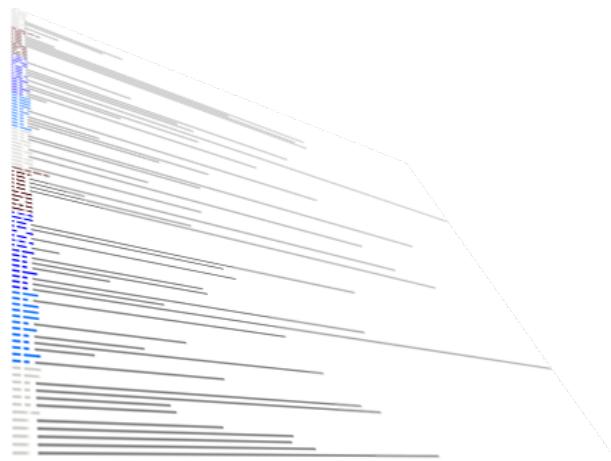


Relatório final

(PIBIC 2012-13)

Visualização Interativa de Dados para Suporte à Atividade de Análise Qualitativa ‘Conteúdo-Temporal’ de Fóruns de Discussão



Discente: João Vítor Antunes Ribeiro

Orientador: Prof. Dr. Milton Hirokazu Shimabukuro

Colaboradora: Profa. Dra. Renata Portela Rinaldi

Presidente Prudente - SP

Outubro de 2013



AGRADECIMENTOS

Agradecimentos à Pró-Reitoria de Pesquisa (PROPe) da UNESP e à Faculdade de Ciências e Tecnologia (FCT - Campus de Presidente Prudente), pela oportunidade de desenvolvimento do trabalho científico, aos professores Milton e Renata, pelo suporte teórico e conceitual, aos amigos Alisson e Leonardo, pelo auxílio técnico, ao CPIDES e seus integrantes, pelo ambiente de desenvolvimento, aos amigos, pela descontração e troca de experiências, à minha família, pelo apoio incondicional e à todos aqueles que ainda mantém vivo o sentimento da curiosidade científica.



SUMÁRIO

| | | |
|----------|---|-----------|
| 1 | Introdução | 1 |
| 1.1 | Objetivo | 2 |
| 1.2 | Cronograma | 3 |
| 2 | Mineração | 5 |
| 2.1 | Introdução | 5 |
| 2.2 | Mineração de Texto | 6 |
| 2.2.1 | Etapas | 7 |
| 2.3 | Tipos de dados | 10 |
| 3 | Visualização | 11 |
| 3.1 | Introdução | 11 |
| 3.2 | Tipos de Visualização | 12 |
| 3.2.1 | Visualização Científica | 12 |
| 3.2.2 | Visualização de Informação | 13 |
| 3.3 | Técnicas de Visualização | 14 |
| 3.3.1 | Visualização Orientada a Pixel | 14 |
| 4 | Aplicação para Análise Qualitativa ‘Conteúdo-Temporal’ | 17 |
| 4.1 | Introdução | 17 |
| 4.2 | Desenvolvimento | 19 |
| 4.3 | Funcionamento | 24 |

| | |
|--------------|----|
| 5 Conclusões | 27 |
| 6 Anexo | 31 |

LISTA DE FIGURAS

| | | |
|-----|---|----|
| 1.1 | Cronograma do desenvolvimento do projeto. | 4 |
| 2.1 | Etapas de KDD segundo Fayyad; adaptado de [Fayyad et al., 1996]. | 7 |
| 2.2 | Exemplo de visualização de dados em 2D e 3D; extraído de [van Zudilova-Seinstra et al., 2009]. | 9 |
| 3.1 | Exemplo de Visualização Científica, Disponível em: < http://www.fab.mil.br/portal/capa/index.php?datan=03/08/2009&page=mostra_notimpol >, Acessado em: 6/12/12. | 13 |
| 3.2 | Exemplo de Visualização de Informação, Disponível em: < http://truthy.indiana.edu/memedetail?id=324&resmin=45&theme_id=4 >, Acessado em: 6/12/12. | 13 |
| 3.3 | Visualização Orientada a Pixel por meio de Gráfico de Barras; adaptado de [Ankerst, 2001]. | 15 |
| 3.4 | Visualização Orientada a Pixel por meio de Segmento de Círculos; extraído de [Ankerst, 2001]. | 15 |
| 4.1 | Representação visual de texto por pontos, extraída do programa desenvolvido por Pacheco Jr. (2011). | 18 |
| 4.2 | Primeiro modelo de Visualização temporal proposto. | 20 |
| 4.3 | Esquema da primeira Visualização temporal desenvolvida. | 20 |
| 4.4 | Modelo de visualização temporal final: múltiplos painéis em funcionamento. | 22 |
| 4.5 | Visualização temporal baseada em calendário. | 23 |
| 4.6 | Destaque de <i>startwords</i> nos textos integrais. | 24 |

| | | |
|-----|-------------------------------|----|
| 4.7 | Menu principal do BoardWords. | 25 |
|-----|-------------------------------|----|



RESUMO

Nesse relatório são abordados os temas: Mineração de Dados, Mineração de Texto e Visualização de Informação. O objetivo é estudar algumas técnicas presentes nessas áreas a fim de desenvolver uma aplicação de Mineração Visual de Texto para auxiliar na análise de uma base de dados de um fórum de discussões. Esse trabalho estende o projeto “Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” (PIBIC, id. 16856), realizado com os mesmos fins. A contribuição principal deste trabalho é a adição de novas funcionalidades das quais a inclusão da Visualização temporal é a mais significativa.

Palavras-chave: KDD, Mineração de Texto, Visualização temporal, fóruns de discussão.

CAPÍTULO 1

INTRODUÇÃO

A interação organizada entre indivíduos é um importante instrumento para disseminação de informações e dados, socialização de experiências e colaboração para a construção coletiva do conhecimento. Registros produzidos nesta atividade constituem uma importante fonte para avaliar a evolução do grupo de indivíduos em relação ao tópico objeto da interação. Para apoiar o especialista na avaliação, técnicas computacionais são úteis para a análise destes registros e possibilitam, além de um processamento rápido, um tratamento adequado para volumes maiores de dados e, também, podem conferir flexibilidade ao processo pela diversidade de abordagens, em relação ao procedimento manual.

Dentre esses registros, estão os textuais. A *internet* é um importante meio para isto, decorrente de sua infraestrutura e de seus serviços, pois provê diversas formas para as pessoas interagirem, tais como correio eletrônico, chats e fóruns de discussão. Tal interação, ou colaboração, pode ocorrer de forma síncrona - os atores estão conectados simultaneamente - ou assíncrona.

Neste projeto, a fonte de informação a ser tratada é formada pelas mensagens contidas em fóruns de discussão provenientes de um ambiente virtual aprendizagem - AVA -, isto é, em um contexto educacional. Consideramos, contudo, que isto não é uma restrição que impacta a generalização da solução. Serão utilizadas técnicas de Visualização Interativa de Dados e aspectos específicos de Mineração de Textos.

No projeto “Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” [Pacheco Jr., 2011], foi desenvolvida uma aplicação para o

processamento dos dados e os resultados foram animadores, isto é, a especialista conseguiu extrair novos conhecimentos em relação a sua primeira abordagem manual, o que fundamenta a sua extensão. Nesta presente proposta, além do conteúdo das mensagens, será considerada a dimensão temporal.

Esse texto está organizado em capítulos que contribuem para o entendimento do projeto realizado. No próximo capítulo (segundo) é feita uma apresentação sobre Mineração de Dados e de Texto, onde são estudadas as etapas de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* ou KDD). O terceiro capítulo discorre sobre Visualização de Informação, dando embasamento em algumas técnicas de Visualização. No quarto, inicia-se a abordagem sobre o desenvolvimento do objetivo final do projeto: o desenvolvimento de uma aplicação estendida do projeto de Pacheco Jr. com a adição de funcionalidades considerando a componente temporal. Nessa etapa, é feita uma análise e avaliação do programa que serviu como base, ressaltando os objetivos a que se propunha. Ainda nesse último capítulo, é relatado o desenvolvimento da nova aplicação, sendo discutido tanto as técnicas utilizadas quanto seu funcionamento e, a seguir, no quinto e último capítulo, é realizada uma análise geral do trabalho e de suas implicações.

1.1 Objetivo

O objetivo geral neste projeto é a investigação do uso de técnicas computacionais para suporte ao processo de análise de documentos textuais pela integração de técnicas de representação gráfica e manipulação interativa – Visualização Interativa de Dados - e de técnicas analíticas empregadas em Mineração de Texto, buscando beneficiar-se da união dos pontos fortes de cada categoria de técnica.

O conjunto de documentos textuais a ser tratado é composto por mensagens em fóruns de discussão provenientes de um ambiente virtual de aprendizagem - AVA [Rinaldi, 2009]. A investigação foi iniciada no projeto “Processo de Visual Analytics para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” [Pacheco Jr., 2011]. Os resultados do projeto inicial foram bastante satisfatórios e animadores por terem trazido novas informações àquelas conclusões que a especialista chegou pela análise “manual” (avaliação pela leitura de mensagem por mensagem). Os resultados, bem como a aplicação

desenvolvida e a técnica de visualização concebida, estão detalhados no relatório final daquele projeto ([Pacheco Jr., 2011]) e motivam a proposta deste projeto atual.

A principal contribuição nesta nova proposta é incluir a dimensão temporal na análise das mensagens, estendendo a aplicação para além da análise de conteúdo do projeto anterior, justificando, assim, o termo ‘Conteúdo-Temporal’ presente no título deste projeto. A partir do componente temporal, é possível extrair informações tais como picos de atividade, período predominante de atuação dos participantes e distância temporal entre participações, as quais, juntamente, com as informações extraídas pela análise do conteúdo, podem ser concretizadas em conhecimento para a estratégia de intervenção do especialista.

Trabalhos sobre a análise de conteúdo no mesmo contexto têm sido realizados e podem ser vistos em [Azevedo et al., 2011], [Azevedo et al., 2009], [Longhi et al., 2009] e [Stavrianou and Chauchat, 2008]. Técnicas visuais para o tratamento de documentos são apresentadas em [Stoffel et al., 2010],

[Keim et al., 2010] e [Strobelt et al., 2009]. O processamento da dimensão temporal em outros domínios de aplicação está descrito em [Yu et al., 2012], [Chittaro et al., 2003] (Medicina) e [Kechadi and Bertolotto, 2006] (Meio ambiente). Desta forma, os objetivos específicos deste projeto são:

- Inclusão da dimensão temporal na análise;
- Ampliação das funcionalidades para a manipulação interativa dos dados;
- Inserção de métricas usadas em Mineração de Texto.

1.2 Cronograma

Esse projeto foi dividido em duas partes, conforme ilustrado na Figura 1.1. Na primeira parte (fase 1), teve início o estudo da revisão e complementação da bibliografia dos temas abordados. Juntamente com esse estudo da teoria referente à Mineração de Texto e Visualização de Informação, também foi iniciada a implementação do aplicativo de Mineração de Texto.

| Atividades | Meses | | | | | | | | | | | |
|---|-------|---|---|---|---|---|---|---|---|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Revisão e Complementação da Bibliografia | | | | | | | | | | | | |
| Implementação fase 1 - estudo da aplicação anterior, temporal, interação e métricas | | | | | | | | | | | | |
| Implementação fase 2 - refinamentos | | | | | | | | | | | | |
| Escrita de relatório | | | | | | | | | | | | |
| Entrega relatório parcial | | | | | | | | | | | | |
| Entrega relatório final | | | | | | | | | | | | |

Figura 1.1: Cronograma do desenvolvimento do projeto.

Nessa segunda parte do projeto (fase 2), foram refinadas as funcionalidades e técnicas empregadas na primeira fase. Dessa maneira, a aplicação desenvolvida tornou-se mais interativa e eficiente, incorporando funções de *zoom*, suporte à múltiplas visualizações sincronizadas, visualização baseada em calendário entre outras, além de métricas de Visualização, como frequência inversa e porcentual de ocorrência das palavras.

CAPÍTULO 2

MINERAÇÃO

2.1 Introdução

Não há muito consenso à respeito de uma definição clara sobre Mineração de Dados. Enquanto alguns autores consideram Mineração de Dados como sinônimo de KDD (*Knowledge Discovery in Databases*), outros a consideram apenas uma etapa de KDD. Para [Fayyad et al., 1996],

“A um nível abstrato, o campo KDD preocupa-se com o desenvolvimento de métodos e técnicas para dar sentido aos dados. O problema básico abordado pelo processo de KDD é um mapeamento de dados de baixo nível (que são tipicamente muito volumosos para compreender e digerir facilmente) em outras formas que possam ser mais compactas (por exemplo, um relatório curto), mais abstratas (por exemplo, uma aproximação descritiva ou o modelo do processo que gerou os dados), ou mais úteis (por exemplo, um modelo preditivo para estimar o valor de casos futuros)”.

“A mineração de dados é uma etapa do processo de KDD que consiste na aplicação de análise de dados e algoritmos de descoberta que, sob limitações aceitáveis de eficiência computacional, produz uma determinada enumeração dos padrões (ou modelos) sobre os dados” (tradução nossa).

Concluímos dessas definições que o principal objetivo do KDD é organizar e reestruturar conjuntos primitivos de dados de forma a descobrir informações úteis a seu respeito.

Com o aumento gigantesco e gradativo do volume de dados armazenados no mundo, a Mineração de Dados pode ser aplicada com o objetivo de tornar a tarefa de análise desses

dados mais rápida e eficiente. Apenas para informação, um recente estudo realizado pelo IDC (*International Data Corporation*), a pedido da empresa *EMC Corporation*, estimou que até 2020 o volume total de dados do mundo chegará a 40 zettabytes (ZB), sendo que atualmente menos de 1% dos dados mundiais são analisados, e menos de 20% são protegidos [Weinzierl, 2012].

2.2 Mineração de Texto

Nos primeiros estudos sobre a Mineração de Texto, os conjuntos de textos eram tratados de forma monolítica e estática, não agregando nenhuma informação a não ser seu conteúdo textual. Avanços na área, contudo, revelaram que informações de data e hora associadas à cada texto também poderiam ser consideradas para aumentar o universo de informações que poderiam ser descobertas. Nesse escopo, surgem conceitos de *Trend Analysis*, *Ephemeral Associations* e *Deviation Detection*, que podem ser definidas respectivamente como: distribuição de comportamentos sobre múltiplos subconjuntos de documentos usando atributos temporais, influência de frequências ou “picos” nas associações dos documentos e detecção de desvios ou anomalias que ocorrem quando as distribuições de probabilidades comparadas entre tópicos do mesmo conjunto de textos diferem [Feldman and Sanger, 2007].

Há diversas áreas em que as técnicas de Mineração de Dados é aplicada, o que depende do tipo de dado que pretende-se processar e das informações que se deseja obter. Essas divisões são compostas por descrição, classificação, estimativa ou regressão, predição e agrupamento [Shimabukuro, 2004][Fayyad et al., 1996][Chen et al., 1996]. Para esse trabalho, utilizamos a Descrição para extrair informações dos dados utilizados, implementada em conjunto com a Visualização temporal.

Uma particularidade muito importante da Mineração de Dados é que ela foi projetada para ser aplicada somente em dados estruturados, como Banco de Dados e *Warehouses* (um tipo particular de base de dados) [Fayyad et al., 1996]. Dados não-estruturados ou semi-estruturados, como arquivos de mídia e de texto, constituem um tipo a parte de estrutura, chamadas Estruturas Complexas [Fayyad et al., 1996]. Na Mineração de Texto, todos os dados são semi-estruturados ou não possuem estrutura sob o ponto de vista de organização, e não morfológico. Esse tipo de visualização tornou-se muito comum devido

ao crescimento de dados textuais provindos principalmente da *web* e da facilidade de edição desses arquivos, que muitas vezes podem ser editados com um simples editor nativo do próprio sistema operacional. Arquivos estruturados, por outro lado, embora resultem em extrações de informações com maior exatidão, possuem o inconveniente de necessitar de um sistema particular para manipulá-los.

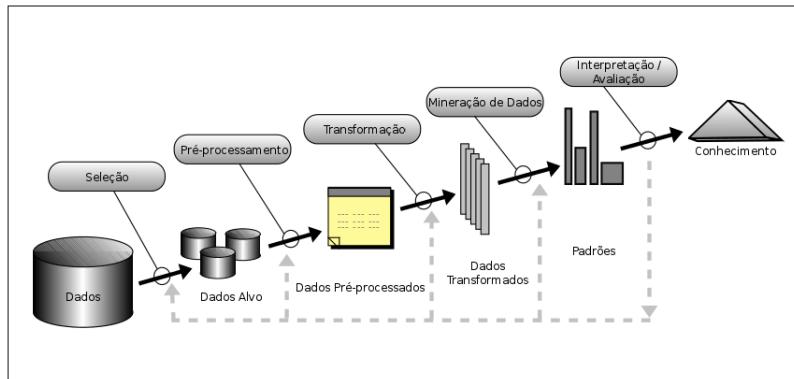


Figura 2.1: Etapas de KDD segundo Fayyad; adaptado de [Fayyad et al., 1996].

Em todas os tipos de Mineração, as etapas de KDD são as mesmas. Para extrair informações de bases de dados faz-se necessário executar todas as etapas mostradas na Figura 2.1. As etapas que compõem esse processo são dadas por 1) seleção dos dados de um conjunto para processar, o 2) pré-processamento (filtragem), a 3) transformação dos dados, baseada em alguma técnica de Mineração, a 4) Mineração de Dados, que deverá extrair padrões, os quais podem ser visualizados por meio de figuras ou tabelas, por exemplo, e a 5) interpretação, para verificar se as informações são úteis, ou seja, se são conhecimento.

2.2.1 Etapas

Seleção

A fase de seleção é uma das etapas mais importantes do processo de KDD. Nela, a partir de uma base bruta de dados devem ser selecionados os dados de onde realmente queremos extrair alguma informação. Esses dados devem ser semânticos e ter alguma padronização de aspecto dependendo da forma que se encontre.

Nos dados utilizados nesse projeto, os textos tem uma forma específica de tratamento. Por terem sido retirados de *posts* de um fórum de discussões, as três primeiras linhas dos textos são desnecessárias ao resultado final, pois os próprios nomes dos arquivos forneciam

as informações temporais. Os nomes dos arquivos possuem o formato *mm-dd-aaaa hh-mm.txt*, e as três primeiras linhas possuem as informações *De:* ou remetente (1^a linha), *Para:* ou destinatário (2^a linha), data e hora de postagem (3^a linha).

Embora dados em formato texto não possuam uma estrutura definida, como no agrupamento de um Banco de Dados, não é muito difícil de analisá-los. O problema ocorre quando esses dados estão em linguagens diferentes, o que pode gerar muito ruído ou interferência na avaliação final das informações. Esse tratamento deve ser feito à priori pelo usuário, que deve verificar essas discrepâncias nos dados.

Pré-processamento e Transformação

Já vimos que, para o conjunto de dados utilizados, as três primeiras linhas de todos os arquivos seriam inúteis para a extração de informações final e que, por esse motivo, deveriam ser extraídas do conjunto inicial. Essa fase de filtragem e melhoramento dos dados é conhecida como “pré-processamento”, que tem por finalidade deixar os dados o mais uniformes possível para serem processados e transformados em visualizações úteis. No pré-processamento, os dados originais são transformados com base em uma coleção de dados que não contribuem para a fase de Mineração. No programa Mineração de Texto, implementado no projeto anterior [Pacheco Jr., 2011], os dados dessa coleção são chamados de *stopwords*, ou seja, são palavras irrelevantes no contexto em que se pretende “descobrir conhecimento”, tais como conectivos, pontuações, espaços em branco ou números. Geralmente, a coleção de *stopwords* é dada na forma de um ou mais arquivos de texto, o que não implica que esse conjunto não possa ser convertido em um conjunto de dados estruturados para agilizar no processamento, como em um Banco de Dados.

A fase de transformação dos dados ocorre quando os dados pré-processados adquirem certa estrutura bem definida para que possam ser computados por máquinas. Nesse processo, o conjunto de textos pode ser rearranjado em tabela virtuais que facilitem o manuseio dos dados, como matrizes ou vetores, posto que o processamento direto dos arquivos é muito custoso em termos de custo computacional, como memória e processamento.

Mineração de Dados

Talvez a fase mais importante do processo de KDD, a Mineração se constitui do casamento de padrões dos dados por meio de algum algoritmo e, a partir dessa informações, gerar modelos gráficos que simbolizam o conteúdo resultante [Fayyad et al., 1996]. Esses modelos são imprescindíveis para a próxima etapa, visto que será a partir deles que a avaliação será realizada e o possível conhecimento extraído.

Por ser tão importante ao processo de KDD, decidimos abordar as técnicas de geração de modelos gráficos em um tópico à parte. Esse estudo de geração de modelos baseado em dados é conhecido como Visualização de Dados, de onde podemos citar como exemplo a geração de gráficos estatísticos [Chen et al., 2008]. Abaixo, a Figura 2.2 ilustra uma comparação entre visualizações 2D e 3D.

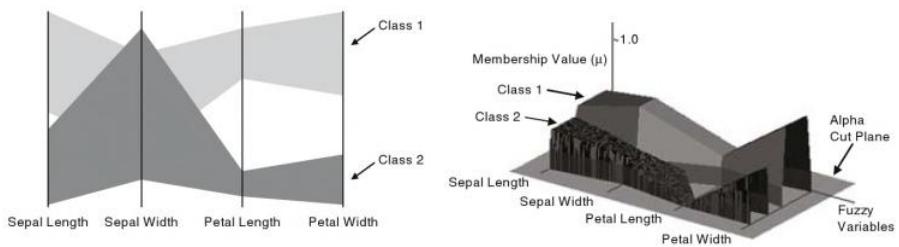


Figura 2.2: Exemplo de visualização de dados em 2D e 3D; extraído de [van Zudilova-Seinstra et al., 2009].

Interpretação ou Avaliação

Última etapa do KDD, é nesse ponto em que são extraídas informações acerca dos dados processados. Geralmente, carece de uma análise feita por um especialista sobre o assunto dos dados, que pode inferir com mais propriedades sobre o resultado encontrado, sendo que programas sob essa perspectiva não costumam apresentar métodos inteligíveis de análise automática dos dados.

A forma como os dados são representados por meio de imagens reflete diretamente na qualidade das informações extraídas. Existem muitas formas em que uma imagem pode ser constituída. Na imagem anterior, para o mesmo conjunto de dados foram geradas imagens diferentes. Na primeira, os resultados são mostrados em uma imagem 2D, que fornece poucas informações sobre o conjunto todo dos dados. Já na segunda, podemos notar uma nítida diferença do volume de informações mostradas; além da dimensão do

conjunto todo pelo volume de cada área, também são representados atributos como textura e coloração, este último também presente na primeira imagem.

2.3 Tipos de dados

Os dados analisados na Mineração de Texto podem ser de dois tipos: estáticos ou dinâmicos. Quando dizemos tipos, não queremos nos referir aos formatos dos dados analisados (*pdf* e *txt*, por exemplo), e sim ao estado em que serão analisados. Dados estáticos representam uma coleção de dados fixa, imutável durante o processo de KDD, enquanto que os dinâmicos são suscetíveis a modificações ao decorrer desse processo.

CAPÍTULO 3

VISUALIZAÇÃO

3.1 Introdução

Segundo a definição de [McCormick et al., 1987],

“A visualização é um método de computação. Ela transforma o (conjunto de dados) simbólico em geométrico, permitindo aos pesquisadores observar sua simulação e computação. Visualização oferece um método para ver o invisível. Ela enriquece o processo de descoberta científica e promove reflexões profundas e inesperadas. Em muitos campos já está sendo revolucionada a forma como os cientistas fazem ciência” (tradução nossa).

Embora não pareça, o estudo de técnicas de Visualização de Dados é relativamente antigo. Segundo [Friendly, 2012] e [Chen et al., 2008], desde o século XIX esse campo tem sido alvo de muitos estudos e pesquisas de desenvolvimento, principalmente nas áreas de Estatística e Cartografia. Nesse sentido, a Visualização age como intermediador entre o conhecimento que pode ser obtido em um conjunto de dados e o indivíduo que deseja obtê-lo. Sua principal tarefa é representar conjuntos de dados por meio de imagens que podem ou não ser abstratas, dependendo do tipo da Visualização.

Geralmente, o processo de Visualização de Dados é aplicado em grandes conjuntos que seriam inviáveis de serem analisados manualmente, isto porque quanto maior for o espaço amostral dos dados coletados, mais precisos serão os resultados ou informações que poderão ser coletados na Visualização. Isso pode proporcionar um ganho considerável na análise de dados, visto que o apoio computacional aliado às técnicas corretas de

Visualização podem levar o pesquisador à uma análise mais apurada dos dados, que por sua vez pode resultar na descoberta de conhecimentos antes não obtidos em uma análise manual.

3.2 Tipos de Visualização

Em 1987, no relatório *Visualization in Scientific Computing*, da NSF (*National Science Foundation*), foi introduzido o termo Visualização Científica. O texto ainda ressalta a importância e necessidade de serem criadas novas técnicas de Visualização de Dados, que deveriam satisfazer os mais variados objetivos de análise, assim como oferecia uma definição e abordava o domínio e aplicações do tema, além de recomendações. McCormick et al. (1987) ainda discorre sobre o principal fator que influencia para que a Visualização seja um bom modelo de análise de dados, salientando uma estimativa de 50% dos neurônios cerebrais humanos estarem associados à visão [McCormick et al., 1987].

3.2.1 Visualização Científica

De acordo com [McCormick et al., 1987], a Visualização Científica foi a primeira a ocorrer no âmbito dos estudos sobre Visualização. No início, apenas dados relacionados às ciências e engenharias eram amplamente utilizados. As áreas mais abrangidas eram a Estatística e Cartografia, que em sua maioria possuíam dados como tabelas de números e códigos; por essa razão, as imagens criadas tinham um caráter mais rigoroso com a realidade, representando os dados de uma forma muito realista.

Em um segundo momento, a Visualização Científica também começou a ser usada em outras áreas, como a Medicina, Biologia, Química e Astronomia [Chen et al., 2008]. No exemplo da Figura 3.1, é exibido um exemplo do uso da Visualização Científica em uma aplicação cartográfica.

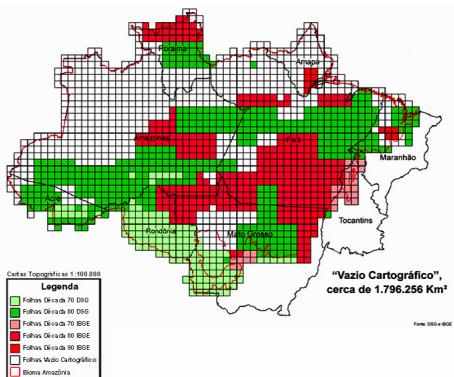


Figura 3.1: Exemplo de Visualização Científica, Disponível em: < http://www.fab.mil.br/portal/capa/index.php?datan=03/08/2009&page=mostra_notimpol >, Acessado em: 6/12/12.

3.2.2 Visualização de Informação

Na Visualização de Informação, a visualização possui um caráter mais abstrato, sem tanto rigor com a representação real e geométrica dos dados analisados. O importante não é representar os dados como eles são na realidade, mas criar uma imagem que represente os dados da melhor forma possível, de maneira que o pesquisador possa conseguir abstrair as informações coletadas e transmiti-las para sua realidade. Essa forma de Visualização lida principalmente com dados que possuem um significado abstrato no mundo real, como mensagens na *web*, códigos-fonte e indexação de sites [Schroeder et al., 2003]. Um exemplo é apresentado na Figura 3.2.

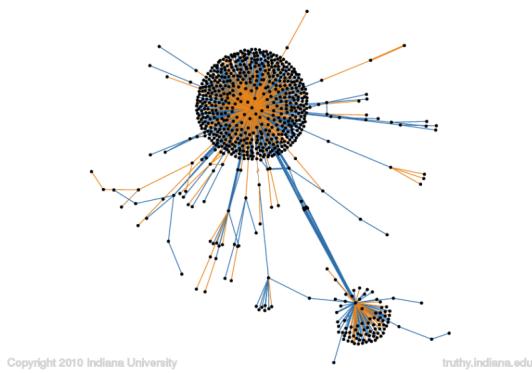


Figura 3.2: Exemplo de Visualização de Informação, Disponível em: < http://truthy.indiana.edu/memedetail?id=324&resmin=45&theme_id=4 >, Acessado em: 6/12/12.

Muitas vezes, por esse tipo de visualização ser mais abstrata do que a Visualização Científica, as imagens geradas podem virar verdadeiras obras de arte. Por um lado, isso é

bom, pois pode ajudar o usuário a compreender melhor os dados exibidos, mas por outro, o excesso de arte pode prejudicar a efetividade da exibição dos dados, o que é ruim para o resultado final. Deve haver um equilíbrio entre a escolha certa da técnica utilizada, o volume de informações mostradas, quais informações devem ser mostradas e a qualidade pictórica da Visualização.

3.3 Técnicas de Visualização

Na literatura existem diversas técnicas de Visualização que são adequadas para tipos específicos de dados. Essas técnicas, contudo, podem muitas vezes não ser o bastante para a representação de uma coleção de dados, o que implica na necessidade de se criar novos modelos inspirados nas principais técnicas para representar esses dados. Durante a fase de implementação de nosso aplicativo de Visualização, optamos por utilizar uma técnica comumente abordada por autores da área para representar os dados analisados, a Visualização Orientada a Pixel [Ankerst, 2001]. Para tanto, fizemos algumas adaptações no modelo usado para que fossem obtidos os resultados desejados.

3.3.1 Visualização Orientada a Pixel

A partir dessa técnica, a visualização é feita com base em um conjunto de atributos que classificam e diferenciam os dados de acordo com seu valor ou natureza. Todos os dados possuem o aspecto de um pixel de tela, que são coloridos de acordo com uma regra específica. Como pixels são componentes muito pequenos na maioria das vezes, optamos por nominar esses supostos pixels como pontos que, por serem adimensionais, podem ser aplicados independente do tamanho dos componentes. Sendo assim, esses pontos possuem um tamanho pré-definido, e são dispostos na tela conforme um mapa de distribuição [Ankerst, 2001].

A diferenciação dos pontos é feita tanto pela sua cor e intensidade, quanto pela disposição geométrica. De acordo com [Shimabukuro, 2004], existem diversos aspectos a serem considerados na aplicação dessa técnica, sendo que os principais são: arranjo dos pontos nas janelas, mapeamento de cor e o formato das janelas. Nesse último aspecto, podem ressaltar dois tipos de Visualização: Gráficos de Barras e Segmentos de Círculos [Shimabukuro, 2004]. No primeiro (ver Figura 3.3), os dados são arranjados em um

retângulo principal, que nada mais é do que o conjunto de janelas de pontos ordenados sequencialmente, sendo que a leitura dos dados é feita de cima para baixo e da esquerda para a direita. No Segmento de Círculos, exemplificado pela Figura 3.4, essas janelas são dispostas em fatias ordenadas circularmente, formando um círculo que se assemelha muito com um gráfico de pizza; os dados são lidos de fora para dentro das janelas.

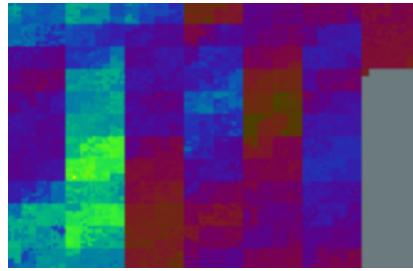


Figura 3.3: Visualização Orientada a Pixel por meio de Gráfico de Barras; adaptado de [Ankerst, 2001].

A característica mais importante da Visualização Orientada a Pixel é a possibilidade de exibir muitos dados simultaneamente em um espaço reduzido de tela [Pacheco Jr., 2011]. Entretanto, há certa diferenciação entre as técnicas de Gráfico de Barras e Segmento de Círculos. A técnica de Segmento de Círculos proporciona maior clareza na relação dos dados do que a técnica de Gráfico de Barras, isto porque há uma disposição de círculos concêntricos na imagem que auxiliam nessa identificação. Por outro lado, essa facilidade se esvai quando esses círculos ficam cada vez maiores, e o que antes era uma vantagem pode se tornar uma desvantagem.

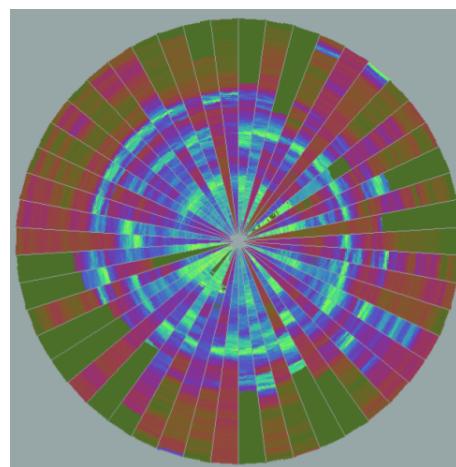


Figura 3.4: Visualização Orientada a Pixel por meio de Segmento de Círculos; extraído de [Ankerst, 2001].

Nessa última imagem (Figura 3.4), podemos notar uma relação bem clara entre os dados do círculo. Há uma grande predominância de pontos verdes e azuis do centro até o meio do círculo, enquanto que a outra metade é composta predominantemente por pontos em tons de roxo e verde, ambos com maior grau de opacidade.

CAPÍTULO 4

APLICAÇÃO PARA ANÁLISE QUALITATIVA ‘CONTEÚDO-TEMPORAL’

4.1 Introdução

O principal foco deste projeto é o desenvolvimento em continuidade de uma aplicação de Mineração Visual de Texto baseado em um trabalho anterior desenvolvido por Pacheco Jr. (2011). O objetivo de ambos os trabalhos é auxiliar na análise de dados de um fórum de discussões, buscando proporcionar a extração de novas informações acerca de um conjunto específico de dados já analisados [Rinaldi, 2009] pela pesquisadora e colaboradora deste projeto.

A base de dados a que nos referimos é composta por arquivos de texto provenientes de um fórum de discussões, onde cada arquivo é um *post* escrito pelos usuários do fórum. Essa base já tinha sido analisada pela pesquisadora [Rinaldi, 2009] sem o suporte de um *software* de Visualização mas, por ser muito trabalhosa, a análise seria inviável para conjuntos de dados muito volumosos. Analisar coisas semelhantes repetidas vezes torna-se cansativo e improdutivo, pois a repetição excessiva de uma mesma atividade pode prejudicar a interpretação dos dados, que pode ser causada por “vista cansada” ou “vista viciada”, que é quando o cérebro identifica padrões em coisas distintas por já ter se habituado à um cenário, amortecendo a capacidade de concentração e avaliação do indivíduo.

Durante a análise manual dos dados, a pesquisadora conseguiu alcançar resultados

importantes sobre a base, mas suspeitou que poderia obter melhora na análise se auxiliada por recursos computacionais. A partir dessa necessidade, iniciou-se um estudo sobre as técnicas de Visualização de Informação que pudessem ser úteis à esse propósito, onde chegou-se à conclusão de que a técnica de visualização por pixel poderia ser utilizada [Pacheco Jr., 2011]. No projeto realizado por Pacheco Jr. (2011), propôs-se uma visualização geral dos dados, para exibir todos os textos analisados em uma única visualização. Todos os textos do fórum foram transformados em linhas de pontos (pequenos quadradinhos), onde cada ponto representava uma palavra do texto. Dessa forma, dispondo as linhas (textos) uma embaixo da outra, pôde-se obter uma visão clara de todos os dados (textos divididos em palavras, representadas por pontos coloridos) ao mesmo tempo, conforme mostrado na Figura 4.1. Esse tipo de Visualização permite ao analista descobrir possíveis relações intertextuais baseando-se na localização, quantidade e frequência de palavras destacadas.

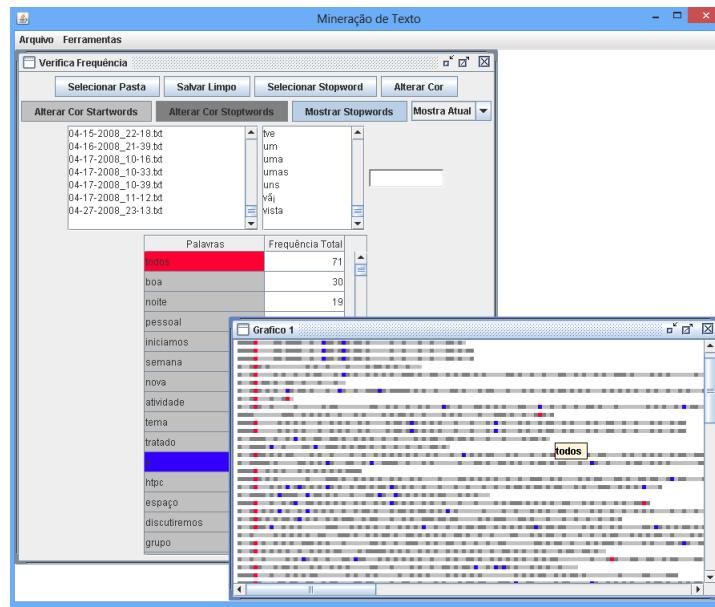


Figura 4.1: Representação visual de texto por pontos, extraída do programa desenvolvido por Pacheco Jr. (2011).

Conforme ilustra a figura acima, o programa foi composto por uma janela de visualização, onde as informações são mostradas, uma janela de configurações, onde está concentrada a maior parte das funcionalidades, e um menu principal, por onde são acessadas essas janelas. Nesse menu, há ainda um editor de textos simples, caso o usuário queira criar ou modificar algum texto usando a própria ferramenta.

Para usar o programa, o usuário deve selecionar através do menu tanto o diretório

que contém os arquivos de entrada (textos) quanto um arquivo de *stopwords* (arquivo que contém as palavras irrelevantes para a Visualização segundo o usuário). O arquivo de *stopwords* é opcional e, quando o diretório dos arquivos de entrada é selecionado, a Visualização dos textos é automaticamente processada. Entre as funcionalidades primárias constam a exibição das frequências das *startwords*, a possibilidade de destacar uma *startword* através da mudança de cor - feita instantaneamente -, e a identificação das palavras ao passar o cursor do mouse sobre cada ponto. Já as secundárias são compostas pela exibição do arquivo original de cada texto, obtida ao clicar sobre um ponto qualquer de cada texto, mudança da cor dos *stopwords* ou *startwords* e salvamento dos arquivos processados.

A pesquisadora conseguiu obter resultados animadores com o programa, tendo conseguido extrair mais informações à respeito da base de dados, informações essas que não tinham sido identificadas na análise manual [Pacheco Jr., 2011]. Contudo, questões como a concentração de *posts* escritos em uma determinada faixa de horário sobre determinado assunto são impossíveis de serem respondidas. Para que isso fosse possível, seria necessário implementar a dimensão temporal à visualização, principal contribuição desse novo programa desenvolvido.

4.2 Desenvolvimento

Para inserir o atributo temporal, optamos por também utilizar a técnica de Visualização Orientada a Pixel, pois a pesquisadora especialista da área já estava familiarizada com a mesma. A linguagem de programação utilizada também continua sendo Java, por motivos de facilidade e portabilidade.

Com o objetivo de tornar mais fácil a referência ao aplicativo criado nesse projeto, optou-se por nomear o *software* com um nome sugestivo, que fizesse referência ao seu propósito. Assim, esse programa ganhou o nome de BoardWords, uma adaptação do termo *board of words* (trad. tábua de palavras), referência direta ao aspecto das visualizações presentes no programa.

Na primeira fase da projeto, foi sugerido uma separação da Visualização em dois tipos: normal e temporal, pois acreditou-se que simplesmente inserir um novo atributo sobre aqueles dados da representação da primeira visualização (Figura 4.1) poderia ocasionar

excesso de informações e, consequentemente, ser ineficaz para o que se propunha.

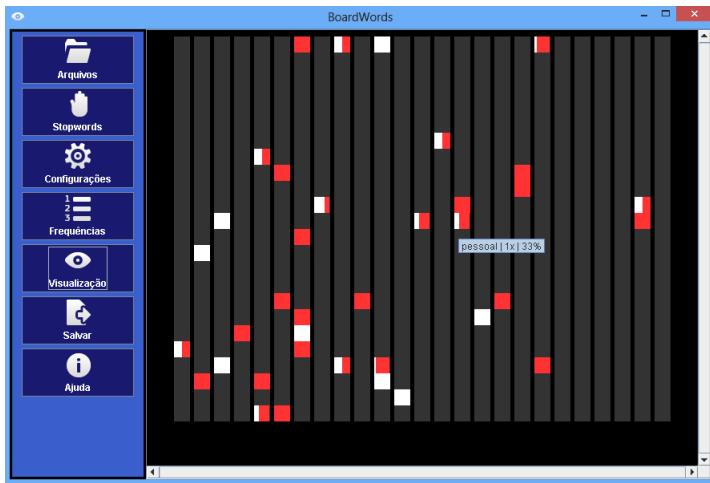


Figura 4.2: Primeiro modelo de Visualização temporal proposto.

No painel principal (à direita) da tela apresentada na Figura 4.2, é mostrado um exemplo de Visualização temporal, cujas variáveis horizontais e verticais são dia e hora, ambas com valor igual a 1. Cada coluna representa um conjunto de dias, meses ou anos, e o espaço vertical das colunas representa o passar do tempo desse conjunto de dias, meses ou anos. Por exemplo, se tivermos um conjunto de textos que foram escritos em um determinado ano, e fizermos uma visualização temporal por dia, com o valor 2 para essa variável, cada coluna representará o conjunto de textos escritos a cada dois dias (por exemplo, 1 e 2, 7 e 8); se a configuração for por mês e o valor for 7, a visualização representará o conjunto de textos escritos a cada 7 meses, e assim por diante, ressaltando que essa contagem se inicia à partir da data do primeiro *post* da base de textos.

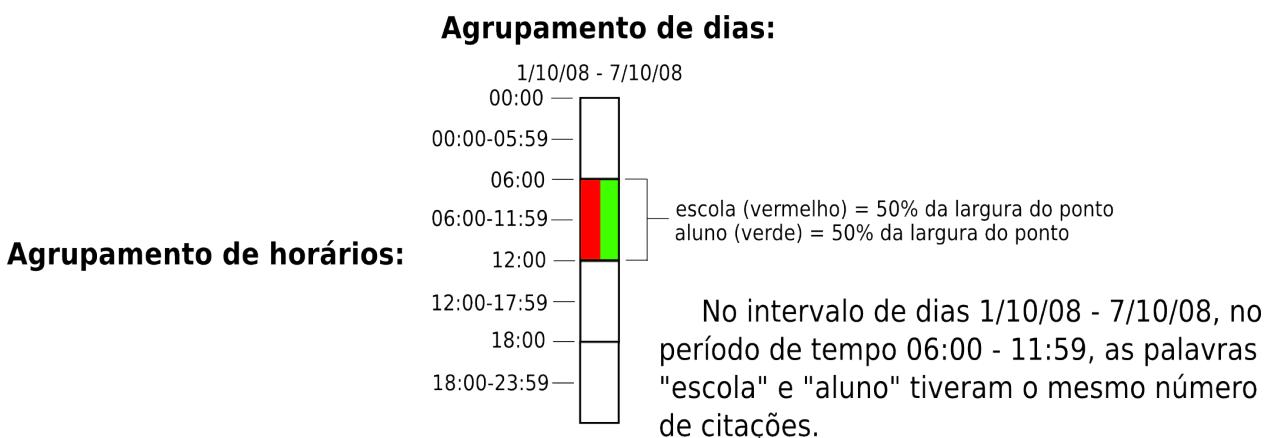


Figura 4.3: Esquema da primeira Visualização temporal desenvolvida.

A Figura 4.3 ilustra o esquema dessa Visualização. O comprimento das colunas (vertical) é relativo ao passar das horas de cada dia do conjunto de arquivos de cada coluna. Digamos que a configuração horizontal seja mês e seu valor seja 9, então, a cada 9 meses temos um conjunto de textos. Supondo que a configuração vertical seja minuto e o valor seja 30, teremos um ponto a cada 30 minutos, começando das 0h de cada dia. Desta forma, teríamos 24 pontos em cada coluna, pois em um dia cabem 24 vezes o período de tempo 30 minutos. A leitura dessa visualização deve ser feita da esquerda para a direita e de cima para baixo, sendo que os pontos representam o conjunto de palavras anteriormente incluídas para serem visualizadas. Dentro de um ponto são exibidos vários retângulos, que podem ter diferentes larguras e cores. A largura de cada retângulo simboliza o percentual aproximado que aquela palavra (retângulo) possui em relação às ocorrências de todas as palavras do conjunto incluído dentro daquele intervalo de tempo e de dias. Para facilitar a identificação, é possível identificar cada palavra posicionando-se o cursor do mouse sobre o retângulo desejado; além da palavra, também é mostrado a quantidade de ocorrências e o percentual da mesma, sendo que ao clicar é exibido o conjunto de textos daquele ponto. Entretanto, foi concluído que o modelo de Visualização temporal proposto (Figura 4.2) era pouco eficiente e incompleto para auxiliar na descoberta de todas as informações que se almejava conhecer, pois não permitia a descoberta de conhecimentos sobre picos de atividades de utilização do fórum ou relacionamentos dos assuntos abordados nos textos. Essa Visualização permitia apenas descobrir o relacionamento entre palavras isoladas nos textos, através de distribuição temporal, orientada por horário e data de criação do texto, e frequências, orientadas pelo tamanho do ponto de representação da palavra.

Em reunião com a colaboradora do projeto, foi constatado que o conceito de agrupamento temporal dos dias e horários, aplicados na proposta de Visualização temporal da primeira fase, poderia ser migrada para a Visualização normal, que segue o modelo proposto por Pacheco Jr.. Essa possível junção de ambas idéias poderia proporcionar a descoberta de associações entre textos que foram escritos em diferentes períodos de data ou horário, abrangendo o universo de associações possíveis, como picos de atividade, período predominante de atuação dos usuários do fórum ou participantes e distância entre participações ou publicações de textos, aliado a informações extraídas pela análise do conteúdo íntegro dos textos.

Foram desenvolvidos dois tipos de Visualização, uma baseada em colunas de pontos

(Figura 4.4), onde podem ser aplicados agrupamentos de data ou horário (não é necessário), e outra baseada em calendário. O primeiro tipo busca responder questões sobre relacionamentos entre os textos, independente do grupo de horário ou data em que corresponde. Esses agrupamentos revelam a afinidade ou relacionamento entre os textos sobre diferentes pontos de vista. Agrupamentos de períodos de datas permitem identificar os textos que pertencem ao mesmo período, não considerando o horário em que foram escritos. Os agrupamentos por horário identificam os textos em que foram escritos na mesma faixa de horário e, por serem ambos os agrupamentos voltados ao atributo temporal, devem ser aplicados em conjunto.

Painéis ou janelas de visualização

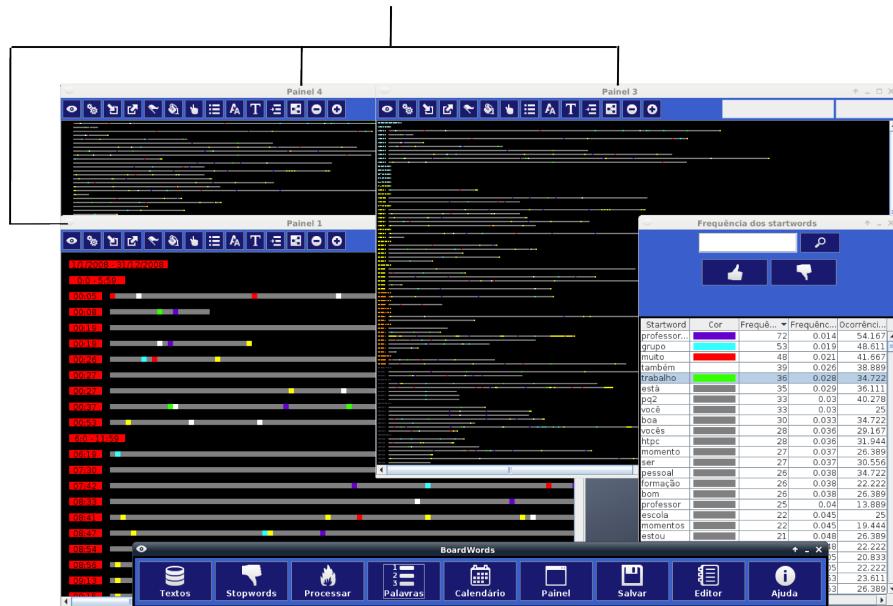


Figura 4.4: Modelo de visualização temporal final: múltiplos painéis em funcionamento.

O segundo tipo de Visualização (Figura 4.5) tem como finalidade mostrar a distribuição de atividade do fórum sobre o ponto de vista de calendário, salientando os dias ou pontos em que houve maior atividade no fórum por intensidade de coloração: quanto maior o número de *posts* publicados, maior a intensidade de cor do ponto. Isso ajuda o analista a identificar momentos em que houve maior atividade no fórum, e assim selecionar as regiões de maior interesse para a análise. Na Figura 4.5, as regiões em azul estão selecionadas, e as regiões em vermelho ilustram a intensidade de *posts* publicados em cada dia (ponto).

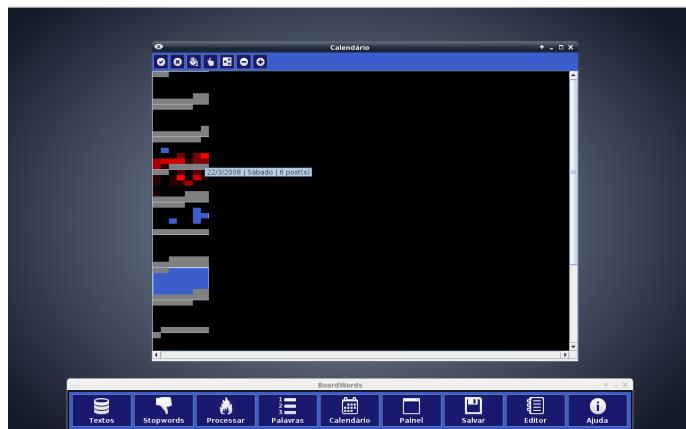


Figura 4.5: Visualização temporal baseada em calendário.

BoardWords apresenta uma série de recursos voltados à interatividade com o usuário, permitindo a realização de *zoom*, alteração das cores dos componentes, seleção de textos manual ou por palavra, destaque de palavras selecionadas (Figura 4.6), entre outros. Houve também uma preocupação em manter a possibilidade do usuário realizar múltiplas visualizações sincronizadas (Figura 4.4), recurso que já existia no projeto de Pacheco Jr. [Pacheco Jr., 2011] mas que ainda não tinha sido adicionado à essa aplicação. A realização de múltiplas visualizações sincronizadas é muito importante para a análise de dados, pois permite ao analista compreender de forma mais clara os relacionamentos dos dados sobre diferentes configurações de visualização. Por exemplo, pode-se comparar uma visualização com agrupamento de textos dispostos de hora em hora, e uma visualização com agrupamento de textos dispostos de semana em semana, permitindo conhecer, em determinada semana, quais dias em que houve maior atividade sobre um assunto específico, identificado pelo relacionamento de palavras associadas. Essa sincronização se deve ao fato dos textos e palavras sempre possuírem as mesmas cores, independente de quantas visualizações estão sendo realizadas (em janelas diferentes) e de quais as suas configurações.

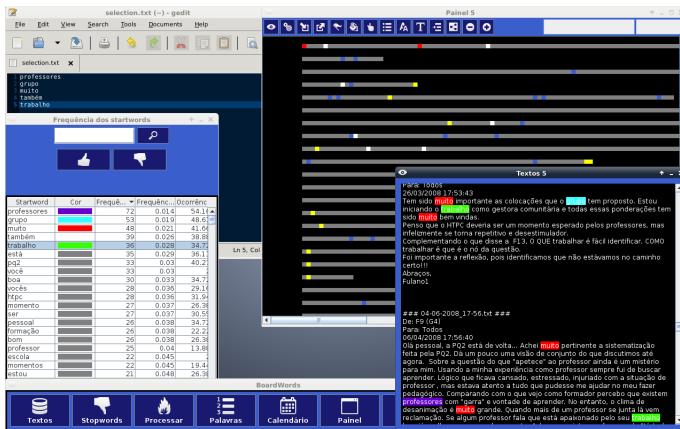


Figura 4.6: Destaque de *startwords* nos textos integrais.

Conforme relatado na proposta de desenvolvimento deste trabalho, foram utilizadas técnicas de Mineração de Texto para complementar o conjunto de funcionalidades da aplicação. Entre as métricas adicionadas estão a distribuição de *posts* baseada em calendário e representada pela intensidade de coloração dos pontos, a frequência absoluta das palavras em relação à todos os textos ou de cada texto (quantidade de vezes em que ocorre), a abrangência das palavras (porcentual em que ocorre em todos os textos), a frequência relativa em relação à cada texto (proporção da frequência absoluta das palavras) e a frequência inversa de cada texto (razão da frequência absoluta pelo total de textos em que o termo aparece). Em todas os cálculos de frequência, deseja-se extrair ou identificar a relevância das palavras, para poder priorizar termos mais significativos ao significado do textos. Conceitos mais detalhados de cálculos de frequências podem ser encontrados em [Morais and Ambrósio, 2007].

4.3 Funcionamento

BoardWords pode ser caracterizado de uma forma bem objetiva como uma aplicação de Mineração Visual de Texto composta por um menu principal independente (Figura 4.7) responsável por controlar suas funcionalidades. Nesse menu principal estão as funcionalidades essenciais necessárias para analisar dados textuais: entrada de dados ou textos através da seleção de um diretório que os contém (botão **Textos**), seleção de um arquivo de *stopwords* (botão **Stopwords**) e processamento dos dados (botão **Processar**), baseado na coleção de *stopwords*, respectivamente. Essa primeira fase é necessária e indispensável para se ter acesso às funcionalidades de análise dos dados.

Os dados de entrada devem ser nomeados seguindo o modelo *mm-dd-aaaa_hh-mm.txt*, e seu conteúdo deve possuir no mínimo três linhas de dados, conforme explicado na subseção **Seleção** do capítulo **Mineração**. O arquivo de *stopwords* também é obrigatório, e cada linha de seu conteúdo deve representar uma *stopword*. Dessa forma, as palavras passam a ser distinguidas entre *stopwords* ou *startwords*. Processados os dados, são habilitadas as funções representadas pelos botões **Palavras**, **Calendário** e **Painel**.



Figura 4.7: Menu principal do BoardWords.

No botão **Palavras** estão reunidas todas as *startwords* dos textos, com respectivas cores de representação, frequência absoluta, frequência inversa e abrangência, sendo todos os cálculos baseados na junção de todos os dados de entrada. Através dessa funcionalidade, é possível alterar as cores das palavras de forma independente ou baseado na categorização de *stopwords* ou *startwords*, como também pesquisar por uma palavra específica na lista de *startwords*.

Calendário é utilizado para identificar a distribuição temporal do conjunto de textos analisados, baseando-se em uma estrutura de calendário. Dessa forma, conforme ilustrado na Figura 4.5, a frequência de atividade dos textos postados é plotada em uma estrutura de colunas que representam anos. No caso do exemplo em questão, os dados analisados correspondem à apenas um ano. Essa coluna é dividida em 12 retângulos que representam os meses, sendo separados por uma linha mais clara. Os pontos visíveis de atividade, identificados em vermelho, são os dias em que houve atividade no fórum. Tomando a cor vermelha como base, a intensidade de atividade no fórum varia do preto (ausência de cor) para o vermelho (cor de referência), obtendo-se um contraste entre os pontos mais significantes. Em cinza, são diferenciados os dias inválidos dos meses, isto é, a continuação dos dias dos outros meses. Através desse calendário, pode-se identificar períodos em que houve maior atividade no fórum, e assim selecionar os dados desse período (em azul, na Figura 4.5) para uma análise mais detalhada. Outras funcionalidades podem ser exploradas no menu de funções da janela Calendário.

Agregando o maior conjunto de funcionalidades da aplicação, **Painel** é o recurso fundamental do BoardWords. Conta com um menu superior de funcionalidades que proporciona ao analista realizar uma análise detalhada dos dados processados. Entre

as opções do menu, podem ser destacadas a de geração de Visualização, configurações da Visualização, importação e exportação de configurações, destaque de palavras, visualização dos termos destacados sobre os textos analisados e projeção de textos destacados para uma nova janela.

No menu principal da aplicação, contamos ainda com a opção de salvamento dos textos processados, edição de textos e exibição de ajuda sobre a aplicação. Em **Ajuda**, o usuário poderá conferir mais detalhadamente o relato de funcionamento de todas as opções do BoardWords. Maiores informações sobre a aplicação podem ser obtidas no manual incluído em Anexo.

CAPÍTULO 5

CONCLUSÕES

Com o objetivo de proporcionar uma análise muito mais rápida e detalhada sobre um conjunto de dados, e para auxiliar pesquisadores na obtenção de informações de alguma forma ocultas à análise manual, com o foco em textos de fóruns de discussão, neste projeto é estendida a aplicação desenvolvida em “Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” [Pacheco Jr., 2011].

Vimos que, a partir da Visualização temporal concebida e implementada neste projeto, o conjunto de dados anteriormente analisados [Rinaldi, 2009] pela pesquisadora proporciona ainda mais informações. Embora a aplicação desenvolvida por Pacheco Jr. (2011) tenha fornecido uma visão geral dos dados, questões como picos de atividade dos usuários do fórum, distribuição temporal dos dados, padrões e perfis de citações de palavras e assuntos não podem ser respondidas ou identificadas apenas com a Visualização implementada na versão anterior da aplicação. Diante disso, nosso objetivo principal foi adicionar esse novo recurso e implementar novas funcionalidades úteis ao usuário.

O foco na criação de modelos de Visualização que fossem úteis às tarefas do processo de análise qualitativa realizada pela especialista e a preocupação com um ambiente de análise interativo, aliados à utilização de algumas métricas básicas de Mineração de Texto, foram imprescindíveis para o desenvolvimento deste trabalho. Acreditamos ainda que, embora a aplicação tenha sido desenvolvida para o propósito específico de analisar dados de um fórum de discussões, isso não seja uma limitação para a generalização da aplicação do programa, que pode ser utilizado para analisar dados de outros ambientes que envolvam textos relacionados à discussão coletiva em torno de um tema.



REFERÊNCIAS BIBLIOGRÁFICAS

- [Ankerst, 2001] Ankerst, M. (2001). Visual data mining with pixel-oriented visualization techniques. *ACM SIGKDD Workshop on Visual Data Mining*.
- [Azevedo et al., 2009] Azevedo, B. F. T., Reategui, E., and Behar, P. A. (2009). Estudo de análise qualitativa em fórum de discussão, novas tecnologias na educação. 7(3).
- [Azevedo et al., 2011] Azevedo, B. F. T., Reategui, E., and Behar, P. A. (2011). *Automatic Analysis of Messages in Discussion Forums*. pages 76–81.
- [Chen et al., 2008] Chen, C.-H., Härdle, W., and Unwin, A. (2008). *Handbook of Data Visualization*. Springer, Berlin.
- [Chen et al., 1996] Chen, M.-S., Han, J., and Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):866–883.
- [Chittaro et al., 2003] Chittaro, L., Combi, C., and Trapasso, G. (2003). Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages and Computing*, 14(6):591–620.
- [Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*.
- [Feldman and Sanger, 2007] Feldman, R. and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, New York, EUA.

- [Friendly, 2012] Friendly, M. (2012). Milestones in the history of thematic cartography, statistical graphics, and data visualization.
- [Kechadi and Bertolotto, 2006] Kechadi, M.-T. and Bertolotto, M. (2006). A visual approach for spatio-temporal data mining. In *Information Reuse and Integration, 2006 IEEE International Conference on*, pages 504 –509, Waikoloa Village, HI, EUA.
- [Keim et al., 2010] Keim, D. A., Oelke, D., and Rohrdantz, C. (2010). Analyzing document collections via context-aware term extraction. In *Proceedings of the 14th international conference on Applications of Natural Language to Information Systems, NLDB'09*, pages 154–168, Berlin, Heidelberg. Springer-Verlag.
- [Longhi et al., 2009] Longhi, M. T., Behar, P. A., Bercht, M., and Simonato, G. (2009). Investigando a subjetividade afetiva na comunicação assíncrona de ambientes virtuais de aprendizagem.
- [McCormick et al., 1987] McCormick, B. H., DeFanti, T. A., and Brown, M. D., editors (1987). *Visualization in Scientific Computing*. ACM SIGGRAPH, New York.
- [Morais and Ambrósio, 2007] Morais, E. A. M. and Ambrósio, A. P. L. (2007). Mineração de textos. *Technical Report, Instituto de Informática, Universidade Federal de Goiás, (INF05/07)*.
- [Pacheco Jr., 2011] Pacheco Jr., J. C. (2011). Processo de visual analytics para a análise qualitativa de conteúdo em fóruns de discussão. *PIBIC, id. 16856*.
- [Rinaldi, 2009] Rinaldi, R. P. (2009). *Desenvolvimento Profissional de formadores em exercício: contribuições de um programa online*. PhD thesis, Universidade Federal de São Carlos, Centro de Educação e Ciências Humanas, Curso de doutorado em Educação, São Carlos.
- [Schroeder et al., 2003] Schroeder, W., Martin, K., and Lorensen, B. (2003). The visualization toolkit: An object-oriented approach to 3d graphics, third edition. *Kitware Inc. (formerly Prentice-Hall)*.
- [Shimabukuro, 2004] Shimabukuro, M. H. (2004). *Visualizações Temporais em uma Plataforma de Software Extensível e Adaptável*. PhD thesis, Universidade de São Paulo,

Instituto de Ciências Matemáticas e de Computação, Curso de doutorado em Ciências de Computação e Matemática Computacional, São Carlos.

[Stavrianou and Chauchat, 2008] Stavrianou, A. and Chauchat, J.-H. (2008). Opinion mining issues and agreement identification in forum texts. pages 51–58.

[Stoffel et al., 2010] Stoffel, A., Spretke, D., Kinnemann, H., and Keim, D. A. (2010). Enhancing document structure analysis using visual analytics. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 8–12, New York, NY, USA. ACM.

[Strobelt et al., 2009] Strobelt, H., Oelke, D., Rohrdantz, C., Stoffel, A., Deussen, O., and Keim, D. (2009). Document cards: A top trumps visualization for documents. *IEEE transactions on visualization and computer graphics (tvcg - infovis)*. 15:1145–1152.

[van Zudilova-Seinstra et al., 2009] van Zudilova-Seinstra, E., Adriaansen, T., and van Liere, R. (2009). *Trends in Interactive Visualization*. Springer, London.

[Weinzierl, 2012] Weinzierl, H. (2012). New digital universe study reveals big data gap: Less than 1% of world's data is analyzed.

[Yu et al., 2012] Yu, C., Yurovsky, D., and Xu, T. (2012). Visual data mining: An exploratory approach to analyzing temporal patterns of eye movements. *Infancy*, 17(1):33–60.

CAPÍTULO 6

ANEXO



BoardWords

Manual do Usuário

Versão 1.0.0



SUMÁRIO

| | | |
|----------|------------------------|-----------|
| 1 | Introdução | 1 |
| 2 | Funcionalidades | 2 |
| 2.1 | Textos | 3 |
| 2.2 | Stopwords | 4 |
| 2.3 | Processar | 4 |
| 2.4 | Palavras | 4 |
| 2.5 | Calendário | 5 |
| 2.6 | Painel | 7 |
| 2.7 | Salvar | 10 |
| 2.8 | Editor | 10 |
| 2.9 | Ajuda | 10 |
| 3 | Funcionamento | 11 |

CAPÍTULO 1

INTRODUÇÃO

BoarWords é uma aplicação da área de Computação Gráfica, que agrupa técnicas de Visualização de Informações e Mineração de Texto conjuntamente. Foi desenvolvido com o objetivo de auxiliar na análise de dados de texto gerados a partir de um fórum de discussões em um AVA (Ambiente Virtual de Aprendizagem). Dessa forma, a aplicação possui interesses específicos, mas acreditamos que isso não seja um fator limitante para sua generalização e utilização em outros ambientes de análise.

O nome da aplicação é uma adaptação do termo “*board of words*” (trad. tábua de palavras), que faz referência direta ao aspecto visual das principais funcionalidades da aplicação. Esse projeto é uma extensão do projeto desenvolvido por Pacheco Jr., intitulado “Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo em Fóruns de Discussão” [Pacheco Jr., 2011], e foi concebido baseado nos bons resultados obtidos no mesmo. Através dele, foi possível extrair novas informações sobre o conjunto de dados analisados por Rinaldi [Rinaldi, 2009], contribuindo para uma compreensão mais ampla do escopo analisado.

A principal contribuição desse projeto é a inserção de novas funcionalidades de Visualização, das quais o atributo temporal é o mais relevante, justificando o termo ‘Conteúdo-Temporal’ usado no título do trabalho. Através disso, será possível realizar uma análise mais detalhada dos dados textuais, disponibilizando ao usuário ou pesquisador um conjunto maior de informações a seu respeito.

CAPÍTULO 2

FUNCIONALIDADES

As principais funcionalidades do BoardWords se referem ao esquema natural do processo de Mineração de Dados definido por Fayyad [Fayyad et al., 1996], conforme ilustrado na Figura 2.1.

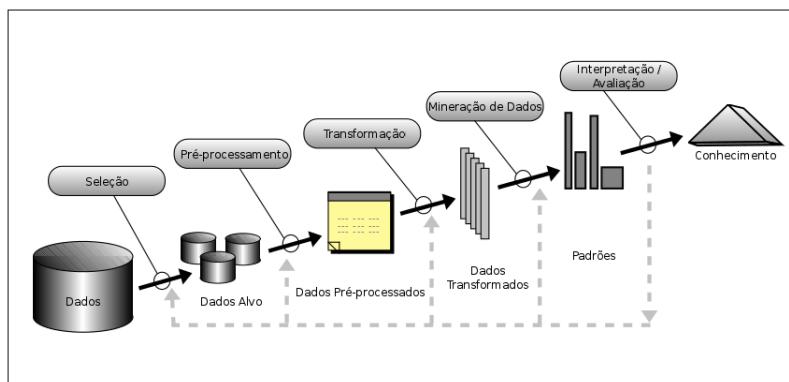


Figura 2.1: etapas de KDD segundo Fayyad; adaptado de [Fayyad et al., 1996].

No menu principal (Figura 2.2) estão localizadas as funções relacionadas às etapas de Seleção, Pré-processamento, Transformação e Mineração de Dados descritas por Fayyad, representadas pelos botões *Textos*, *Stopwords*, *Processar* e *Painel*. As demais funcionalidades servem como apoio à esse processo, oferecendo recursos de interatividade para trabalhar com os dados.

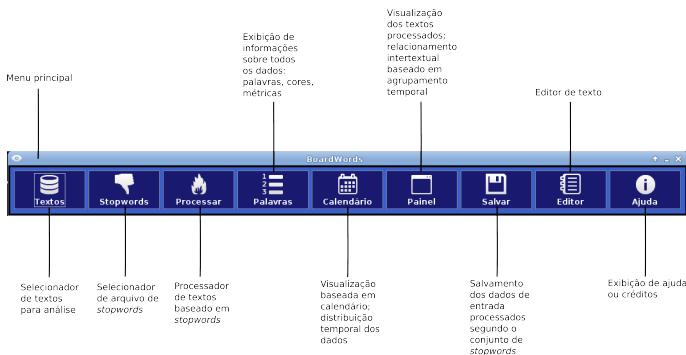


Figura 2.2: menu principal.

Observação: para o correto funcionamento do programa, todos os arquivos utilizados devem necessariamente ser do formato *txt* e possuir codificação de caracteres ISO 8859-1. Essa aplicação também depende da instalação e configuração correta do *software JRE*, na versão 6 ou superior.

2.1 Textos

Textos é o meio de seleção do diretório contendo os dados ou arquivos de texto a serem analisados. Os arquivos devem ser nomeados de acordo com o esquema *mmddaaaa_hh-mm.txt*, que representam respectivamente o mês, dia, ano, hora e minuto de criação dos textos. Pelo fato dos dados analisados nesse projeto serem originários de um fórum de discussões em um AVA específico, o conteúdo dos textos segue o padrão abaixo:

```
De: <remetente>
Para: <destinatário>
dd/mm/aaaa hh:mm:ss
<texto>
```

O padrão utilizado nos arquivos de texto não é necessário para se utilizar a aplicação, pois os seus dados não são aproveitados, já que os atributos temporais são derivados apenas do nome dos arquivos, que obrigatoriamente devem seguir o modelo já citado. Dessa forma, a aplicação pode ser aplicada mais facilmente em outros ambientes, aumentando seu campo de atuação. Entretanto, levando em conta o objetivo inicial proposto para esse trabalho, os arquivos de texto devem possuir no mínimo três linhas de dados (relativas ao cabeçalho dos arquivos analisados) que, por serem desprezadas, podem possuir qualquer conteúdo.

2.2 Stopwords

Stopwords é o meio de seleção do arquivo que contém as *stopwords* ou palavras a serem filtradas dos dados de entrada, selecionados pelo botão *Textos*. O arquivo deve ser do tipo *txt* e seu conteúdo deve possuir o formato abaixo, onde cada $\langle \text{palavra}_n \rangle$ representa a *stopword* da n -ésima linha do arquivo:

```
<palavra_1>
<palavra_2>
...
...
```

2.3 Processar

Processar realiza o processamento dos dados de entrada baseado no arquivo de *stopwords*, separando as palavras ou em *startwords* ou em *stopwords* e removendo o conjunto de caracteres apresentados na Figura 2.3, considerados nulos nessa análise textual, assim como espaços em branco.

Figura 2.3: conjunto de caracteres considerados nulos.

2.4 Palavras

Palavras é o meio de exibição das *startwords* com suas respectivas cores de representação e dados das métricas aplicadas. A função é ilustrada pela Figura 2.4, e o conjunto de suas funcionalidades é apresentado na Tabela 2.1.

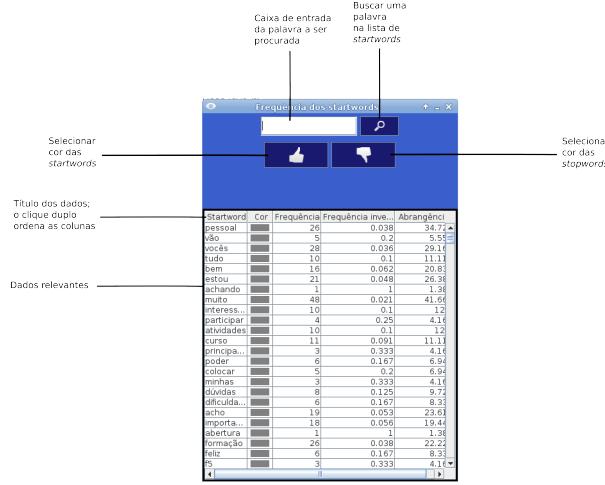


Figura 2.4: janela de exibição das frequências das palavras.

| Ícone | Texto | Ação | Função |
|--------------------|--------------|---|--------|
| | Clique único | Realiza buscas por <i>startwords</i> na lista de palavras. | |
| | Clique único | Seleciona a cor das <i>startwords</i> . | |
| | Clique único | Seleciona a cor das <i>stopwords</i> . | |
| Startword | Clique único | Exibi uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente. | |
| Cor | Clique único | Exibi a cor de representação de uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente. | |
| Frequência | Clique único | Exibi a quantidade absoluta (em todos os textos) de ocorrências de uma palavra. Se o título for clicado, ordena os valores em ordem crescente/descrescente. | |
| Frequência inversa | Clique único | Exibi o inverso da quantidade absoluta de ocorrências de uma palavra, isto é, $ft_i = \frac{1}{ft_a}$, onde ft_a é a frequência total absoluta e ft_i é a frequência total inversa dessa palavra em todos os textos. Se o título for clicado, ordena os valores em ordem crescente/descrescente. | |
| Abrangência (%) | Clique único | Exibi o percentual de textos em que uma palavra é citada. Se o título for clicado, ordena os valores em ordem crescente/descrescente. | |

Tabela 2.1: subfuncionalidades de *Palavras*.

2.5 Calendário

Calendário é o meio de exibição da distribuição da data de criação dos dados sobre a forma de um calendário, ilustrado pela Figura 2.5. O conjunto de suas funcionalidades é apresentado na Tabela 2.2.

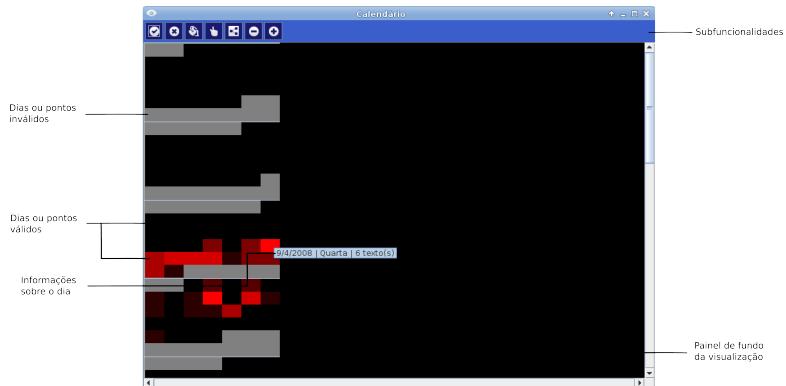


Figura 2.5: janela de exibição da distribuição dos textos.

| Ícone | Texto | Ação | Função |
|-------|--|--------------|--|
| ✓ | | Clique único | Seleciona a cor dos dias válidos no calendário, onde será aplicado o percentual relativo à quantidade de textos presentes em determinado dia, representado por pontos ou retângulos. |
| ✗ | | Clique único | Seleciona a cor dos dias inválidos (por exemplo, dia 32). |
| ⌚ | | Clique único | Seleciona a cor do painel de fundo da visualização. |
| 👉 | | Clique único | Seleciona a cor de seleção dos dias. |
| ➡ | | Clique único | Exporta os dados contidos nos dias selecionados para uma nova janela de Visualização. |
| – | | Clique único | Diminui a escala da visualização. |
| + | | Clique único | Aumenta a escala da visualização. |
| | Rolagem do scroll do mouse para baixo | | Diminui a escala da visualização. |
| | Rolagem do scroll do mouse para cima | | Aumenta a escala da visualização. |
| | Clique único sobre um dia válido | | Seleciona/deseleciona o dia. |
| | Clique duplo sobre um dia válido | | Seleciona/deseleciona o mês referente ao dia. |
| | Clique duplo sobre o painel de fundo | | Deseleciona os dias selecionados. |
| | Posicionamento do cursor sobre um dia válido | | Exibe as informações sobre o dia, no formato $dd/mm/aaaa < dia_da_semana > n \text{ texto(s)}$, onde $dd/mm/aaaa$ é o formato de data, $< dia_da_semana >$ corresponde ao dia da semana daquele dia (segunda, por exemplo), e n é a quantidade de textos criados naquele dia. |

Tabela 2.2: subfuncionalidades de *Calendário*.

2.6 Painel

Painel é o meio de criação de um ambiente de Visualização contendo todos os dados processados. O conjunto de suas funcionalidades é apresentado na Tabela 2.3, na Tabela 2.4 e na Tabela 2.5.

| Ícone | Texto | Ação | Função |
|-------|--------------|------|---|
| | Clique único | | Processa e exibe a Visualização baseado nas configurações de entrada. |
| | Clique único | | Configura os parâmetros da Visualização. Através dessa função, é exibida uma janela de configuração da Visualização a ser realizada, que tem como princípio o conceito de agrupamento de dados ou textos. O usuário pode escolher entre agrupar os textos por períodos de dias, meses ou anos, com valores entre [1-999], denominado <i>agrupamento de datas</i> , e períodos de minutos e horas, com valores variáveis [1-1440] e [1-24], respectivamente, denominado <i>agrupamento de horários</i> . |
| | Clique único | | Importa um arquivo de configurações de Visualização, no formato: <code>grouping:<boolean></code> <code>legend:<boolean></code> <code>frequencies:<boolean></code> <code>grouping_dates:<day month year></code> <code>value_grouping_dates:<integer></code> <code>grouping_schedules:<minute hour></code> <code>value_grouping_schedules:<integer></code> |
| | Clique único | | Exporta as configurações de Visualização atuais para um arquivo. |
| | Clique único | | Seleciona a cor dos identificadores de horário de criação dos textos e dos agrupamentos. |
| | Clique único | | Seleciona a cor de fundo do painel de visualização. |
| | Clique único | | Seleciona a cor de seleção dos textos e palavras da Visualização. |
| | Clique único | | Seleciona palavras no texto, baseado em um arquivo de entrada com aspecto similar ao arquivo de <i>stopwords</i> , onde cada linha possui uma palavra a ser destacada. |
| | Clique único | | Seleciona a cor da fonte dos textos onde são exibidas as palavras selecionadas. |

Tabela 2.3: subfuncionalidades de *Painel*.

| Ícone | Texto | Ação | Função |
|---|---|--|--|
| T | | Clique único | Exibe as palavras selecionadas sobre os arquivos de texto originais. |
|  | | Clique único | Seleciona os textos, baseado em um arquivo de entrada com aspecto similar ao arquivo de <i>stopwords</i> , onde cada linha possui uma palavra a ser destacada. |
|  | | Clique único | Exporta os textos selecionados para uma nova janela de Visualização. |
|  | | Clique único | Diminui a escala da visualização. |
|  | | Clique único | Aumenta a escala da visualização. |
| | Rolagem do <i>scroll</i> do mouse para baixo | | Diminui a escala da visualização. |
| | Rolagem do <i>scroll</i> do mouse para cima | | Aumenta a escala da visualização. |
| | Clique duplo na área de exibição dos textos originais | Deseleciona todas as palavras destacadas. | |
| | Clique duplo sobre o painel de Visualização | Deseleciona os textos ou palavras selecionados. | |
| | Posicionamento do cursor sobre o identificador do horário de criação dos textos | Exibe a data de criação do texto. | |
| | Clique duplo sobre as palavras do texto | Exibe o texto original em que está contida a palavra, sendo possível visualizar os demais textos analisados pelos botões de próximo texto e texto anterior na janela de exibição dos textos originais. | |
| | Posicionamento do cursor sobre as palavras do texto | Exibe a data e hora de criação do texto no menu da janela (lado direito) e as informações do ponto; se <i>Frequência</i> não estiver selecionado, exibe apenas a palavra representada pelo ponto, senão, exibe as seguintes informações: | |
| | | <p>p:<palavra></p> <p>fa:<fp_a></p> <p>fr:<fp_r></p> <p>fi:<fp_i></p> | |

Tabela 2.4: subfuncionalidades de *Painel*.

| Ícone | Texto | Ação | Função |
|-------|-------------|--------------|--|
| | Dia | Clique único | Agrupa os textos por grupos cronológicos de dias. |
| | Mês | Clique único | Agrupa os textos por grupos cronológicos de meses. |
| | Ano | Clique único | Agrupa os textos por grupos cronológicos de anos. |
| | Minuto | Clique único | Agrupa os textos por subgrupos cronológicos de minutos. |
| | Hora | Clique único | Agrupa os textos por subgrupos cronológicos de horas. |
| | Agrupamento | Clique único | Insere/remove o agrupamento de dados para realizar a Visualização. |
| | Legenda | Clique único | Exibe/omite os identificadores dos agrupamentos e horários de criação dos textos. |
| | Frequência | Clique único | Exibe/omite as frequências do cálculo de relevância das palavras no texto em que se encontram. As métricas calculadas são frequência absoluta (fpa – frequência parcial absoluta), frequência relativa (fp_r – frequência parcial relativa) e frequência inversa de documentos (fpi – frequência parcial inversa), definidas a seguir [Morais and Ambrósio, 2007]: <ul style="list-style-type: none"> • fpa: quantidade de vezes em que a palavra é citada no texto. • fp_r: razão entre a fpa e a quantidade de palavras relacionadas no texto processado, dada pela função $fp_r = \frac{fpa}{n}$, onde n é a quantidade total de palavras do texto processado. Através desse cálculo é possível ter a percepção do percentual de ocorrência que da palavra sobre o texto. • fpi: razão entre a fpa e o número de documentos em que a palavra ocorre ou a_a (abrangência absoluta), de onde é derivado o cálculo da abrangência das palavras em relação a todos os textos. A função que define fpi é dada por $fpi = \frac{fpa}{a_a}$. Através desse cálculo é possível destacar termos ou palavras que aparecem em menor escala nos textos e diminuir a importância dos termos mais freqüentes, que aparecem mais comumente nos textos. |

Tabela 2.5: subfuncionalidades das configurações de Visualização de *Painel*.

2.7 Salvar

Salvar é o meio de armazenamento dos arquivos de entrada processados com base no arquivo de filtro contendo as *stopwords*. Quando os arquivos de texto são processados, são removidas as três primeiras linhas relativas ao cabeçalho do texto e todas as palavras que são *stopwords*. Os textos são comprimidos em uma única linha, onde as palavras, transformadas em minúsculas, são separadas por espaçamento único.

2.8 Editor

Editor é o meio de criação de um ambiente de edição de textos, que pode ser utilizado para várias finalidades, inclusive para o usuário registrar informações extraídas dos dados analisados. O conjunto de suas funcionalidades é apresentado na Tabela 2.6.

| Ícone | Texto | Ação | Função |
|-------|---|--|--------|
| | Clique único | Seleciona um arquivo de texto para edição. | |
| | Clique único | Cria um novo arquivo de texto. | |
| | Clique único | Salva o arquivo de texto atual, sobrescrevendo-o se este já existir. | |
| | Clique único | Salva o arquivo de texto atual com nome e localização variáveis. | |
| | Clique único | Destaca uma palavra no texto. | |
| | Clique duplo na área de exibição dos textos | Deseleciona todas as palavras destacadas. | |

Tabela 2.6: subfuncionalidades de *Editor*.

2.9 Ajuda

Ajuda é o meio de exibição do manual da aplicação e dos créditos de desenvolvimento.

CAPÍTULO 3

FUNCIONAMENTO

O BoardWords se baseia no modelo de Mineração de Dados definido por Fayyad (ver Figura 2.1). Dessa forma, as etapas necessárias para analisar dados através da aplicação são:

1. Seleção do diretório contendo os dados ou arquivos de entrada, através do botão *Textos* (menu principal);
2. Seleção do arquivo de *stopwords*, através do botão *Stopwords* (menu principal);
3. Processamento dos dados de entrada baseado no arquivo de *stopwords*, através do botão *Processar* (menu principal);

O usuário também pode utilizar as funções de *Editor* e *Ajuda*, independentes de outros procedimentos. Realizados os procedimentos necessários, todas as funcionalidades da aplicação tornam-se acessíveis ao usuário. Através do botão *Palavras* (menu principal), podem ser visualizadas informações sobre todos os dados de entrada, como *startwords*, frequências das palavras e abrangência do termo em todos os textos.

No botão *Calendário* (menu principal), pode ser realizada uma análise mais superficial ou geral do conjunto de dados de entrada, sendo possível identificar e selecionar regiões de maior interesse para uma análise mais detalhada, como picos de atividade, destacados pela maior intensidade de coloração.

O *Painel* é a principal funcionalidade da aplicação, pois através dele são realizadas as análises mais detalhadas do conjunto de dados. Os relacionamentos intertextuais são

obtidos pelo princípio de agrupamento temporal, no qual o usuário pode organizar os textos por períodos de tempo, que podem ser por datas (dias, meses ou anos) ou por horários (minutos ou horas). No caso do agrupamento de horários, a entrada de valores divisores de 6 (1 minuto ou 6 horas, por exemplo) implica na coloração dos identificadores de agrupamento baseado na transição simbólica da luminosidade do sol sobre o dia, representada por 4 faixas de coloração. Esse artifício permite identificar mais rapidamente períodos relacionados pela proximidade.

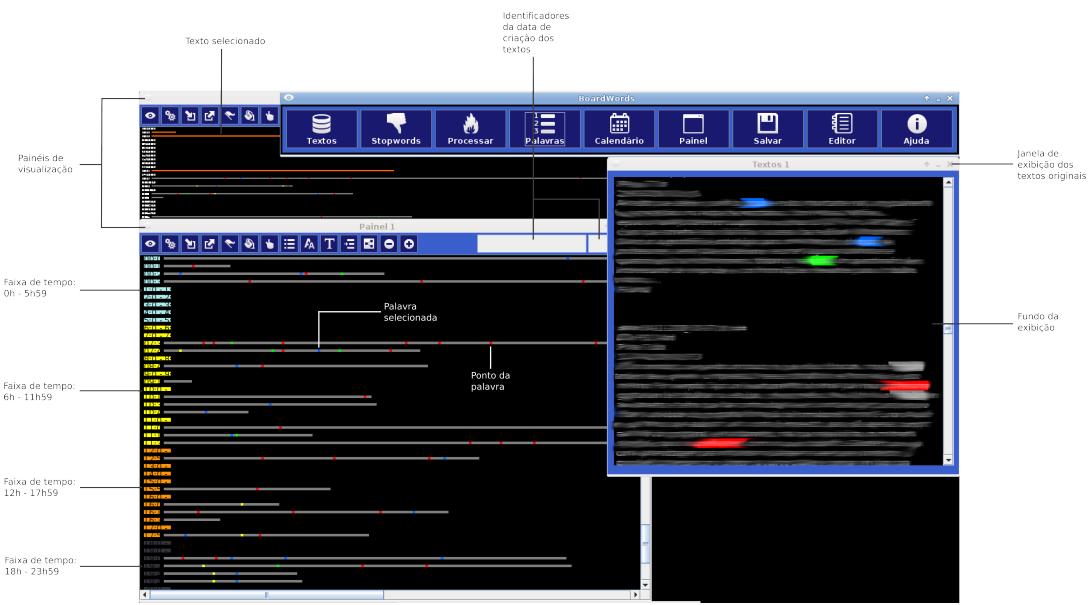


Figura 3.1: ambiente de análise múltipla.

Na Figura 3.1, é exibido um exemplo de análise múltipla da aplicação. Através da janela de exibição dos textos originais, o usuário pode relacionar a visualização com os textos originais (sem processamento), destacando as palavras que achar conveniente para a análise. Nesse exemplo, os textos originais aparecem borrados para manter o sigilo dos dados analisados.



REFERÊNCIAS BIBLIOGRÁFICAS

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence.*

[Morais and Ambrósio, 2007] Morais, E. A. M. and Ambrósio, A. P. L. (2007). Mineração de textos. *Technical Report, Instituto de Informática, Universidade Federal de Goiás, (INF_05/07)*.

[Pacheco Jr., 2011] Pacheco Jr., J. C. (2011). Processo de visual analytics para a análise qualitativa de conteúdo em fóruns de discussão. *PIBIC, id. 16856.*

[Rinaldi, 2009] Rinaldi, R. P. (2009). *Desenvolvimento Profissional de formadores em exercício: contribuições de um programa online.* PhD thesis, Universidade Federal de São Carlos, Centro de Educação e Ciências Humanas, Curso de doutorado em Educação, São Carlos.

PROJETO DE PESQUISA

Programa Institucional de Bolsas de Iniciação Científica – PIBIC 2012/2013

Orientador: Prof. Dr. Milton Hirokazu Shimabukuro (DMC).

Aluno: João Vítor Antunes Ribeiro (Bacharelado em Ciência da Computação).

Colaboração: Profa. Dra. Renata Portela Rinaldi (Departamento de Educação).

Unidade: Faculdade de Ciências e Tecnologia – Campus de Presidente Prudente
Departamento de Matemática e Computação - DMC.

Área: Ciência da Computação.

Título: Visualização Interativa de Dados para Suporte à Atividade de Análise Qualitativa ‘Conteúdo-Temporal’ de Fóruns de Discussão.

Agradecimentos: Alisson Fernando Coelho do Carmo
Leonardo Tadashi Nozawa