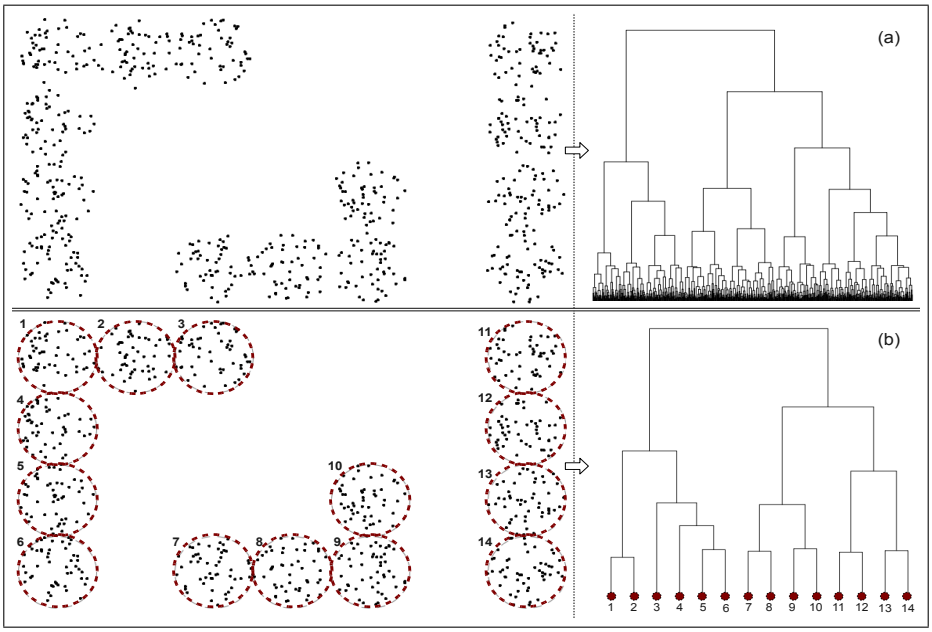

Figura 1. (a) Agrupamento hierárquico. (b) Agrupamento hierárquico resultante da representação condensada dos dados.

A técnica utilizada para obter a representação condensada dos dados, em cenários dinâmicos como os diretórios web, deve construir e atualizar esta representação maneira incremental. Para tal, são utilizados algoritmos de agrupamento incremental conhecidos como "*single-pass*", capazes de obter uma estrutura de agrupamento com poucas iterações sobre os dados, geralmente em tempo linear [Bradley et al. 1998, Farnstrom et al. 2000, Xu and Wunsch 2008]. Em um algoritmo de agrupamento "*single-pass*", cada novo objeto inserido é alocado ao grupo existente mais próximo, por exemplo, computando a similaridade entre o novo objeto e os centroides dos grupos existentes. Se este valor de similaridade não satisfizer um critério predeterminado, então um novo grupo é criado para alocar o objeto. O número máximo de grupos a ser criado define o nível de condensação da representação, e pode ser controlado de acordo com a memória disponível para a aplicação. O algoritmo Leader [Jain et al. 1999, Xu and Wunsch 2008] é um dos mais simples nesta categoria, e é utilizado em vários trabalhos que envolvem agrupamento incremental. No Leader, o agrupamento é obtido de maneira incremental com complexidade de tempo $O(n \cdot k)$ e de espaço $O(k)$, em que n é o número de objetos e k o número máximo de grupos. Ao usar o Leader para obter a representação condensada dos dados, a complexidade do agrupamento hierárquico a partir desta representação é de $O(k^2)$, em contraste com $O(n^2)$ quando aplicado em todo o conjunto de dados (em geral $n \gg k$).

A interpretação do agrupamento é uma dificuldade neste processo. Em geral, é necessário aplicar uma técnica após o agrupamento para seleção dos termos mais discriminantes e utilizá-los como descritores do agrupamento [Manning et al. 2008], por exemplo, pela seleção dos termos mais importantes em relação ao centroide de cada grupo [Moura and Rezende 2010]. A abordagem proposta neste trabalho obtém descritores juntamente com o processo de agrupamento incremental.

3. Construção Automática de Diretórios Web

Na seção anterior foi discutido um processo de agrupamento no qual similaridades entre documentos são o foco central na obtenção dos grupos. Neste trabalho, é explorada uma abordagem focada no agrupamento de termos da coleção textual. O objetivo é extrair grupos de termos correlacionados de acordo com suas distribuições nos documentos. Os grupos de termos identificam conceitos e tópicos implícitos nos textos, de forma que documentos relacionados a um mesmo tópico possam ser alocados em um mesmo grupo.

O método IHTC (*Incremental Hierarchical Term Clustering*), proposto inicialmente em [Marcacini and Rezende 2010a], permite organizar coleções textuais utilizando uma estratégia de agrupamento incremental de termos. A principal vantagem do IHTC é obter o agrupamento de documentos com descritores associados, permitindo a análise exploratória de coleções textuais.

Neste trabalho, o IHTC foi adaptado para obter um representação condensada dos dados e, assim, apoiar a construção automática de diretórios web. Uma visão geral da abordagem proposta é apresentada na Figura 2. A abordagem é dividida em quatro fases: (1) construção da rede de coocorrência de termos; (2) agrupamento de termos; (3) agrupamento de documentos e (4) extração da organização hierárquica. As três primeiras fases são responsáveis por obter uma representação condensada da coleção textual, de maneira incremental, que é utilizada na organização hierárquica dos textos na última fase.

Uma **rede coocorrência de termos** é definida como um grafo no qual os vértices

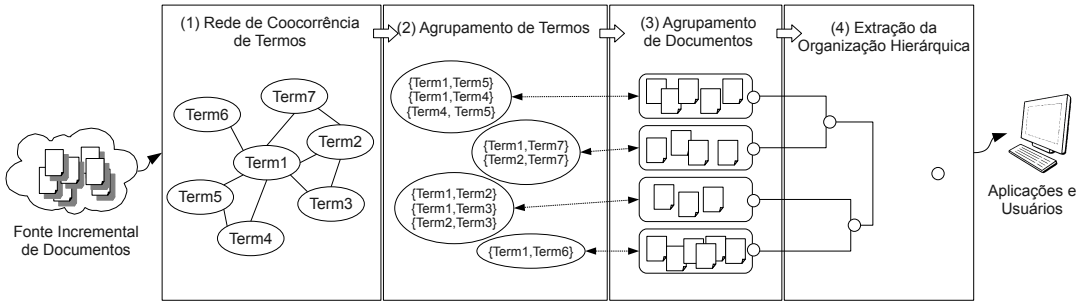

Figura 2. Visão geral da abordagem proposta

são os termos existentes na coleção textual, mais especificamente, termos selecionados para representação de cada documento no modelo espaço-vetorial. No entanto, nem todos os termos são utilizados como vértices. Para um termo ser utilizado, é necessário que o mesmo participe de uma relação de coocorrência com outro termo da coleção, ou seja, ocorrerem juntos em múltiplos documentos da coleção textual. A coocorrência entre dois termos também identifica as arestas do grafo. Assim, dois termos são conectados por uma aresta se existe frequência de coocorrência significativa entre eles, que é dependente da coleção textual. Em cada aresta, é associado um centroide que identifica relação entre dois termos t_i e t_j, por meio de uma função $w(e) = C(t_i \cap t_j)$, em que $C(t_i \cap t_j)$ é o centroide que representa o subconjunto de documentos com ambos os termos t_i e t_j. Desta forma, a rede de coocorrência de termos, conforme aplicada neste trabalho, pode ser vista como uma estrutura com duas características principais:

1. Capaz de identificar relações significativas entre os termos da coleção textual, baseada na frequência de coocorrência; e
2. Capaz de extrair subconjuntos de documentos (representados por centroides), em que os pares de termos (arestas) podem ser utilizados como descritores.

Algoritmo 1: Construção incremental da rede de coocorrência de termos

Parâmetros:
$X_{inc} = \{x_1, x_2, ...\}$: fonte incremental de documentos
r: número máximo de arestas para a rede de coocorrência de termos

```
1  M ← inicializar lista global com r contadores;
2  para cada novo documento x ∈ X_inc faça
3      E_x ← pares de termos obtidos do documento x;
4      para cada par de termos e = {t_i, t_j} ∈ E_x faça
5          se e = {t_i, t_j} está na lista de contadores M então
6              m_e ← contador de e na lista M;
7              m_e ← m_e + 1;
8              atualizar centroide da aresta e com o documento x;
9          senão
10             m_old ← menor contador da lista M;
11             m_new ← novo contador para e;
12             m_new ← m_old + 1;
13             substituir m_old por m_new na lista M;
14         fim
15     fim
16 fim
```

No IHTC, a rede de coocorrência de termos é construída e atualizada de forma

incremental, conforme apresentado no Algoritmo 1 [Marcacini and Rezende 2010a]. O algoritmo proposto é inspirado em uma técnica de detecção de elementos frequentes em "*data streams*" [Metwally et al. 2005]. Neste algoritmo, cada par de termos (aresta) extraído dos textos é monitorado por r contadores. Se um par de termos $e = \{t_i, t_j\}$ é frequente, então os termos t_i e t_j podem ser conectados na rede de coocorrência. Para tal, cada novo documento apresentado é processado individualmente. Os diferentes pares de termos do documento x são extraídos e armazenados em E_x (Linha 3). Para cada par de termos $e = \{t_i, t_j\} \in E_x$, é verificado se o mesmo já está sendo monitorado por algum contador. Se e já estiver sendo monitorado (Linha 5), significa que foi adicionado por um documento anterior, então o valor de frequência de coocorrência de e é incrementado (Linha 7) e a rede é atualizada (Linha 8). Caso contrário, busca-se o contador com menor valor de frequência m_{old} (Linha 10) e, em seguida, m_{old} é substituído pelo contador de e, identificado como m_{new} (Linha 13). Um detalhe desta substituição é que o valor para m_{new} é calculado como $m_{old} + 1$ (Linha 12), ou seja, o valor é superestimado a fim de garantir que cada elemento frequente seja monitorado. Desta forma, ao longo do tempo a lista de contadores mantém os pares de termos mais frequentes, enquanto os menos frequentes são gradativamente removidos. O número de contadores r é um parâmetro do algoritmo e define o tamanho máximo de arestas da rede de coocorrência de termos.

O objetivo da fase de **agrupamento de termos** é sumarizar as relações existentes na rede de coocorrência de termos. Para cada atualização ocorrida na rede, o respectivo centroide (aresta) é visto como um novo objeto por um algoritmo do tipo "*single-pass*". Assim, mantém-se o agrupamento de termos de forma incremental, no qual cada grupo possui pares de termos correlacionados que identificam tópicos existentes nos textos.

Na fase de **agrupamento de documentos**, os documentos da coleção são mapeados aos grupos de termos mais próximos. Assim, após o documento x passar pela fase de construção da rede de coocorrência e da fase de agrupamento de termos, calcula-se a similaridade cosseno entre x e o centróide de cada grupo de termos existente naquela iteração. Em seguida, o documento x é mapeado ao grupo de termos com maior valor de similaridade. Desta forma, o agrupamento final é formado por grupos do tipo $G_i = (G_i^d, G_i^t)$, ou seja, composto por um grupo de documentos G_i^d e um grupo de termos G_i^t; e os termos existentes em G_i^t são utilizados como descritores de G_i^d.

Por fim, a **extração da organização hierárquica** é feita com base na representação condensada dos dados obtida até o momento. Para tal, pode ser aplicado um algoritmo de agrupamento hierárquico sobre a representação condensada.

A complexidade de tempo para manter a representação condensada dos dados é linear em termo de números de documentos [Marcacini and Rezende 2010a], o que torna a abordagem útil em cenários dinâmicos. A organização hierárquica final possui grupos e descritores associados. Em diretórios web, cada grupo representa um diretório e os descritores do grupo são selecionados para indicar o conteúdo dos documentos ali agrupados.

4. Avaliação Experimental

A abordagem proposta neste trabalho para construção de diretórios web foi avaliada experimentalmente em uma base de dados real proveniente do projeto Dmoz ("*Open Directory Project*"). O Dmoz é um diretório público gerenciado manualmente por uma comunidade global de editores voluntários. Atualmente, várias ferramentas de busca na internet, como

o Google e AOL Search, incorporam o Dmoz em seus sistemas.

Em vista da relevância e da disponibilidade pública da sua base de dados, o Dmoz foi utilizado neste trabalho para um estudo exploratório sobre a construção automática de diretórios web. O objetivo é avaliar a abordagem proposta com o método IHTC e compará-la com uma estratégia similar da literatura baseada no algoritmo Leader [Jain et al. 1999, Xu and Wunsch 2008], analisando-se a qualidade do agrupamento hierárquico e dos descritores dos grupos.

4.1. Base de dados

Para a avaliação experimental, foi selecionado um subconjunto dos dados com 5 diretórios, provenientes do projeto Dmoz: "*Business*", "*Computers*", "*Health*", "*Science*" e "*Sports*". Cada diretório possui subdiretórios e documentos relacionados. Ao final, obteve-se uma base de dados com 566.234 documentos organizados em 43.817 diretórios e subdiretórios. Na Tabela 1 é apresentada uma visão geral da base de dados.

Tabela 1. Visão geral da base de dados selecionada a partir da base do Dmoz

Características da base de dados	
Documentos da Coleção	566234
Número de Termos	246280
Número de Termos com DF ≥ 2	84039
Número de (Sub)Diretórios	43817
Altura da Hierarquia	11
Média de Termos por Documento	12,41

O número total de termos (após remoção de *stopwords*[4] e aplicação de *stemming*) é de 246.280. Deste total, há 84.039 termos que ocorrem em dois ou mais documentos (DF ≥ 2) e que podem ser úteis para a tarefa de agrupamento. Os documentos associados aos diretórios da hierarquia representam uma página na web. Cada documento é composto por quatro campos: (1) o título da página, (2) a URL com o endereço de internet para a página, (3) uma breve descrição com 25 à 30 palavras sobre o conteúdo da página; e (4) o diretório da hierarquia na qual a página está alocada (removido para a realização dos experimentos). Por fim, foram selecionados 5 subconjuntos de dados, contendo diretórios de referência para serem utilizado na etapa de validação. Assim, é possível avaliar a eficácia da abordagem proposta por meio de critérios objetivos.

4.2. Critério de Avaliação

O índice FScore [Zhao and Karypis 2002] é uma medida que utiliza as idéias de *precisão* e *revocação*, da recuperação de informação, para avaliar a eficácia de recuperação em uma organização hierárquica de documentos. É empregada como critério externo de validação, pois utiliza o conhecimento prévio (informação externa) sobre categorias ou tópicos existentes. A ideia básica é verificar o quanto uma organização hierárquica de uma base de dados consegue recuperar a informação de categoria associada a cada documento. Para o cálculo do índice FScore, considere que

- H é um agrupamento hierárquico obtido por um determinado algoritmo;

[4]Os textos do DMOZ são em inglês, então foi utilizada uma lista de *stopwords* padrão para a língua inglesa (disponível em `http://sites.labic.icmc.usp.br/marcacini/ihtc`).

- L_r é uma determinada categoria/diretório (informação externa) representando um conjunto de documentos de um mesmo tópico; e
- G_i é um determinado grupo, e seu respectivo conjunto de documentos, pertencente ao agrupamento hierárquico H.

Assim, dada uma categoria L_r e um grupo G_i, calcula-se as medidas de precisão P e revocação R conforme a Equação 1 e Equação 2, respectivamente. Em seguida, é obtida a média harmônica F (Equação 3), um balanceamento entre a precisão e revocação.

$$P(L_r, G_i) = \frac{|L_r \cap G_i|}{|G_i|} \qquad (1) \qquad\qquad R(L_r, G_i) = \frac{|L_r \cap G_i|}{|L_i|} \qquad (2)$$

$$F(L_r, G_i) = \frac{2 * P(L_r, G_i) * R(L_r, G_i)}{P(L_r, G_i) + R(L_r, G_i)} \qquad (3)$$

A medida F selecionada para uma determinada categoria L_r é o maior valor obtido por algum grupo da hierarquia *H*, conforme a Equação 4. Finalmente, o valor FScore global de um agrupamento hierárquico com *n* documentos e *c* categorias, é o somatório do valor F das categorias ponderado pelo número de documentos (Equação 5).

$$F(L_r) = \max_{G_i \in H} F(L_r, G_i) \qquad (4) \qquad\qquad FScore = \sum_{r=1}^{c} \frac{|L_r|}{n} F(L_r) \qquad (5)$$

Conforme o agrupamento hierárquico consegue recuperar a informação das categorias predeterminadas de uma coleção, o valor de FScore se aproxima de 1. Caso contrário, a FScore tem valor 0.

O mesmo processo pode ser aplicado para avaliar os descritores selecionados para o agrupamento. Neste caso, o conjunto de documentos associado ao grupo G_i é substituído pelo conjunto de documentos T_i recuperados utilizando os descritores de G_i em uma expressão de busca. Assim, o índice FScore é utilizado tanto para medir a eficácia de recuperação do agrupamento hierárquico H quanto a eficácia de recuperação dos descritores selecionados para H.

4.3. Experimentos Realizados e Análise dos Resultados

Os experimentos foram realizados com auxílio da ferramenta *Torch - Topic Hierarchies* [Marcacini and Rezende 2010b], que atualmente disponibiliza vários algoritmos de agrupamento incremental de documentos, além de técnicas de pré-processamento de textos.

Foram realizados dois experimentos de construção automática de diretórios web. O primeiro com a abordagem proposta, representada por IHTC, e o segundo com base no algoritmo Leader. Para o IHTC, foi definido o parâmetro $r = 20000$ para o tamanho máximo da rede de coocorrência de termos. No caso do Leader, o parâmetro de dissimilaridade mínima foi definido como $\gamma = 0.8$. O agrupamento hierárquico final foi obtido por meio do *Bisecting-kmeans*, a partir de uma representação condensada em 3 mil grupos. A seleção de descritores para o agrupamento obtido com base no Leader é feita por um método baseado em centroides. Os parâmetros utilizados em ambos os algoritmos foram definidos de acordo com experimentos preliminares.

A seguir, é realizada uma análise dos resultados obtidos de acordo com dois critérios calculados com o índice FScore: (1) a eficácia de recuperação do agrupamento

hierárquico; e (2) a eficácia de recuperação dos descritores selecionados para o agrupamento. Para determinar o grau de confiança das comparações realizadas, foi aplicado o teste estatístico não paramétrico de Wilcoxon conforme descrito em [Demšar 2006].

Em relação ao critério de **eficácia da recuperação do agrupamento hierárquico**, o IHTC e o Leader obtiveram resultados similares. Na Figura 3a é ilustrada uma comparação entre a medida FScore em 5 subconjuntos de validação. O Leader apresentou resultados superiores em relação ao agrupamento hierárquico, entretanto, não foi possível indicar diferença estatisticamente significativa na comparação deste critério.

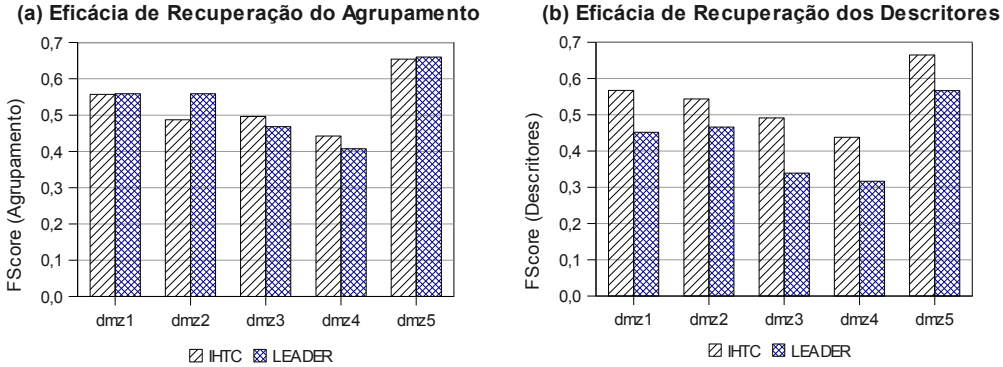

Figura 3. Eficácia de recuperação do agrupamento hierárquico e dos descritores entre o IHTC e Leader

Por outro lado, o IHTC obteve resultados superiores em relação ao critério de **eficácia da recuperação dos descritores**, com diferença estatisticamente significativa. Conforme descrito anteriormente, o IHTC permite fornecer grupos de termos como descritores para o agrupamento. Ao utilizar esses grupos de termos, os descritores selecionados para a hierarquia obtiveram maior eficácia de recuperação do que a estratégia baseada em centroide adotada no algoritmo Leader. Na Figura 3b é ilustrada a comparação da eficácia da recuperação dos descritores.

Este resultado está conformidade com a proposta do IHTC, que visa auxiliar a tarefa de interpretação do agrupamento. Para isto, o IHTC utiliza as relações existentes na rede de coocorrência de termos para identificação de possíveis tópicos na coleção textual. Na Tabela 2 são apresentados alguns dos diretórios de referência extraídos do Dmoz (construída por humanos). Para cada tópico de referência, também são apresentados os respectivos descritores obtidos pelos experimentos realizados com o IHTC e Leader.

É importante observar que o IHTC possui a vantagem de utilizar pares de termos (arestas da rede de coocorrências) como descritores dos grupos. Em geral, uma expressão ou conceito identificado por conjuntos de termos relacionados tem maior poder discriminativo do que termos simples. Mesmo em situações em que termos simples apresentam melhor eficácia de recuperação, o uso de conjunto de termos pode auxiliar mais os usuários na interpretação dos grupos, por exemplo, no caso de "*Cancers, Treatments*" e "*Cancers, Researchs*" em vez do termo simples "*Cancers*".

Os diretórios web completos obtidos com a realização deste trabalho estão disponíveis *online* no endereço http://sites.labic.icmc.usp.br/marcacini/ihtc, incluindo a base de dados utilizada, a ferramenta *Torch*, e um módulo computacional para apoiar a construção e publicação de diretórios web.

Tabela 2. Exemplo de diretórios obtidos automaticamente pelo IHTC e Leader em comparação com o diretório original da Dmoz

Origem	Diretório (Tópico)	Eficácia da Recuperação	
		Agrupamento	Descritores
DMOZ	Cancer	-	-
IHTC	{Cancers,Treatments}, {Cancers,Researchers}, {Cancers,Information}	**0,335**	0,499
Leader	Cancers, Breasts, Nci	0,332	**0,582**
DMOZ	Veterinarians	-	-
IHTC	{Animals,Services}, {Animals,Hospital}, {Hospital,Services}	0,706	**0,691**
Leader	Veterinary, Hours, Hospital	**0,782**	0,430
DMOZ	Translation	-	-
IHTC	{English,Translations}, {Germans,Translations}, {Frenchs,Translations}	**0,706**	**0,684**
Leader	Translations, Interpreting, English	0,691	0,481
DMOZ	Robotics	-	-
IHTC	{Researchers,Robotics}, {Mobile,Robotics}, {Controls,Robotics}	0,418	0,379
Leader	Robotics, Estimators, Autonomous	**0,524**	**0,486**
DMOZ	Money_Managers	-	-
IHTC	{Financial,Investment}, {Investment,Planning}, {Advisory,Investment}	**0,406**	**0,228**
Leader	Investment, Advisors, Advisory	0,260	0,041
DMOZ	Search_and_Rescue	-	-
IHTC	{Rescue,Searches}, {Searches,Volunteer}, {Searches,Teams}	0,428	**0,737**
Leader	Rescue, Trucks, Volunteer	**0,450**	0,184
DMOZ	Accounting	-	-
IHTC	{Services,Taxes}, {Accounting,Services}, {Accounting,Taxes}	0,593	0,611
Leader	Taxes, Accounting, Cpas	**0,655**	**0,653**

5. Considerações Finais e Trabalhos Futuros

Neste trabalho, foi descrita uma abordagem para construção automática de diretórios web que ataca dois desafios atuais: o agrupamento incremental e a interpretação dos resultados do agrupamento. A abordagem é baseada no método IHTC, desenvolvido para organizar coleções de textos em grupos de documentos com descritores associados.

As avaliações experimentais realizadas indicam que a abordagem proposta é competitiva. O diretório web construído com base no IHTC possui qualidade de agrupamento similar ao Leader. No entanto, o IHTC tem a capacidade de obter melhores descritores para os diretórios. Isto auxilia usuários em tarefas de busca exploratória, o que o torna especialmente útil na construção de diretórios web.

Como trabalho futuro, espera-se realizar uma avaliação mais profunda dos resultados, por meio da participação de usuários, utilização de um maior número de conjuntos de validação e emprego de outros critérios para análise do agrupamento. Ainda, explorar a rede de coocorrência de termos para extração de *n-gramas* na seleção de descritores.

Agradecimentos

Os autores gostariam de agradecer à FAPESP e ao CNPq pelo apoio financeiro.

Referências

Bradley, P. S., Fayyad, U. M., and Reina, C. (1998). Scaling Clustering Algorithms to Large Databases. In *Knowledge Discovery and Data Mining*, pages 9–15.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Farnstrom, F., Lewis, J., and Elkan, C. (2000). Scalability for clustering algorithms revisited. *ACM SIGKDD Explorations Newsletter*, 2:51–57.

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fung, B. C. M., Wang, K., and Ester, M. (2008). *The Encyclopedia of Data Warehousing and Mining*, chapter Hierarchical Document Clustering, pages 970–975. Idea Group.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Kim, H. J. (2006). On text mining algorithms for automated maintenance of hierarchical knowledge directory. In *Knowledge Science, Engineering and Management*, Lecture Notes in Computer Science, pages 202–214.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press.

Marcacini, R. M. and Rezende, S. O. (2010a). Incremental construction of topic hierarchies using hierarchical term clustering. In *SEKE'2010: Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering*, pages 553–558. KSI - Knowledge Systems Institute.

Marcacini, R. M. and Rezende, S. O. (2010b). Torch: a tool for building topic hierarchies from growing text collection. In *WFA'2010: IX Workshop de Ferramentas e Aplicações - XVI Webmedia*, pages 1–3.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of ACM*, 49(4):41–46.

Metwally, A., Agrawal, D., and Abbadi, A. E. (2005). Efficient computation of frequent and top-k elements in data streams. In *ICDT'05: Proceedings of 10th International Conference on Database Theory*, pages 398–412.

Moura, M. F. and Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In *IAI'10: Proceedings of the 10th International Conference on Artificial Intelligence and Applications*, pages 363–371, Acta Press, 2010.

Nassar, S., Sander, J., and Cheng, C. (2004). Incremental and effective data summarization for dynamic hierarchical clustering. In *SIGMOD'04: Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 467–478.

Premalatha, K. and Natarajan, A. (2010). A Literature Review on Document Clustering. *Information Technology Journal*, 9(5):993–1002.

Xu, R. and Wunsch, D. (2008). *Clustering*. Wiley-IEEE Press, IEEE Press Series on Computational Intelligence.

Yang, H. C. and Lee, C. H. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications*, 27(4):645–663.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM '02: Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 515–524.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.