UNIVERSIDADE FEDERAL DO PARÁ

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANDRE LUIZ DA SILVA KAUER

Utilizando Múltiplas Análises Visuais e Interativas de Dados para Continuidade de Transações Eletrônicas

Prof. Dr. Bianchi SeriqueMeiguins
Orientador

Belém-PA
2012

André Luiz da Silva Kauer

Utilizando Múltiplas Análises Visuais e Interativas de Dados para Continuidade de Transações Eletrônicas

Dissertação de Mestrado apresentada para obtenção do grau de Mestre em Ciência da Computação.
Programa de Pós Graduação em Ciência da Computação.
Instituto de Ciências Exatas e Naturais.
Universidade Federal do Pará.
Orientador Prof. Dr. Bianchi SeriqueMeiguins

Belém-PA
2012

Kauer,Andre Luiz da Silva
Utilizandomúltiplas análises visuais e interativas de dados paracontinuidade de negócios eletrônicos / (Andre Luiz da SilvaKauer); orientador, Bianchi SeriqueMeiguins. - 2012.
80 f. il. 28 cm
Dissertação (Mestrado) – Universidade Federal do Pará. Instituto de Ciências Exatas e Naturais. Programa de Pós-Graduação em Ciência da Computação. Belém, 2012.

1. Tecnologia da Informação. 2. Visualizaçãode Informações.3.Comercio eletrônico. I. Meiguins, Bianchi Serique, orient. II. Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed.004

André Luiz da Silva Kauer

Utilizando Múltiplas Análises Visuais e Interativas de Dados para Continuidade de Transações Eletrônicas

Dissertação de Mestrado apresentada para obtenção do grau de Mestre em Ciência da Computação.
Programa de Pós Graduação em Ciência da Computação.
Instituto de Ciências Exatas e Naturais.
Universidade Federal do Pará.

Data da aprovação: Belém-PA,21/09/2012.

Banca Examinadora

_____
Prof. Dr. Bianchi SeriqueMeiguins
Instituto de Computação – UFPA – Orientador

_____
Prof. Dr. Eloi Luiz Favero
Instituto de Computação – UFPA – Membro

_____
Prof. Dr. Mario MassakuniKubo
Instituto de Computação – Faculdade Alvorada/DF– Membro

Visto:
Prof. Sandro Ronaldo Bezerra Oliveira, Dr. (UFPA)
Coordenador do PPGCC – UFPA

Belém-PA
2012

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

PUBLICAÇÕES

da Silva Kauer, A.L; Meiguins, B.S. ; do Carmo, R.M.C. ; de Brito Garcia, M. ; Meiguins, A.S.G. An Information Visualization Tool with Multiple Coordinated Views for Network Traffic Analysis .IV '08.12th International ConferenceInformation Visualisation, 2008. Page(s): 151 - 156

da Silva Kauer, A.L.; Meiguins, B.S.; Pires, A.H.I.; Garcia, M.; GoncalvesMeiguins, A.S. Business Transactions Analysis Based on Multiple Data Coordinated Views. InformationVisualisation, 2008. IV '08. 12th InternationalConference. Page(s): 258 - 263

RESUMO

Conhecer melhor os clientes de uma empresa e suas atividades permite oferecer serviços mais adequados e aumentar a fidelização dos seus clientes, possibilitando oportunidades de negócio para o crescimento da empresa. Uma das maneiras de aprender mais sobre o comportamento dos clientes é através da análise dos históricos de transações realizadas, com informações sobre os tipos de operações, tempo médio das operações, operações não autorizadas, operações iniciadas e não finalizadas, etc. Dentre as operações iniciadas e não finalizadas, há aquelas motivadas por falta de recursos computacionais. Neste universo é importante analisar se o comportamento interno da empresa, para uso dos serviços de Internet, por exemplo, esta possibilitando uma competição de recursos disponíveis para as transações eletrônicas. Como estudo de caso, foi utilizando a ferramenta de visualização PRISMA para análise de dados de transações eletrônicas e de redes. A ferramenta PRISMA é extensível, desenvolvida em Java, e possibilita a visualização de múltiplas visões de dados coordenadas. Foram realizadas melhoriasna ferramenta PRISMA quanto a técnica de coordenadas paralelas e adicionada a técnica heatmap, somadas as técnicas de dispersão de dados, treemap e outras visualizações mais comuns já existentes.

Palavra-chave: Visualização da Informação, Visões Coordenadas, Coordenadas Paralelas, HeatMap, Transações Eletrônicas.

ABSTRACT

Having a solid knowledge about clients and their enterprise activities allows a company to offer more suited services, improving clients fidelity and creating new business opportunities. One way to gather a better understanding on client behavior is the analysis of transactions history, using information on the types and duration of operations, average of unauthorized and unfinished operations, etc. Among the unfinished operations there are some motivated by insufficient computer resources. From this universe it is important to analyze if the internal enterprise behavior related to the use of internet services, for example, is causing a completion for computer resources. In this study, the PRISMA visualization tool was used to analyze data originated from electronic transactions and network data. PRISMA is an extensible Java tool that supports multiple coordinated views. In this work, the parallel coordinates were improvementand heatmap visualization techniques added to the PRISMA tool, in addition to the scatterplot, treemap and other existing visualizations.

Keywords: Information Visualization, Multiple Coordinated Views, Parallel Coordinates, Heatmap, Electronic Transactions.

1 INTRODUÇÃO

A prosperidade de um determinado negócio pode ser avaliada por vários critérios, sendo o principal deles a satisfação dos clientes, e em muitos momentos pode-se até traduzir satisfação do usuário por lucro do negócio. A satisfação dos clientes não é apenas um simples dado, mas uma sequência de informações e decisões sobre a qualidade dos serviços prestados (prever melhoria da infraestrutura baseada no crescimento dos clientes), os produtos oferecidos (definindo expansão, com variedade ou especialidade), conhecermelhor os clientes (que perfil de clientes há e que serviços e produtos eles adquirem e de que forma, regularmente ou esporadicamente), entre outras coisas.

Uma análise analítica de um crescente conjunto de dados não é uma tarefa trivial, relatórios gerenciais nem sempre respondem todas as questões, em muitas ocasiões o tomador de decisão não tem dados suficientes, pois os dados estão muito sumarizados ou se perde nos dados em função de muitos detalhes. Os relatórios quase sempre estáticos em sua forma dificultam o processo de correlação de informações, e o tempo quase sempre é um fator crítico para o sucesso da decisão a ser tomada.

Automatizar a geração de informação e manipulá-la de forma fácil e intuitiva é o que todo tomador de decisão procura. Nos últimos anos técnicas de mineração de dados, entre ouras, têm sido utilizadas para gerar essas novas informações sobre o negócio, porém nem sempre é uma tarefa fácil manipular e entender os resultados gerados. Para contornar essas dificuldades, cada vez mais ferramentas de visualização de informação estão surgindo para preencher a laguna deixada por estas técnicas, principalmente para os quesitos exploração e análise dos dados.

Uma ferramenta de visualização da informação tem como objetivo principal criar representações visuais para dados abstratos, e organizá-los espacialmente sob a ótica de alguma metodologia, e permitir de forma interativa que usuário possa explorar esse

espaço de itens visuais com determinada atividades, tais como: diminuindo ou aumentando a quantidade de itens visíveis, alterando características dos itens visuais, ou escolhendo outra metodologia para apresentação (rearranjo) dos itens visuais, correlacionado itens visuais, tudo isso para possibilitar ao usuário melhor percepção na exploração e análise dos dados.

Dentre as formas de transações de negócio que usuários realizam com empresas, uma que tem se destacado são as transações realizadas pela Internet ou eletronicamente. E que ainda apresentam um grande potencial de crescimento, ver Tabela 1, uma vez que aproximadamente 37,4% da população brasileira tem acesso a Internet, o que equivale a 75,98 milhões de internautas, e que desses apenas 31,7 milhões já realizaram alguma transação pela Internet (**Erro! Fonte de referência não encontrada.**), o equivalente a 41,7%.

Tabela 1: Quantidade de pessoas conectadas a Web no Brasil

Data da Pesquisa	População Total IBGE	Internautas (milhões)	% da População Brasileira	Fontes de pesquisa Internautas
2011/jun	203,4	75,98	37,4	InternetWorldStats
2008/dez	196,3	67,51	34,4	InternetWorldStats
2007/dez	188,6	42,6	22,6	InternetWorldStats
2006/dez	186,7	30,01	16,1	InternetWorldStats
2005/jan	185,6	25,9	14,0	InternetWorldStats
2004/jan	178,4	20,05	11,2	Nielsen NetRatings
2003/jan	176	14,32	8,1	Nielsen NetRatings
2002/ago	175	13,98	8,0	Nielsen NetRatings
2001/set	172,3	12,04	7,0	Nielsen NetRatings
2000/nov	169,7	9,84	5,8	Nielsen NetRatings
1999/dez	166,4	6,79	4,1	Computer Ind. Almanac
1998/dez	163,2	2,35	1,4	IDC
1997/dez	160,1	1,3	0,8	Brazilian ISC
1997/jul	160,1	1,15	0,7	Brazilian ISC

Compilado por www.e-commerce.org.br / fonte: pesquisas diversas / população: variações anuais estimadas. / Internautas refere-se a quantidade de pessoas que tem acesso à Internet nas residências, no trabalho ou locais públicos.

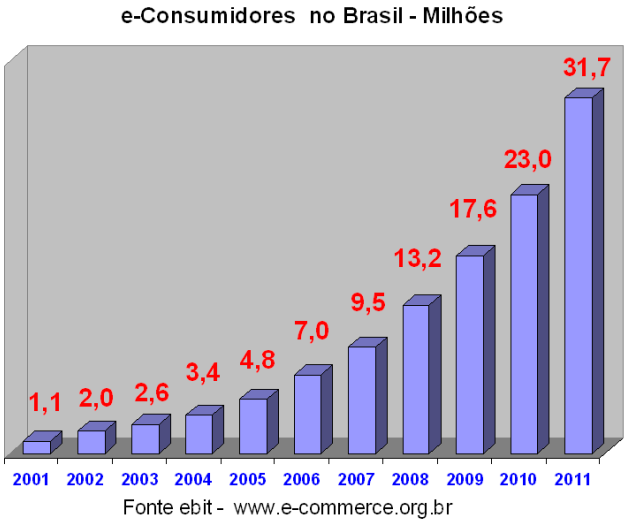

Figura 1: Número de e-consumidores no Brasil (Fonte: http://www.e-commerce.org.br)

Para uma empresa prosperar, é fundamental conhecer o comportamento de seu cliente, tanto para buscar melhorias quanto oportunidades de negócio. Para buscar oportunidades de negócio, uma das maneiras é entender mais sobre o comportamento dos clientes através da análise do histórico de transações eletrônicasque ele fez com a empresa e com empresas parceiras. Normalmente a quantidade de transações é enorme, demandando ferramentas que auxiliem a exploração e análise dos dados em busca por padrões ou exceções, em dados do tipo de operações realizadas, tempo médio das operações, operações não autorizadas, operações iniciadas e não finalizadas, etc. Na busca por melhorias, deve-se analisar dados que permitam responder por que a infraestrutura da empresa foi responsável pelo cliente não conseguir completar uma transação. Destacam-se possíveis origens do problema: uso indevido da infraestrutura TI de forma a concorrer com as tentativas de transações dos clientes, ataques internos e externos a infraestrutura de TI da empresa que possibilite lentidão ou paralização dos serviços aos clientes, falhas ou falta de infraestrutura de TI, etc. Entre os possíveis problemas foi escolhida apossibilidade do uso indevido da infraestrutura de TI pela empresa. Essa abordagem foi escolhida pelo impacto que o comportamento interno da empresa pode ocasionar ao próprio negócio, sendo possível solução de contorno do problema com custos baixo, normalmente envolvendo medidas mais restritivas no uso da infraestrutura de TI, e consequentemente mudanças na política de segurança da empresa, e facilidade na obtenção de dados para análise. Para isso análise de dados de

tráfego de redes, logs de firewall, logs de sistemas de detecção de intrusos, logs de servidores proxys, etc, são potenciais conjunto de dados a serem analisados.

Para análise e exploração de logstransacionais e de rede foi utilizada a ferramenta PRISMA (Godinho, P.; Meiguins, B.; Carmo, C.; Garcia, M.; Almeida, L.; Lourenço, R., 2007). PRISMA é uma ferramenta de visualização de informação que possibilita múltiplas visões coordenadas. A ferramenta foi desenvolvida totalmente em Java, sendo extensível em diversos tipos e números de técnicas de visualização de informação. Além disso, também possui recursos importantes para uma boa ferramenta de visualização da informação, como seleção, filtros dinâmicos, zoom, configurações dos atributos, gráficos estatísticos, relatórios customizados, acesso a diversas fontes de dados, entre outros.

Esta dissertação aperfeiçoouna ferramenta PRISMA odesempenho da técnica de coordenadas paralelas, que é uma técnica de visualização de informação multidimensional, permitindo a análise de grandes volumes de dados. Esta técnica exibe as dimensões de forma paralelas umas com as outras em um plano, representadas por eixos, e relacionar informações entre si a partir de linhas que interceptam cada eixo(Inselberg & Dimsdale, 1990). Posteriormente, visando agregar o aspecto temporal para análise mais facilitada foi proposta a inclusão na ferramenta PRISMA da técnica Heatmap. A técnica de HeatMap, ou mapa de calor, é uma representação gráfica dos dados, onde um determinado dado esta organizado em uma matriz e seu valor é representado por uma cor (Wilkinson & Friendly, 2008).

Como estudo de caso, foi utilizado registros de logs de dados de uma empresa. Apesar da autorização de uso, os dados de redes, ou aplicações, que possam identificar os usuários e a empresa foram alterados. Foram disponibilizados alguns registros de logs de determinados sistemas para análise, as características das bases e as descobertas da exploração serão apresentadas ao longo da dissertação.

1.1 Objetivos

O objetivo principal desta dissertação é aperfeiçoar uma ferramenta de visualização da informação com múltiplas técnicas de visualização coordenadas de dados entre si, agregando a visualização de dados temporais com a técnica Heatmap, facilitando a exploração e a análise de logs de transações eletrônicas e logs de eventos de redes computadores.

1.1.1 Objetivos Específicos

Como objetivos específicos destacam-se:

- Aprimorar a performance da técnica Coordenadas Paralelas na ferramenta PRISMA;
- Agregar na ferramenta PRISMA a técnica HeatMap para analise de dados temporal;
- Coordenar as técnicas desenvolvidas as já existentes na ferramenta PRISMA;
- Definir cenários de análise de transações eletrônicas e logs de eventos de rede;
- Avaliar com especialistas a ferramenta PRISMA, com os cenários definidos, e os padrões e exceções encontradas.

1.2 ORGANIZAÇÃO

O texto da dissertação está organizado na forma que segue:

No Capítulo 2, são apresentados conceitos sobre visualização de informação e coordenação de visões de dados. Nele, são feitas explanações sobre definições e termos relacionados à área. É apresentada uma visão geral das principais recomendações para uma boa ferramenta de visualização de informação com múltiplas visões coordenadas.

No Capítulo 3, é apresentado o conceito de visualização de segurança da informação, sendo presentados também as características do log do ISASERVER e trabalhos relacionados.

No Capítulo 4, é apresentada a ferramenta PRISMA e os aspectos de implementação de coordenadas paralelas e HeatMap, assim como as funcionalidades das ferramentas.

No Capítulo 5, são apresentados estudos de caso de análise de transações eletrônicas e logs do ISASERVER. Também é apresentado uma analise de log pelo HeatMap.

No Capítulo 6, são realizadas as considerações finais e citados os trabalhos futuros.

2 VISUALIZAÇÃO DA INFORMAÇÃO

2.1 Definição

A Visualização daInformação (VI), às vezes chamada de visualização de negócios, ou simplesmente visualização, é uma representação visual interativa que transforma dados abstratos em uma representação visual que é compreendida prontamente por um usuário, podendo então gerar um novo conhecimento da relação entre os dados. Pode ser usada para tarefas como identificação, correlação multivariada, procura, consulta, exploração e comunicação. Os dados são tipicamente quantitativos ou categorizados, mas também podem incluir: texto não estruturado, tipos de mídias diferentes e objetos estruturados (SPENCE, 2007) (CARD, 1999).

Há um campo relacionado, e algumas vezes sobreposto, à visualização de informação chamada de "visualização científica". A visualização científica se preocupa em representar visualmente uma simulação tridimensional de algo real. Por exemplo, nuvens fluindo através de uma cadeia de montanhas, dada certa condição do vento (SPENCE, 2007). Este trabalho não trata de visualização científica, entretanto muitas das técnicas que serão apresentadas são pertinentes às duas áreas.

Para exemplificar a potencialidade da representação gráfica é mostrada na Tabela 2 dados sobre as despesas de um estudante. Sugere-se que sejam respondidas algumas questões com relação a análise da Tabela 2:

- Qual é a maior despesa durante o referido período?
- Qual é a segunda maior despesa?
- Qual despesa oscilou constantemente com o tempo?
- Qual despesa apresentou uma tendência de redução?

Tabela 2: Tabela de despesas do estudante (Nascimento & Ferreira, 2005)

	Condomínio	Luz	Conta Telefone	Aluguel Telefone	Aluguel Apartamento	Faxina	Multicanal	Total
AGO	179,61	14,58	51,40	40,00	267,08	52,40		**605,07**
SET	183,81	23,50	38,35	40,00	267,08	52,40		**605,14**
OUT	201,21	30,24	149,00	40,00	267,08	52,40		**739,93**
NOV	219,73	35,94	143,95	40,00	232,08	52,40		**724,10**
DEZ	238,10	27,30	164,10	40,00	232,08	52,40		**753,98**
JAN	168,90	24,19	126,68	40,00	217,08	52,40		**629,25**
FEV	160,10	15,89	25,49	40,00	225,00	52,40		**518,88**
MAR	148,00	21,60	148,88	40,00	243,55	52,40		**654,43**
ABR	170,35	23,84	174,76	40,00	267,08	52,40		**728,43**
MAI	152,55	27,13	132,51	40,00	267,08	52,40		**671,67**
JUN	157,70	24,19	56,90	40,00	319,00	52,40		**650,19**
JUL	162,25	26,09	254,52	40,00	267,08	52,40		**802,34**
AGO	171,25	21,25	185,74	40,00	267,08	52,40	59,90	**797,62**
SET	155,85	29,55	114,42	40,00	267,08	52,40	59,90	**719,20**
OUT	148,90	28,68	171,74	40,00	265,00	52,40	59,90	**766,62**
NOV	150,35	15,38	98,16	40,00	265,00	52,40	57,90	**679,19**
DEZ	132,20	49,77	183,39	40,00	267,08	82,40	59,90	**814,74**
JAN	148,32	26,44	114,57	40,00	267,08	52,40	59,90	**708,71**

As primeiras duas perguntas com certo esforço podem ser respondidas com base nos dados da tabela, mas podem ser mais facilmente identificadas em um gráfico de pizza (Figura 2).

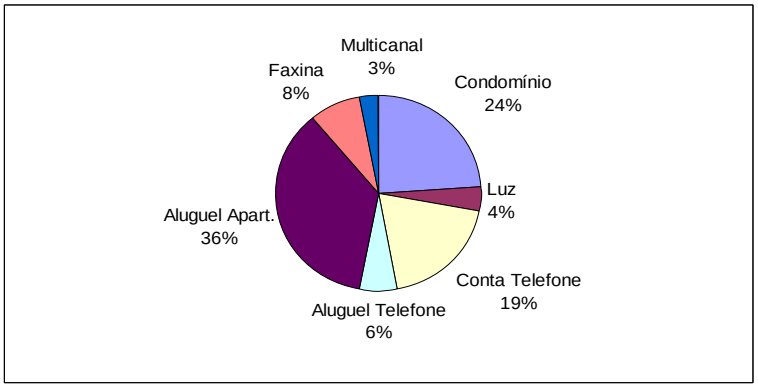

Figura 2: Gráfico de pizza das despesas(Nascimento & Ferreira, 2005)

As duas últimas perguntas reforçam a potencialidade do uso de uma técnica de visualização para representar os dados (Figura 2: Gráfico de pizza das despesasFigura 3). As respostas podem ser mais rapidamente encontradas, pois a percepção do usuário em relação aos dados e seus relacionamentos melhoram ao analisar um gráfico.

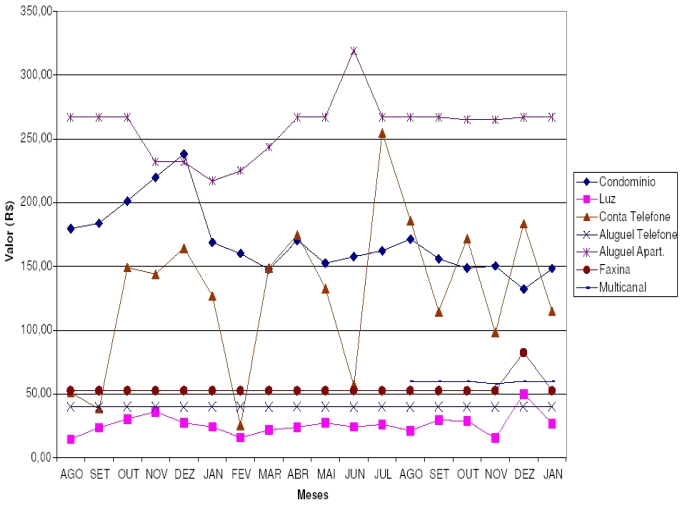

Figura 3: Gráfico de linhas das despesas(Nascimento & Ferreira, 2005)

2.2 Visualização de Informação sem o Computador

Mesmo antes do advento do computador, as técnicas de visualização de informação eram utilizadas para descobrir ou demonstrar algum relacionamento implícito entre os dados. Uma das primeiras ocorrências do uso destas técnicas visuais ocorreu por volta de 1861 quando MonsierMinard demonstrou através de um mapa a campanha fracassada de invasão a Rússia por Napoleão em 1812-1813 (Figura 4), contendo informações de temperatura, topografia, quantidade de soldados, distância percorridas, direção do deslocamento, tudo isto em relação ao tempo e percurso(Spence, 2007).

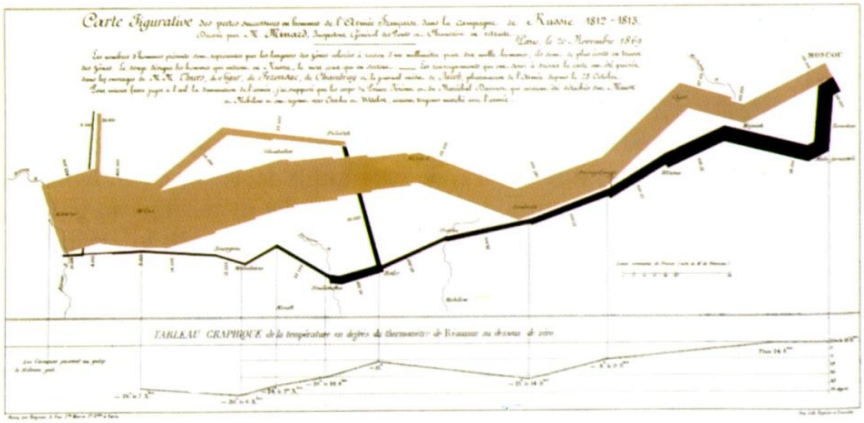

Figura 4: Mapa feito por Minard sobre a Marcha de Napoleão a Moscou (Spence, 2007)

Outro exemplo foi utilizado por Florence Nightingale, uma heroica enfermeira que cuidava de soldados da linha de frente nos hospitais do exército britânicos na guerra da

Crimea (1855). O que não era muito conhecido é que como resultado das observações das precárias condições hospitalares feitas por ela para atendimento dos soldados feridos, conseguiu através de sua técnica de visualização de informação, persuadir uma comissão sanitária da Inglaterra a empreender melhorias nos hospitais na frente de batalha para salvar mais vidas. Demonstrou que com o passar do tempo, morriam mais soldados nos hospitais do que no campo de batalha(Figura 5) devido as precárias condições hospitalares(Spence , 2007).

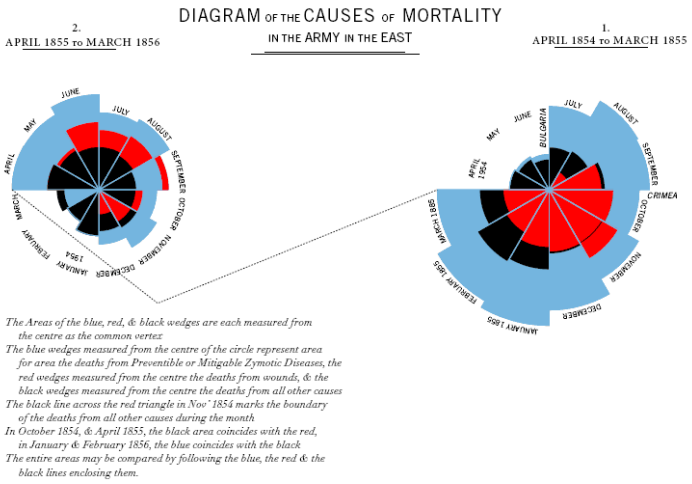

Figura 5: Visualização de Florence Nightingale envolvendo comportamento ao longo do tempo e comparações (Spence, 2007)

Estatísticos em 1873 criaram uma matriz para visualização resumindo 40 mapas de Paris demostrando as características de nacionalidade, profissões, idade, entre outros, de 20 distritos, usando uma escala de cor do branco(Figura 6), através do amarelo e azul, até ao vermelho(alto).

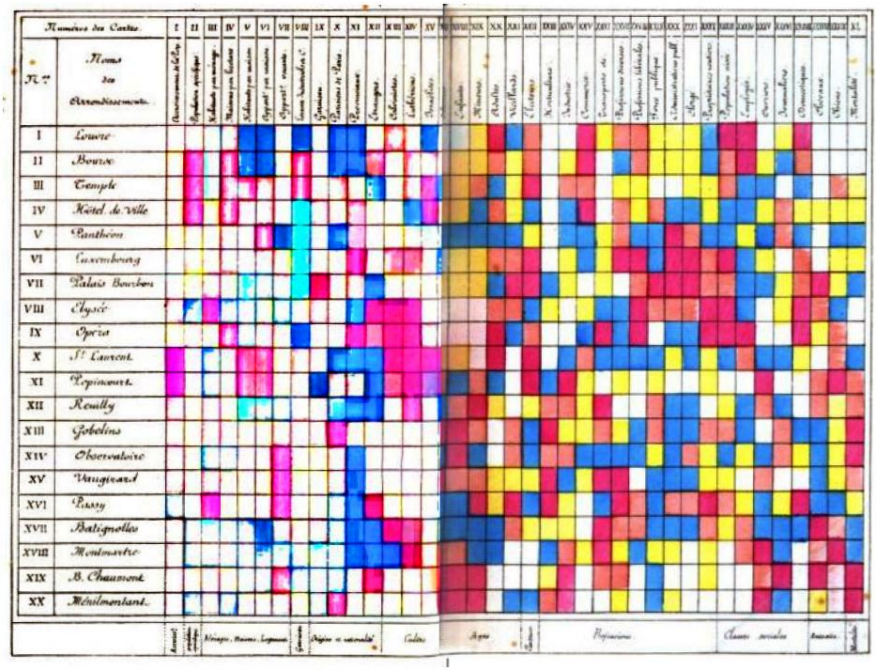

Figura 6: Gráfico de Matriz de 1873 (Wilkinson & Friendly, 2008)

Um avanço relevante para visualização da informação foi a representação abstrata para eixos apresentadas por Willian Playfair (Friendly, 2009), permitindo que outros dados pudessem ser utilizados como coordenadas. Desta forma, pode-se ter densidade em um dos eixos e temperatura em outro. Por exemplo, a (Figura 7), desenvolvida por Playfair, apresenta o debito de uma nação ao longo do tempo. Outro exemplo de Playfair inclui o gráfico de série temporal sobre a balança comercial entre Inglaterra e Noruega/Dinamarca ao longo de alguns anos (Figura 8).

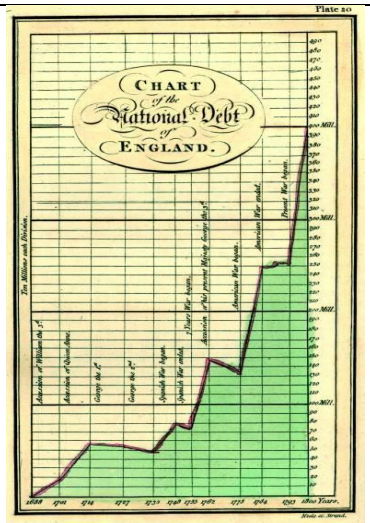

Figura 7: Playfair - Uso de valores abstratos nos eixos

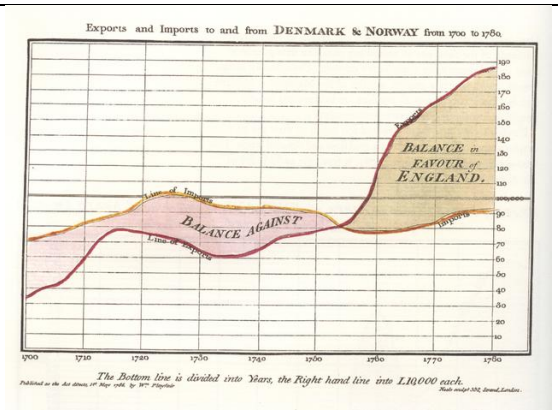

Figura 8: Gráfico de série temporal sobre a balança comercial

2.3 Visualização de Informação com o uso do Computador

Embora nenhum computador estivesse envolvido nos exemplos anteriores, a simplicidade das ilustrações nos permite, não obstante, identificar assuntos significantes associados com a VI, características que ainda são mais pertinentes quando o poder do computador está disponível. São essas:

- Seleção: dentre todos os dados disponíveis o usuário deve selecionar aqueles que achar pertinente para tarefa que será desenvolvida;
- Representação: o desenvolvedor de uma ferramenta de visualização tem que representar dados abstratos utilizando cores, formas e direções, entre outros;
- Apresentação: é a forma de disposição dos dados ao usuário, isso se torna mais complexo se os recursos gráficos forem escassos como em celulares, handhelds, etc.;
- Escala e Dimensionalidade: a quantidade de dados armazenados eletronicamente é enorme e a percepção do usuário em relação a esses dados é pequena, então a questão é: como representar mais atributos na visualização e melhorar a percepção do usuário simultaneamente? Que técnicas suportam alta dimensionalidade?;
- Rearranjo, interação e exploração: uma nova visão sobre os dados sempre gera uma nova percepção sobre os mesmos. Quanto melhor o conhecimento sobre os dados e seus relacionamentos, melhor será a tomada de decisão. Essas novas visões são possíveis através de rearranjos dos dados na visualização, filtros, seleção, entre outros. A interface deve permitir isso de forma fácil, rápida e intuitiva;
- Modelos mentais: a visualização tem sido considerada uma atividade essencialmente humana, embora apoiada efetivamente pelo computador. Se as pesquisas puderem identificar como esse processo ocorre, os desenvolvedores poderão projetar ferramentas de visualização mais eficientes e eficazes. Infelizmente a compreensão da mente humana ainda é muito limitada.

Outra mudança que se deve destacar com o uso do computador em visualização de informação é em relação à interação do usuário com este processo. Antes do advento do computador em VI, o autor da visualização realizava a seleção, representação e apresentação dos dados de acordo com a sua compreensão da tarefa a ser executada, compreensão essa que deveria ser igual à do espectador, o que nem sempre era simples por serem pessoas diferentes. Assim, o usuário ficava limitado à visão do autor, a uma representação estática (Figura 9). Agora, com a disponibilidade de computadores com alto poder de processamento, permite-se a possibilidade do usuário interferir em todas as etapas do processo de visualização (Figura 10), com uma liberdade definida pelo autor da visualização, atuando em cima de uma visualização dinâmica (SPENCE, 2007).

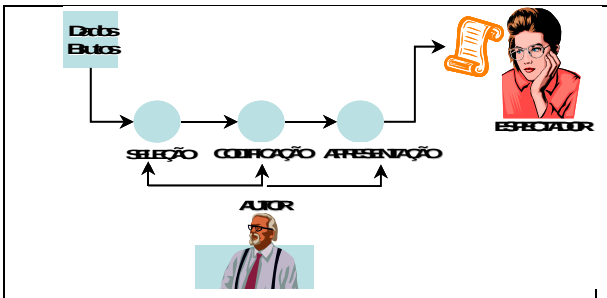

Figura 9: Usuário não participa da concepção da visualização da informação

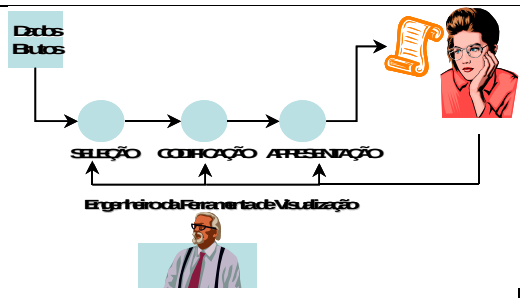

Figura 10: Usuário participa da concepção da visualização da informação

As possibilidades de aplicações de visualizações são inúmeras, no dia a dia já é possível encontrar comumente mapas, grafos e gráficos simples, utilizados para as mais diversas tarefas, como por exemplo: traçado de roteiros, análises econômicas, estatísticas, tendências de mercado, etc. Os tipos de objetos que podem ser visualizados são os mais diversos: dados, algoritmos, resultados de cálculos, processos, entre outros. A ideia principal é que o usuário possa obter prontamente informação da representação visual, mas que ela seja interativa, gerando novas informações e percepções do objeto de estudo.

2.4 Características de uma boa ferramenta de visualização

Segundo Shneiderman (1996) uma ferramenta de Visualização de Informação deve realizar as seguintes tarefas:

- Visão geral: o usuário precisa ganhar uma noção sobre todos os dados que serão analisados. Essa noção está baseada nos parâmetros que o usuário escolheu para a visualização, nos limites do dispositivo gráfico usado e de sua percepção. Os atributos gráficos mais usados são: posição, cor, tipo de representação e tamanho.
- Zoom: a técnica de zoom é importante porque permite focar em certo subconjunto dos dados para análise, ou seja, analisar um determinado contexto. Além disso, conforme se aplica o zoom, mais detalhes sobre uma determinada visão dos dados são mostrados, o que se chama de zoom semântico.
- Filtro: usuários frequentemente precisam reduzir o tamanho do conjunto de dados, eliminando itens com base em seus atributos. Uma das maneiras mais eficientes é o uso de consultas dinâmicas.
- Detalhes sob demanda: quando o usuário está explorando um conjunto de dados, ele necessita ver detalhes sobre um item em particular. Isto é normalmente feito usando o clique do mouse, onde as informações adicionais podem aparecer em uma janela auxiliar, ou na própria visão dos dados (visualização).

Adicionalmente, pode-se incluir mais duas características (SPENCE, 2007):

- Relacionamentos: Se o usuário descobre um item de interesse, ele pode precisar identificar outros itens com atributos similares.
- Histórico: o usuário precisa de suporte para desfazer uma ação e mostrar os passos até aquele ponto.

2.5 Múltiplas Visões Coordenadas

Sistemas de múltiplas visões usam duas ou mais representações visuais distintas para auxiliar o processo de investigação de uma única entidade conceitual

(BALDONADO, 2000). Uma visão é considerada distinta das outras se permitir ao usuário aprender sobre diferentes aspectos da entidade conceitual, ou pela apresentação de informações diferentes, ou enfatizando diferentes aspectos da mesma informação, por exemplo, utilizando representações diferentes ou técnicas de visualização diferentes.

De acordo com North andShneiderman (2000) e Baldonado (2000), o uso de sistemas de múltiplas visões coordenadas apresenta algumas vantagens na análise, entre elas destacam-se: melhora do desempenho do usuário na percepção dos dados, facilita a descoberta de relacionamentos não triviais entre os dados, minimiza o overhead cognitivo de uma única visão ou de uma visão mais complexa, entre outras.

Os sistemas de visualização de informação que utilizam múltiplas visões coordenadas podem ser classificados por níveis de flexibilidade em relação aos dados, visões e coordenação. São eles:

- Dados: usuários podem utilizar diferentes conjunto de dados em suas visualizações;
- Visões: usuários podem escolher diferentes conjuntos de visualização para determinado conjunto de dados;
- Coordenação: usuário poderá escolher diferentes tipos de coordenação entre pares de visões para auxiliar sua necessidade de exploração dos relacionamentos entre os dados.

Para o desenvolvimento de sistemas de visualização de informação com múltiplas visões coordenadas, as recomendações mais frequentes de uso são (BALDONADO, 2000):

- Quando há uma diversidade atributos, modelos, perfis de usuário, níveis de abstração ou gênero;
- Quando as visões diferentes destacam correlações ou disparidades;
- Quando há necessidade de diminuir a complexidade do conjunto de dados, utilizando múltiplas visões mais simples;
- Usar múltiplas visões minimamente, justificar o uso de múltiplas visões versus custo de aprendizado do usuário e espaço de visualização.

Pillat (2005) destaca como principais possibilidades de coordenação de múltiplas visões:

- Seleção: itens de dados selecionados em uma visão são destacados em outras visões;
- Filtro: reduzir o conjunto de dados para análise em todas as visões;
- Cor, Transparência e Tamanho: características visuais para representar a variação de valores de um dado atributo dos dados em todas as visões;
- Ordenação: valores de um atributo definem a ordem das representações visuais dos dados;
- Rótulo: determina que conteúdo os rótulos exibirá para cada item de dados das visões;
- Manipulação de Atributos: permite ao usuário adicionar/remover atributos das visões de dados.

Dado o contexto de múltiplas visões coordenadas, destacam-se os principais desafios no desenvolvimento de sistemas de múltiplas visões coordenadas:

- Os mecanismos de coordenação;
- Requisitos computacionais para renderização das visões;
- Disposição da interface – layout, com espaço normalmente muito reduzido para novas visões;
- Interação do usuário entre as diversas formas de visualização;
- Aspectos cognitivos relacionados ao uso de sistemas de múltiplas visões coordenadas:
    - Tempo e esforço necessário para o aprendizado do sistema;
    - Sobrecarga de informações na memória de trabalho do usuário;
    - Esforço necessário para comparação;
    - Esforço necessário para troca de contexto.

2.6 Tipos de Dados Versus Tipos de Visualização

É natural pensar que um ambiente tridimensional seja um ambiente melhor para a representação de dados. Contudo, nem sempre três dimensões são necessariamente

melhor do que duas dimensões para visualização de dados. Um dos critérios para essa escolha é o tipo de dado que se quer visualizar. De acordo com Shneiderman (SHNEIDERMAN, 1996), há sete tipos diferentes dados que descrevem diferentes tipos de visualizações. São eles:

- 1-Dimensão: este tipo de dado é representado por texto ou dados similares, como linhas de código. Podem haver outras informações associadas a ele, como data da criação, tamanho, data da última modificação, etc. Uma técnica bastante associada a esse tipo de dado é o uso de linhas com cores e larguras variadas, representando outros atributos;
- 2-Dimensões: este tipo de dado inclui dados geográficos, plantas de engenharia, etc. Pode-se associar uma grande quantidade de atributos com uso de cores, tamanhos e formas diferentes.
- 3-Dimensões: o volume de um objeto torna-se importante, um tributo a mais. Se o contexto do mundo real puder ser incluído para melhorar a percepção do usuário é mais indicado ainda. Não se deve deixar de mencionar problemas inerentes a uma visualização 3D, como a oclusão, quando parte de um dado esconde outro. Para isso, técnicas de visões diferenciadas, transparência e slicing são necessárias.
- Temporal: este tipo de dado reúne todas as características dos dados citados acima mais o atributo tempo. Para o atributo tempo o mais indicado é formar uma dimensão. Os gráficos "tempo versus algum atributo" são bastante utilizados e conhecidos. A animação deve ser considerada quando há uma grande quantidade de dados.
- Multidimensional: base de dados relacional ou estatística pode ser considerada como pontos em um espaço multidimensional. Técnicas como consultas dinâmicas e diagramas de dispersão são bastante úteis.
- Hierárquico: útil para classificação de dados. Normalmente é representado por diagramas com nós, com ligações entre os mesmos.
- Rede: dados de rede são nós conectados por links previamente definidos. Esses links podem ser organizados em árvores ou em hierarquias, e a melhor maneira de manipulação é permitindo mudar o foco sobre os nós.

Entre eles, destacamos o tipo de dados temporal que reúne as características dos dados citados acima mais o atributo tempo. Para o atributo tempo o mais indicado é formar uma ou duas dimensões. Os gráficos tempo x atributo são bastante utilizados e conhecidos. O processo de visualização de um domínio temporal (Figura 11) esta baseado em quatro passos denominados: tempo, ponto da visão no tempo, espaço-tempo, ponto da visão no espaço-tempo. (Daassi, 2008)

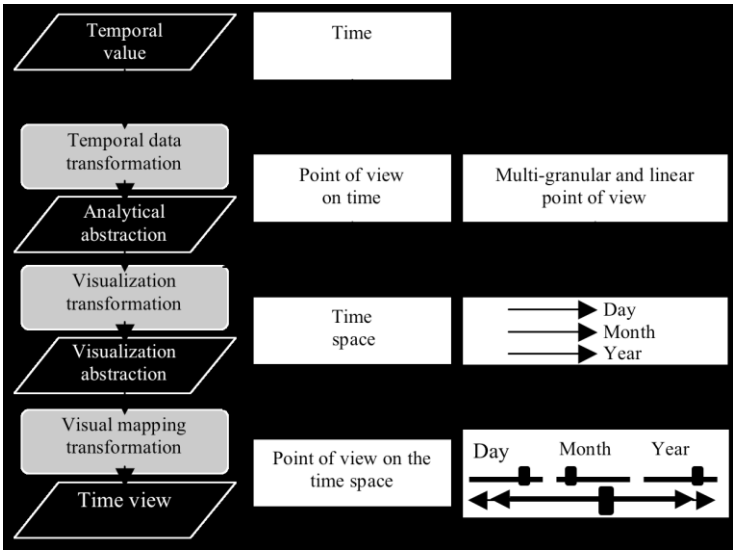

Figura 11: Processo de visualização de domínio temporal (Daassi,2008)

2.7 Técnicas de visualização de informação

As técnicas de visualização apresentadas nesta seção foram escolhidas por serem comentadas nas seções posteriores, e representam um subconjunto pequeno do universo de técnicas de visualizações de informação. Existem diversas técnicas de visualização da informação, e o uso de uma visualização específica depende da característica dos dados da análise a ser feita. O uso do conjunto das técnicas que será apresentado nas próximas sub-seções, a princípio, abrange uma vasta possibilidade de análise de dados como arquivos de logs de transações eletrônicas e de tráfego de rede, principalmente se forem analisados através da técnica de múltiplas visões coordenadas, para destaque de correlações ou disparidades.

RaffaelMarty (Marty,2008) relacionou um esquema (Tabela 3) para facilitar a identificação da técnica certa de acordo com alguns parâmetros, conforme abaixo:

- O numero máximo de dimensões que podem ser visualizados;
- O numero máximo de dados a serem visualizados;
- A melhor técnica em acordo com o tipo de dados;
- O cenário de caso de uso;
- Aplicação com segurança da informação.

Tabela 3: Tabela de tipos de gráficos adaptadas de Marty (Marty,2008)

Técnica de Visualização	Dimensão	Máximo de valores de dados	Tipo de dados	Caso de uso
Grafico de Pizza	1	Aproximadamente 10	Categórico	Comparar valores com proporção ou percentual
Gráfico de Barra	1	Aproximadamente 50	Categórico	Representar a frequência ou agregação de valores
Grafico de linha	1	Aproximadamente 50	Ordinal, Intervalo	Representar a frequência ou agregação de valores
Stacked Pie	2	Aproximadamente 10 series de 5	Categórico	Comparar valores em duas dimensões
Stacked Bar	2	Aproximadamente 50 séries de 5	Categórico	Representar a frequência ou agregação de valores em duas dimensões
StackedLine	2	Aproximadamente 50 séries de 10	Ordinal, intervalo para cada série de dados	Representar a frequência ou agregação de valores para duas dimensões
Histograma	1	Aproximadamente 50	Ordinal, contínuo	Usado para mostrar a distribuição de valores
Box Plot	2	Aproximadamente 10	Contínuo, categórico	Usado para mostrar a distribuição de valores com comparações.
Scatterplot	2 ou 3	Milhares	Contínuo	Detectar cluster e tendência nos dados
Coordenadas Paralelas	N	Milhares	Qualquer	Visualizar múltiplas dimensões
Link graph	2 ou 3	1000 (sem agregação)	Qualquer	Visualizar relacionamentos de valores em uma dimensão e através de múltiplas dimensões
Map	1	100	Coordenadas, Qualquer	Visualização de localização física
TreeMap	N	10.000	Categórico, qualquer	Visualizar estrutura hierárquica entre os dados

Acrescenta-se ao estudo acima a exemplificação da técnica de visualização HeatMap (seção 2.7.4) com o uso de 2 dimensões, com representação de milhares de registros de forma agregadas, com dados numéricos ou intervalo, para visualização do comportamento dos dados, por exemplo, relacionados ao aspecto temporal.

2.7.1 Treemap

A técnica Treemap (Figura 12)de visualização, desenvolvida durante a década de 1990, consiste na visualização de estruturas hierárquicas, através de métodos de visualização de preenchimento de espaço (Schneiderman, 1992), capazes de representar grandes coleções hierárquicas de dados quantitativos permitindo a usuários comparar valores de nós e sub-árvores através da profundidade da árvore(Schneiderman, 2009), usando cores e tamanhos, para demonstrar padrões e exceções.

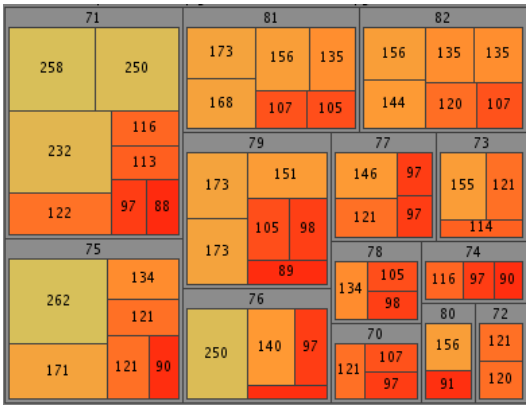

Figura 12: Exemplo de Visualização da Técnica Treemap

2.7.2 Dispersão de Dados

A técnica de dispersão de dados é uma das técnicas mais conhecidas. Ela consiste em relacionar os atributos da base com os eixos cartesianos ‘x’ e ‘y’, no caso bidimensional ou ‘x’, ‘y’ e ‘z’, no caso tridimensional (Spence, 2007) (Bachthaler&Weiskopf, 2008). Assim, é fácil perceber as associações entre as variáveis dentro do plano cartesiano. Além disso, um objeto pode representar vários atributos através de características como forma, tamanho e cor (Figura 13) e uso da técnica de

Brushing(Becker & Cleveland, 1987). Um dos principais problemas dessa técnica é a oclusão de um objeto por outro.

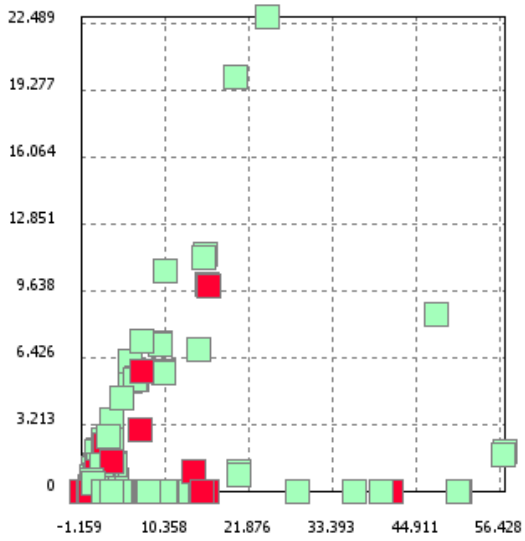

Figura 13: Técnica de dispersão de Dados

2.7.3 Coordenadas Paralelas

A técnica Coordenadas paralelas apresentada por Inselberg(Inselberg, 1985), demonstra em um plano bidimensional várias dimensões de atributos através da representação em eixos (Figura 14). Cada eixo no gráfico demonstra uma dimensão e o relacionamento dos dados entre as dimensões são representadas pelas linhas. A técnica é normalmente usada para demonstrar o relacionamento entre as dimensões.

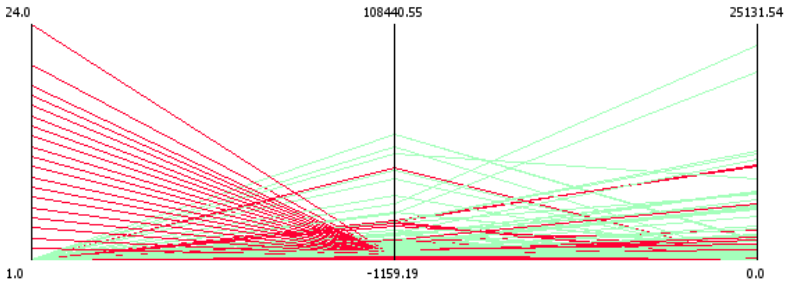

Figura 14: Exemplo da Técnica de Coordenadas Paralelas

2.7.4 HeatMap

Wilkinson e Friendly no artigo The historyofthe cluster HeatMap (Wilkinson&Friendly, 2008) define HeatMap como uma visualização que simultaneamente revela uma estrutura de agrupamento hierarquica entre linhas e colunas em uma matriz de dados, desenvolvida e aperfeiçoada durante séculos por estatísticos. Um HeatMap (mapa de calor) é uma representação gráfica de dados onde os valores assumidos por uma variável em uma tabela bidimensional são representados com cores, organizados em uma matriz de dados.

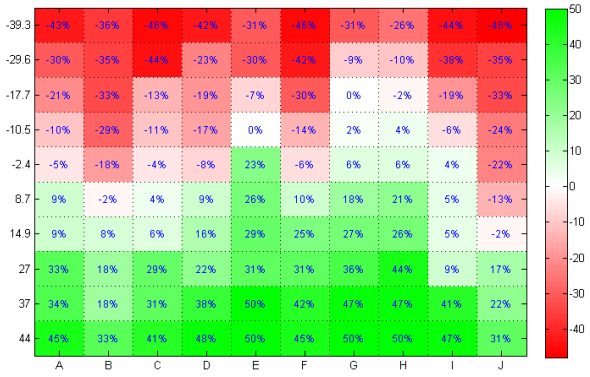

Figura 15: Exemplo da Técnica HeatMap

2.8 Trabalhos Relacionados

Os trabalhos relacionados apresentados nesta seção estão mais focados na utilização de visões coordenadas de dados.

- Snap-Together (North &Shneiderman, 2000) (Figura 16) é um sistema de propósito geral sistema que suporta vários tipos de conjunto de dados. Como permite diversos tipos de dados a coordenação é limitada a seleção de itens ou navegação.

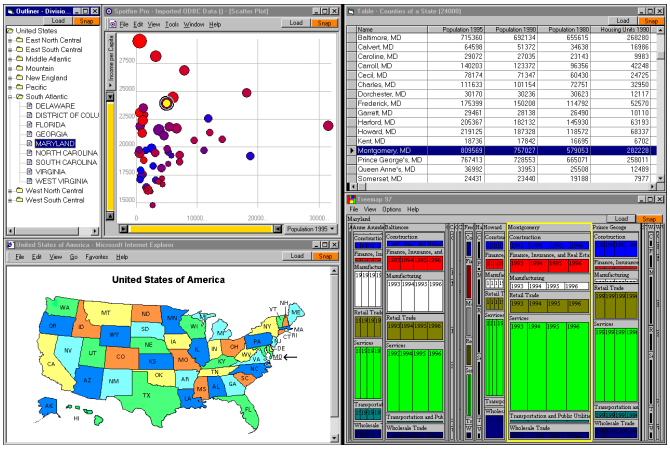

Figura16: Snap-Together (North &Shneiderman, 2000)

- GeoVista Studio (Figura 17) é uma ferramenta para visualização e análise de dados geo-científicos. Embora a ferramenta entregue técnicas de visualização interessantes para exibir dados multidimensionais, a mesma é limitadas nas formas de coordenação: seleção, destaque e consultas dinâmicas em mapas de cor(Takatsuka, 2002).

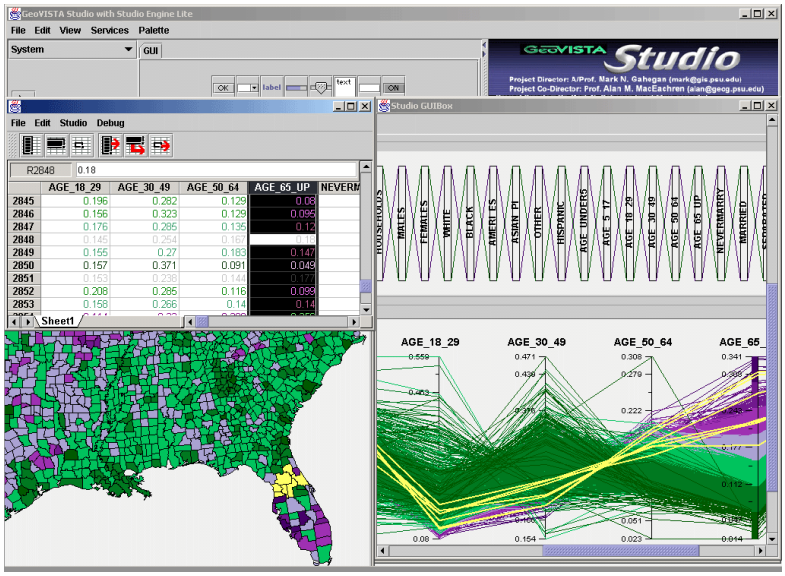

Figura 17: GEOVista Studio (http://www.geovista.psu.edu)

- Improvise (Figura 18) é um softwareque permite aos usuários construir e manipular interativamente múltiplas visões com alto nível de coordenação. A ferramenta possibilita o controle de funções simples, como navegação e seleção de objetos nas múltiplas visões(Weaver, 2004).

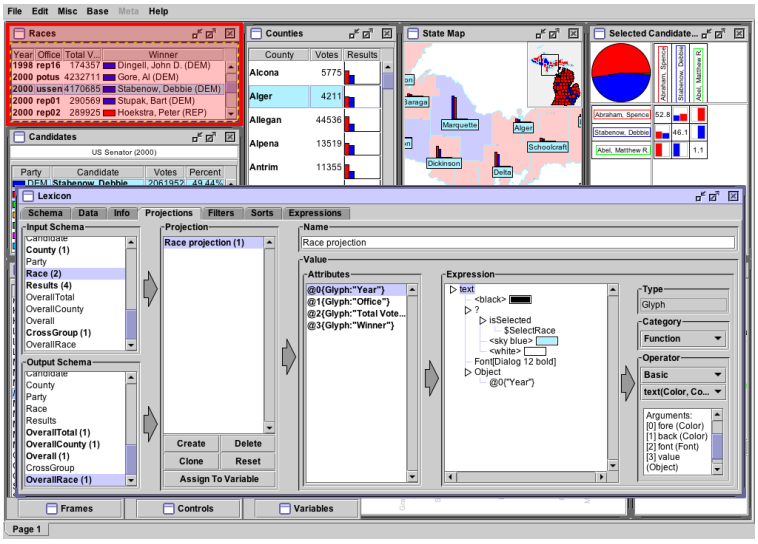

Figura18: Improvise(Weaver, 2004)

- Xmdv (Figura 19) representa uma matriz de dispersão de dados, permitindo coordenar brushing de n-dimensões, bem como a manipulação de dimensões.

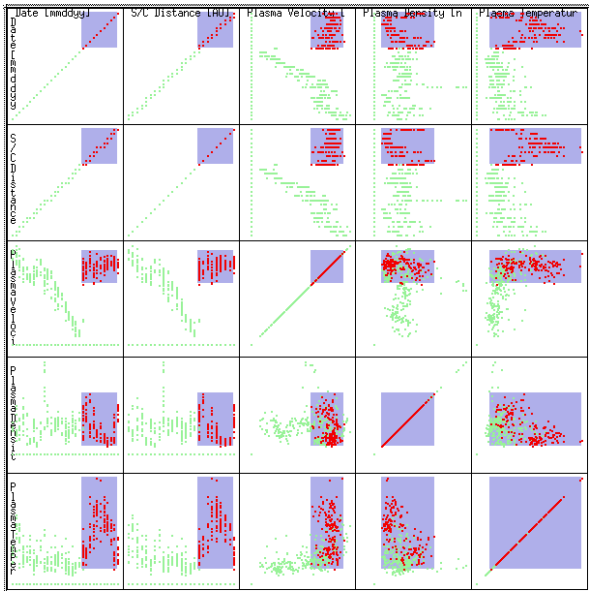

Figura 19: Xmdv - http://davis.wpi.edu/~xmdv/vis_parcoord.html

Observadas as características e comparações de algumas ferramentas, o próximo passo neste trabalho naturalmente seria a escolha da ferramenta que possa contribuir ou que possamos adaptá-la para analisar o comportamento de registros de logs. Contudo, antes de selecionarmos uma ferramenta própria a ser utilizada na pesquisa, convém ressaltar que a análise de logs vem crescendo em importância seja pela necessidade de conhecer o comportamento das transações, seja para detectar ataques e anomalias quanto à segurança da informação. Neste contexto de análise de logs, surgiu nos últimos anos uma área nova que consiste na Visualização de Segurança (Marty, 2008). Desta forma, no próximo capítulo, será apresentado o tema Visualização da Segurança e trabalhos relacionados que utilizam ferramentas com intuito de visualizar informações no contexto da segurança da informação e por consequência darmos ênfase na análise de logs de transações eletrônicas e de redes para analisar o comportamento das transações e do ambiente interno de uma determinada empresa, com intuito de identificar problemas e propor soluções para a continuidade das transações eletrônicas.

3 VISUALIZAÇÃO DA SEGURANÇA DA INFORMAÇÃO

A segurança da Informação é importante para uma organização, seja pelo fator de competitividade do negócio, seja pela confidencialidade, integridade e disponibilidade dos dados armazenados ou trafegados.

Apesar da existência de várias normas sobre o assunto Segurança da Informação (Norma ISO/IEC 27001, Norma ISO/IEC 27002,Norma ISO/IEC 27003, Norma ISO/IEC 27004, Norma ISO/IEC 27005), que contribuem na adoção de boas práticas para área, a identificação de eventos e riscos ainda é um trabalho árduo que compete aos analistas de segurança, realizado através de inúmeros mecanismos e procedimentos de monitoração, gerando grandes quantidades de logs para análise. Como logs de mecanismos de segurança incluem-se: os logs de sistemas operacionais, logs de firewall, logs de sistemas de detecção de instrução, coletas de tráfego da rede, logs de WebProxy, logs de correio eletrônico, traces de banco de dados, logs de auditoria e ações executadas de usuários dentro de sistemas legados entre outros.

Estes mecanismos e dispositivos possuem ferramentas para geração de logs, filtros e sumarização, contudo geralmente estas ferramentas não são eficientes para análise de grandes quantidades de dados. Um dos objetivos da visualização de segurança é agregar as técnicas de visualização para explorar e descobrir eventos e incidentes com intuito de permitir a tomada de decisão para melhoria do Sistema de Gestão de Segurança da Informação (SGSI).

Existem diversos desafios sobre a visualização de segurança da informação. Um dos principais pontos é simplesmente entender a estrutura dos dados, de como estão armazenados e o que significa cada registro dentro da estrutura de armazenamento ou na captura do tráfego de pacotes. Dados de segurança podem estar associados ao tempo ou não. O fato de não estarem associados ao tempo dificulta bastante a análise do comportamento de um eventual ataque. Os logs de segurança devem possuir na sua

estrutura além do entendimento do que significam, deveriam também responder as questões tais como o quê, como, quando e onde. Existem diversos problemas comuns para análise de dados, apesar de esforços de grupos para padronizar os logs, os trabalhos ainda estão no início (Marty, 2008). Com a falta de padrão da indústria de dispositivos de segurança da informação para os logs, existe grande dificuldade na automatização do processo de *parsing*dos mesmos e por consequência sua integração com demais logs de outros dispositivos.

Outros problemas significantes são a ocorrência de informações incompletas nos logs e a falta de sincronismo. Normalmente um incidente de segurança é registrado por vários dispositivos, como por exemplo, um firewall, por um sistema de detecção de intrusão e por um log de auditoria do sistema operacional da máquina afetada pelo incidente. A falta de sincronismo ou a geração de logs incompletos dificulta bastante a analise do analista de segurança da informação para entender o incidente de segurança.

Não é intuito deste trabalho resolver estes problemas através de um padrão, contudo entender que eles existem e que devem ser tratados de acordo com cada tipo de log analisado, é importante para construção de uma visualização adequada, a partir da característica de dados a ser avaliada.

Uma forma de conhecer o comportamento de usuários na rede é a análise de logs de serviços que utilizam servidores proxy que realizam o papel de filtrar e acelerar acesso através de ambiente web. Entre os servidores de proxy podemos citar o Internet Security andAcceleration Server (ISA Server) que registra centenas de linhas de logs.

3.1 Log do ISA Server

.Uma das funções do ISA Serveré ser um filtroque analisa o cabeçalho dos pacotes de IP e analisa os dados na busca de tráfego não permitido.

O ISA Server possui no seu módulo gerencial relatórios pré-definidos que ajudam o analista a analisar o tráfego na rede dentro dos parâmetros estipulados, agrupados em cinco categorias subdivididas conforme a Tabela 4: Tipos de dados disponíveis nos logs do ISA ServerTabela 4:

Tabela 4: Tipos de dados disponíveis nos logs do ISA Server

Categoria	SubCategoria
Sumary	Protocols Top Users Top Websites Cache Performance Traffic Daily Trafic
Web Usage	Top Web Users Top WebSites Protocols Http Responses Object Types Top Browsers Operating Systems
ApplicationUsage	Protocols Top Application Users Top Applications Operating Systems Top Destination
Traffic&Utilization	Protocols Traffic Cache Performance Connections Processing Time Daily traffic Errors
Security	Authorization Failures Dropped Packets

Os relatórios padrões do ISA Server apresentam a análise dos requisitos de segurança previamente estabelecidos e conhecidos, e o que faz com eficiência, tais como: a confirmação de dados conhecidos em relação aos padrões esperados e exceções. O que se deve pensar no momento da análise é se realmente estas formas de apresentação dos dados em forma de relatórios e gráficos simples (Figura 20) são suficientes relacionar categorias e subcategorias, e proporcionar a exploração dos dados

e a descoberta de novas informações, sem a necessidade de analisar milhares de linhas de logs, além de permitir análise do comportamento dos dados em determinada situação em correlacionada com algum outro atributo, como por exemplo, o tempo.

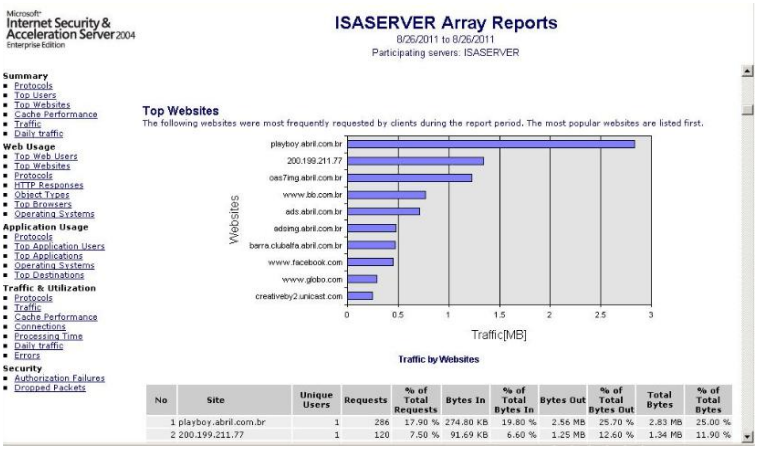

Figura 20: Relatório gerado pela Isaserver.

3.2 Trabalhos Relacionados

Os trabalhos relacionados nesta seção estão mais focados em ferramentas de visualização aplicadas na área de segurança da informação.

Komlodi (Komlodi, A. Goodall, J. R. Lutters, W. G, 2004) propôs um framework para visualização de informação para detecção de intrusão (Figura 21), onde foram definidas as etapas de monitoramento, análise e resposta, e para cada etapa foram definidas as tarefas dos analistas e características de técnicas de visualização são mais adequadas.

Phase	Analyst Tasks	Visualization Needs
Monitoring	• Monitoring all attack alerts • Identifying potentially suspicious alerts	• An overview of the alert data • Simple displays • Support for pattern and anomaly recognition • Flexibility • Speed of processing
Analysis	• Analyzing alert data • Analyzing other related data • Diagnosing attack	• Multiple views, zoom, drill down, focus + context solutions • Correlation between displays, linked views • Filtering and data selection
Response	• Responding to attack • Documenting and reporting attack • Updating IDS	• Suggestion for response action • Incident reporting • Annotation/feedback to facilitate future analysis • Saving views • Historical display • Reporting data transfer

Figura 21: Etapas e Características de tarefas de análise (Komlodi, A. Goodall, J. R. Lutters, W. G, 2004)

Koike e Ohno (2004) desenvolveram um módulo de visualização para agregar ao Snort, chamado SnortView(Figura 22), que tem o objetivo de visualização em tempo real para detecção de ataques analisando o tráfego da rede, disponibilizando um visualização de dados 2D multidimensional. A ferramenta é dividida em três partes: origem, quadro de alertas e matriz origem-destino.

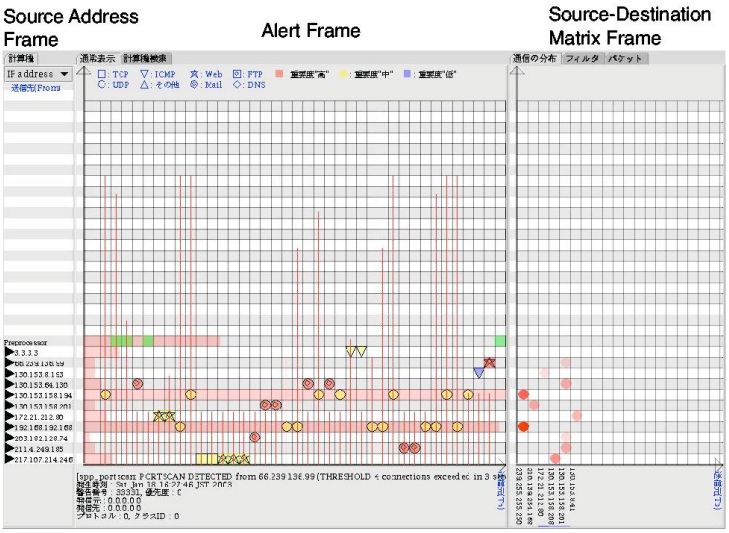

Figura 22: SnortView (Koike e Ohno, 2004)

Malécot (Malécot, 2006) desenvolveu uma ferramenta para monitoração e análise de pacotes baseada em múltiplas visões (duas), sendo uma bidimensional e outro tridimensional, analisando o tráfego na rede onde os cubos são representações gráficas de sub-redes (Figura 23).

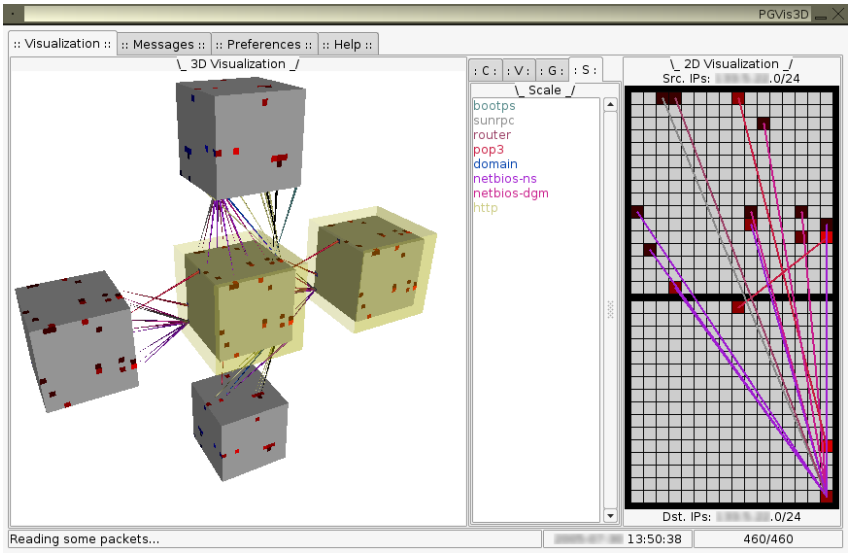

Figura 23: Visualização de trafego de rede (Malécot , 2006)

Takada e Koike (2006) desenvolveram ferramenta de visualização de logs denominada TUDUMI (Figura 24) para detecção de anomalias no tráfego através de visualização tridimensional, sumarizando log de dados.

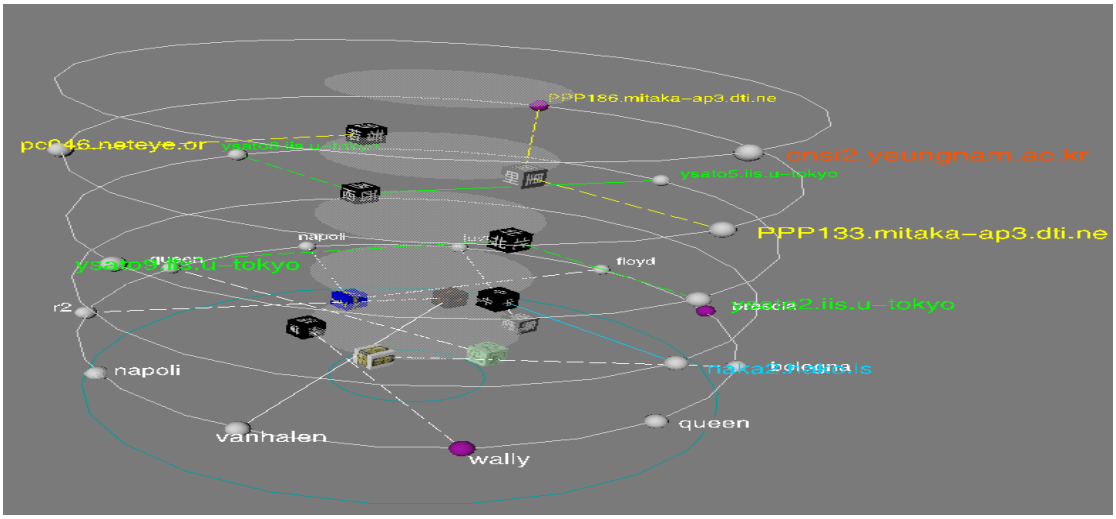

Figura 24: Visualização no TUDUMI (Takada e Koike,2006)

4 PRISMA

A ferramenta de visualização de informação PRISMA foi utilizada como ferramenta base para o desenvolvimento desta dissertação, a contribuição no desenvolvimento da mesma se deu no melhoramento de maneira geral da arquitetura, no melhoramento da técnica de coordenadas paralelas, e adição de umanova técnica de visualização denominada HeatMap. Assim, as principais técnicas de visualização de informação que compõe a ferramenta PRISMA são: treemap, coordenadas paralelas, dispersão de dados e heatmap, além de outros gráficos auxiliares.

PRISMA (Godinho, 2007) (Figura25) é uma ferramenta de visualização de informação com múltiplas visões coordenadas. A ferramenta foi desenvolvida totalmente em Java, dando total portabilidade entre plataformas. Além disso, também possui recursos importantes para uma boa ferramenta de VI, como seleção, filtros dinâmicos, zoom, configurações dos atributos, gráficos estatísticos, relatórios customizados, acesso a diversas fontes de dados, entre outros.

Os principais pontos são:

- Ser genérico para atender os requisitos de qualquer base de dados e técnica de visualização;
- Desenvolver uma interface com boa usabilidade para a execução das tarefas do usuário;
- Permitir a visualização e interação com dados em mais de uma técnica de visualização simultaneamente ou individualmente;
- Desenvolver mecanismos de coordenação entre as técnicas de visualização de informação implementadas, e também com os mecanismos de interação com os dados;

- Permitir a portabilidade para diversas plataformas e extensibilidade de software, desenvolvendo-o utilizando a linguagem Java e as técnicas de orientação objeto;
- Permitir diversas fontes de dados de entrada, desde arquivos textos pré-formatados e estruturas XML a banco de dados relacionais.

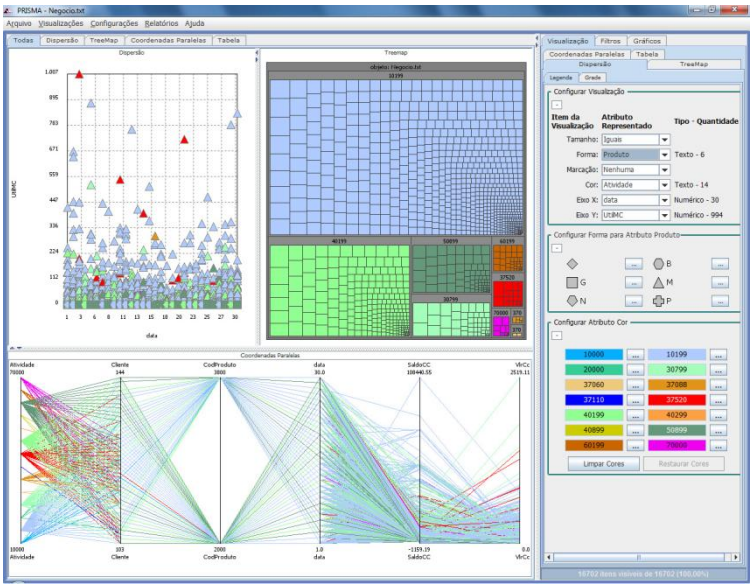

Figura25: PRISMA

4.1 Arquitetura

A arquitetura do PRISMA foi projetada baseada no modelo MVC, apresentando uma distribuição bem definida de responsabilidades entre seus módulos. A arquitetura é composta por três módulos principais: Núcleo, ModuloVis e Apresentação (Figura 26: Principais módulos da arquitetura do PRISMA), sendo cada um destes subdividido em módulos especializados. Núcleo e ModuloVis possuem estruturas de controle e modelos estruturais de suas funcionalidades particulares, enquanto os módulos da camada de Apresentação compõem-se principalmente por interfaces e mecanismos de interação com usuário.

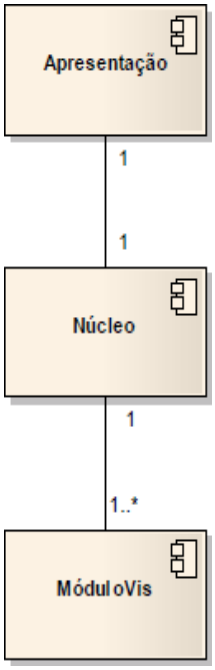

Figura 26: Principais módulos da arquitetura do PRISMA

4.1.1 Núcleo

O Núcleo é o módulo central da aplicação bem como o ponto de entrada da mesma (Figura 27). Ele mantém a responsabilidade de ligação com a camada de apresentação, geração e manutenção do ModuloVis. Entre algumas de suas funcionalidades pode-se listar:

- Inicializa o carregamento da base de dados;
- Inicialização da camada de apresentação;
- Execução de determinadas chamadas provenientes da interface com o usuário, como abrir nova base de dados;
- Entrelaçamento entre os componentes gráficos do módulo de Apresentação com os modelos contidos no ModuloVis;

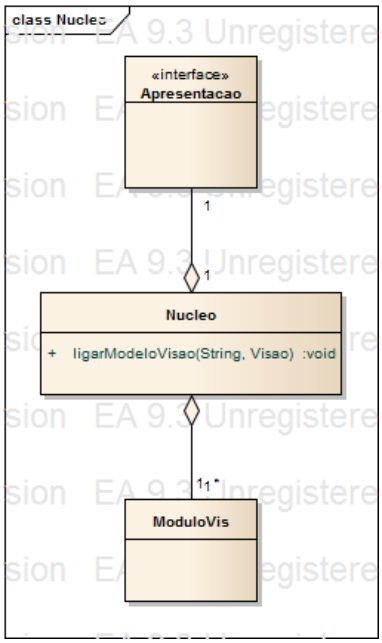

Figura 27: Módulo Núcleo do PRISMA

4.1.2 ModuloVis

O ModuloVis mantém as informações de configuração e gera a imagem das visualizações a partir destas configurações. Neste módulo estão os controles de filtro de dados, de cores e de detalhes que atuam sobre as visualizações. Dessa forma é também o ModuloVis o responsável pela coordenação dos elementos visuais entre as visualizações, notificando as visualizações quando um Controle realiza uma alteração.

A Figura 28 apresenta a arquitetura interna do ModuloVis.

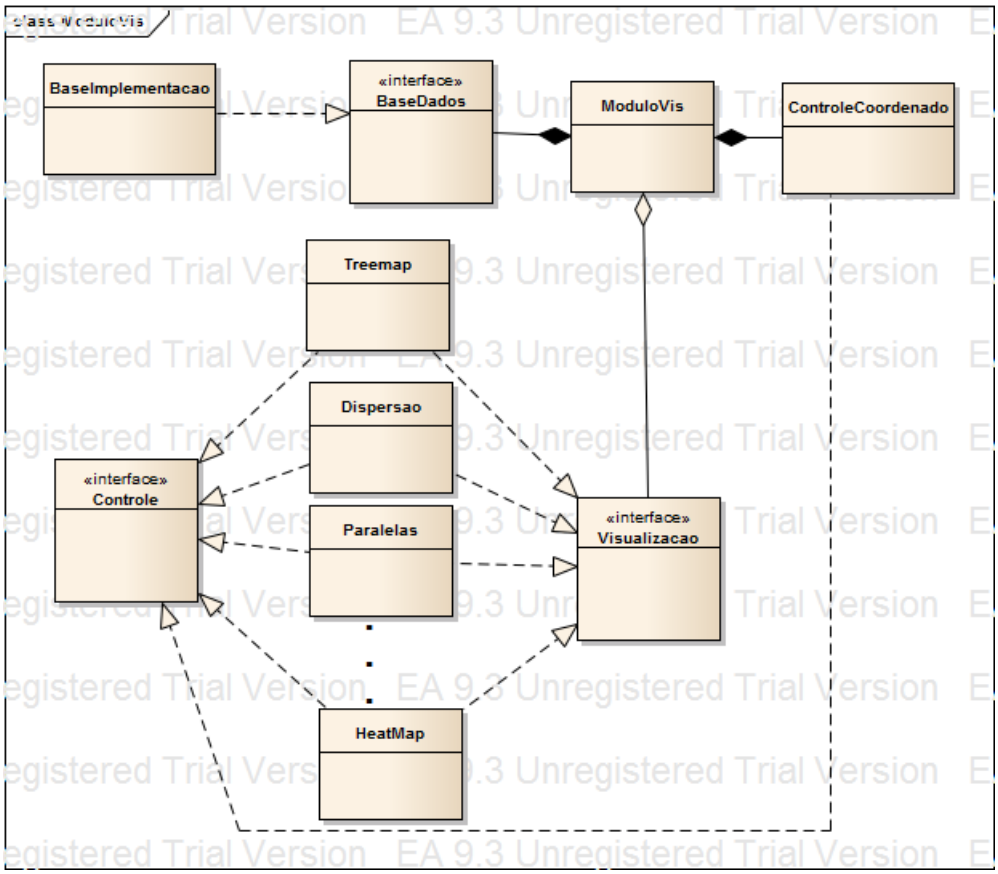

Figura 28: Arquitetura interna do ModuloVis

A seguir serão apresentadas uma breve descrição dos principais componentes do ModulosVis:

- **BaseImplementação** é um componente responsável pela carga, acesso e gerencia dos dados que pode ser realizada de diversas maneiras, de acordo com a fonte dos dados.A única restrição é obedecer as regras impostas pela interface de comunicação com o MóduloVis. Isto possibilita implementações específicas e otimizadas para a estrutura em que os dados se encontram.
- **ControleCoordenado** é responsável por gerenciar interações que se reflitam em todas as visualizações. Dentre estas estão:
    - **Gerenciador de Filtros**: que tem o propósito de auxiliar na redução dos registros visualizados, visando focar a análise do usuário em um subconjunto de dados mais restrito. Estes filtros podem ser: contínuos com múltiplos intervalos, para atributos numéricos; categórico de múltipla seleção, para atributos categóricos com pequeno número de valores possíveis, é composto de uma lista com os valores selecionados,

e categóricos ordinais, para atributos categóricos com grande número de valores possíveis, que possam ser ordenados por algum critério e limitados por duas letras ou palavras que servirão de limiares de seleção.

- **Gerenciador de Cores**: para gerenciar a coloração exibida nas visualizações utiliza-se uma abordagem similar a utilizada no gerenciador de filtros. De acordo com o tipo e características do atributo da base de dados um determinado modelo é utilizado. São dois possíveis casos: coloração em degrade, para atributos numéricos, é composta de uma cor para o menor valor do atributo e uma cor para o maior valor do atributo. Com essas duas cores de referências são obtidas as cores para os valores intermediários do atributo, e coloração por valor: para atributos categóricos com pequeno número de valores possíveis, é composto de um mapeamento valor-cor que contempla todos os possíveis valores do atributo.
- **Gerenciador de Detalhes sob Demanda**: este gerenciador guarda a lista de atributos que deve ser exibida quando a representação gráfica de um registro ou grupo de registros da base de dados sofre uma determinada interação ditada pela técnica de visualização, a interação mais comum é a ação de posicionar o mouse sobre a representação gráfica do registro. O modelo deste gerenciador é disponibilizado as técnicas de visualização por intermédio do MóduloVis durante a etapa de desenho da técnica de visualização.

- **Técnicas de Visualização de Informação**: as técnicas de visualização de informação têm com objetivo gerar representações gráficas e interativas de determinados contextos. Estes contextos são compostos por:
  - registros da base de dados que satisfazem as condições de filtragem;
  - mapeamento de cores para os registros;
  - atributos que devem ser exibidos em detalhes sob demanda;
  - configurações específicas da visualização.

Neste trabalho utilizou-se este padrão para desenvolver uma visualização nova na ferramenta PRISMA, a técnica do HeatMap. Devido à arquitetura base da ferramenta, não foi necessário tratar um conjunto comum de características, a exemplo do

tratamento da base de dados. Uma vez que a lógica interna do HeatMap está encapsulada na visualização e integrada ao PRISMA através de suas interfaces, é possível reutilizar o HeatMap em outras ferramentas de Visualização em Java.

4.1.3 Apresentação

O módulo Apresentação (Figura 29) cria e gerencia o grupo de componentes de interface gráfica de usuário. Estes componentes são criados de acordo com as características a base dados de análise e podem ser adicionados com a necessidade do usuário. Apresentação tem integração com o Núcleo, sempre que a Apresentação cria uma nova visão (seja esta um painel de configuração ou um painel de desenho) é feita uma requisição ao Núcleo para entrelaçar a visão recém-criada com seu respectivo módulo. Esta abordagem foi escolhida visando trabalhos futuros com interfaces web, em que o módulo de Apresentação poderá se tornar um WebService provendo informações e um caminho de comunicação para qualquer plataforma compatível.

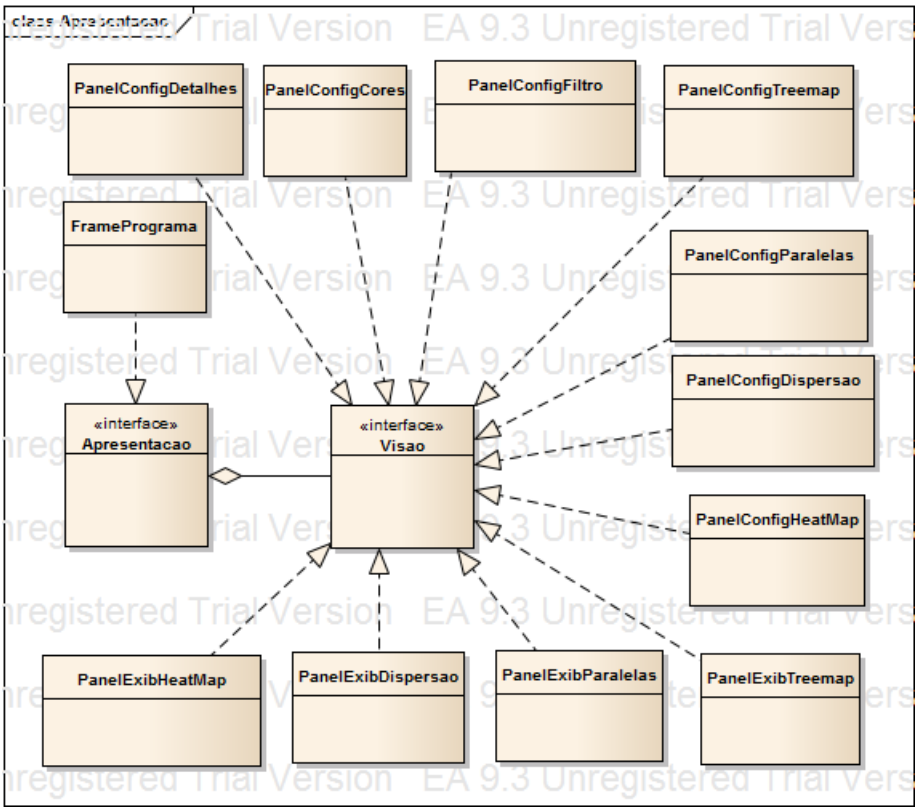

Figura 29: Componentes internos da Apresentação

4.2 Aspectos de Implementação de Coordenadas Paralelas

A técnica é composta por eixos verticais que representam atributos da base de dados, estes eixos são dispostos paralelamente em um plano, além disto, têm-se segmentos de retas entre os eixos que representam os valores dos registros analisados.

4.2.1 Processamento e renderização

Neste trabalho, o processo de criação de uma representação gráfica da técnica de Coordenadas Paralelas ocorre em dois momentos distintos.

A primeira etapa consiste em analisar as configurações e interações realizadas pelo usuário e utilizar esta informação para recalcular as propriedades de desenho da visualização (posicionamento e escala de: eixos, segmentos de linha, rótulos e elementos de interação). Uma das grandes preocupações desta etapa é evitar recálculos desnecessários, principalmente pela grandeza de processamento estar ligada diretamente ao número de eixos visualizados. Para cada novo eixo adicionado na visualização, são criados n novos segmentos de linha em que n é o número de registros selecionados para exibição.

As prevenções de reprocessamento são:

1. Para as operações de configurar a escala de um eixo (alterar mínimo, alterar máximo, inversão), apenas são recalculados os segmentos de reta que estão ligados a este eixo;
2. Para as interações de filtro e seleção, são apenas recalculados os segmentos de reta para os registros que não participavam da visualização e como resultado do filtro passaram a ser exibidos;
3. Para a operação de reordenação de eixos, apenas são recalculados os segmentos de reta afetados pela reordenação;
4. Para ajustes de cor os segmentos de reta não sofrem recalculo;
5. Detalhes sob demanda não implicam em recalculo.

Para o segundo momento, três buffers de desenho que são utilizados como camadas da representação gráfica final. Estas camadas utilizam transparência e para exibição no buffer de vídeo são mescladas em uma única saída. Estas camadas são:

1. Fundo: camada de desenho principal em que a visualização é renderizada, esta deve ser preservada sempre que possível;
2. Brushing: camada para renderização de registros afetados pela operação de brushing;
3. Interação: camada para renderização de artifícios voláteis, como uma borda de destaque para elementos sob omouse, e os detalhes sob demanda.

Este conceito é utilizado para evitar reprocessamento gráfico, já que, a divisão em camadas faz com que os artifícios de interação e brushing não danifiquem a camada de fundo, ou seja, em sucessíveis operações de detalhes sob demanda somente a camada de interação sofrerá processamento, as outras camadas são apenas copiadas para a saída.

4.2.2 Otimizações de interação

Como pré-requisito de funcionamento a técnica de coordenadas paralelas necessita de pelo menos dois eixos selecionados para exibição. Com dois eixos é desenhado um segmento de reta para cada registro, com três eixos são desenhados dois segmentos de reta por registro, e assim quantos mais dimensões são disponibilizadas para análise, maior o grau de complexidade computacional para gerar uma visão.

Para amenizar o impacto de performance e diminuir o tempo de resposta em interações com o mouse na visão, primeiro é feita a classificação da posição do mesmo em relação a cena. A classificação é dada em zonas entre dois eixos, com isso a busca por qual segmento de reta está sob o mouse é realizada apenas nos segmentos de reta da zona de classificação. Este conceito também é utilizado nas interações de Brushing e Seleção, nestes casos é permitida a classificação de múltiplas zonas por tratar-se de uma área de seleção e não de um ponto sob o mouse.

4.3 Aspectos de Implementação do HeatMap

Como visto nas seções anteriores, HeatMap é uma técnica de visualização da informação, que está organizado na forma de matriz, onde cada coordenada da matriz

(quadrado ou retângulo) representa um valor contínuo ou discreto codificado por uma coloração.

Muitas das estratégias utilizadas para concepção da técnica de coordenadas paralelas foram utilizadas na concepção da técnica de HeatMap, tais como:

4.3.1 Processamento e renderização

O redesenho da técnica HeatMap, que é dependente da interação do usuário, esta concentrada na alteração dos eixos que forma a matriz, e adicionalmente a coloração dos eixos de acordo com seleção do atributo feita pelo usuário. A complexidade e reprocessamento são menores quanto comparados à técnica de coordenadas paralelas, e estão ligadas diretamente quantidade de valores nos eixos.

As prevenções de reprocessamento são:

- Para as operações de configurar a escala de um eixo (alterar mínimo, alterar máximo, inversão), apenas são recalculados os segmentos de reta que estão ligados a este eixo;
- Para as interações de filtro e seleção são apenas recalculadas a coloração dos itens da matriz para os registros que não participavam da visualização e como resultado do filtro passaram a ser exibidos.
- Para ajustes de cor, a matriz não sofre recalculo.
- Detalhes sob demanda não implicam em recalculo.

Da mesma forma que a técnica de Coordenadas paralelas, três buffers de desenhosão utilizados como camadas da representação gráfica fina utilizando transparência, e mescladas em uma única saída no buffer de vídeo, são elas: Fundo, Brushing e Interação.

4.3.2 Otimizações de interação

Como pré-requisito de funcionamento a técnica Heatmap necessita de pelo menos dois atributos para serem representados nos eixos da matriz, e outro atributo para coloração das células ou itens da matriz.

Para amenizar o impacto de performance e diminuir o tempo de resposta em interações com o mouse na visão, primeiro é feita a classificação da sua posição em relação a cena. A classificação é dada por células. Este conceito também é utilizado nas interações de Brushing e Seleção, nestes casos é permitida a classificação de múltiplas zonas por tratar-se de uma área de seleção e não de um ponto sob o mouse.

4.3.3 Outras Considerações

Outras considerações sobre a técnica HeatMap, são:

- Células

  As célulasda matriz da técnica HeatMap representam um conjunto de dados do mesmo tipo. Neste caso é feito a soma das ocorrências de uma coluna da base de dados, que contenham os mesmos eixos. No controle (interface) é possível personalizar qual coluna do banco de dados pode ser somada.

- Cores

  A cor que cada célula apresenta depende do calculo dos dados para redefinição da célula, e a faixa de valor que foi pré-definida para classificação. Existem três classes de dados pré-definidas. Classe alta, onde os dados estão entre 75% e 100% do valor mais alto da base de dados. Classe média, onde os dados que estão entre 25% e 75% do maior valor. Classe baixa, os dados estão entre o menor valor da base de dados e até 25% do maior valor. Esses valores são representados por uma cor fixa, mas com transparência entre 10% e 100%. No controle é possível fazer a personalização dessas cores.

- Controles Adicionais

  Em relação a controle adicionais, foram implementados controle para exibição dos rótulos em todas as células, para personalização dos eixos e para os dados nulos na base.

4.4 Protótipo e Funcionalidades

4.4.1 Aspectos Gerais da Interface

A interface gráfica de interação da ferramenta (Figura 30) é composta de três áreas gerais de interação:

1. Menu de aplicação: utilizado para carregar uma nova base de dados, salvar/restaurar configurações, exibir tópicos de ajuda, emitir relatórios e sair da aplicação;
2. Organizador de visões: exibe e organiza as visões gráficas configuradas;
3. Controles: conjunto de controles relacionados a base de dados e técnicas de visualização carregadas, mais especificamente o lugar em que será configurado: filtro, detalhes sob demanda, cores, coordenação as técnicas de visualização (cada técnica com seu controle próprio).

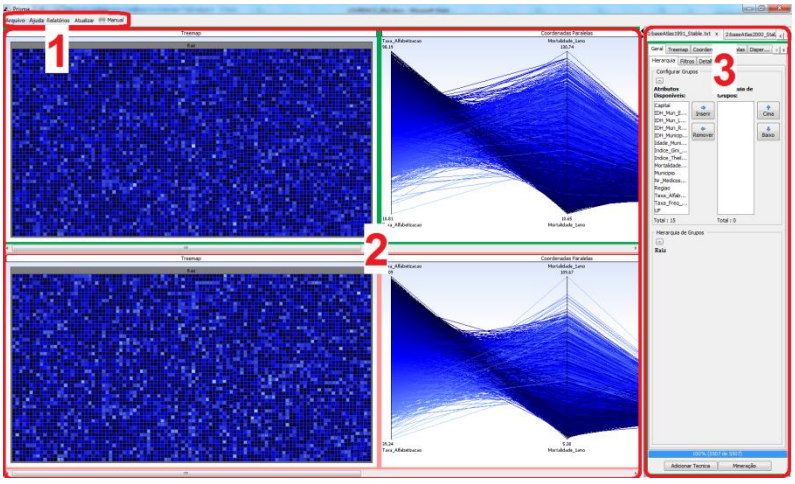

Figura30: Interface principal PRISMA

4.4.2 Controles de Configuração

O controle de configuração está dispostoem abas (Figura 31) e em dois grandes grupos:

- Geral, que refletem em todas as visualizações;
- Específico por técnica de visualização.

No primeiro,configuram-se todas as características que estão coordenadas, por exemplo, o atributo visual COR, ou mecanismo de filtro. Nas abas específicas de técnicas de visualização estão as configurações das mesmas, por exemplo, para a técnica de coordenadas paralelas tem-se (Figura 32): a inserção e remoção dos eixos de análise, a ordenação dos valores dos eixos, ou a definição da posição do eixo na sequencia dos eixos visíveis.

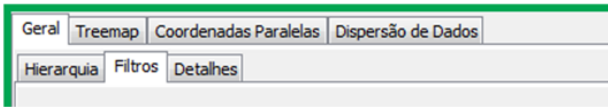

Figura 31: Abas de Configurações

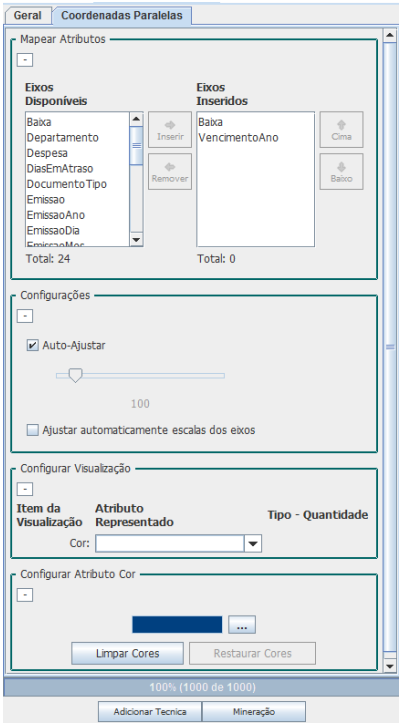

Figura 32: Configuração Coordenadas Paralelas

4.4.3 Configuração das Cores

Para configuração de cores da aplicação utiliza-se tanto atributo categórico quanto contínuo. Cada um destes tipos implica em uma abordagem diferente na atribuição de cor aos dados.

- Cores Discretas ou Categóricas: para os atributos categóricos é realizado um mapeamento de cada possível valor do atributo com uma cor equivalente (Figura 33).
- Cores Contínuas: em atributos contínuos utiliza-se o mapeamento de intervalos de cores. Na implementação três valores são mapeados, o limite inferior do atributo, o ponto médio e o limite superior, a partir desse mapeamento pode-se inferir as cores paras os valores intermediários (Figura 34).

Figura 33: Configuração cor atributo categórico

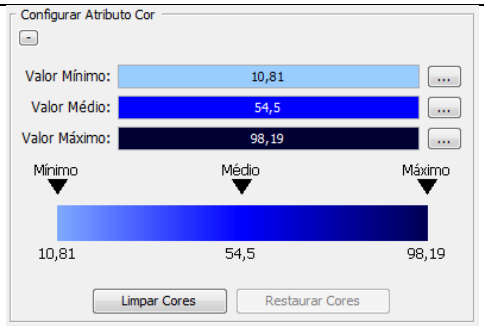

Figura 34: Configuração cor atributo contínuo

4.4.4 Filtros

Há necessidade de selecionar o atributo para interagir com filtro desejado, objetivo não sobrecarregar o usuário com informações não úteis. Há dois tipos de filtros: filtros categóricos e filtros contínuos.

- Filtros Categóricos: Para atributos categóricos os controles são compostos de caixas de seleção, informando se ele está ativo ou não da visualização(Figura35).
- Filtro contínuo: para atributos contínuos utiliza-se um controle de seleção múltipla de intervalos. Há um gráfico de barras, apresentando a distribuição dos valores, e as partes do gráfico em amarelas estão visíveis (Figura36).

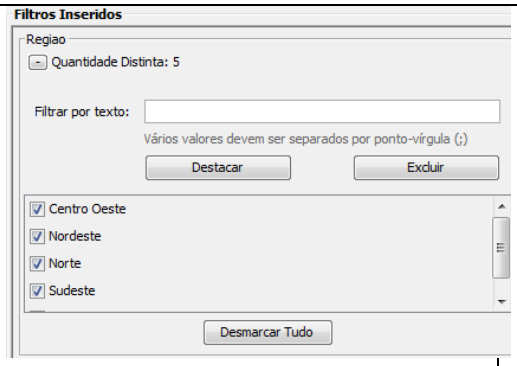

Figura35: FiltroDiscreto

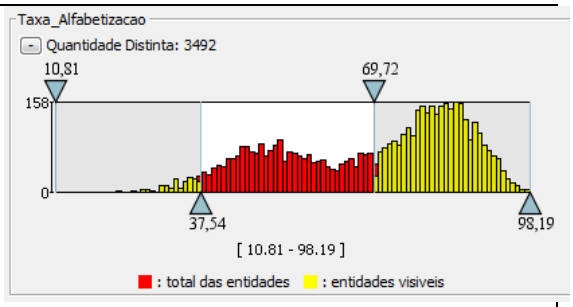

Figura36: Filtrocontínuo

4.4.5 Seleção e Brushing

As ações de seleção e brushing (Figura 37) não dispõem de um painel de controle lateral. A Seleção permite gerar ações para isolar ou remover itens de uma visão. Brushing permite, através do destaque, correlacionar duas visões de dados.

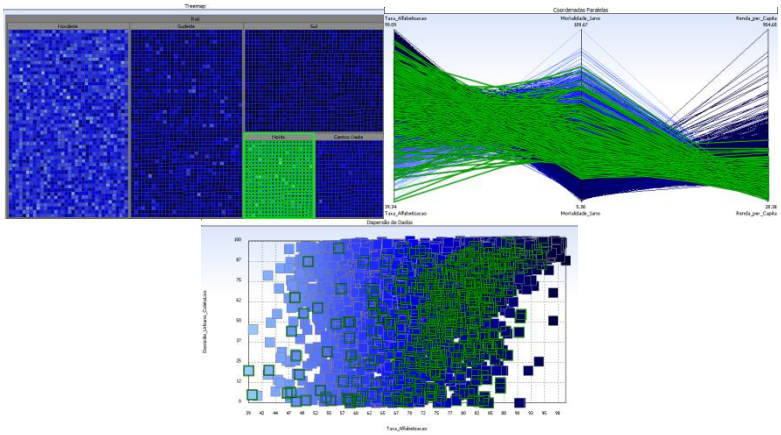

Figura 37: Exemplo de brushing em diversas técnicas de visualização de informação

4.4.6 Coordenação Entre Visões

A coordenação entre visões de uma mesma base de dados (mesmo MóduloVis) está presente e ocorre de forma automática ao configurar-se cor, filtro, seleção, brushing e a hierarquia dos dados.

5 ESTUDO DE CASO

A análise de transações eletrônicas e monitoramento de eventos de segurança da informação demandam a análise e correlação de diversos conjuntos de dados, sendo esta uma tarefa não trivial, principalmente pela quantidade diária de informações acumuladas eletronicamente.A ferramenta PRISMA foi utilizada neste trabalho pelas seguintes motivações:

- Disponibiliza múltiplas técnicas de visualização da informação, isso possibilita ao usuário criar várias visões dos mesmos dados em representações diferentes buscando um melhor entendimento;
- Permite trabalhar com várias visões simultâneas, possibilitando várias visões do mesmo dado, ou visões que se complementem objetivando analisar simultaneamente mais dados e correlaciona-los;
- Foi disponibilizado o código fonte para estudo, e isso permitiu amelhoria do módulo de Coordenadas Paralelas e desenvolvimento do módulo Heatmap.

A análise desses dados seguiu a metodologia abaixo:

- Levantamento e avaliação de que dados poderiam ser relevantes para as análises em questão, e que fontes poderiam fornecer esses dados. Essas perguntas foram feitas junto a especialistas;
- Definição conceitual dos cenários de análise;
- Pré-processamento dos dados coletados;
- Concepção das visualizações e configurações na ferramenta PRISMA;
- Avaliação junto a especialistas.

Os cenários para análise foram concebidos primeiramente sem a técnica Heatmap na ferramenta PRISMA. A motivação para incorporação da técnica Heatmap ao PRISMA se deu após as entrevistas com especialistas, que apontaram da necessidade de análise temporal e a fragilidade do PRISMA em relação a esse tipo de análise.

5.1 Pré-Processamento

O processo de pré-processamento é utilizado para reduzir a quantidade de linhas de um arquivo sem perder a informação, para conjunto de dados que possua atributos temporais de datas e horas.

A motivação do pré-processamento é o grande volumes de dados que afetam a performance, e as vezes até a impossibilidade, de ferramentas codificarem os dados e visualizá-los através de técnicas de visualização.

A Figura 38: Potenciais dados de análise em sistemas e monitoramentoapresenta potenciais dados encontrados nos principais sistemas de monitoramento.

Event Type	Data Source	Devices
Network Traces	–Raw Packets –Netflow Records	Tcpdump, Tshark Cisco NefFlow NDE, Cisco NSEL Netflow
Security Events	–Intrusion Detection Systems –Intrusion Prevention Systems –Firewalls –Virtual Private Networks –Anti-virus	Cisco CSA, Cisco IDS, Enterasys Dragon, Fortinet Fortigate, Juniper ISG, SNORT, Niksun NetVCR, SourceFire Intrusion Sensor ForeScout ConterACT, Juniper NetScreen IDP, McAfee Intrushield, Radware Defense Pro, FireEye, Tipping Point X, IPAngel Check Point, Linux Iptables, PaloAlto PA, Cisco ACE ArraySP, Nortel VPN Gateway, Checkpoint VPN-1, Cisco ASA Mcafee, Sophos, Symantec, Trend Micro
Network Activity Context	–Layer 7 application context	Q1 Labs QFlow, Foundry SFlow, Juniper JFlow, Packeteer FDR
User/Asset Context	–Vulnerability Scanners –Identity and Access Management	NMap, eEye REM, Nessus, Rapid7 NeXpose, SecureScout nCircle IP360, Patchlink Scan, Qualys, Saint Microsoft ForeFront Identity Manager, Identity Forge Quest Identity Manager One, EmpowerID
Network Events	–Switches –Routers –Servers –Hosts	Cisco CatOS, Cisco Catalyst, 3Com 8800 Series Cisco Routers, Enterasys Router, Juniper Router, Nortel Router Apache, BlueCoat SG, Cisco Ironport, IIS, Sun Sendmail Windows, Linux, Solaris, IBM AIX RACF, HP Tandem
Application Logs	–Application Databases –Workflow –Enterprise Resource Planning –Management Platforms	IBM DB2, SQL Server, Oracle, Imperva SecureSphere, Sybase

Figura 38: Potenciais dados de análise em sistemas e monitoramento

Como base de teste para criação dos cenários de eventos de segurança foi utilizado um arquivo de log do proxyweb gerado com um conjunto de dados contendo 24

atributos, sendo destes 20 strings, 1 campo tempo(time), um campo data e 3 campos numéricos.

O campo temporal(time hh:mm:ss:sss) foi decodificado e somente o campo hora, minuto e segundo foram aproveitados, devido a característica da base. Após, o conjunto de dados foi ordenado pelos campos strings e pelo campo hora. Um campo que registrava um registro de url do tipo www.terra.com.br/principal/index.html foi descartado devido a existência de outro campo de url contendo a informação apenas da parte da url principal www.terra.com.br. O agrupamento foi realizado pelos campos strings e hora, somando os registros numéricos e criando um atributo de contagem para cada linha do conjunto de dados anterior, que no caso denominamos de cont_registro. Com isto há a redução significativa sem perder a informação de quantidade de linhas e sem perder a característica do conjunto de dados.

O arquivo sem pré-processamento levava aproximadamente 30 minutos para carregar na ferramenta, sem contar com a lentidão a cada iteração que o usuário realizava.

O mesmo arquivo após o pré-processamento ficou com aproximadamente 6(seis) vezes menos registros e levava aproximadamente 1 minuto para carregar, e não foi detectado lentidão na iteração com a ferramenta.

O conjunto de dados selecionado com informações de transações eletrônicas possuía data, tempo, produto, status da transação (sucesso, negada, retornada ou em espera).

5.2 Análise de Transações Eletrônicas

Um número crescente de empresas fornece aos seus clientes um canal de relacionamento através da Internet, oferecendo os mesmos serviços ou produtos encontrados nas lojas tradicionais.

A análise dessas transações eletrônicas ajuda a entender a relação entre os clientes e a empresa. Transação do cliente pode ser dividida nas seguintes categorias: concluído com êxito; negado devido a uma regra de negócio; em espera; cancelado por falha no sistema, e desfeitas (automaticamente ou manualmente).

A vantagem competitiva pode ser alcançada pela análise adequada das transações que foram recusados por causa de uma regra de negócio. O entendimento do comportamento do cliente ajuda a medir a satisfação do cliente com um produto ou da própria empresa.

A transação negada do cliente pode representar uma ocorrência simples e típica relacionado a principal característica do negócio da empresa, como uma senha inválida ou permissões insuficiente para realizar uma transação. Por outro lado, a mesma transação pode indicar uma oportunidade para melhorar um produto, ou criar um novo produto novo mais apropriado para um grupo específico de clientes, partindo da hipótese de identificar clientes que tentam utilizar um produto, mas não o fazem por não possuir perfis adequados.

Por exemplo, se o número médio de transações negadas está entre 15% e 25%, e ele aumenta consideravelmente em um mês específico do ano, isto pode indicar que os clientes não tenham a informação apropriada sobre o produto, ou que a interface eletrônica é de difícil interação, ou que um produto similar deve ser fornecido para o grupo de clientes interessados.

O objetivo deste estudo é analisar os problemas que podem estar ocorrendo com transações eletrônicas e melhorá-las ou, em alguns casos, determinar a necessidade de criar novos produtos e serviços.

A análise do status de transações (O - sucesso, E - negado, C - cancelado, 0 - desfeita, e P - pendente), em um primeiro momento, foiiniciadana categorização por cores, e cada status pode ser vista nas múltiplas técnicas disponíveis no PRISMA (Figura 39). A visão Treemap foi organizada pelo status da transação e código de erro. Na visão de dispersão de dados foram configurados a quantidade e o montante solicitado da transação. A visão de coordenadas paralelasincluiu o valor da transação eo número de parcelas. A visão do gráfico do Treemapapontou que 74% das transações foram bem sucedidos, 23% foram negados, 2% foram cancelados por falha de sistema ou de tempo limite e os outros somaram 1%. Estes eram esperados percentis confirmados pela análise.

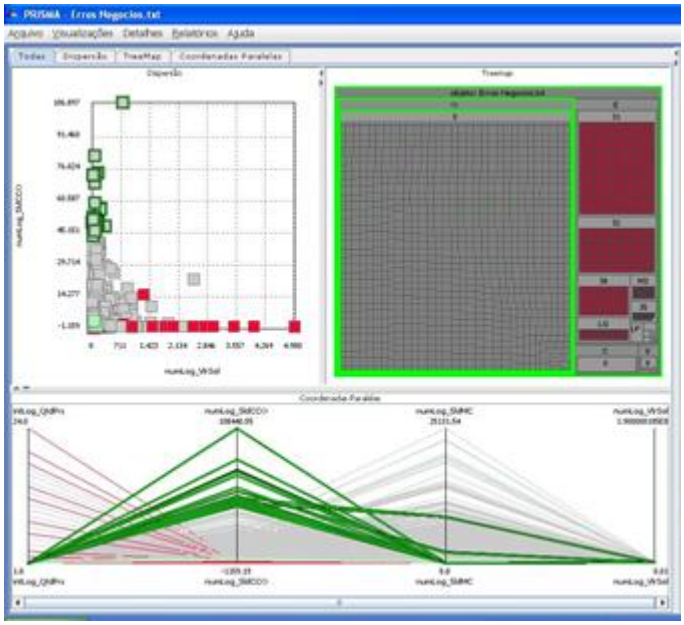

Figura 39: Cor e brushing relacionadosas várias visões do PRISMA

Após a identificação da esperada distribuição dostatus da transação, outros cenários foram explorados. Primeiro, uma seleção de filtro foi aplicado para apresentar somente as transações negadas e canceladas. Na visão Treemap o código de erro principal foi relacionado à quantidade insuficiente de valorespara o cliente. Esta percepção é confirmada na visão de dispersão destacada em vermelho (Figura 40). Um padrão inesperado, no entanto, indicou um número de transações negadas para os clientes que poderiam ser classificados como perfis para usar determinado serviço ou produto(destaque em azul). A seleção dessas ocorrências chamou a atenção para a mensagem de erro onde clientes estavam tentando realizar transações de produtos indisponíveis para suas categorias de clientes. Isso foi interpretado como uma oportunidade de negócio dirigido ao grupo identificado de clientes.

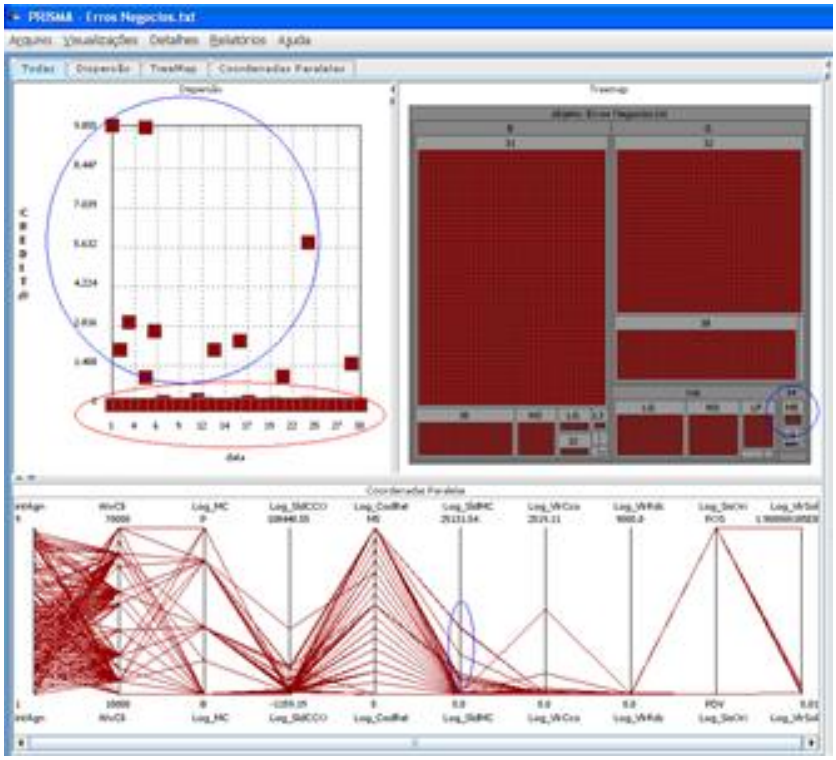

Figura 40: Transações com erros e negadas

Em um cenário diferente, as transações canceladas foram adicionadas à visualização (Figura 41). Foi então possível perceber que essas transações, embora em pequeno número, são significativas quando consideradas o valor total. Estes números podem indicarpossíveis necessidades de melhoramento da infraestrutura de TI. Além disso, o montante total disponível para transações com estes clientes representam 12,06% das transações canceladas (Figura 41) e 2,48% das transações negadas (Figura 42). O montante total indisponível para estes clientes chega a 14,54%.

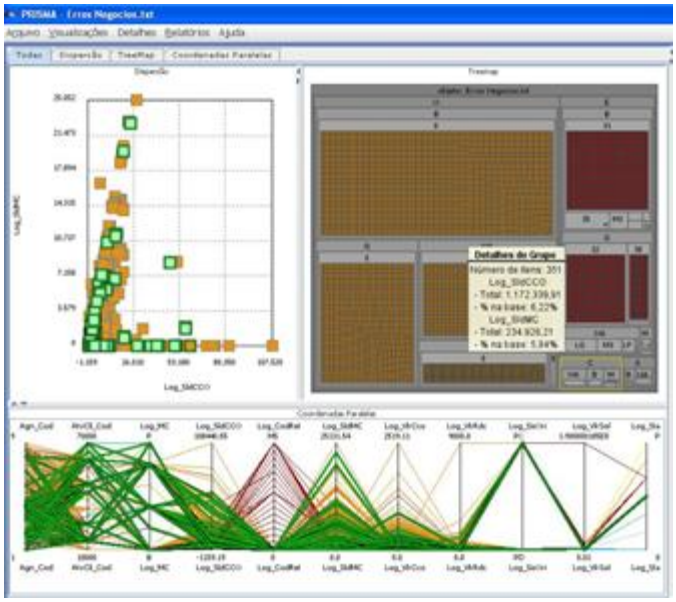

Figura 41: Cenário adicionado com transações canceladas

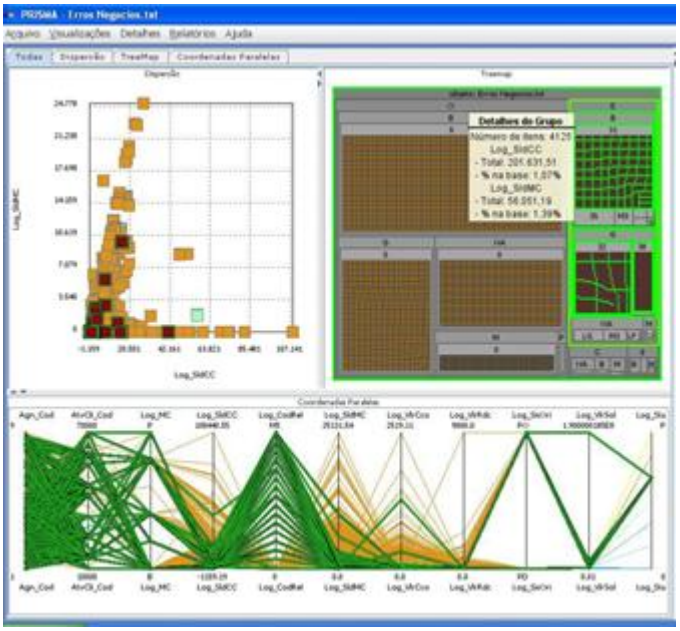

Figura 42: Seleção de transações negadas por regra de negócio

A visualização foi então rearranjada para apresentar o Treemap agrupado por produto e status. O atributo status foi filtrado para mostrar apenas as operações bem sucedidas e negadas. Um padrão interessante mostra que o produto codificado como M está associado a 7,86% das transações e quase nenhuma rejeição. Ao analisar o gráfico de dispersão e coordenadas paralelas é possível identificar o potencial dos clientes associados,sendo o produto em questão, um candidato de destaque para uma maior promoção da empresa.

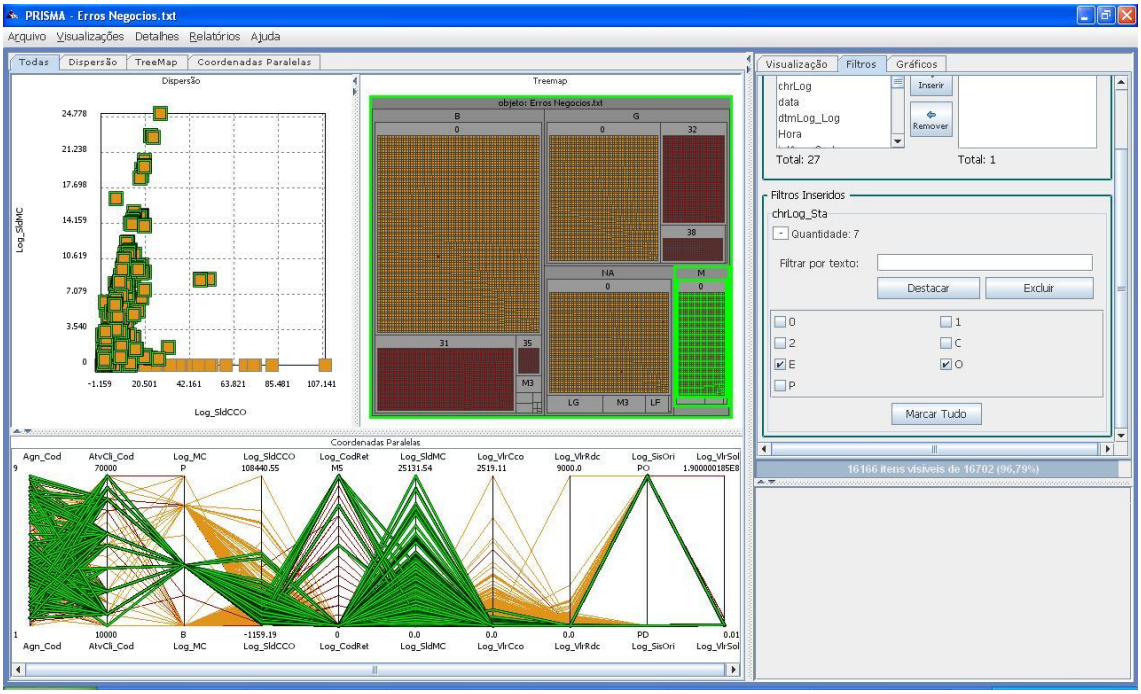

Figura 43: Percepção do Produto

5.3 Análise de Logs de Rede

Empresas vêm adotando melhorias nas práticas de segurança para tornar o uso dos recursos de informação de modo mais seguro. Apesar dos grandes investimentos em equipamentos de segurança e softwares um ponto importante é o monitoramento e análise de tráfego de rede comomecanismos para restringir ações que podem tornar indisponíveis os serviços prestados a clientes externos. A ferramenta PRISMA foi usada para ajudar a análise de situações em que um grande volume de dados, utilizando múltiplas visões coordenadas para uma melhor percepção dos dados e suas relações com o ambiente interno das transações. Com a avaliação do ambiente interno é possível identificar melhorias e soluções para manter as transações externas disponíveis para os clientes.

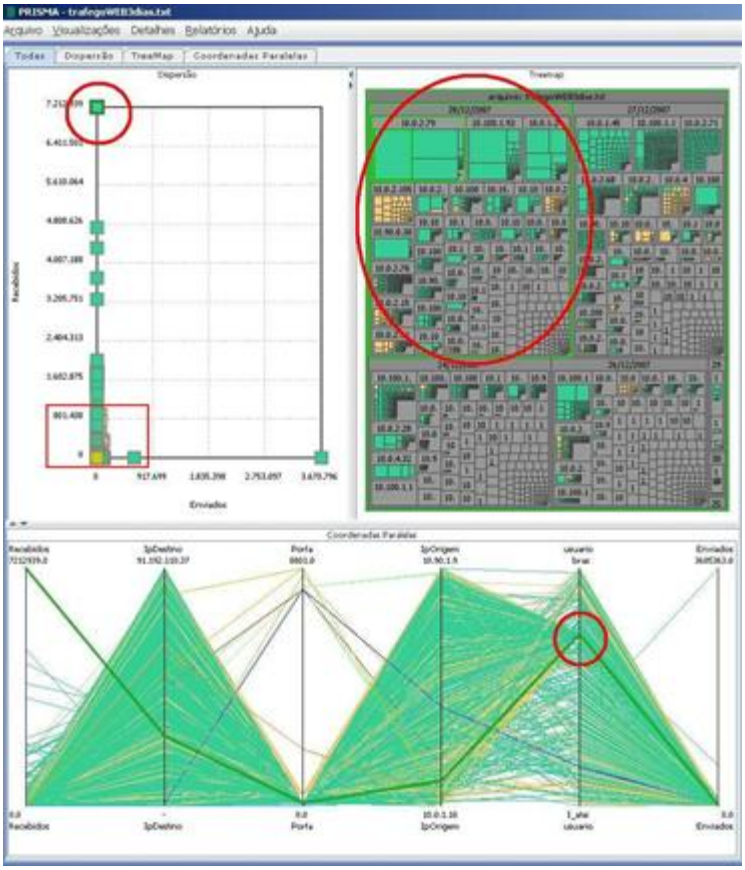

Figura 44: Configuração inicial do PRISMA

No primeiro momento, um registro de uma semana de eventos foi carregado. No módulo Treemap, a informação foi agrupado por dia e IP de origem. Com esse layout de informações, foi rápido descobrir que algumas máquinas tinham comportamento diferenciado, com grande número de arquivos recebidos. A Figura 44 mostra a utilização da técnica de Brushing, a qual destaca os mesmos dados em três técnicas, permitindo a confirmação de uma hipótese de utilização abusiva da rede de uma forma mais precisa.

A seleção de um item visual no diagrama de dispersão (Figura 44) mostra que um grande número de bytes recebidos (outlier) permitidos por meio da técnica de brushing identificar notreemap o IP associado e datas, bem como comparar os pontos, em termos de quantidades de dados.

Para os mesmos itens de dados a técnica de Coordenadas Paralelas mostra se eles estão associados a um ou muitos sites, ou a um ou muitos usuários. Neste exemplo, foi possível identificar que o mesmo usuário realizou vários downloads de endereços IP de

diferentes origens. Também é possível associar um usuário específico para um endereço de IP ou porta.

Para uma análise mais criteriosa, zoom foi aplicado para a técnica de dispersão, tal como apresentado em vermelho naFigura 44. O comportamento diferenciado na Figura 45 é identificado na técnica de coordenadas paralelas e realçado em outras técnicas. O analista de segurança então é capaz de identificar a porta e realizar uma pesquisa mais detalhada em outros registros. Neste caso, o analista de segurança identifica uma tentativa de usar aplicações P2P proibidas pela política de rede da empresa, bem como a data e IP relacionadas a esse evento. Esta informação ajuda o gerente entrar em contato com os usuários responsáveis e reforçar a política da empresa sobre o comportamento do mesmo. Outro exemplo é o agendamento de execução em horário inapropriado de atualizações permitidas, tais como anti-vírus que consomem recursos da rede.

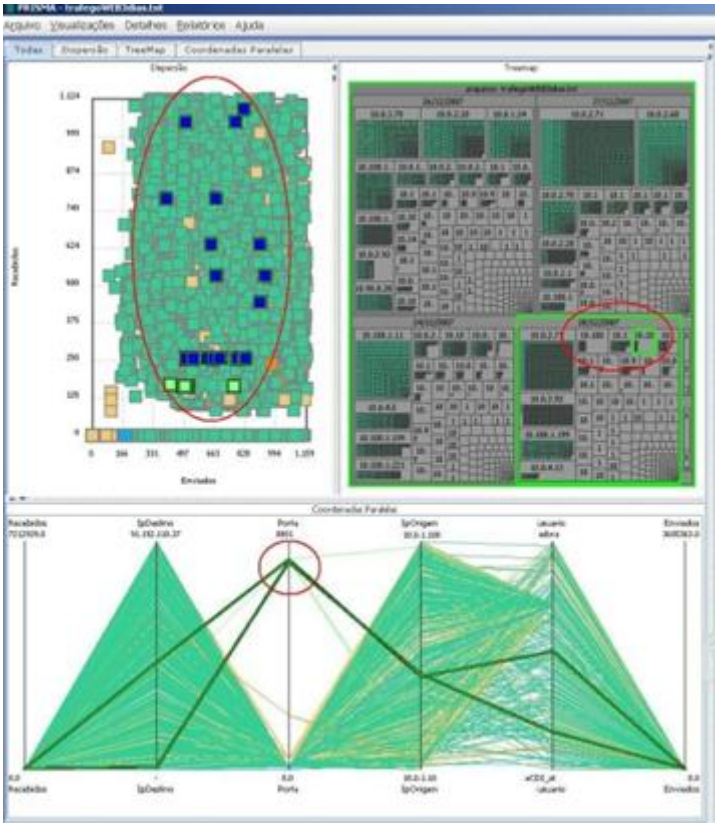

Figura 45: Aplicação da técnica de zoom no PRISMA

Em outro cenário, foi filtrado um dia específico, resultando na detecção de uso indevido de recursos de informação, tais como a utilização de sites para download de vídeos (Figura 46), degradando a rede em momentos em que há um grande número de

transações e acessos externos a sites proxy que funcionam como uma ponte de acesso a sites não permitidas sob as regras do software de proxy da empresa, facilmente detectadas através dos filtros.

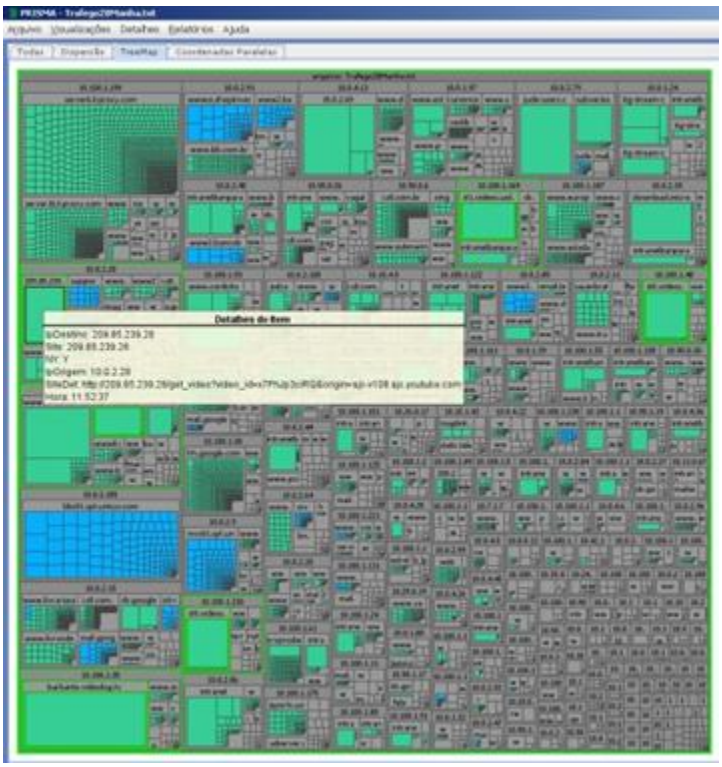

Figura 46: Tráfego de vídeo na rede

5.4 Avaliação da ferramenta PRISMA para análise delogs

Foi aplicado um questionário após a utilização da ferramenta PRISMA pelos analistas de segurança. Os analistas possuíam de 6 (seis) meses a 3 (três) anos de experiência em segurança da informação, sendo perguntado o quão foi possível identificar o comportamento dos dados temporais em relação as três técnicas (Treemap, Dispersão, Coordenadas Paralelas). Observa-se naFigura 47 que apesar dos usuários terem avaliadoa ferramenta PRISMA quanto a exploração dos dados com conceito superior a 9 (nove), foram observadas notas médias inferiores a 6 quando a mesma pergunta é em relação ao atributo temporal.

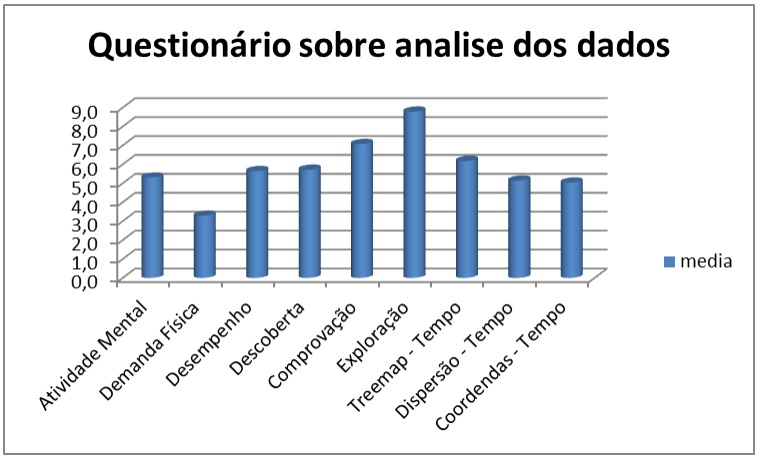

Figura 47: Avaliação PRISMA em relação análise temporal

Segue exemplo modelo de questionário aplicado.

Questionário

- Entrevistado:
- Tempo de experiência com Segurança da Informação:
- Tarefas
    - * Valores de 0 - Baixo a 10 Alto (intervalos de 0,5)
    - Identificar Padrões nos dados.
    - Identificar Exceções nos dados.
    - Identificar comportamento dos dados.
- Demanda Mental:

  Quanto de atividade mental e perceptiva foi necessário (raciocínio, decisão, cálculo, lembrança, busca, identificação)?

- Demanda Física:

  Quanto de esforço físico foi necessário (pressionamento de botões, movimentos com o mouse...) para realizar as tarefas?

- Desempenho:

O quão bem sucedido você acha que foi no alcance dos objetivos das tarefas?

- O quão foi possível identificar pelo uso em de Multiplas visões coordenadas alguma informação nova?
- O quão foi possível comprovar pelo uso das técnicas alguma informação de conhecimento prévio?
- O quão foi possível explorar os dados pelo uso das técnicas
- O quão foi possível identificar o comportamento nos dados em relação ao tempo na Técnica de Treemap?
- O quão foi possível identificar o comportamento em relação ao tempo na técnica de dispersão?
- O quão foi possível identificar o comportamento em relação ao tempo na técnica de coordenadas paralelas?
- Espaço aberto para sugestões, comentários:

Com os cenários construídos anteriormente para análise de transações e tráfego de rede foi possível identificar padrões e outliers que possibilitassem novas oportunidades de negócio ou novas propostas para política de segurança da informação. Contudo, os analista de negócio e segurança apontaram a necessidade de explicitar mais o aspecto temporal dos dados de análise.

5.5 Análise de Logs com HeatMap

Em resposta a necessidade elencada de explicitar o aspecto temporal dos dadosanalisados foi proposta uma nova técnica de visualização para PRISMA(Figura 48), o HeatMap, já apresentado nas seções anteriores. O pré-processamento também foi uma fase importante para melhorar análise temporal dos dados, e foram realizados os seguintes passos:

- Decodificar campo de tempo em ano,mês,dia,hora,minuto e segundo.

- Selecionar os atributos para os eixos e para quantificação e sumarização
- Ordenar
- Agrupar
- Retirar a média
- Montar gráfico
- Colorir as partes usando o degrade de cores de acordo com os limites estipulados
- Criar iteração temporal

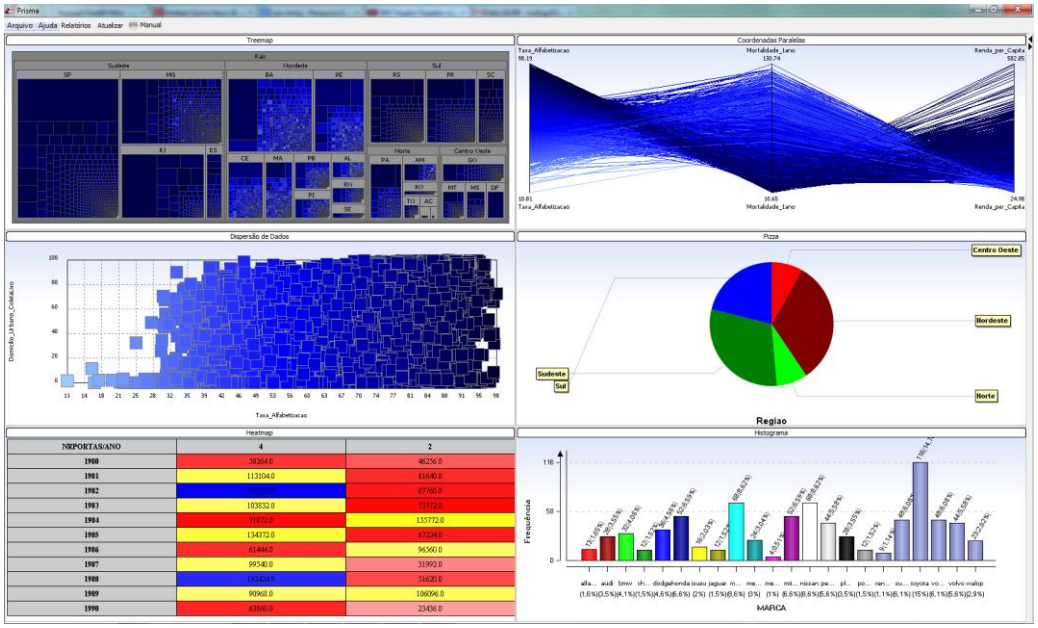

Figura 48: PRISMA com Heatmap integrado

Os cenários utilizados foram semelhantes aos já apresentados, o uso do HeatMap, de maneira geral, melhorou o tempo de percepção do mesmo padrão. Além disso, a técnica de HeatMap é bem fácil de interpretar, motivando, como passo inicial a exploração em outras técnicas.

A Figura 49 permite ao analista de segurança através do Treemap analisar o comportamento dos usuários da empresa, e para cada usuário detalhar os sites navegados, bem como o tipo de tráfego e quantidade do mesmo. O HeatMap auxilia, ou complementa, a análise detalhando a hora do dia em que um determinado evento acontece, e com que intensidade.

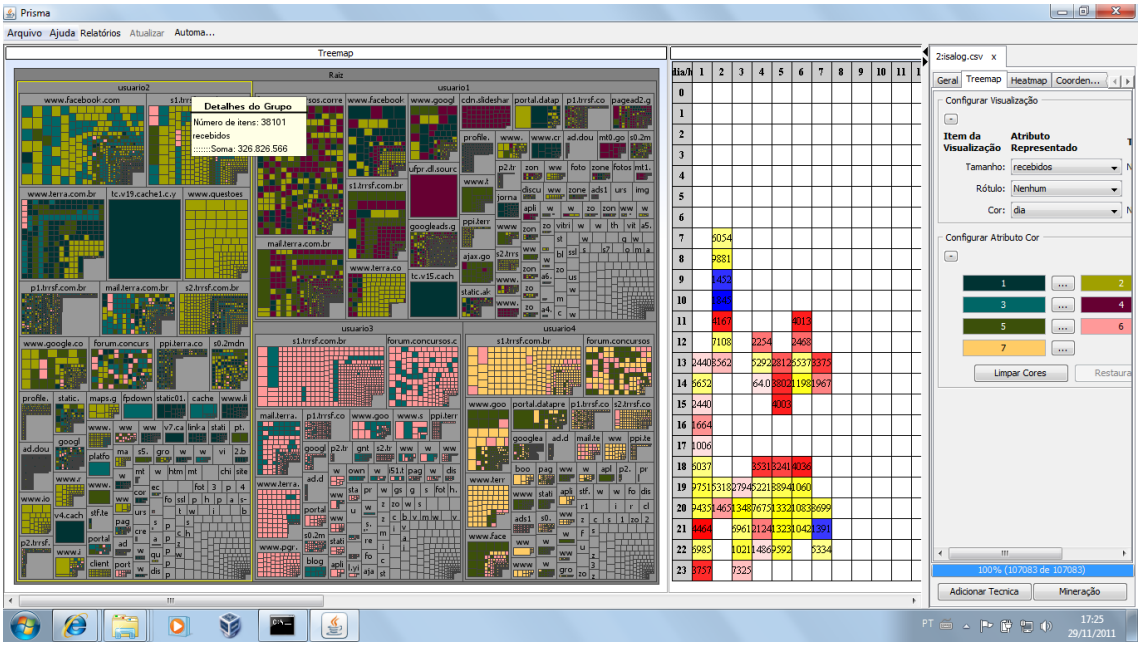

Figura 49: Tráfego por usuário x site, detalhado no HeatMap por dia x hora

Desta forma, a técnica HeatMap agrega a ferramenta PRISMA uma visão do comportamento da transação em relação ao tempo e a quantificação do fluxo de dados em determinado período, permitindo por exemplo, mostrar em uma visão a quantificação de transações de um mês dia-a-dia e hora-a-hora (dia x hora) permitindo a detecção de gargalos, interrupções e picos de transações.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Cada vez mais as empresas prestam serviços através de canais de relacionamentos eletrônicos. As análises do comportamento destas transações ajudam a entender o relacionamento do cliente com a empresa, e sugerir produtos e serviços mais adequados aos clientes, criando oportunidades de negócio.

A cada dia esses serviços digitais são utilizados pelos clientes, o que torna imprescindível que esses serviços estejam disponíveis sempre que forem demandados. Então, o que se quer descobriré porque não foi possível atender a uma determinada demanda do cliente. Uma transação não concluída pelo cliente pode representar uma simples ocorrência, tais como, senha inválida ou acesso negado a determinado serviço. Por outro lado, esta simples ocorrência pode indicar a oportunidade de realizar treinamentos ou de oferecer serviços sob demanda a alguns clientes. Outro possibilidade para transações canceladas pode estar relacionada a falta de recursos operacionais disponíveis para realiza-la.

Um dos objetivos desta dissertação foi a criação de cenário de análises para os contextos anteriormente apresentados. E um dos cenários explorados foi a análise de transações não concluídas quanto a falta de recursos operacionais, primeiramente analisando a quantidade de transações não concluídas e posteriormente a utilização do recurso operacionais através da análise do trafego de rede.

A disponibilização de serviçoseletrônicos ao usuário está intrinsecamente ligada à segurança da informação seja pela confidencialidade, integridade ou pela disponibilidade. Estes três pilares da segurança da informação ajudam a garantir a qualidade do serviço ofertada ao usuário. O entendimento do usuário sobre a importância da disponibilidade da informaçãocontribui para a diminuição de incidentes e a continuidade dos serviços.

Para o contexto da visualização de segurança da informação também foram propostos cenários que objetivavam identificar padrões de comportamento dos usuários e avaliar os recursos utilizados na rede.

A ferramenta de visualização da informação PRISMA foi utilizada para análise dos cenários propostos. As características da ferramenta PRISMA, tais como: múltiplas técnicas de visualização da informação, múltiplas visões coordenadas, facilidade de interação com os dados através de filtros, podem facilitar a análise e monitoramento de grandes volumes de dados gerados por logs de sistemas e redes.

Apesar da descoberta de padrões relevantes no contexto de análise, de acordo com os analistas de segurança, foi apontada a necessidade de melhoria de recursos para visualização de eventos em relação ao tempo.

Com intuito de atender essa necessidade foi proposto a visualização de HeatMap para ser incorporada a ferramenta PRISMA. Esta técnica foi escolhida em função da facilidade de percepção, facilidade de representação do aspecto temporal dos dados, e possibilidade de monitoria dos dados em fluxo contínuo.

Assim, este trabalho esteve focado nos aspectos de dar diretrizes na utilização de ferramentas de visualização de informação para análise de logs de transações eletrônicas e tráfego de rede, implementação de técnica de visualização de informação que possibilitasse de maneira facilitada a análise do aspecto temporal para dados de log de sistemas e rede.

6.1 Trabalhos Futuros

Os trabalhos futuros sugeridos são:

- Melhoramentos no módulo Heatmap, tais como:
    - Aplicação de um painel flutuante para a visualização textual dos dados.
    - Criação de ferramentas para zoom e seleção dos dados.
    - Criação de uma linha do tempo, para visualização dinâmica dos dados.

    - Criação de um filtro específico para a técnica.
- Concepção de novos cenários;
- Detalhar análise dos cenários propostos;
- e, utilizar outras bases de logs.

7 BIBLIOGRAFIA

Bachthaler, S., &Weiskopf, D. (2008).Continuous Scatterplots. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1428-1435.

Baldonado, M. Q., Woodruff, A., &Kuchinsky, A. (2000).Guidelines for using multiple views in information visualization. Proceedings of the working conference on Advanced visual interfaces (pp. 110 - 119). Palermo, Italy: ACM Press.

Becker, R. A., & Cleveland, W. S. (1987).Brushing scatterplots.Technometrics, 29(2), 127 - 142.

Card, S. K., Mackinlay, J. D., &Shneiderman, B. (Eds.). (1999). Readings in Information Visualizationn: Using Vision to Think. San Francisco: Morgan Kaufmann Publishers.

Carr, D. A. (1999).Guidelines for Desegning Information Visualization Applications. Proceedings of Ericsson Conference in Usability Engineering, (pp. 1-7).

Daassi, C.; Nigay, L.; Fauvet, M. (2008).A taxonomy of temporal data visualization techniques.

Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis.

Friendly, M.,(2009). Milestones in the history of thematic cartography, statistical graphics, and data visualization.

Godinho, P. I.,Meiguins, B., Meiguins, A., do Carmo, R., Garcia, M., Almeida, L., & Lourenço, R. (2007). PRISMA - A Multidimensional Information Visualization Tool Using Multiple Coordinated Views. Proceedings of 11th International Conference Information Visualization (IV '07), (pp. 23-32). Zurich.

IBGE, P. N. (2009). Síntese de Indicadores. IBGE. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística.

Inselberg, A. (2009). Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer.

Inselberg, A. (1985) The Plane with Parallel Coordinates, In The Visual Computer, pages 69-91.

Inselberg, A., &Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multi-dimensional geometry. Proceedings of the 1st conference on Visualization '90 (pp. 361 - 378). San Francisco: IEEE Computer Society Press.

Koike, H.; Ohno, K. (2004).SnortView: Visualization System of Snort Logs

Komlodi, A. Goodall, J. R. Lutters, W. G. (2004). An Information VisualizationFramework for Intrusion Detection.CHI 2004, Abril 24–29, 2004, Vienna, Austria.

Malécot, E. L (2006) Interactively Combining 2D and 3D Visualization for Network

Marty, R (2008).Applied Security Visualization.

Nascimento, H. A., & Ferreira, C. B. (2005). Visualização de Informação: Uma Abordagem Prática. In: XXIV JAI - Jornada de Atualização em Informática (pp. 1262-1312). São Leopoldo: SBC.

North, C., &Shneiderman, B. (2000). Snap-Together Visualization: A User Interface for coordinating Visulizations via Relational Schemata. Advanced visual interfaces , (pp. 23-26).

Pillat, R. M., Valiati, E. R., &Freitas, C. M. (2005).Experimental Study on Evaluation of Multidimensional Information Visualization Techniques. CLIHC'05, (pp. 20-30). Cuernavaca - Mexico.

Pillat, R. M.; Freitas, C. D. S. Coordinating Views in the InfoVis Toolkit.

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. Proceedings of the IEEE Symposium on Visual Languages, (pp. 336 -343).

Shneiderman, B. (2009). Treemaps for Space-Constrained Visualization of Hierarchies.Acessoem 01 de 05 de 2012, disponívelem HCIL - Department of Computer Science - University of Maryland: http://www.cs.umd.edu/hcil/treemap-history

Spence, R. (2007). Information Visualization - Design for interaction (2ª ed.). Essex, UK: Pearson Education Limited.

Takada, T. Koike, H. (2006). Tudumi: Information Visualization System forTraffic Monitoring VizSEC'06, Novembro 2006, Alexandria, Virginia, USA.

Takatsuka, M. Gahegan, M. (2002).GeoVISTA Studio: A Codeless Visual Programming Environment ForGeoscientific Data Analysis And Visualization.

Thomas,J. J.,& CookK. A., (2005). Iluminating the Path: The research and development agenda for visual analytics (I ed.). National Visualization and Analytics Center.IEEE CS Press.

Ward, M. O., Grinstein , G., &Keim, D. (2010). Interactive Data Visualization: Foundations, Techniques, and Applications.

Ware, C. (2004). Information Visualization: Perception for Design (2ª ed.). San Francisco, CA, USA: Elsevier.

Weaver, C. (2004). Building Highly-Coordinated Visualizations In Improvise. Proceedings of the IEEE Symposium on Information Visualization .

Wilkinson, L; Friendly, M. (2008).The History of Cluster Heat Map.