Resumo

O trabalho de investigação criminalista é de fundamental importância para o estado moderno. Em um mundo globalizado, onde a troca de informações entre os que compõem o crime organizado é feita com eficiência, faz-se necessário que os investigadores disponham de ferramentas avançadas para ajudar na elucidação dos crimes. Os relatórios policiais geralmente são armazenados sob a forma de texto não estruturado, sendo necessário que um analista revise tais materiais de maneira manual. Este trabalho é lento e tem um alto custo para os cofres públicos. A proposta de estudo deste trabalho científico consiste na concepção de uma ferramenta de apoio à investigação, cuja finalidade é otimizar o processo de leitura e interpretação dos relatórios policiais, reduzindo conjuntamente o tempo e os recursos gastos em uma investigação. A redução do tempo de investigação proporcionará um ganho percentual na elucidação dos crimes, uma vez que as primeiras horas após a ocorrência do crime são críticas para a conclusão da investigação. A ferramenta é alimentada por arquivos de relatórios policiais, através do qual será possível gerar uma rede de relacionamentos dos elementos envolvidos numa investigação. A rede de relacionamentos em questão mostra o grau de relação entre os alvos da investigação bem como suas respectivas conexões e delitos. A geração da rede de relacionamentos é feita através de um algoritmo que foi desenvolvido ao longo deste trabalho. O *software* foi desenvolvido em Java, utilizando conceitos de arquitetura em camadas e visa aplicar técnicas baseadas em mineração de textos para facilitar as investigações criminais.

Abstract

The criminal investigation is very important for the modern state. In a globalized world, where the information exchange between those who participate from the organized crime is done efficiently, it's necessary that the investigators have advanced tools to improve the solution of crimes. The police reports are usually stored in non-structured text form, been necessary that an analyst to review these materials manually. This work is slow and promotes a high cost for the state. The propose of this scientific study is to create a tool for help the investigations, whose finality is to optimize the reading and interpretation process of police reports. The reduction of time will make a percent gain in solution of crimes, because the first's hours after a crime occurs are critics for the conclusion of the investigations. The tool proposed works by the insertion of police reports archives that will be possible to generate the relationship networks from the elements involved in an investigation. The relationship network shows the relationship degree between the targets of the investigation as well as its respective connections and crimes. The generation of the relationship network is done by an algorithm that was developed during this work. The software was developed in Java, using concepts of architecture and uses text mining techniques to make the criminal investigations easier.

Sumário

Índice de Figuras

Tabela de Símbolos e Siglas

KDD - *Knowledge Discovery in Databases*

EUA – Estados Unidos da América

SMS - *Short Message* Service

MMS - *Multimedia Messaging Service*

VOIP – *Voice over IP*

E-MAIL – *Eletronic mail*

CSV - *Comma-separated values*

FBI - *Federal Bureal of Investigation*

IP – *Internet Protocol*

Capítulo 1
Introdução

A criminologia refere-se ao estudo da compreensão da realidade criminal em todos os seus aspectos [1]. No inicio, a criminologia preocupava-se apenas com a figura do delinqüente, e utilizava fundamentos da biologia e psiquiatria para explicar os crimes. Numa segunda fase da criminologia, as atenções foram voltadas para o meio social no qual ocorreu um delito. Compreende-se que não existe sociedade sem crimes, pois transgressão à norma é algo inerente aos seres humanos que convivem em sociedade, logo não é cabível investigar um crime sem evocar o meio social no qual este está envolvido [1].

As organizações criminosas estão se tornando cada vez mais sofisticadas no Brasil. A facilidade de comunicação entre os delinqüentes está ajudando a fortalecer ainda mais este tipo de organização. Os crimes passaram a ter uma maior quantidade de pessoas envolvidas, o que torna o trabalho do analista ainda mais difícil. Para esclarecer a realidade de um crime em todos os seus aspectos, o analista deve então realizar um estudo em um grupo de pessoas que apresentam relações sociais, interagem e cometem crimes.

1.1 Motivação.

O avanço dos meios de comunicação e da internet tem facilitado a ação de grupos terroristas. As redes de comunicação estão altamente acessíveis à maioria da população, o que possibilita uma maior troca de informações entre os criminosos e grupos terroristas.

Os ataques terroristas estão se tornando mais violentos e freqüentes. No dia 11 de setembro de 2001, os EUA (Estados Unidos da América) sofreram ataques

terroristas coordenados e sincronizados. Prédios comerciais localizados na cidade de Manhattann, conhecidos como World Trade Center, e a sede do pentágono nos Estados Unidos foram atingidos por aviões comerciais. Estima-se que o número de vítimas nestes ataques ultrapassa os 2900. Uma investigação realizada pelo FBI (*Federal Bureal of Investigation*) nos EUA, composta por mais de 7 mil agentes, concluiu que os ataques foram realizados por uma organização terrorista chamada *Al-Qaeda*. Ataques coordenados como estes só foram possíveis com o avanço da tecnologia na área da comunicação.

O avanço da comunicação é uma necessidade na sociedade contemporânea. Os benefícios proporcionados pela democratização dos meios de comunicação são inúmeros. A comunicação permitiu a redução de custos para as empresas, o que possibilita uma redução dos preços dos produtos. No meio acadêmico, permitiu a troca de conhecimento entre as mais distantes universidades. A necessidade de comunicação é de vital importância para a sociedade. A diminuição do acesso à comunicação para a população causaria grandes problemas para a sociedade.

Para permitir o acesso à comunicação e evitar que criminosos tirem proveito das vantagens oferecidas, faz-se necessário realizar uma análise do conteúdo das informações trafegadas. As agências de inteligência dos governos devem ter meios eficientes de analisar as informações que são transmitidas nos meios de comunicação.

A quantidade de informação e o tempo necessário para a análise das informações trafegadas nos meios de comunicação inviabilizam a utilização de seres humanos para esta tarefa. Uma vasta variedade de dados e voz podem ser interceptados e analisados por computadores para auxiliar no processo de investigação criminal. Isto permitirá que os criminosos sejam encontrados em tempo hábil e consequentemente evitar que novos crimes aconteçam.

1.2 Trabalhos relacionados: X-TRACT Verint

A Verint Analytics and Communications Solutions é uma empresa privada Israelense que produz *softwares* para governos e agências de inteligência com a

finalidade de ajudar no processo investigativo de terroristas e criminosos que exploram os meios de comunicação disponíveis à população.

O *software* extrai informações críticas de uma vasta variedade de bases de dados que podem ser interceptados por qualquer rede de comunicação. No portfólio de serviço oferecidos pela empresas constam: FAX, MMS (*Multimedia Messaging Service)*, SMS (*Short Messaging Service*), mensagens instantâneas, e-mail, navegação na *web*, VOIP (*Voice over* IP), conversas por voz, etc. As informações interceptadas são analisadas em tempo hábil e podem rapidamente detectar ameaças à segurança e constituir evidências para acusações.

Devido à natureza das operações realizadas pela empresa, não existe divulgação de informações sobre clientes, o que dificulta a análise dos métodos utilizados para a obtenção dos resultados.

O X-TRACT é um sistema de apoio à investigação fim-a-fim. Possui acesso a diversas bases de dados estruturadas (bancos de dados) e não estruturadas (textos). O fato de permitir a utilização de bases de dados de texto possibilita uma grande abrangência da informação. A maior parte das informações disponibilizadas está armazenada sob a forma de texto não estruturado. O *software* apresenta níveis de segurança da informação, permitindo a utilização em organizações que apresentam níveis de hierarquia.

O sistema possui um motor baseado em regras, para que se possam produzir respostas rapidamente, devido ao alto volume de informação. Um sistema de pontuação permite a integração de múltiplas fontes de dados. Isto permite que várias regras sejam aplicadas a múltiplas fontes de dados diferentes, sendo elas estruturadas ou não.

A interface gráfica projetada de maneira intuitiva e de fácil compreensão é um dos pontos fortes do *software*. Os resultados são apresentados por meio de uma rede de relacionamentos. Os elementos desta rede estão conectados por pesos, que indicam a relação entre as pessoas e os eventos. Com base nestas informações, um analista poderá ter uma melhor visão sobre as relações entre as pessoas e os eventos.

O sistema X-TRACT possui características semelhantes ao trabalho desenvolvido nesta monografia. A utilização de bases de dados não estruturadas, a adoção de um sistema baseado em regras e a facilidade na visualização do conhecimento produzido foram características incorporadas ao projeto.

1.3 Extração de conhecimento em bases de dados

A extração de conhecimentos em bases de dados, que é conhecida como KDD (*Knowledge Discovery in Databases*), é o processo de obtenção de conhecimento a partir de bases de dados. A característica mais importante de KDD é a extração de conhecimentos não triviais, implícitos, desconhecidos e potencialmente úteis a partir de bases de dados.

O número e o tamanho das bases de dados estão crescendo continuamente a cada dia que se passa. Estima-se que o volume de informações no mundo dobra a cada 20 meses [5]. A automação dos negócios, o conteúdo da internet, imagens de satélite produzem uma imensa quantidade de informação.

A quantidade de informação produzida pelas empresas e pelo governo excede bastante a capacidade de compreensão do ser humano. As limitações do homem o impedem de encontrar, nestas bases de dados, regularidades implícitas, regras e leis de formação dos dados devido ao volume de informações gerados. O resultado disto é um aumento do vazio entre a geração de dados e geração de informação. Grande parte dos dados armazenados pelas empresas não são utilizados por falta de ferramentas eficientes de geração de informação.

Ferramentas de KDD visam analisar dados e produzir como saída padrões de comportamento dos dados, ou conhecimento. A saída de um programa que monitora dados em uma base de dados e produz padrões é considerado conhecimento descoberto.

Em geral, este processo pode ser dividido em sete etapas:

1) Limpeza dos dados: Onde serão removidos dados inconsistentes.

2) Integração de dados: Fontes de dados diferentes serão integradas.

3) Seleção de dados: Os dados relevantes serão separados.

4) Transformação de dados: Os dados serão transformados em um formato apropriado para a fase de mineração de dados.

5) Mineração de dados: Fase onde os dados são explorados por procedimentos de mineração.

6) Avaliação dos padrões: Nesta etapa, serão validados os padrões fornecidos pela etapa anterior.

7) Apresentação do conhecimento: Apresentação dos resultados validados ao usuário.

Em geral, ferramentas de KDD buscam fazer identificação de classes conhecido como *clustering*, fazer previsões sobre os dados e descoberta de associações entre os dados. Este processo deve acontecer de maneira automatizada e o conhecimento obtido deve ser apresentado de forma clara para o usuário.

1.4 Mineração de dados

A mineração de dados é um campo de pesquisa relacionado à exploração de grandes volumes de dados para encontrar padrões e relacionamentos entre as variáveis. A mineração de dados é parte de um processo de *Knowledge Discovery in Databases*, não sendo o próprio processo em si. Em geral, o processo de mineração de dados pode ser dividido em três etapas: A exploração dos dados, definição do padrão e verificação.

Hoje em dia, as empresas estão focadas em armazenar a maior quantidade de informações possível. Estas empresas conseguem coletar dados de maneira eficiente, mas ainda não existe um aproveitamento adequado para os dados que foram coletados. Os dados armazenados podem gerar conhecimento para a melhoria no atendimento aos clientes e para a identificação de padrões de

comportamento de compra. Com base nestas informações, uma empresa poderá fazer um melhor planejamento das suas atividades para atender o cliente de forma eficiente.

A mineração de dados é um campo de pesquisa multidisciplinar, pois envolve áreas como banco de dados, aprendizado de máquina, estatística, reconhecimento de padrões, recuperação de informações, inteligência artificial, etc.

A grande capacidade de armazenamento de dados, que são coletados e armazenados em vários repositórios, ultrapassa a capacidade humana de entendimento e identificação de padrões. Ferramentas automatizadas para ajudar na análise de grandes volumes de dados são muito importantes para realizar um processamento adequado das informações armazenadas.

1.5 Mineração de texto

A mineração de textos refere-se ao processo de extração de padrões não triviais e interessantes, ou conhecimento a partir de bases de dados textuais não estruturadas [2]. A mineração de texto foi definida como a extração de informação não trivial implícita, desconhecida e potencialmente útil para um conjunto de dados [3].

A mineração de textos é idealizada por muitos como sendo uma nova forma de obtenção de conhecimento. A forma mais natural de armazenamento de informações é o texto. Um estudo recente indica que 80% das informações das companhias estão armazenadas em forma de texto não estruturado [2]. Acredita-se que a mineração de texto tem um potencial de desenvolvimento comparável ao potencial da mineração de dados, que apresenta um alto valor comercial.

Bases de dados tradicionais armazenam informações de maneira estruturada e provêem métodos de pesquisa para obter registros que satisfazem determinadas condições [3]. A maioria dos trabalhos relacionados à KDD está concentrada em bases de dados estruturadas. É fato que existem poucos trabalhos relacionados à mineração de texto, pois a maioria dos trabalhos concentra-se em realizar o

refinamento de bases de dados estruturadas. Trabalhos relacionados a esta área, geralmente envolvem aprendizado de máquina e técnicas de análise estatísticas aplicadas ao reconhecimento automático de padrões.

1.5.1 O que caracteriza a mineração de textos

A mineração de textos é a descoberta, por meio de um computador, de informações novas e desconhecidas pela exploração de fontes de texto.

Mineração de texto é diferente do que se está familiarizado com busca na internet. Na busca na internet, o usuário procura por um conteúdo que ele já conhece, mas que foi escrito por terceiros. Já na mineração de texto, o objetivo é conhecer informação desconhecida, que ninguém conhece e consequentemente que não poderia ter sido obtida.

A mineração de texto é uma variação de um campo conhecido como mineração de dados. Um exemplo de mineração de texto é utilizar as informações de venda dos clientes de uma loja para fazer previsões de venda. Um cliente que compra uma lanterna provavelmente vai precisar comprar pilhas para poder utilizá-la.

A diferença entre a mineração de texto e a mineração de dados é a forma de aquisição dos dados. A mineração de textos tem o objetivo de extrair padrões a partir de textos em linguagem natural. Os bancos de dados foram criados para serem acessados por programas de computador, já o texto, é escrito para pessoas lerem. Não existem programas que lêem textos, pois pesquisadores da área acreditam que para fazer programas que possam interpretar textos, será necessário fazer uma simulação completa da mente humana [4].

Existe um campo chamado lingüística computacional que também é conhecido como processamento de linguagem natural. Está área de pesquisa está evoluindo em tarefas de análise de texto. Um exemplo de uma aplicação é fazer um programa que aparenta resumir as idéias centrais de um livro ou artigo. Separando as palavras que ocorrem com mais freqüência e subtraindo as palavras comuns

como artigos e pronomes, pode-se fazer uma ferramenta capaz de extrair as idéias centrais de um texto.

O grande problema enfrentado pela área de mineração de textos é a falta de ferramentas mais sofisticadas para análise da linguagem natural. Existem vários trabalhos em aberto relacionados à análise semântica de partes de texto. A tarefa de leitura de um texto pode ser dividida em etapas [4], fragmentando o texto para que o processo se torne mais eficiente.

Atualmente, não existem programas que façam a interpretação completa de um texto. Acredita-se que isto não será possível por um bom tempo [4]. Textos não estruturados podem ser escritos de diversas maneiras diferentes. Um texto escrito para um artigo é bem diferente do conteúdo de texto de um *e-mail*, um relatório, um livro de ação, etc. A natureza da informação também causa problemas para um analisador semântico.

1.5.2 O estado da arte e desafios

A mineração de textos pode ser considerada uma subárea da mineração de dados [2]. É uma nova forma de obtenção de conhecimento, pois a forma mais comum de armazenamento de dados é a forma textual. Embora a base de dados de texto seja maior que bases de dados convencionais, a tarefa de mineração de texto é bem mais complexa se comparada à mineração de dados. Os textos apresentam dados armazenados de maneira desordenada e de difícil entendimento, o que torna a mineração de textos uma campo multidisciplinar, envolvendo várias áreas do conhecimento como análise de texto, extração de informação, categorização, aprendizado, etc.

Os algoritmos de mineração de dados geralmente são divididos em duas etapas: Uma fase que transforma um texto comum em um texto padrão de forma intermediária, que pode ser um grafo ou dados relacionais. Em uma segunda fase, os dados armazenados numa forma intermediária são transformados em informação útil.

A obtenção de conhecimento deve estar associada a um domínio específico. Uma contextualização do texto é de fundamental importância para uma boa geração de conhecimento. Por exemplo, dado um conjunto de novos artigos científicos, espera-se que as informações obtidas a partir destes documentos sejam relativas ao meio acadêmico.

Uma pesquisa realizada entre as principais ferramentas comerciais de mineração de texto mostra que os *softwares* podem ser divididos em classificadores de documentos e ferramentas de análise e entendimento. Os *softwares* de classificação têm a finalidade de organizar documentos de texto de acordo com a sua similaridade. Já as ferramentas de análise de texto são baseadas no processamento de linguagem natural e realizam análise, categorização e extração de informações.

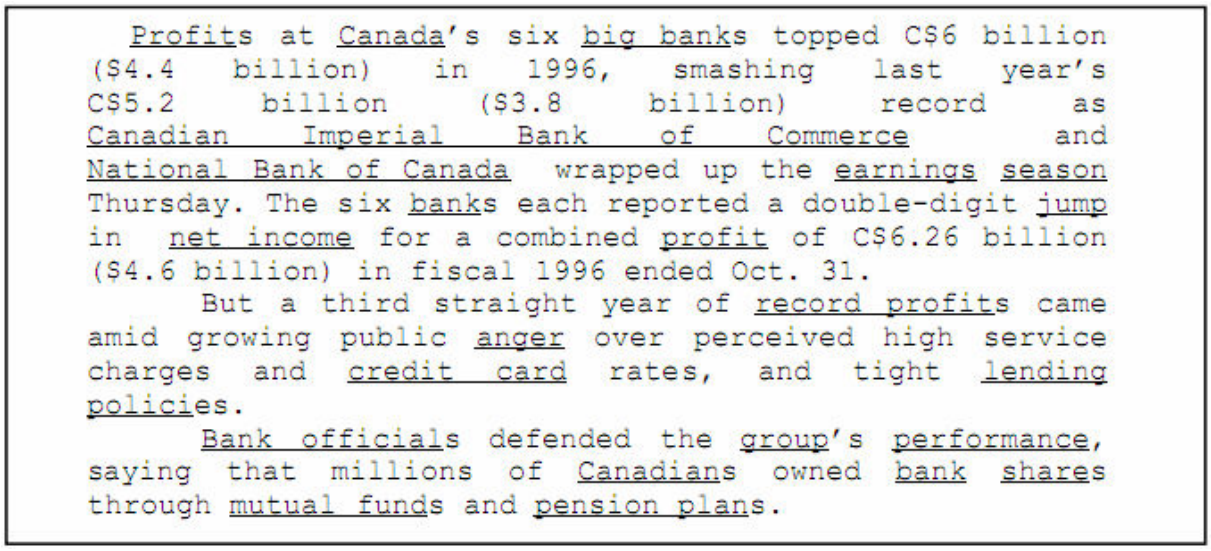
A mineração de textos apresenta alguns problemas em aberto, como a geração de documentos na forma intermediária, que podem apresentar vários níveis de complexidade. A análise semântica tem um alto custo computacional, com taxa de processamento de algumas palavras por segundo. As ferramentas foram aprimoradas para funcionar com textos escritos em inglês, o que gera um problema potencial para textos escritos com palavras em mais de um idioma. A qualidade da classificação dos textos influencia na geração do conhecimento, pois o analisador sintático pode ser mais eficiente quando se sabe o domínio do texto em questão.

1.5.3 Mineração de texto em nível de termos

É uma técnica de mineração de dados baseada na extração de termos significativos a partir de documentos de texto não estruturado. Os sistemas padrões de mineração de texto geralmente operam sobre documentos categorizados. A exploração de texto requer um pré-processamento para que o módulo de extração automática de termos possa encontrar elementos mais complexos do que apenas simples palavras.

Após a fase de pré-processamento, o texto recebe uma rotulação dos termos identificados pela fase anterior. Em seguida, o texto e entidades de alto nível identificadas são usados para dar suporte a uma variedade de operações de KDD.

A freqüência de co-ocorrência dos termos pode ser utilizada para prover uma base nas operações de KDD em coleções de documentos de texto. Uma outra aplicação é encontrar subconjuntos de documentos que diferem significativamente dos outros elementos da coleção. A Figura 1 mostra um exemplo da aplicação de um módulo de extração de termos. Os termos sublinhados serão rotulados com marcações que os identifica unicamente. A partir deste ponto, ferramentas de KDD podem ser utilizadas para gerar conhecimento.



Profits at Canada's six big banks topped C$6 billion ($4.4 billion) in 1996, smashing last year's C$5.2 billion ($3.8 billion) record as Canadian Imperial Bank of Commerce and National Bank of Canada wrapped up the earnings season Thursday. The six banks each reported a double-digit jump in net income for a combined profit of C$6.26 billion ($4.6 billion) in fiscal 1996 ended Oct. 31.

But a third straight year of record profits came amid growing public anger over perceived high service charges and credit card rates, and tight lending policies.

Bank officials defended the group's performance, saying that millions of Canadians owned bank shares through mutual funds and pension plans.

Figura 1. Exemplo de aplicação de um módulo de extração de termos

A arquitetura de um módulo de extração de termos pode ser dividida em três estágios: Uma fase de processamento lingüístico, a segunda de geração de termos e uma terceira de filtragem de termos.

Na primeira fase, onde é feito o processamento lingüístico, o texto é transformado em *tokens.* Em um segundo momento, deverá ser feita uma classificação das palavras e armazenamento numa estrutura de *tokens*, destacando elementos como nomes, verbos e adjetivos.

Na fase de geração de termos, serão levantados os possíveis candidatos a termos. Composições de "nome nome", "nome preposição nome", "adjetivo nome" são exemplos de composições de termos. Em cada candidato a termo, também é gerado um coeficiente. Este coeficiente é utilizado em casos de concorrência entre os elementos de um termo. No caso de um texto contendo "T1 T2 T3", pode haver

uma geração de termos candidatos das seguintes formas concorrentes: “T1 T2”, “T2 T3”. Neste caso o par com melhor associação de coeficientes será escolhido primeiro.

Na fase de filtragem de termos, devem-se definir os critérios para a eliminação dos termos gerados pela etapa anterior. A fase de geração de termos produz conjuntos de associações de termos, sem a preocupação com a relevância dos termos. Isto termina provocado um excesso de termos gerados, cabendo a realização de um refinamento dos termos gerados, para que se possam aplicar ferramentas de mineração de texto. Os termos serão escolhidos de acordo com o coeficiente gerado na fase anterior. Isto permite que os termos com maior relevância sejam escolhidos primeiro.

A forma de composição de padrões de geração de termos é uma área de pesquisa em aberto. Diversas formas de geração de termos podem ser definidas para obtenção de resultados diferentes. A geração automatizada de termos é de grande importância para um processo de mineração de dados, pois possibilita que um documento seja avaliado de maneira mais natural, pois os termos são mais representativos que as palavras.

1.6 Estrutura da monografia

A monografia está organizada da seguinte forma:

- Capitulo 1: Apresenta uma introdução, trabalhos relacionados ao tema desta monografia e fundamentação teórica.
- Capitulo 2: Descreve o sistema de apoio à investigação proposto, mostra os requisitos e detalha o funcionamento do algoritmo desenvolvido.
- Capitulo 3: Estudo de caso realizado com livros clássicos da literatura e um depoimento real prestado à polícia com a finalidade de testar o sistema desenvolvido.

- Capitulo 4: São apresentadas as conclusões obtidas através da análise do sistema e apresentam-se propostas de trabalhos futuros.

Capítulo 2

O sistema de apoio à investigação

Neste capítulo, será apresentada a concepção e o desenvolvimento de um sistema de apoio à investigação criminalista. Esta ferramenta será capaz de gerar uma estrutura denominada rede de relacionamentos a partir da análise de documentos de texto não estruturado. Esta rede de relacionamentos é uma matriz ponderada formada pelos elementos de investigação alinhados na primeira linha e na primeira coluna da matriz.

2.1 Visão geral do sistema

O sistema de apoio à investigação foi projetado para agilizar e melhorar o processo investigativo. O analista poderá obter informações relativas a determinados documentos de texto não estruturado de maneira automatizada. A rede de relacionamentos fornece ao analista uma visão complementar dos casos que estão sendo investigados. O grau de relação entre as pessoas e os eventos será mostrado em uma matriz para que futuramente se possa montar uma rede de relacionamentos.

Com base nas informações fornecidas pelo *software*, será possível encontrar os criminosos de maneira mais rápida e eficiente. Os crimes também serão analisados no processo de investigação. Ao final, será obtida a relação entre os suspeitos, conexões e os delitos envolvidos.

O fluxo de execução do *software* é dado da seguinte maneira: O analista insere no sistema textos não estruturados, nomes de suspeitos, nomes de envolvidos em um crime, palavras que descrevem delitos e o sistema monta uma rede de relacionamentos entre as pessoas envolvidas em um crime e os delitos. A saída do sistema é uma rede de relacionamentos, que servirá de base para o analista dar andamento aos processos de investigação.

2.2 Definições do sistema

Esta seção apresenta as estruturas que foram definidas para a manipulação da informação e funcionamento do sistema de investigação. O entendimento destas estruturas é de fundamental importância para o que se possa utilizar o sistema de apoio à investigação de maneira correta.

2.2.1 Alvos e conexões

Nesta subseção, será apresentado como os elementos de um processo investigativo serão representados no *software*. O analista será responsável pela classificação destes elementos, com base em suas conclusões ao longo do processo. O analista poderá modificar a classificação dos elementos de investigação durante a investigação, caso julgue necessário.

Alvo

Esta classificação representa os principais elementos de investigação. O analista deverá guiar as suas investigações através da análise dos alvos. Os alvos de investigação podem ser classificados como pessoas ou eventos.

A classificação uma pessoa como alvo, representa que esta pessoa é suspeita de realizar um determinado crime, ou teve algum tipo de participação direta no delito.Os eventos classificados como alvos serão os acontecimentos que fazem parte de um determinado crime. Exemplos de eventos são: assassinato, homicídio, extorsão, etc.

Conexão

Esta classificação representa os elementos que tiveram uma participação indireta em um determinado crime. As conexões serão classificadas como pessoas ou eventos.

A classificação de uma pessoa como conexão, indica que esta pessoa teve uma participação indireta em um determinado crime. Exemplos de pessoas

conexões são fornecedores de equipamentos ilegais. Os eventos rotulados conexões são acontecimentos que possuem algum tipo de relação com o crime.

2.2.2 Relatório

A base de dados do sistema de investigação são relatórios. Os relatórios são documentos de texto não estruturado que foram escritos por pessoas. Servirão de base para a aplicação da ferramenta desenvolvida neste trabalho, com a finalidade de extração do conhecimento.

Os relatórios, no projeto em questão, são documentos de texto puro. Não existe nenhuma restrição quanto à escrita. Os relatórios podem ser textos de qualquer natureza, cabendo ao analista identificar quais textos são relevantes para o caso em questão.

No estudo de caso que foi feito para analisar a ferramenta desenvolvida, a base de dados de texto escolhida foram livros da literatura clássica. Estes livros estão armazenados sob a forma de arquivos de texto puro.

2.2.3 Conectivos textuais

Os conectivos textuais são palavras que tem a função de intensificar ou de atenuar a relação entre dois elementos de investigação. Os conectivos textuais são palavras que possuem um peso associado. Um peso associado maior do que zero, representa uma palavra aglutinadora e um peso menor do que zero, representa uma palavra separadora.

A função que calcula a relação de peso entre os elementos de investigação faz uma busca por palavras aglutinadoras e separadoras, entre as palavras que representam os elementos de investigação. Caso haja conectivos textuais, o algoritmo leva em consideração o peso associado aos conectivos textuais e calcula a nova relação entre os elementos de investigação.

Para cada projeto, será definida uma nova lista de conectivos textuais. Esta definição permite que o analista tenha maior flexibilidade na escolha de conectivos

textuais. O analista pode chegar à conclusão que uma palavra pode ser aglutinadora em um determinado caso e em outro não.

2.2.4 Rede de relacionamentos

A estrutura denominada rede de relacionamentos representa o grau de relação entre os alvos e conexões. Esta está armazenada sob a forma de uma matriz ponderada. O peso associado a cada elemento da matriz, representa o grau de ligação entre os elementos de investigação.

A primeira linha e a primeira coluna desta matriz são formadas pela concatenação dos elementos de investigação. Desta maneira é possível obter uma estrutura que represente a relação entre todos os elementos.

Esta matriz é simétrica, pois o cálculo das relações entre os elementos de investigação são feitos utilizando um algoritmo padrão. Por exemplo, o peso de relação entre um personagem X e um personagem Y é igual ao peso da relação entre Y e X. Isto garante a propriedade de simetria da matriz.

Por convenção, o peso relacionado à relação entre um determinado elemento com ele mesmo foi definido como zero. A conseqüência desta definição é que a matriz possui a diagonal principal nula. Esta convenção possibilitou calcular uma normalização na matriz para tornar mais fácil a compreensão do grau de relação entre os elementos. A Figura 2 mostra um exemplo de rede de relacionamentos

null	Capitu	Bentinho	Cosme	Escobar	Ezequiel	José
Capitu	0.0	4.305265	2.7481067	9.298723	4.937892	10.780688
Bentinho	4.3052673	0.0	0.99009746	1.1510603	0.3080875	2.5192177
Cosme	2.7480912	0.99009854	0.0	2.4074557	0.22856133	5.369362
Escobar	9.298748	1.1510625	2.4074612	0.0	2.8271084	3.5273027
Ezequiel	4.9378963	0.30808794	0.22856125	2.8271008	0.0	0.92855704
José	10.7807045	2.519221	5.369351	3.5273104	0.9285577	0.0

Figura 2. Exemplo de rede de relacionamentos

A rede de relacionamentos pode não mostrar uma visão clara das relações entre os elementos de investigação, pois esta é composta por números reais que podem assumir valor entre zero e infinito. Uma solução proposta foi aplicar uma normalização na rede encontrada. Esta normalização permite que o analista tenha uma visão mais objetiva do peso entre as relações.

A forma de realizar a normalização foi definida como dividir cada elemento da matriz pelo somatório de todos os elementos da matriz, e em seguida multiplicar por cem. Com uma matriz normalizada, é possível ter uma visão de como as relações entre um elemento de investigação e os outros elementos estão distribuídas. A Figura 3 mostra um exemplo de uma matriz de uma rede de relacionamentos normalizada.

A	Capitu	Bentinho	Cosme	Escobar	Ezequiel	José
Capitu	0.0	4.113768875377105	2.625872230553819	8.88512025581543	4.718256929497627	10.301168162598922
Bentinho	4.113771073074206	0.0	0.9460583993175629	1.0998616893088422	0.29438391559932864	2.407164103617105
Cosme	2.625857419986393	0.9460594312796803	0.0	2.300373223833887	0.21839503154133258	5.130535350607352
Escobar	8.885144143827405	1.0998637914538965	2.3003784791965223	0.0	2.7013599727861504	3.370409965772984
Ezequiel	4.7182610382356875	0.29438433602833947	0.21839495509969425	2.701352710830509	0.0	0.8872552677162248
José	10.301183928686829	2.4071672568346862	5.130524839882082	3.3704173232806727	0.8872558983597412	0.0

Figura 3. Exemplo de rede de relacionamentos normalizada

A normalização da matriz que representa a rede de relacionamentos não remove a propriedade de simetria da matriz.

2.3 Método utilizado para calcular o peso do relacionamento

A parte mais importante do sistema de apoio à investigação é a forma de se calcular o peso entre os elementos de investigação. A base de dados a ser analisada pelo sistema é composta por textos não estruturados.

A análise de documentos de texto não estruturado por ferramentas de processamento de linguagem natural foi descartada porque este processo requer um custo computacional bastante elevado. Métodos de processamento de linguagem natural em geral envolvem inteligência computacional, o que poderia levar o sistema a gastar um tempo relativamente alto no processamento dos relatórios. O sistema precisará analisar uma grande quantidade de informação e prover um resultado em um tempo hábil.

As soluções propostas foram pensadas a partir de sistemas baseados em regras, devido à restrição no tempo de resposta do algoritmo. Nas próximas seções serão apresentados os algoritmos que foram criados para o cálculo do peso das relações.

2.3.1 O método do inverso da distância entre as palavras

A solução inicial para resolver o problema do cálculo do peso entre dois elementos de investigação foi utilizar o somatório do inverso da distância textual entre os elementos de investigação.

Este método está fundamentado na hipótese de que quanto menor for a distância em palavras entre dois elementos de investigação, maior será o peso associado ao relacionamento. Se os elementos de investigação aparecem no texto a uma distância menor, o peso associado ao relacionamento será máximo. Quanto maior for a distância entre os elementos, menor será a influencia no cálculo do peso da relação.

A Figura 4 mostra um fragmento de texto que foi retirado do livro Dom Casmurro, de Machado de Assis. As palavras significativas para o algoritmo estão sublinhadas com finalidade de proporcionar uma melhor visualização. O texto é então segmentado em palavras numeradas. A palavra Bentinho ocorre nas posições 10 e 73, a palavra Capitu ocorre na posição 78. Neste caso o peso do relacionamento entre Bentinho e Capitu será a soma do inverso das distâncias entre as palavras, que são 68 e 5. O peso será 1/68 somado a 1/5, que é igual a 0.2147 aproximadamente.

—É um modo de falar. Em segredinhos, sempre juntos. **Bentinho** quase que não sai de lá. A pequena é uma desmiolada; o pai faz que não vê; tomara ele que as cousas corressem de maneira, que... Compreendo o seu gesto; a senhora não crê em tais cálculos, parece-lhe que todos têm a alma cândida...
— Mas, Sr. **José Dias**, tenho visto os pequenos brincando, e nunca vi nada que faça desconfiar. Basta a idade;
Bentinho mal tem quinze anos. **Capitu** fez quatorze à semana passada...

Figura 4. Fragmento de texto extraído do livro Dom Casmurro

Para que o sistema apresentasse um resultado mais preciso, os conectivos textuais passaram a atuar no calculo da distância entre as palavras. Os conectivos textuais são palavras que tem um peso associado. Caso haja palavras conectivas

entre os elementos de investigação, o sistema soma o peso dos conectivos com o sinal invertido.

A introdução dos conectivos textuais possibilita que a distância entre as palavras torne-se zero ou negativa. Isto se tornou um problema, pois a função F(X) = 1 / X não está definida no ponto x = 0. Outro fato importante é quando x assume valores menores do que zero, essas distâncias menores do que zero passam a influenciar negativamente no cálculo do peso entre as relações. A Figura 5 mostra o comportamento da função F(X) = 1 / X.

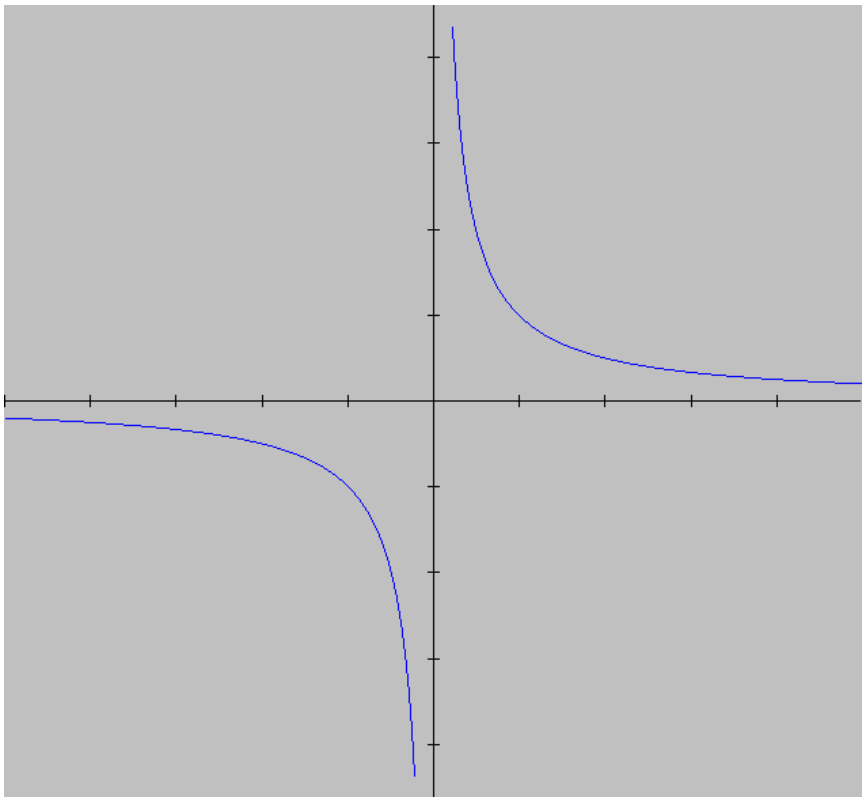


Figura 5. Função F(X) = 1 / X

A conclusão obtida foi que a função F(X) = 1 / X não poderia ser utilizada no calculo do peso entre os elementos de investigação porque não apresentava resultados desejáveis no intervalo em que x é menor ou igual a zero.

2.3.2 Utilização da função exponencial

Uma segunda proposta foi elaborada por causa de problemas encontrados com a função inverso da distância para o calculo do peso. A função F(X) = exp (-X)

tem características da função inverso da distância, é decrescente e não assume valores menores do que zero. A Figura 6 mostra o gráfico da função F(X) = exp(-X).

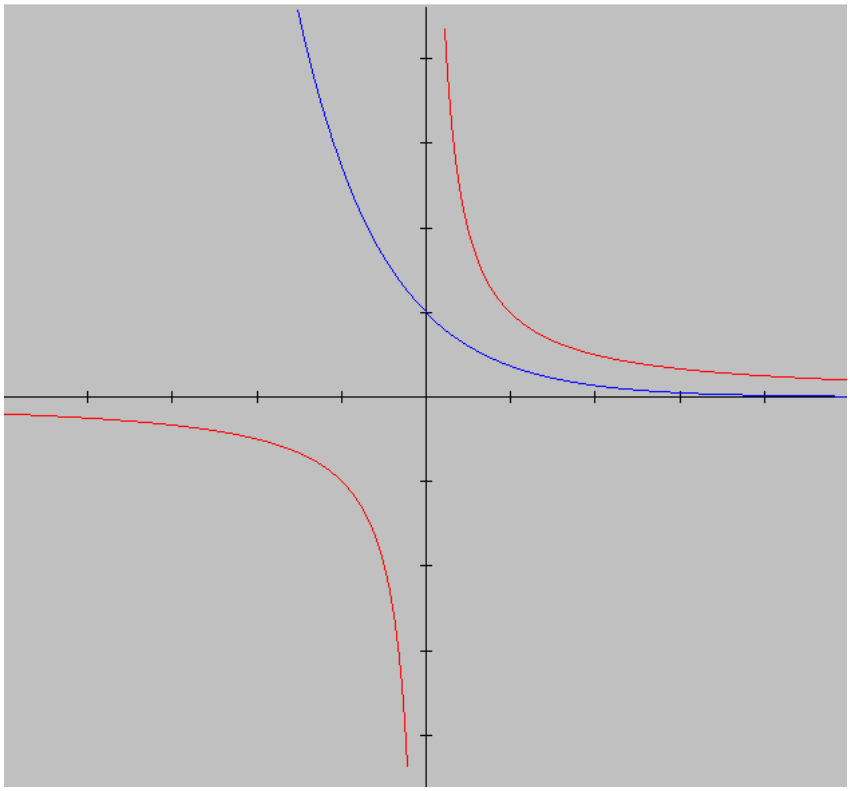


Figura 6. Comparativo entre a Função F(X) = exp(-X) e função F(X) = 1/X

Esta função possui um decaimento comparável ao da função inverso da distância e está definida para valores menores ou iguais a zero, como mostrado na Figura 6.

Esta função foi escolhida para ser adotada no projeto, pois apresenta resultados satisfatórios quando comparada à função inverso da distância e estádefinida para qualquer valor de x real.

2.4 Requisitos funcionais

2.4.1 Projeto

O sistema de investigação foi concebido com a idéia de possibilitar que o analista possa estar trabalhando em mais de um projeto ao mesmo tempo. Neste contexto, foi definida a idéia de projeto. Em um projeto, o analista poderá organizar

listas de alvos, lista de conexões, redes de relacionamentos, listas ponderadas de conectivos textuais.

A funcionalidade de criação de projeto define um novo projeto, com lista de alvos, lista de conexões, redes de relacionamentos e lista de conectivos textuais zerados. Cabendo ao analista carregar as informações relativas às novas listas, bem como a criação de novas redes de relacionamentos.

O usuário poderá salvar as alterações feitas em um projeto ou não, caso julgue conveniente. Exemplos de alterações no projeto são: alteração de suspeito para conexão, alteração na lista de conectivos textuais, etc.

2.4.2 Manipulação de relatórios

Esta seção apresenta as funcionalidades relativas aos relatórios, que são documentos de texto não estruturados que servirão de base de dados para a ferramenta de mineração de dados.

O sistema oferece ao analista uma funcionalidade que permite a leitura do relatório na tela do sistema. Este recurso é útil quando o analista quiser ter uma visão detalhada do documento a ser analisado. O analista poderá visualizar o relatório em uma tela desenvolvida para esta finalidade.

Com base nas informações de conectivos textuais e elementos de investigação, o analista poderá gerar as redes de relacionamentos a partir de relatórios. Primeiramente o analista deverá informar ao sistema quais são os elementos de investigação, diferenciando os alvos e as conexões. Uma segunda classificação dos elementos é aplicada no momento da inserção dos elementos de investigação, cabendo ao analista classificá-los como pessoas ou eventos. Os conectivos textuais deverão estar carregados no sistema para que o sistema tenha uma melhor precisão no cálculo do peso das relações.

O produto final da inclusão de um relatório é uma matriz que representa a rede de relacionamentos. Esta rede deverá ser armazenada em um meio persistente, para que o analista possa visualizar posteriormente. O formato do arquivo gerado é CSV (*comma separated values*), que representa uma matriz de

rede de relacionamentos. A cada relatório que o analista inclui, o sistema gera uma matriz ponderada de pesos.

2.4.3 Apresentação da rede de relacionamentos

A apresentação dos resultados é a etapa onde o analista poderá visualizar as informações extraídas dos relatórios pelo sistema. O analista deverá escolher o arquivo de rede de relacionamentos armazenado na etapa de inclusão de relatório e terá uma visualização gráfica da rede de relacionamentos.

A proposta inicial do sistema seria mostrar a rede de relacionamentos de maneira gráfica. O analista poderia ter uma visão de um grafo com pesos associados a cada elemento. Por restrições de tempo de projeto, o sistema limitou-se a mostrar a rede de relacionamentos em forma de uma matriz ponderada, cabendo ao analista extrair as informações geradas e usá-las em uma ferramenta gráfica.

A apresentação da rede de relacionamentos é feita através da matriz da rede e da matriz normalizada da rede. A matriz normalizada facilita a visualização de como estão distribuídos os relacionamentos de um determinado elemento de investigação.

Com base nas informações de pesos associadas aos elementos de investigação, o analista terá uma visão complementar das relações entre os alvos e conexões envolvidas em um processo investigativo.

A apresentação da rede de relacionamentos de forma gráfica torna a visualização mais clara e intuitiva. O analista poderá ter um melhor entendimento do caso quando este visualiza a rede em uma representação visual.

A aplicação desenvolvida neste trabalho acadêmico limita-se a apresentar a rede de relacionamentos de maneira matricial, cabendo ao analista coletar os dados e inseri-los em ferramentas de geração de redes de relacionamentos.

2.5 Detalhamento da implementação

Esta seção tem a finalidade de detalhar o algoritmo utilizado para calcular a rede de relacionamento. Será apresentada uma visão de como o algoritmo foi modificado para funcionar de maneira mais eficiente, e consequentemente proporcionar uma redução no tempo de resposta.

A Figura 7, abaixo, mostra como é feito o processamento dos arquivos de texto não estruturados até a geração de uma estrutura que contem uma lista de ocorrência de palavras associadas aos elementos de investigação.

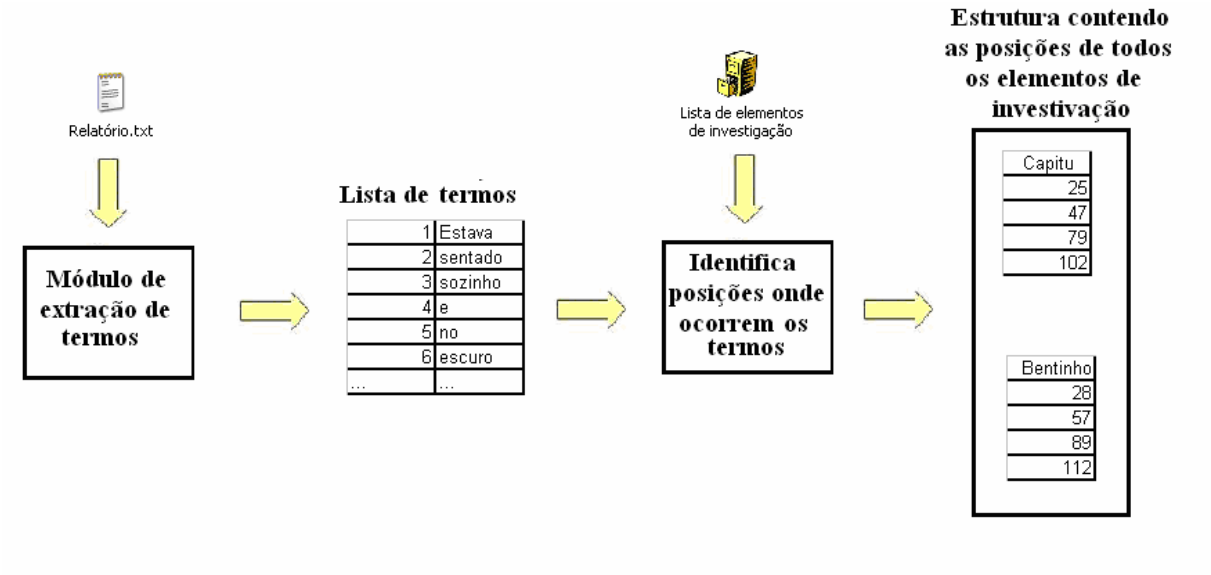


Figura 7. Detalhes de implementação do algoritmo

O módulo de extração de termos tem a finalidade de transformar o relatório policial, que é um arquivo de texto puro, em um *array* de *string*, removendo a pontuação e outros elementos que não são considerados palavras.

O módulo que identifica onde a ocorrência dos elementos de investigação recebe uma lista de palavras e uma lista de elementos de investigação. Os objetos do tipo elemento de investigação contêm um atributo de lista encadeada de inteiros, os números representam a posição de ocorrência das palavras associadas ao elemento de investigação. Este módulo atualiza as listas de ocorrências. Esta parte do algoritmo é a mais demorada, pois é necessário realizar comparação entre as

palavras. A saída desta fase produz uma estrutura que permite que o cálculo entre os pesos dos vários elementos de investigação seja realizado rapidamente.

A Figura 8 mostra como foi realizado o cálculo do peso de relacionamento dos elementos de investigação. A distância textual entre dois elementos é definida como o módulo da diferença entre as posições das palavras.

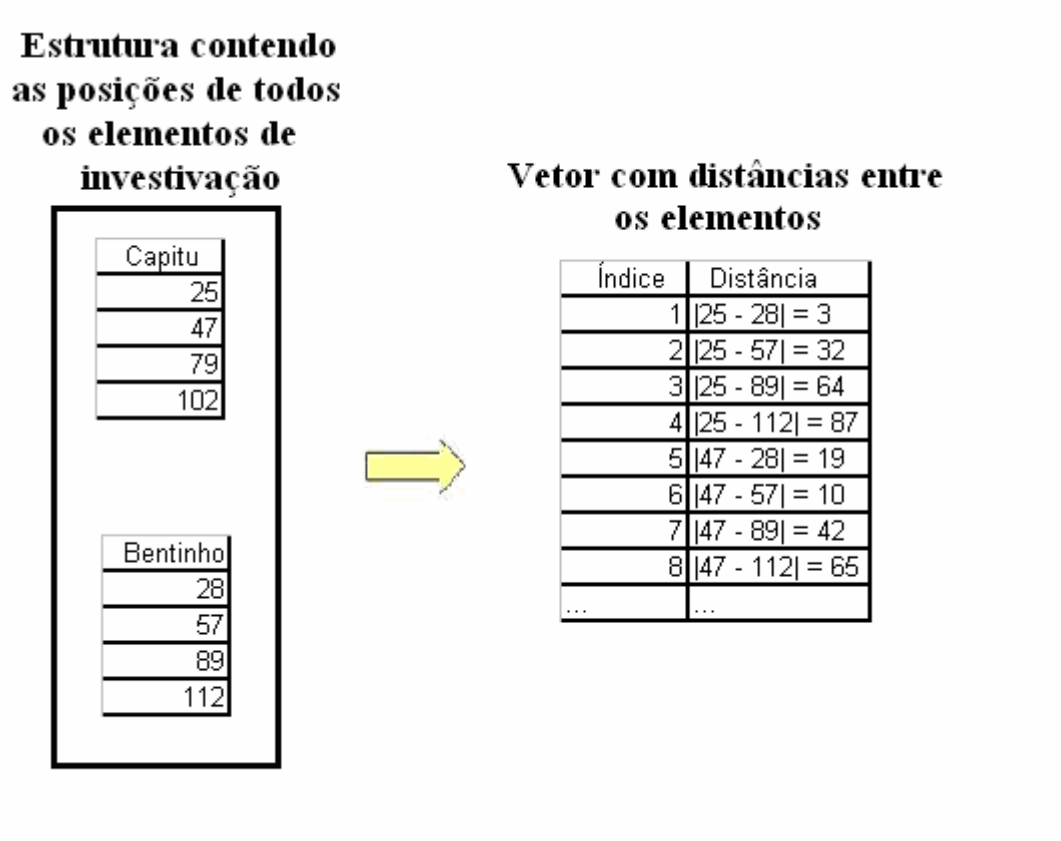


Figura 8. Cálculo do vetor de distâncias

Para calcular o peso da relação entre dois elementos é preciso aplicar a função exponencial exp(-x) a cada elemento do vetor de distâncias e realizar uma soma.

Capítulo 3

Estudo de caso

Neste capítulo, será apresentado um estudo de caso para a avaliação da ferramenta de apoio à investigação desenvolvida. Para fins de avaliação da ferramenta, serão abordados livros clássicos da literatura, pois são de conhecimento público e permitem que o leitor possa analisá-los com clareza.

3.1 Caso 1 - Livro Dom Casmurro

3.1.1 Dados sobre o autor e obra

Autor: Machado de Assis

Título da Obra: Dom Casmurro

Editora: Ática

Edição: 32

3.1.2 Razões para utilizar este texto

Este livro foi analisado neste estudo de caso porque apresenta um mistério em seu enredo. A conclusão do livro não define de maneira clara o que realmente aconteceu na história, cabendo ao leitor interpretar a história e tirar suas próprias conclusões.

Este pode ser analisado para fins de investigação, pois apresenta semelhanças com textos que a ferramenta de estudo deste trabalho se propõe a investigar.

Esta obra é escrita em primeira pessoa, mostra a visão do autor sobre um possível caso de adultério envolvendo os personagens do livro. Espera-se que o sistema de apoio à investigação forneça informações sobre o grau de ligação entre os personagens do livro. Com base nas informações relativas ao grau de relação entre

as pessoas, o leitor do livro poderá ter uma visão mais clara do que aconteceu de fato.

3.1.3 Personagens

Os personagens expostos neste livro serão classificados como alvos e conexões. Os alvos serão personagens que estão no centro das investigações.

- Capitu
- Bentinho
- Escobar
- Ezequiel

3.1.4 Resumo da investigação

O estudo de caso destina-se a investigar um possível caso de traição entre os personagens Capitu e Escobar. No decorrer da história, Capitu casou-se com Bentinho e eles tiveram um filho chamado Ezequiel. Bentinho começa a perceber que o seu filho Ezequiel era muito parecido com seu amigo Escobar. Bentinho começa imaginar que seu melhor amigo teve um caso com a sua esposa, e que o seu filho era fruto desta relação extraconjugal.

A investigação concentra-se em determinar se houve ou não um caso de adultério nesta história. A história foi narrada em primeira pessoa pelo próprio autor. A visão da história é parcial e mostra o que o autor percebeu da história. No final, o livro não chega a uma conclusão sobre o acontecido, deixando o leitor tirar as suas próprias conclusões.

3.1.5 Elementos de investigação

Através do resumo do livro, pode-se concluir que os alvos da investigação são Capitu e Escobar, pois estes são os suspeitos de terem tido um relacionamento extraconjugal. Bentinho e seu filho Ezequiel são classificados como conexões, pois estes não participaram do delito diretamente, mas tem ligações com os alvos.

3.1.6 Resultados

O livro Dom Casmurro foi dividido em três partes para que se possa ter uma noção da evolução da rede de relacionamentos durante a apresentação dos capítulos.

A primeira parte é a que tem início no capítulo I e segue até o capítulo XLI. A Figura 9 abaixo mostra a rede de relacionamentos e a Figura 10 mostra a representação de forma matricial da rede de relacionamentos. A Figura 11 mostra a rede de relacionamentos normalizada.

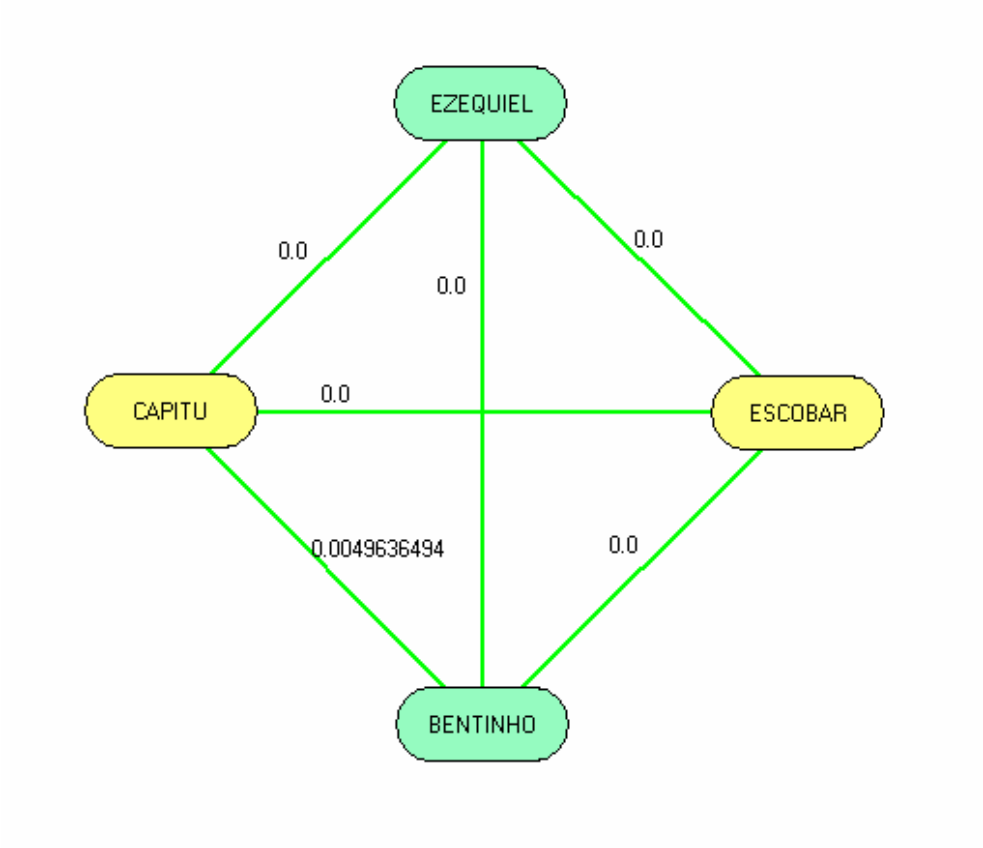


Figura 9. Rede de relacionamento parte 1, livro Dom Casmurro.

A	Capitu	Escobar	Bentinho	Ezequiel
Capitu	0.0	0.0	0.0049636494	0.0
Escobar	0.0	0.0	0.0	0.0
Bentinho	0.0049636494	0.0	0.0	0.0
Ezequiel	0.0	0.0	0.0	0.0

Figura 10. Rede de relacionamentos matricial parte 1, livro Dom Casmurro.

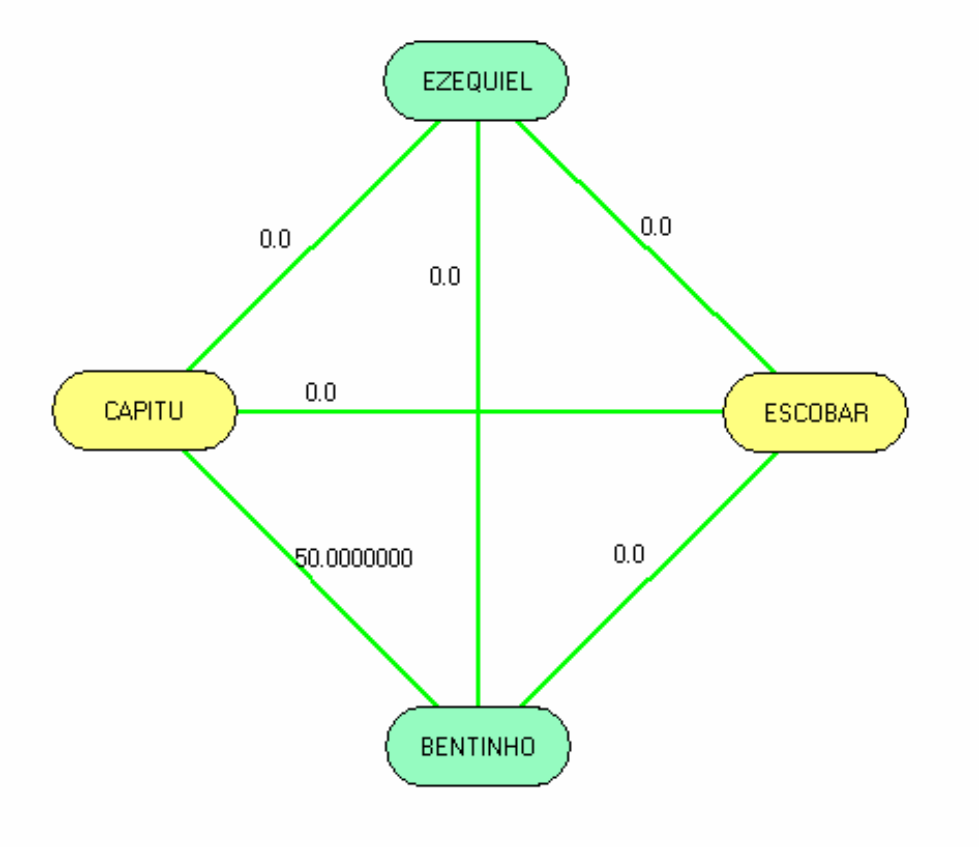


Figura 11. Rede de relacionamentos normalizada parte 1, livro Dom Casmurro.

A segunda parte tem inicio no capítulo XLII e segue até o capítulo LXXXIV. A Figura 12 abaixo mostra a rede de relacionamentos e a Figura 13 mostra a representação em uma forma matricial da rede de relacionamentos. A Figura 14 mostra a rede normalizada.

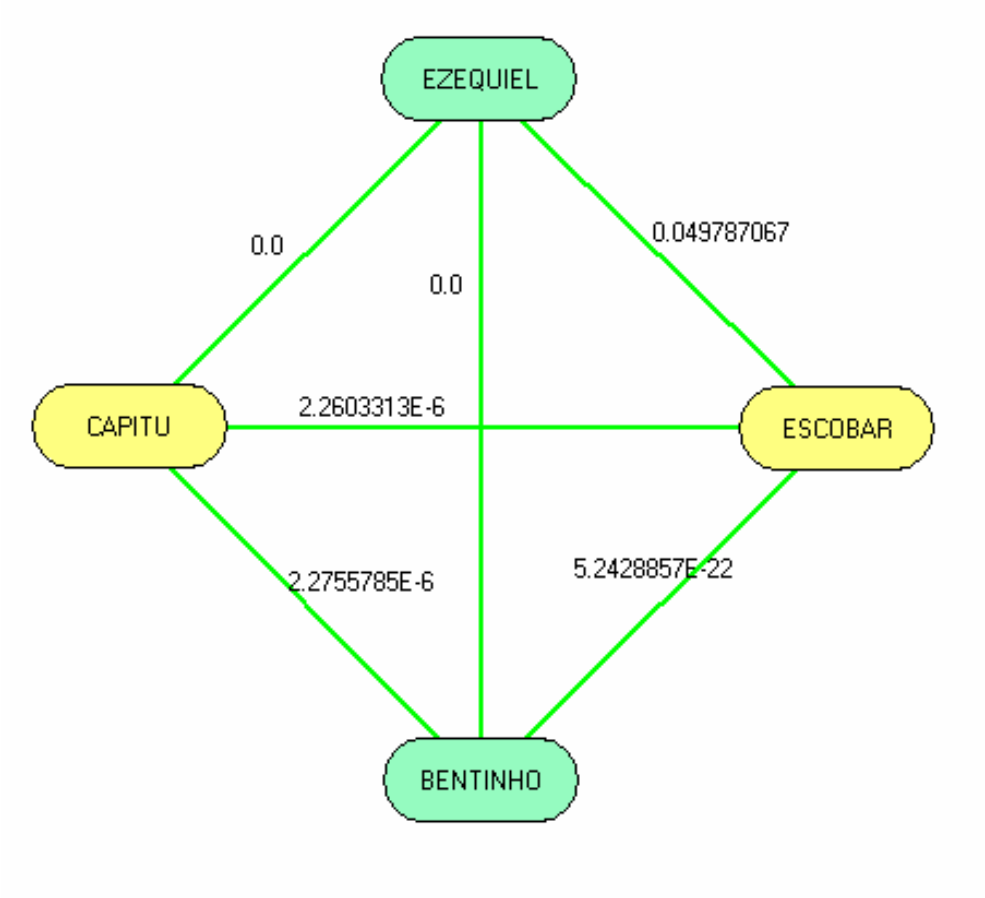


Figura 12. Rede de relacionamento parte 2, livro Dom Casmurro.

A	Capitu	Escobar	Bentinho	Ezequiel
Capitu	0.0	2.2603313E-6	2.2755785E-6	0.0
Escobar	2.2603313E-6	0.0	5.2428857E-22	0.049787067
Bentinho	2.2755785E-6	5.2428857E-22	0.0	0.0
Ezequiel	0.0	0.049787067	0.0	0.0

Figura 13. Rede de relacionamentos matricial parte 2, livro Dom Casmurro.

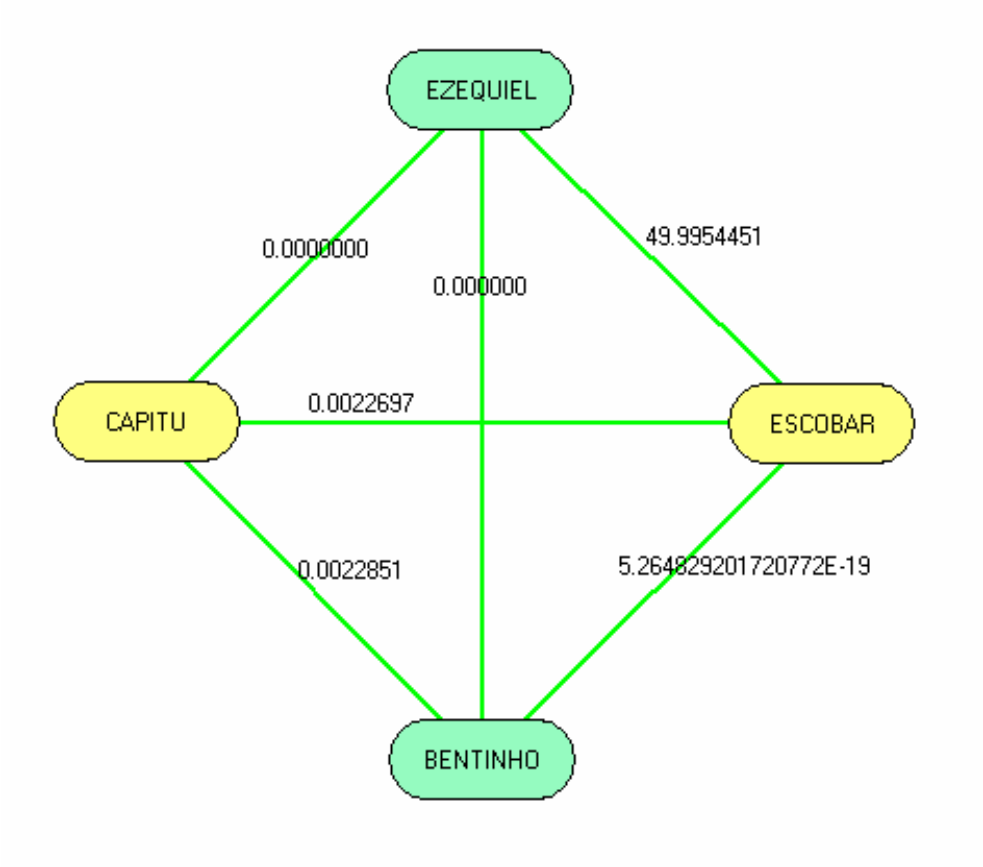


Figura 14. Rede de relacionamentos normalizada parte 2, livro Dom Casmurro.

A terceira parte começa no capítulo LXXV se estende até o capítulo XLVIII. O resultado da aplicação da ferramenta será mostrado nas Figuras 15 e 16. A Figura 17 mostra a rede normalizada.

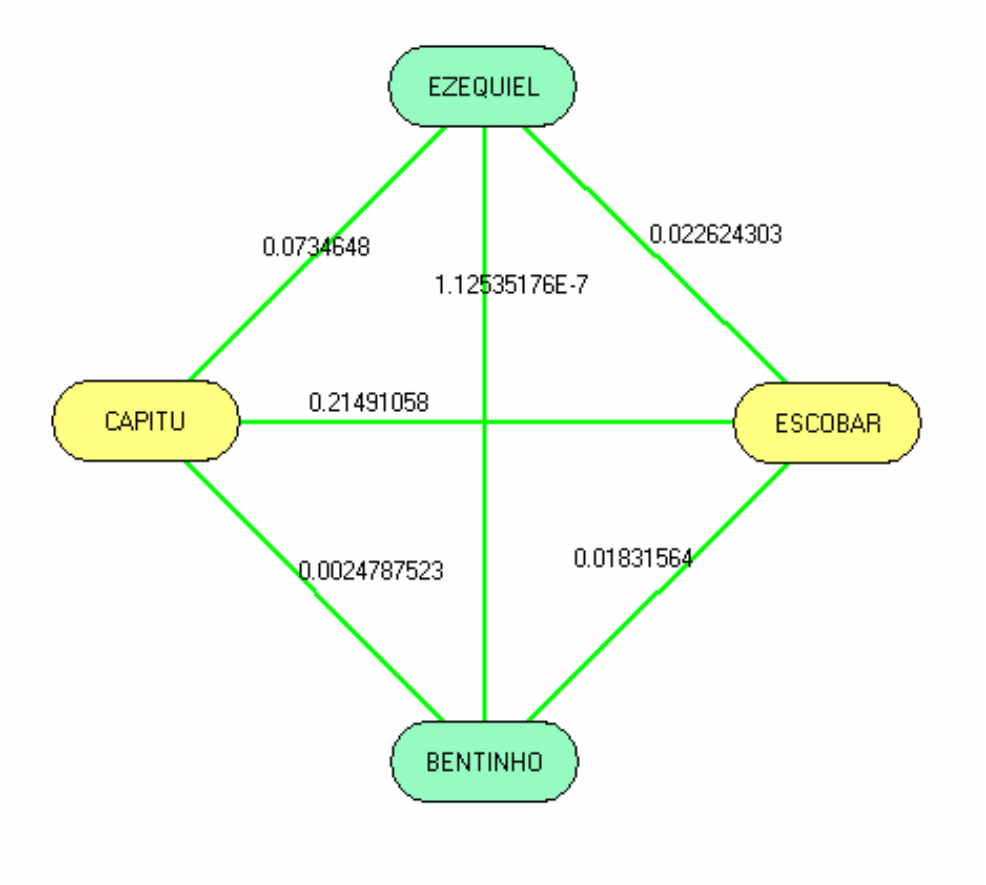


Figura 15. Rede de relacionamento parte 3, livro Dom Casmurro.

A	Capitu	Escobar	Bentinho	Ezequiel
Capitu	0.0	0.21491058	0.0024787523	0.0734648
Escobar	0.21491058	0.0	0.01831564	0.022624303
Bentinho	0.0024787523	0.01831564	0.0	1.12535176E-7
Ezequiel	0.0734648	0.022624303	1.12535176E-7	0.0

Figura 16. Rede de relacionamentos matricial parte 3, livro Dom Casmurro.

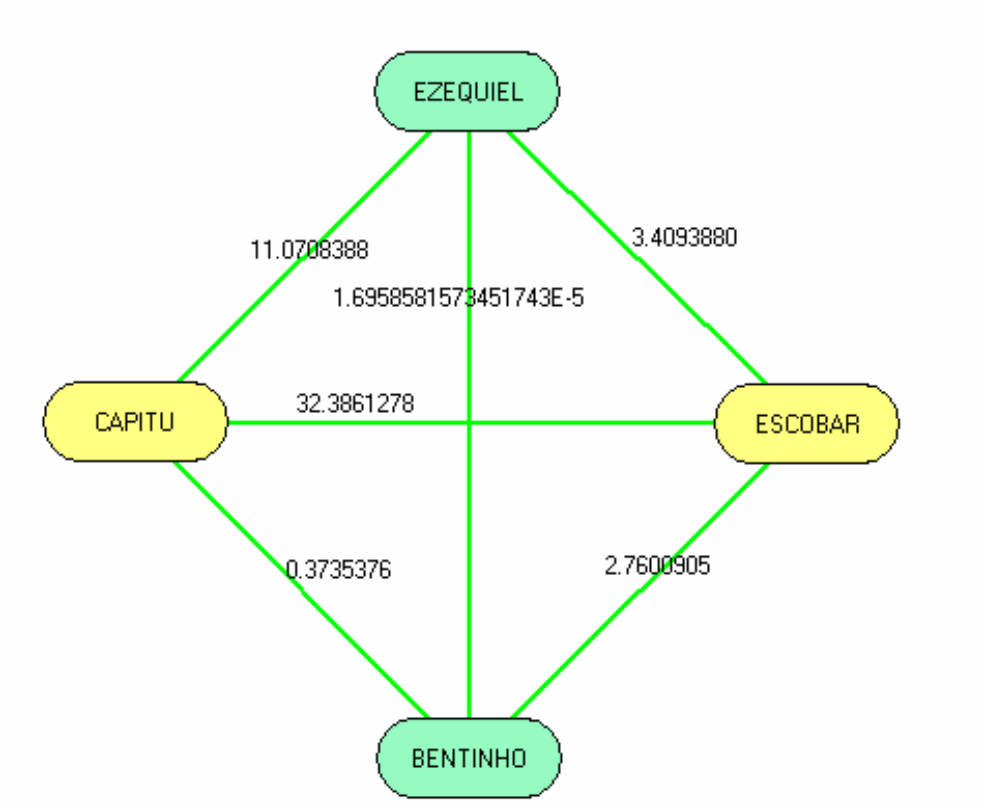


Figura 17. Rede de relacionamentos normalizada parte 3, livro Dom Casmurro.

A Figura 18 mostra o resultado da aplicação da ferramenta utilizando o livro inteiro com os alvos e conexões que foram descritos na subseção anterior.

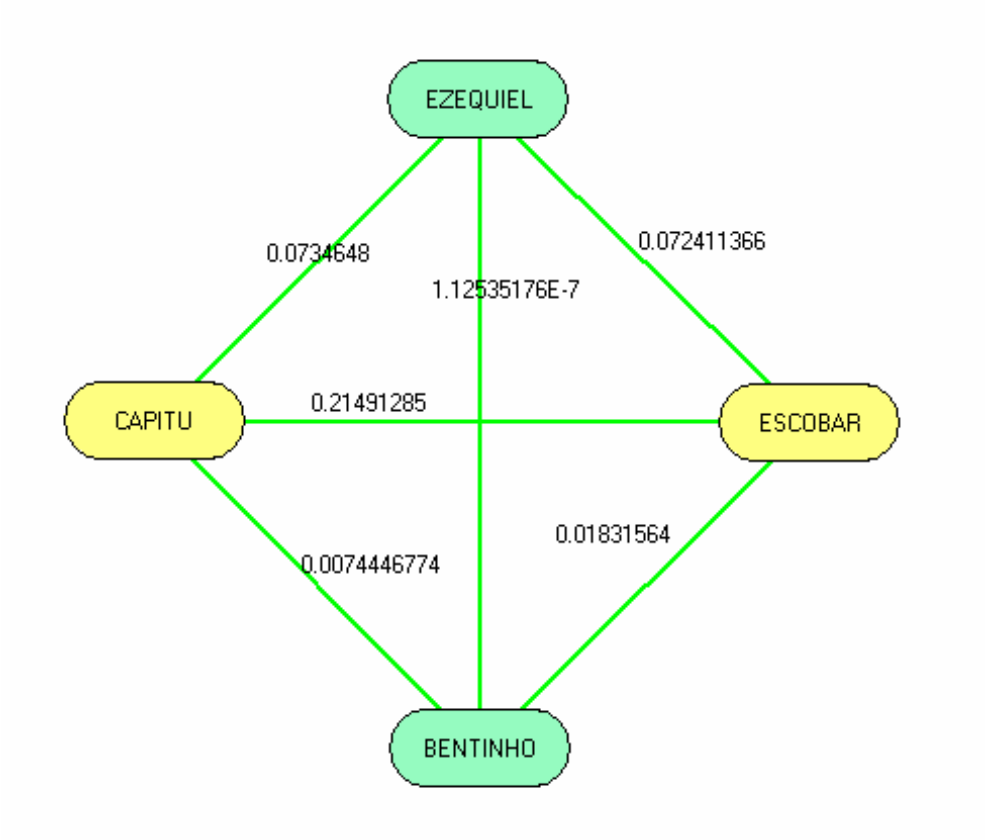


Figura 18. Rede de relacionamentos resultante do livro Dom Casmurro.

A representação matricial será mostrada na Figura 19. Esta representação é uma matriz simétrica que contem o peso associado a todos os elementos de investigação envolvidos. A Figura 20 mostra a rede normalizada.

A	Capitu	Escobar	Bentinho	Ezequiel
Capitu	0.0	0.21491285	0.0074446774	0.0734648
Escobar	0.21491285	0.0	0.01831564	0.072411366
Bentinho	0.0074446774	0.01831564	0.0	1.12535176E-7
Ezequiel	0.0734648	0.072411366	1.12535176E-7	0.0

Figura 19. Representação matricial da rede de relacionamentos caso 1

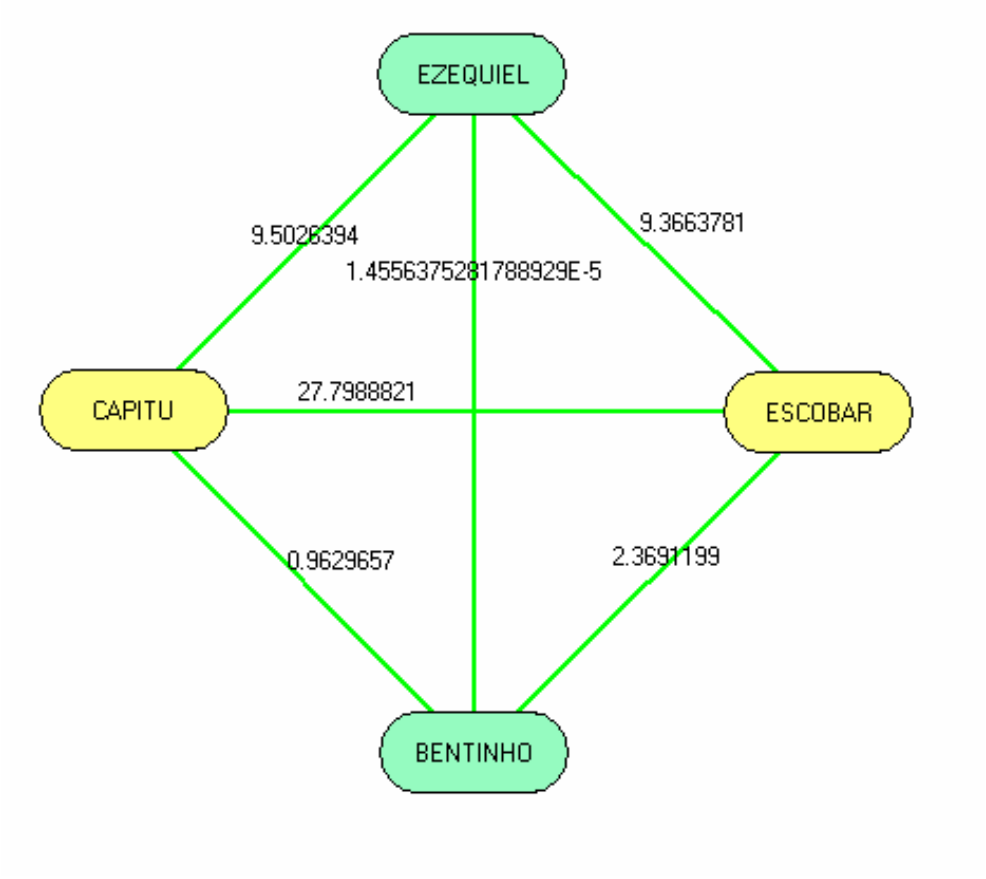


Figura 20. Representação normalizada da rede de relacionamentos caso 1

3.1.7 Conclusão

Com base na análise dos resultados obtidos através da aplicação da ferramenta, é possível inferir que os alvos da investigação têm um grau de relacionamento relativamente acentuado, na visão do autor do livro.

O peso do relacionamento entre Capitu e Escobar é cerca de oito vezes maior que o peso entre o seu marido Bentinho. Com base nestas informações pode-se concluir que a ferramenta pode identificar uma visão tendenciosa do personagem Bentinho sobre a possível traição de sua esposa Capitu.

Quanto à questão da paternidade de Ezequiel, a ferramenta aponta que o peso da relação entre Ezequiel e Escobar é superior ao peso da relação entre Ezequiel e Bentinho. Outro ponto importante é o fato de que o peso da relação entre Capitu e seu filho Ezequiel é bastante próximo ao valor associado ao peso entre Escobar e Ezequiel, o que aponta uma simetria de pai e mãe. O autor Machado de

Assis soube construir uma obra verossímil, que mostra na visão tendenciosa de Bentinho que a relação entre Ezequiel e seus possíveis pais acontece de maneira equilibrada.

3.2 Caso 2 – Livro A herdeira

3.2.1 Dados sobre o autor e obra

Autor: Sidney Sheldon

Título da Obra: A Herdeira

Editora: RECORD.

Ano de Edição: 1999

3.2.2 Razões para utilizar este texto

Este livro é um romance policial que apresenta uma história que parte de um assassinato e tem uma história envolvida em mistério. No decorrer da história, o autor apresenta os suspeitos e uma seqüência de acontecimentos que levam o leitor a idealizar quem realmente é o culpado pelos crimes. E livro visa emular o ambiente de uma investigação policial, pois o autor só revela o culpado no final do livro.

O autor desta obra consegue apresentar os fatos ao leitor de maneira imparcial, pois os acontecimentos são simplesmente expostos sem uma tendência definida. Esta característica do autor é semelhante aos relatórios policiais, que visam ser imparciais.

3.2.3 Personagens

- Elizabeth Roffe
- Sam Roffe
- Charles Martin

- Ivo Pallazi
- Alec Nichols
- Rhys William

3.2.4 Resumo da investigação

O livro começa com um acidente durante uma eslacalada a uma montanha que termina na morte Sam Roffe, um bilionário que era presidente da industria farmacêutica Roffe & Sons. Elizabeth, filha de Sam, herda um império bilionário com morte de seu pai. Elizabeth assume o controle das empresas de seu pai, contrariando sócios da empresa que tinham a intenção de tornar a empresa de capital aberto. Uma investigação realizada pelo inspetor Max Hornung, deduz que Sam havia sido assassinado. As suspeitas caem sobre os primos de Sam. Em uma viagem à Europa, Elizabeth sofre atentados contra a sua vida e torna-se um alvo do criminoso.

Todos aqueles que têm interesse que a empresa torne-se uma sociedade anônima tornam-se suspeitos de ter cometido o crime. Enquanto o inspetor Max investiga o caso, mais ataques são feitos à Elizabeth, mas mesmo assim ela continua com a presidência da empresa. Se Elizabeth morresse, a decisão de abrir capital da empresa seria unânime. Todos querem o dinheiro que seria levantado com a venda das ações. Enquanto isto, um homem se relacionava com prostitutas e colocava uma fita vermelha em seu pescoço. Esta fita era utilizava para sufocar as suas vítimas. Toda a cena era gravada. O inspetor acredita que existe uma ligação entre o assassino em série e a morte de Sam.

3.2.5 Elementos de investigação

Os suspeitos de terem cometido o assassinato de Sam são os primos Charles Martin, Ivo Pallazi e Alec Nichols, porque estes estavam sendo ameaçados e precisavam do dinheiro para pagar dívidas. Rhys William também é classificado como alvo porque sabia de um segredo industrial da empresa que poderia gerar lucros, caso a empresa fosse negociada no mercado de ações.

Ivo Pallazi é um italiano, primo de Sam Roffe e casado com Simonetta Pallazi. Ivo tem três filhos com uma amante chamada Donatella. Ivo está sendo chantageado por Donatella. Caso ele não consiga o dinheiro, ela ameaça contar a tudo para Simonetta. Ivo ficou contente ao saber da morte de Sam, pois finalmente ele conseguiria o dinheiro para pagar.

Charles Martin é casado com Hélène Martin. Charles estava sendo pressionado por causa de uma dívida. Em um momento de desespero, Charles chegou a trocar uma jóia da esposa por uma jóia falsa para levantar dinheiro.

Alec Nichols é casado com Viviam Nichols e também estava sendo pressionado por causa de dinheiro. Sua esposa era bem mais jovem que ele. Sua esposa ficou sabendo que Alec não tinha dinheiro para pagar uma dívida, e disse que ele estava com problemas de caixa.

3.2.6 Resultados

A Figura 21 mostra o resultado da aplicação da ferramenta em todo o livro com a lista de suspeitos Rhys Williams, Ivo Pallazi, Charles Martin e Alec Nichols. A representação matricial da rede de relacionamentos é mostrada na Figura 22. A Figura 23 mostra a rede normalizada.

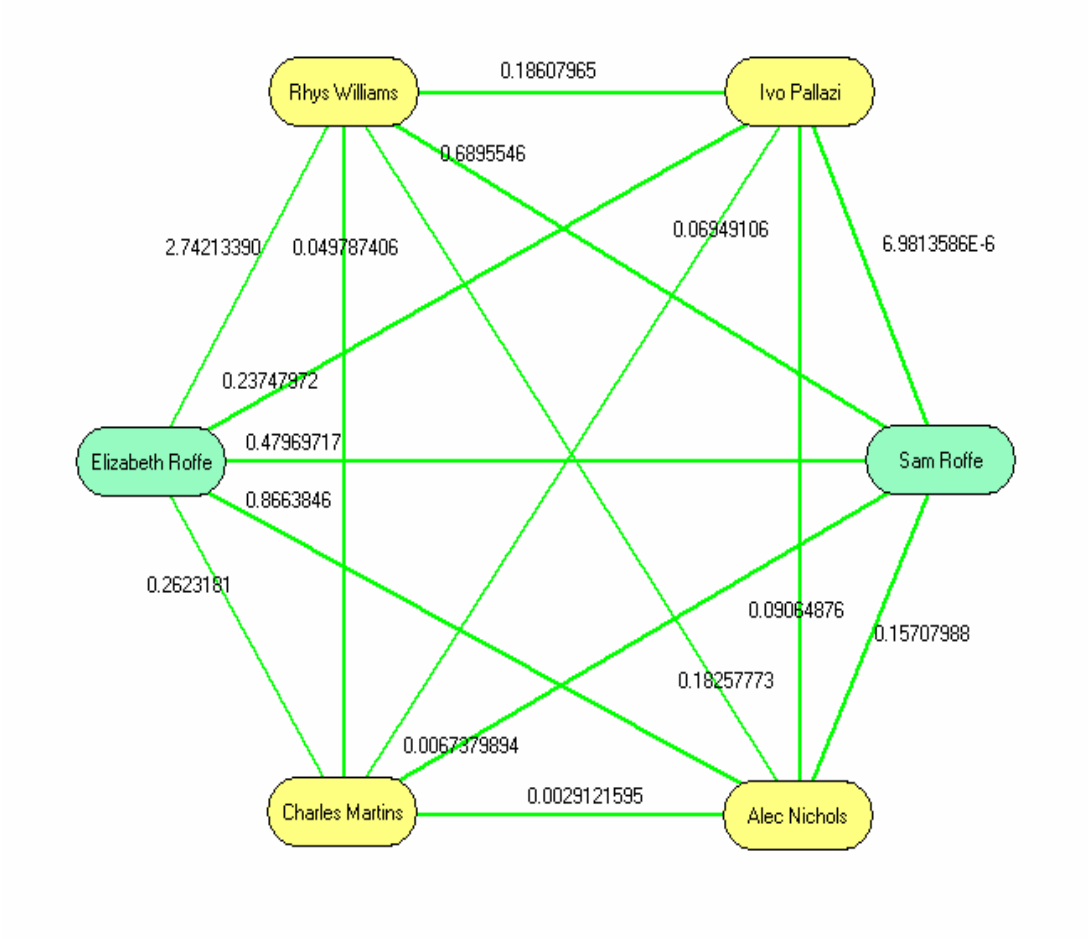


Figura 21. Rede de relacionamentos caso 2, livro completo.

null	Sam	Elizabeth	Rhys	Ivo	Charles	Alec
Sam	0.0	0.47969717	0.6895546	6.9813586E-6	0.0067379894	0.15707988
Elizabeth	0.47969717	0.0	2.7421339	0.23747972	0.2623181	0.8663846
Rhys	0.6895546	2.7421339	0.0	0.18607965	0.049787406	0.18257773
Ivo	6.9813586E-6	0.23747972	0.18607965	0.0	0.06949106	0.09064876
Charles	0.0067379894	0.2623181	0.049787406	0.06949106	0.0	0.0029121595
Alec	0.15707988	0.8663846	0.18257773	0.09064876	0.0029121595	0.0

Figura 22. Representação matricial da rede de relacionamentos caso 2, livro completo.

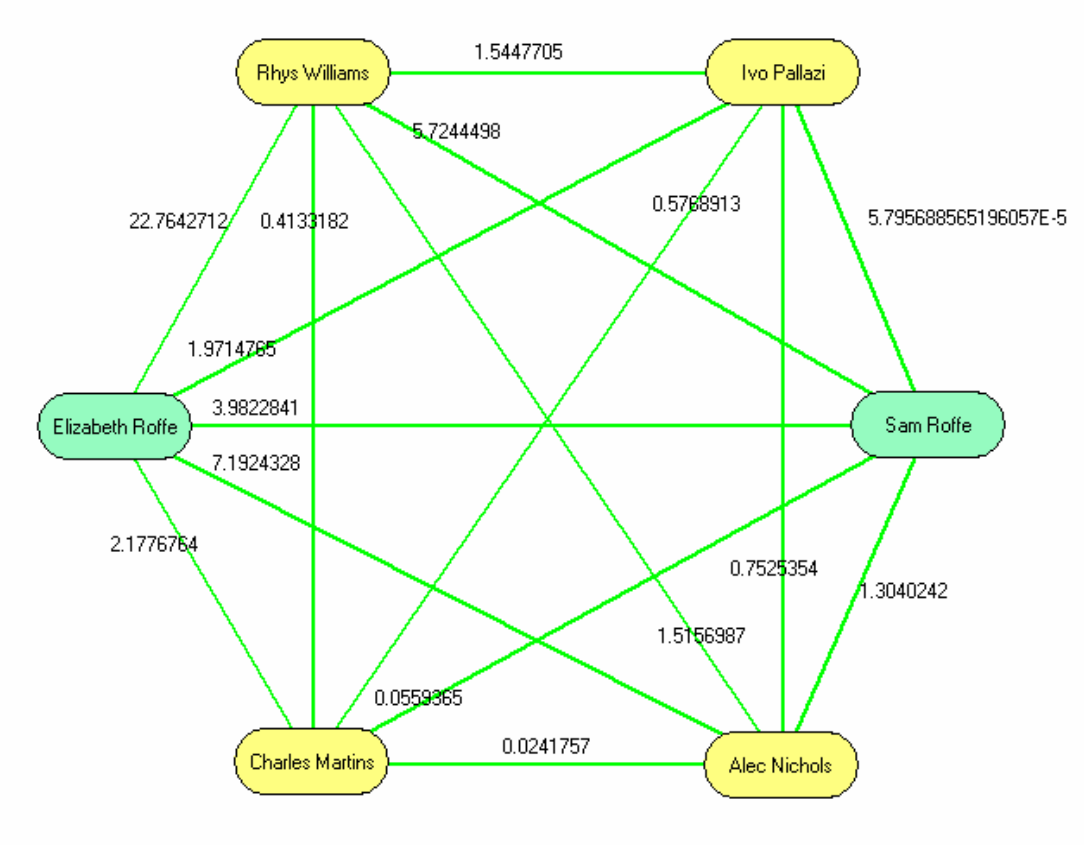


Figura 23. Representação normalizada da rede de relacionamentos caso 2, livro completo.

3.2.7 Conclusão

Através da análise da rede de relacionamentos apresentada na Figura 23, pode-se concluir que a ligação entre Elizabeth e Rhys é bastante forte. Isto pode ser justificado porque existia uma relação afetiva entre eles e conseqüentemente uma maior interação.

As três pessoas que têm uma maior relação com Sam Roffe são Elizabeth, Rhys e Alec. Eliminando-se Elizabeth e seu amante Rhys, temos Alec como um forte candidato a suspeito de ter praticado o crime, devido à forte ligação com Sam.

Alec Nichols é um dos suspeitos que possui uma relação mais acentuada com Elizabeth e foi culpado pela morte de Sam. Devido a esta relação mais intensa com Elizabeth, ele teria mais informações sobre onde ela estaria. Isto poderia explicar os

atentados sofridos, a sabotagem ao carro de Elizabeth e atentados sofridos em uma viagem à Europa.

A personagem Elizabeth é o elemento central da história, porque possui uma relação mais acentuada com todos os outros personagens. Este fato pode ser comprovado através da leitura da obra completa.

A inclusão de eventos na geração da rede de relacionamentos pode ajudar na resolução do caso. Seria possível investigar a relação entre as pessoas e eventos principais, permitindo que o analista tenha uma visualização da associação entre os personagens e os eventos.

3.3 Caso 3 – Depoimento de testemunha à polícia

3.3.1 Dados sobre o caso

Tipo de documento: Depoimento de vítima à delegacia

Data: 24 de outubro de 2008

O documento foi obtido através do site do jornal Estadão em 17 de maio de 2009, no endereço eletrônico <http://www.estadao.com.br/noticias/cidades,leia-a-integra-do-depoimento-de-nayara-sobre-o-caso-eloa,265677,0.htm>. A matéria, que contem o depoimento na integra, foi convertida para o formato de arquivo de texto puro.

3.3.2 Razões para utilizar este texto

Este documento apresenta a versão da testemunha Nayara no caso de um seqüestro que aconteceu em Santo André, no estado de São Paulo. O texto foi escrito em terceira pessoa e mostra a visão de uma testemunha que esteve sob a custódia de um seqüestrador e que sobreviveu.

Espera-se que a ferramenta aplicada a este documento mostre o peso da relação entre os envolvidos no crime. Deve-se observar que o texto trata-se de um depoimento que foi feito por uma vítima do seqüestro, e que logo tem uma opinião parcial sobre o caso.

Este documento é um depoimento de um caso real de seqüestro, logo a ferramenta poderá ser testada com dados reais de um inquérito policial. Este caso poderá mostrar uma visão de como esta ferramenta pode auxiliar no processo de investigação policial.

3.3.3 Personagens

Os envolvidos no caso em questão serão classificados como alvos e conexões. Os alvos serão personagens que estão no centro das investigações.

- Lindemberg
- Eloá
- Nayara

3.3.4 Resumo da investigação

O seqüestro de Eloá Pimentel foi o caso mais longo de cárcere privado registrado pela polícia brasileira. O material utilizado para a análise pela ferramenta foi o depoimento da sobrevivente do seqüestro, Nayara, que era amiga da vítima.

O seqüestro teve inicio em 13 de outubro de 2008 e se estendeu até o dia 18 de outubro. Lindemberg Fernandes invadiu o apartamento de Eloá Pimentel e a manteve refém juntamente com sua amiga Nayara em um seqüestro que durou vários dias.

O seqüestro teve fim com a morte de Eloá. Nayara, a amiga de Eloá, foi baleada mas sobreviveu. Alguns dias depois, Nayara prestou depoimento à delegacia informando o que aconteceu durante o seqüestro.

3.3.5 Elementos de investigação

De acordo com dados referentes à investigação, pode-se concluir que o responsável pelo crime é Lindemberg, que receberá a classificação de Alvo. As vítimas, Eloá e Nayara, serão classificadas como conexões.

3.3.6 Resultados

A Figura 24 abaixo, mostra a rede de relacionamentos obtida através da aplicação da ferramenta em todo o depoimento prestado por Nayara à polícia. A Figura 25 mostra a representação matricial da rede obtida diretamente da aplicação. A Figura 26 mostra a rede normalizada.

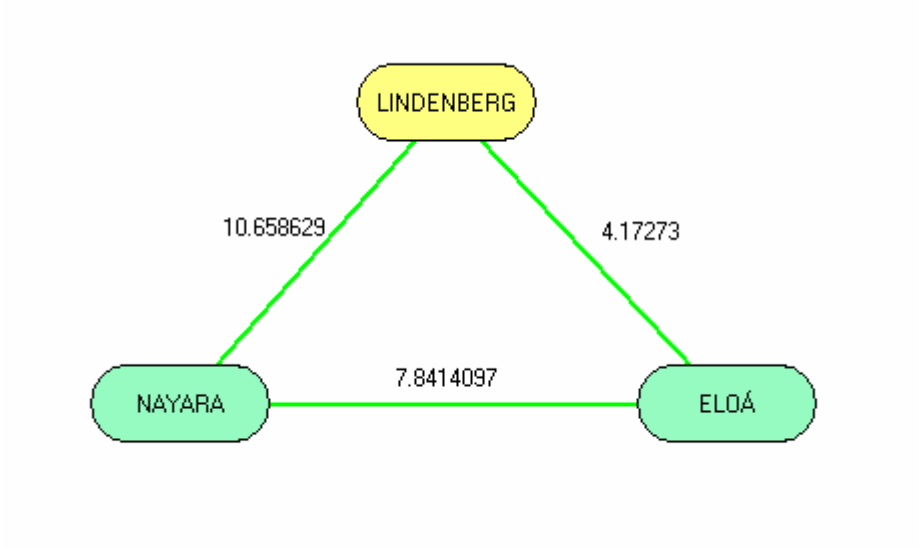


Figura 24. Rede de relacionamentos caso 3, depoimento completo.

A	Nayara	Eloá	Lindemberg
Nayara	0.0	7.8414097	10.658629
Eloá	7.8414097	0.0	4.17273
Lindemberg	10.658629	4.17273	0.0

Figura 25. Representação matricial da rede de relacionamentos do caso 3.

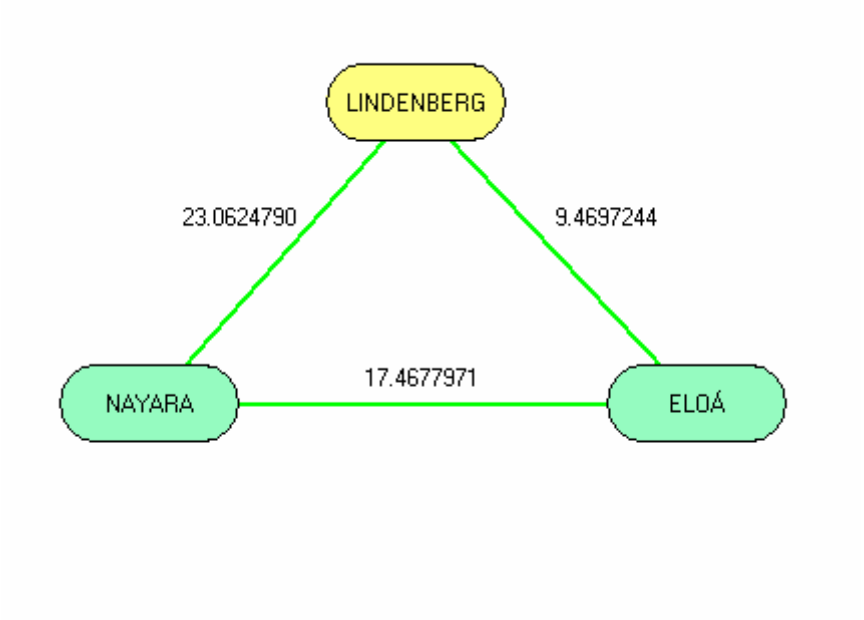


Figura 26. Representação normalizada da rede de relacionamentos do caso 3.

3.3.7 Conclusão

Com base na rede de relacionamentos, pode-se concluir que Nayara tinha uma ligação muito forte com o seqüestrador Lindemberg. Este fato pode ser justificado porque no depoimento apresentado por Nayara, ela conta que interagiu diversas vezes com o seqüestrador.

Comparando o peso das relações relativos a Eloá, percebe-se que ela tinha uma maior relação com Nayara. O peso da relação entre Eloá e Lindemberg é inferior a todos os outros pesos, o que mostra a fragilidade na relação entre os dois. Este foi o motivo pelo qual o seqüestrador estava praticando o crime.

A aplicação desta ferramenta em um caso de investigação real pôde mostrar o funcionamento do sistema com fontes seguras de informação.

Capítulo 4

Conclusão e trabalhos futuros

Este capítulo é dedicado à apresentação da conclusão do trabalho desta monografia.

4.1 Conclusões gerais

O sistema de apoio à investigação pode ajudar a reduzir o tempo gasto em processos de investigação criminalista, proporcionar ao analista uma visão complementar da relação entre os elementos envolvidos em determinado crime e consequentemente reduzir os custos associados à investigação.

A utilização de ferramentas computacionais para combater o crime é de fundamental importância para uma sociedade. Os crimes estão se tornando cada vez mais complexos e envolvendo uma maior quantidade de pessoas, o que justifica a utilização de meios computacionais para auxiliar no processo investigativo.

As ferramentas de mineração de texto é uma área de conhecimento que tem um grande potencial de desenvolvimento, visto que cerca de 80% das informações armazenadas em computadores estão sob a forma de texto não estruturado [2]. A extração de conhecimento de maneira automática a partir deste tipo de base de dados é uma ótima fonte de informação devido a quantidade de informação disponibilizada.

A construção do algoritmo que foi desenvolvido para calcular o peso entre os elementos de investigação foi baseada em conceitos de técnicas de mineração de textos, como a mineração de textos em nível de termos. Estudos realizados sobre o estado da arte em mineração de textos levaram à conclusão de que a utilização de processamento de linguagem natural não é uma boa alternativa para o cálculo dos pesos na rede de relacionamentos, pois o tempo necessário para a avaliação dos documentos de texto é significativamente alto. A aplicação se propõe a processar

um alto volume de dados, o que justifica a utilização de um sistema baseado em regras.

A ferramenta desenvolvida ao longo deste trabalho tem a finalidade de facilitar o processo de investigação e prover aos analistas um sistema que ofereça uma visão sobre os casos que estão sendo analisados. O sistema pode atuar em qualquer tipo de fonte de informação que possa ser convertida em texto. Isto permite que a ferramenta seja utilizada em várias aplicações, pois grande parte das informações armazenadas pode ser convertida em texto.

Um analista poderá gerar uma rede de relacionamentos ponderada entre os elementos envolvidos em uma investigação a partir de documentos de texto de qualquer espécie e de uma lista contendo alvos e conexões. A visualização gráfica da rede de relacionamentos, como mostrado no capítulo anterior, possibilita uma melhor forma de visualizar as relações entre os envolvidos em um crime.

Os resultados obtidos através dos capítulos do livro Dom Casmurro puderam emular os relatórios gerados em uma investigação criminalista. A segmentação do livro em três partes pôde evidenciar a evolução da rede de relacionamentos ao longo do tempo. Isto permite que o analista tire conclusões sobre os capítulos isoladamente, até a formação da rede de relacionamentos que contempla todo o livro.

Os resultados obtidos mostram que a aplicação desenvolvida pode ser bastante útil para auxiliar as investigações, pois otimiza a extração de informações a partir de práticas já existentes.

O sistema de apoio à investigação não tem o objetivo de substituir o trabalho do analista. A ferramenta tem a vantagem de ser dependente da análise de pessoas, o que permite que as decisões tomadas levem em consideração fatores que as máquinas não podem identificar. Dessa forma, o sistema dará um suporte para a tomada de decisões.

4.2 Trabalhos futuros

O sistema poderá ser aprimorado em alguns aspectos para oferecer mais robustez e comodidade para o analista. As subseções descritas abaixo, mostram exemplos de aperfeiçoamentos sugeridos para um melhor aproveitamento da ferramenta.

4.2.1 Modelagem em banco de dados

A utilização de banco de dados para armazenamento dos dados produzidos durante o processo de investigação poderá tornar a aplicação mais robusta. Gerenciadores de bancos de dados incorporam técnicas sofisticadas para armazenamento de dados, o que permite uma maior confiabilidade ao sistema.

O sistema desenvolvido utiliza o armazenamento de seus dados em arquivos de texto e em arquivos no formato CSV que é um padrão adotado para armazenamento de estruturas em formas de tabelas.

4.2.2 Geração automática do gráfico das redes de relacionamentos

A aplicação limita-se a gerar a representação matricial da rede de relacionamentos. A visualização gráfica permite uma maior clareza para o analista e pode ser apresentada no próprio ambiente da aplicação.

4.2.3 Inclusão de dicionários de pessoas e eventos

Realizar um pré-processamento do texto para segmentá-lo em termos ajudará a ferramenta a criar uma identificação única para cada entidade envolvida. A ferramenta está limitada a reconhecer identificadores compostos por uma única palavra.

4.2.4 Portar a solução para a *web*

Os investigadores poderiam acessar dados referentes a investigações em qualquer lugar que houvesse conectividade com a internet. O sistema pode aceitar vários formatos de arquivos disponíveis no mercado e permitir o envio de relatórios através da rede.

Bibliografia

[1] Carlos, J.A. **O crime segundo a perspectiva de Durkheim**. Disponível em:< http://www.buscalegis.ufsc.br/revistas/index.php/buscalegis/article/viewFile/10554/10119>. Data de acesso: 3 de abril de 2009.

[2] Tan, A.H. Text mining: The state of the art and challenges. **Proceedings of the PAKDD 1999 Workshop on Knowledge,** 1999, Singapore City, Singapore.

[3] Feldman, R.; Fresko, M.; Kinar, Y.; Lindell, Y.; Liphstat, O.; Rajman, M; Schler, Y e Zamir, O. Text mining at the term level. **Lecture Notes In Computer Science**, 1998, Seattle, USA.

[4] Hearst, M. **What is text mining**? Disponível em: < http://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf>. Data de acesso 22 de abril de 2009.

[5] Frawley, W. J.; Piatetsky-Shapiro, G.**;** Matheus, C.J. Knowledge Discovery in Databases: An Overview. **AI Magazine,** 13, 3, 03. 1992. <http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011/929>. Data de acesso: 11 de abril de 2009.