Mapeamento de Dados Multi-dimensionais – Integrando Mineração e Visualização*

Fernando V. Paulovich e Rosane Minghim

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400, São Carlos/SP, Brasil

{paulovic,rminghim}@icmc.usp.br

Abstract. *The projection or point placement techniques, which seek to map multi-dimensional data onto visual spaces, have been the interest of the visual data analysis community for a long time due to their ability for exploratory tasks based on similarity and correlation. However, many problems still persist, impairing their application. The main contribution of this paper is a further understanding of the problems found on the current techniques and the development of projection techniques which are fast, appropriately define groups of highly similar data instances, separate these groups on the final layout, and allow the data exploration on different levels of detail. In addition, we integrate some data mining features to the process of multidimensional visualization, mainly on the application of projections to the visualization of document collections.*

Resumo. *As técnicas de projeção ou posicionamento de pontos no plano, que servem para mapear dados multi-dimensionais em espaços visuais, sempre despertaram grande interesse da comunidade de visualização e análise de dados por representarem uma forma útil de exploração baseada em relações de similaridade e correlação. Apesar disso, muitos problemas ainda são encontrados em tais técnicas, limitando suas aplicações. Como principal contribuição deste trabalho propomos um entendimento mais profundo dos problemas encontrados nas técnicas de projeção vigentes e o desenvolvimento de técnicas de projeção (ou mapeamento) que são rápidas, tratam adequadamente a formação visual de grupos de dados altamente similares, separam satisfatoriamente esses grupos no layout, e permitem a exploração dos dados em vários níveis de detalhe. Além disso, incorporamos alguns aspectos de mineração integrados ao processo de visualização multi-dimensional, principalmente na aplicação de projeções para visualização de coleções de documentos.*

1. Introdução

Com a diminuição do custo e a melhoria das tecnologias para armazenamento, distribuição e recuperação de dados, a quantidade de informação produzida, ou disseminada, tem crescido substancialmente tanto em volume quanto em complexidade. Apesar de muita informação não relevante poder ser filtrada, a geração de informação útil ainda continua sendo muito maior que a capacidade das ferramentas de interpretação vigentes.

*A tese de doutorado a que esse trabalho se refere pode ser encontrada em http://www.teses.usp.br/

Daí a necessidade de métodos e ferramentas capazes de sintetizar esse aglomerado de dados para que seja possível interpretá-lo, apresentando-o de forma simples e amigável.

Dentre as possíveis áreas que estudam métodos para apoiar a interpretação de tal informação, a fusão de duas, a *Mineração de Dados* e a *Visualização de Informação*, tem despertado grande interesse da comunidade científica. Da união de ambas áreas surge uma nova área onde a mineração e a visualização co-existem na busca de soluções para interpretação de conjuntos de dados complexos. Tal área é denominada *Mineração Visual de Dados* [Wong 1999].

Umas das abordagens de visualização de informação que vem sendo empregada com sucesso no processo de mineração visual de dados são as *Técnicas de Posicionamento de Pontos*, também conhecidas como *Técnicas de Projeção Multi-dimensional* [Tejada et al. 2003]. Essas técnicas buscam criar representações visuais possibilitando que usuários empreguem suas habilidades visuais para reconhecer estruturas e padrões presentes nos dados. Nesta representação, cada instância de dados é mapeada em um elemento visual, tal como um círculo, ponto ou esfera em um espaço de visualização (1D, 2D ou 3D). As posições relativas desses elementos refletem algum tipo de relacionamento entre as instâncias de dados, as mais comuns sendo relações de similaridade ou vizinhança [Paulovich and Minghim 2008]. Nesse caso, se pontos forem posicionados próximos no *layout* produzido, isso indica que os objetos que esses representam são similares de acordo com a distância (dissimilaridade) escolhida, e se pontos forem projetados distantes, isso indica que os objetos que os mesmos representam são pouco relacionados.

Apesar do sucesso das técnicas de projeção para o suporte a interpretação de dados multi-dimensionais, alguns problemas persistem que ainda precisam ser tratados de forma a torná-las ferramentas efetivas para o processo de mineração visual de dados. Com base em um estudo comparativo entre várias técnicas delineamos alguns objetivos de forma a tratar esses problemas [Paulovich 2008]: (1) o desenvolvimento de uma técnica pautado na redução de complexidade deve também levar em consideração a qualidade do *layout* gerado; (2) uma técnica de projeção deve buscar preservar as relações de vizinhança entre as instâncias individuais de dados (informação local) tanto quanto as relações entre grupos de instâncias similares (informação global); (3) em conjuntos de dados que apresentem relações não-lineares entre seus atributos é preferível trabalhar com vizinhanças locais já que nesse caso essas relações passam a ser lineares [Merino and Muñoz 2004].

Com base nesse estudo, desenvolvemos novas técnicas visando aperfeiçoar e tornar mais efetiva a abordagem de posicionamento de pontos no plano, com especial atenção à mineração visual de coleções de documentos. Essas técnicas são capazes de lidar com grandes conjuntos de dados, mantendo a qualidade das representações gráficas geradas. Além disso, elas são capazes de prover exploração de conjuntos de dados em diferentes níveis de detalhamento, reduzindo assim os problemas relacionados a carga cognitiva imposta aos usuários no processo de interpretação das representações visuais.

As seguintes seções apresentam essas técnicas, seguindo de outras contribuições desenvolvidas e finalizando com as conclusões, destacando as principais contribuições desse trabalho.

2. Convergindo para Mineração Visual de Dados

No trabalho aqui apresentado, o domínio principal de aplicação considerado foi o da construção de representações visuais de coleções de documentos. Nesse caso, para se calcular a dissimilaridade entre documentos nós usamos a abordagem mais comumente empregada que é a representação vetorial de coleções de documentos [Salton 1991]. Nessa representação, primeiro é escolhido um conjunto de m termos representativos. Em seguida, cada documento é transformado em um vetor m-dimensional onde cada dimensão representa um termo, e as coordenadas desses vetores são dadas pelas freqüências de ocorrência (ponderadas) dos termos dentro dos documentos, resultando em uma matriz de "documentos x termos".

Normalmente, o espaço resultante desse processo será esparso, o que torna os objetos (documentos) bastante dissimilares entre si – independente da métrica de distância empregada –, estando um objeto somente relacionados à um número limitado de vizinhos mais próximos dentro de um mesmo sub-espaço local [Merino and Muñoz 2004].

Levando-se essa característica em consideração, desenvolvemos uma técnica de projeção especificamente para o mapeamento de coleções de documentos, denominada *Least Square Projection (LSP)* [Paulovich et al. 2008a], descrita a seguir.

2.1. Least Square Projection (LSP)

A *LSP* adota uma estratégia diferente das técnicas de projeção convencionais. Ela busca preservar relações de vizinhança entre os objetos m-dimensionais no espaço projetado, ao invés de tentar preservar relações de similaridade. Assim, quando um conjunto de objetos multi-dimensionais é projetado o que se busca é garantir que os objetos vizinhos no espaço multi-dimensional sejam projetados dentro de uma mesma vizinhança no espaço visual.

Dois passos principais são executados nesse processo de projeção. Primeiro, um subconjunto de objetos multi-dimensionais, chamados de "pontos de controle", é cuidadosamente escolhido e projetado no espaço visual usando-se uma técnica que preserva as relações de distância com precisão. Depois, fazendo-se uso das relações de vizinhança dos objetos no espaço multi-dimensional, e das respectivas coordenadas cartesianas dos pontos de controle projetados, é construído um sistema linear cuja solução visa posicionar os objetos restantes no fecho convexo de seus k vizinhos mais próximos.

Para se construir esse sistema, considere $V_i = \{p_1, \ldots, p_{k_i}\}$ como o conjunto dos k_i vizinhos mais próximos de p_i e $\tilde{p}_i$ as coordenadas cartesianas de p_i no espaço projetado. Se $\tilde{p}_i$ forem dadas pela equação $\tilde{p}_i - \sum_{p_j \in V_i} \frac{1}{k_i} \tilde{p}_j = 0$, p_i residirá no fecho convexo de seus k_i vizinhos mais próximos, mais precisamente no centróide desses vizinhos [Sorkine and Cohen-Or 2004].

Se essa equação for resolvida para todos os p_i do conjunto de dados, o resultado será um sistema linear $A\mathbf{x} = b$, onde A é normalmente chamada de matriz *Laplaciana*. Considerando conexo o grafo de vizinhança dos objetos multi-dimensionais – o grafo que conecta cada objeto com seus vizinhos mais próximos – esse sistema passa a admitir solução não-trivial. Contudo, para tornar essa solução mais atrativa, informação geométrica é embutida no sistema adicionando novas linhas referentes aos pontos de controle. Esse novo sistema apresenta *rank completo* e pode ser resolvido aplicando-se mínimos quadrados, ou seja, encontrando-se $\mathbf{x}$ que minimize $||Ax - b||^2$. O sistema

$A^T A\mathbf{x} = A^T\mathbf{b}$ que deve ser resolvido é simétrico e esparso, permitindo que métodos eficientes de resolução possam ser empregados [Sorkine and Cohen-Or 2004].

A idéia da preservação de vizinhanças para lidar com dados não-lineares já foi empregada em outras técnicas, como por exemplo na técnica de redução de dimensionalidade *Local Linear Embedding (LLE)* [Roweis and Saul 2000]. Porém, diferentemente da *LLE*, a *LSP* busca preservar também informação global por meio dos pontos de controle, isto é, informação sobre os possíveis grupos de objetos dentro do conjunto de dados. Assim, comparativamente, os resultados apresentados pela *LSP* são normalmente superiores em termos da separação e agrupamento dos objetos multi-dimensionais, produzindo representações visuais mais coerentes [Paulovich 2008].

2.2. Extração de Tópicos por Covariância

Embora uma projeção multi-dimensional seja útil para revelar relações de similaridade entre objetos e grupos de objetos, nenhuma informação é dada sobre o motivo de certos grupos terem se formado no *layout* final. No caso específico da exploração de coleções de documentos, um mecanismo que pode auxiliar no processo de análise de uma projeção é a extração automática e semi-automática de tópicos. Aqui, nós definimos tópicos como conjuntos de termos relacionados que buscam identificar o assunto comum a um determinado grupo de documentos [Lopes et al. 2007].

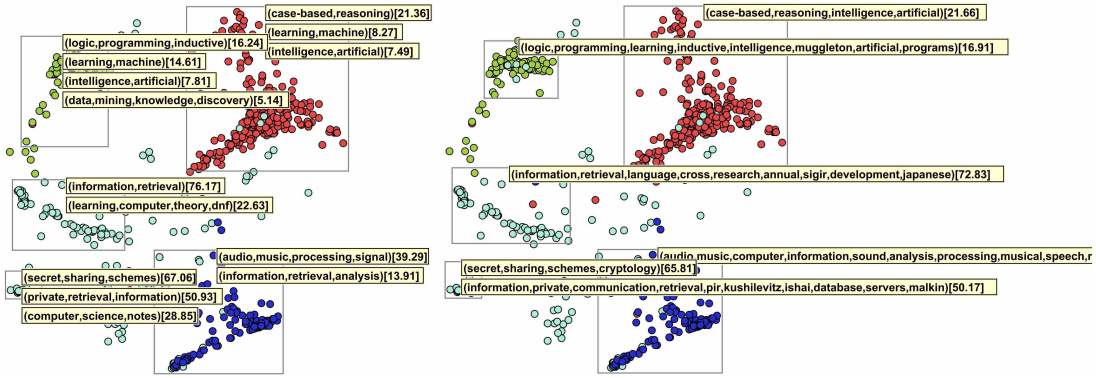
O processo de extração de tópicos aqui definido, denominado de *Tópicos por Covariância* [Cuadros et al. 2007], se inicia com o usuário selecionando uma determinada área de uma projeção. Após isso, uma matriz de "documentos x termos" é criada considerando-se somente os documentos escolhidos. Com base nessa matriz, os dois termos que apresentarem a maior covariância são inicialmente escolhidos e adicionados ao tópico. Uma vez que os dois termos com maior covariância tenham sido encontrados, a covariância média entre esses e os demais é calculada. Se a razão do valor da covariância média pela maior covariância for igual ou ultrapassar um limiar pré-estabelecido (α), o termo é adicionado ao tópico. Além disso, permitimos que múltiplos tópicos sejam criados para o grupo de documentos selecionados. Para tal, qualquer par de palavras cuja razão de sua covariância pela maior covariância ultrapassar ou for igual a um limiar pré-estabelecido (β), acaba gerando um novo tópico.

A Figura 1 apresenta projeções *LSP* com alguns tópicos extraídos variando os valores de α e β para o conjunto de dados **CBR-ILP-IR-SON**, uma coleção de 675 documentos composta por título, autores, afiliação, resumo e referências de artigos científicos em quatro diferentes áreas: *Case-Based Reasoning (CBR)*, *Inductive Logic Programming (ILP)*, *Information Retrieval (IR)* e *Sonification (SON)* [1]. Para a criação desses tópicos, as cinco áreas que apresentam conjuntos de documentos claramente identificáveis foram selecionadas. É possível notar que apesar de ser um processo simples de detecção e extração de tópicos, as quatro grandes áreas dentro desse conjunto de documentos foram satisfatoriamente identificadas.

2.3. Hierarchical Point Placement (HiPP)

Pela literatura recente sobre projeções multi-dimensionais é possível notar uma crescente preocupação com a capacidade dessas de lidar com grandes conjuntos de dados, ou seja,

[1] Esse conjunto de dados e as ferramentas e código produzido referentes a esse projeto de doutorado se encontram disponíveis em `http://infoserver.lcad.icmc.usp.br/`



(a) $\alpha = 0.5$ e $\beta = 0.25$. (b) $\alpha = 0.25$ e $\beta = 0.75$.

Figura 1. Tópicos extraídos, variando-se os parâmetros α e β, de uma projeção *LSP* do conjunto CBR-ILP-IR-SON.

com a escalabilidade. Nesse sentido, muito do quem sido desenvolvido se concentra na redução da complexidade computacional, possibilitando o processamento de volumes de dados cada vez maiores. Apesar dessa ser uma preocupação legítima, outro fator de escalabilidade também deve ser considerado: a *escalabilidade visual*. Por escalabilidade visual entende-se a capacidade da representação visual, e das ferramentas de visualização, de efetivamente apresentar grandes conjuntos de dados [Eick and Karr 2002]. Nesse aspecto, a metáfora visual empregada nas projeções – um elemento gráfico representando um objeto multi-dimensional – tende a apresentar problemas para grandes conjuntos de dados devido a forte sobreposição que ocorre entre os elementos e a desordem visual (*cluttering*) [Paulovich 2008].

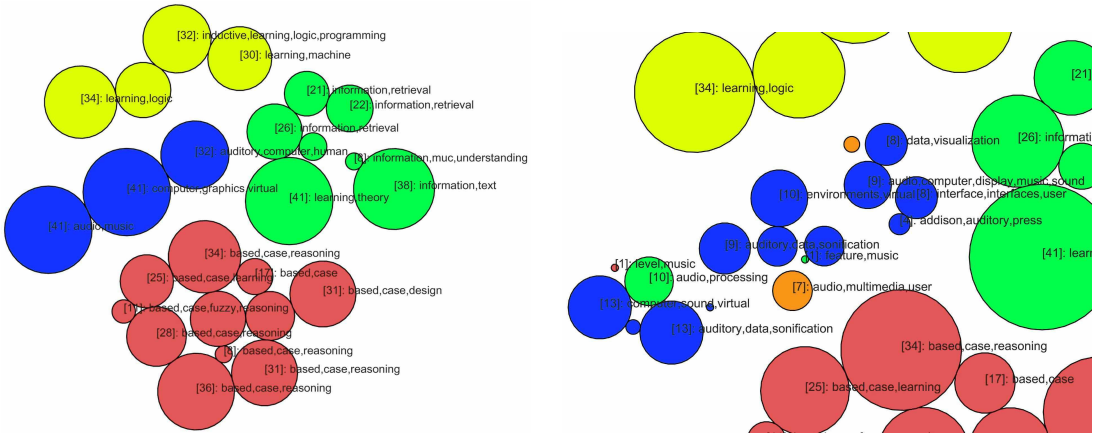
De forma a apontar uma possível solução para esse problema, e partindo-se da observação de que uma projeção multi-dimensional normalmente é empregada para se extrair grupos de objetos multi-dimensionais similares, relações entre esses, e entre objetos individuais, desenvolvemos uma técnica chamada *Hierarchical Point Placement (HiPP)* [Paulovich and Minghim 2008]. Na *HiPP*, os elementos visuais podem representar tanto instâncias individuais de objetos multi-dimensionais quanto grupos de instâncias.

Em síntese, podemos dividir essa técnica em dois grandes passos: (1) uma hierarquia de agrupamentos é criada usando um processo recursivo de particionamento onde os nós internos são agrupamentos e as folhas são instâncias individuais de dados (outros métodos também podem ser usados para esse fim, como os de agrupamento hierárquico, ou mesmo uma hierarquia pré-existente [Paulovich 2008]); e (2) os elementos dessa árvore são mapeados para o espaço bi-dimensional para se criar a representação visual.

O resultado da *HiPP* é uma representação gráfica interativa onde um usuário parte de uma visão geral dos dados (agrupamentos de mais alto-nível), e concentrando-se nos agrupamentos de interesse (clicando e expandindo em sub-agrupamentos), alcança as instâncias individuais de dados. Assim, menos informação é imposta ao usuário para a interpretação de uma primeira visão de um conjunto de dados, e o usuário pode mais facilmente se concentrar na informação de interesse.

A Figura 2 apresenta o mapa de documentos para o conjunto de dados **CBR-**

ILP-IR-SON. Na Figura 2(a) é apresentada a visão de mais alto nível do mapa, onde os círculos representam os agrupamentos de documentos. A lista de palavras colocadas sobre cada nó são os tópicos extraídos deles. É possível notar que para cada uma das quatro áreas de artigos científicos existem nós representando-as, e que os grupos de artigos com tópicos similares são posicionados próximos entre si na representação visual. A Figura 2(b) apresenta o resultado se alguns agrupamentos em azul (em tom mais escuro) forem expandidos em sub-agrupamentos (referente a área de sonificação).



(a) Agrupamentos de mais alto nível. (b) Visão de alguns sub-agrupamentos.

Figura 2. Mapa de documentos do conjunto CBR-ILP-IR-SON usando a *HiPP*.

2.3.1. Re-Arranjando a Hierarquia

Para a composição da hierarquia de agrupamentos, agrupamentos de instâncias de dados são extraídos recursivamente usando-se alguma heurística. Assim, dependendo-se de qual heurística é empregada e o que é entendido como um agrupamento, diferentes resultados podem ser alcançados [Tan et al. 2005]. Portanto, é possível que a hierarquia resultante contenha agrupamentos diferentes dos esperados pelo usuário, possivelmente apresentando grupos com instâncias que não são adequadamente relacionadas, ou colocando instâncias altamente relacionadas em diferentes grupos.

De forma a superar esse problema, nós definimos uma estratégia de unir agrupamentos baseada na inspeção visual do *layout* gerado pela *HiPP*. Nessa estratégia, após o usuário ter selecionado um conjunto de agrupamentos para ser unido, um novo nó é criado contendo as instâncias de dados desses nós e ligado ao ancestral comum desses. Em seguida, o novo nó é projetado no plano, seu tópico é extraído, e todos os nós que tiveram instâncias removidas são reduzidos e seus tópicos renovados. Desta forma, um usuário pode reconstruir a hierarquia de agrupamentos usando sua própria interpretação, reorganizando o conjunto de dados baseado em seu próprio conhecimento e necessidade.

Com isso, a *HiPP* se torna uma técnica que efetivamente une conceitos de mineração de dados com visualização de informação, podendo ser usada em um processo que tira proveito das características especificas dessas abordagens na tarefa de interpretação de coleções de documentos.

3. Outras Contribuições e Resultados

Além das contribuições apresentadas anteriormente, que configuram os principais resultados desse trabalho, podemos destacar outros resultados produzidos ou provenientes da cooperação com outros pesquisadores, entre eles: (1) a criação de uma técnica de projeção de propósito geral [Paulovich and Minghim 2006]; (2) o desenvolvimento de diferentes ferramentas para a criação e exploração de projeções multidimensionais [Paulovich et al. 2007, Paulovich et al. 2008b]; (3) uma técnica baseada em métodos de compressão como forma de cálculo da similaridade entre documentos [Telles et al. 2007]; e (4) uma técnica baseada em regras de associação para a extração de tópicos de coleções de documentos [Lopes et al. 2007]

4. Conclusões

De forma geral, além das técnicas e ferramentas desenvolvidas, este trabalho contribui para um melhor entendimento das características desejáveis de uma técnica de projeção multi-dimensional, principalmente as específicas para a projeção de coleções de documentos, de forma a possibilitar suas aplicações ao processo de mineração visual de dados. Nesse caso específico, o que se pode concluir é que embora muitas das técnicas vigentes para a projeção de coleções de documentos busquem preservar as relações de similaridade entre todos os documentos, a preservação de vizinhanças locais leva a resultados mais precisos e coerentes. Nesse sentido desenvolvemos a *Least Square Projection (LSP)*.

Outra contribuição foi mostrar que, tão importante quanto buscar a redução da complexidade computacional das técnicas de projeção de forma a ser possível tratar grandes conjuntos de dados, deve ser a preocupação com a escalabilidade visual das representações gráficas produzidas. Isso porque, além das limitações físicas de espaço visual para o desenho dessas representações (normalmente o monitor de um computador) existe também o fato de as habilidades e capacidades humanas não serem tão escaláveis [Thomas and Cook 2005]. Do contrário a metáfora, normalmente empregada nas representações visuais de projeções, que mapeia cada instância de dados em um elemento gráfico, poderia prejudicar o processo de exploração visual dos dados já que muita informação é simultaneamente apresentada ao usuário.

Como resultado dessa observação, desenvolvemos a *Hierarchical Point Placement (HiPP)*. Essa técnica permite que um usuário inspecione uma coleção de documentos em diferentes níveis de detalhamento, iniciando com uma visão mais abstrata dos grupos de documentos que podem ocorrer em uma coleção, e concentrado-se nos grupos de interesse ir detalhando a visão que se tem da coleção, diminuindo-se a carga cognitiva imposta ao usuário para se interpretar a representação visual. Além disso, essa técnica permite que a hierarquia de agrupamentos de documentos seja modificada conforme a representação visual é inspecionada, unindo de fato estratégias de mineração de dados com uma abordagem de visualização de informação.

Agradecimentos

Este trabalho recebeu suporte financeiro da *FAPESP* (proc. no. 04/07866-4) e da *CAPES* (proc. no. 2214-07-5).

Referências

Cuadros, A. M., Paulovich, F. V., Minghim, R., and Telles, G. P. (2007). Point placement by phylogenetic trees and its application for visual analysis of document collections. In *Proc. of IEEE VAST 2007*, pages 99–106.

Eick, S. G. and Karr, A. F. (2002). Visual scalability. *Journal of Computational & Graphical Statistics*, 11(1):22–43.

Lopes, A. A., Pinho, R., Paulovich, F. V., and Minghim, R. (2007). Visual text mining using association rules. *Computer & Graphics*, 31(3):316–326.

Merino, M. M. and Muñoz, A. (2004). A new sammon algorithm for sparse data visualization. In *Proceedings of ICPR'04*, pages 477–481, Washington, USA. IEEE CS.

Paulovich, F. V. (2008). *Mapeamento de Dados Multi-Dimensionais – Integrando Mineração e Visualização*. Tese Doutorado, ICMC/USP.

Paulovich, F. V. and Minghim, R. (2006). Text map explorer: a tool to create and explore document maps. In *Proc. of IV'06*, pages 245–251, Washington, USA. IEEE CS.

Paulovich, F. V. and Minghim, R. (2008). HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Trans. on Vis. and Comp. Graph. (Proc. of InfoVis 2008)*, 14(6):1229–1236.

Paulovich, F. V., Nonato, L. G., Minghim, R., and Levkowitz, H. (2008a). Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Trans. on Vis. and Comp. Graph.*, 14(3):564–575.

Paulovich, F. V., Oliveira, M. C. F., and Minghim, R. (2007). The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proc. of SIBGRAPI 2007*, pages 27–36, Washington, USA. IEEE CS.

Paulovich, F. V., Pinho, R., Botha, C. P., Heijs, A., and Minghim, R. (2008b). Pex-web: Content-based visualization of web search results. In *Proc. of IV'08*, pages 208–214, Los Alamitos, USA. IEEE CS.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Salton, G. (1991). Developments in automatic text retrieval. *Science*, 253:974–979.

Sorkine, O. and Cohen-Or, D. (2004). Least-squares meshes. In *Proc. of SMI'04*, pages 191–199, Washington, USA. IEEE CS.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, USA.

Tejada, E., Minghim, R., and Nonato, L. G. (2003). On improved projection techniques to support visual exploration of multidimensional data sets. *Inf. Vis.*, 2(4):218–231.

Telles, G. P., Minghim, R., and Paulovich, F. V. (2007). Normalized compression distance for visual analysis of document collections. *Computer & Graphics*, 31(3):327–337.

Thomas, J. J. and Cook, K. A., editors (2005). *Illuminating the path: The Research and Development Agenda for Visual Analytics*. IEEE CS, Los Alamitos, USA.

Wong, P. C. (1999). Visual data mining. *IEEE Comp. Graph. and App.*, 19(5):20–21.