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

João Vítor Antunes Ribeiro[1], Milton Hirokazu Shimabukuro[1], Renata Portela Rinaldi[2]

[1]Departamento de Matemática e Computação, Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista –UNESP
j.antunes.cc@gmail.com, miltonhs@fct.unesp.br
[2]Departamento de Educação, Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista – UNESP
renata.rinaldi@fct.unesp.br

Resumo*: Este artigo tem por objetivo a investigação do uso de recursos computacionais para suporte ao processo de análise de documentos textuais pela integração de técnicas analíticas empregadas em Mineração de Texto e Visualização de Dados, buscando beneficiar-se da união dos benefícios de cada abordagem. Nesse sentido, descreve o processo de desenvolvimento de uma ferramenta de Mineração Visual de Texto (MVT) para auxiliar na análise de uma base de dados de um fórum de discussões em AVA. Está vinculado ao projeto "Processo de Visual Analytics para a Análise Qualitativa de Conteúdo em Fóruns de Discussão" (PIBIC, id. 16856), realizado com os mesmos fins. A principal contribuição desse novo trabalho é a adição de novas funcionalidades, das quais a Visualização temporal é a mais significativa, ressaltando ser papel fundamental para que a analista pudesse extrair novas informações à respeito de sua base de dados.*
Palavras chave: *KDD, Mineração de Texto, Visualização temporal, fóruns de discussão.*

Abstract: *This article aims to investigate the use of computational resources to support the process of analyzing textual documents by integrating analytical techniques employed in Text and Data Visualization Mining, seeking to benefit from the union of each approach. In this sense, it describes the development process of a Visual Text Mining Tool (MVT) to assist in the analysis of a database of a forum for discussions on AVA. It is linked to the project "Visual Analytics Process for the Qualitative Analysis of Content in Forums" (PIBIC, id. 16856), held for the same purposes. The main contribution of this new work is the addition of new features, of which the temporal visualization is the most significant, underscoring be essential for the analyst could extract new information about its database.*
Keywords*: KDD, Text Mining, temporal Visualization, discussion forums.*

1 Introdução

A evolução da tecnologia tem influenciado diretamente a interação entre os indivíduos, tanto pelo conteúdo gerado quanto pelo modo de se comunicar. De forma geral, meios digitais de comunicação, como e-mails, redes sociais e fóruns de discussão, são amplamente utilizados em razão de envolver um processo mais prático, rápido e com menos custo do que abordagens clássicas, como cartas ou telefonemas. Essa migração para o conteúdo digital é fundamental para ampliar a

disseminação dessas informações geradas, facilitar a socialização de experiências e favorecer a colaboração, para a construção coletiva de conhecimento.

No âmbito da manipulação de informações digitais, o Ambiente Virtual de Aprendizagem (AVA) se destaca como importante recurso de desenvolvimento interativo e colaborativo de conhecimento entre seus participantes. Esse tipo de ambiente gera informações úteis sobre relações interpessoais, como opiniões a respeito de determinados temas e horários das postagens em fóruns de discussão, possibilitando conhecer perfis de utilização e aprimorar abordagens educacionais. Essas informações podem ser manipuladas e processadas por ferramentas que permitem a extração de conhecimento em bases de dados (Knowledge Discovery in Databases - KDD), permitindo a descoberta de informações novas acerca de um conjunto de dados.

Neste estudo, a fonte de informação a ser tratada é formada por mensagens contidas em fóruns de discussão provenientes de um AVA, isto é, informações presentes em um contexto educacional. No entanto, a abordagem utilizada e empregada neste trabalho não está restrita aos dados encontrados em um AVA, fato que não impacta a generalização dessa solução de extração de conhecimento de dados digitais para ser aplicada em outro contexto. Para permitir a extração de informação e facilitar a interpretação dos significados, foram utilizadas técnicas de Visualização Interativa de Dados e aspectos específicos de Mineração de Textos. A Visualização Interativa de Dados pode ser vista como um processo da KDD, geralmente posterior ao pré-processamento, organização ou Mineração de Dados. A Mineração de Texto é um escopo específico do contexto de Mineração de Dados, no qual a fonte de informações é encontrada em conjuntos de textos que apresentam forma não estruturada.

Este trabalho está organizado em seções que contribuem para o entendimento das ações do projeto realizado com o mesmo nome deste artigo. Assim, na próxima Seção é feita uma breve descrição dos objetivos gerais do projeto, seguido por uma Seção sobre os materiais e métodos e outra sobre resultados e discussões, e por fim uma Seção com as principais conclusões.

2 Objetivos

No projeto "Processo de *Visual Analytics* para a Análise Qualitativa de Conteúdo de Fóruns de Discussão" (PACHECO JR., 2011) foi desenvolvida uma aplicação para auxiliar a pesquisadora na análise de dados textuais gerados a partir de um fórum de discussões de um AVA. Os resultados do projeto inicial foram bastante satisfatórios e animadores por terem trazido novas informações àquelas conclusões que a especialista obteve pela análise "manual" (avaliação pela leitura de mensagem por mensagem), tendo contribuído para a descoberta de novos conhecimentos sobre o conjunto de textos (RINALDI, 2009). Os resultados obtidos com o projeto inicial fundamentaram sua extensão. Para o novo projeto, houve o desenvolvimento de uma nova ferramenta computacional, abordada neste trabalho, para possibilitar a visualização textual do conteúdo das mensagens agregada com os aspectos temporais referentes à data e horário de postagem.

Dessa forma, configura-se como objetivo geral neste trabalho a investigação do uso de recursos computacionais para suporte ao processo de análise de documentos textuais pela integração de técnicas analíticas empregadas em Mineração de Texto e Visualização de Dados, buscando beneficiar-se da união dos benefícios de cada abordagem. 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 possibilitar estratégias de intervenção do especialista.

Trabalhos sobre a análise de conteúdo no mesmo contexto educacional de um AVA têm sido realizados e produzidos resultados positivos em relação à interpretação dos dados e extração de informações valiosas (AZEVEDO *et al.*, 2009, 2011), (LONGHI *et al.*, 2009) e (STAVRIANOU & CHAUCHAT, 2008). Além disso, técnicas visuais para o tratamento de documentos são constantemente utilizadas e aplicadas em diferentes cenários, principalmente em razão dos benefícios relacionados à capacidade de percepção visual e facilidade de interpretação de representações mais significativas (STOFFEL *et al.*, 2010), (KEIM *et al.*, 2010) e (STROBELT *et al.*, 2009). O processamento da

dimensão temporal é uma tarefa essencial e que pode ser utilizada em diferentes outros domínios de informação, como medicina (YU *et al.*, 2012), (CHITTARO *et al.*, 2003) e meio ambiente (KECHADI & BERTOLOTTO, 2006). Diante deste cenário, os objetivos específicos deste trabalho podem ser resumidos em: inclusão da dimensão temporal da análise, ampliação das funcionalidade para a manipulação interativa dos dados e verificação da utilidade de métricas usadas em Mineração de Texto.

3 Material e Métodos

Para desenvolver este trabalho, realizou-se uma revisão e complementação da bibliografia utilizada no trabalho desenvolvido por Pacheco Jr.. Foi preciso estudar a ferramenta construída para o projeto inicial (PACHECO JR., 2011), com o propósito de possibilitar a inserção de novas funcionalidades, as quais a projeção temporal é a mais relevante. O uso de métricas também foi inserido no escopo do presente trabalho, fato que valoriza a análise interativa dos dados.

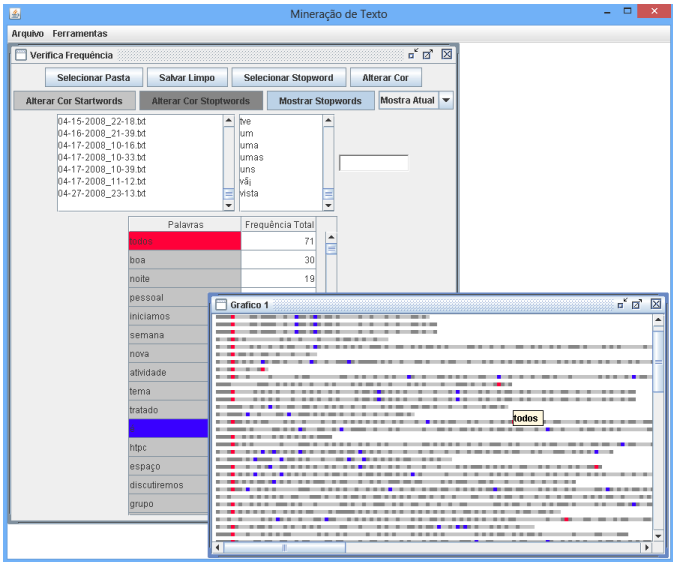

Figura 1 – Representação visual de texto por pixel, extraída do programa desenvolvido por Pacheco Jr. (2011).

A partir da necessidade da pesquisadora em aplicar recursos computacionais para extrair mais informações do conjunto de dados em relação à análise manual dos mesmos, iniciou-se um estudo

sobre as técnicas de Visualização de Informação que pudessem ser úteis à esse propósito. No projeto realizado por Pacheco Jr. (2011), propôs-se uma representação geral dos dados por meio de uma técnica de visualização de dados baseada em pixel, onde todos os textos do fórum de discussões foram transformados em linhas de pixels (célula correspondente à menor unidade de representação em um monitor), onde cada pixel representava uma palavra do texto. Dessa forma, dispondo as linhas (textos) uma embaixo da outra, pôde-se obter uma visão clara de todos dados simultaneamente, conforme ilustrado na **Figura 1**.

É possível observar que o programa foi composto por uma janela de visualização, onde as informações são mostradas, uma janela de configurações, onde estava concentrada a maior parte das funcionalidades, e um menu principal, por onde eram acessadas essas janelas. Nesse menu, há ainda um editor de textos simples, para o caso em que o usuário queira criar ou modificar algum texto usando a própria ferramenta.

Para usar o programa, o usuário deve selecionar através do menu o diretório que contém os arquivos de entrada (textos) e 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* (palavras relevantes), 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 obteve 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 postagens escritas em 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 da nova ferramenta desenvolvida.

Para inserir o atributo temporal, foi decidido manter 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 continuou sendo Java para *Desktop*, por motivos relacionados à portabilidade.

Com o objetivo de tornar mais fácil a referência à ferramenta desenvolvida nesse trabalho, 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* (do português "tábua de palavras"), referência direta ao aspecto das representações visuais presentes na ferramenta.

Na primeira fase do projeto, foi sugerida uma separação da representação visual em dois tipos: normal e temporal. Acreditou-se que simplesmente inserir um novo atributo sobre aqueles dados da representação da primeira visualização (**Figura 1**) poderia ocasionar excesso de informações e, consequentemente ser ineficaz para o que se propunha.

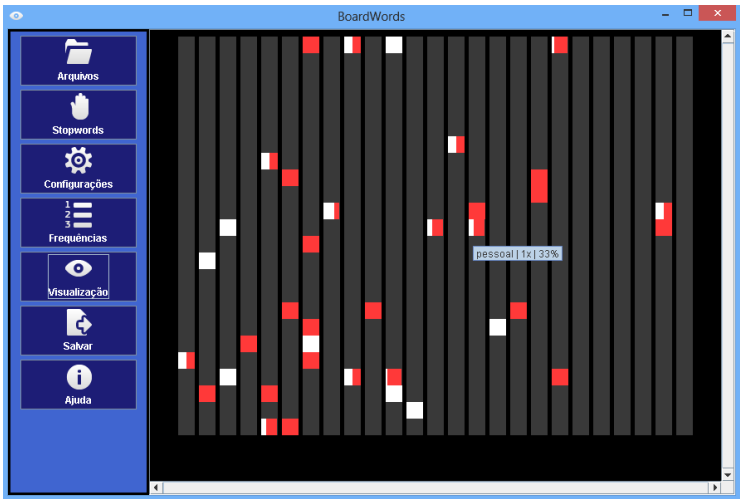

Figura 2 – Primeiro modelo de Visualização temporal proposto.

Na **Figura 2** é ilustrado um exemplo da representação visual proposta. Cada coluna representa um conjunto de dias, meses ou anos, e o espaço vertical das colunas representa a "linha do tempo" no

respectivo intervalo temporal. Por exemplo, para um conjunto de textos que foram escritos em um determinado ano, e ao construir a 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. Vale ressaltar que essa contagem se inicia a partir da data do primeiro *post* da base de textos. Esse esquema pode ser compreendido através da **Figura 3**.

Figura 3 – Esquema da primeira Visualização temporal desenvolvida.

Na **Figura 3**, o comprimento das colunas (vertical) é relativo ao passar das horas de cada dia do conjunto de arquivos de cada coluna. Em uma configuração horizontal por mês com valor 9, têm que “a cada 9 meses existirá um conjunto de textos”. Supondo que a configuração seja minuto e o valor seja 30, “existirá uma célula a cada 30 minutos, começando das 0h00 de cada dia”. Desta forma, existirá 48 vezes o período de tempo de 30 minutos. A leitura dessa Visualização deve ser feita da esquerda para a direita e de cima para baixo, na qual os blocos representam o conjunto de palavras previamente incluídas para serem visualizadas. Dentro de um bloco são exibidos vários retângulos, que podem ter diferentes larguras e cores. A largura de cada retângulo simboliza o percentual aproximado que aquela palavra (representada pelo retângulo) possui em relação às ocorrências de todas as palavras do conjunto incluído dentro daquele intervalo de tempo e de dias. Para facilitar a interpretação é possível identificar cada palavras posicionando-se o cursor do mouse sobre o retângulo desejado; além da palavra, também

é exibida a quantidade de ocorrências e o percentual da mesma, ao clicar é exibido o conjunto de textos daquele bloco.

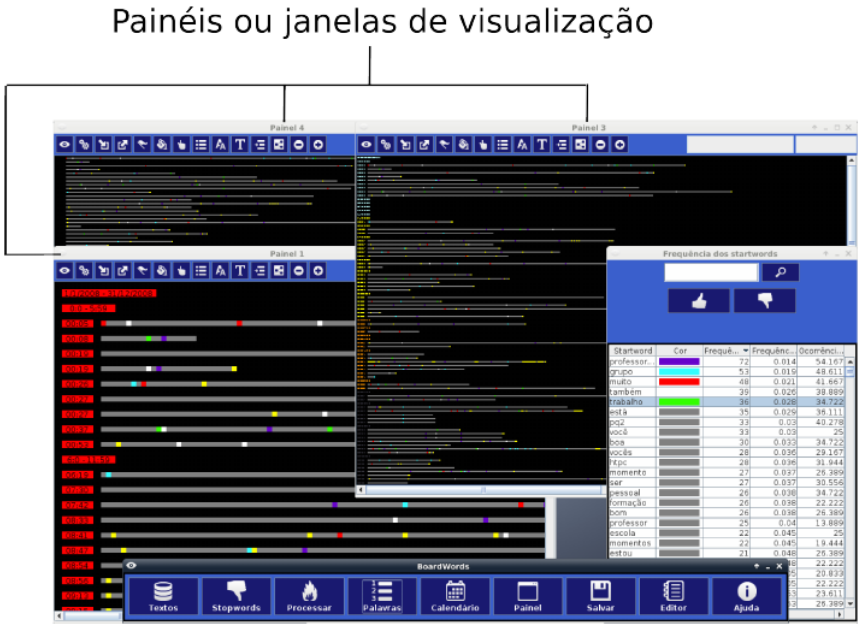

Figura 4 – Modelo de Visualização temporal final: múltiplos painéis em funcionamento simultâneo.

Diante da técnica desenvolvida, foi concluído que esse modelo de Visualização (**Figura 2**) era pouco eficiente e incompleto para auxiliar na descoberta de todas as informações que se almejava conhecer, pois ele não permitia a descoberta de conhecimento 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 da distribuição temporal orientada por horário de criação do texto e frequências orientadas pelo tamanho do bloco de representação da palavra.

No desenvolvimento desse trabalho foi constatado que o conceito de agrupamento temporal dos dias e horários, aplicados na proposta de Visualização temporal da primeira fase do projeto, poderia ser migrada para a Visualização normal, que segue o modelo proposto por Pacheco Jr. (2011). Essa possível junção de ideias poderia proporcionar a descoberta de associações entre textos que foram

escritos em diferentes períodos temporais, abrangendo todo o universo de associações possíveis, como picos de atividade, período predominante de atuação dos usuários do fórum e distância entre participações ou publicações de textos, aliado a informações extraídas pela análise do conteúdo íntegro dos textos.

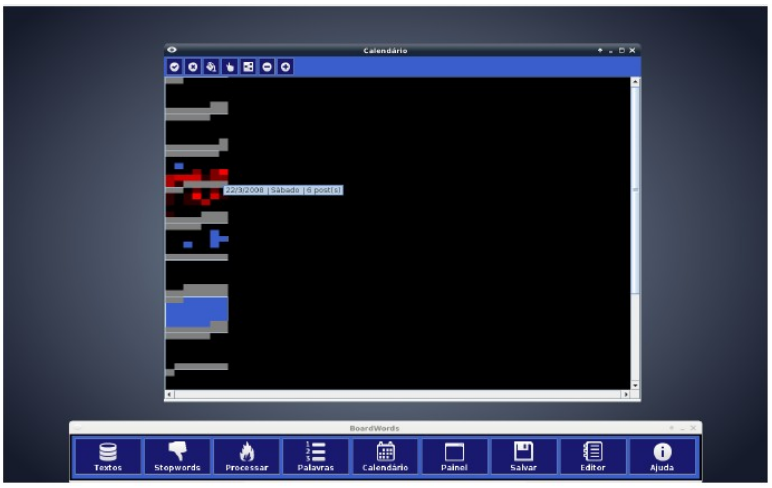

Figura 5 – Visualização temporal baseada em calendário.

Assim, foram desenvolvidas dos tipos de Visualização, uma baseada em colunas de pontos (**Figura 4**) e outra baseada em calendário (**Figura 5**). 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. O agrupamento de períodos temporais permite identificar os textos que pertencem ao mesmo intervalo, sem considerar o horário em que foram escritos. O agrupamento por horário identifica os textos que foram escritos na mesma faixa de horário. Em razão de ambos os agrupamentos serem voltados ao atributo temporal, devem ser aplicados em conjunto. Na **Figura 4** é possível identificar diferentes pontos de vista do mesmo conjunto de dados, exibidos em diferentes painéis. O usuário também conta com ferramentas de aproximação (*zoom*), seleção de textos para exportar em outros painéis, importar/exportar configurações, mudança de cores dos termos ou textos e também de um identificar do progresso de tempo do dia, variando da cor mais clara ao tom mais escuro – ilustrado no painel mais

à direita da **Figura 4 -**, entre outras.

O segundo tipo de Visualização (**Figura 5**) tem como finalidade mostrar a distribuição de atividades do fórum de discussões sobre o ponto de vista de calendário, salientando por intensidade de cor os dias ou pontos em que houveram maior atividade no fórum: quanto maior o número de postagens publicadas, maior a intensidade de cor do ponto (mais vermelho, no caso da **Figura 5**, onde o preto representa a ausência de postagens). Essa funcionalidade ajuda o analista identificar momentos em que houve maior atividade no fórum e assim selecionar as regiões de maior interesse para a análise. Na **Figura 5**, as regiões em azul estão selecionadas, e as regiões em vermelho ilustram a quantidade de postagens publicadas em cada dia (ponto). Além dos recursos de Visualização, como *zoom* e seleção, BoardWords permite múltiplas visualizações sincronizadas, que permite ao analista compreender de forma mais clara os relacionamento entre os dados sobre diferentes configurações de visualização e pontos de vista, além da identificação dos termos destacados nos textos integrais, como ilustra a **Figura 6**.

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 alguns cálculos de frequências comumente utilizados e discutidos na literatura (MORAIS & AMBRÓSIO, 2007), como a distribuição de postagens baseada em calendário e representada pela intensidade de coloração dos pontos, a frequência absoluta das palavras em relação a todos os textos ou de cada texto (quantidade de vezes em que ocorre), a abrangência das palavras (percentual 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 todos os cálculos de frequência, deseja-se extrair ou identificar a relevânca das palavras, para poder priorizar termos mais significativos ao significado dos textos.

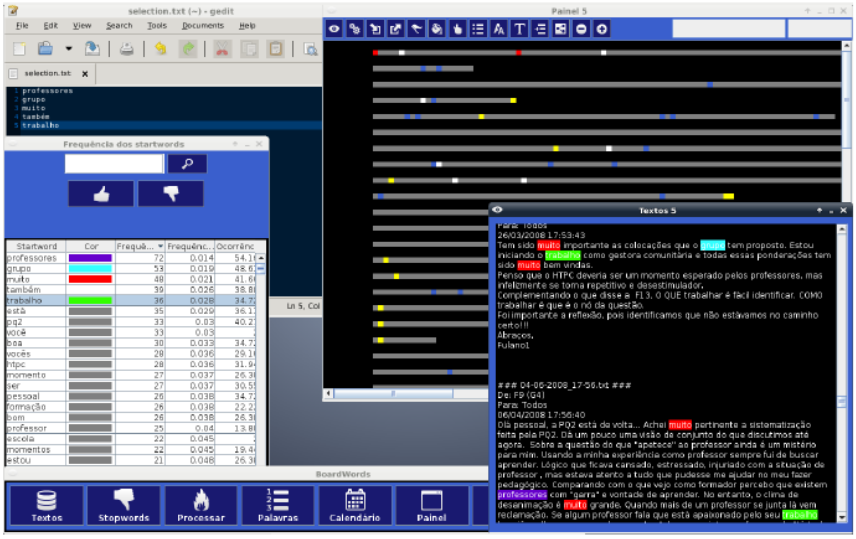

Figura 6 – Destaque de *startwords* nos textos integrais.

4 Conclusões

O desenvolvimento deste trabalho permite concluir 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 ferramenta. Diante disso, o objetivo principal foi adicionar esse novo recurso e implementar novas funcionalidade úteis para o 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. Embora a aplicação tenha sido desenvolvida para o propósito

específico de analisar dados de um fórum de discussões, não é uma limitação para a generalização da aplicação da ferramenta, que pode ser utilizada para analisar dados de outros ambientes que envolvam textos relacionados à discussão coletiva em torno de um tema.

Referências

PACHECO JR., J. C. (2011). Processo de visual analytics para a análise qualitativa de conteúdo em fóruns de discussão. PIBIC, id. 16856.
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.
AZEVEDO, B. F. T., REATEGUI, E., e BEHAR, P. A. (2009). Estudo de análise qualitativa em fórum de discussão, novas tecnologias na educação. 7 (3).
AZEVEDO, B. F. T., REATEGUI, E., e BEHAR, P. A. (2011). Automatic Analysis of Messages in Discussion Forums. páginas 76-81.
LONGHI, M. T., BEHAR, P. A., BERCHT, M., e Simonato, G. (2009). Investigando a subjetividade afetiva na comunicação assíncrona de ambientes de aprendizagem.
STRAVIANOU, A., e CHAUCHAT, J.-H. (2008). Opinion mining issues and agreement identification in fórum texts. Páginas 51-58.
STOFFEL, A., SPRETKE, D., KINNEMANN, H., e KEIM, D. A. (2010). Enhancing document structure analysis using visual analytics. In Proceedings of the 2010 ACM Symposium on Applied Computing, SAC'10, Páginas 8-12, New York, NY, USA. ACM.
KEIM, D. A., OELKE, D., e 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, Páginas 154-168, Berlin, Heidelberg. Springer-Verlag.
STROBELT, H., OELKE, D., ROHRDANTZ, C., STOFFEL, A., DEUSSEN, O., e KEIM, D. (2009). Document cards: A top trumps visualization for document ieee transactions on visualization and computer graphics (tvcg - infovis). 15:1145-1152.
YU, C., YUROVSKY, D., e XU, T. (2012). Visual data mining: An exploratory approach to analyzing temporal patterns of eye movements. Infancy, 17(1):33-60.
CHITTARO, L., COMBI, C., e 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.
KECHADI, M.-T. e BERTOLOTTO, M. (2006). A visual approach for spatio-temporal data mining. In Information Reuse and Integration, 2006 IEEE international Conference on, Páginas 504-509, Waikoloa Village, HI, EUA.
MORAIS, E. A. M. e AMBRÓSIO, A. P. L. (2007). Mineração de textos. Technical Report, Instituto de Informática, Universidade Federal de Goiás, ($INF_0$05/07.).