Refinamento interativo de mapas de documentos apoiado por extração de tópicos

Renato Rodrigues Oliveira da Silva

Refinamento interativo de mapas de documentos apoiado por extração de tópicos

Renato Rodrigues Oliveira da Silva

Orientador: ***Profa. Dra. Maria Cristina Ferreira de Oliveira***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. .

USP – São Carlos
Outubro/2010

Agradecimentos

Agradeço à professora Maria Cristina Ferreira de Oliveira pela excelente orientação, paciência e apoio na realização deste projeto. Espero seguir seu exemplo de dedicação, responsabilidade e comprometimento.

À professora Rosane Minghim pelos conselhos, apoio e amizade. Sua ajuda foi fundamental para prosseguir meus estudos.

Aos pesquisadores Alípio Jorge, Alneu Lopes, Fernando Paulovich, Luciana Nedel e Roberto Pinho pelas sugestões e ajuda. A colaboração de vocês enriqueceu muito este trabalho.

Aos amigos do laboratório LCAD pela colaboração direta e indireta para a conclusão deste projeto. Agradeço também aos amigos que fiz no ICMC, em especial ao Ricardo Cerri, Valter Messias, Tácito Trindade, Paulo Gabriel e muitos outros pelo compaheirismo e troca de idéias.

Aos colegas que ajudaram na avaliação de usabilidade: Francisco Souza, Frizzi Alejandra, Hitoshi Mizobuchi, Paulo Gabriel, Ricardo Cerri, Rodrigo Córdoba e Tácito Trindade.

Ao Instituto de Ciências Matemáticas e de Computação (ICMC-USP) pela oportunidade de realizar o curso de mestrado. Aos professores pelos ensinamentos desde minha graduação até os dias atuais, e aos funcionários pelos excelentes serviços prestados. Gostaria de destacar as funcionárias Lhaís Visentin, Ana Paula e Laura Aparecida pela presteza, paciência e dedicação em sanar minhas incontáveis dúvidas.

Aos amigos de república Francisco Souza, Rafael Beraldo e Rodrigo Córdoba.

À Talita Justel, por todo apoio e carinho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro concedido que viabilizou a realização deste trabalho, processo 132099/2008-0.

À minha família, pelo apoio ao longo de toda minha vida.

Resumo

Silva, R. R. O. **Refinamento Interativo de Mapas de Documentos apoiado por Extração de Tópicos**. 2010. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010.

Mapas de documentos são representações visuais que permitem analisar de forma eficiente diversas relações entre documentos de uma coleção. Técnicas de projeção multidimensional podem ser empregadas para criar mapas que refletem a similaridade de conteúdo, favorecendo a identificação de agrupamentos com conteúdo similar. Este trabalho aborda uma evolução do arcabouço genérico oferecido pelas projeções multidimensionais para apoiar a análise interativa de documentos textuais, implementado na plataforma PEx. Foram propostas e implementadas técnicas que permitem ao usuário interagir com o mapa de documentos utilizando tópicos extraídos do próprio corpus. Assim a representação visual pode gradualmente evoluir para refletir melhor os interesses do usuário, e apoiá-lo de maneira mais efetiva em tarefas exploratórias. A interação foi avaliada utilizando uma técnica de inspeção de usabilidade, que visa identificar os principais problemas enfrentados pelos usuários ao interagir com as funcionalidades desenvolvidas. Adicionalmente, a utilidade das funcionalidades foi avaliada pela condução de dois estudos de caso, em que foram definidas tarefas a serem conduzidas pelo usuário sobre os mapas de documentos. Os resultados mostram que com o auxílio das visualizações foi possível conduzir as tarefas satisfatoriamente, permitindo manipular de forma eficiente milhares de documentos sem a necessidade de ler individualmente cada texto.

Palavras-chave: Mineração visual de textos, análise de textos, visualização de documentos, extração de tópicos.

Abstract

Silva, R. R. O. **Interactive Refinement of Document Maps supported by Topic Extraction**. 2010. Dissertation (Master) – Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, 2010.

Content-based document maps are visualizations that help users to identify and explore relationships among documents in a collection. Multidimensional projection techniques have been employed to create similarity-based maps that can help identifying documents of similar content. This work aims to enhance the generic framework offered by the multidimensional projection techniques in the PEx visualization platform to support interactive analysis of textual data. Several interaction functions and visual representations have been proposed and implemented that allow users to interact with document maps aided by topics automatically extracted from the corpus. By exploring the topics and maps in an integrated manner, users can refine and evolve the visual representations gradually to better reflect their needs and interests, enhancing support to exploratory tasks. The proposed interaction functions were evaluated employing a usability inspection technique, seeking to detect interface problems. Moreover, two illustrative case studies were conducted to evaluate the usefulness of the proposed interactions, based on typical user tasks defined over different document collections. They illustrate how the developed visualizations can assist the proposed tasks, allowing users to interactively explore large document corpora and refine document maps.

Keywords: Visual text mining, text analytics, document visualization, topic extraction.

Lista de Figuras

Lista de Tabelas

Sumário

CAPÍTULO

1

Introdução

1.1 Contexto e Motivação

As dificuldades associadas à análise efetiva de dados – hoje amplamente disponíveis, caracterizados por enorme heterogeneidade e grandes volumes – permeiam diferentes disciplinas e domínios, tanto no cenário da pesquisa acadêmica como nos setores de serviços, industrial e agro-pastoril. A gestão da informação em grandes volumes de dados multimídia distribuídos integra os Grandes Desafios de pesquisa em Computação identificados no documento da Sociedade Brasileira de Computação[1] sendo que a obtenção de informações úteis contidas em grandes massas de dados pode ser vista como um dos problemas associados à sua gestão.

A existência de ferramentas de análise mais efetivas tem um papel fundamental nesse contexto, uma vez que é preciso ampliar a capacidade cognitiva do ser humano em processos de exploração de dados, favorecendo o acesso à informação. Essa é, justamente, a proposta da Visualização Computacional, disciplina que surgiu a partir da disseminação dos recursos computacionais capazes de gerar gráficos interativos (Card *et al.*, 1999; Chen, 2004; Oliveira e Levkowitz, 2003). Pesquisadores do grupo de Visualização, Imagens e Computação Gráfica do ICMC-USP têm atuado, já há algum tempo, nas áreas de Visualização Científica (Berger *et al.*, 2010; Cuadros-Vargas *et al.*, 2009; Ferreira *et al.*, 2009; Siqueira *et al.*, 2009) e Visualização de Informação (Eler *et al.*, 2009a; Moraes *et al.*, 2010; Pinho *et al.*, 2010, 2009).

[1] Grandes Desafios da Computação - http://www.sbc.org.br

Visualização Científica é a disciplina que objetiva criar representações visuais a partir de dados que possuem uma informação geométrica inerente, por exemplo, dados provenientes de aparelhos de ressonância magnética (representando o corpo humano), dados provenientes de um satélite (relativos a tempestades ou correntes marítimas). A Visualização de Informação, por sua vez, objetiva criar representações visuais de dados que não possuem essa característica geométrica, como dados do censo, o relacionamento entre tópicos de pesquisas de uma área científica, etc. Portanto, uma diferença fundamental entre as visualizações científica e de informação é a arbitrariedade na escolha de uma representação visual significativa para os dados analisados.

A representação visual proporciona uma visão ampla das relações existentes nos dados, facilitando a identificação de padrões, a detecção de anomalias, discrepâncias e obtenção de *insight*. *Insight* pode ser caracterizado como descobertas inesperadas, um novo modo de entender o objeto de estudo, ou ainda um conhecimento mais profundo sobre comportamentos e relações analisados nos dados. Este ponto é particularmente explorado na área conhecida como *Visual Analytics* (Análise Visual) que integra as áreas de visualização de informação, análise estatística, ciência cognitiva, entre outras (Wong e Thomas, 2004).

Grande parte das informações disponíveis atualmente é representada na forma de documentos. Jornais, revistas, páginas da internet e mensagens eletrônicas são exemplos de meios que utilizam textos para transmitir informações a seus leitores. Com a popularização da internet, um espaço fértil para a disseminação de idéias é traduzido na criação de incontáveis páginas eletrônicas, grupos de discussão, fóruns, revistas eletrônicas, bibliotecas digitais, entre outros. No entanto, essa enorme disponibilidade de informação traz à tona os seguintes problemas:

- Como um leitor será capaz de distinguir qual informação lhe é realmente útil?
- Como tratar a grande heterogeneidade de informação?
- De que forma procurar os documentos que realmente lhe interessam, em tempo hábil?

Naturalmente não é possível responder a essas questões por meio da análise individual de cada documento. É necessário criar ferramentas que reduzam a complexidade de análise, capazes de propor abstrações para ocultar detalhes não significativos e apresentar inicialmente apenas informações gerais. As ferramentas não objetivam substituir a leitura dos documentos, mas fornecer uma visão ampla das relações existentes em uma coleção, visando facilitar o seu entendimento e diminuir a carga cognitiva sobre o leitor. Dessa forma, o leitor pode dedicar mais tempo para analisar cuidadodamente os documentos que considera relevantes.

O grupo de pesquisa tem se concentrado no desenvolvimento de técnicas de visualização que permitem criar mapas visuais de conjuntos de dados multidimensionais em geral (Eler *et al.*, 2009a; Moraes *et al.*, 2010), e de coleções de documentos textuais em particular (Eler *et al.*, 2009b; Felizardo *et al.*, 2009; Paulovich *et al.*, 2008). Esses mapas são representações visuais criadas utilizando diferentes técnicas de projeção multidimensional.

As projeções geram uma imagem 2D dos documentos que reflete sua similaridade de conteúdo, e os mapas obtidos podem ser explorados por usuários em busca de grupos de documentos de interesse. Com isso, um usuário pode explorar uma grande coleção de documentos sem necessidade de percorrer – e ler – sequencialmente todos os documentos em busca dos que realmente lhe interessam. Trabalhos anteriores do grupo de pesquisa (Lopes *et al.*, 2007; Paulovich, 2008; Pinho *et al.*, em preparação) abordam, também, técnicas capazes de extrair tópicos de documentos de forma automática. Esses tópicos são formados por termos (palavras) representativos que, de certa forma, identificam os assuntos abordados pelo conjunto de documentos. Os tópicos são úteis para rotular os agrupamentos do mapa de documentos, e dessa forma fornecerem informações sobre o seu conteúdo.

Os mapas são criados considerando um conjunto de documentos fornecidos como entrada. No entanto, um pesquisador pode, após uma exploração inicial, decidir que está interessado em apenas um sub-conjunto do corpus, tornando os documentos restantes, portanto, um "ruído" na representação gráfica. A interferência do usuário no processo de refinamento do mapa é um cenário ainda pouco explorado.

1.2 Objetivos e Justificativa

Este trabalho aborda, especificamente, uma evolução do arcabouço genérico oferecido pelas projeções multidimensionais para apoiar a análise interativa de documentos textuais. Pretende-se investigar, propor e validar técnicas e estratégias que permitam ao usuário interagir com os mapas de documentos, utilizando informações adicionais extraídas dos próprios documentos para refinar os mapas gerados, de modo que o mapa possa gradualmente evoluir, com a interferência do usuário, para refletir mais precisamente o seu foco e os seus interesses.

Como ponto de partida, foram utilizados resultados anteriores em extração automática de tópicos em mapas de documentos, obtidos em um projeto de doutorado (Pinho, 2009). Os tópicos extraídos fornecem pistas para identificar o conteúdo dos documentos que compõem o mapa. Cada tópico cobre um conjunto específico de documentos, rotulando uma região delimitada pelo usuário no mapa. Neste trabalho foram propostas e implementadas diversas funcionalidades de interação que atuam sobre estes tópicos e seus documentos correspondentes. Essas funciona-

lidades permitem que o usuário identifique relações entre os tópicos, de forma que tenha mais informações para julgar quais documentos são pertinentes aos focos de sua pesquisa.

A avaliação das estratégias de interação implementadas é realizada aplicando-se inspeção de usabilidade, que visa identificar eventuais falhas de *design* e propor correções para melhorar a usabilidade[2] do sistema.

A validação foi feita por meio da condução de estudos de caso, propostos sobre um corpus de notícicas e outro de artigos científicos. Cada estudo de caso explora tarefas aplicadas aos usuários[3], como, identificar grupos de documentos semelhantes, excluir documentos pouco representativos em um contexto exploratório, identificar fronteiras entre áreas distintas. A validação tem o objetivo de verificar se as funcionalidades desenvolvidas foram úteis para a conclusão das tarefas propostas.

1.3 Organização do Documento

Este documento está organizado do seguinte modo:

- O **capítulo 2** apresenta uma introdução sobre a visualização de documentos, especificamente aspectos sobre a criação de mapas de documentos empregando técnicas de projeção multidimensional.
- O **capítulo 3** aborda conceitos de Mineração Visual de Documentos (*Visual Text Mining*), área que integra o processo de visualização e mineração de dados para ampliar a eficiência na análise de grandes coleções textuais.
- O **capítulo 4** apresenta as funcionalidades de interação desenvolvidas para possibilitar a manipulação e análise dos tópicos extraídos de um mapa de documentos.
- No **capítulo 5** são apresentados os resultados obtidos com a avaliação e validação das estratégias propostas.
- Por fim, o **capítulo 6** apresenta a conclusão do trabalho e sugestões de trabalhos futuros.

[2]Usabilidade é um termo bastante amplo que descreve um conjunto de caracerísticas em sistemas interativos, envolvendo facilidade de uso, facilidade de aprendizado, quão agradável e eficiente é o seu uso etc.

[3]As tarefas foram conduzidas por pesquisadores envolvidos no desenvolvimento do *software* utilizado pelo grupo de pesquisa

CAPÍTULO

2

Mapas de Documentos

A busca de informações relevantes em uma grande base de documentos textuais é uma tarefa custosa e cansativa. Naturalmente existem ferramentas de busca que, por meio de consultas por palavras-chave, se propõem a facilitar o acesso aos documentos 'filtrando' aqueles que não correspondem ao interesse do usuário. Essas técnicas, no entanto, apenas fornecem um universo menor no seu espaço de busca, sendo necessário ainda percorrer sequencialmente a lista de documentos fornecida como saída. Outra consideração importante é o fato de que essas ferramentas não fornecem informações de alto nível sobre o relacionamento entre os documentos retornados, como, por exemplo, suas áreas e sub-áreas. Adicionalmente, em uma busca exploratória, o usuário não tem uma idéia precisa sobre o que está pesquisando a ponto de formular uma consulta sobre o universo de documentos.

A visualização de informação dispõe de técnicas para auxiliar a identificar relações entre conjuntos de documentos. Algoritmos de redução de dimensionalidade permitem criar mapas de documentos[1] com base na similaridade de conteúdo, que fornecem uma visão intuitiva sobre a similaridade entre documentos e as áreas a que pertencem. Nesses mapas cada documento é representado como um único ponto no plano, sendo que a proximidade entre os pontos sugere a similaridade de seus conteúdos. Pontos próximos indicam documentos altamente similares

[1]Essa representação está inserida no contexto de diversas pesquisas investigadas pelo grupo: http://infoserver.lcad.icmc.usp.br

em conteúdo, e vice-versa. Assim, agrupamentos de pontos sugerem conjuntos que abordam assuntos semelhantes.

Ao explorar a capacidade do ser humano de reconhecer padrões visuais, os mapas de documentos possibilitam que uma pesquisa em uma base de textos não seja conduzida apenas pela leitura exaustiva de cada documento, pois a visualização do mapa permite focar apenas em determinados agrupamentos de documentos.

2.1 Etapas da visualização

Segundo Börner *et al.* (2003), o processo de criação de mapas de documentos envolve seis etapas principais:

A **extração de dados** define que informações são efetivamente extraídas de um conjunto de documentos para compor os dados a serem visualizados. A qualidade da visualização é initmamente ligada à qualidade dos dados, e portanto deve-se definir com cuidado que informações contidas no textos são relevantes para o processo de análise.

Definir uma **unidade de análise** implica escolher qual o foco da visualização, ou seja, por qual ângulo deseja-se analisar o conjunto de documentos. A unidade de análise pode ser, por exemplo, autores, termos, documentos, revistas etc. Ao definir a unidade como sendo os autores, pode-se visualizar a relação entre os trabalhos de determinados grupos de pesquisa e definir a fronteira entre áreas do conhecimento. Definindo-se a unidade como documentos, pode-se visualizar a relação de similaridade entre conjuntos de documentos e identificar agrupamentos de documentos que abordam assuntos relacionados.

A **seleção de medidas** e **cálculo de similaridade** definem que estratégia será empregada sobre a unidade de análise para relacionar uma medida de distância entre seus elementos. Documentos similares podem ser definidos como aqueles que são citados em conjunto por outros documentos (co-citação). Outra estratégia é representar documentos por um modelo vetorial (o mesmo empregado em Recuperação de Informação), no qual cada documento é definido como um vetor e cada termo presente no documento corresponde a uma coluna desse vetor, e então aplicar uma função para medir a similaridade entre os vetores (como a distância dos cossenos, por exemplo). O resultado desse processo é a criação de uma matriz de distâncias.

A etapa de **arranjo** tem o objetivo de representar os dados multidimensionais, obtidos nas etapas anteriores, em um número reduzido de dimensões espaciais (tipicamente duas ou três) e então apresentar os documentos em um espaço 2D ou 3D. Para esse fim são empregadas técnicas como *Multidimensional Scaling* (MDS) (Xu e Wunsch, 2008), *Self-Organizing Maps* (SOMs) (Skupin, 2002) etc.

Por fim, a etapa de **exibição** apresenta o mapa ao usuário, que deve ser capaz de visualizar todo o conjunto de documentos simultaneamente ou focar em regiões específicas. Para tal deve interagir com o mapa de documentos, por meio de operações de *zoom* e seleção, por exemplo.

2.2 Medidas de Similaridade

A medida de similaridade entre dois documentos deve refletir, idealmente, a mesma noção de similaridade compreendida por seres humanos analisando os mesmos documentos. No entanto, como o modo de análise empregado por seres humanos ainda não é completamente conhecido, as medidas de similaridade podem ser consideradas apenas aproximações (Navarro e Lee, 2001).

Uma abordagem amplamente utilizada para representar documentos na etapa de seleção de medidas e cálculo de similaridade é o modelo de espaço vetorial. Esse modelo será então adotado para ilustrar o tratamento de similaridade entre documentos neste trabalho. Nesse modelo os documentos são representados por vetores, nos quais cada dimensão representa uma palavra distinta do corpus, e o valor de uma posição do vetor registra uma contagem da frequência dessa palavra no respectivo documento.

Na próxima a etapa devem ser atribuídos pesos aos termos existentes na representação vetorial. Um modelo utilizado para o cálculo de pesos é o $tf - idf$ (*term frequency-inverse document frequency*) (Salton e Buckley, 1987). Nele, maior peso é atribuído aos termos que aparecem com maior frequência apenas em um conjunto pequeno de documentos. A sigla tf simboliza a contagem da frequência de um determinado termo em um documento em particular, e a sigla idf a contagem do número de documentos que também possui o termo em seu conteúdo. A Expressão 2.1 é utilizada para calcular o peso de um determinado termo:

$$w_t = f \times \log \frac{N}{n} \tag{2.1}$$

O peso do termo é representado então pela contagem de sua frequência (f) no documento analisado, multiplicado pelo logaritmo da razão entre o número total de documentos do corpus (N) e o número de documentos que contêm o termo (n).

Documentos grandes possuem maior quantidade de termos do que documentos pequenos. Como consequência, um documento grande terá uma contagem de termos relativamente maior, o que irá aumentar o peso (importância) de seus termos. É desejável que todos os documentos relevantes sejam tratados de forma igual para fins de comparação, independente de seu tamanho. É necessário, portanto, normalizar os pesos dos termos. Assumindo que w_t representa o peso

do termo t, em um documento com k dimensões, o peso final atribuído pode ser calculado pela Expressão 2.2 (Salton e Buckley, 1987):

$$w_t = \frac{w_t}{\sum_{i=1}^{k} w_i} \tag{2.2}$$

Os pesos são utilizados então pelas técnicas de representação de similaridade que utilizam o modelo de representação vetorial de documentos. Existem medidas que não são baseadas no modelo vetorial, como o caso da medida baseada em aproximação de Kolmogorov (Telles *et al.*, 2007), por exemplo. Nesse caso o processo é feito diretamente sobre o corpo do documento, comparando texto contra texto. A seguir serão ilustradas duas métricas aplicáveis ao modelo vetorial: Distância Euclidiana, que representa a noção intuitiva de distância nos eixos cartesianos, e a Distância dos Cossenos, empregada para medir distância em conjuntos de dados esparsos.

2.2.1 Distância Euclidiana

A medida Euclidiana calcula a distância entre documentos usando o mesmo princípio empregado no cálculo da distância entre vetores nos eixos cartesianos.

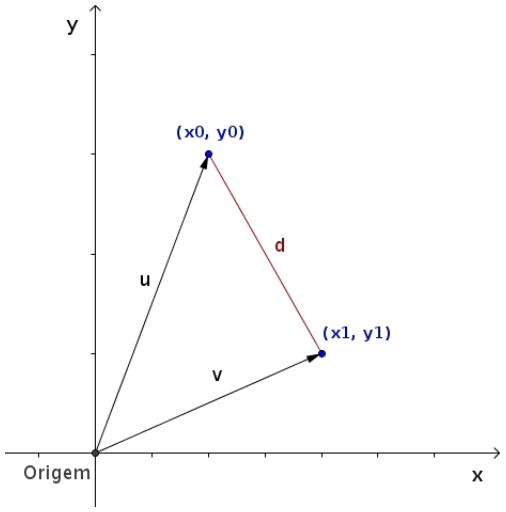


Figura 2.1: Representação da distância Euclidiana entre vetores

No exemplo ilustrado na Figura 2.1, supondo duas dimensões, a distância entre os vetores u e v é calculada pela expressão 2.3:

$$d_{uv} = \sqrt[2]{(x_1 - x_0)^2 + (y_1 - y_0)^2} \tag{2.3}$$

Esta expressão pode ser extendida para n dimensões ao incluir o quadrado da diferença dos termos correspondentes a cada dimensão. Dessa forma, a equação utilizada para medir a distância euclidiana d entre dois documentos i e j, de n dimensões (palavras), é dada pela expressão 2.4:

$$d_{ij} = \sqrt[2]{(w_{i,1} - w_{j,1})^2 + \cdots + (w_{i,n} - w_{j,n})^2} \tag{2.4}$$

O resultado final da equação indica a distância. Assim, quanto menor o resultado, maior é a similaridade entre os documentos. No entanto essa métrica não é adequada para calcular a distância entre documentos, pois se tratam de dados esparsos, ou seja, muitas posições do vetor não possuem informação.

2.2.2 Distância dos Cossenos

O princípio dessa métrica é adotar que medir o ângulo formado por dois vetores dá uma idéia de sua similaridade. A Figura 2.2 ilustra um exemplo da medida dos ângulos entre os vetores u, v e w, no domínio 2D. Conforme o ângulo entre eles diminui, o cosseno do ângulo se aproxima de 1, indicando que a sua distância é menor (Garcia, 2006).

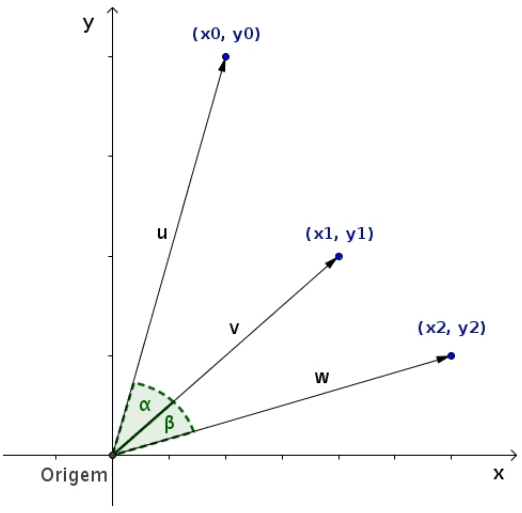


Figura 2.2: Medida dos ângulos entre vetores

A equação empregada para calcular o cosseno do ângulo entre os vetores u e v, de duas dimensões, ilustrados na Figura 2.2, é exibida na expressão 2.5:

$$\cos\alpha = \frac{u \times v}{|u| \times |v|} = \frac{(x_0 \times x_1) + (y_0 \times y_1)}{\sqrt[2]{(x_0^2 + y_0^2)} \times \sqrt[2]{(x_1^2 + y_1^2)}} \tag{2.5}$$

Essa expressão pode ser estendida para várias dimensões. Por exemplo, ao ser empregada para calcular a distância entre os documentos i e j, de n dimensões, a expressão assume a forma ilustrada em 2.6:

$$\cos\alpha = \frac{(w_{i,1} \times w_{j,1}) + \cdots + (w_{i,n} \times w_{j,n})}{\sqrt[2]{w_{i,1}^2 + \cdots + w_{i,n}^2} \times \sqrt[2]{w_{j,1}^2 + \cdots + w_{j,n}^2}} \tag{2.6}$$

Essa técnica é usualmente utilizada para calcular a similaridade entre documentos.

2.3 Técnicas de Projeção Multidimensional

As técnicas de projeção multidimensional procuram mapear um espaço multidimensional n em um espaço k, no qual k tem tamanho menor do que n. Essa redução de dimensionalidade procura manter as relações de distância e similaridade entre os elementos projetados, embora a perda de informação seja inevitável. Elas podem ser classificadas em duas categorias, de acordo com a função de projeção utilizada: técnicas de projeção lineares e técnicas de projeção não-lineares.

As técnicas lineares trabalham criando combinações lineares dos atributos originais, definindo uma nova base ortogonal de dimensionalidade reduzida. Técnicas lineares trabalham bem com dados normalmente distribuídos, mas são incapazes de reconhecer características importantes em estruturas não lineares, como agrupamentos de formatos arbitrários.

Já as técnicas não-lineares procuram minimizar uma função de perda de informação decorrente da projeção. Essa função normalmente atua apenas sobre as distâncias calculadas no espaço original, não necessitando, portanto, trabalhar sobre os vetores originais. Outra vantagem é a natureza iterativa das técnicas não-lineares, que requerem apenas iterações adicionais para acrescentar novos dados à projeção. Para acrescentar novos dados as técnicas lineares requerem que a projeção seja re-calculada desde o início.

Serão descritas a seguir algumas técnicas de projeção não-lineares implementadas na ferramenta *Projection Explorer* (Paulovich *et al.*, 2007), utilizada no projeto para a elaboração dos mapas de documentos por similaridade. Para ilustrar essas técnicas foi utilizado um corpus de 574 artigos científicos, distribuídos em três áreas de aprendizado de máquina. Nesses mapas a cor do documento aponta a classe a qual ele pertence: *Case-Based Reasoning* (vermelho), *Inductive Logic Programming* (verde) e *Information Retrieval* (azul).

2.3.1 IDMAP - *Interactive Document Map*

A técnica IDMAP projeta os objetos inicialmente utilizando uma técnica rápida de projeção, como a *Fastmap*, e depois aplica um processo de ajuste aos pontos projetados para compensar a informação perdida durante a projeção. O processo de ajuste consiste em calcular as distâncias sobre o espaço original e o espaço projetado, e então aplicar forças de atração e repulsão sobre cada ponto para compensar o erro introduzido na projeção. Essa estratégia de simulação de forças é chamado *Force-Directed Placement* (FDP). Um algoritmo que implementa essa estratégia é o *Force Scheme* (Tejada *et al.*, 2003).

A técnica Fastmap, proposta por (Faloutsos e Lin, 1995), projeta os pontos, originalmente em um espaço de dimensão n, em um espaço de dimensão k, com $k < n$. Inicialmente são escolhidos os dois pontos mais distantes entre si, de acordo com a matriz de distâncias. Os dois pontos definem uma reta no espaço n-dimensional e um hiperplano de dimensionalidade $n - 1$ perpendicular à reta. Cada um dos demais objetos é projetado nesse plano. O processo é repetido até que $n - 1$ atinja o valor desejado de k. (Pinho, 2009). A Figura 2.3 exibe um mapa de documentos projetado com IDMAP.

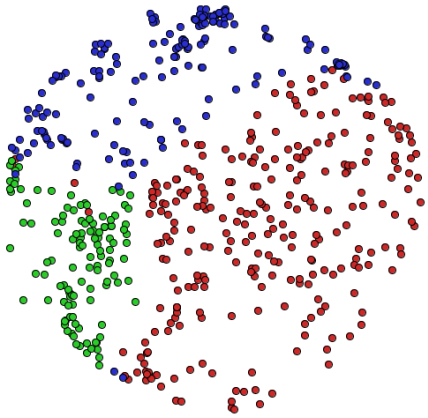


Figura 2.3: Mapa de documentos projetado com a técnica IDMAP

2.3.2 ProjClus - *Projection by Clustering*

O princípio da técnica ProjClus (Paulovich e Minghim, 2006) é agrupar os elementos da projeção e depois aplicar a técnica de FDP *Force Scheme* sobre cada agrupamento. Inicialmente o centróide de cada agrupamento é calculado e projetado no plano, para depois projetar separadamente cada conjunto de pontos pertencente a cada agrupamento. Por fim o desenho final do mapa é ajustado para refletir a posição de cada agrupamento de acordo com a localização de seus centróides. A Figura 2.4 exibe um mapa de documentos projetado com ProjClus.

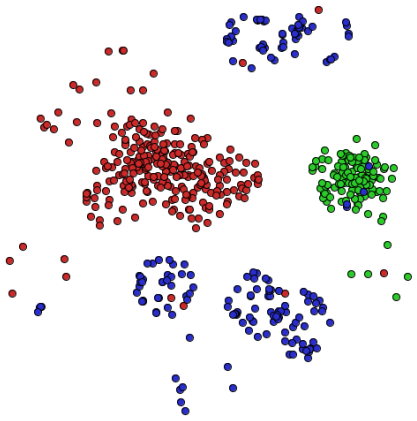


Figura 2.4: Mapa de documentos projetado com a técnica ProjClus

2.3.3 LSP - *Least Squares Projection*

Considerando um conjunto de pontos $S = \{p_1, \cdots, p_m\}$ inicialmente em $\mathbb{R}^n$, a técnica LSP inicialmente projeta em $\mathbb{R}^k$ ($k < n$) um sub-conjunto de S (chamado de pontos de controle). Os pontos de controle são projetados utilizando uma técnica de projeção que procura preservar as relações de distância no espaço original, como a *Fastmap*.

Utilizando as relações de vizinhança dos pontos em $\mathbb{R}^n$, e das coordenadas dos pontos de controle em $\mathbb{R}^k$, é calculado um sistema linear cuja solução posiciona o restante dos pontos em $\mathbb{R}^k$ (Paulovich *et al.*, 2008). A Figura 2.5 exibe um mapa de documentos projetado utilizando a técnica LSP.

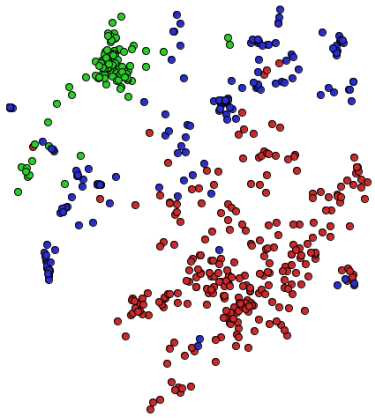


Figura 2.5: Mapa de documentos projetado com a técnica LSP

2.4 Plataformas de Visualização de Documentos

2.4.1 Projection Explorer

A plataforma *Projection Explorer* (PEx) (Paulovich *et al.*, 2007), desenvolvida pelo grupo de pesquisa e disponibilizada gratuitamente sob a licença GNU/GPL, permite criar mapas visuais de conjuntos multidimensionais em geral, como documentos, dados estruturados (tabelas), buscas na web etc. Os mapas são construídos utilizando técnicas de projeção multidimensional, como discutido na seção 2.3. A Figura 2.6 exibe a janela de visualização do *software* contendo um mapa de documentos.

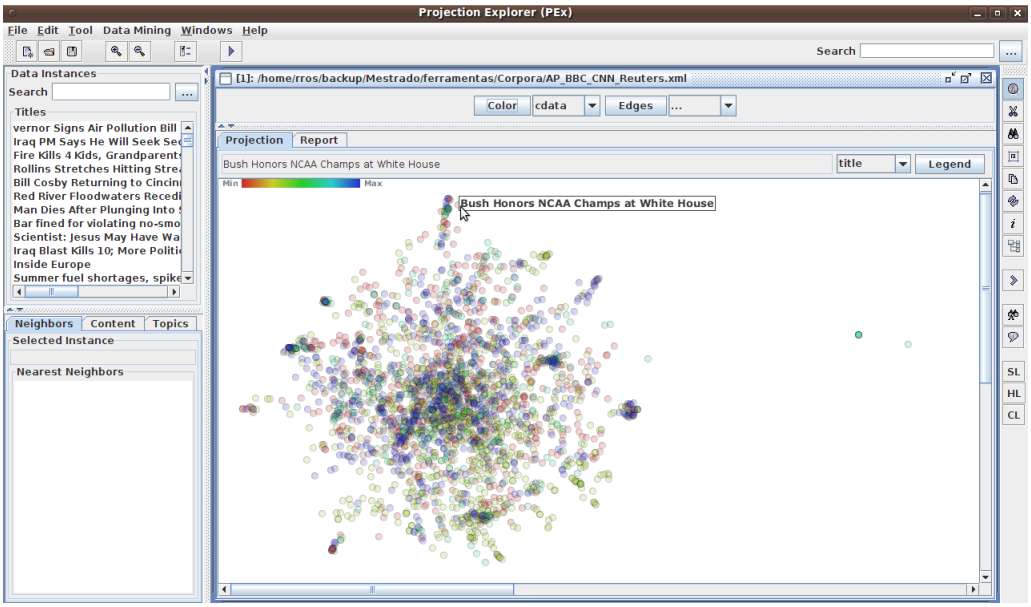


Figura 2.6: Janela de visualização do *Projection Explorer*

O processo para a visualização de mapas de documentos é ilustrado na Figura 2.7. Inicialmente o conjunto de documentos é submetido a um processo de **pré-processamento** no qual é realizada a contagem das palavras por documento e é construído o modelo de representação vetorial. Nesse processo são descartadas palavras consideradas não significativas (*stopwords*), como artigos, preposições, por exemplo. Opcionalmente pode-se aplicar cortes de Luhn (Luhn, 1958) para eliminar palavras muito ou pouco frequentes, que normalmente não são significativas. Pode-se, também, aplicar um algoritmo de lematização para reduzir as palavras ao seu radical, e assim reduzir a quantidade de dimensões. Posteriormente é aplicado um algoritmo que associa um respectivo peso a cada palavra, como o *tf-idf* (Salton e Buckley, 1987). O resultado final do pré-processamento é uma matriz esparsa, em que cada linha representa um documento do corpus e cada coluna o peso de determinada palavra.

A próxima etapa é o **cálculo de distâncias** entre os documentos, objetivando estabelecer relações entre os elementos analisados, que deverão ser traduzidas como relações de distância durante a montagem dos mapas. O resultado é, usualmente, uma matriz de distâncias em um espaço n-dimensional, no qual n representa a quantidade de documentos processados (Pinho, 2009).

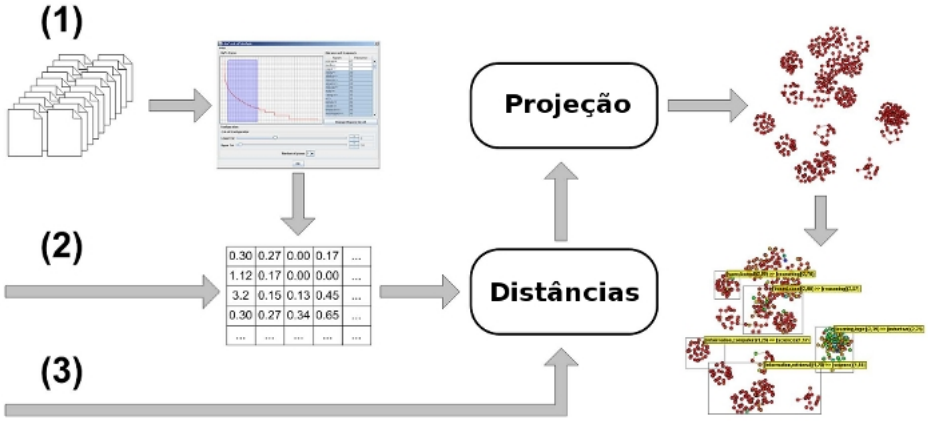


Figura 2.7: Processo de geração de um mapa de documentos por meio de projeções multidimensionais: (1) pré-processamento (extração de termos relevantes), geração do modelo vetorial, (2) cálculo de distâncias e (3) depois projeção no espaço 2D. O layout gerado serve de base para diferentes representações visuais, cujas propriedades gráficas podem ser alteradas interativamente conforme o objetivo da visualização. Fonte: Paulovich *et al.* (2007).

Por fim, utiliza-se uma técnica de **projeção** para reduzir a dimensionalidade da matriz. Foram desenvolvidas e incorporadas ao PEx diversas técnicas de projeção (ver seção 2.3). Ao reduzir a dimensionalidade para duas ou três dimensões, é possível visualizar os dados como grafos 2D, 3D ou superfícies. A partir da representação visual é possível mapear atributos (características) a cada elemento projetado. Valores como a quantidade de um determinado termo em um documento podem ser mapeados na representação visual como cor, ou altura da superfície. Com a representação visual já definida, o usuário pode interagir com a imagem para iniciar a exploração do mapa.

O *software* possui ainda diversas funcionalidades para auxiliar a exploração de mapas de documentos. Uma barra de tarefas, disposta à direita da janela de visualização, possui opções para selecionar documentos e visualizar o seu conteúdo, extrair tópicos a partir da seleção de um grupo de documentos, coordenar diferentes visualizações do mesmo conjunto de documentos etc. É possível, também, mapear atributos à cor e ao tamanho de cada ponto do mapa, como por exemplo a frequência de determinado termo nos documentos, ou a classe a que pertence cada documento.

2.4.2 InfoSky

O InfoSky (Andrews *et al.*, 2002) permite explorar conjuntos de documentos por meio de uma representação visual que se assemelha à observação de estrelas no espaço. Os documentos são representados como estrelas, que são posicionadas no plano 2D segundo um algoritmo de FDP cujas forças de atração e repulsão são baseadas na similaridade de conteúdo de cada documento. Dessa forma o mapa de documentos é visualizado como conjuntos de pequenos pontos em um plano 2D, e o usuário pode explorar os grupos de documentos similares, posicionados próximos uns aos outros, utilizando uma metáfora visual que se assemelha a um telescópio (Figura 2.8).

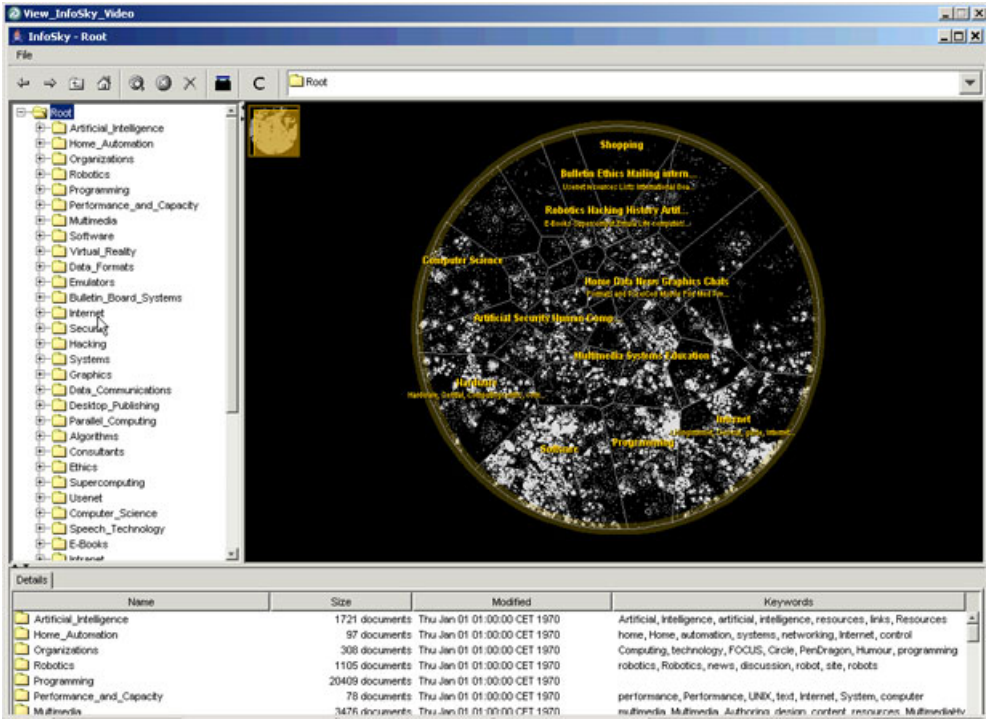
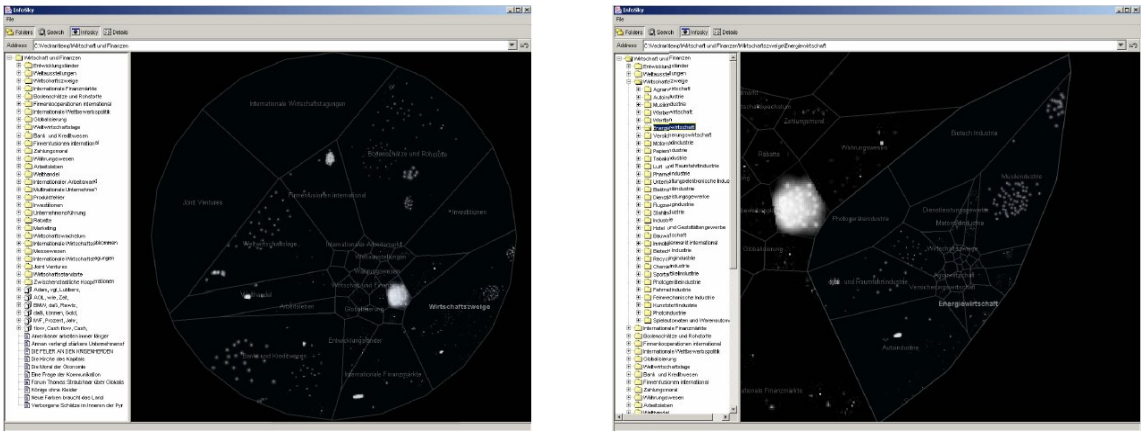


Figura 2.8: Janela de visualização do InfoSky. Documentos representados como estrelas, em uma visualização que se assemelha a um telescópio observando o espaço. Fonte: Andrews *et al.* (2002)

A construção da visualização considera relações de hierarquia, definidas em repositórios de documentos. Essas relações são representadas como polígonos que delimitam as áreas e sub-áreas dos agrupamentos (constelações) de documentos. Os polígonos são calculados sobre o centróide de cada agrupamento, utilizando diagramas Voronoi modificados, sendo a área do polígono definida de acordo com a quantidade de documentos contemplada.

O usuário pode navegar pela hierarquia de documentos utilizando um controle, localizado em um painel à esquerda da janela de visualização, que exibe a organização do corpus em diferentes pastas. Conforme uma pasta é selecionada, o *software* focaliza e magnifica a região

no mapa e o próximo nível de hierarquia da coleção é visualizado. O processo pode se repetir até que o último nível da hierarquia dos documentos seja atingido e apenas um grupo de estrelas focado. Adicionalmente o usuário pode pesquisar por documentos por meio de consultas por palavras chave. Os documentos que satisfazem a condição de busca são enfatizados no mapa, permitindo ao usuário focar a pesquisa na região correspondente.



(a) Exibição da hierarquia de grupos e sub-grupos de documentos: visão geral do corpus

(b) Exibição dos documentos (estrelas): maior detalhes das relações de grupos específicos

Figura 2.9: Visualização da hierarquia no mapa de documentos do InfoSky. Múltiplos níveis de hierarquia de podem ser focados. Fonte: Andrews *et al.* (2002)

2.4.3 Self-Organizing Maps

São mapas que utilizam redes neurais artificiais para definir a localização dos documentos (pontos) no plano 2D. O processo de construção desse mapa é ilustrado na Figura 2.10. Inicialmente é definida uma quantidade de neurônios que deve receber uma amostra de documentos para treinamento. Os documentos são estruturados no modelo de representação vetorial, e cada neurônio recebe como entrada um vetor dessa dimensão. A região de visualização é então dividida igualmente entre os neurônios, de forma que a representação inicial se assemelhe a uma matriz **(a)**. Os documentos do corpus são atribuídos à região do mapa correspondente ao neurônio considerado mais similar (comparando o vetor do documento ao vetor do neurônio). Um único neurônio pode ser associado a múltiplos documentos **(b)**. Posteriormente os documentos são posicionados aleatoriamente dentro da região da celula do neurônio **(c)**. Por fim, a cada documento é atribuído uma área do mapa 2D usando polígonos Voronoi ou Thiessen **(d)**.

Para visualizar uma grande quantidade de documentos, deve ser aplicado um algoritmo de agrupamento para evitar oclusão visual e permitir agrupar os documentos por áreas e sub-áreas. Os agrupamentos definem novos polígonos, formados pela união dos polígonos dos respectivos

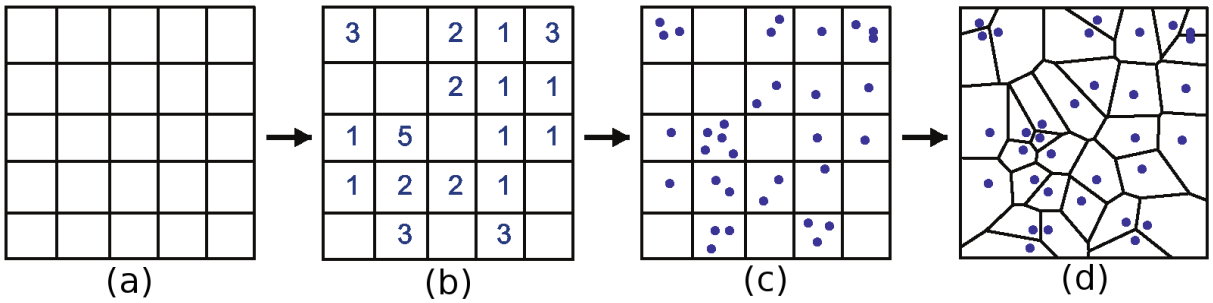


Figura 2.10: Processo de construção da visualização de um *self-organizing map*. Fonte: Skupin (2002)

documentos, conforme ilustrado na Figura 2.11. Ao explorar um mapa de documentos, o usuário pode inicialmente visualizar uma quantidade reduzida de grupos e obter uma visão geral do corpus. Ao definir uma área de interesse, ele pode focar em um grupo específico, de forma que o mapa exiba uma quantidade maior de agrupamentos para essa região do mapa.

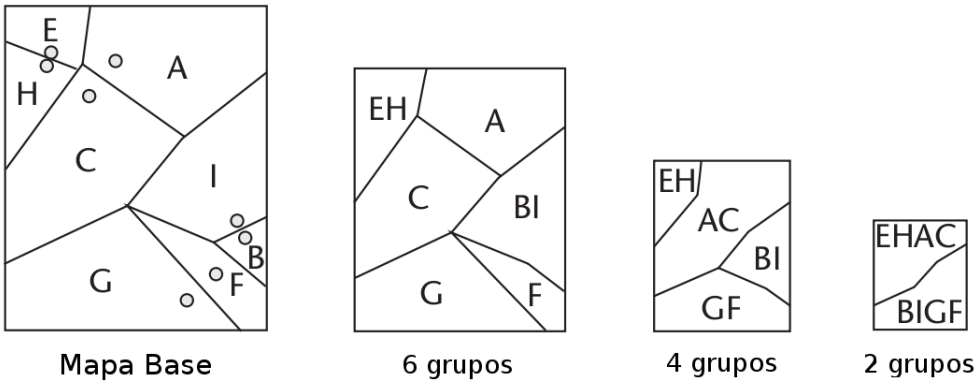


Figura 2.11: Polígonos que delimitam as regiões no mapa, definidas de acordo com os agrupamentos de documentos. Fonte: Skupin (2002)

O trabalho de Skupin (2002) ilustra uma aplicação de *self-organizing maps* ao visualizar um corpus de 2.200 artigos, contendo os *abstracts* submetidos ao *Annual Meeting of the Association of American Geographers* (AAG) de 1999. O mapa, ilustrado na Figura 2.12, define os domínios de conhecimento contemplados no artigos do evento em três níveis de detalhamento. O primeiro nível, delimitado com bordas vermelhas, exibe 10 grupos de documentos e fornece uma visão geral do domínio. O segundo nível, com bordas verdes, particiona o corpus em 25 grupos, fornecendo mais detalhes sobre cada área. O terceito nível, com bordas cinzas, fornece o maior nível de detalhes ao particionar o corpus em 100 grupos.

Figura 2.12: *Self-organizing map* de um corpus de 2.200 artigos sobre temas relacionados a geografia. Fonte: Skupin (2002)

2.5 Considerações Finais

Este capítulo apresentou alguns desafios encontrados na análise de grandes volumes de dados textuais e como as técnicas de visualização de informação, especificamente mapas de documentos, podem ser utilizadas para auxiliar a análise exploratória desses dados.

O próximo capítulo trata a questão de Mineração Visual de Textos (*Visual Text Mining*), área que integra visualização e mineração de dados para apoiar processos de extração de conhecimento de conjuntos de documentos.

CAPÍTULO

3

Mineração Visual de Textos

A quantidade crescente de informação atualmente disponível para análise torna indispensável o uso de ferramentas que auxiliem a sua interpretação. Os dados podem ser provenientes de diversas fontes: bases de dados, planilhas, internet, repositórios de documentos, etc.

Dados provenientes de planilhas e bases de dados podem ser considerados estruturados, pois seguem uma estrutura pré-definida em que cada instância de informação é estruturada segundo um conjunto fixo de atributos. Por exemplo, em uma dada base de dados, as informações sobre o cadastro de uma pessoa podem ser organizadas em campos de tamanho determinado, contendo o seu nome, endereço, telefone, cep etc. A análise desses dados é relativamente simples, pois como os campos são bem definidos, bastaria ordenar os dados segundo algum critério ou realizar alguma consulta à base de dados (utilizando uma linguagem específica para esse fim, como a SQL (*Standard Query Language*)) para elaborar relatórios detalhados.

Já os dados provenientes de textos, como documentos e páginas web, são considerados desestruturados. A informação está espalhada no corpo do texto e cabe ao usuário ler, interpretar e extrair a informação que realmente o interessa. Nesse cenário, conforme o volume de documentos aumenta, a análise pode se tornar uma tarefa exaustiva e pouco eficiente. Estudos estimam que cerca de 85% de toda informação utilizada pelas corporações são armazenadas na forma de textos (Hotho *et al.*, 2005). Muitas decisões estratégicas dependem da interpretação em tempo hábil desse grande volume de informações.

3.1 Mineração de Textos

Mineração de textos pode ser definido como o processo de extração de conhecimento no qual um usuário interage com uma coleção de documentos com o auxílio de ferramentas de análise. É uma área que objetiva extrair informação útil a partir de coleções de documentos, por meio da identificação e exploração de padrões interessantes. Como os documentos são fontes de dados desestruturadas, os esforços da mineração de textos são voltados a identificar e extrair características representativas para encontrar um formato estruturado intermediário (Feldman e Sanger, 2006).

Uma das técnicas de mineração utilizadas para descrever padrões são as **regras de associação**. Essa abordagem baseia-se na identificação de itens que co-ocorrem em um conjunto analisado, sendo um exemplo típico a utilização em bases de dados de compras de supermercado. É possível analisar um conjunto de transações de compras de diversos clientes e verificar a existência de padrões em itens que são adquiridos em conjunto.

As regras são implicações definidas na forma $X \Longrightarrow Y$ na qual X determina o antecedente e Y o consequente da regra (Agrawal *et al.*, 1993). Tanto o antecedente quanto o consequente podem possuir um ou mais itens do conjunto analisado (*itemset*), mas não há elementos em comum entre os dois, ou seja, $X \cap Y = \emptyset$. No contexto de mineração de textos, os itens são os termos provenientes da coleção de documentos (*bag of words*).

Segundo Agrawal e Srikant (1994), a aplicação de regras de associação em textos é definida da seguinte forma: seja $I = i_1, i_2, \ldots, i_m$ um conjunto de itens representando os termos da *bag of words*. Seja T um conjunto de transações, na qual cada transação D é um conjunto de itens ($D \subseteq I$) que representa um documento. Uma regra de associação é uma implicação na forma $X \Longrightarrow Y$, na qual $X \subseteq I$, $Y \subseteq I$ e $X \cap Y = \emptyset$.

Duas medidas empregadas em regras de associação são o suporte e confiança. A regra $X \Longrightarrow Y$ possui suporte s no conjunto de transações (documentos) T se $s\%$ das transações possuirem os termos de $X \cup Y$. A mesma regra possui confiança c se $c\%$ dos documentos em T que contenham X também contenham Y. Em outras palavras o suporte indica a proporção de documentos do conjunto que possuem todos os termos presentes na regra de associação. Já a confiança da regra ($X \Longrightarrow Y$) indica, dentre as transações que contêm X, a proporção das transações que também contêm Y. Uma regra com um suporte alto indica que todos os seus termos são encontrados em um grande conjunto de documentos. Se uma regra possuir uma alta confiança, a presença do antecedente indica uma alta probabilidade de ocorrência do consequente (Minghim e Levkowitz, 2007).

Devido à complexidade envolvida na mineração de textos, as técnicas de visualização de informação são fundamentais para ampliar a capacidade de análise. A integração entre mineração de textos e visualização, denominada mineração visual de textos, insere o usuário no processo de construção de um modelo mental para um conjunto de dados particular (Lopes *et al.*, 2007). A intervenção do usuário pode guiar o processo de mineração e medidas subjetivas podem ser melhor especificadas pelo usuário, de forma que as técnicas possam atingir melhores resultados.

O restante desse capítulo apresenta técnicas de mineração de textos para extração automática de tópicos de documentos. Também discute como os tópicos extraídos podem ser utilizados para guiar o processo de exploração e refinamento de um mapa de documentos.

3.2 Extração de Tópicos em Documentos

Tópicos são representações sucintas sobre conteúdo de um conjunto de documentos. A extração de tópicos é uma técnica de mineração útil para viabilizar a tarefa de exploração de um conjunto grande de documentos, pois a existência de palavras que possam definir grupos e sub-grupos do corpus evita que se faça a leitura de cada texto individualmente para identificar o seu conteúdo.

Se, em um processo de exploração, o usuário definir previamente o assunto desejado, ele pode então se dedicar exclusivamente a estudar o grupo rotulado com o assunto de seu interesse e descartar o restante. Dessa forma, o trabalho de leitura pode ser drasticamente reduzido, pois o conjunto de documentos que efetivamente interessa ao usuário pode ser uma pequena fração do conjunto inicial. Algoritmos de classificação podem, também, utilizar a informação dos tópicos para definir a classe de um conjunto de documentos, organizando-os de forma automática em uma base de dados.

São apresentadas a seguir duas técnicas utilizadas para a extração automática de tópicos de conjuntos de documentos, ambas utilizadas na ferramenta *Projection Explorer* (PEx[1]) para rotular mapas de documentos: uma técnica baseada em uma análise de covariância de termos, e outra que utiliza indução de regras de associação.

3.2.1 Covariância de Termos

A estratégia de extração de tópicos por covariância procura identificar termos que possuem alta dependência na coleção de documentos. A medida de covariância tem o objetivo justamente de medir a dependência entre elementos, e, quanto maior o valor, maior é a relação.

[1]Disponível em http://infoserver.lcad.icmc.usp.br

Essa estratégia identifica como tópicos os termos que apresentem uma covariância acima de um limiar previamente definido. A covariância termo a termo é calculada pela Equação 3.1 (Paulovich, 2008):

$$cov(t_i, t_j) = \frac{1}{n-1} \sum_{k=1}^{n} (t_{ki} - \bar{t}_i)(t_{kj} - \bar{t}_j) \tag{3.1}$$

onde $\bar{t}_i$ e $\bar{t}_j$ representam a média das frequências contabilizadas do i-ésimo e j-ésimo elementos do vetor de termos. As variáveis t_{ki} e t_{kj} representam os valores de frequência do i-ésimo e j-ésimo elementos, respectivamente, ambos do k-ésimo documento.

Inicialmente o tópico é definido pelos dois termos que apresentarem a maior covariância. Para cada termo restante é calculada a média da covariância relativa aos dois termos inicialmente escolhidos. Se o valor calculado para o novo termo ficar acima de um limiar previamente estipulado (α), a nova palavra é adicionada ao tópico.

No entanto, as palavras que formam o tópico dependem apenas de sua relação com os dois termos que inicialmente possuem a maior covariância. Dessa forma muitos tópicos interessantes podem não ser revelados, como no caso de, por exemplo, outros dois termos possuirem um valor de covariância ligeiramente menor que os termos iniciais. Para evitar que isso ocorra, é definido que quaisquer valores de covariância entre dois termos que sejam iguais ou maiores que um valor estipulado β, geram um novo tópico.

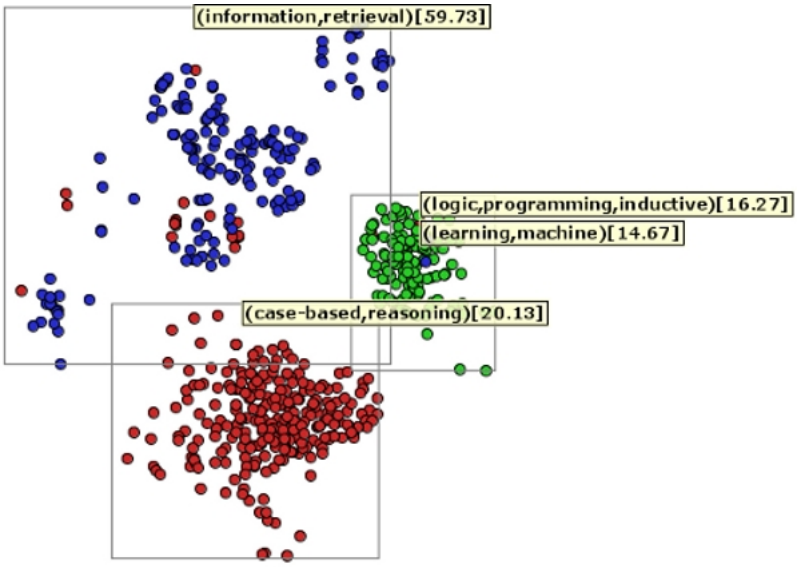


Figura 3.1: Mapa rotulado por tópicos criados por co-variância de termos. As regiões retangulares (em cinza) indicam as regiões selecionadas pelo usuário e o respectivo tópico, representativo dos documentos selecionados.

A Figura 3.1 exemplifica tópicos extraídos por covariância de termos. A imagem exibe um mapa de documentos composto por artigos científicos de três áreas de conhecimento, e é possível delimitar os agrupamentos relativos a cada área. No entanto, não é possível identificar o assunto de cada agrupamento sem o auxílio dos rótulos, que indicam os assuntos principais: *Inductive Logic Programming* (verde), *Information Retrieval* (azul), *Case-based Reasoning* (vermelho).

Esta é uma técnica útil para identificar os tópicos principais de um grupo, mas não é capaz de identificar, por exemplo, sub-grupos de documentos que abordam sub-tópicos específicos. Essa estratégia não leva em conta a relação de causalidade entre os termos, sendo apenas uma medida da relação de dependência que indica os termos que variam em conjunto.

3.2.2 Regras de Associação: Algoritmo LWR

Outra estratégia utilizada na ferramenta PEx para extrair tópicos é a indução de regras de associação sobre o conjunto de termos que compõem os documentos do mapa. Essa técnica pode ser empregada tanto em uma seleção de documentos conduzida pelo usuário, quanto em todo o mapa automaticamente. A produção de boas regras de associação fornece termos representativos, que podem ser utilizados como tópicos para auxiliar a identificar o conteúdo de conjuntos de documentos.

Um dos problemas encontrados em aplicações que utilizam regras de associação é a produção de um grande número de regras redundantes e pouco representativas. Conforme o número de regras cresce, aumenta também a dificuldade em analisa-lás e identificar associações interessantes. Uma situação em que muitas regras possuem poucos termos distintos indica uma alta redundância. Outra situação, em que regras com poucos termos distintos cobrem uma grande parcela dos documentos, indica tópicos genéricos e pouco representativos. Esses problemas devem ser tratados para que boas regras sejam criadas, ou seja, regras que descrevem os temas abordados por um conjunto de documentos, com pouca redundância e diferenciando-os dos demais.

O algoritmo LWR (*Locally Weighted Rules*) (Lopes *et al.*, 2007) estende e adapta o clássico algoritmo *Apriori* (Agrawal e Srikant, 1994) para extrair boas regras de associação. Dado um sub-conjunto de documentos – por exemplo, uma seleção do usuário no mapa de documentos – o LWR induz regras de associação que identificam os termos representativos abordados pelos documentos da seleção. As regras extraídas empregam ao menos um termo considerado altamente representativo (chamado semente). Os termos sementes são aqueles que aparecem com grande frequência no grupo selecionado e em menor frequência no restante do corpus. Outro

ponto atacado é a redundância do conjunto de regras criado. O algoritmo impõe apenas a adição de regras que contenham documentos ainda não cobertos por regras já existentes.

O algoritmo é estruturado em dois blocos aninhados. O bloco interno produz as regras mais relevantes de acordo com os termos de maior peso encontrados no conjunto selecionado de documentos. Apenas as regras que possuem ao menos uma semente são selecionadas. O bloco externo remove os documentos cobertos pelas regras já definidas e repete o processo, até que nenhuma nova regra seja produzida. As regras extraídas são então ordenadas de acordo com a medida de confiança. Dessa forma, as regras melhor classificadas podem ser obtidas e priorizadas. Se duas regras cobrem o mesmo subgrupo de documentos, então a regra de maior peso é escolhida.

A Figura 3.2 exibe um mapa de documentos com algumas seleções rotuladas por tópicos extraídos pelo algoritmo LWR. O mapa foi criado a partir de um corpus que contém 2684 notícias, coletadas durante dois dias em abril de 2006. Pode-se verificar que, de acordo com os tópicos extraídos, um grupo de documentos aborda notícias relacionadas à gripe aviária (*flu, bird*). Outro grupo aborda notícias sobre o torneio de golfe *Master's Augusta* (*masters, augusta*). Um terceiro grupo exibe notícias sobre o presidente do Peru (*president, peru*). O primeiro termo do tópico representa o antecedente da regra de associação, e o termos restantes representam o consequente. Em seguida é informado o suporte do tópico, em que o primeiro número indica a quantidade de elementos suportados pelo tópico, seguida da porcentagem de elementos cobertos pelo tópico no agrupamento selecionado. Por fim, o tópico exibe a confiança da regra.

É adotado também um processo para extração de regras em todo o mapa, que consiste em repetir o procedimento local sobre todas as regiões do mapa de documentos. Isso requer uma definição de como subdividir o mapa em regiões. São adotadas duas estratégias: **(i)** simular uma grade sobreposta ao mapa de documentos, e aplicar o procedimento local em cada célula dessa grade; e **(ii)** aplicar um algoritmo de agrupamento – o K-Médias, por exemplo – para particionar o mapa em grupos de documentos similares e, em seguida, executar o procedimento de extração de regras sobre cada um dos grupos definidos. A Figura 3.3 mostra um exemplo de mapa rotulado por regras de associação, utilizando a abordagem que extrai automaticamente regras de todas as regiões do mapa, por meio da estratégia de participnamento por agrupamentos.

Entretanto, ambas estratégias possuem o efeito indesejado de gerar um conjunto de regras de associação diferente conforme os parâmetros utilizados para definir a grade ou os agrupamentos, variação esta que não é desejável. O problema é contornado repetindo-se o processo utilizando, em cada interação, versões sutilmente diferentes da grade inicial, ou gerando vários agrupamentos com diferentes números de elementos.

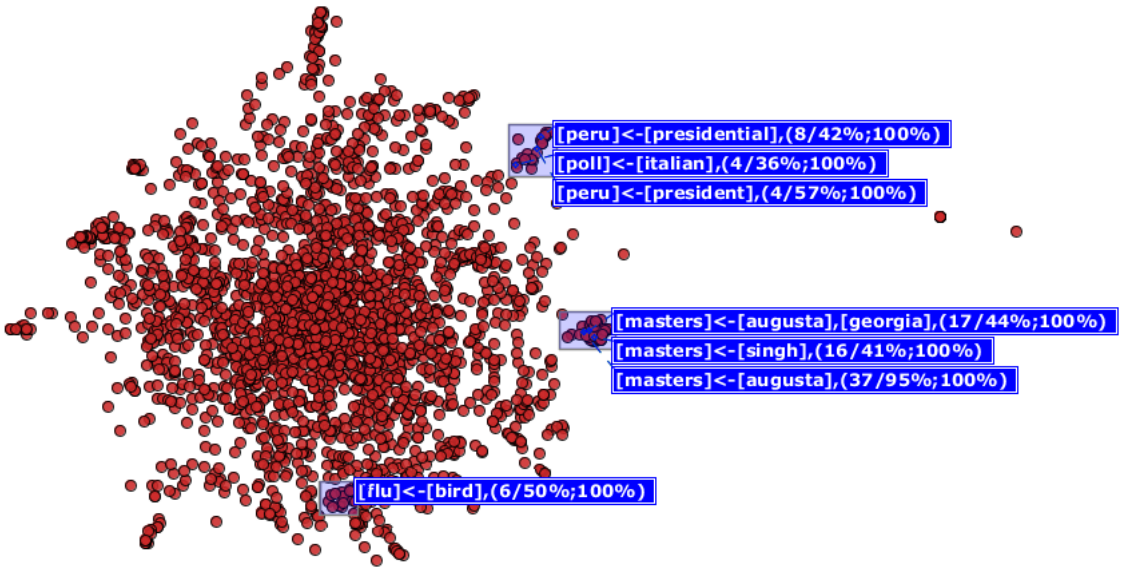


Figura 3.2: Mapa com grupos de documentos rotulados por regras de associação. As regiões destacadas em azul indicam o conjunto de documentos selecionados pelo usuário e os respectivos tópicos.

A estratégia de definir uma grade no mapa implica em dividir o domínio em l linhas e c colunas, e cada iteração incrementa o valor de l ou c. Outra possibilidade é de mover a posição s da grade no mapa. O processo se repete até que todas as células da grade sejam processadas, ou até que a condição de parada seja atingida. Na estratégia de particionamento por agrupamento deve-se definir uma quantidade inicial de grupos e o incremento (no número de grupos desejado) que será incorporado a cada iteração do algoritmo. É definido, também, um número máximo de grupos a ser utilizado. Em ambos os casos, as iterações prosseguem até que a quantidade de documentos cobertos por novas regras seja suficientemente pequena, ou até que as novas regras não promovam a cobertura de novos documentos.

É disponibilizada[2] uma API (*Application Programming Interface*), desenvolvida na linguagem Java, que implementa as técnicas de extração de regras de associação do LWR. Um aplicativo, que ilustra um exemplo de uso da API, é invocado pela linha de comando e utiliza como entrada uma matriz – que simboliza a representação vetorial do conjunto de documentos – e um conjunto de parâmetros que define o comportamento do algoritmo – como o suporte mínimo a ser considerado para cada regra. A saída do processo é o conjunto de regras gerado para o arquivo de entrada. A API pode ser utilizada para facilitar tarefas de avaliação, como determinar o suporte mínimo ideal para determinado conjunto.

[2]Disponível no endereço http://infoserver.lcad.icmc.usp.br/infovis2/LWR

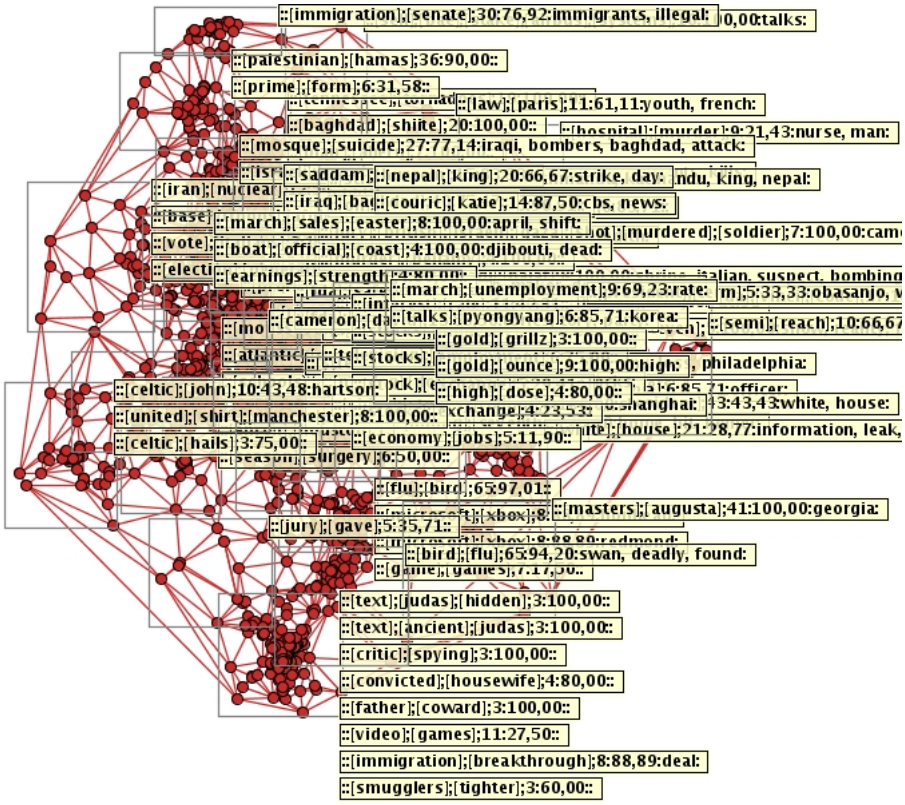


Figura 3.3: Mapa de documentos rotulado por regras de associação

A Figura 3.3 evidencia a dificuldade em interagir com um mapa densamente rotulado. A quantidade de rótulos pode ser tão grande que dificulta a visualização do mapa de documentos e até mesmo dos próprios tópicos. É necessário, portanto, desenvolver estratégias para exibir os tópicos em um espaço visual diferente do mapa.

3.3 Refinamento de Mapas de Documentos

Uma representação visual realmente informativa e útil precisa contar com o auxílio do conhecimento do usuário, a quem cabe direcionar o foco do processo exploratório, indicando os parâmetros necessários para que a ferramenta gere uma representação visual ajustada dinamicamente e mais efetiva. Portanto o usuário parte de uma visão global da coleção de documentos, fornecida pelo mapa inicial, mas direciona sua busca e focaliza gradualmente em sub-grupos de documentos de interesse. Esses grupos são então visualizados por mapas atualizados e refinados que refletem melhor o foco atual.

Para que o usuário direcione o processo exploratório é necessário que a ferramenta de visualização forneça informações sobre a representação. Os mapas de documentos inicialmente

informam ao usuário apenas a representação de distâncias entre os documentos. É possível então verificar a existência de agrupamentos, que indicam documentos com grande similaridade de conteúdo, mas não possibilita saber a priori os assuntos abordados por cada agrupamento.

Conforme descrito anteriormente, existem técnicas de mineração de texto capazes de fornecer tópicos que, de certa forma, sumarizam o conteúdo de um grupo de documentos. Esses tópicos podem ser, então, apresentados no mapa para identificar o conteúdo principal de cada agrupamento. Adicionalmente, os mesmos tópicos podem ser manipulados pelo usuário, que utilizando ferramentas de interação adequadas, saberia exatamente quais são os documentos referenciados pelos tópicos e julgaria se o conteúdo o interessa. Os documentos que não forem considerados relevantes poderiam ser gradualmente descartados e uma nova representação visual criada apenas com os documentos restantes, revelando eventualmente a presença de novos sub-grupos a serem explorados.

Supondo, por exemplo, que um usuário queira explorar o corpus de notícias apresentado na Figura 3.2 em busca de notícias sobre **Israel** e **Palestina**. Ele deve então procurar por documentos que abordem esse conteúdo e descartar os demais. A ferramenta PEx possui uma funcionalidade para pesquisar o mapa de documentos por termos e, ao informar os termos "*israel*" e "*palestin*", foram destacados 73 documentos no mapa. Estes foram selecionados manualmente e exportados como um novo corpus, afim de gerar um novo mapa de documentos, exibido na Figura 3.4. O novo mapa foi gerado empregando a técnica de projeção *ProjClus*, com os cortes de Luhn inferior e superior ajustados em 3 e 84, e fator de agrupamento 6. Ao comparar o mapa original (que contém 2684 documentos) com o novo mapa (com 73 documentos), é visível a expressiva redução no domínio de exploração.

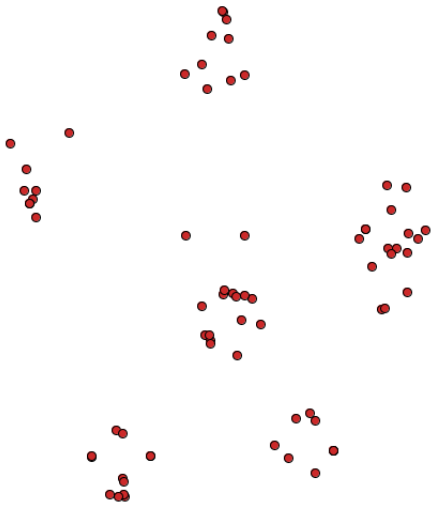


Figura 3.4: Mapa de documentos refinado, contendo notícias relacionadas a Israel e Palestina.

O mapa refinado indica a presença de seis agrupamentos de documentos similares, que podem ser novamente explorados pelo usuário e revelar novas relações entre os documentos, permitindo focar em assuntos mais específicos. Esse exemplo ilustra a utilidade de refinar um mapa de documentos, mas o modo como foi conduzido mostra a necessidade de desenvolver novas funcionalidades para auxiliar o usuário a refinar o mapa de modo mais eficiente.

As palavras empregadas na busca por documentos relevantes utilizando as palavras *israel* e *palestin* retornou 73 documentos. No entanto, o mapa pode conter outros documentos que abordam notícias sobre esse assunto, mas que não necessariamente contém essas palavras em seu conteúdo. Isso evidencia, também, a necessidade de encontrar formas de relacionar documentos que não compartilham os termos empregados em uma busca, mas que ainda assim abordem assuntos relacionados aos focos da pesquisa do usuário.

3.4 Considerações Finais

Este capítulo apresentou conceitos de mineração visual de textos, campo que une as áreas de visualização e mineração de textos. Foram abordadas técnicas de extração de tópicos e como essas técnicas podem auxiliar o usuário a aprimorar a representação visual. A representação visual amplia a capacidade de analisar grandes conjuntos de documentos, mas o mapa pode não ser adequado para atender aos objetivos do usuário. O problema, então, é como ajudá-lo no processo de refinamento da representação e busca por assuntos e temas de interesse.

O próximo capítulo aborda a criação de funcionalidades para auxiliar um usuário a manipular o mapa de documentos e a identificar grupos de documentos relacionados, por meio da análise dos tópicos previamente extraídos.

CAPÍTULO

4

Interação Baseada em Tópicos

Ao inspecionar um mapa de documentos, criado por similaridade de conteúdo, é possível observar a formação de diversos agrupamentos de documentos que supostamente abordam temas em comum. Inicialmente o mapa não informa o conteúdo dos documentos, mas "fornece pistas" sobre onde documentos similares podem ser encontrados. O próximo passo é selecionar regiões do mapa e extrair tópicos que, de certa forma, sumarizam o conteúdo dos documentos pertencentes ao grupo selecionado. Com base nesses tópicos, posicionados próximos à região selecionada, um usuário pode então ter informação suficiente para identificar quais assuntos são tratados nesses documentos. A extração de tópicos pode ser empregada de duas formas: manualmente, ao selecionar regiões específicas do mapa, ou automaticamente, ao utilizar uma estratégia de cobertura que simula essa seleção manual de grupos sobre todas as regiões do mapa.

No entanto, esse processo apresenta limitações. Até então a versão do *software Projection Explorer* [1] não possibilita interagir com os tópicos. Não é possível, por exemplo, selecionar um tópico e saber que documentos efetivamente contribuíram para criá-lo. Outro problema é a oclusão visual causada quando uma grande quantidade de tópicos é extraída. A quantidade de tópicos apresentados simultaneamente no mapa pode ser tão grande que impossibilita identificar os documentos e até mesmo a leitura dos próprios tópicos (ver na Figura 3.3).

[1] *Projection Explorer 1.6.3*

Esses problemas evidenciam a necessidade de funcionalidades que possibilitem interagir com os tópicos e eliminar a necessidade de exibi-los no mesmo espaço visual do mapa de documentos. O foco dessas funcionalidades, também, deve ser fornecer informações adicionais a respeito do relacionamento entre os tópicos e os respectivos documentos, de forma que o usuário tenha mais informações para explorar o corpus e identificar grupos de documentos relevantes à sua pesquisa. Ao manipular os tópicos, o usuário pode identificar documentos não interessantes, excluí-los do mapa, e assim priorizar o foco apenas nos documentos restantes.

Neste trabalho foram desenvolvidas cinco ferramentas de interação, descritas a seguir. Cada uma oferece funcionalidades complementares, e proporcionam diferentes visões sobre o conjunto de tópicos e documentos. O trabalho dá continuidade ao trabalho de doutorado de Roberto Dantas de Pinho (Pinho, 2009), ao implementar um processo interativo e iterativo baseado em tópicos, para a exploração visual de mapas de documentos, sugerido naquele trabalho.

4.1 Árvore de Tópicos

A árvore de tópicos oferece meios de manipular e exibir os tópicos em uma área visual adequada, independente do mapa de documentos. Cada nó da árvore pode representar um tópico ou documento. Os documentos são sempre exibidos como nós folha (não possuem filhos) e não há limitações sobre a quantidade máxima de filhos suportados pelos tópicos. A Figura 4.1 exibe uma árvore de tópicos. Os documentos do mapa que não são cobertos por nenhum tópico são associados ao nó *"Uncovered"*.

Na árvore também são incluídos sub-tópicos, definidos como tópicos que contemplam apenas uma parcela dos documentos contidos em um tópico já existente na árvore (o tópico "pai"). Foram definidas duas estratégias para caracterizar um sub-tópico: cobertura de documentos e cobertura de termos. A estratégia de **cobertura de documentos** considera um sub-tópico aquele tópico **"A"** que cobre apenas uma fração dos documentos cobertos por outro tópico **"B"**. Assim, **"B"** torna-se pai de **"A"** e essa hierarquia é representada na árvore. Já a estratégia de **cobertura de termos** considera que um tópico **"C"** é sub-tópico de um tópico **"D"** caso o segundo possua todos os termos (palavras) empregadas na formação do primeiro. Por exemplo, considerando a extração dos tópicos *[presidente, brasil]* e *[presidente, brasil, visita]*, segundo a estratégia de cobertura de termos o segundo tópico será considerado um sub-tópico do primeiro, pois o primeiro seria uma generalização do segundo. Ao manipular a árvore, o usuário pode escolher qual das duas estratégias será empregada.

Admitir as duas estratégias para caracterizar sub-tópicos possibilita que o usuário escolha como deseja focar a análise da árvore. Caso escolha a estratégia por cobertura de termos, fo-

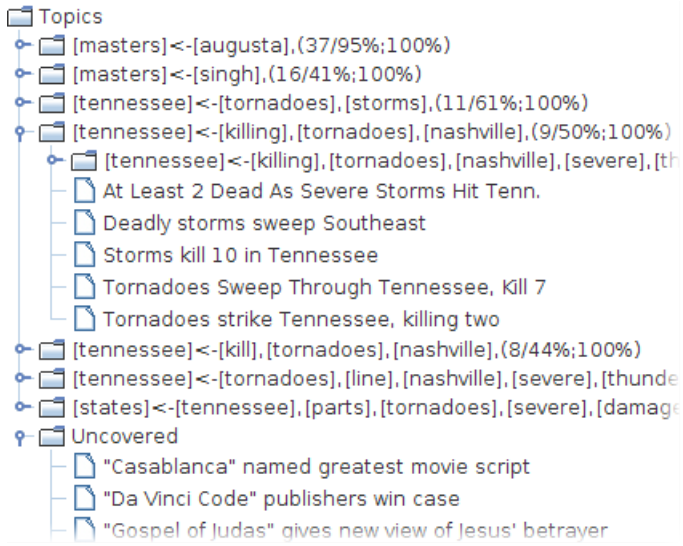


Figura 4.1: Árvore de tópicos. Os tópicos são representados como pastas (em azul) e os documentos contemplados pelos tópicos são representados como nós filhos (em branco).

cará sobre tópicos específicos, como *[brasil, presidente], [brasil, presidente, política], [brasil, presidente, acordo]*, por exemplo. Todos os tópicos que contenham os termos relacionados ao presidente do brasil são relacionados ao nó *[brasil, presidente]* e isso pode facilitar a análise desse assunto. Já a escolha por cobertura de documentos agrega diferentes tópicos, que não necessariamente compartilham termos em comum. Ao pesquisar um conjunto de documentos específico, coberto por um determinado tópico, o usuário pode obter os assuntos cobertos por sub-grupos ao analisar os sub-tópicos, e dessa forma descobrir outros caminhos para focar sua análise.

4.1.1 Interação

Ao selecionar um nó da árvore, os documentos cobertos pelo item são enfatizados no mapa de documentos. Dessa forma é possível verificar, por exemplo, que tópicos correspondem a agrupamentos de documentos bem definidos no mapa, quantos e quais documentos não são cobertos por nenhum tópico etc. O usuário pode selecionar diversos nós da árvore ao manter as teclas *"Shift"* ou *"Ctrl"* pressionadas ao clicar sobre o novo nó.

Foram desenvolvidos dois componentes com controles para manipular a árvore de tópicos: **um menu flutuante** e uma **barra de tarefas**. Os dois componentes são descritos a seguir.

Menu Flutuante

Após selecionar um conjunto de tópicos, o usuário pode manipulá-los por meio de um menu flutuante, disponível ao clicar com o botão direito do *mouse* sobre a árvore. Esse menu é exibido na Figura 4.2 e a funcionalidade de cada item é descrita a seguir:

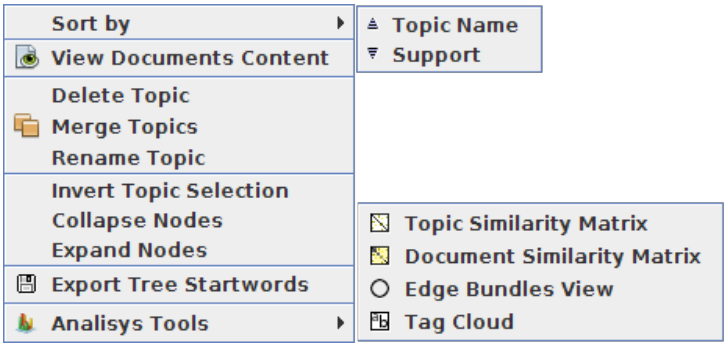


Figura 4.2: Menu flutuante da árvore de tópicos

- ***Sort by:*** Permite ordenar os tópicos da árvore. Um sub-menu possibilita ordenar por ordem decrescente de suporte (quantidade de documentos cobertos) ou ordem alfabética.
- ***View Documents Content:*** Exibe uma janela que mostra o conteúdo dos documentos cobertos pelos tópicos selecionados. As palavras que compõem os tópicos são enfatizadas nos textos (Figura 4.3).

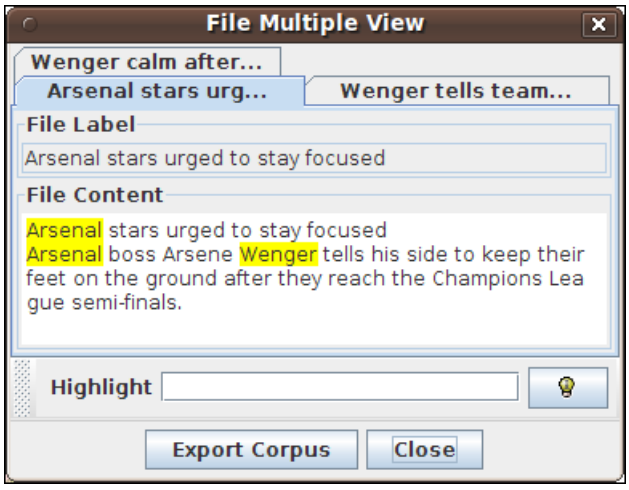


Figura 4.3: Janela de visualização do conteúdo de documentos

- ***Delete Topic:*** Exclui da árvore os tópicos selecionados, juntamente com os respectivos documentos do mapa, caso estes documentos não sejam cobertos por mais nenhum tópico.

O usuário pode optar por excluir apenas os tópicos – e manter no mapa os documentos – ao manter pressionada a tecla *"Shift"* quando escolher essa ação.

- ***Merge Topics:*** Une os tópicos selecionados e forma um novo tópico que contempla todos os documentos cobertos pelos tópicos selecionados. Os tópicos unidos são excluídos da árvore. O novo tópico é exibido na cor amarela (Figura 4.4) para alertar o usuário que trata-se de um tópico diferenciado. O nome do novo tópico é composto pela concatenação das palavras que formavam os tópicos selecionados.

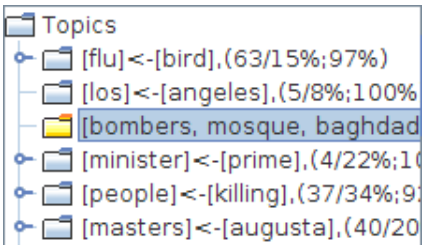


Figura 4.4: Tópico composto, enfatizado na árvore de tópicos

- ***Rename Topic:*** Permite alterar o nome do tópico selecionado. Ao escolher essa opção, um cursor é posicionado no início do rótulo do tópico, e o usuário pode então alterar o nome do tópico e pressionar a tecla *Enter* para concluir o processo ou a tecla *Esc* para cancelar.

- ***Invert Topic Selection:*** Inverte a seleção de tópicos, ou seja, ativa a seleção complementar dos nós da árvore. Por exemplo, se nenhum tópico estiver selecionado, ao escolher essa opção, todos os tópicos da árvore são selecionados.

- ***Collapse Nodes:*** Recolhe os nós expandidos da árvore, e dessa forma, apenas os tópicos do nível mais alto da árvore são exibidos.

- ***Expand Nodes:*** Expande os nós da árvore, exibindo todos os níveis simultaneamente.

- ***Export Tree Startwords:*** Possibilita salvar, em um documento de texto, as palavras contidas nos tópicos da árvore. Essas palavras podem posteriormente ser utilizadas como *bag-of-words* na etapa de pré-processamento da criação de um mapa de documentos.

- ***Analysis Tools:*** Apresenta uma lista de ferramentas de visualização, em um sub-menu, que atuam sobre os tópicos selecionados. O funcionamento de cada visualização será detalhado adiante.

A árvore é a principal ferramenta de interação com os tópicos. É por meio dela que o usuário é capaz de visualizar, editar e excluir os tópicos extraídos do mapa de documentos. Outras ferramentas, acessíveis por meio do item *Analysis Tools* do menu flutuante, permitem observar relações entre os tópicos e são descritas com detalhes mais adiante.

Barra de Tarefas

Outras ações sobre os tópicos são disponibilizadas por meio de uma barra de tarefas, localizada acima da árvore (Figura 4.5). Essa barra possui cinco botões, cujas finalidades são descritas a seguir:

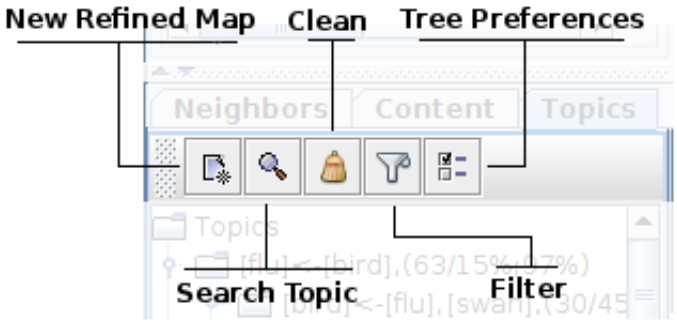


Figura 4.5: Barra de tarefas da árvore de tópicos

- ***New Refined Map:*** Possibilita criar um mapa de documentos refinado, tomando como entrada os documentos do mapa atual. O novo mapa vai incorporar apenas os documentos presentes no mapa atual para gerar a nova representação vetorial da coleção. Essa representação é então reintroduzida no processo de construção do novo mapa. Dessa forma os documentos excluídos anteriormente pelo usuário são descartados do processo e novas relações de similaridade, contendo apenas os documentos de interesse, poderão ser observadas.

  Ao clicar nesse botão, o usuário é direcionado ao assistente de projeção, no qual são definidos os parâmetros para a construção do novo mapa. Entre os parâmetros definidos pelo usuário estão a técnica de projeção multidimensional, parâmetros de pré-processamento etc. Assim, a ferramenta permite o teste de diversas configurações diferentes para a criação do novo mapa.

- ***Search Topic:*** Permite pesquisar por termos de interesse, destacando os nós que contiverem ao menos um dos termos digitados pelo usuário. Ao selecionar essa opção, o programa exibe uma janela com uma caixa de texto na qual o usuário pode escrever os

termos que deseja pesquisar na árvore de tópicos. Os termos devem ser separados por espaços. Se algum tópico possuir ao menos uma das palavras fornecidas pelo usuário, ele será destacado na árvore, assim como serão destacados no mapa os documentos cobertos por ele.

- ***Clean:*** Remove todos os tópicos da árvore, sem excluir os respectivos documentos do mapa.

- ***Filter:*** Permite selecionar um conjunto de regras para filtrar o conteúdo da árvore, e assim exibir apenas os tópicos que satisfaçam as regras definidas. Ao clicar nessa opção, é exibida uma janela com parâmetros para definir as regras do filtro (Figura 4.6).

  O primeiro item a ser definido é o campo (*Field*) da regra. Atualmente é possível escolher os atributos ***Support***, ***Confidence*** e ***Cross-Support***. O segundo item (*Condition*) especifica que condição será aplicada ao filtro, ou seja, como o filtro deve comparar o valor a ser informado no terceiro item (*Value*). As condições existentes são: ***Equals***, ***Different than***, ***Smaller than*** e ***Greater than***.

  É possível especificar regras adicionais para o filtro ao clicar no botão "**+**" localizado abaixo do seletor de regras. Caso o usuário queira remover regras, ele deve selecioná-las e clicar no botão "**-**".

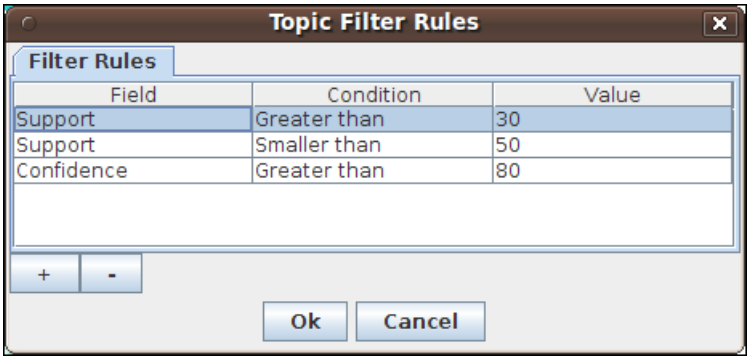


Figura 4.6: Janela com a seleção de regras para filtrar tópicos da árvore

- ***Tree Preferences:*** Apresenta um formulário (Figura 4.7) no qual é possível definir os parâmetros visuais da árvore de tópicos.

  Os controles exibidos na região ***Tree Nodes*** definem características visuais dos elementos da árvore. A caixa de seleção *Show Node Tooltip* permite escolher o uso de *tooltips*[2] sobre

[2]Caixas de texto que surgem na interface gráfica ao posicionar o mouse sobre determinados componentes. Nesse contexto, a sua utilidade é facilitar a visualização dos tópicos, pois determinados tópicos podem ser compostos por diversas palavras, o que pode dificultar a sua visualização conforme a resolução de tela do usuário

os nós da árvore ao posicionar o mouse sobre cada nó. A caixa se seleção *Show Document Nodes* possibilita exibir ou ocultar a exibição dos nós que representam os documentos contidos em cada tópico.

Os controles exibidos na região ***Sub Topics*** definem o como como a árvore interpreta a relação de hierarquia dos tópicos criados. Ou seja, definem como é analisada a relação que dita quais nós devem ser tópicos e quais nós devem ser sub-tópicos. As opções possíveis são *Document-Coverage Based* (cobertura de documentos) e *Topic-Term Based* (cobertura de termos).

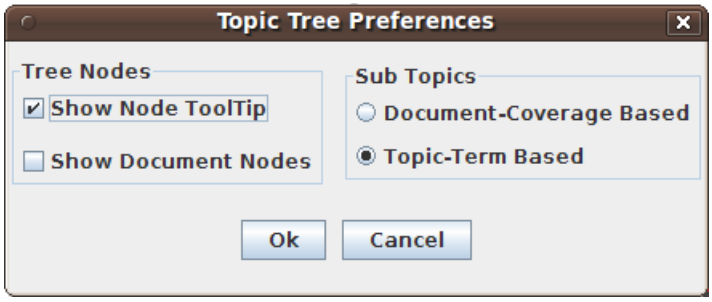


Figura 4.7: Janela com a seleção das preferências da árvore de tópicos

Os itens da árvore de tópicos fornecem um panorama dos assuntos identificados no mapa de documentos. No entanto, em uma busca exploratória, o conjunto de palavras que compõe cada tópico pode não ser suficiente para identificar que tópicos abordam assuntos relacionados. Apenas com o auxílio da árvore de tópicos não é possível analisar essas relações de forma eficiente, pois seria necessário ler os documentos de cada tópico individualmente.

Foram propostas então ferramentas adicionais para verificar a relação entre diversos grupos de tópicos simultaneamente, facilitando a exploração do mapa de documentos. Foram propostas e implementadas 4 ferramentas de análise: Matriz de Similaridade de Tópicos, Matriz de Similaridade de Documentos, *Edge Bundles* e *Tag Cloud*.

4.2 Matrizes de Similaridade

São representações gráficas que permitem visualizar relações entre os tópicos em um formato matricial, como ilustrado na Figura 4.8. Cada linha e coluna da matriz representa um tópico selecionado da árvore. As células, formadas pelo cruzamento das linhas e colunas, possuem cores que representam o relacionamento entre os respectivos tópicos. Quanto mais escura a cor da célula, maior a similaridade. Dessa forma a diagonal principal da matriz tem a cor preta, que indica a total similaridade entre tópicos (obviamente) iguais.

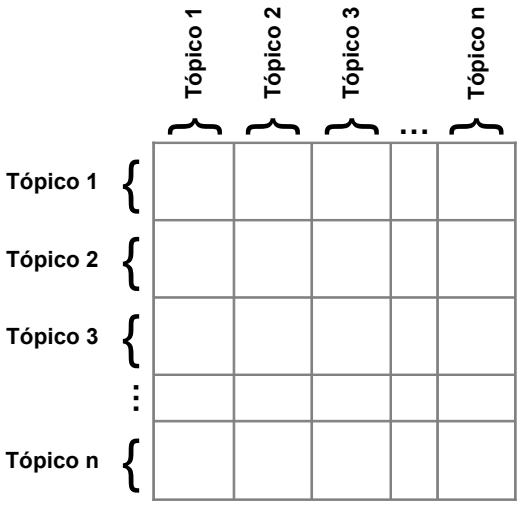


Figura 4.8: Matriz de similaridade

Uma consequência dessa representação é a redundância de informações, pois como os mesmos tópicos são representados tanto nas linhas quanto nas colunas, as relações entre os tópicos são repetidas acima e abaixo da diagonal principal de forma simétrica. A ordenação das linhas e colunas respeita a ordem dos itens selecionados na árvore de tópicos. A primeira linha/coluna da matriz representa o tópico selecionado na árvore mais próximo da raiz, e assim sucessivamente.

Um índice é posicionado em cada linha e coluna da matriz e auxilia a identificar que tópico está representado. Ao percorrer a matriz com o mouse, a janela de visualização exibe em uma caixa de texto a relação entre o índice e o respectivo tópico. Foram implementadas duas matrizes de similaridade: Matriz de Similaridade de Tópicos e Matriz de Similaridade de Documentos. Cada uma adota uma estratégia diferente para apoiar a análise do relacionamento entre os tópicos.

4.2.1 Matriz de Similaridade de Tópicos

O objetivo dessa matriz é informar o compartilhamento de documentos entre tópicos, pois tópicos distintos podem conter documentos em comum. A coloração das células dessa matriz informa a similaridade entre dois tópicos em relação à quantidade de documentos compartilhados. A Figura 4.9 exibe uma matriz de similaridade de tópicos com 10 dimensões, ou seja, relativa a 10 tópicos.

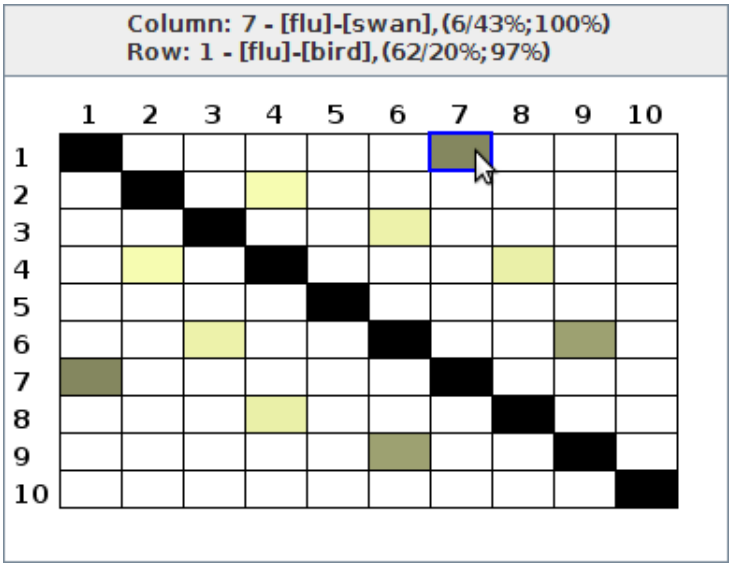


Figura 4.9: Matriz de similaridade de tópicos

A medida de similaridade adotada é o coeficiente de Jacquard. Assim, a similaridade entre os conjuntos de documentos DA e DB , suportados pelos tópicos A e B, é computada pela expressão 4.1. Ao mapear os valores da expressão para uma tabela de cores, o valor 1 corresponde à cor preta e o valor 0 à cor branca. Valores intermediários são mapeados em tons de amarelo, sendo que quanto mais próximo de 0, mais claro é o tom de amarelo e vice-versa.

$$jcoef(A, B) = \frac{|DA \cap DB|}{|DA \cup DB|} \tag{4.1}$$

Ao analisar essa visualização, pode-se verificar como conjuntos de documentos com elementos repetidos podem formar tópicos que empregam termos distintos. De acordo com a Figura 4.9, os tópicos 1 (cujo rótulo é *[flu, bird]*) e 7 (rótulo *[flu, swan]*) possuem coberturas bastante similares, ou seja, cobrem muitos documentos em comum, mas seus tópicos empregam palavras diferentes. Esse comportamento permite descobrir novas informações sobre um tópico específico. No exemplo, o tópico *[flu, bird]* trata de notícias sobre a gripe aviária, que ocorreu em 2006. Ao verificar a similaridade com o tópico *[flu, swan]*, o termo *swan* (cisne) indica que essa ave teve uma participação importante nas notícias que cobrem o evento da gripe. De fato, ao analisar os documentos que abordam esse assunto, verificou-se que um cisne foi encontrado morto na Escócia, vítima dessa doença, o que indica que a gripe aviária chegou àquele país.

4.2.2 Matriz de Similaridade de Documentos

Essa matriz objetiva exibir a similaridade entre os conjuntos de documentos que compõem os tópicos considerando, no entanto, a similaridade entre cada documento individualmente. Cada linha e coluna representa um documento, e a cor das células indica a similaridade medida pela distância dos cossenos (a mesma empregada na geração da projeção a partir da representação vetorial dos documentos). Assim, quanto mais escura a cor, maior a proximidade de conteúdo. Novamente as células que compõem a diagonal principal são exibidas na cor preta, pois representam a similaridade entre cada elemento consigo mesmo.

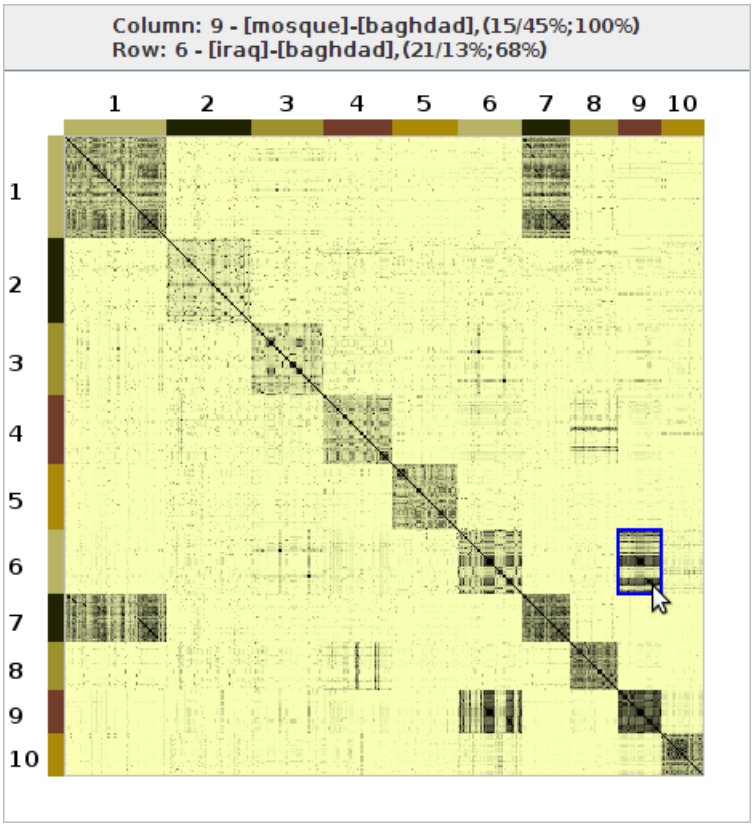


Figura 4.10: Matriz de similaridade de documentos

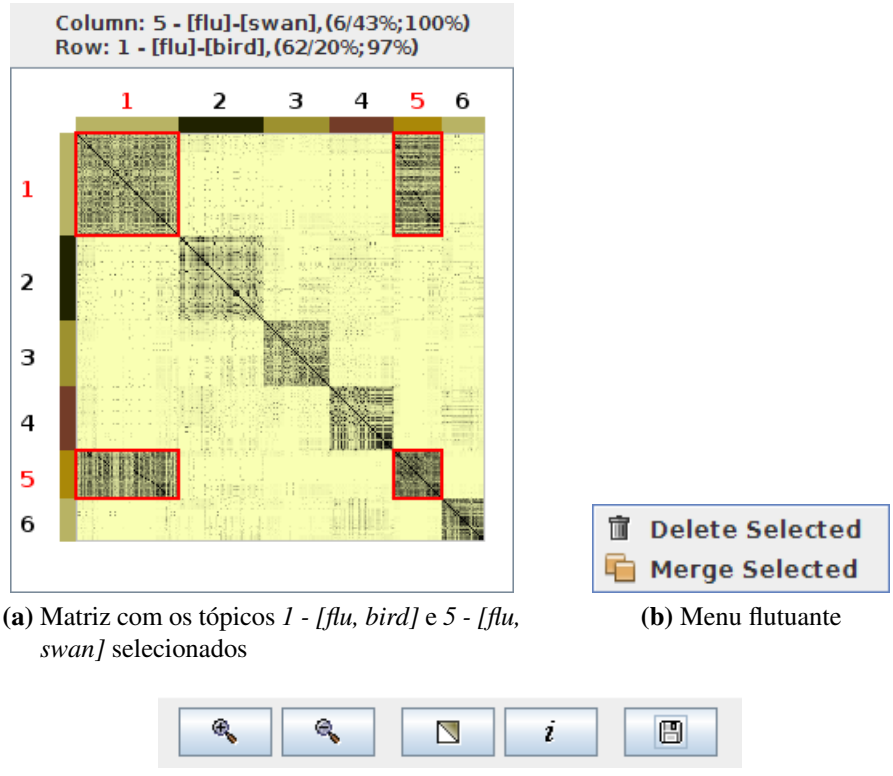
A matriz é exibida com bordas (Figura 4.10) que indicam a região correspondente a cada tópico. A alternância das cores indica somente os limites da região alocada na matriz para um determinado tópico. Tópicos distintos podem cobrir documentos repetidos, isto é, um documento pode ser associado a múltiplos tópicos, e portanto podem ocorrer repetições de elementos nas linhas e colunas da matriz.

Essa visualização permite verificar, também, a similaridade entre os documentos que compõem um mesmo tópico. Regiões escuras evidenciam que os documentos cobertos por um

determinado tópico são similares em conteúdo (de acordo com a medida de distância dos cossenos). Tópicos representados por regiões claras podem indicar que os seus documentos abordam assuntos desconexos, e foram agrupados no mesmo tópico apenas por satisfazerem os critérios do algoritmo de extração de tópicos.

4.2.3 Interação

A interação do usuário com as matrizes ocorre ao percorrer as células com o *mouse*, sendo possível, inclusive, realizar algumas operações presentes na árvore de tópicos. Conforme o usuário percorre o cursor sobre a matriz, a célula posicionada exatamente sob o cursor é enfatizada. Ao clicar sobre a célula enfatizada, a respectiva linha e coluna da matriz são selecionadas e a região correspondente é enfatizada (Figura 4.11a). Uma caixa de texto, localizada na região superior da janela da visualização, informa que tópicos são representados na linha e a coluna da célula localizada sob o cursor.



(a) Matriz com os tópicos *1 - [flu, bird]* e *5 - [flu, swan]* selecionados

(b) Menu flutuante

(c) Barra de ferramentas

Figura 4.11: Interação com a matriz de similaridade

Ao clicar com o botão direito do mouse sobre a matriz, é apresentado um menu flutuante com as opções de unir e excluir os tópicos selecionados (Figura 4.11b), de maneira similar ao menu existente na árvore de tópicos. Caso o usuário opte por realizar uma dessas ações, a matriz

de similaridade e a árvore são atualizadas para refletir a operação. A janela de visualização possui ainda uma barra de ferramentas (Figura 4.11c) com opções adicionais para interagir com a matriz, descritas a seguir:

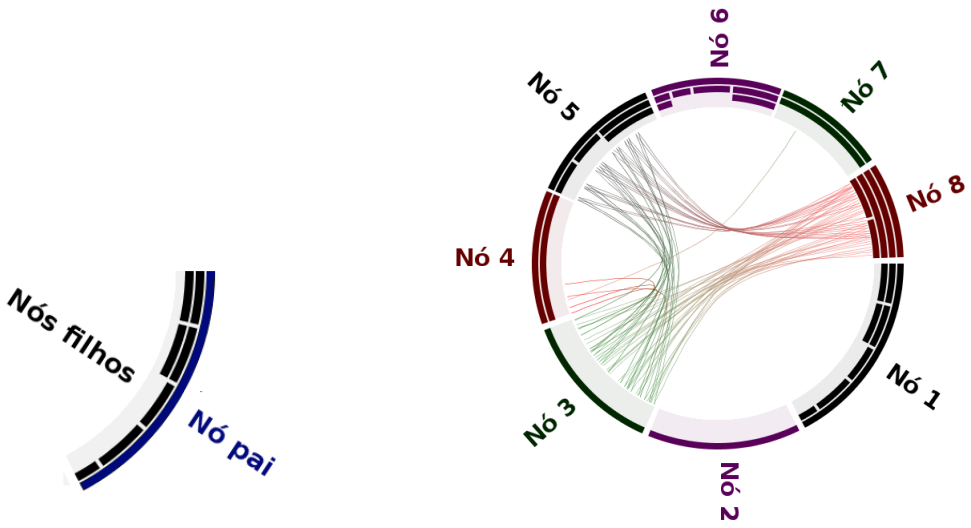
- ***Zoom In / Zoom Out:*** Possibilita ampliar ou reduzir o tamanho da imagem.
- ***Triangular View:*** Alterna entre as visões **completa** e **triangular** da matriz. Como a matriz é simétrica, a mesma informação é exibida acima e abaixo da diagonal principal. Ao selecionar a visão triangular, apenas as células posicionadas acima da diagonal principal serão exibidas.
- ***Change Labels Font:*** Permite alterar a fonte empregada nos índices da matriz. Ao pressionar esse botão é exibida uma janela na qual o usuário pode selecionar as fontes disponíveis no sistema, o estilo (normal, negrito, itálico, negrito + itálico) e o tamanho da fonte.
- ***Export Image:*** Permite salvar a imagem da matriz em um arquivo no disco.

4.3 Edge Bundles

A visualizaçao *Edge Bundles*, proposta por Holten (2006), tem o objetivo de representar simultaneamente informações de hierarquia e adjacência entre elementos de uma base de dados. A hierarquia é representada como uma árvore circular invertida, na qual os nós são apresentados como setores circulares (fatias de um círculo). Os nós de mais alta hierarquia são dispostos na borda mais externa da visualização, e os de hierarquia mais baixa na borda mais interna (Figura 4.12a). Setores circulares maiores representam nós com mais filhos. A raiz da árvore não é considerada nessa visualização.

Os nós-folha não são exibidos explicitamente, mas sua localização é evidenciada na representação das relações de adjacência. As relações de adjacência podem ser quaisquer relações investigadas nos dados analisados, como vizinhança, similaridade etc, e são apresentadas como linhas que ligam os nós-folha da árvore que obedecem à relação analisada. Dessa forma, a visualização é composta por setores circulares que possuem ligações entre si (Figura 4.12b). As cores utilizadas na visualização objetivam apenas diferenciar cada tópico e ajudar a identificar a quais tópicos cada ligação pertence, pois a cor de cada ligação é definida como uma transição entre as cores dos dois tópicos relacionados.

No contexto desse projeto, essa visualização foi empregada para visualizar relações entre os tópicos da árvore de tópicos. O propósito é evidenciar que tópicos compartilham documentos entre si, sendo essa relação representada pelas ligações. Assim, cada linha, representa um



(a) Representação da hierarquia **(b)** Representação das relações de adjacência

Figura 4.12: Visualização Edge Bundles: Representação de hierarquia e adjacência.

caminho que liga dois nós-folhas (documentos) da árvore de tópicos que cobrem documentos em comum. Essa visualização pode ser considerada um complemento à matriz de similaridade de tópicos, pois além de representar a mesma relação de uma forma diferente (e mais intuitiva), o *Edge Bundles* consegue representar também a hierarquia dos tópicos de forma compacta. Baseado nesse raciocínio, é possível verificar na Figura 4.12b que os tópicos (nós) 3, 5 e 8 compartilham diversos documentos e que os dois últimos possuem sub-tópicos que podem conter termos com informações relevantes ao usuário.

A idéia principal dessa visualização é "empacotar" as ligações de uma mesma hieraquia para evitar oclusão visual e possibilitar a visualização de diversos elementos simultaneamente. Essa idéia é análoga à de agrupar fios elétricos que percorrem as mesmas regiões em placas eletrônicas, ou cabos de rede agrupados até determinada região, para então se espalharem novamente. Para atingir esse objetivo, as ligações são representadas como *B-Splines* cuja localização dos pontos de controle é definida pela Equação 4.2.

$$P'_i = \beta.P_i + (1-\beta)(P_0 + \frac{i}{N-1}(P_{N-1} - P_0)), \tag{4.2}$$

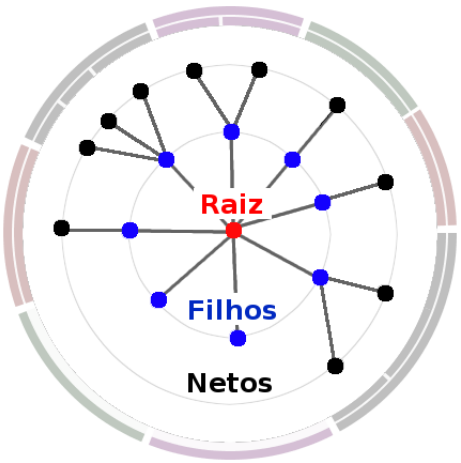
onde

N: número de pontos de controle,

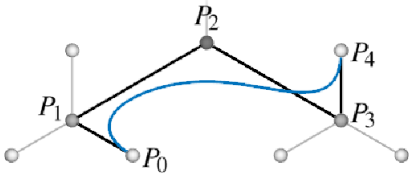
i: índice do ponto de controle, $i \in \{0, \ldots, N-1\}$,

β: força de agrupamento, $\beta \in [0, 1]$.

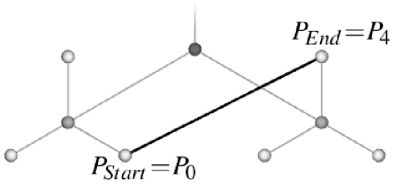
Os pontos de controle (P_i) são definidos como os nós de uma árvore circular localizada na região central da visualização (Figura 4.13a), árvore essa não exibida na visualização e cujo propósito é apenas guiar a localização dos pontos de controle. Cada linha usa como pontos de controle os nós pertencentes ao caminho da árvore que liga os respectivos nós-folha. O parâmetro β atua como um parâmetro de "força de agrupamento" e, conforme é alterado, os pontos de controle são re-definidos (P_i').



(a) Pontos de controle localizados na árvore circular



(b) *B-Spline* com $\beta = 1$



(c) *B-Spline* com $\beta = 0$

Figura 4.13: Pontos de controle das *B-Splines* na visualização *Edge Bundles*

Quanto maior o valor de β, mais próximo da localização dos nós são posicionados os pontos de controle (Figura 4.13b), e quanto menor o valor desse parâmetro, menos influência terá cada nó para formar a *B-Spline* (Figura 4.13c). Dessa forma, valores altos de β tendem a agrupar as ligações que compartilham uma mesma hierarquia, pois essas ligações compartilham os mesmos pontos de controle.

4.3.1 Interação

Ao posicionar o cursor do *mouse* sobre um tópico, é apresentada uma caixa de texto que contém informações sobre o tópico explorado, como o nome do tópico, sua hierarquia, o número de documentos cobertos e o número de conexões com outros tópicos. Ao clicar sobre um tópico, apenas as linhas que correspondem ao tópico selecionado permanecem visíveis, e assim pode-se verificar com mais clareza quais tópicos se relacionam com um tópico específico. É possível selecionar múltiplos tópicos se a tecla *Ctrl* estiver pressionada.

Se o usuário clicar com o botão direito do *mouse* sobre a visualização, um menu flutuante é exibido com as opções de **unir**, **excluir** ou **alterar a cor** dos tópicos selecionados. Se as operações de união ou exclusão forem acionadas, a visualização e a árvore de tópicos serão atualizadas para refletir as alterações.

A janela de visualização possui uma barra de tarefas (Figura 4.14) com controles adicionais para interação, descritos a seguir:

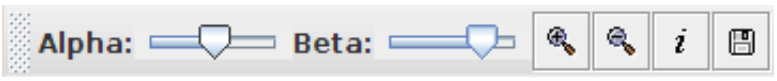


Figura 4.14: Barra de ferramentas da visualização Edge Bundles

- ***Alpha:*** Altera a opacidade das ligações entre os tópicos. Quanto maior o valor desse atributo, maior é a opacidade das linhas. A opacidade é afetada de forma diferente em cada relação, sendo que as linhas que possuem um comprimento maior são mais afetadas. A explicação para esse comportamento é o fato de que linhas maiores ocupam um espaço maior na visualização, dificultando a observação das linhas menores quando muitas relações são exibidas simultaneamente.
- ***Beta:*** Permite alterar o valor do parâmetro β utilizado na equação que define a localização dos pontos de controle das *B-Splines*. Quanto maior o valor de β, maior a "força de agrupamento" das relações que partem de uma mesma hierarquia.
- ***Zoom In / Zoom Out:*** Permite aumentar e diminuir o tamanho da imagem.
- ***Change Font:*** Permite alterar a fonte do rótulo do tópico.
- ***Export Image:*** Permite gravar uma imagem da visualização em um arquivo.

4.4 Tag Cloud

Muito popular em páginas na internet, a visualização *Tag Cloud* exibe um conjunto de palavras relacionadas, dispostas lado-a-lado em um painel. O tamanho de cada palavra está relacionado a alguma medida adotada, normalmente a frequência de ocorrência de cada palavra. No contexto desse projeto a *Tag Cloud* representa as palavras contidas nos textos que compõem o tópico analisado. A Figura 4.15 apresenta um exemplo dessa visualização para o tópico *[suicide, attack]*, extraído de um corpus de notícias.

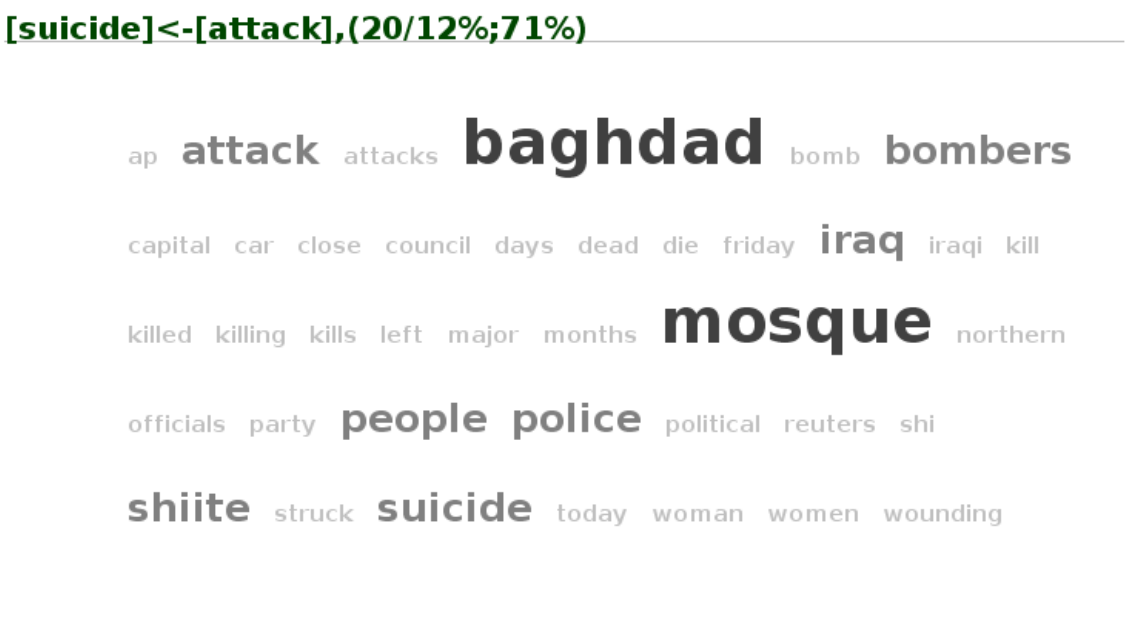


Figura 4.15: Visualização *Tag Cloud*

A visualização exibe inicialmente um cabeçalho que contém os rótulos dos tópicos selecionados. Em seguida exibe o conjunto de palavras extraídas dos documentos que compõem os tópicos, no qual o tamanho da palavra reflete a frequência do termo. Termos com maior frequência são exibidos usando uma fonte maior. Por fim, um sumário exibe a quantidade de palavras visualizadas e a frequência máxima e mínima do conjunto.

Os termos do exemplo sugerem que os documentos tratam de ataques suicidas e a *Tag Cloud* fornece mais detalhes, exibindo os termos utilizados pelos documentos de forma que aqueles com maior frequência (e talvez mais representativos) tenham maior destaque. Nota-se que as palavras *baghdad* e *mosque* são as mais citadas, pois são escritas com tamanho de fonte maior, seguidas pelas palavras *attack, bombers,iraq, shiite* e *suicide*.

Essa visualização possibilita verificar quais palavras foram consideradas na construção do modelo vetorial de documentos, pois são exibidas apenas as palavras que respeitam parâmetros como os cortes de Luhn, *stopwords*, etc. Esses parâmetros podem ser alterados na própria PEx, por meio do ajuste das preferências da ferramenta.

4.4.1 Interação

Ao posicionar o cursor do *mouse* sobre um termo, uma caixa de texto informa a quantidade de ocorrências do termo no tópico (considerando a contagem em todos os documentos cobertos). Se o usuário desejar visualizar mais de um tópico simultaneamente, a visualização exibirá apenas os termos comuns aos documentos de todos os tópicos. Assim, é possível obter pistas para identificar que assuntos os tópicos considerados abordam em comum.

A janela da visualização conta com um menu principal, localizado acima do painel de visualização, que possui os seguintes itens:

- ***Sort:*** Possibilita ordenar as palavras por ordem alfabética ou por ordem decrescente de frequência.
- ***Title Font:*** Permite alterar a fonte empregada no título do cabeçalho e do sumário.
- ***Preferences:*** Exibe uma nova janela onde é possível definir quais componentes devem aparecer na visualização: cabeçalho, tags e sumário.
- ***Export Image:*** Permite salvar a visualização como uma imagem no disco.

4.5 Considerações Finais

Esse capítulo apresentou o conjunto de ferramentas interativas desenvolvido com o propósito de auxiliar a manipulação e análise dos tópicos extraídos do mapa de documentos. Cada ferramenta possui diferentes características, cada uma com vantagens e desvantagens, e juntas oferecem ao usuário diferentes visões sobre o mesmo conjunto de tópicos e sua relação com os documentos do mapa.

A árvore de tópicos é a ferramenta principal, que dispõe em uma só visualização todos os tópicos extraídos do mapa de documentos, e possibilita manipulá-los e visualizar seus respectivos documentos. No entanto a árvore não oferece, sozinha, meios de verificar o relacionamento entre os tópicos. As matrizes de similaridade podem exibir um grande número de relações entre tópicos simultaneamente, permitindo verificar que tópicos compartilham documentos (matriz de similaridade de tópicos) ou que tópicos possuem documentos similares (matriz de similaridade de documentos). No entanto as matrizes não detalham as relações entre tópicos e sub-tópicos, como faz a visualização *Edge Bundles*. A *Edge Bundles*, por sua vez, não possui escalabilidade suficiente para exibir simultaneamente um conjunto muito grande de tópicos, pois nesse caso a interação é prejudicada pelo elevado custo computacional de renderização das *B-Splines*.

Por fim, a *Tag Cloud* possibilita identificar que termos são empregados nos documentos que compõem um determinado tópico, e com que frequência. Dessa forma pode-se analisar com mais detalhes os assuntos abordados por esses documentos, sem a necessidade de ler individualmente cada documento. Essas ferramentas podem ser empregadas simultaneamente, em diversas janelas, limitando-se apenas pelo tamanho da área de trabalho do usuário.

O próximo capítulo descreve como as funcionalidades desenvolvidas nesse projeto podem ser usadas para facilitar a resolução de tarefas de mineração visual. É descrito também como a usabilidade das funcionalidades foi avaliada, visando identificar os problemas enfrentados pelos usuários ao utilizar o *software*.

CAPÍTULO

5

Avaliação e Validação

Este capítulo descreve como as funcionalidades desenvolvidas e integradas ao *software Projection Explorer* foram avaliadas e validadas. A avaliação de interfaces humano-computador tem o objetivo de verificar que problemas são encontrados ao utilizar a interface, o que possibilita sanar ou minimizar as dificuldades e proporcionar uma experiência melhor e mais produtiva ao usuário. A validação é feita para verificar se os resultados obtidos utilizando a nova ferramenta coincidem com os resultados esperados, conhecidos anteriormente a partir de outros estudos.

No caso específico desse projeto, a avaliação foi empregada para identificar como melhorar o uso da interface, identificar problemas enfrentados pelo usuário e sugerir melhorias. A validação é mais complexa, pois os "resultados esperados" não estão disponíveis para comparar com os resultados obtidos nesse trabalho. Assim, a estratégia empregada na validação foi identificar tarefas de usuário típicas no cenário de exploração de documentos e verificar se a ferramenta foi útil e contribuiu para a execução dessas tarefas.

Para a avaliação foi utilizado um método de inspeção de usabilidade, conduzido por alunos do grupo de pesquisa familiares com ferramenta. O método adotado é a avaliação heurística, no qual um conjunto de recomendações de usabilidade é estabelecido e as interfaces são observadas para verificar se seguem essas recomendações. Eventuais problemas de usabilidade são relatados pelos avaliadores e devem ser sanados em versões posteriores do *software*.

Para a validação, foram conduzidos estudos de caso empregando um corpus de artigos científicos e um corpus de notícias *on-line*. O estudo no corpus de artigos científicos foi conduzido

por um usuário especialista no domínio. O objetivo desse estudo é encontrar o conjunto de documentos que abordam pesquisas sobre determinada sub-área de interesse. O estudo de caso empregado sobre o corpus de notícias objetiva verificar quais são os assuntos mais frequentes abordados pelos documentos analisados no período observado e foi conduzido por um usuário familiar com o *software*.

5.1 Avaliação Heurística

Um dos focos deste trabalho foi a elaboração de interfaces gráficas interativas. Portanto, é necessário dedicar um esforço no sentido de avaliar as interfaces e conhecer eventuais problemas encontrados pelos usuários.

Segundo Rocha e Baranauskas (2003), a avaliação de interfaces tem três grandes objetivos:

1. Avaliar a funcionalidade do sistema
2. Avaliar o efeito da interface junto ao usuário
3. Identificar problemas específicos do sistema

O **primeiro** item objetiva verificar aspectos de desempenho, ou seja, se a funcionalidade do sistema se enquadra aos requisitos da tarefa do usuário, de modo que o sistema seja projetado para que o usuário utilize os recursos eficientemente. O **segundo** item objetiva verificar a usabilidade da interface. Usabilidade é um termo amplo, que reflete o quão "usável" é uma interface, ao considerar fatores como facilidade de aprendizado, baixa ocorrência de erros e eficiência. Já o **terceiro** item, intimamente relacionado com os dois anteriores, objetiva verificar que tipo de problemas as interfaces avaliadas apresentam, quando empregadas no contexto do sistema.

Um método de avaliação de usabilidade amplamente utilizado é a avaliação heurística, proposta por Nielsen (1993). Nesse método, fatores da interface são inspecionados por usuários do sistema, que seguem um conjunto de heurísticas (princípios) estabelecidos a priori e tentam identificar se a interface segue essas heurísticas. Aos problemas encontrados são atribuídos pesos (graus de severidade), arbitrariamente estipulados pelos avaliadores. O resultado final desse processo é um documento contendo os problemas encontrados, que deve ser usado posteriormente pela equipe de desenvolvimento em futuras versões do *software*. Esse método possui como vantagens a fácil aplicação, rapidez e não exigir um grande número de avaliadores.

5.1.1 Desenvolvimento da Avaliação

Nesse contexto foi avaliada a usabilidade da ferramenta *Projection Explorer*, incorporando as funcionalidades desenvolvidas neste trabalho. Foram utilizandas as heurísticas propostas por Rocha e Baranauskas (2003) (ilustradas na Tabela 5.1), que representam uma versão revisada das heurísticas definidas por Nielsen (1993).

Nesse processo foram empregados cinco avaliadores, alunos de pós-graduação da USP. Individualmente foi oferecido um treinamento básico sobre a ferramenta a cada avaliador, no qual o desenvolvedor ilustrou as diversas opções do *software*, em que cenários são empregadas, as vantagens e desvantagens de cada funcionalidade.

Após sanar as possíveis dúvidas sobre o uso do *software*, foram apresentadas as 10 heurísticas presentes na Tabela 5.1. Notou-se que a maioria dos avaliadores teve dificuldades em entender as heurísticas baseando-se apenas nas descrições da tabela. Foi necessário ilustrar exemplos de violação de cada uma em aplicações familiares ao avaliador (como editores de texto).

Para cada problema de usabilidade deve ser atribuído um grau de severidade, que indica o impacto do problema, sob a ótica do avaliador. Essa informação também pode ser interpretada como sendo a prioridade que os desenvolvedores devem alocar para solucionar tal problema. Os graus de severidade podem assumir os valores “Baixo”, “Médio”, “Grave”. O grau **Baixo** contempla problemas “cosméticos”, ou seja, que não afetam significativamente a usabilidade da interface e devem ser corrigidos apenas se os desenvolvedores tiverem tempo disponível. O grau **Médio** indica problemas que afetam a usabilidade da interface, e devem ser corrigidos em versões posteriores do *software*. Por fim, o grau **Grave** contempla problemas que afetam sensivelmente a usabilidade do sistema, e devem ser tratados com prioridade pela equipe de desenvolvedores.

Por fim, foi entregue a cada avaliador um formulário de avaliação, contendo os campos “problema”, “grau de severidade”, “heurística violada” e “descrição”. O campo **problema** apenas enumera o erro encontrado. O campo **grau de severidade** pode assumir os valores B (Baixo), M (Médio) e G (Grave). O campo **heurística violada** deve assumir um dos valores correspondentes ao índice de uma heurística listada na Tabela 5.1 (variando de 1 a 10). O campo **descrição** deve conter detalhes sobre o problema encontrado. A Tabela 5.2 ilustra um exemplo de formulário de avaliação. Como cada avaliador pode ter diferentes considerações sobre o impacto de um mesmo problema, os valores de impacto podem ser conflitantes nos formulários dos avaliadores. Nesse caso foi adotado que o grau de severidade do problema será considerado aquele definido pela maioria dos avaliadores.

Tabela 5.1: Versão revisada das heurísticas. (Rocha e Baranauskas, 2003)

1. **Visibilidade do status do sistema**: o sistema precisa manter os usuários informados sobre o que está acontecendo, fornecendo feedback adequado dentro de um tempo razoável.
2. **Compatibilidade do sistema com o mundo real**: o sistema precisa falar a linguagem do usuário, com palavras, frases e conceitos familiares ao usuário, ao invés de termos orientados ao sistema.
3. **Controle do usuário e liberdade**: usuários frequentemente escolhem, por engano, funções do sistema e precisam ter claras saídas de emergência para sair do estado indesejado sem ter que percorrer um extenso diálogo. Prover funções *undo* e *redo*.
4. **Consistência e padrões**: usuários não precisam adivinhar que diferentes palavras, situações ou ações significam a mesma coisa. Seguir convenções de plataforma computacional.
5. **Prevenção de erros**: melhor que uma boa mensagem de erro é um design cuidadoso o qual previne o erro antes que ele aconteça.
6. **Reconhecimento ao invés de relembrança**: tornar objetos, ações e opções visíveis. O usuário não deve ter que lembrar informação de uma para outra parte do diálogo. Instruções para uso do sistema devem estar visíveis e facilmente recuperáveis quando necessário.
7. **Flexibilidade e eficiência de uso**: usuários novatos se tornam peritos com o uso. Prover aceleradores de forma a aumentar a velocidade da interação. Permitir a usuários experientes "cortar caminho" em ações frequentes.
8. **Estética e design minimalista**: diálogos não devem conter informação irrelevante ou raramente necessária. Qualquer unidade de informação extra no diálogo irá competir com unidades relevantes e diminuir sua visibilidade relativa.
9. **Ajudar os usuários a reconhecer, diagnosticar e corrigir erros**: mensagens de erro devem ser expressas em linguagem clara (sem códigos) indicando precisamente o problema e construtivamente sugerindo uma solução.
10. **Help e documentação**: embora seja melhor um sistema que possa ser usado sem documentação, é necessário prover *help* e documentação. Essas informações devem ser fáceis de encontrar, focalizadas na tarefa do usuário e não muito extensas.

Vale ressaltar que esse método de avaliação é bastante flexível, e não exige que os avaliadores reservem um tempo pré-estabelecido para executar as suas tarefas. Naturalmente quanto

mais tempo o avaliador reservar para utilizar a interface, maior a chance de encontrar erros. Nesse processo um dos avaliadores necessitou da presença do desenvolvedor para conduzir sua avaliação, pois não possuía experiência em *softwares* de visualização e, por se tratar de uma ferramenta de pesquisa, o *software* possui muitos termos específicos de especialistas no domínio. Apesar dessa dificuldade, com o auxílio do desenvolvedor, o avaliador pôde sanar suas dúvidas enquanto explorava as funcionalidades da ferramenta. Essa flexibilidade do método permite que avaliadores com diferentes níveis de experiência possam conduzir as tarefas, aumentando a chance de encontrar problemas de usabilidade tanto para usuários leigos quanto avançados.

Tabela 5.2: Formulário de avaliação

Data: _____	**Projeto:** _____	**Versão:** _____
Avaliador: _____		

Problema	Grau de Severidade	Heurística Violada	Descrição
1			
2			
⋮	⋮	⋮	⋮
n			

5.1.2 Resultados

As fichas dos avaliadores foram comparadas, e os problemas de usabilidade foram agregados em um único formulário, cada um com seu respectivo grau de severidade. Em diversos casos, foram atribuídos graus de severidades distintos a problemas similares encontrados por diferentes avaliadores. Esse conflito foi solucionado atribuindo-se o grau definido pelo maior número de avaliadores. Os problemas de usabilidade encontrados nessa avaliação são enumerados na Tabela 5.3.

Tabela 5.3: Resultados da Avaliação Heurística

1. **Grau de Severidade:** G **Heurística Violada**: 3
   Não é possível refazer ou desfazer as ações (exclusão, edição, união etc) sobre os tópicos em nenhuma das visualizações.

2. **Grau de Severidade:** G **Heurística Violada**: 10
   O sistema não possui documentação, apenas "hints" nos controles visuais.

3. **Grau de Severidade:** M **Heurística Violada**: 2
   Tópicos apresentam números e porcentagens que podem confundir o usuário com porcentagens de similaridade.

4. **Grau de Severidade:** M **Heurística Violada**: 5
   O sistema deveria pedir a confirmação do usuário antes de excluir tópicos.

5. **Grau de Severidade:** M **Heurística Violada**: 6
   As ações "Excluir apenas Tópicos" ou "Excluir Tópicos e Documentos" deveriam ser acessadas também por itens de menu distintos, não obrigando o usuário unicamente a lembrar teclas de atalho.

6. **Grau de Severidade:** M **Heurística Violada**: 6
   Os botões da barra lateral direita do sistema são pouco intuitivos. É díficil associar o desenho com a funçao.

7. **Grau de Severidade:** M **Heurística Violada**: 7
   Não existe uma opção para permitir que o usuário salve as suas preferências sobre cada visualização.

8. **Grau de Severidade:** M **Heurística Violada**: 7
   Não há teclas de atalho para executar as ações sobre árvore de tópicos e demais visualizações.

9. **Grau de Severidade:** B **Heurística Violada**: 1
   Enquanto são geradas as visualizações, não é informado o que o programa está fazendo.

10. **Grau de Severidade:** B **Heurística Violada**: 1
    Ao carregar o arquivo da visualização, o sistema não exibe o que está fazendo.

11. **Grau de Severidade:** B **Heurística Violada**: 4
    Apenas alguns ítens do menu da árvore de tópicos possuem ícones. Padronizar.

12. **Grau de Severidade:** B **Heurística Violada**: 7
    A visualização *Edge Bundles* poderia ser melhor visualizada se pudesse ser rotacionada, pois facilitaria a leitura dos tópicos.

5.2 Estudos de Caso

Os estudos de caso propostos têm a finalidade de validar as funcionalidades propostas e apresentar cenários em que elas possam ser utilizadas para explorar mapas de documentos. A validação é obtida ao especificar um conjunto de tarefas ao usuário e verificar se, apoiado pelas funcionalidades desenvolvidas, ele foi capaz de concluir as tarefas.

Ao explorar os tópicos extraídos do mapa de documentos, o usuário pode optar por excluir tópicos – e seus respectivos documentos no mapa – que julgar não interessantes em sua pesquisa. Esse processo interativo possibilita criar novos mapas de documentos, contendo apenas grupos de documentos de interesse, e dessa forma refinar o processo de análise.

5.2.1 Classificação de Artigos Científicos

Este estudo foi conduzido por um pesquisador do ICMC, especialista em Aprendizado de Máquina e Mineração de Dados, sobre um conjunto de 574 artigos científicos (contendo apenas o título, resumo e referências). Os artigos foram extraídos de periódicos que abordam as áreas de *Case-Based Reasoning* (CBR), *Inductive Logic Programming* (ILP) e *Information Retrieval* (IR), sendo cada artigo rotulado manualmente em uma dessas três sub-áreas de Aprendizado de Máquina. Essa rotulação foi conduzida com base apenas na fonte do artigo, não sendo empregada, portanto, uma inspeção cuidadosa do conteúdo dos artigos. Especificamente os artigos da sub-área *IR* foram rotulados com base em uma busca na *web* utilizando esses termos.

O objetivo desse estudo é identificar os documentos que abordam temas dentro da mesma sub-área por meio da análise dos tópicos extraídos do mapa de documentos, para evitar o processo exaustivo de ler cada documento para fazer a classificação. O resultado então é comparado ao da rotulação manual.

Inicialmente o conjunto de artigos passa pela etapa de pré-processamento, pela qual é obtida a representação vetorial da coleção (*bag-of-words*). Os cortes de Luhn inferior e superior empregados nessa etapa foram 13 e 2745, respectivamente. A dissimilaridade entre os documentos foi calculada utilizando a distância dos cossenos sobre a representação vetorial, que resultou em uma matriz de distâncias em que cada linha e coluna representam um documento, e o valor da célula a distância entre os respectivos documentos. Essa matriz é então utilizada como entrada por uma técnica de projeção multidimensional, que possibilita representar os documentos como pontos no plano. A técnica de projeção utilizada nesse estudo foi a *ProjClus*, e o mapa resultante é ilustrado na Figura 5.1. A cor de cada círculo reflete a rotulação manual, realizada previamente: artigos classificados como *CBR* são representados em vermelho, *ILP* em verde e

IR em azul. No entanto, propositalmente essa coloração não foi utilizada no estudo, pois o objetivo é rotular os documentos apoiado apenas pelos tópicos extraídos e pelas funcionalidades desenvolvidas.

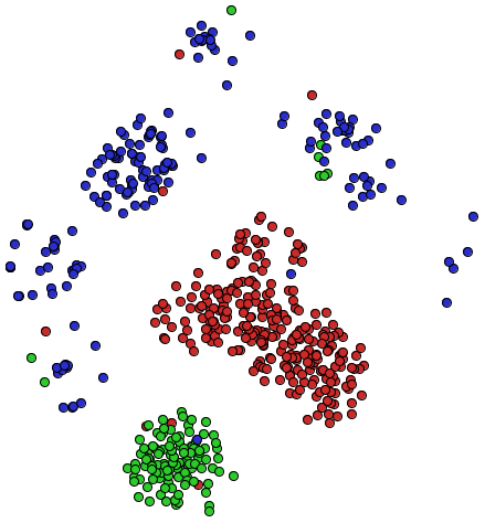
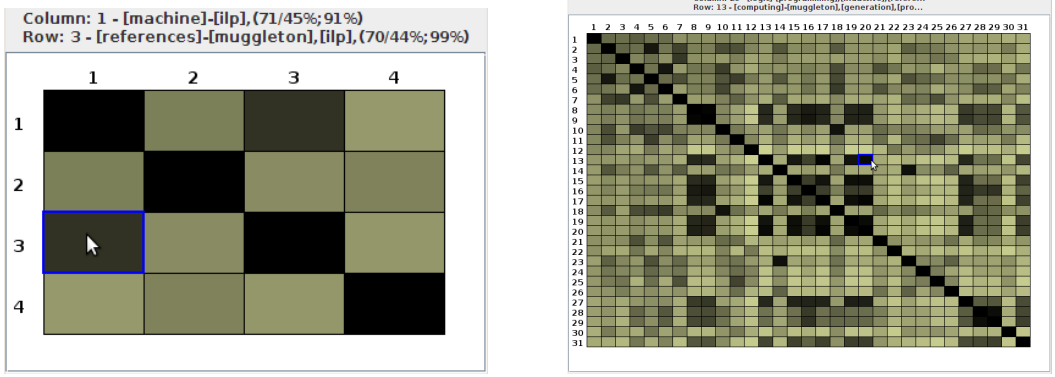


Figura 5.1: Mapa de similaridade de artigos científicos, relacionados a três áreas: *Inductive Logic Programming* (verde), *Case Based Reasoning* (vermelho) e *Information Retrieval* (azul)

A partir do mapa de documentos, os tópicos são extraídos empregando o algoritmo LWR, tendo sida adotada a estratégia de seleção por agrupamentos com múltiplo reinício. Os mesmos parâmetros para os cortes de Luhn empregados na etapa de pré-processamento da construção do mapa foram utilizados no algoritmo para extrair os tópicos. O número de agrupamentos inicial e máximo escolhidos foram 7 e 85, respectivamente, com incremento de 39 a cada iteração. Esses valores foram escolhidos após experimentos anteriores indicarem que valores próximos de $\sqrt{tamanho\ do\ corpus}$ apresentam bons resultados. O processo finalizou após três iterações, e resultou na extração de 311 tópicos que cobrem 570 artigos (4 artigos não foram associados a nenhum tópico).

O especialista, familiar com as áreas contempladas pelo corpus, explorou o mapa de documentos e sua árvore de tópicos, com foco nos tópicos que chamaram sua atenção. Ao seguir a evidência apresentada pela matriz de similaridade de tópicos (Figura 5.2a), ele percebeu que os tópicos *[machine, ilp]*, *[learning, logic, machine, programming, inductive, references, raedt]*, *[references, muggleton, ilp]*, e *[logic, programming, inductive, references, dzeroski]* cobrem conjuntos de documentos similares. Estes tópicos foram unificados, e o tópico resultante renomeado como *[ilp muggleton dzeroski raed]*.



(a) Matriz de Similaridade com 4 tópicos da área *Inductive Logic Programming*

(b) Matriz de Similaridade com 31 tópicos relacionados aos autores Dzeroski, Muggleton e Raedt

Figura 5.2: Matrizes de similaridade de tópicos sobre artigos científicos

O pesquisador sabia que Muggleton, Dzeroski e Raedt são autores importantes da área de *Inductive Logic Programming* e concentrou-se em procurar tópicos cujos termos empregam os nomes desses autores. Após utilizar a ferramenta de busca da árvore de tópicos, e informar os nomes dos autores procurados, foram localizados e selecionados 31 tópicos. O pesquisador utilizou novamente a matriz para visualizar a similaridade entre esses tópicos (Figura 5.2b). A matriz indica que todos os 31 tópicos possuem documentos em comum. Após analisar alguns documentos individualmente, o usuário unificou os 31 tópicos.

O tópico resultante, que cobre 118 documentos, foi renomeado para *[Inductive Logic Programming]*. Desses 118 documentos, 117 pertencem de fato ao grupo rotulado como ILP, que contém 119 documentos no corpus. Na Figura 5.3, os documentos cobertos pelo tópico *[Inductive Logic Programming]* são enfatizados com bordas mais grossas e maior opacidade. É possível observar uma grande semelhança entre os documentos destacados e aqueles classificados manualmente (em verde).

5.2.2 Seleção de Notícias

Esse estudo foi conduzido em um corpus de 2.684 artigos de notícias, coletados dos servidores RSS[1] das agências *Associated Press, BBC, CNN* e *Reuters*, durante dois dias em Abril de 2006. O objetivo do estudo é encontrar os assuntos mais abordados por essas agências, no período considerado, com o apoio dos recursos visuais e dos tópicos extraídos do mapa de documentos.

Iniciamente o corpus de notícias foi pré-processado, e a matriz de distâncias resultante foi utilizada como entrada na técnica de projeção multidimensional LSP (*Least Square Projection*)

[1] *Really Simple Syndication* – Arquivos textuais com pequenos resumos de notícias

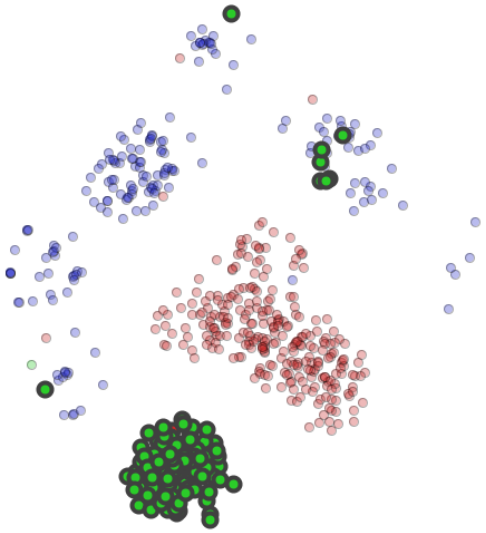


Figura 5.3: Artigos do tópico *[Inductive Logic Programming]* enfatizados no mapa de similaridade

para representar os documentos como pontos no plano. Tópicos foram extraídos do mapa utilizando o algoritmo LWR, que empregou a mesma estratégia adotada no estudo anterior. O número de agrupamentos inicial e máximo escolhidos foram 17 e 189, respectivamente, com incremento de 86 a cada iteração. O processo resultou na extração de 525 tópicos, com a cobertura de 1402 documentos.

O usuário iniciou a análise pela ordenação da árvore de tópicos, por ordem decrescente de suporte. Essa é uma opção disponível no menu flutuante da árvore. O resultado é ilustrado na Figura 5.4.

Ao verificar os tópicos exibidos na árvore, o usuário observou que o tópico que cobre mais documentos (maior suporte) refere-se à gripe aviária – *[flu, bird]* –, seguido pelo tópico *[los, angeles]*, *[minister, prime]* e assim por diante. Foi utilizada então a visualização *Edge Bundles* sobre os 20 tópicos de maior suporte, para verificar a hierarquia dos tópicos e o compartilhamento de documentos entre eles. Essa visualização (Figura 5.5) inicialmente exibe todas as relações simultaneamente, mas permite também focar a análise sobre tópicos específicos. Ao clicar no tópico de maior suporte *[flu, bird]*, que cobre 66 documentos, apenas as relações deste tópico com os demais são exibidas, e o usuário verificou que este tópico compartilha todos os 30 documentos cobertos pelo tópico *[flu, swan]*. Essa semelhança indica que estes dois tópicos podem ser unidos. Dessa forma a quantidade de tópicos analisada é gradualmente reduzida, o que facilita a análise do conjunto, agrega mais termos significativos ao novo tópico (gerado pela união dos outros) e elimina a redundância de termos repetidos nesses tópicos.

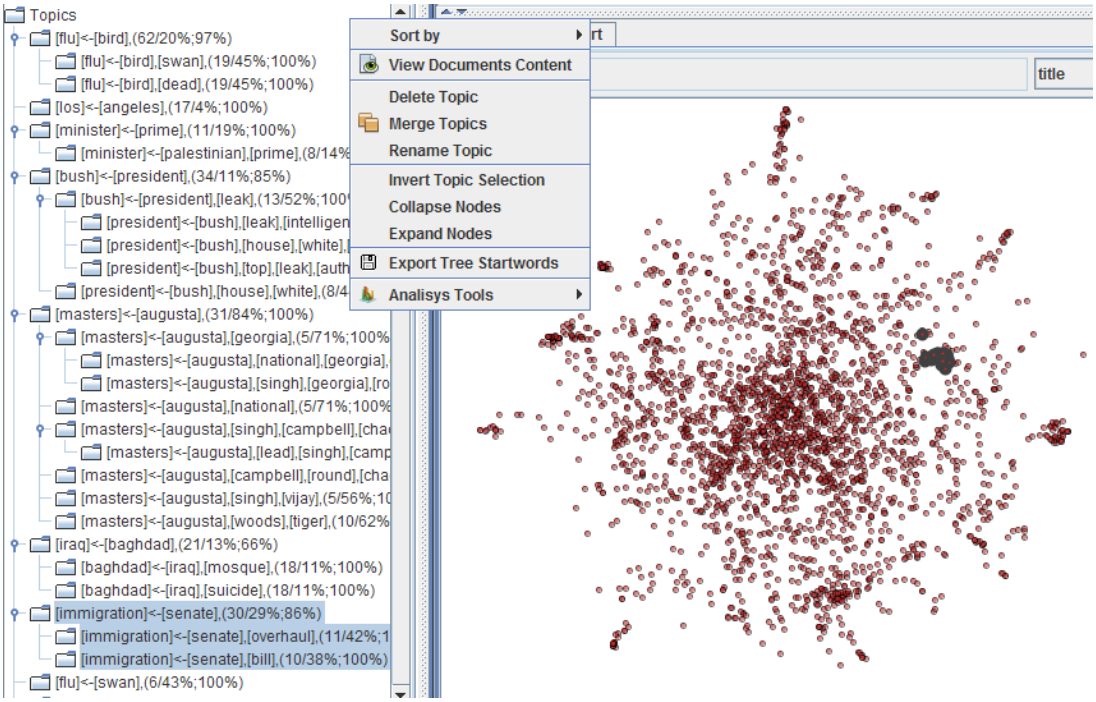


Figura 5.4: Mapa de similaridade de notícias *on-line* e árvore de tópicos associada

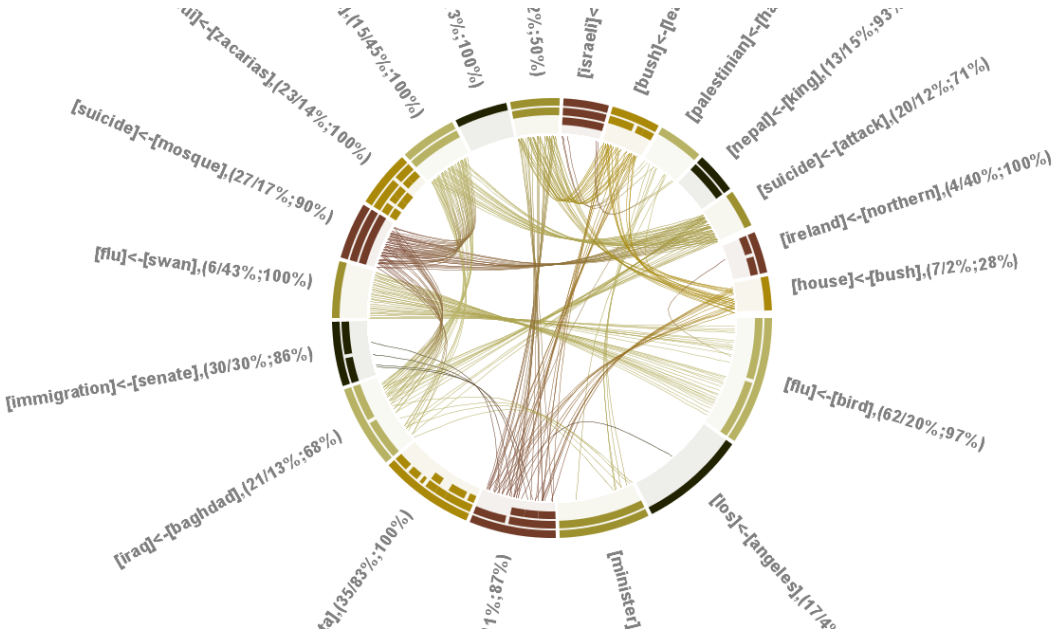


Figura 5.5: Visualização *Edge Bundles* sobre 20 tópicos de notícias

Ao posicionar o cursor sobre um tópico na visualização, alguns detalhes são exibidos em uma caixa de texto: o nome do tópico, a quantidade de documentos cobertos por ele e a quantidade de documentos compartilhados com outros tópicos. Dessa forma a hierarquia de tópicos

também é analisada. O tópico *[flu, bird]*, por exemplo, possui dois sub-tópicos: *[flu, bird, dead]* e *[flu, bird, swan]*, que cobrem 28 e 30 documentos respectivamente. Assim, pode-se partir de uma análise mais ampla (tópicos) para uma análise mais detalhada (sub-tópicos). Os sub-tópicos também são considerados no processo de união de tópicos. Ao unir dois ou mais tópicos, todas as palavras compartilhadas pelos tópicos e sub-tópicos selecionados são consideradas para formar o novo tópico (o seu nome e cobertura são atualizados). Os tópicos *[flu, bird]* e *[flu, swan]*, considerados relevantes, foram unidos e o tópico resultante foi nomeado *[bird, flu, swan, dead]*.

Outra constatação é a grande abrangência do tópico *[bush, president]*. Esse é o que mais se relaciona com outros , pois possui ligações com outros seis tópicos. Outro ponto digno de atenção é a alta concentração de ligações entre quatro tópicos: *[suicide, attack], [iraq, baghdad], [suicide, mosque]* e *[mosque, baghdad]*. Uma rápida inspeção individual de alguns documentos revela que tratam-se de tópicos relacionados a ataques suicidas na capital do Iraque. Os quatro tópicos foram unidos, o que resultou na criação do tópico *[mosque, baghdad, iraq, attack, iraq, suicide]*.

A matriz de similaridade de documentos (Figura 5.6) foi empregada para visualizar os 16 tópicos restantes. Regiões escuras representam relações de alta similaridade entre tópicos. As células escuras na diagonal principal indicam que os tópicos cobrem documentos altamente similares entre si. Portanto, células muito claras na diagonal principal podem evidenciar tópicos não representativos, que abordam documentos desconexos. Esse é o caso dos tópicos *[los, angeles]* (2^o), *[minister, prime]* (4^o), *[bush, president]* (5^o) e *[san, francisco]* (9^o).

Ao analisar alguns documentos cobertos pelo tópico *[los angeles]*, o usuário notou que o tópico trata de notícias desconexas, que relatam eventos diversos relativos à cidade de Los Angeles. Ao aplicar a visualização *Tag Cloud* ao tópico, observa-se que as palavras Los e Angeles são as que aparecem com maior frequência nos documentos, pois são exibidas usando fonte maior. As outras palavras são exibidas com uma fonte pequena, o que indica que os documentos não abordam assuntos semelhantes e foram associadas ao tópico apenas por apresentarem em comum o nome da cidade. Comportamento semelhante foi observado nos tópicos *[minister, prime], [bush, president]* e *[san, francisco]* e os quatro foram excluídos.

Ainda apoiado pela matriz de similaridade de documentos, o usuário percebeu que os tópicos *[bush, iraq], [bush, leak]* e *[bush, house]* são similares em termos de cobertura de documentos. Ao inspecionar alguns documentos, constatou que noticiam um vazamento de informações sobre a guerra do Iraque. Os três tópicos foram unidos no tópico *[information, cheney, leak, classified, bush, house, iraq, intelligence]*.

Ao final do processo, apenas nove tópicos restaram na árvore (dos 20 iniciais). Os demais foram excluídos, juntamente com os respectivos documentos. Os tópicos restantes cobrem

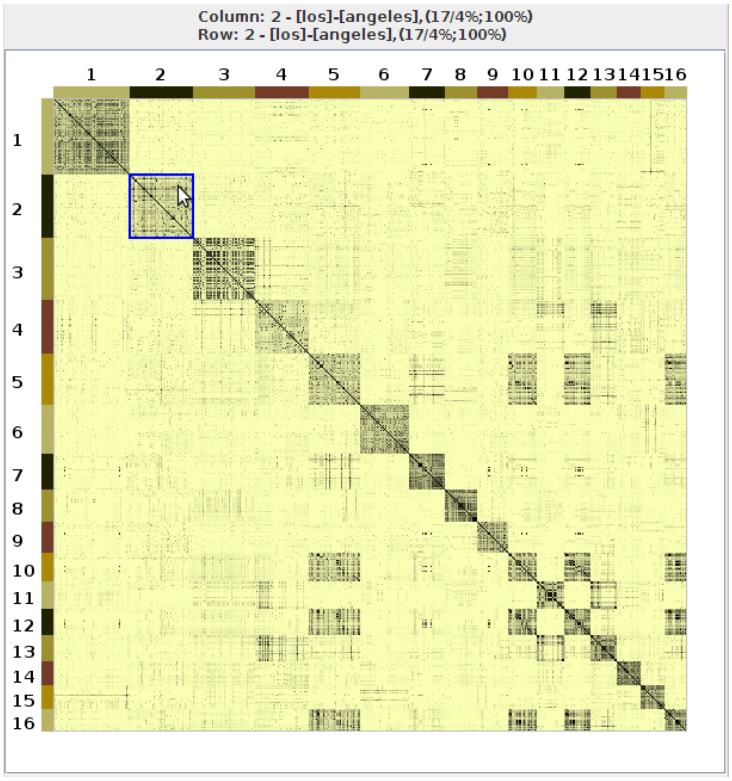


Figura 5.6: Matriz de similaridade de documentos exibindo 16 tópicos de notícias

Figura 5.7: Visualização *Tag Cloud* sobre o tópico *[los, angeles]*

330 documentos e um novo mapa de notícias foi construído a partir desses documentos. O resultado é observado na Figura 5.8. As cores dos artigos representam os diferentes tópicos

restantes na árvore de tópicos do mapa original. Os itens representados na cor laranja, por exemplo, localizados no centro do mapa, são documentos pertencentes ao tópico que aborda ataques suicidas no Iraque. Ao analisar o novo mapa, pode-se verificar que sete tópicos formam agrupamentos bem definidos. Outros dois tópicos formaram cinco agrupamentos.

O tópico que aborda notícias sobre conflitos entre palestinos e israelenses, representado em azul escuro, se espalhou e formou três agrupamentos. Após inspecionar alguns documentos contidos em cada um desses três agrupamentos, contatou-se que o agrupamento central trata especificamente de notícias sobre conflitos na região da Faixa de Gaza e sobre o recém-eleito governo palestino formado por membros do grupo Hamas. O agrupamento localizado mais à direita trata da morte de um jornalista, supostamente assassinado por um soldado israelense. O agrupamento localizado na região inferior esquerda trata da suspensão de ajuda financeira por parte da União Européia e dos Estados Unidos ao governo palestino, após as eleições palestinas.

O outro tópico, representado em verde e que trata de notícias sobre a Irlanda do Norte, formou dois agrupamentos. Um deles, localizado na região inferior direita do mapa, trata de questões políticas da Irlanda do Norte. O outro, localizado na região superior, está próximo ao agrupamento que trata de notícias relativas ao torneio de golfe *Master's Augusta* (em vermelho). Ao inspecionar esses documentos, verificou-se que, de fato, as notícias abordam o desempanho de atletas ingleses e irlandeses na competição de golfe.

Esse estudo confirma que o mapa refinado permite identificar diferentes relações entre as notícias, que no mapa anterior eram ofuscadas pela grande quantidade de elementos supostamente não interessantes ao usuário. O novo mapa, ao considerar a interferência do usuário na escolha dos documentos, trabalha com um modelo de representação vetorial menor. Assim, a técnica de projeção é menos sucetível à interferências dos documentos que o usuário não considera importantes, o que pode favorecer a formação de agrupamentos mais relevantes.

O processo de refinamento pode ser executado tantas vezes quanto o usuário julgar necessário. Novos tópicos podem ser extraídos, pois os diferentes agrupamentos podem favorecer a identificação de novos tópicos e sub-tópicos, o que pode revelar informações mais detalhadas sobre os focos da pesquisa do usuário. Esse estudo é ilustrado também em um vídeo, disponível no endereço http://infoserver.lcad.icmc.usp.br/infovis2/TopicPEx.

5.3 Considerações Finais

A avaliação de usabilidade do sistema evidenciou diversos problemas, apontados pela violação das heurísticas empregadas. Esses problemas devem ser considerados nas próximas versões do

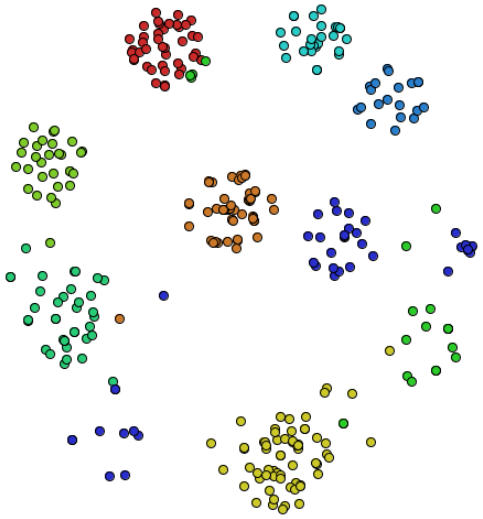


Figura 5.8: Mapa refinado de documentos de notícias

software, para sanar ou ao menos reduzir as dificuldades encontradas. No entanto, cabe ressaltar que a avaliação foi conduzida por um número reduzido de avaliadores, por limitações de tempo.

O estudo de caso que envolve classificação de artigos científicos, conduzido por um pesquisador especialista no domínio, empregou apenas três visualizações: árvore de tópicos, matriz de similaridade e mapa de documentos. Esse estudo atingiu bons resultados pois contou com o conhecimento prévio do pesquisador, que sabia os nomes de autores importantes de uma determinada área de pesquisa. O processo encerrou ao classificar apenas uma área, mas poderia ser expandido para classificar todo o corpus.

O outro estudo de caso ilustra uma pesquisa exploratória que contou com o apoio de todas as visualizações desenvolvidas neste trabalho. O pesquisador conduz o estudo com base nos tópicos que julga mais relevantes, e emprega as visualizações para obter diferentes visões do mesmo conjunto de dados. Ao final do processo, um novo mapa de documentos evidencia novas relações a serem exploradas.

CAPÍTULO

6

Conclusões

Neste capítulo são apresentadas as principais contribuições deste trabalho para a exploração e refinamento de mapas de documentos, suas limitações e uma discussão sobre possíveis trabalhos futuros.

Com a popularização da internet, cada dia mais informações são disponibilizadas na rede – na forma de e-mails, blogs, relatórios etc – em volume tão grande que inviabiliza a pesquisa pela leitura sequencial de cada texto. Os mapas de documentos oferecem alternativas para identificar grupos de documentos com conteúdo similar, e permitem visualizar um "apanhado geral" das áreas contempladas pelo domínio analisado.

O principal objetivo deste trabalho foi propor e desenvolver estratégias que permitam interagir com mapas de documentos, de forma a oferecer meios de facilitar a investigação e descoberta de informações em tarefas de exploração de coleções de documentos. Estendendo o trabalho de extração de tópicos de Pinho (2009) foram criadas, adaptadas e integradas cinco visualizações e interações associadas que possibilitam identificar diferentes relações entre os tópicos abordados pelos documentos de um mapa de similaridade.

6.1 Contribuições

Tópicos sumarizam conjuntos de documentos do mapa, e as visualizações permitem verificar quais tópicos são similares, considerando aspectos diferentes em cada visualização. Dessa

forma um pesquisador pode focar em tópicos pertinentes aos interesses de sua pesquisa e descartar os tópicos e documentos não relevantes. Os documentos remanescentes podem ser re-introduzidos no processo de construção do mapa de documentos, refinando o mapa de acordo com os critérios definidos pelo pesquisador. O mapa refinado pode revelar novas relações interessantes entre os documentos, favorecendo o surgimento de novos focos de pesquisa. Esse processo interativo pode ser empregado tantas vezes quanto o usuário julgar necessário.

Este trabalho introduz visualizações interativas e funcionalidades associadas à plataforma de visualização *Projection Explorer*, agregando novas funcionalidades para a exploração de mapas de documentos. Essas visualizações são coordenadas e poderiam ser estendidas para analisar relações entre outros tipos de dados, como mapas de imagens, volumes, redes sociais, disponíveis em outras versões do *software*. A árvore de tópicos permite que o pesquisador manipule e visualize um grande conjunto de tópicos simultaneamente, em uma área visual independente do mapa de documentos, eliminando a oclusão visual ocasionada pela presença de um grande número de tópicos, ilustrada na Figura 3.3. As matrizes de similaridade possibilitam visualizar diferentes relações entre tópicos, e possibilitam que o pesquisador tenha mais informações para julgar quais tópicos estão relacionados aos temas que lhe interessam. O *Edge Bundles*, além de oferecer uma visualização alternativa à matriz de similaridade de tópicos, permite visualizar e explorar as relações de hierarquia da árvore de tópicos. A visualização *Tag Cloud* destaca as palavras que ocorrem com maior frequência nos documentos cobertos por um determinado tópico, o que facilita a observação dos assuntos mais abordados do conjunto de documentos.

Foi conduzido um estudo sobre as dificuldades de interação enfrentadas pelos usuários do sistema. Ao entender as dificuldades enfrentadas pelo usuário, as novas versões do sistema podem ser adaptadas para suprir as deficiências de usabilidade, possibilitando um aprendizado mais rápido, mais conforto e eficiência no uso. Essa característica é particularmente importante em visualização, por se tratar de uma área multidisciplinar, em que os usuários muitas vezes não são especialistas em determinados domínios.

Os estudos de caso apresentados ilustram potenciais usos do sistema, e podem ser reproduzidos em outros corpora para auxiliar tarefas que demandam um esforço excessivo como a exploração de grandes volumes de documentos. O trabalho implementa o processo introduzido por Pinho (2009) para criar mapas refinados, ou seja, mapas de documentos construídos a partir de sub-conjuntos de documentos de um mapa pre-existente. Nesse processo podem ser especificados novos parâmetros para a construção do mapa, auxiliando o usuário a testar como novas configurações alteram o posicionamento dos documentos no mapa e evidenciam novas relações entre esses documentos.

6.2 Limitações

A avaliação de usabilidade realizada foi limitada por restrições de tempo e quantidade de avaliadores. Como a técnica de avaliação empregada não estipula um limite mínimo de tempo a ser empregado na avaliação, cada avaliador alocou uma pequena parte do seu tempo livre para conduzir a avaliação. Boa parte dos avaliadores se queixou de não dispor de tempo livre para conduzir detalhadamente sua avaliação. Outro ponto importante seria empregar avaliadores de diferentes áreas, não necessariamente das ciências exatas, para verificar as dificuldades enfrentadas por esses usuários.

As visualizações interativas podem ser aprimoradas. A *Edge Bundles*, por exemplo, pode mapear um atributo à cor das arestas que ligam dois tópicos, e essa característica atualmente não é utilizada. O mesmo pode ser aplicado à *Tag Cloud*, que não mapeia cor como um atributo de seus termos.

Outro ponto a ser atacado é a grande quantidade de tópicos extraídos do corpus pela estratégia de extração automática. Apesar de ajudarem a entender os assuntos principais abordados no corpus, a quantidade de tópicos extraída pode ser tão grande que a sua análise pode demandar um grande esforço. Como ponto de partida, poderia ser adotada a estratégia de definir, no início do processo de extração de tópicos, que palavras são relevantes à pesquisa do usuário. Assim, o algoritmo LWR poderia considerar válidos apenas os tópicos que contenham ao menos uma dessas palavras.

6.3 Trabalhos Futuros

Devem ser conduzidas novas avaliações de usabilidade sobre o *software*. Técnicas como o percurso cognitivo, ou o teste de usabilidade, abordam o problema sob novas perspectivas e, aliadas aos resultados já obtidos, devem fornecer mais informações sobre os problemas enfrentados pelos usuários. Ao tratar esses problemas, o *software* torna-se acessível a mais pessoas.

Outros estudos de caso devem ser conduzidos para entender melhor as necessidades do usuário e identificar padrões de uso da plataforma de visualização. É possível, com o auxílio das funcionalidades desenvolvidas, explorar um grande volume de documentos, e identificar documentos que podem ser utilizados como conjuntos de treinamentos para classificadores de texto. Esse cenário ainda não foi explorado nas investigações realizadas até o momento.

Seria interessante comparar diferentes técnicas de extração de tópicos com o algoritmo LWR_2. Esse estudo pode comparar aspectos como a cobertura dos tópicos sobre os documentos do corpus, quantidade de tópicos extraídos, escalabilidade, tempo de execução etc.

Os parâmetros envolvidos no processo de extração de tópicos também devem ser investigados. Os cortes de Luhn, *stopwords* e suporte mínimo, por exemplo, são definidos empiricamente no início do processo, sendo ajustados conforme o usuário verifica que a quantidade de tópicos extraídos e sua cobertura ainda não atingiu um nível aceitável.

Referências Bibliográficas

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, New York, NY, USA: ACM, 1993, p. 207–216.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

ANDREWS, K.; KIENREICH, W.; SABOL, V.; BECKER, J.; DROSCHL, G.; KAPPE, F.; GRANITZER, M.; AUER, P.; TOCHTERMANN, K. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, v. 1, n. 3/4, p. 166–181, 2002.

BERGER, M.; NONATO, L. G.; PASCUCCI, V.; SILVA, C. T. Fiedler trees for multiscale surface analysis. *Computers & Graphics*, v. 34, n. 3, p. 272 – 281, shape Modelling International (SMI) Conference 2010, 2010.
Disponível em `http://www.sciencedirect.com/science/article/B6TYG-4YN5PH1-3/2/c3eafe0d9ae761c7124d5af3694abb2d`

BÖRNER, K.; CHEN, C.; BOYACK, K. Visualizing Knowledge Domains. *Annual Review of Information Science and Technology(ARIST)*, v. 37, p. 179–255, 2003.

CARD, S. K.; MACKINLAY, J.; SHNEIDERMAN, B. *Readings in information visualization: Using vision to think*. Morgan Kaufmann, 1999.

CHEN, C. *Information visualization: Beyond the horizon*. Springer, 2004.

CUADROS-VARGAS, A.; LIZIER, M.; MINGHIM, R.; NONATO, L. Generating segmented quality meshes from images. *Journal of Mathematical Imaging and Vision*, v. 33, p. 11–23, 10.1007/s10851-008-0105-2, 2009.
Disponível em `http://dx.doi.org/10.1007/s10851-008-0105-2`

ELER, D.; NAKAZAKI, M.; PAULOVICH, F.; SANTOS, D.; ANDERY, G.; OLIVEIRA, M.; BATISTA NETO, J.; MINGHIM, R. Visual analysis of image collections. *The Visual Computer*, v. 25, p. 923–937, 10.1007/s00371-009-0368-7, 2009a.
Disponível em http://dx.doi.org/10.1007/s00371-009-0368-7

ELER, D.; PAULOVICH, F.; DE OLIVEIRA, M.; MINGHIM, R. Topic-based coordination for visual analysis of evolving document collections. 2009b, p. 149 –155.

FALOUTSOS, C.; LIN, K.-I. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets, p. 163–174. 1995.

FELDMAN, R.; SANGER, J. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, MA, USA: Cambridge University Press, 2006.

FELIZARDO, K.; MARTINS, R. M.; MALDONADO, J. C.; LOPES, A. A.; MINGHIM, R. Content based visual mining of document collections using ontologies. *Proc of the 2nd International Workshop on Web and Text Intelligence*, p. 1–8, 2009.

FERREIRA, V.; KUROKAWA, F.; OISHI, C.; KAIBARA, M.; CASTELO, A.; CUMINATO, J. Evaluation of a bounded high order upwind scheme for 3d incompressible free surface flow computations. *Mathematics and Computers in Simulation*, v. 79, n. 6, p. 1895 – 1914, applied and Computational MathematicsSelected Papers of the Sixth PanAmerican WorkshopJuly 23-28, 2006, Huatulco-Oaxaca, Mexico, 2009.
Disponível em http://www.sciencedirect.com/science/article/B6V0T-4NMWRB9-1/2/51e2b6c5cfb3f79a66709323352cce57

GARCIA, D. E. Cosine similarity and term weight tutorial. *http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html. Último acesso em 05/02/2009.*, 2006.
Disponível em http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html

HOLTEN, D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, v. 12, n. 5, p. 741–748, 2006.

HOTHO, A.; NURNBERGER, A.; PAASS, G. A brief survey of text mining. 2005.

LOPES, A. A.; PINHO, R.; PAULOVICH, F. V.; MINGHIM, R. Visual text mining using association rules. *Comput. Graph.*, v. 31, n. 3, p. 316–326, 2007.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, p. 159–165, 1958.

MINGHIM, R.; LEVKOWITZ, H. Visual mining of text collections. In: *Tutorial Notes 11*, Eurographics, 2007.

MORAES, M. L.; MAKI, R. M.; PAULOVICH, F. V.; RODRIGUES FILHO, U. P.; DE OLIVEIRA, M. C. F.; RIUL, A.; DE SOUZA, N. C.; FERREIRA, M.; GOMES, H. L.; OLIVEIRA, O. N. Strategies to optimize biosensors based on impedance spectroscopy to detect phytic acid using layer-by-layer films. *Analytical Chemistry*, v. 82, n. 8, p. 3239–3246, pMID: 20334387, 2010.
Disponível em `http://pubs.acs.org/doi/abs/10.1021/ac902949h`

NAVARRO, D.; LEE, M. Spatial visualisation of document similarity. *Proceedings of the Defence Human Factors Special Interest Group Meeting*, p. 39–44, 2001.

NIELSEN, J. *Usability engineering*. Academic Press, 1993.

OLIVEIRA, M. C. F.; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 9, p. 378 – 394, 2003.

PAULOVICH, F. V. *Mapeamento de dados multi-dimensionais - integrando mineração e visualização*. Tese de Doutoramento, Instituto de Ciências Matemáticas e de Computação, 2008.

PAULOVICH, F. V.; MINGHIM, R. Text map explorer: a tool to create and explore document maps. In: *Proocedings of the 10th International Conference on Information Visualisation - IV*, London - UK: IEEE CS Press, 2006, p. 245–251.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 3, p. 564–575, 2008.

PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The projection explorer: A flexible tool for projection-based multidimensional visualization. In: *Proc. XX Brazilian Symposium on Computer Graphics and Image Processing SIBGRAPI 2007*, 2007, p. 27–36.

PINHO, R.; DE OLIVEIRA, M.; DE ANDRADE LOPES, A. An incremental space to visualize dynamic data sets. *Multimedia Tools and Applications*, v. 50, p. 533–562, 10.1007/s11042-010-0483-5, 2010.
Disponível em `http://dx.doi.org/10.1007/s11042-010-0483-5`

PINHO, R.; DE OLIVEIRA, M. C. F.; DE A. LOPES, A. Incremental board: a grid-based space for visualizing dynamic data sets. In: *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*, New York, NY, USA: ACM, 2009, p. 1757–1764.

PINHO, R. D. *Espaço incremental para a mineração visual de conjuntos dinâmicos de documentos*. Tese de Doutoramento, Instituto de Ciências Máticas e de Computação - USP, 2009.

PINHO, R. D.; SILVA, R. R. O.; LOPES, A. A.; MINGHIM, R.; DE OLIVEIRA, M. C. F. User-centered visual exploration of document collections with rule-based topic mining, em preparação.

ROCHA, H. V.; BARANAUSKAS, M. C. C. *Design e avaliação de interfaces humano-computador*. NIED, 224 p., 2003.

SALTON, G.; BUCKLEY, C. *Term weighting approaches in automatic text retrieval*. Relatório Técnico, Ithaca, NY, USA, 1987.

SIQUEIRA, M.; XU, D.; GALLIER, J.; NONATO, L. G.; MORERA, D. M.; VELHO, L. A new construction of smooth surfaces from triangle meshes using parametric pseudo-manifolds. *Computers & Graphics*, v. 33, n. 3, p. 331 – 340, iEEE International Conference on Shape Modelling and Applications 2009, 2009.
Disponível em `http://www.sciencedirect.com/science/article/B6TYG-4VTVR0F-7/2/d1b06d771429c1a1fc26b4445fec1d2c`

SKUPIN, A. A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications*, v. 22, p. 50–58, 2002.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, v. 2, n. 4, p. 218–231, 2003.

TELLES, G. P.; MINGHIM, R.; PAULOVICH, F. V. Normalized compression distance for visual analysis of document collections. *Computers & Graphics*, v. 31, n. 3, p. 327–337, 2007.

WONG, P. C.; THOMAS, J. Visual analytics. *IEEE Computer Graphics and Applications*, v. 24, p. 20–21, 2004.

XU, R.; WUNSCH, D. *Clustering*. Wiley-IEEE Press, 2008.