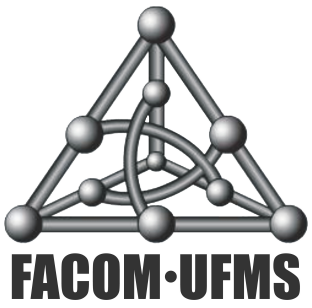
Titulo não definido ainda, trabalho em desenvolvimento

Qualificação (proposta da Dissertação de Mestrado)

Eder Petrica

Orientação: Prof. Dr. Paulo Aristarco Pagliosa

Área de concentração: Visualização



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul

Campo Grande, Dezembro de 2015

Sumário

Capítulo 1

Introdução

Visualização é uma área da comunicação de informações que utiliza representações gráficas para extrair informações de conjuntos de dados[1], ela facilita a interpretação das informações, pois, fornece ao usuário modelos gráficos dos dados que podem ser interpretados mais rápido do uma analise dos dados brutos. Ela pode contribuir com diversas áreas de estudo como, finanças, esporte, politica, acadêmica, entre outras.

O foco deste projeto é a área acadêmica, pois, explorar coleções de documentos com o objetivo de identificar e extrair informações de interesse, como, uma revisão bibliográfica sistemática para uma pesquisa, exige muito tempo devido a necessidade de fazer uma seleção entre inúmeros artigos para descobrir quais são os mais relevantes na área. Em função ao aumento da quantidade da produção textual nos últimos anos, como por exemplo no banco de dados de artigos da IEEE, fazendo-se uma busca rápida pelo termo *"visualisation"*, até 2010 haviam aproximadamente 89 mil artigos científicos publicados, já no inicio de 2016 são mais de 159 mil [2], com base nesta informação há um aumento percentual de aproximadamente 79%, com isso ter uma ferramenta que possa trabalhar com uma quantidade grande de artigos se torna necessária.

A efetividade da ferramenta utilizada depende muito da metáfora de visualização empregada para sintetizar e transmitir informações que se deseja conseguir a partir da visualização é muito importante para a interpretação pelo usuário [3]. Dentre diversas técnicas algumas são melhores para determinadas situações do que

outras, como exemplo, nuvens de palavras é uma técnica eficaz em aplicações que se tem por finalidade fornecer uma visualização do resumo do conteúdo de documentos enquanto métodos que se baseiam em estruturas hierárquicas podem permitir uma exploração mais detalhada da relação entre documentos de acordo com a semelhança entre eles.

Varias metodologias propõem uma combinação de metáforas com o intuito de proporcionar um conjunto mais completo de informações em um único *layout*. Embora algumas metáforas favoreçam a apresentação simultânea de informações com naturezas distintas, algumas abordagens existentes tem de fato sido bem sucedidas na criação de *layouts* compostos que proporcionam efeitos visuais significativos, evitando representações que distraiam ou que sejam muito poluentes visualmente[3]. Em particular, a combinação de layouts de pontos dinâmicos de dados textuais com base em conteúdo de sumarizações, tais como, nuvens de palavras é um problema que foi abordado por [3], mas, que precisa ser melhorado e estendido.

Este trabalho tem como finalidade fornecer uma ferramenta eficaz para realizar

1.1 Objetivos

O objetivo geral deste projeto é desenvolver uma estrutura teórica computacional para exploração visual de uma base de dados composta por artigos científicos que permita o usuário realizar consultas e explorar o conteúdo resultante da visualização.

1.1.1 Objetivos Especificos

1. Desenvolver uma ferramenta de visualização interativa baseada em projeções multidimensionais.
2. Desenvolver uma aplicação capaz de trabalhar com uma base de dados extensível.
3. Desenvolver uma ferramenta que permita pesquisa, interação pelo usuário e navegação na área de visualização.

4. Estender

5.

1.2 Visão Geral para uso interno

O propósito deste trabalho é desenvolver um arcabouço teórico e computacional para exploração visual da base de dados $\mathcal{D}$. A partir de consultas feitas por um usuário (formulada em termos dos atributos dos dados em $\mathcal{D}$), a ferramenta resultante deve, primeiramente, selecionar as n_q instâncias oriundas do processamento da consulta para início da exploração visual. Em seguida, as $n_a \leq n_q$ instâncias selecionadas mais relevantes deverão ser mapeadas para um espaço visual bidimensional, através do emprego de uma técnica de projeção multidimensional. Como decorrência da projeção, cada instância selecionada x_i será mapeada para um ponto $p_i \in \mathbb{R}^2$. O próximo passo é atribuir a cada ponto p_i alguma forma geométrica de área a_i de acordo com ranqueamento de x_i, isto é, se $r_{x_i} > r_{x_j}$, então $a_i > a_j$. O número n_a é definido tal que

$$\sum_{i=1}^{n_a} a_i < s_a\, A,$$

em que $0 < s_a \leq 1$ é um fator especificado pelo usuário e A é a área da janela do espaço visual no qual a visualização será exibida. Em virtude de formas geométricas serem atribuídas aos pontos projetados, essas podem se interceptar no espaço visual. Por isso, deverá ser empregado algum método de repulsão que, ao mesmo tempo em que evita a sobreposição das formas geométricas, mantém, tanto quanto possível, a relação de vizinhança entre os pontos da projeção. As $n_q - n_a$ instâncias selecionadas restantes (isto é, que não têm área associada) deverão ter seus pontos projetados exibidos utilizando-se alguma metáfora representativa de alguma medida de *densidade* no espaço visual. Assim, áreas da janela de visualização com maior "densidade" de pontos representarão regiões com um número maior de instâncias a serem exploradas pelo usuário.

Usuário poderá interagir com os resultados da visualização, além de poder realizar consultas através de querys. Após o resultado visual da busca o usuário também poderá fazer seleção de áreas de interesse para explorar melhor a coleção de dados na tela.

As formas geométricas decorrentes da visualização deverão conter informações sobre os dados que elas representam com tags mostrando nome, autor e outras informações relevantes.

Sumario de um agrupamento

Este trabalho será baseado em uma aplicação ja existente, desenvolvida por [3] o MIST, este por sua vez tem funções limitadas que serão exploradas e poderão ser desenvolvidas com objetivos ampliados.

1.3 MIST

O MIST *(Multiscale Information and Summaries of Texts)* é uma ferramenta que permite a visualização simultânea de documentos individuais, bem como um resumo do conteúdo de coleções de documentos, permite uma exploração multi escalar de subconjuntos de documentos por conteúdo[3].

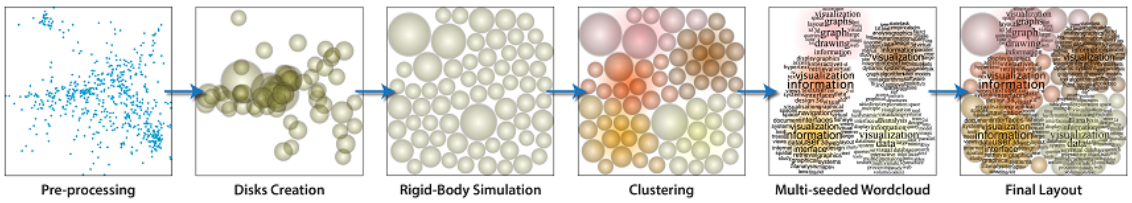


Figura 1.1: Pipeline de visualizações do MIST [3]

A técnica MIST compreende três etapas principais: pré-processamento, criação de discos, simulação de corpo rígido e geração de nuvem de palavras, como ilustrado na Fig. 1.1. Três tarefas são realizadas durante o pré-processamento. A primeira tarefa é um processo de extração de palavras-chave, para gerar a nuvem de palavras no terceiro passo do pipeline. Palavras-chave são também usados para calcular a similaridade entre documentos. Esta semelhança é usada no segundo passo, como entrada para um processo de projeção multidimensional que mapeia os documentos em um espaço visual 2D. A importância de cada documento na coleção também é computada como uma pré-tarefa de processamento e é baseada na ligação entre os documentos individuais, dadas por um usuário ou aplicação definida. Na segunda etapa do MIST, uma simulação de motor de corpo rígido organiza um conjunto de discos que representam os documentos, com o seu tamanho determinado pela im-

portância do documento, evita também sobreposições e ainda preserva estruturas de vizinhança fornecidas pela projeção multidimensional inicial [3]. Na terceira e última etapa do pipeline, os documentos são agrupados de acordo com sua vizinhança, após, nuvens de palavras são gerados e harmoniosamente fundidas para produzir o layout final.

Mas, a aplicação tem limitações como por exemplo a quantidade de artigos que podem ser processados (*deadline:* 2 mil artigos), não sendo viável tratar uma quantia maior de artigos, entre outras que serão exploradas ao decorrer do trabalho. Para que a ferramenta seja usada para processar uma quantia maior de artigos terá de ser feita uma reestruturação de seus algoritmos e métodos.

Contexto e relevância (o que já existe atualmente: MIST).

Lacunas do MIST a serem investigadas no projeto:

- Tamanho da base de dados $\mathcal{D}$ pode se tornar "grande". Como consequência, à medida que novas instâncias são adicionadas a $\mathcal{D}$, tornam-se necessário o uso de um método de ranqueamento e de projeção multidimensional *incrementais*. Atualmente, MIST emprega projeções que podem manipular somente conjuntos "pequenos" de dados.
- MIST não admite consultas iniciais.
- As formas geométricas em MIST são restritas a círculos.
- O método de repulsão é baseado em impulsos aplicados às formas geométricas, consideradas como sendo corpos rígidos bidimensionais. O método não consegue fazer uma escala eficiente para um número "grande" de pontos.
- MIST não emprega uma metáfora para densidade de pontos.
- MIST não efetua agrupamentos no espaço $\mathbb{R}^m$, mas sim no espaço visual $\mathbb{R}^2$. O número de grupos é especificado arbitrariamente pelo usuário e computado através de *k-means*. Como consequência, as instâncias em um agrupamento não mantêm necessariamente outra relação entre si que não seja a de proximidade de suas projeções no espaço visual.
- MIST sumariza um agrupamento através de nuvens de palavras, as quais são exibidas como textura de fundo do retângulo envolvente do agrupamento no

espaço visual. Tal procedimento restringe o emprego de sumários. Embora a visualização contenha vários elementos informativos, pode-se contestar a real eficácia de tal esquema de sumarização, além da confusão visual por parte de usuários.

Capítulo 2

Trabalhos Relacionados e Fundamentos

2.1 Introdução

Neste capítulo são apresentados os alguns trabalhos mais relevantes, métodos que são relacionados e conceitos teóricos necessários para o desenvolvimento deste trabalho ...

2.2 Trabalhos Relacionados

Nesta seção são expostos alguns trabalhos que utilizar métodos de visualização e técnicas pertinentes a temática deste trabalho como nuvens de palavras, river methaphor, estruturas linguísticas, estruturas hierárquicas.

Koh et. al. [4] apresenta uma ferramenta de visualização de palavras baseada em Wordle (aplicação usada para gerar nuvens de palavras) o ManiWordle. Ele renova as interações com o layout, apoiando manipulações personalizadas, permite a manipulação de tipografia, cor e composição, não só para o layout como um todo, mas, também para as palavras individuais, permitindo ter um melhor controle sobre o resultado do layout.

Wu et. al. [5] apresenta um método para criar nuvens de palavras preservando a semântica, apresentando uma metodologia que calcula primeiro as relações semânticas e em seguida usa escalonamento multidimensional para colocar as palavras-chave no espaço visual.

Métodos, como SparkClouds apresentado em [7] e Tag Clouds Paralel [8], aumentam as nuvens de palavras com recursos visuais adicionais, tais como linhas de ignição e coordenadas paralelas, a fim de melhor transmitir o conteúdo do resumo de documentos. Embora muito eficaz para descobrir informações essenciais contidas em uma coleção de documentos, o paradigma da nuvem de palavras por si só não identifica a associação entre palavras para um determinado documento ou grupo de documentos e não permite o exame de similaridade entre eles.

Em [6] é proposta uma técnica de visualização multidimensional com projeção que se baseia em exemplos representativos para definir agrupamentos no espaço visual. Exemplos representativos são selecionados por um sistema de amostragem determinista derivada da decomposição da matriz que é sensível à variabilidade de dados capaz de lidar com as classes com um pequeno número de casos. Além disso, o mecanismo de amostragem pode facilmente ser adaptado para selecionar atributos relevantes de cada agrupamento.

Outro mecanismo muito eficiente para visualizar variações temáticas de coleções de documentos em uma linha do tempo são as metáforas de rio (River Metaphor). Com essas metáforas pode-se visualizar variações nos estilos comparando com outros documentos através de uma linha do tempo oriundas de eventos externos. Introduzido inicialmente pelo sistema ThemeRiver [9], as metáforas foram melhoradas com mecanismos sofisticados para fazer a derivação em relação aos tópicos de detecção de tempo [10] e visualizações em camadas capazes de descrever o nascimento, morte e divisão de classes de eventos [11].

Existe também o EventRiver [12] que faz uso de um esquema de *cluster* para grupos de notícias similares que tenham conteúdos próximos ao longo do tempo, como colunas de jornal, ele usa metáfora de bolha cuja espessura representa o número de documentos e o comprimento a duração de um evento. Em [13] é apresentada uma ferramenta para fazer uma análise exploratória de dados, o fluxo histórico de visualização, também pode ser visto como metáfora de rio concebido para visualizar edições de um documento (ou uma coleção de documentos, tais como a Wikipédia)

feito por diferentes autores, enfatizando partes que sobrevivem ao longo do tempo. Metáforas de rio proporcionar uma visualização agradável e intuitiva quanto ao comportamento temporal de uma coleção de documentos, mas, semelhante a nuvens de palavras, a técnica não permite a identificação imediata de documentos específicos, a sua relevância dentro da coleção ou o seu contribuição para um tópico. Além disso, interagir com o esquema de rio para realizar alterações na perspectiva do utilizador conjunto de dados é claramente não viável.

Outro método utilizado para construção visual é o baseado em estruturas linguísticas semânticas como o Word Tree [14], por exemplo, que utiliza um esquema de árvore para visualizar a ocorrência de termos juntamente com a frases que o compõem, eles são dispostos em ramos descendentes da árvore. Existe também o Phrase Nets [15] que emprega um layout baseado em grafos onde os nós (*nodes*) correspondem a um subconjunto de palavras e arestas correspondem à relação semântica ou léxica entre as palavras. O tamanho da fonte e espessura de borda são usados para mapear visualmente os atributos como o número de ocorrências de um conjunto de palavras e seus relacionamentos. Uma análise linguística mais sofisticada é aplicada pelo DocuBurst [16], ele faz uso de uma base de dados léxica eletrônica e um layout de árvore com preenchimento radial no espaço para visualizar o conteúdo do documento de uma forma léxica. Keim e Oelke [17] desenvolveram um método que emprega regras semânticas para segmentar um documento em blocos e funções de palavras para mapear os blocos em vetores de características. O principal componente de cada recurso do vetor é usado para colorir os blocos, resultando em uma imagem como uma impressão digital do documento. Em contraste com os outros métodos baseados em linguística acima descritos, o método de Keim e Oelke permite identificar e comparar os documentos específicos no conjunto de dados, mas, comprometem a legibilidade do seu conteúdo.

Existem algumas técnicas que se baseiam em estruturas hierárquicas, elas permitem um nível de detalhamento, exploração e navegação diferente das demais apresentadas encontradas na literatura. Topic Island [18], por exemplo, cria uma hierarquia através da aplicação de uma transformada *wavelet* em sinais customizados extraídos de palavras do documento. A hierarquia permite a visualização com alterações de tema e partes importantes da coleção de documentos relacionados com o conteúdo total do documento. InfoSky [19] visualiza documentos hierarquicamente organizados, subdividindo o espaço visual usando um diagrama de Voronoi recursivo. A navegação em toda a hierarquia é ativada por um mecanismo tipo um zoom

telescópio. Hipp [20] faz uso de um conjunto de árvores para organizar hierarquicamente documentos de acordo com a sua semelhança, realizando a visualização da hierarquia, o resultado da visualização é uma árvore. Mao et al. [21] apresentam um técnica para visualizar documentos usando curvas construídas a partir de uma generalização de n-grams e médias locais, a construção da hierarquia, alterando o apoio dos grãos usados no cálculo da média. Embora eficaz para construir sumarizações visuais, bem como para identificar as estruturas nos tópicos do documento, técnicas hierárquicas não são eficazes para associar conteúdo e documentos quando a hierarquia é feita sobre os temas. Além disso, a visualização da estrutura hierárquica e a importância de cada documento, simultaneamente, não é uma tarefa simples.

Em [22] é proposto um um método baseado em topologia que evita a oclusão estrutural para preservar recursos de agrupamentos primário e propriedades geométricas negligênciadas que não podem ser preservadas em representações de baixa dimensionalidade. Ele abstrai os pontos de entrada nas regiões com as propriedades de cada um e fornece ao usuário visualizações intuitivas tipo paisagem que ilustram a estrutura de alta dimensão do agrupamento livre de oclusão.

Capítulo 3

Metodologia

Neste capítulo apresentamos uma proposta para o trabalho de pesquisa a ser desenvolvido, mas, ainda esta incompleto, pois, ainda estão sendo investigadas metodologias a serem trabalhadas no trabalho.

3.1 Aplicativo Web

3.2 Cronograma previsto

O projeto terá as seguintes etapas,

Etapa	Jan 2016	Fev 2016	Mar 2016	Abr 2016	Mai 2016	Jun 2016	Ago 2016	Set 2016	Out 2016	Nov 2016
implementação das ferramentas										
estudo aprofundado										
instanciações em projetos específicos										
escrita e defesa da dissertação										

Figura 3.1: Cronograma previsto para o projeto.

Referências Bibliográficas

[1] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.

[2] IEEE, "Ieee," IEEE, 2016.

[3] P. Pagliosa, R. M. Martins, D. Cedrim, A. Paiva, R. Minghim, and L. G. Nonato, "Mist: Multiscale information and summaries of texts," in *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*, pp. 91–98, IEEE, 2013.

[4] K. Koh, B. Lee, B. Kim, and J. Seo, "Maniwordle: Providing flexible control over wordle," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 1190–1197, 2010.

[5] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma, "Semantic-preserving word clouds by seam carving," in *Computer Graphics Forum*, vol. 30, pp. 741–750, Wiley Online Library, 2011.

[6] P. Joia, F. Petronetto, and L. G. Nonato, "Uncovering representative groups in multidimensional projections," in *Computer Graphics Forum*, vol. 34, pp. 281–290, Wiley Online Library, 2015.

[7] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "Sparkclouds: Visualizing trends in tag clouds," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 1182–1189, 2010.

[8] C. Collins, F. B. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 91–98, IEEE, 2009.

[9] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 1, pp. 9–20, 2002.

[10] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 25, 2012.

[11] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2412–2421, 2011.

[12] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim, "Eventriver: Visually exploring text collections with temporal references," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 1, pp. 93–105, 2012.

[13] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 575–582, ACM, 2004.

[14] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1221–1228, 2008.

[15] F. Van Ham, M. Wattenberg, and F. B. Viégas, "Mapping text with phrase nets," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1169–1176, 2009.

[16] C. Collins, S. Carpendale, and G. Penn, "Docuburst: Visualizing document content using language structure," in *Computer graphics forum*, vol. 28, pp. 1039–1046, Wiley Online Library, 2009.

[17] D. Keim, D. Oelke, *et al.*, "Literature fingerprinting: A new method for visual literary analysis," in *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pp. 115–122, IEEE, 2007.

[18] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote, "Topic islands tm-a wavelet-based text visualization system," in *Visualization'98. Proceedings*, pp. 189–196, IEEE, 1998.

[19] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, "The infosky visual explorer: exploiting hierarchical structure and document similarities," *Information Visualization*, vol. 1, no. 3-4, pp. 166–181, 2002.

[20] F. V. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1229–1236, 2008.

[21] Y. Mao, J. V. Dillon, and G. Lebanon, "Sequential document visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1208–1215, 2007.

[22] P. Oesterling, C. Heine, G. H. Weber, and G. Scheuermann, "A topology-based approach to visualize the thematic composition of document collections," in *Text Mining*, pp. 63–85, Springer, 2014.