Instituto de Ciências Matemáticas e Computação

ISSN – 0103-2569

Uso da mineração de textos na análise exploratória de artigos científicos

Geraldo Nunes Corrêa

Ricardo Marcondes Marcacini

Solange Oliveira Rezende

RELATÓRIOS TÉCNICOS DO ICMC

Nº XXX

São Carlos

Setembro/2012

Resumo

Neste relatório técnico é apresentada uma descrição do uso de mineração de textos como busca exploratória na fase de levantamento bibliográfico na área de pesquisa médica, mais precisamente sobre câncer de cabeça e pescoço. Entre das diversas técnicas de mineração de textos existentes, foram adotados métodos não supervisionados para extração de hierarquias de tópicos. As hierarquias de tópicos são modelos de dados úteis para organizar e classificar textos de um domínio específico. Para suportar o aprendizado não supervisionado de hierarquias de tópicos a partir dos textos, foi utilizada a ferramenta *TORCH – Topic Hierarchies*. Os experimentos realizados a partir de artigos reais provenientes de bases médicas (*PubMED*), demonstrou que a aplicabilidade de métodos não supervisionados é eficaz para extrair tópicos de níveis mais genéricos. No entanto, para identificar tópicos mais específicos e inovadores aos especialistas de domínios, foi observado que é necessário realizar um pré-processamento dos textos mais apropriado, contando inclusive com a interação de um especialista humano.

Palavras-chave: Mineração de Textos, Extração de Conhecimento, Hierarquias de Tópicos, Agrupamento Hierárquico.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

As hierarquias de tópicos desempenham um papel importante na recuperação e organização de informação, principalmente em tarefas de busca exploratória. Nesse tipo de tarefa, o usuário geralmente tem pouco domínio sobre o tema de interesse, o que dificulta expressar o objetivo diretamente por meio de palavras-chave (Marchionini, 2006).

Assim, torna-se interessante disponibilizar previamente algumas opções para guiar o processo de busca da informação. Para tal, cada grupo possui um conjunto de descritores que contextualizam e indicam o significado dos documentos ali agrupados. Essa organização está relacionada com a hipótese de que se um usuário está interessado em um documento específico pertencente a um determinado tópico, deve também estar interessado em outros documentos desse tópico e de seus subtópicos (Manning et al., 2008).

A Mineração de Textos é um processo que busca descobrir conhecimento útil a partir de coleções textuais, o que viabiliza sobremaneira a análise exploratória de documentos científicos. O processo de Mineração de Textos pode ser dividido em cinco fases principais: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Uso do Conhecimento (Rezende et al., 2003). Esse trabalho explora a fase de Extração de Padrões, no qual métodos de agrupamento de documentos podem ser utilizados para a organização de coleções textuais de maneira não supervisionada.

Em tarefas de agrupamento, o objetivo é organizar um conjunto de objetos em grupos, em que objetos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos objetos de outros grupos. Em outras palavras, o agrupamento é baseado no princípio de maximizar a similaridade intragrupo e minimizar a similaridade intergrupos (Everitt et al., 2001). Os métodos de agrupamento também são conhecidos como algoritmos de aprendizado por observação ou análise exploratória dos dados, pois a organização obtida é realizada por observação de regularidades nos dados, sem uso de conhecimento externo (Xu e Wunsch, 2008).

Os algoritmos de agrupamento podem ser classificados em hierárquicos ou particionais, de acordo com a estratégia de geração dos grupos (Jain et al., 1999). Nos métodos hierárquicos, um conjunto de objetos é organizado em uma hierarquia de grupos e subgrupos enquanto nos métodos particionais os objetos são divididos em uma partição com k grupos, em que o valor de k deve ser definido pelo usuário. Os métodos de agrupamento hierárquico têm sido utilizados para apoiar o aprendizado não supervisionado de hierarquias de tópicos, pois organizam coleções de documentos em grupos e subgrupos, permitindo busca exploratória em diversos níveis de granularidade (Zhao et al., 2005).

Em vista disso, neste trabalho é aplicado um método de agrupamento de documentos, visando o aprendizado não supervisionado de hierarquias de tópicos em coleções textuais envolvendo artigos científicos na área médica com o objetivo de busca exploratória sobre um determinado tema de pesquisa.

O campo de estudo escolhido para a realização deste trabalho foi o Hospital de Câncer – HC - na cidade de Barretos-SP, que realiza mais de 3000 atendimentos diários nos mais diferentes tipos da doença. Além de tornar-se hospital de referência para atendimento pelo Sistema Único de Saúde, o HC de Barretos mantém o Instituto de Ensino e Pesquisa, formado por uma equipe multidisciplinar preparada para oferecer suporte aos colaboradores e alunos de mestrado e doutorado da Fundação Pio XII e que tem por objetivo promover o desenvolvimento da pesquisa científica na instituição. O Instituto possui 24 docentes e mais de 60 alunos de mestrado e doutorado, que atuam nas seguintes linhas de pesquisa.

1. Biologia Tumoral
2. Cuidados Paliativos e Qualidade de Vida
3. Epidemiologia Clinica e Molecular em Oncologia
4. Fatores Ambientais e Câncer
5. Cirurgia Experimental e Minimamente Invasiva

2 Conceitos Básicos Envolvidos

A Mineração de Textos (MT) pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais (Ebecken et al., 2003). Em um contexto no qual grande parte da informação corporativa, como e-mails, memorandos internos e blogs industriais, é registrada em linguagem natural, a MT surge como uma poderosa ferramenta para gestão do conhecimento.

O processo de Mineração de Textos pode ser dividido em cinco grandes etapas, formando um ciclo no qual, ao final, obtém-se o conhecimento acerca dos dados analisados. As etapas são: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (Figura 2.1). Cada etapa pode ser instanciada de acordo com a necessidade dos usuários e de cada aplicação.

Figura 2.1: Etapas do processo de mineração de textos (Rezende et al., 2003)

Para o aprendizado não supervisionado de hierarquias de tópicos, a grande diferença está na etapa de extração de padrões, na qual são utilizados métodos de agrupamento hierárquico que organizam coleções de documentos em grupos e subgrupos. Em seguida, são aplicadas algumas técnicas de seleção de descritores para os agrupamentos formados, ou seja, palavras e expressões que auxiliam a interpretação dos grupos. Após validação dos resultados, o agrupamento hierárquico e seus descritores podem ser

utilizados como uma hierarquia de tópicos para tarefas de análise exploratória dos textos (Cutting et al., 1992; Sanderson e Croft, 1999; Moura e Rezende, 2010), além de apoiar sistemas de recuperação de informação (Zeng et al., 2004; Carpineto et al., 2009).

2.1 Identificação do Problema

Nessa etapa, delimita-se o escopo do problema, definindo o objetivo da aplicação do processo de MT. Basicamente, é necessário selecionar a base de textos com a qual irá trabalhar, o que se espera obter com a análise dos textos, as restrições existentes, a tarefa de extração de padrões apropriados para o problema e como o resultado da análise pode ser utilizado.

No contexto deste trabalho, o objetivo do processo de mineração de textos é extrair uma hierarquia de tópicos a partir de textos de um determinado domínio de conhecimento na área médica. A hierarquia de tópicos obtida pode ser aplicada para auxiliar usuários nas tarefas de organização dos documentos e análise dos tópicos da coleção, além de apoiar a decisão de direcionamento em um projeto de pesquisa.

Um aspecto importante identificado nessa etapa é a restrição ao uso de informação externa, ou seja, o problema analisado não considera o uso de conhecimento de mundo ou especialistas de domínio. Diante disso, uma solução é o uso de aprendizado não supervisionado para extração do conhecimento implícito nos textos. Em especial, os métodos de agrupamento de documentos podem ser utilizados, uma vez que permitem obter uma organização dos textos de forma não supervisionada. Essa organização em grupos de documentos similares possibilita segmentar a coleção em tópicos, obtendo-se resultados de acordo com os objetivos iniciais.

Os resultados obtidos por algoritmos de agrupamento de documentos auxiliam diversas tarefas de organização da informação textual, partindo-se da hipótese que se um usuário está interessado em um documento específico pertencente a um grupo, deve também estar interessado em outros documentos desse grupo (Chakrabarti, 2002; Manning et al., 2008).

Essa hipótese é utilizada em atividades de busca exploratória que ocorrem quando um usuário pode não ter certeza do que ele está procurando até que as opções disponíveis sejam apresentadas e/ou o objetivo não pode ser expresso por palavras-chave, como nos sistemas de busca tradicionais (Marchionini, 2006). Outro aspecto de grande importância é a definição da coleção de documentos a ser utilizada, devendo-se selecionar textos que sejam mais relevantes ao domínio e à aplicação do conhecimento a ser extraído. Essa é uma atividade crítica, uma vez que os textos podem não estar disponíveis no formato adequado, como documentos não-digitalizados.

Ainda, mesmo após a digitalização dos textos, é necessário convertê-los em um padrão de texto puro para que possam ser processados de maneira mais fácil. Assim, é possível iniciar a etapa conhecida como pré-processamento, conforme discutida a seguir.

2.2 Pré-processamento dos Textos

Na etapa de pré-processamento se encontra a principal diferença entre os processos de MT e processos de mineração de dados: a estruturação dos textos em um formato adequado para a extração de conhecimento. Muitos autores consideram essa etapa a que mais tempo consome durante todo o ciclo do processo de MT. O objetivo do pré-processamento é extrair de textos escritos em língua natural, inerentemente não estruturados, uma representação estruturada, concisa e manipulável por algoritmos de agrupamento de documentos.

Para tal, são executadas atividades de tratamento e padronização da coleção de textos, seleção dos termos (palavras) mais significativos e, por fim, representação da coleção textual em um formato estruturado que preserve as características necessárias aos objetivos definidos na etapa de identificação do problema (Feldman e Sanger, 2006).

Os documentos da coleção podem estar em diferentes formatos, uma vez que existem diversos aplicativos para apoiar a geração e publicação de textos eletrônicos. Dependendo de como os documentos foram armazenados ou gerados, há a necessidade de padronizar as formas em que se encontram. Na **padronização dos textos**, geralmente, os documentos são convertidos para a forma de texto plano sem formatação.

Um dos maiores desafios do processo de MT é a alta dimensionalidade dos dados. Uma pequena coleção de textos pode facilmente conter milhares de termos, muitos deles redundantes e desnecessários, que tornam lento o processo de extração de conhecimento e prejudicam a qualidade dos resultados.

A **seleção de termos** tenta solucionar esse desafio e tem o objetivo de obter um subconjunto conciso e representativo de termos da coleção textual. O primeiro passo é a eliminação de *stopwords*, que são os termos que nada acrescentam à representatividade da coleção ou que sozinhos nada significam, como artigos, pronomes e advérbios. O conjunto de *stopwords* é a *stoplist*. Essa eliminação reduz significativamente a quantidade de termos diminuindo o custo computacional das próximas etapas (Manning et al., 2008).

Posteriormente, busca-se identificar as variações morfológicas e termos sinônimos. Para tal, pode-se, por exemplo, reduzir uma palavra à sua raiz por meio de processos de *stemming* ou mesmo usar dicionários ou *thesaurus*. Além disso, é possível buscar na coleção a formação de termos compostos, ou *n-gramas*, que são termos formados por mais de um elemento, porém com um único significado semântico (Manning et al., 2008; Conrado et al., 2009).

Outra forma de realizar a seleção de termos é avaliá-los por medidas estatísticas simples, como a frequência de termo, conhecida como TF (do inglês *term frequency*), e frequência de documentos, conhecida como DF (do inglês *document frequency*). A frequência de termo contabiliza a frequência absoluta de um determinado termo ao longo da coleção textual. A frequência de documentos, por sua vez, contabiliza o número de documentos em que um determinado termo aparece.

O método de Luhn (Luhn, 1958) é uma técnica tradicional para seleção de termos utilizando a medida TF. Nesse método, o autor baseou-se na Lei de Zipf (Zipf, 1932), também conhecida como Princípio do Menor Esforço. Em textos, ao contabilizar a frequência dos termos e ordenar o histograma resultante em ordem decrescente, forma-se a chamada Curva de Zipf, na qual o k-ésimo termo mais comum ocorre com frequência inversamente proporcional a k. Os termos de alta frequência são julgados

não relevantes por geralmente aparecerem na grande maioria dos textos, não trazendo, em geral, informações úteis para discriminar este texto. Já os termos de baixa frequência são considerados muito raros e não possuem caráter discriminatório. Assim, são traçados pontos de corte superior e inferior da Curva de Zipf, de maneira que termos com alta e baixa frequência são descartados, considerando os termos mais significativos os de frequência intermediária (Figura 2.2).

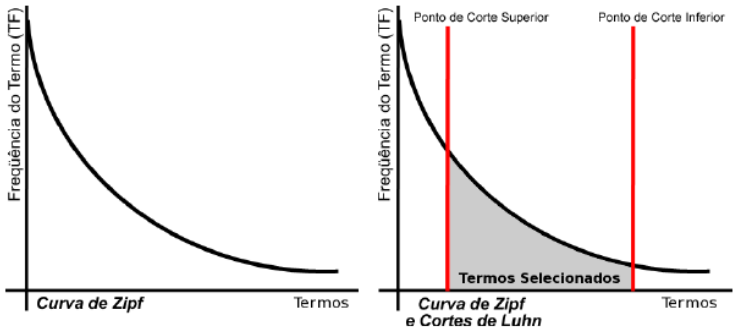

Figura 2.2: Método de Luhn para seleção de termos (adaptado de Soares et al. (2008))

Dado o baixo processamento demandado por esse método, ele é facilmente escalável para coleções textuais muito grandes (Soares et al., 2008). Entretanto, os pontos de corte superior e inferior sugeridos pelo autor não são exatos, sendo a subjetividade da escolha desses pontos a principal desvantagem do método. Uma vez selecionados os termos mais representativos da coleção textual, deve-se buscar a estruturação dos documentos, de maneira a torná-los processáveis pelos algoritmos de agrupamento que são utilizados para apoiar o aprendizado de hierarquias de tópicos.

O modelo mais utilizado para representação de dados textuais é o modelo espaço-vetorial, no qual cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção (Feldman e Sanger, 2006). Para tal, pode-se estruturar os textos em uma *bag-of-words*, na qual os termos são considerados independentes, formando um conjunto desordenado em que a ordem de ocorrência das palavras não importa. A *bag-of-words* é uma tabela documento-termo, como ilustrado na Tabela 2.1 na qual d_i corresponde ao i-ésimo documento, t_j representa o j-ésimo termo e a_{ij} é um valor que relaciona o i-ésimo documento com o j-ésimo termo. Observe que nessa representação não há informação de classe, uma vez que a tarefa de aprendizado com métodos de agrupamento é não supervisionada.

	t_1	t_2	...	t_M
d_1	a_{11}	a_{12}	...	a_{1M}
d_2	a_{21}	a_{22}	...	a_{2M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	a_{N1}	a_{N2}	...	a_{NM}

Tabela 2.1: Tabela documento-termo: representação da matriz atributo-valor

Por meio da tabela documento-termo, cada documento pode ser representado como um vetor $di = (a_{i1}, a_{i2},..., a_{iM})$. Geralmente, o valor da medida a_{ij} é obtido de duas formas:

- um valor que indica se um determinado termo está presente ou não em um dado documento; e
- um valor que indica a importância ou distribuição do termo ao longo da coleção de documentos, por exemplo, o valor de TF. Outras formas, baseadas em critérios de ponderação e normalização, podem ser encontradas em Salton e Buckley (1988) e Liu et al. (2005). Entre elas, destaca-se o critério TF-IDF (*Term Frequency Inverse Document Frequency*), que leva em consideração tanto o valor de TF quanto o valor de DF (Salton et al., 1996).

A representação por meio da tabela documento-termo permite o emprego de um grande leque de algoritmos de agrupamento de documentos, além de outras técnicas de extração de conhecimento. Deve-se ressaltar que essa etapa de Pré-Processamento pode ser redefinida e então repetida após as próximas etapas, uma vez que a descoberta de alguns padrões pode levar a estabelecer melhorias a serem empregadas sobre a tabela documento termo, como, ponderar a importância de cada termo ou até mesmo refinar a seleção dos termos (Rezende et al., 2003).

2.3 Extração de Padrões usando Agrupamento de Documentos

Após a identificação/delimitação do problema e representação dos textos, o processo avança para a etapa de extração de padrões usando agrupamento de documentos.

Em tarefas de agrupamento, o objetivo é organizar um conjunto de objetos em grupos, baseado em uma medida de proximidade, na qual objetos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos objetos de outros grupos

(Everitt et al., 2001). Em outras palavras, o agrupamento é baseado no princípio de maximizar a similaridade interna dos grupos (intragrupo) e minimizar a similaridade entre os grupos (intergrupos) (Everitt et al., 2001). A análise de agrupamento também é conhecida como aprendizado por observação ou análise exploratória dos dados, pois a organização dos objetos em grupos é realizada apenas pela observação de regularidades nos dados, sem uso de conhecimento externo (Xu e Wunsch, 2008). Assim, ao contrário de métodos supervisionados, como algoritmos de classificação, em processos de agrupamento não há classes ou rótulos predefinidos para treinamento de um modelo, ou seja, o aprendizado é realizado de forma não supervisionada (Jain et al., 1999; Han e Kamber, 2006).

O processo de agrupamento depende de dois fatores principais: (1) uma medida de proximidade e (2) uma estratégia de agrupamento. As medidas de proximidade determinam como a similaridade entre dois objetos é calculada. Sua escolha influencia a forma como os grupos são obtidos e depende dos tipos de variáveis ou atributos que representam os objetos. Existe uma variedade de medidas de proximidade e as principais adotadas em dados textuais são discutidas na Seção 2.3.1. As estratégias de agrupamento são os métodos e algoritmos para definição dos grupos (Seção 2.3.2). Em geral, pode-se classificar os algoritmos de agrupamento em métodos particionais e métodos hierárquicos. Por fim, ainda na etapa de extração de padrões, é importante encontrar descritores que indicam o significado do agrupamento obtido para os usuários. Os conjuntos de descritores de cada grupo formam os possíveis tópicos na coleção. Os principais métodos de seleção de descritores são descritos na Seção 2.3.3.

2.3.1 Medidas de Proximidade

A escolha da medida de proximidade para calcular o quão similar são dois objetos é fundamental para a análise de agrupamentos. Essa escolha depende das características do conjunto de dados, principalmente dos tipos e escala dos dados. Assim, existem medidas de proximidade para dados contínuos, discretos e mistura entre dados contínuos e discretos. As medidas de proximidade podem calcular tanto a similaridade quanto dissimilaridade (ou distância) entre objetos. No entanto, as medidas de similaridades podem ser, geralmente, convertidas para medidas de dissimilaridade, e vice-versa.

A seguir, serão descritas duas medidas de similaridade comumente utilizadas em dados textuais: Cosseno e Jaccard. Para tal, considere dois documentos $x_i = (x_{i1}, x_{i2},..., x_{im})$ e $x_j = (x_{j1}, x_{j2},..., x_{jm})$, representados no espaço vetorial m-dimensional, no qual cada termo da coleção representa uma dessas dimensões.

A medida de similaridade Cosseno é definida de acordo com ângulo cosseno formado entre os vetores de dois documentos, conforme a Equação 2.1 (Tan et al., 2005; Feldman e Sanger, 2006).

$$cosseno(x_i, x_j) = \frac{x_i \bullet x_j}{\mid x_i \mid\mid x_j \mid} = \frac{\sum_{l=1}^{m} x_{il} x_{jl}}{\sqrt{\sum_{l=1}^{m} {x_{il}}^2} \sqrt{\sum_{l=1}^{m} {x_{jl}}^2}}$$

O valor da medida está no intervalo [0,1] quando aplicada em dados textuais[1]. Assim, se o valor da medida de similaridade Cosseno é 0, o ângulo entre x_i e x_j é 90º, ou seja, os documentos não compartilham nenhum termo. Por outro lado, se o valor da similaridade for próximo de 1, o ângulo entre x_i e x_j é próximo de 0º, indicando que os documentos compartilham termos e são similares. É importante observar que essa medida não considera a magnitude dos dados para computar a proximidade entre documentos.

Em algumas situações os vetores são representados por valores binários, ou seja, indicam a presença ou ausência de algum termo. O cálculo da proximidade entre dois documentos representados por vetores binários pode ser realizado pela medida Jaccard. Seja x_i e x_j dois documentos, a medida Jaccard pode ser derivada a partir das seguintes contagens:

- f_{11} = número de termos presentes em ambos documentos;
- f_{01} = número de termos ausentes em x_i e presentes em x_j ; e
- f_{10} = número de termos presentes em x_i e ausentes x_j .

[1] A medida de cosseno pode variar no intervalo [1,1] quando são utilizados valores negativos na representação dos atributos. Para dados textuais, geralmente utiliza-se valores baseados em frequência que são maiores ou igual a zero.

A partir das contagens, a medida Jaccard é definida na Equação 2.2 (Tan et al., 2005; Feldman e Sanger, 2006).

$$jaccard(x_i, x_j) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

O valor da medida Jaccard fica no intervalo [0,1]. Quanto mais próximo de 1 maior a similaridade entre os dois documentos. Pode-se observar que as medidas Cosseno e Jaccard são medidas de similaridade. Conforme comentado anteriormente, as medidas de similaridade podem ser transformadas em medidas de dissimilaridade (ou distância).

2.3.2 Métodos de Agrupamento

Após a escolha de uma medida de proximidade para os documentos, é selecionado um método para o agrupamento. Os métodos de agrupamento podem ser classificados considerando diferentes aspectos. Jain et al. (1999) organizam os métodos de agrupamento de acordo com a estratégia adotada para definir os grupos. Uma análise de diferentes métodos de agrupamento considerando o cenário de Mineração de Dados é apresentada em Berkhin (2006).

Em geral, as estratégias de agrupamento podem ser organizadas em dois tipos: agrupamento particional e agrupamento hierárquico. No **agrupamento particional** a coleção de documentos é dividida em uma partição simples de *k* grupos, enquanto no **agrupamento hierárquico** é produzida uma sequência de partições aninhadas, ou seja, a coleção textual é organizada em grupos e subgrupos de documentos (Feldman e Sanger, 2006). Além disso, o agrupamento obtido pode conter sobreposição, isto é, quando um documento pertence a mais de um grupo ou, até mesmo, quando cada documento possui um grau de pertinência associado aos grupos. No contexto deste trabalho, são exploradas as estratégias que produzem agrupamento sem sobreposição, também conhecidas como estratégias rígidas ou *crisp* (Everitt et al., 2001). Assim, se o conjunto $X = \{x_1, x_2, ...; x_n\}$ representa uma coleção de n documentos, uma partição rígida $P = \{G_1, G_2, ..., G_k\}$ com k grupos não sobrepostos é tal que:

- $G_1 \cup G_2 \cup ... \cup G_k = X$;
- $G_i \neq \emptyset$ para todo $i \in \{1, 2, ..., k\}$; e
- $G_i \cap G_j = \emptyset$ para todo $i \neq j$.

As diversas estratégias de agrupamento são, na prática, algoritmos que buscam uma solução aproximada para o problema de agrupamento. Para exemplificar, um algoritmo de força bruta que busca a melhor partição de um conjunto de n documentos em k grupos, precisa avaliar $k^n/k!$ possíveis partições (Liu, 1968). Enumerar e avaliar todas as possíveis partições é inviável computacionalmente. A seguir, são descritos alguns dos principais algoritmos que são utilizados para agrupamento de documentos.

Agrupamento Particional

O agrupamento particional também é conhecido como agrupamento por otimização. O objetivo é dividir iterativamente o conjunto de objetos em k grupos, na qual k geralmente é um valor informado previamente pelo usuário. Os grupos de documentos são formados visando otimizar a compactação e/ou separação do agrupamento.

O algoritmo k-means (MacQueen, 1967) é o representante mais conhecido para agrupamento particional e muito utilizado em coleções textuais (Steinbach et al., 2000). No k-means utiliza-se um representante de grupo denominado centroide, que é simplesmente um vetor médio computado a partir dos demais vetores do grupo. A Equação 2.5 define o cálculo do centroide C para um determinado grupo G, em que x representa um documento pertencente a G e o número total de documentos no grupo é |G|.

$$C = \frac{1}{|G|} \sum_{x \in G} x$$

Dessa forma, o centroide mantém um conjunto de características centrais do grupo, permitindo representar todos os documentos que pertencem a este grupo. Ainda, é importante observar que o k-means só é aplicável em situações na qual a média possa ser calculada. O pseudocódigo para o k-means, contextualizado para agrupamento de documentos, está descrito no Algoritmo 1.

Algoritmo 1: O algoritmo k-means

Entrada:

$X = \{x_1, x_2, ..., x_n\}$: conjunto de documentos

k: número de grupos

Saída:

$P = \{G_1, G_2, ..., G_k\}$: partição com k grupos

```
1 selecionar aleatoriamente k documentos como centroides iniciais;
2 repita
3     para cada documento x ∈ X faça
4         computar a (dis)similaridade de x para cada centroide C ;
5         atribuir x ao centroide mais próximo ;
6     fim
7     recomputar o centroide de cada grupo;
8 até atingir um critério de parada;
```

O critério de parada do k-means é dado quando não ocorre mais alterações no agrupamento, ou seja, a solução converge para uma determinada partição. Outro critério de parada pode ser um número máximo de iterações.

Agrupamento Hierárquico

Os algoritmos de agrupamento hierárquico podem ser aglomerativos ou divisivos. No agrupamento hierárquico aglomerativo, inicialmente cada documento pertence a um grupo e, em cada iteração, os pares de grupos mais próximos são unidos até se formar um único grupo (Feldman e Sanger, 2006). 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 (grupo com apenas um documento) (Steinbach et al., 2000; Zhao et al., 2005).

Tanto os métodos aglomerativos quanto os divisivos organizam os resultados do agrupamento em uma árvore binária conhecida como dendrograma (Figura 2.3). Essa representação é uma forma intuitiva de visualizar e descrever a sequência do agrupamento. Cada nó do dendrograma representa um grupo de documentos. A altura dos arcos que unem dois subgrupos indica o grau de compactação do grupo formado por eles. Quanto menor a altura, mais compactos são os grupos. No entanto, também espera-

se que os grupos formados sejam distantes entre si, ou seja, que a proximidade de objetos em grupos distintos seja a menor possível. Essa característica é representada quando existe uma grande diferença entre a altura de um arco e os arcos formados abaixo dele (Metz, 2006).

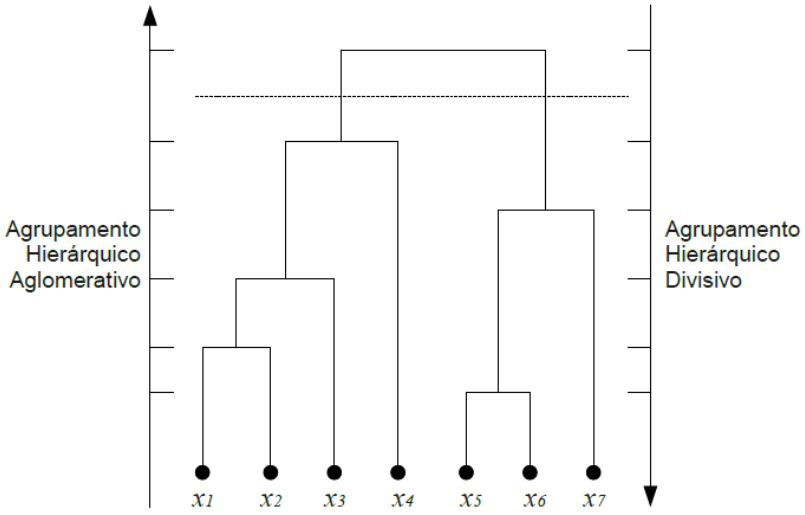

Figura 2.3: Exemplo de um dendrograma (adaptado de Xu e Wunsch (2008))

A partir do dendrograma também é possível obter uma partição com um determinado número de grupos, como nos métodos particionais. Por exemplo, a linha tracejada na Figura 2.3 indica uma partição com dois grupos de documentos: $\{x_1, x_2, x_3, x_4\}$ e $\{x_5; x_6; x_7\}$.

A maioria dos trabalhos relacionados com agrupamento hierárquico na literatura referenciam as estratégias aglomerativas, mostrando pouco interesse nas estratégias divisivas. A possível causa é a complexidade das estratégias divisivas, que cresce exponencialmente em relação ao tamanho do conjunto de dados, proibindo sua aplicação em conjuntos de dados grandes (Kaufman e Rousseeuw, 1990). Para lidar com esse problema, Steinbach et al. (2000) propuseram o algoritmo Bisecting k-means, que basicamente utiliza agrupamento particional baseado no k-means sucessivamente, possibilitando sua aplicação em conjuntos de dados maiores, inclusive em coleções textuais.

2.3.3 Seleção de Descritores para Agrupamento

Uma vez obtido o agrupamento (particional ou hierárquico) de documentos, deve-se selecionar descritores que auxiliam a interpretação dos resultados. Essa é uma tarefa importante, pois o agrupamento geralmente é utilizado em atividades exploratórias para descoberta de conhecimento e, assim, é necessário indicar o significado de cada grupo para que usuários e/ou aplicações possam interagir com o agrupamento de forma mais intuitiva (Manning et al., 2008).

Conforme comentado anteriormente, o centroide mantém o conjunto de características centrais do grupo, permitindo representar todos os documentos pertencentes a este grupo. Por este motivo, algumas técnicas utilizam o centroide como ponto de partida para seleção dos descritores de um grupo. Uma estratégia simplista é selecionar os termos mais frequentes de um grupo, porém, a literatura indica que os resultados obtidos por essa estratégia não são satisfatórios (Chuang e Chien, 2004). Outra estratégia é selecionar os termos dos *j* documentos mais próximos ao centroide como descritores (Cutting et al., 1992). Manning et al. (2008) discute que as técnicas existentes de seleção de atributos em tarefas de aprendizado de máquina podem ser aplicadas na seleção de descritores de agrupamento. Assim, é possível obter um ranking dos termos que melhor discriminam um determinado grupo.

2.4 Pós-processamento

A etapa de pós-processamento é responsável pela validação do conhecimento extraído. A avaliação pode ser realizada de forma subjetiva, utilizando um conhecimento de um especialista de domínio, ou de forma objetiva por meio de índices estatísticos que indicam a qualidade dos resultados. Nesta seção, serão abordados alguns desses índices para validação objetiva.

No contexto deste trabalho, a qualidade da hierarquia de tópicos está diretamente relacionada com a qualidade do agrupamento na extração de padrões. Assim, a validação do conhecimento extraído é realizada por meio de índices utilizados na análise de agrupamentos.

A validação do resultado de um agrupamento, em geral, é realizada por meio de índices estatísticos que expressam o “mérito” das estruturas encontradas, ou seja, quantifica

alguma informação sobre a qualidade de um agrupamento (Faceli et al., 2005; Xu e Wunsch, 2008). O uso de técnicas de validação em resultados de agrupamento é uma atividade importante, uma vez que algoritmos de agrupamento sempre encontram grupos nos dados, independentemente de serem reais ou não (Halkidi et al., 2001). Em geral, existem três tipos de critérios para realizar a validação de um agrupamento: critérios internos, relativos e externos (Everitt et al., 2001).

Os critérios internos obtêm a qualidade de um agrupamento a partir de informações do próprio conjunto de dados. Geralmente, um critério interno analisa se as posições dos objetos em um agrupamento obtido correspondem a matriz de proximidades. Já os critérios relativos comparam diversos agrupamentos para decidir qual deles é o mais adequado aos dados. Finalmente, os critérios externos avaliam um agrupamento de acordo com uma informação externa, geralmente uma intuição do pesquisador sobre a estrutura presente nos dados ou um agrupamento construído por um especialista de domínio. Por exemplo, um critério externo pode medir se o agrupamento obtido corresponde com uma partição dos dados já agrupados manualmente.

Alguns trabalhos na literatura descrevem e comparam técnicas e índices de validação. No trabalho de Milligan e Cooper (1985), trinta índices de validação são comparados na tarefa de estimar o número de grupos em conjuntos de dados. Uma avaliação similar é realizada por Vendramin et al. (2010), com uma comparação de diversos índices de validade relativa de agrupamento. Uma revisão geral de diversas abordagens para validação de agrupamento é encontrada em (Jain e Dubes, 1988; Halkidi et al., 2001; Xu e Wunsch, 2008).

2.5 Uso do Conhecimento

Na etapa de uso do conhecimento, os resultados estão validados e aptos a serem utilizados para apoiar algum processo de tomada de decisão, de acordo com os objetivos estabelecidos na etapa de Identificação do Problema.

Entre as aplicações que se beneficiam das hierarquias de tópicos, destacam-se os trabalhos envolvendo bibliotecas digitais (Krowne e Halbert, 2005; Marcacini et al., 2007; Zhang e Wu, 2008), sistemas de gestão de conhecimento (Bedford, 2008; Zhong

e Liu, 2010), web mining (Song, 2009) e sistemas de recuperação de informação (Zeng et al., 2004; Carpineto et al., 2009).

3 Descrição Mineração de Textos para Apoiar o Levantamento Bibliográfico para Pesquisa Médica no Hospital do Câncer de Barretos

Conforme já mencionado anteriormente, a pesquisa médica no Hospital do Câncer de Barretos é extensa, justificando o uso da Mineração de Textos em atividades exploratórias de diferentes temas de pesquisa.

Dentro deste contexto, uma área de pesquisa específica foi selecionada para aplicar o processo de Mineração de Textos descrito na seção 2. Abaixo está instanciado o processo utilizado nesta primeira atividade com o Hospital do Câncer.

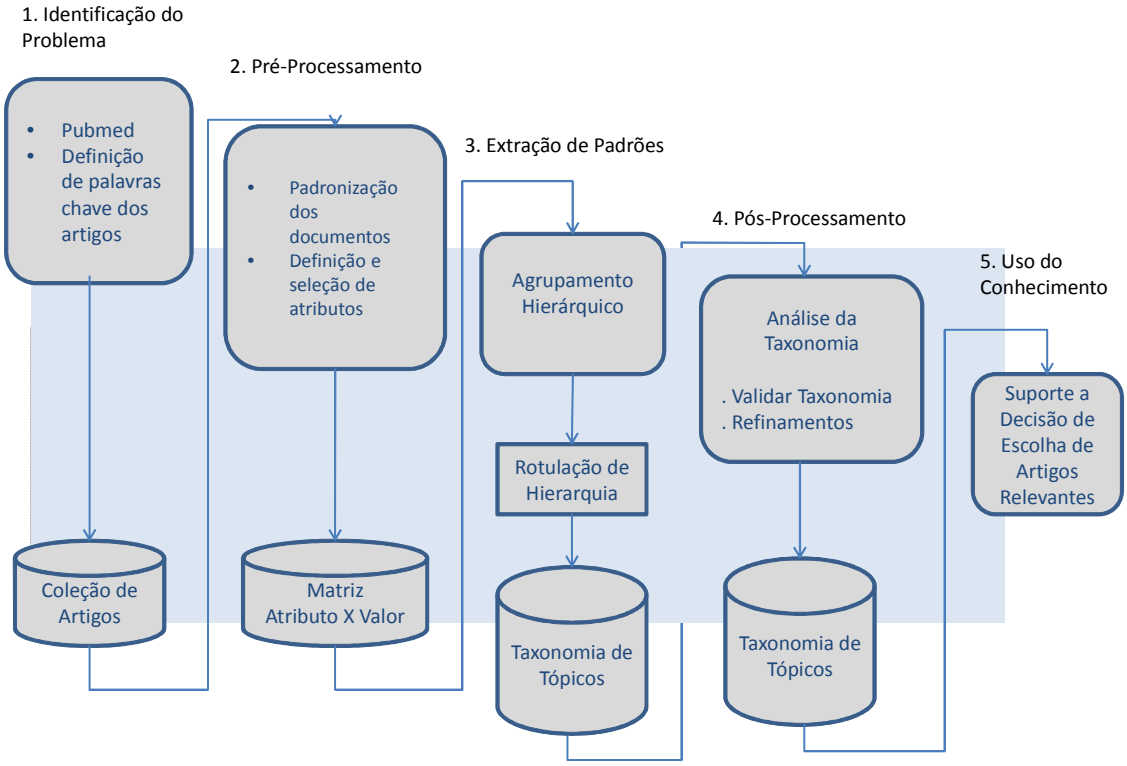

Figura 3.1: : Processo de Mineração de Texto Aplicado no Hospital de Câncer

3.1 Identificação do Problema

Nesta etapa inicial foi definido que o tema de pesquisa a ser explorado no processo de mineração de texto é o câncer de cabeça e pescoço. Dentro deste contexto de pesquisa, existem vários artigos científicos e o problema é a identificação de artigos relevantes ao

pesquisador. Neste sentido, o uso de métodos de agrupamento hierárquico é adequado para a resolução de problemas, uma vez que aborda a busca exploratória de documentos de interesse ao pesquisador.

Conforme orientação do médico pesquisador do Hospital do Câncer, foi indicada uma biblioteca digital relevante dentro da área de pesquisa, a *US National Center for Biotechnology Information*[2]. Dentro desta biblioteca foi utilizada a base de dados Pubmed, que inclui artigos referentes à área médica. Para se ter uma ideia da dimensão da quantidade de artigos, uma simples pesquisa usando as palavras chaves *head and neck cancer* retornou mais 230.000 artigos, conforme pode ser observado na figura seguinte.

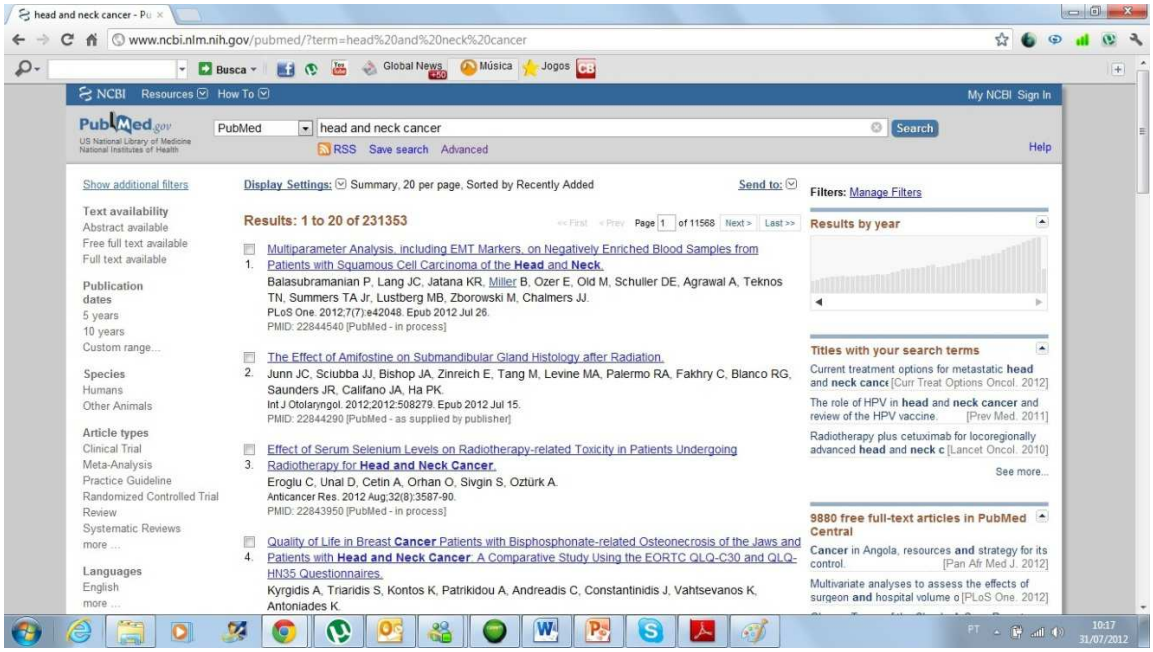

Figura 3.2: Pesquisa inicial de artigos sobre câncer de cabeça e pescoço

No entanto, um recurso de busca avança da biblioteca pôde ser utilizado, reduzindo o número total de artigos de interesse. Outro recurso importante da biblioteca digital Pubmed é o mecanismo de exportação dos artigos. A partir do resultado de uma consulta, tanto o artigo completo como o resumo. Entre os formatos que podem ser escolhidos estão o pdf e xml.

[2] http://www.ncbi.nlm.nih.gov/

3.2 Pré-Processamento

Em posse da coleção de arquivos, seja eles completos ou resumos, o passo seguinte é a **padronização dos textos**, ou seja, os documentos são convertidos para a forma de texto plano sem formatação. Para isso, foram desenvolvidos dois scripts, um para converter arquivos do formato pdf e outro do formato xml.

Em seguida, são realizados os processos de retirada de stopwords e de stemming para viabilizar a geração da matriz atributo X valor. Para a seleção de termos foi utilizada a medida estatística simples de frequência de termo, conhecida como TF (do inglês *term frequency*), usando o método de Luhn.

O código abaixo demonstra o parser dos arquivos xml extraídos da Pubmed.

```
<?php
$xml = simplexml_load_file("pubmed_result.xml");foreach($xml->PubmedArticle as
$PubmedArticle){
     $titulo = $PubmedArticle->MedlineCitation->Article->ArticleTitle;
     $id = $PubmedArticle->MedlineCitation->PMID;
     $abstract = $PubmedArticle->MedlineCitation->Article->AbstractText;
     $dia = $PubmedArticle->MedlineCitation->Article->ArticleDate->Day;
     $mes = $PubmedArticle->MedlineCitation->Article->ArticleDate->Month;
     $ano = $PubmedArticle->MedlineCitation->Article->ArticleDate->Year;
     $data = $ano.$mes.$dia;
     if(strlen($titulo) > 20 && strlen($abstract) > 100 && strlen($id) == 8 &&
strlen($data)==8){
         echo "Title: ".$titulo."\n";
         echo "Abstract: ".$abstract."\n";
         echo "ID: ".$id."\n";
         echo "Date: ".$data."\n";
     }
}
?>
```

3.3 Extração de Padrões

Nesta etapa, o objetivo é organizar o conjunto de artigos científicos em grupos, baseado em uma medida de proximidade, na qual artigos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos artigos de outros grupos. Ainda nesta etapa, a análise de agrupamento também é conhecida como aprendizado por observação ou análise exploratória dos dados, pois a organização dos objetos em grupos é realizada apenas pela observação de regularidades nos dados, sem uso de conhecimento externo, ou seja, não supervisionada.

A medida de proximidade para o processo de agrupamento foi a do cosseno e a estratégia de agrupamento foi o agrupamento hierárquico. Ambos os conceitos foram descritos na seção 2.

Para selecionar descritores que auxiliam a interpretação dos resultados a estratégia adotada foi a de selecionar os termos mais dos documentos mais próximos ao centroide como descritores, gerando um ranking dos termos que melhor discriminam um determinado grupo de artigos.

Esta etapa de extração de padrões é apoiada por uma ferramenta desenvolvida no Labic, TORCH (Topic Hierarchy)[3], que realiza a tarefa de agrupamento utilizando os conceitos citados.

3.4 Pós-processamento

A avaliação do conhecimento extraído é realizada de forma subjetiva, utilizando o conhecimento do pesquisador. No contexto deste trabalho, a qualidade da hierarquia de tópicos está diretamente relacionada com a qualidade do agrupamento na extração de padrões. Neste trabalho não são utilizados índices estatísticos para expressar o "mérito" das estruturas encontradas, ou seja, para quantificar alguma informação sobre a qualidade de um agrupamento. Tal tarefa é realizada pelo especialista de domínio.

3.5 Uso do Conhecimento

Na etapa de uso do conhecimento, os resultados são validados pelo médico especialista tornam-se aptos a serem utilizados para apoiar a decisão de escolha dos artigos a serem utilizados na pesquisa científica, conforme os objetivos estabelecidos na etapa de Identificação do Problema.

[3] http://sites.labic.icmc.usp.br/torch

4 Resultados e Discussão

O processo descrito na seção 3 foi aplicado a uma determinada área de pesquisa dentro do Hospital do Câncer, que é o câncer de cabeça e pescoço. Foi utilizada base de a dados da Pubmed e foram coletados 91 artigos completos no formato pdf.

Na sequencia, esta coleção de artigos foram transformados para o formato txt, sendo que destes 5 não foram convertidos por problemas do OCR[4]. Desta forma 84 artigos representaram a base inicial de documentos para a etapa de pré-processamento. Na tabela 1 são apresentados mais detalhes da etapa de pré-processamento.

Tabela 4.1: Características Gerais da Base de Textos

Documentos:	84
Termos:	629
Média Termo/Doc:	7.49

Tendo como base esta coleção de documento, foi utilizada a ferramenta TORCH para realização da etapa de pré-processamento e geração da hierarquia de tópicos. A ferramenta foi configurada para a apresentação de 7 níveis, sendo cada nível rotulado por 3 descritores conforme algoritmo para esta finalidade.

Como pode ser observado na Figura 4.1, a rotulação dos 7 primeiros níveis ficou da seguinte maneira:

1. Cells, Cancers, Expressed;
2. Patients, Treatments, Carcinomas;
3. Thyroids, Patients, Cancers;
4. Tumors, Rnas, Cancers;
5. Cancers, Patients, Necks;
6. Patients, Cancers, Survive;

[4] *Optical Character Recognition*

7. Cancers, Patients, Treaments;

Observando a rotulação de cada nível, percebe-se a repetição de termos em diferentes níveis. Além disso, dois níveis, 2 e 7, podem ser considerados iguais, uma vez que o termo Carcinomas é sinônimo do termo Cancers.

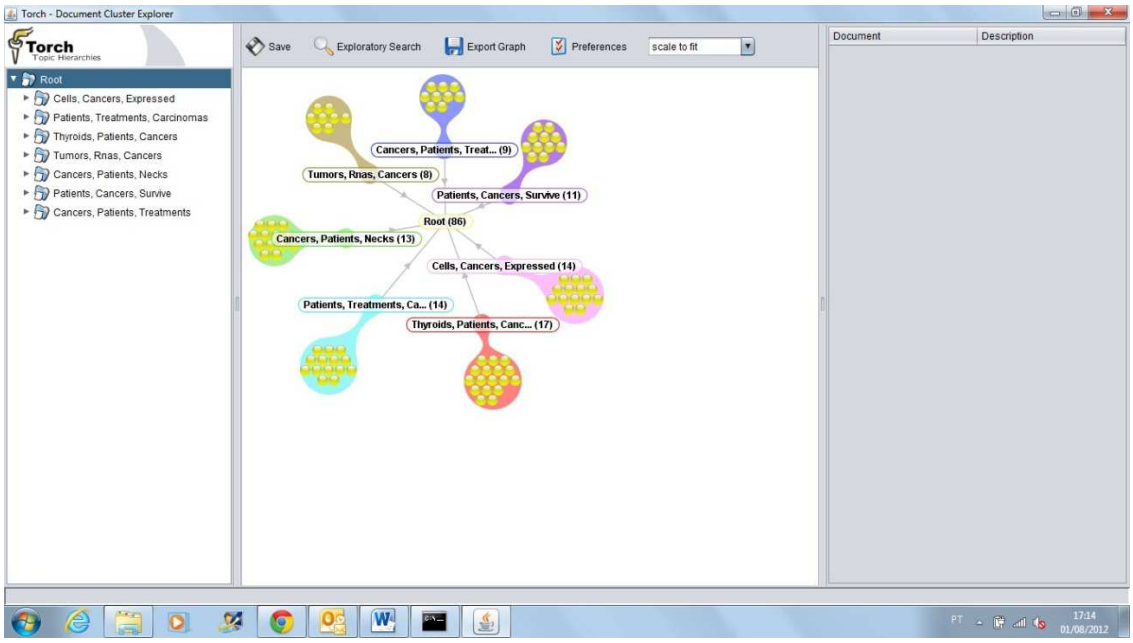

Figura 4.1: Geração do agrupamento dos artigos completos em 7 níveis

Partindo da identificação deste problema na geração do agrupamento, a ferramenta foi utilizada novamente utilizando-se diferentes configurações de níveis para verificar o agrupamento gerado. A seguir é apresentada a tela de configuração da ferramenta TORCH.

Figura 4.2: Tela de configuração da Ferramenta TORCH

Na tela acima são configurados a pasta com a coleção dos artigos (documentos), a remoção de stopwords em inglês, stemming e o algoritmo de agrupamento x-secting-means. Na configuração do algoritmo é possível indicar a quantidade de níveis (branch fator) para a geração do agrupamento, conforme a tela abaixo.

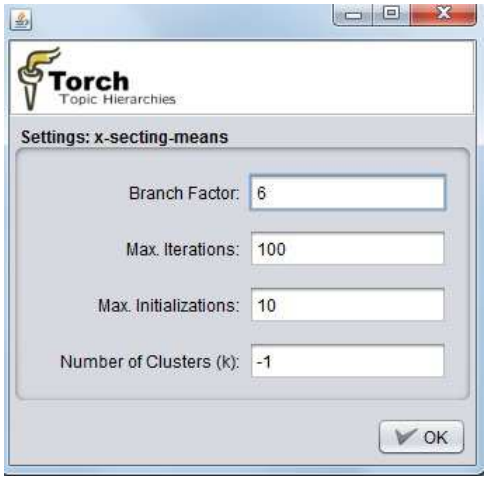

Figura 4.3: Tela de configuração dos níveis de agrupamentos

Desta maneira foram gerados os gráficos de 4, 5 e 6 níveis para analisar a qualidade do agrupamento antes de uma reunião com o médico pesquisador. As telas são apresentadas a seguir.

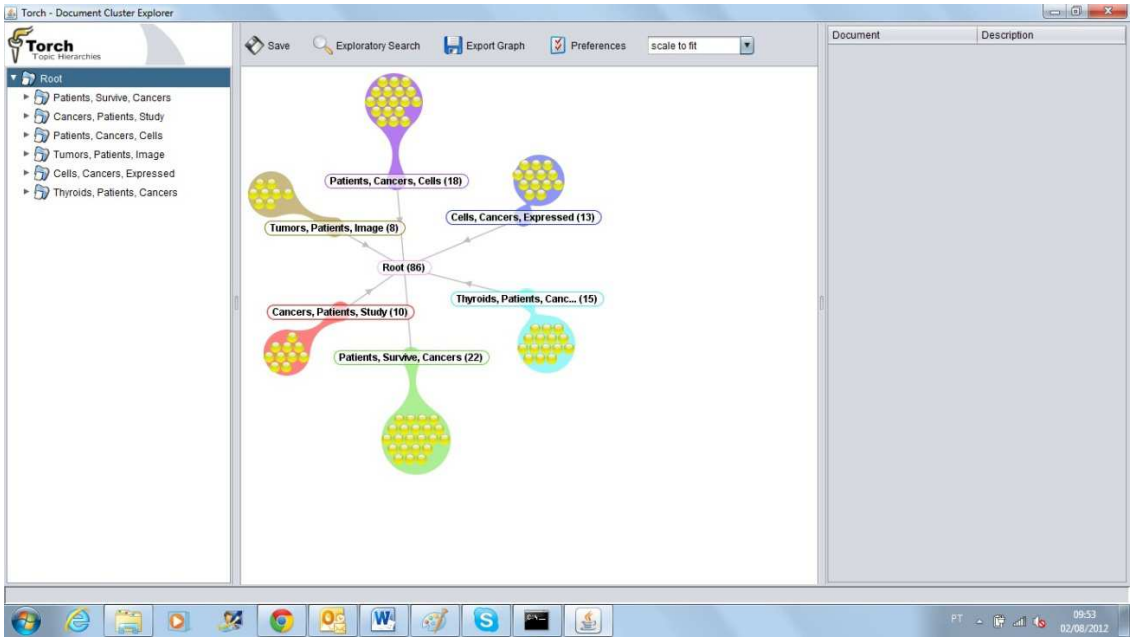

Figura 4.4: Geração do agrupamento dos artigos completos em 6 níveis

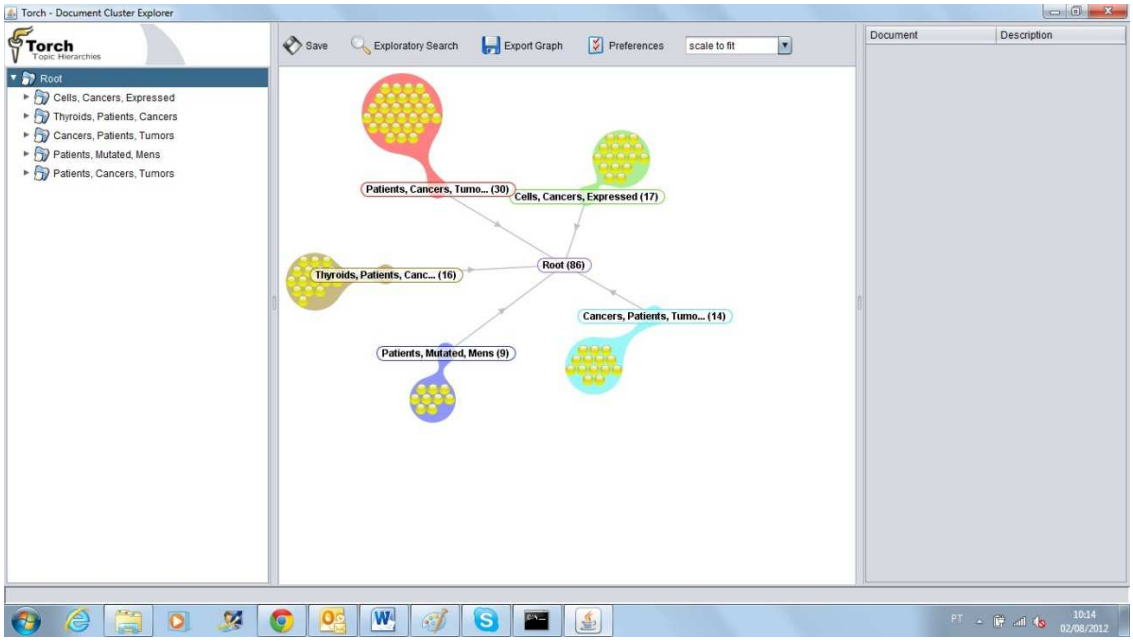

Figura 4.5: Geração do agrupamento dos artigos completos em 5 níveis

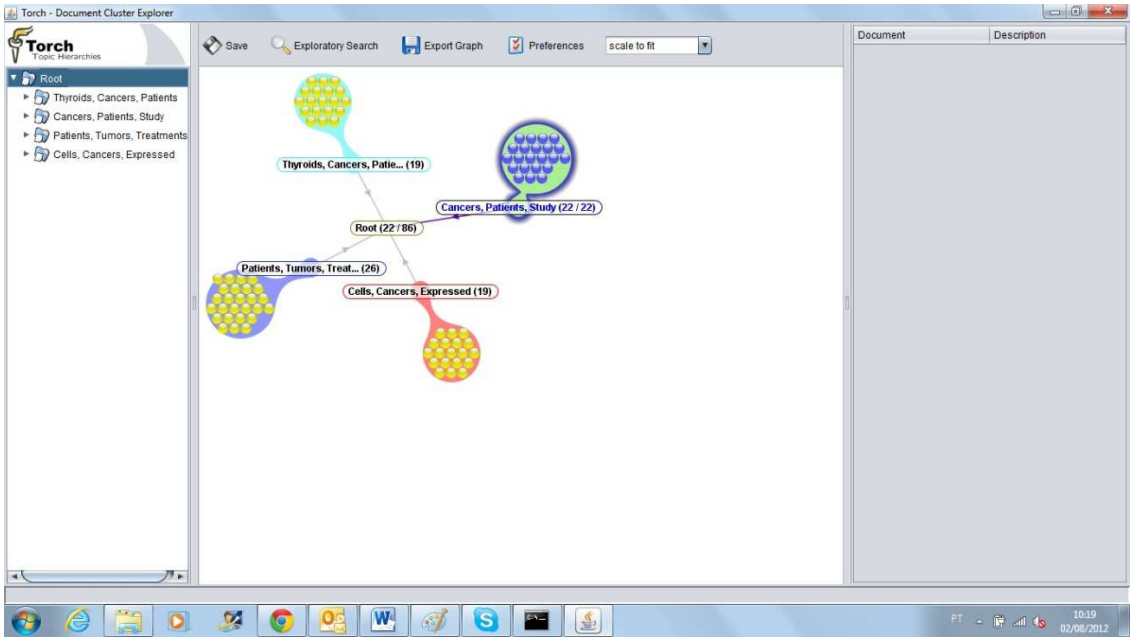

Figura 4.6: : Geração do agrupamento dos artigos completos em 4 níveis

Após a geração dos agrupamentos com diferentes níveis foi realizada uma reunião com o médico pesquisador. A questão da rotulação dos níveis foi discutida e um outro problema foi identificado. Como o pré-processamento foi realizado em artigos completos, um componente é comum a todos os documentos é o tópico de introdução. Em tal tópico sempre existem palavras (termos) que são comuns em vários artigos, possuindo, assim, uma alta frequência, gerando resultados distorcidos na etapa de pré-processamento.

Desta forma, outra estratégia foi adotada para a busca exploratória na área de pesquisa selecionada neste projeto. Ao invés de artigos completos, o médico pesquisador solicitou a realização da tarefa de pré-processamento sobre os abstracts dos artigos, nos quais estão presentes as palavras (termos) que representam a essência da proposta do autor.

Na sequencia, foram gerados em formato xml os abstracts dos mesmos artigos selecionados na estratégia anterior. O resultado alcançado a partir de tal coleção de documentos é apresentado na figura seguinte.

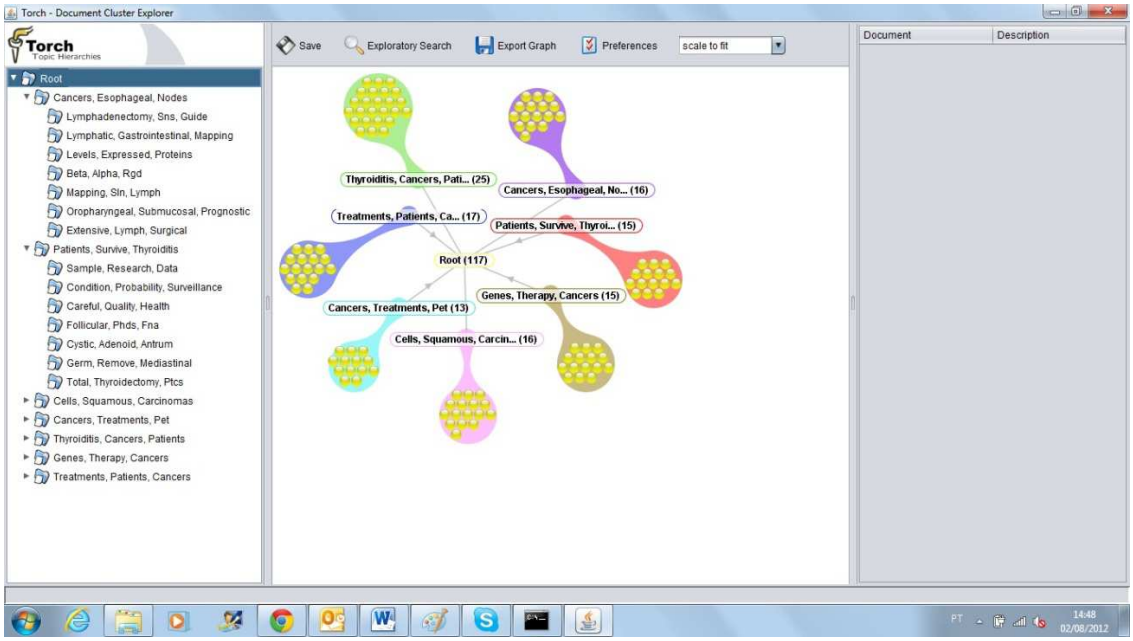

Figura 4.7: Geração do agrupamento a partir dos resumos dos artigos

A rotulação dos níveis se alterou comparado com a estratégia de se utilizar os artigos completos como coleção inicial de documentos. No entanto, em outra reunião com o médico pesquisador foi observado que:

1. A rotulação dos níveis ainda se apresentava com problemas para a identificação dos grupos de artigos (documentos);

2. Analisando os artigos presentes em cada grupo, o médico pesquisador informou que esperava uma outra forma de agrupamento, ou seja, a similaridade entre os artigos não foi perfeitamente ajustada pelo algoritmo de agrupamento;

5 Conclusões e Trabalhos Futuros

Em um cenário em que grande parte das informações do mundo digital são armazenadas em formato textual, é de extrema importância investigar técnicas computacionais que permitem extrair conhecimento útil desses textos de forma automática. Entre as diversas técnicas existentes, a extração e análise do conhecimento por meio de hierarquias de tópicos foram adotadas neste trabalho. Hierarquias de tópicos são modelos de dados eficientes para descrever e categorizar documentos textuais.

No decorrer do desenvolvimento deste trabalho, um dos objetivos principais foi a extração de hierarquias de tópicos a partir bases de artigos médicos sobre câncer. As

bases de artigos científicos disponíveis na PubMed representam a evolução do conhecimento da área ao longo do tempo. Ainda, a quantidade de informação publicada na PubMed excede a capacidade humana de analisá-la manualmente, incentivando o uso de técnicas de mineração de textos.

A partir da análise das hierarquias de tópicos extraídas da base de artigos coletadas, é possível concluir que métodos não supervisionados foram eficazes para extrair tópicos mais genéricos sobre os dados (níveis mais altos da hierarquia). Estes tópicos são úteis para realizar uma primeira análise exploratória com usuários que não possuem conhecimento aprofundado sobre o assunto descrito nos artigos. Por outro lado, os tópicos mais genéricos não representam conhecimento inovador para usuários especialistas do domínio.

Assim, um ponto importante no método de extração de hierarquias de tópicos é determinar o conjunto de atributo (palavras-chave) dos textos para formação dos tópicos. Métodos não supervisionados tendem a selecionar os termos mais frequentes dos textos, o que a leva a formação de tópicos mais genéricos. A inclusão de um dicionário ou ontologia de domínio para apoiar a seleção de termos mais específicos, bem como técnicas de aprendizado ativo que permitam a inclusão de especialistas de domínio na extração de tópicos, é uma direção de pesquisa promissora e será abordada em trabalhos futuros.

Referências Bibliográficas

Bedford, D. A. D. (2008). Knowledge management in practice: connections and context. American Society for Information Science and Technology.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Kogan, J., Nicholas, C., e Teboulle, M., editors, Grouping Multidimensional Data, chapter 2, páginas 25-71. Springer-Verlag, Berlin, Heidelberg.

Carpineto, C., Osinski, S., Romano, G., e Weiss, D. (2009). A survey of web clustering engines. ACM Computing Surveys, 41:1-17.

Chakrabarti, S. (2002). Mining the web: discovering knowledge from hypertext data. Science & Technology Books.

Chuang, S.-L. e Chien, L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. In CIKM '04: Proceedings of the thirteenth ACM

International Conference on Information and Knowledge Management, páginas 127-136, New York, NY, USA. ACM.

Conrado, M. S., Marcacini, R. M., Moura, M. F., e Rezende, S. O. (2009). O efeito do uso de diferentes formas de geração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa. In WTI'09: II International Workshop on Web and Text Intelligence, páginas 1-10.

Cutting, D. R., Pedersen, J. O., Karger, D., e Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In SIGIR'92: Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval, páginas 318-329.

Ebecken, N. F. F., Lopes, M. C. S., e de Aragão Costa, M. C. (2003). Mineração de textos. In Rezende, S. O., editor, Sistemas Inteligentes: Fundamentos e Aplicações, chapter 13, páginas 337_370. Manole, 1ª edição.

Everitt, B. S., Landau, S., e Leese, M. (2001). Cluster Analysis. Arnold Publishers.

Faceli, K., Carvalho, A. C. P. L. F., e Souto, M. C. P. (2005). Validação de algoritmos de agrupamento. Relatório Técnico 254, Instituto de Ciências Matemáticas e de Computação - ICMC - USP.

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

Halkidi, M., Batistakis, Y., e Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3):107-145.

Han, J. e Kamber, M. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann, 2ndª edição.

Jain, A. K. e Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Upper Saddle River, NJ, USA.

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

Kaufman, L. e Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Interscience, New York.

Krowne, A. e Halbert, M. (2005). An initial evaluation of automated organization for digital library browsing. In JCDL'05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, páginas 246-255, New York, NY, USA. ACM.

Liu, L., Kang, J., Yu, J., e Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. In NLP-KE '05. Proceedings of 2005 International Conference on Natural Language Processing and Knowledge Engineering, páginas 597-601.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal os Research and Development, 2(2):159-165.

MacQueen, J. B. (1967). Some methods for classi_cation and analysis of multivariate observations. In Cam, L. M. L. e Neyman, J., editors, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, páginas 281-297. University of California Press.

Marcacini, R. M., Moura, M. F., e Rezende, S. O. (2007). Biblioteca Digital do IFM: uma Aplicação para a Organização da Informação por meio de Agrupamentos Hierárquicos. In WDL'07: III Workshop on Digital Libraries do Simpósio Brasileiro de Sistemas Multim ídia e Web (Webmedia), páginas 1-10, Gramado - RS. SBC.

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

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

Metz, J. (2006). Interpretação de clusters gerados por algoritmos de clustering hierárquico. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação – ICMC - Universidade de São Paulo - USP.

Milligan, G. e Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2):159-179.

Moura, M. F. e Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In IAI'10: Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications, páginas 363-371, Anaheim, Calgary, Zurich : Acta Press, 2010.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., e Paula, M. F. (2003). Mineração de dados. In Rezende, S. O., editor, Sistemas Inteligentes: Fundamentos e Aplicações, chapter 12, páginas 307_335. Manole, 1ª edição.

Salton, G., Allan, J., e Singhal, A. (1996). Automatic text decomposition and structuring. Information Processing & Management, 32(2):127_138.

Sanderson, M. e Croft, B. (1999). Deriving concept hierarchies from text. In SIGIR'99: Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, páginas 206-213, New York, NY, USA. ACM.

Salton, G. e Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. An International Journal of Information Processing and Management, 24(5):513-523.

Soares, M. V. B., Prati, R. C., e Monard, M. C. (2008). Pretext ii: Descrição da reestrutura ção da ferramenta de pré-processamento de textos. Relatório Técnico 333, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.

Song, M. (2009). Handbook of Research on Text and Web Mining Technologies. Information Science Reference.

Steinbach, M., Karypis, G., e Kumar, V. (2000). A comparison of document clustering techniques. In KDD'2000: Workshop on Text Mining, páginas 1-20.

Tan, P.-N., Steinbach, M., e Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley Longman Publishing, Boston, MA, USA.

Vendramin, L., Campello, R. J. G. B., e Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining, 3(4):209-235.

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

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

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., e Ma, J. (2004). Learning to cluster web search results. In SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval , páginas 210-217, New York, NY, USA. ACM.

Zhang, C. e Wu, D. (2008). Concept extraction and clustering for topic digital library construction. In WI-IAT'08: Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology, páginas 299-302, Los Alamitos, CA, USA. IEEE Computer Society.

Zhong, J. e Liu, J. (2010). Automatic construction of knowledge tree based on text clustering. Application Research of Computers, 27:475-478.

Zipf, G. K. (1932). Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press.