4
Extração de Entidades e Modelagem de Associações

Este capítulo trata da extração de entidades úteis de históricos policiais e de sua aplicação na modelagem de cenários representados por grafos direcionados. As entidades extraídas dos textos representam os vértices e seus relacionamentos os arcos direcionados de um grafo. O valor das associações entre os vértices do grafo é calculado através de um método de co-ocorrências que pesquisa relacionamentos entre entidades extraídas e de seus relacionamentos com os documentos tratados em uma coleção.

O capítulo encontra-se organizado na seguinte seqüência: referencial teórico, método, algoritmos e sistemas utilizados.

4.1
Referencial Teórico

As atividades criminosas, gradualmente, crescem em sofisticação, tecnologia e planejamento. Empregando recursos e métodos tecnologicamente avançados os criminosos conectam-se em redes de relacionamento e utilizam sistemas de comunicação modernos, como Internet, telefonia e rádio. Crimes relacionados com fraudes eletrônicas, tráfico de drogas ou lavagem de dinheiro tornam a investigação ou rastreamento de quadrilhas uma atividade de inteligência policial complexa e sofisticada.

Analistas e investigadores procuram produzir conhecimento criminal a partir de informações pesquisadas em volumosas bases de ocorrências policiais (Hauck et al., 2002). O maior desafio para os analistas, investigadores e departamentos de inteligência policial é alcançar eficiência e exatidão diante da crescente demanda de dados não tratados nas bases de informações criminais (Chen et al., 2004). A demanda potencial de ocorrências policiais sem autoria identificada representa, em todos os países de forma geral e particularmente no Brasil, a caracterização da impunidade.

Baluja et al. (1999) e Hauck et al. (2002) citam o emprego massificado de entidades, em pesquisas voltadas para busca de evidências criminais. Chau et al. (2002) citam um experimento para avaliação de sistemas extratores, realizado com 3 coleções de 12 relatórios policiais pertencentes ao banco de informações do Departamento da Polícia de Phoenix, de onde foram extraídas 429 entidades relevantes, de um total de 600 relevantes, usando como fonte os relatórios do experimento.

Seria desejável poder extrair automaticamente, a partir dos históricos policiais, os atores, objetos e circunstâncias relevantes envolvidos na investigação que possam contribuir para elucidação dos fatos e compreensão dos relacionamentos, modelando de forma automática e eficiente os Mapas de Inteligência para análise. A modelagem de Mapas de Inteligência compreende a construção de uma estrutura topológica em formato de rede que permita inter relacionar as diversas entidades úteis extraídas dos textos tratados, estabelecendo um valor de associação entre estas diversas entidades. A estrutura resultante tem como propósito auxiliar a investigação e reconhecimento de padrões criminais aplicados nos delitos (Chau et al., 2002).

Usualmente em uma rede as entidades são representadas por nós ou vértices e as associações entre as entidades são representadas por conexões ou arcos. Em uma análise criminal os nós são pessoas, organizações, veículos e referências geográficas. Uma associação entre um par de entidades é considerada existente se estas entidades aparecem juntas em uma mesma ocorrência policial. Quanto mais freqüente estas entidades aparecem juntas, maior será o valor ou peso de suas associações (Xu & Chen, 2004).

O peso relacionado às associações entre entidades extraídas pode ser computada por diversas heurísticas. Uma destas heurísticas utiliza regras de termos associativos entre entidades como “membro de”, sugerindo uma associação entidade-entidade. A palavra “prender” e “prisão” indicam uma associação evento-entidade, sugerindo uma associação entre um indivíduo e um evento, envolvendo um aprisionamento. O método de co-ocorrências pesquisa relacionamentos, onde duas entidades aparecem juntas em um mesmo documento. É importante observar que a extração pode captar também entidades sem interesse para a análise ou investigação criminal.

Xu & Chen (2004) citam que o sucesso na construção de mapas para investigação criminal depende do emprego extensivo de técnicas que possam automatizar ao máximo as operações de extração e identificação de entidades úteis, tais como pessoas, locais, organizações ou objetos nos históricos pesquisados.

Chen et al. (2004) citam que entidades extraídas podem oferecer informações importantes para aprofundamento da análise criminal, entretanto toda eficiência desta análise depende essencialmente da limpeza obtida por ocasião da apresentação dos dados na entrada no processo de extração.

4.1.1
Similaridade e Co-ocorrência

O valor das associações entre entidades representadas em um Mapa de Inteligência é de grande relevância para a investigação criminal. Este expressa a intensidade segundo a qual as entidades estão mais próximas ou mais distantes entre si. O valor atribuído aos relacionamentos mapeados auxilia a visibilidade dos vínculos existentes, identifica envolvimentos entre os atores presentes no cenário e produz conhecimento para gerar conclusões e laudos sobre os fatos das ocorrências analisadas.

A determinação dos fatos que ocorrem simultaneamente com probabilidade razoável (co-ocorrência) ou identificação dos itens de uma coleção de dados que estão presentes de forma conjunta (correlação) são tarefas representativas em um modelo de estimativas de grandezas para associações entre entidades em massas de informações (Vidal, 2005).

Destacam-se como conceitos de relevância nas associações entre entidades extraídas de um texto o critério para medida de distância e quantificação de similaridade entre duas entidades. As medidas podem ser categorizadas em medidas de semelhança ou similaridade e medidas de dessemelhança ou dissimilaridade.

Um coeficiente de correlação pode servir como modelo para uma medida de similaridade, enquanto uma distância euclidiana pode representar uma medida de dissimilaridade.

Em bases de dados policiais o conceito de similaridade e dissimilaridade entre entidades é aplicado para identificar associações em históricos criminais, por exemplo, para comparar atributos ou características físicas entre indivíduos, como cabelo ou cor dos olhos (Brown & Hagen, 2002). Coeficientes de similaridade são obtidos como produto de coincidências entre pares distintos de palavras presentes em documentos pertencentes à mesma coleção (Chen & Lynch, 1992).

Vários tipos de distância são usados para calcular medidas específicas de distância entre entidades no espaço vetorial. Algumas medidas usam a distância euclidiana simples enquanto outras usam a distância euclidiana quadrada ou absoluta, onde a distância é a soma das distâncias dos quadrados, evitando o cálculo da raiz quadrada, o que oferece vantagem para velocidade computacional nos cálculos aplicados (Vidal, 2005).

Segundo Kohonen (1997), as palavras em um texto podem ser entendidas como padrões contidos no espaço vetorial. A tarefa primária do modelo de extração deverá estabelecer uma medida de distância razoável entre as palavras extraídas.

Segundo Vidal (2005), a maioria dos algoritmos de análise de agrupamentos estão programados para operar com o conceito de distância (dissimilaridade). Medidas de correlação de similaridade observam a correspondência dos padrões através das características variáveis, não se prendendo à magnitude dos valores dos dados, que é raramente usado nas aplicações de análise de agrupamentos.

Kohonen (1997) cita como a mais simples das medidas de similaridade a **distância de Hamming**, mais apropriadamente definida como medida de dissimilaridade, que consiste na verificação de semelhanças ou diferenças entre conjuntos de símbolos (letras ou números), estabelecendo uma comparação dos conjuntos, computando-se as diferenças de símbolos entre as seqüências comparadas.

Exemplo:

Um valor para a distância de Hamming [dH] entre os conjuntos x=[**pattern**] e y=[**western**] seria :

x= [**p, a, t,** t, e, r, n]

y = [**w, e, s**, t, e, r, n] **dH** (x,y) = 3

Segundo Kohonen (1997), a Distância de Levenshtein ou Distância Editada tornou-se a mais conhecida medida para o conceito de distância, que calcula o número mínimo de operações necessárias para transformação de um conjunto de caracteres em outro, considerando substituições, inserções e eliminações de símbolos entre as duas seqüências de caracteres. Desta forma, a definição da distância de Levenshtein [dL] para as seqüências entre [A e B] será;

dL (A,B) = min (a(i) + b(i) + c(i))

Onde **A** é obtido de **B** através de

a substituições de caracteres;

b inserções de caracteres;

c eliminações de caracteres.

Co-Ocorrência

Originalmente, a abordagem **análise de co-ocorrências** voltava-se para geração automática de dicionários baseados em documentos textuais, computando-se a freqüência com que duas frases apareciam juntas em um mesmo documento. A abordagem estatística define co-ocorrência como a análise de freqüência entre entidades baseada em estatísticas léxicas. Assumindo que duas entidades apareçam juntas em um mesmo documento, poderá existir uma associação e um envolvimento entre estas entidades. Uma co-ocorrência com valor diferente de zero indicará o peso da aproximação entre entidades, tão fortemente associadas quanto maior apresente-se o valor representado pela respectiva co-ocorrência (Schroeder et al., 2007)

Co-ocorrências estatísticas estão relacionadas com as palavras úteis encontradas no texto analisado. O conceito **co-ocorrência** apóia-se na proposição apresentada por Chen & Lynch (1992) para o cálculo de co-ocorrências estatísticas entre palavras extraídas de documentos textuais. O termo co-ocorrência indica uma medida de associação entre quaisquer duas entidades pesquisadas em textos, cujos algoritmos calculam a co-ocorrência entre os pares de entidades presentes nos documentos pesquisados.

Xu & Chen (2004) definem **co-ocorrência** ou **relacionamento associativo** como a relação existente entre um par de entidades, na medida em que surjam juntas em um mesmo documento.

4.1.2
Representação da Rede Semântica

A análise de conexões tem sido usada no âmbito da justiça para pesquisa de associações entre entidades com objetivo de facilitar as investigações criminais (Schroeder et al., 2007).

Dados podem ser representados de forma natural como objetos conectados, por exemplo, **documentos** representados por **nós,** conectados por citações ou referências de hipertextos. De forma similar, organizações podem ser representadas como pessoas (nós) conectadas por relacionamento social ou por padrões de relacionamentos (conexões). A análise específica das conexões entre entidades tem adquirido importância em diversos campos de atividades, como pesquisas criminais, detecção de fraudes, epidemias e recuperação de informações (Jensen, 1999).

4.1.3
Nível das Representações

Uma coleção de documentos pode ser tratada com menor ou maior nível de profundidade pelo algoritmo extrator, variando-se a quantidade de dicionários de apoio e regras associadas à extração de entidades.

O nível de profundidade desejado na análise do Mapa de Inteligência relaciona-se com a classificação das entidades no dicionário especialista e quantidade de dicionários empregados para extração de entidades úteis. Como regra geral, quanto mais diversificadas forem as classes de dicionários utilizadas para apoio à extração (pessoas, veículos, localizações, objetos), maior e mais complexa será a rede resultante. Os dicionários e regras empregados na extração podem atuar como fatores de expansão ou limitação da modelagem do Mapa de Inteligência.

Xiang et al. (2005) citam a importância de limitar o emprego de informações relevantes no apoio à extração (alocação de dicionários e regras aplicadas), de forma reduzir o tempo e recursos computacionais aplicados na busca da solução do problema criminal investigado.

Pesquisas abrangentes, envolvendo grandes volumes de documentos e redes criminais complexas são alimentadas por históricos densos, geralmente envolvendo massificação de entidades, sendo usualmente voltadas para pesquisas de clusters, identificação de sub-grupos criminais, interseção de atividades, vínculos entre sub-grupos e pesquisa de padrões.

Em abordagens onde o objetivo da pesquisa é a descoberta de autorias e evidências, materialidade, envolvimento de cúmplices, uso de objetos e veículos e outras classes mais específicas de entidades, a extração geralmente requer uma mineração de dados mais profunda e abrangente através do uso intensivo de dicionários especialistas e regras amplas para a extração de entidades, cuja coleção apresenta com freqüência um volume limitado de documentos

Em abordagens abrangentes, como pesquisa de quadrilhas especializadas ou pesquisa de subgrupos criminais, o emprego de dicionários não essenciais à análise pode tornar a busca operacionalmente ineficiente. O excesso ou inadequação de entidades na modelagem, sem uma justificada presença, poderá aumentar a complexidade da pesquisa, em decorrência do volumes de entidades e associações conectadas.

Lee (1998) apresenta na Tabela 4.1 um exemplo associando particularidades e características associadas a entidades envolvidas em um caso de trafico de drogas internacional.

"João Pablo, Piloto, Colombiano, chegou ao Aeroporto Internacional de Miami, Passaporte 23456, em 27 de junho de 1997. João Pablo foi detido pelos agentes alfandegários quando 300 kg de cocaína foi encontrado no compartimento destinado ao combustível reserva do seu avião Cessna Station, prefixo PT171"

Tabela 4.1 – Exemplo de propriedades associadas a uma pessoa

Identificação	Documento	Veículo	Entorpecente
Nome: João Pablo Ocupação: Piloto Nacionalidade: Colombiano	Tipo: Passaporte Número: 23456	Tipo: Aéreo Fabricante: Cessna Modelo: Station Identificação: PT-171	Tipo: Cocaína Quantidade: 300 Unidade: Kg

Fonte: Lee (1998)

O exemplo apresentado desmembra a entidade em propriedades associadas que constituem uma identificação mais completa para cada entidade extraída. Esta complexidade entretanto torna-se operacionalmente ineficiente para tratamentos massivos de informações devido ao excesso de conexões a serem geradas pelos algoritmos extratores, sendo recomendada, entretanto, para identificação de autorias criminais, problemas específicos de seqüestros, casos envolvendo múltiplos componentes e objetos, tráfico de entorpecentes, pedofilia, terrorismo ou pesquisas de padrões criminais.

4.1.4
Problemas na extração de entidades de históricos policiais

Diversos erros podem ser encontrados durante a mineração de dados, particularmente quando os tratamentos da extração é executado em históricos textuais, que pode tornar a modelagem inconsistente ou contribuir para resultados distorcidos ou inconclusivos.

Kohonen (1997) cita que freqüentemente são verificados erros na conversão do texto em entidades, produzindo incorreções nos cálculo das distâncias.

Goldberg & Senator (1995) citam que diversas bases de informações apresentam inconsistências, dados incompletos ou múltiplas identificações para as mesmas referências (dualidades) extraídas.

Shimizu & Florentino (2002) recomendam uma consistência progressiva dos dados durante a mineração devido a existência de informações incompletas, erradas e contraditórias que são freqüentemente encontradas nas bases de dados pesquisadas.

Han & Kamber (2001) citam que muitas inconsistências podem ocorrer em bases de informações, tais como violação de restrições ou redundâncias que podem ser removidas através de rotinas integradoras de dados. Alguns atributos podem assumir diferentes nomes em bases de informações heterogêneas. Os erros e inconsistências podem ser eliminados manualmente através de referências externas, imposição de dependências entre atributos, existência de parâmetros para correção ou criação de críticas contra violação de restrições.

Problemas de inadequação funcional ao modelo de extração utilizado poderão ocorrer com alguns documentos presentes nas bases pesquisadas. As rotinas de tratamento de dados deverão identificar, depurar ou descartar os documentos incompatíveis como providência primária para redução de erros e desvios que possam minimizar os resultados esperados pela extração (Lifschitz, 2002).

Xu & Chen (2008) destacam que problemas mais freqüentes detectados na extração de entidades referem-se a dados incompletos, incorretos ou inconsistentes nos registros pesquisados.

- **Dados Incompletos** - as redes criminais operam em modo invisível, dissimuladas ou pouco percebidas. Os criminosos minimizam as interações, objetivando não atrair a atenção policial. Os dados captados tornam-se incompletos, provocando a perda de conexões entre os nós e perda de nós na estrutura da rede.
- **Dados Incorretos** - incorreções referentes a identificações, características físicas ou endereços podem resultar em erros na transcrição das informações que são gerados também de forma intencional pelos próprios criminosos, com intuito de confundir as investigações policiais. Diversos criminosos mentem com respeito às suas identidades quando capturados ou investigados, introduzindo ambigüidades e incorreções nas bases de registros policiais.
- **Inconsistência** - as informações sobre criminosos podem proceder de múltiplas fontes de entradas que alimentam simultaneamente registros policiais, não necessariamente de forma consistente. O criminoso poderá figurar nos históricos policiais com múltiplas identificações, apresentando-

se como indivíduos diferentes, gerando incorreções nas consultas processadas.

Goldberg & Senator (1995) citam como causas mais freqüentes das divergências em registros de bases históricas:

- **Erros na geração dos dados** - são cometidos normalmente por erros introduzidos na identificação do indivíduo durante a entrada de dados cadastrais, decorrentes de declarações abreviadas ou truncadas ou simplesmente cometidos por erros comuns de digitação.
- **Mudanças inesperadas nas especificações dos sistemas** - alterações no formato original do sistema ou alterações promovidas em seu escopo podem gerar divergências nas chaves de acesso aos indivíduos, por exemplo, em uma hipotética mudança nas especificações de entrada.
- **Redução da base de dados** - por razões de custos, diversas informações relevantes utilizadas como chave de acesso podem ser truncadas ou descaracterizadas, perdendo-se importante vínculo nos relacionamentos consolidados entre indivíduos e transações.

  Bancos de dados combinados - são recomendados procedimentos para consolidação de referências ambíguas em bases de dados contendo chaves referenciais divergentes, evitando-se a perda de informações relevantes nos históricos armazenados.

4.1.5
Problemas de consolidação de referências e indicativos de fraudes

As co-ocorrências são computadas pela freqüência com que entidades são identificadas em documentos da coleção pesquisada. Quando extraídas de grandes massas de informações as referências nominais podem apresentar-se incompletas, com descrições alternativas (vulgos ou abreviaturas) ou contendo divergências ortográficas. Tais problemas provocam dificuldades para identificação eficiente de entidades e causam imprecisões no tratamento estatístico das co-ocorrências, particularmente quando as referências envolvem bases de múltiplos indivíduos ou logradouros.

Goldberg & Senator (1995) citam que problemas causados por divergências nos conteúdos das bases de informações provocam limitações no

potencial de utilidade da extração e reduzem a sua confiabilidade. A necessidade da aplicação de rotinas de consolidação para tratamento das conexões depende dos objetivos da pesquisa e das características da base de históricos utilizada. A consolidação é recomendada geralmente quando os dados manipulados provêm de múltiplas fontes, com alto potencial de ruídos nas chaves de acesso, apresentando problemas ortográficos, abreviações ou referências alternativas como abreviaturas ou apelidos.

Constitui-se uma prática comum nos registros policiais a adoção de referências alternativas para o mesmo indivíduo, que utilizam em seus registros o nome completo, nome abreviado ou vulgos, conforme progressivamente torna-se conhecido.

Pressupondo como chave regular de acesso aos históricos policiais o nome completo do indivíduo, as freqüências estatísticas computadas para nomes e vulgos hipotéticos como “Fernando Souza e Costa”, “Fernando Beira-Rio”, “FBR”, ou “Fernandinho Beira-Rio” ou ainda, “Beira-Rio” estariam dispersas e desvinculadas. Estas referências precisam estar consolidados em uma única entidade, considerando que todas remetem para o mesmo e único indivíduo “Fernando Souza e Costa”.

Divergências podem ser solucionados pela supressão dos acentos das palavras durante a fase de pré-processamento do texto. Entretanto, tal recurso potencialmente poderá gerar homônimos, causando convergências indesejáveis para referências nominais. O tratamento de homônimos causados por redução da acentuação ou por referências com grafia semelhante é complexo, exigindo tratamentos com maior profundidade na identificação das propriedades associadas às entidades, como por exemplo, características particulares das entidades ou identificação de padrões e referências específicos associados.

Goldberg & Senator (1995) citam o uso intencional de divergências nominais para se cometer fraudes em sistemas de saúde, ressarcimento de prêmios de seguros, lavagem de dinheiro, uso indevido de serviços por dependentes inexistentes e fraudes de impostos. A descoberta de informações indicativas de fraudes em grandes bases de dados que possam conduzir a indícios conclusivos dos delitos cometidos exige a modelagem de indivíduos devidamente identificados, o conhecimento de seus padrões de atividades, uso de serviços e hábitos de consumo.

Han & Kamber (2001) citam técnicas de pré-processamento empregados para solução de problemas e implementação da qualidade na mineração de dados e redução do tempo requerido para obtenção de resultados processados.

- **Normalização** pode ser implementada para aumento da eficiência nos algoritmos pesquisando para cálculo de distâncias.
- **Redução de Dados** empregada para reduzir dados massivos, através da eliminação de redundâncias ou do emprego de técnicas de clusterização.
- **Integração de Dados** empregado para fusão de múltiplas fontes em um banco coerente de dados.
- **Limpeza de dados** utilizado para remoção de ruídos[6] e eliminação de dados inconsistentes e / ou indesejáveis.

Segundo Goldberg & Senator (1995), as teorias de consolidação automática de informações encontram-se em processo ainda embrionário de desenvolvimento e exigem alto potencial de recursos computacionais para a sua consecução.

4.2
Construção de cenários criminais representados por grafos

Uma representação topológica de vértices e arcos é desenvolvida a partir de documentos textuais com base em dicionários de apoio. A estrutura matemática resultante é representada por uma matriz de nós e arcos direcionados. Cada elemento da matriz (vértices) corresponde a um conceito extraído do texto.

Segundo Chen & Lynch (1992), a base de conhecimento é representada pela rede semântica, onde os nós são representados por palavras, conceitos ou frases e as conexões representam os relacionamentos entre os nós.

Para Baluja et al. (1999) para determinar se certa entidade (*Token*) é relevante basta confirmar a sua existência em um dicionário de apoio ou uma lista de entidades úteis (pessoas, locais ou organizações).

[6]Ruídos são erros aleatórios ou exceções encontradas nas mensurações das variáveis (Han & Kamber, 2001)

4.2.1
Extração de entidades

Chau et al. (2002) apresentam os seguintes métodos para extração de entidades em textos:

- **Pesquisa Léxica** - recurso presente na maioria dos sistemas extratores. Utiliza ferramentas manuais contendo listas de nomes populares para entidades de interesse. Este método pesquisa frases em textos e suas identificações com itens presentes em dicionários de apoio.
- **Pesquisa baseada em regras** - método baseado em regras definidas manualmente. As regras podem ser estruturais, contextuais ou léxicas. Exemplo: letras maiúsculas no primeiro e último nome representa um possível nome de pessoa.
- **Baseada em Estatística** - sistemas que utilizam modelos estatísticos para identificação de ocorrências e padrões associados às entidades relevantes.
- **Máquina de aprendizado**- sistema utiliza algoritmos de aprendizado em substituição à regras para extração de conhecimento e identificação de +padrões em documentos textuais. Exemplos incluem algoritmos para aprendizado de máquina e redes neurais (Baluja, 1999).

Chau & Chen (2002) citam o uso de um método que combina a aplicação de pesquisa léxica, máquina de aprendizado e regras manuais para extração de entidades. Palavras extraídas são verificadas através de técnicas neurais de uma rede do tipo backpropagation para determinação do tipo mais provável do tipo de entidade analisada.

Segundo Hauck et al. (2002), não há limites para construção de uma rede de entidades associadas, que pode conter qualquer volume ou tipos de pessoas, nomes, organizações, lugares e tipos de crimes. Na prática, entretanto, o tamanho da base de dados, o tempo requerido para processamento da modelagem das associações e o tempo requerido para respostas às consultas representam as maiores restrições do método, impondo limitações ao volume envolvido de entidades na análise. Com intuito de equilibrar e tornar eficiente, tanto o processamento, quanto a compreensão dos resultados, torna-se recomendável limitar as entidades envolvidas a um universo restrito, mantendo os volumes

selecionados em níveis reduzidos e restrito apenas a aqueles tipos de entidades de uso mais freqüente pelos usuários.

4.3
Algoritmos pesquisados

Este tópico trata dos algoritmos pesquisados para modelagem de associações em representação de grafos e cálculo de co-ocorrências entre entidades extraídas de textos.

4.3.1
Algoritmo Chen & Lynch

Chen & Lynch (1992) usaram dois algoritmos para identificação dos pesos entre as conexões das entidades extraídas, ambos baseados em freqüências estatísticas.

O resultado do conhecimento extraído é capturado em uma rede semântica, onde os nós representam os distintos tipos de conceitos e o peso das conexões indica a força de sua relevância.

O primeiro algoritmo é baseado na normalização computacional do **Coseno** e o segundo, desenvolvido é baseado na probabilidade das freqüências (co-ocorrências) identificadas na coleção de documentos, denominado de **Algoritmo Cluster**.

A importância das entidades extraídas pode variar de documento para documento. Uma medida denominada **Análise de Cluster** aplica o conceito de **Freqüência de Entidades** (Term Frequency), que indica o número de vezes em que uma determinada entidade é identificada em toda coleção de documentos pesquisados, e **Freqüência em Documentos** (Document Frequency) identificando o numero de documentos em uma coleção de **N** documentos em que certa entidade é encontrada (Hauck et al., 2002).

Houston et al. (2000) citam que a análise através do algoritmo **Cluster** é usado para converter dados brutos indexados e pesos entre entidades em uma matriz de similaridades, indicando um termo de aproximação ou distanciamento através do cálculo assimétrico de distância computacional.

Os dois algoritmos estão fundamentados no mesmo princípio que conceitos relevantes freqüentemente estão presentes no mesmo documento. Entretanto a diferença fundamental entre os dois algoritmos encontra-se na representação da estrutura produzida pelos resultados.

Chen & Lynch (1992) limitaram em três a menor freqüência necessária para participação da entidades na rede semântica, eliminando as entidades menos importantes que participam apenas de forma eventual da extração.

A diferença básica no cálculo entre os dois algoritmos selecionados por Chen e Lynch (1992) encontra-se no terceiro passo do algoritmo, que trata os valores finais dos pesos, através de diferentes modelos para os denominadores das funções no cálculo das associações.

$$\text{Weight}(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} \times d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2 \times \sum_{i=1}^{n} d_{ik}^2}}.$$

Coseno

Cluster

$$\text{Weight}(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} \times d_{ik}}{\sum_{i=1}^{n} d_{ij}}$$

$$\text{Weight}(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ij} \times d_{ik}}{\sum_{i=1}^{n} d_{ik}}.$$

Onde:

T_i - representa a entidade **i**

T_j - representa a entidade **j**

n - número documentos na coleção

d_{ij} - presença da entidade T_j no documento i (0 ou 1)

d_{ik} - presença da entidade T_k no documento i (0 ou 1)

T_j ,T_k - representa o peso entre T_j e T_k

T_k ,T_j - representa o peso entre T_k e T_j

Chen & Lynch (1992) concluíram pelo Algoritmo Cluster como mais eficiente e melhor representativo do conhecimento humano além de suas propriedades assimétricas que podem melhor ajustar-se à práticas e particularidades da extração contendo arcos direcionados.

4.3.2
Descrição do algoritmo Hauck para cálculo de co-ocorrências

Hauck et al. (2002) adaptou a metodologia desenvolvida por Chen & Lynch (1992) para um algoritmo de tratamento de co-ocorrências em rotinas de mineração de dados. O algoritmo estabelece níveis relativos de importância entre as entidades extraídas dos documentos da coleção pesquisada, calculando pesos para os relacionamentos entre cada par de entidade extraída. Os pesos são calculados com base nas freqüências estatísticas que correspondem a um valor para as associações co-relacionadas.

O algoritmo **Hauck** é mais complexo que o algoritmo original desenvolvido por Chen & Lynch (1992) e consome maior capacidade computacional para apuração de seus resultados. A construção o algoritmo exige a especificação de parâmetros referenciais entre os descritores associados, que representam a importância das palavras nos domínios de informações analisadas

O algoritmo **Hauck** identifica dois valores de referência entre entidades e documentos e entre cada par de entidades extraídas. Para processamento do algoritmo são especificados pesos referenciais que determinam a importância de cada classe de descritor (tipo de entidade) na extração. Um destes pesos é fornecido manualmente e o outro é calculado através de heurística incorporada ao algoritmo.

Algoritmo Hauck - Passo 1

O algoritmo Hauck calcula o peso relativo entre cada Entidade em cada documento da coleção ($\mathbf{d_{ij}}$ - entidade-documento).

A Figura 4.1 apresenta a fórmula para cálculo da co-ocorrência d_{ij} em cada documento da coleção:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right)$$

Figura 4.1 – Co-Ocorrências Entidade-Documento - Hauck

Onde:

i - representa cada documento da coleção

j - representa cada entidade encontrada no documento **i**

N - número de documentos da coleção

df_j - número de documentos nos quais **j** está presente

tf_{ij} - número de ocorrências da entidade **J** em cada documento **i** nos quais a entidade **j** foi localizada**.**

w_j - fator que representa a importância da entidade **j** na extração (valor relativo que pode assumir maior ou menor expressão, de acordo com a importância da entidade na extração)

Algoritmo Hauck - Passo 2

O algoritmo calcula o valor da co-ocorrência entre cada par de entidades encontradas juntas nos documentos da coleção (**W_{jk} e W_{kj}**), através da função assimétrica apresentada na Figura 4.2

$$W_{jk} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(k) \qquad W_{kj} = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(j)$$

Figura 4.2 – Cálculo do peso relativo entre Entidades (Hauck,2002)

Onde:

j - representa a primeira entidade, em cada par analisado no documento **i**

k - representa a segunda entidade, em cada par analisado no documento **i**

W_{ij} - representa o peso final entre a entidade **j** e a entidade **k**

W_{kj} - representa o peso final calculado entre a entidade **k** e a entidade **j**

d_{ij} - peso da entidade j, calculado conforme apresentado no passo 2 deste tópico.

df_{jk} - representa o número de documentos na coleção **N**, nos quais as entidades **j** e **k** são reveladas em conjunto.

d_{ijk} - O algoritmo Hauck calcula o peso combinado de cada par de entidades encontradas juntas em cada documento da coleção. A Figura 4.3 apresenta a fórmula de cálculo do peso combinado do par de entidades jk no documento i e cálculo do peso combinado de kj no documento i. A diferença entre estas funções encontra-se no fator de importância relativa (**W_i** / **W_j**) no cálculo da função.

tf_{ijk} - representa o total de ocorrências nos quais as entidades **j** e **k** são reveladas em conjunto no documento i.

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \qquad d_{ikj} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_k\right)$$

Figura 4.3 – Pesos combinados das entidades jk / kj no documento i (Hauck, 2002)

WeightingFactor j / k - Fator de influência que reduz o valor das ocorrências genéricas muito freqüentes, reduzindo o valor de suas respectivas importâncias.

WeightingFactor j / k é obtido através do cálculo apresentado na Figura 4.4:

$$WeightingFactor(j) = \frac{\log\frac{N}{df_j}}{\log N} \qquad WeightingFactor(k) = \frac{\log\frac{N}{df_k}}{\log N}$$

Figura 4.4 – Função WeightingFactor (fator de amortecimento) (Hauck, 2002)

Onde:

WeightingFactor $_k$ - fator de redução ou amortecimento para a entidade **k**

WeightingFactor $_j$ fator de redução ou amortecimento para a entidade **j**

4.3.3
Limitações e problemas relatados com algoritmos para modelagem de associações

Chen & Lynch (1992) citam que o algoritmo **Cluster** é superior ao algoritmo **Coseno** para capturar conhecimento em grandes bases de informações.

Os valores dos pesos produzidos pelo algoritmo **Coseno** são simétricos e apresentam igual valor entre as associações das entidades, enquanto os produzidos pelo algoritmo **Cluster** são assimétricos. A existência de uma conexão entre duas entidades quaisquer **i** e **j** não necessariamente implica na existência de uma conexão entre **j** e **i** ou que contenha o mesmo valor de associação.

Segundo Chen & Lynch (1992), as funções de cálculos de co-ocorrências dos algoritmos **Coseno** e **Cluster** foram desenvolvidas para tratamento de grandes volumes de documentos, não sendo adequadas ou precisas para tratamentos em conjuntos reduzidos de documentos.

O algoritmo proposto por Hauck et al. (2002), de forma semelhante ao algoritmo **Cluster**, produz valores assimétricos para as associações entre entidades, entretanto penaliza com um fator de redução o valor final das palavras mais freqüentemente encontrados nos textos pesquisados. Este fator de redução é utilizado como forma de minimizar a importância de termos genéricos freqüentemente extraídos (Hauck et al., 2002).

O fator de amortecimento (Weighting Factor é computado pelo inverso da freqüência da entidade na coleção, reduzindo o peso do relacionamento entre o vínculo das entidades. No algoritmo Hauck entidades concentradoras (*clusterizadoras*) tendem perder importância como entidades de vínculos fortes, por exemplo, entidades representativas de lideranças ou entidades referenciais geográficas importantes.

4.4
Método para Construção da Matriz de Relacionamentos Criminais

Neste tópico tratamos do método proposto para extração de entidades e modelagem da rede semântica.

O produto obtido pelo método descrito é uma matriz ponderada de relacionamentos onde cada elemento da matriz representa uma entidade extraída e os pesos representam a relevância computada entre estes relacionamentos.

A representação de uma rede em formato matricial fornece um meio para descrever um grafo, dispensando a existência de uma listagem de vértices e arcos ou a construção de um desenho representativo de uma rede (Evans & Mineka, 1992).

Seja **N** a matriz ponderada com **m** linhas e **n** colunas, correspondentes a cada entidades extraídas (vértices). Seja $\mathbf{n}_{ij}$ a representação do elemento na **i**ésima linha e **j**ésima coluna (Evans & Mineka, 1992).

Cada elemento $\mathbf{n}_{ij}$ da matriz é correspondente a um arco (**i,j**) e refere-se a um valor de associação entre as entidades **i** e **j,** caso estas entidades apresentem relacionamento na modelagem extraída. Denominamos a estrutura resultante de **Matriz de Relacionamentos Criminais,** que segue os seguintes passos para sua construção:

Passo I.

Este passo executa a extração de entidades úteis de históricos policiais. O método para extração utiliza como suporte de apoio os dicionários criados no passo anterior, conforme descrito na Seção 3.3 desta Tese.

As entidades extraídas dos históricos textuais são organizadas em uma estrutura indexada por documento, mantendo os dados disponíveis para acesso do algoritmo. Cada par de entidades extraídas é analisado segundo freqüências computadas em cada documento, sendo posteriormente consolidadas de acordo com os totais processados em toda coleção.

Passo II.

Neste passo, são calculados os coeficientes de similaridade entre entidades e documentos, que considera as freqüências computadas de cada entidade em cada documento e o numero de documentos em que a entidade foi localizada.

Passo III.

Neste passo são processados coeficientes de similaridade (co-ocorrência) entre todos os pares de entidades extraídas. Este índice é obtido computando-se todas as freqüências acumuladas entre cada par de

entidades em cada documento e o número de documentos onde o respectivo par de entidades foi localizado.

A Figura 4.5 apresenta o método de extração de entidades úteis e modelagem da rede semântica.

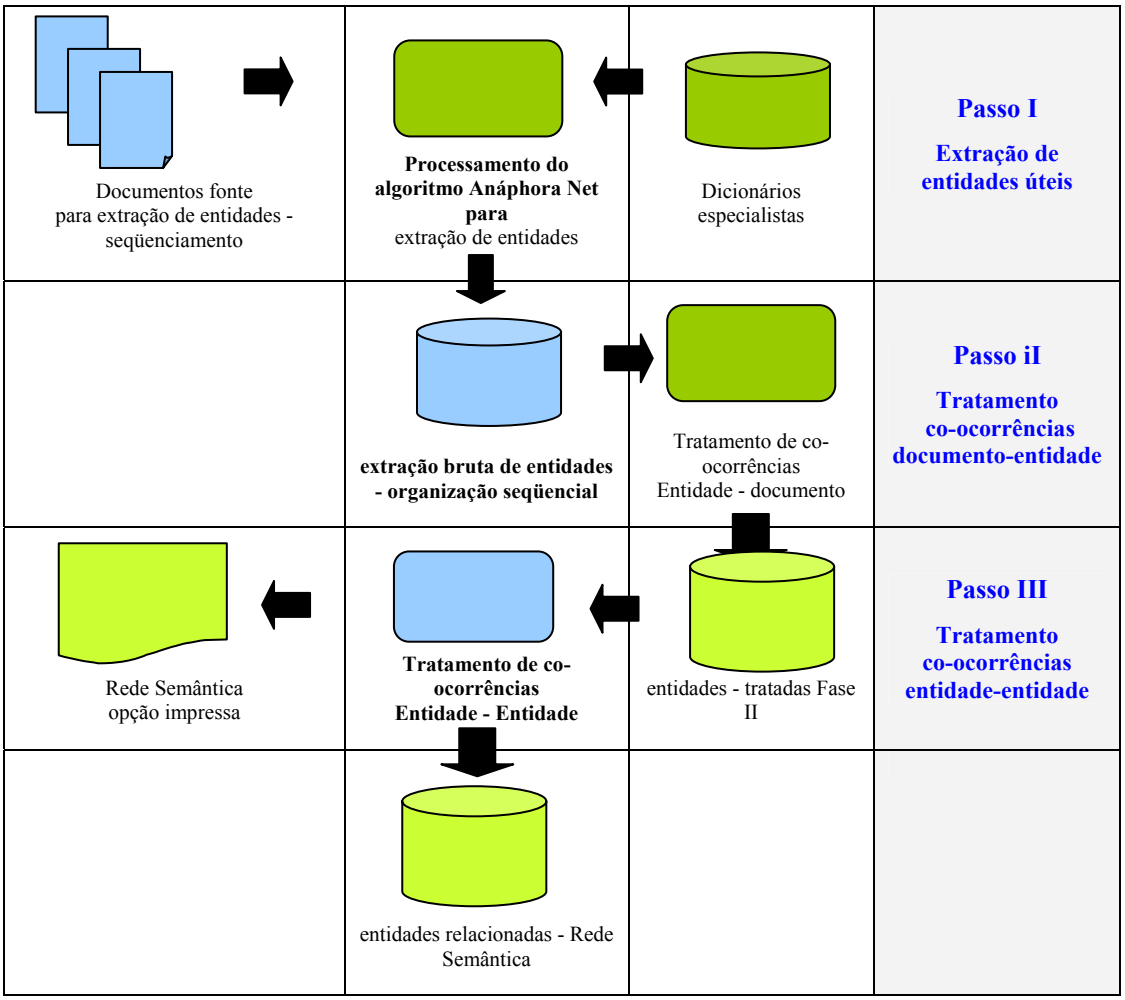

Figura 4.5 – Algoritmo AnaphoraNET para extração de Entidades e Modelagem de Associações

4.5
Alternativas para Modelagem de Cenários

Este tópico descreve alternativas para desenvolvimento de cenários para análise de Mapas de Inteligência. O método proposto neste tópico não limita nem impõe restrições à possíveis combinações de dicionários temáticos ou a diferentes tipos de entidades selecionados para um mesmo Mapa de Inteligência.

Os cenários criminais podem ser produzidos com extrações que envolvem três tipos de modelagem: **elementar**, **moderado** e **profundo.** Esta diferenciação

de cenários relaciona-se diretamente com o nível de detalhes que será extraído do texto para a modelagem do Mapa Criminal e que será resultante do processamento do algoritmo AnaphoraNET.

A seleção do tipo de cenário desejado é informado na etapa inicial de processamento do algoritmo AnaphoraNET. Geralmente, quanto maior for o volume de documentos envolvidos, mais superficial será o nível de detalhes pesquisados nos textos e menor será o volume de entidades extraídas de cada documento processado.

O volume total de entidades extraídas dos textos pesquisados com apoio do Dicionário Especialista relaciona-se com a estrutura das palavras-chave utilizadas na extração, que são classificadas como palavras-chave tipo 1 (orientadas para uma extração com nível de detalhamento elementar), palavras-chave tipo 2, (orientadas para uma extração com nível de detalhamento moderado) e palavras-chave tipo 3 (orientadas para uma extração com nível de detalhamento profundo).

Como regra geral, quanto mais detalhado for configurado o nível da extração, maior será o volume de entidades extraídas e mais complexo será o Mapa de Inteligência a ser analisado.

Nível elementar.

Utiliza geralmente palavras-chave classificadas como Tipo 1 no Dicionário Especialista que são aplicadas para uma extração superficial nos textos explorados. Como resultado da extração são produzidos volumes expressivos de entidades, alimentada por grandes massas de documentos. A extração elementar geralmente envolve nomes, logradouros, tipificação criminal e organizações criminais. O nível elementar tem como objetivos básicos a descoberta de vínculos entre pessoas, identificação de quadrilhas, conexões entre subgrupos e vínculos geográficos.

Nível moderado

Utiliza geralmente palavras-chave classificadas como Tipo 1 e Tipo 2 no Dicionário Especialista que são aplicadas para uma extração moderada de detalhes nos textos explorados. Como resultado da extração são produzidos, de médios a grandes volumes de entidades, alimentada por pequenos a médios volumes de

documentos. A extração elementar geralmente envolve, além das entidades relacionadas no nível elementar, armas, drogas, departamentos, organizações, atividades criminais, e operações policiais. O nível moderado tem como objetivos básicos a descoberta de cumplicidade, vínculos com objetos e atividades criminais. O nível moderado prevê a extração de nomes próprios durante a execução do algoritmo.

Nível profundo

Utiliza geralmente palavras-chave classificadas como Tipo 1, Tipo 2 e Tipo 3 no Dicionário Especialista que são aplicadas para uma extração com grande nível de detalhamento dos textos explorados. Como resultado da extração são produzidos, de pequenos a médios volumes de entidades, alimentada por pequenos a médios volumes de documentos. A extração elementar geralmente envolve, além das entidades relacionadas no nível moderado, objetos, documentos, valores e datas. O nível profundo tem como objetivo a descoberta de autorias, descoberta de cumplicidades, vinculo com documentos e eventos, descoberta de vínculos entre pessoas, cronologia e vínculos com entidades numéricas. O nível profundo prevê a extração de nomes próprios, datas, valores e eventos durante a execução do algoritmo.

4.6
Algoritmos utilizados para construção da Matriz de Relacionamentos Criminais

A extração e modelagem da Matriz de Relacionamentos Criminais é obtido como produto do algoritmo **AnaphoraNET** que trata uma coleção de documentos, extrai entidades úteis orientado por dicionários de apoio, calcula as co-ocorrências entre as entidades extraídas e modela uma Matriz de Relacionamentos Criminais.

4.6.1
Descrição do algoritmo AnaphoraNET para cálculo de co-ocorrências

O algoritmo denominado **AnaphoraNET** foi adaptado para cálculos de co-ocorrências, originalmente baseado na metodologia desenvolvida por Chen & Lynch (1992) e posteriormente recebendo contribuições do algoritmo Hauck (2002).

WeightFactor é um parâmetro intermediário usado durante o tratamento do algoritmo AnaphoraNet para cálculo de co-ocorrências entre entidades extraídas para construção da Matriz de Relacionamentos Criminais. No algoritmo AnaphoraNet o parâmetro WeightFactor apresenta duas funções básicas: dar ênfase à relação entidade / documento, quando a entidade surge com mais freqüência nos documentos pesquisados e reforçar a relação entidade/entidade, como conseqüência das entidades que surgem com maior freqüencia na modelagem na rede.

O Algoritmo **AnaphoraNET** calcula pesos para as associações entre cada par de entidade extraída, com base nas freqüências encontradas em cada documento e surgimento de cada entidade em cada documento da coleção.

O algoritmo AnaphoraNET é baseado no Algoritmo Hauck e segue os passos seguintes:

Algoritmo AnaphoraNET - Passo 1

Inicialmente são extraídas as entidades de relevância na coleção de documentos.

Cada documento é registrado em uma lista de referência:

Cada entidade extraída é registrada em uma lista nominal

As Entidades extraídas são indexadas em uma estrutura contendo dois vínculos:

- **Documentos -** são computadas as freqüências de cada entidade, em cada documento.
- **Par de Entidades -** são computadas as freqüências de cada par de entidades localizadas em conjunto, em cada documento.

Algoritmo AnaphoraNET - Passo 2

O algoritmo AnaphoraNET calcula o peso relativo entre cada Entidade e cada documento da coleção (**d_{ij}** - entidade-documento).

A Figura 4.6 apresenta a fórmula para cálculo da co-ocorrência d_{ij} em cada documento da coleção:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right)$$

Figura 4.6 – Co-Ocorrências Entidade-Documento - AnaphoraNet

Onde:

i - representa cada documento da coleção

j - representa cada entidade encontrada no documento **i**

N - número documentos da coleção

df_j - Número de documentos nos quais **j** está presente

tf_{ij} - número de ocorrências da entidade **j** em cada documento **i** nos quais a entidade **j** foi localizada.

w_j - fator que representa a importância da entidade **j** na extração (valor relativo que pode assumir maior ou menor expressão, de acordo com a importância da entidade na extração)

A Figura 4.7 apresenta a fórmula para cálculo do WeightFactor no algoritmo AnaphoraNET .

$$WeightingFactor(j) = \log\left(df_j * tf_{ij}\right) \qquad WeightingFactor(k) = \log\left(df_j * tf_{ik}\right)$$

Figura 4.7 – Fator intermediário WeightFactor na Função AnaphoraNET

Algoritmo AnaphoraNET - Passo 3

O algoritmo calcula o valor das co-ocorrências entre cada par de entidade extraída na coleção (co-ocorrência W_{jk}), através da função assimétrica apresentada na Figura 4.8:

$$W_{kj} = \sum_{i=1} d_{ikj} \times \sum_{i=1}^{n} d_{ik} \times \log (tf_{ij} \times df_j) \qquad W_{jk} = \sum_{i=1}^{n} d_{ijk} \times \sum_{i=1}^{n} d_{ij} \times \log (tf_{ik} \times df_k)$$

Figura 4.8 – Co-ocorrência Entidade-Entidade algoritmo AnaphoraNET

Onde:

i - representa cada documento da coleção

j - representa a primeira entidade, em cada par analisado no documento i

k - representa a segunda entidade, em cada par analisado no documento i

dij - peso da entidade j, calculado conforme apresentado no passo 2

tfij - número de ocorrências da entidade **j** em cada documento **i** nos quais a entidade **j** foi localizada**.**

dijk - representa o total de ocorrências nos quais as entidades j e k são reveladas em conjunto no documento i.

O algoritmo **AnaphoraNET** é executado, de acordo com os passos definidos na Figura 4.9:

Início	**Passo 1**		**Passo 2**
Para cada Documento da coleção até EOF	Lê dicionários de apoio contendo entidades úteis	Organiza estruturas: documento, entidade, freqüência	Calcula co-ocorrência entre Documentos e entidades
Passo 3			
Para cada estrutura da coleção (Documento a documento) até EOF	Calcula freqüências totalizadas de cada par de entidades em cada documento	Calcula Freqüências de cada par de entidades em cada documento	Calcula co-ocorrência entre cada par de entidade em cada Documentos

Figura 4.9 – Processamento do Algoritmo AnaphoraNET

4.6.2
Algoritmo para Consolidação de Chaves de Acesso

Múltiplas referências divergentes para uma mesma entidade, tais como abreviaturas, apelidos ou acentuação serão resolvidas pelo algoritmo através de uma chave única de referência registrada no dicionário temático. Através desta heurística todos os acessos para uma entidade que eventualmente apresente múltiplas entradas com chaves de acesso divergentes e previamente conhecidas permanecerão ao fim da extração consolidados como uma única chave de acesso à entidade.

No exemplo que ilustra o método descrito, (Tabela 4.2) múltiplas chaves para um único indivíduo foram convertidas para a sua chave única (ilustrada na terceira coluna da Tabela 4.2) consolidadas pelo nome do indivíduo

Tabela 4.2 – Chaves convergentes - Dicionário de nomes

Código Numérico	**Referência Alfabética**	**Chave Única**
0	Colsinho da Vila Vintom	Colso Luiz Rodriguos
2122	Colsinho	Colso Luiz Rodriguos

Fonte - Fotocrim -Sinpol, 2003

4.7
Sistemas

Para execução do Algoritmo AnaphoraNET, foi desenvolvida uma adaptação do sistema DicTools (descrito no capítulo 3), denominado DataAssociations, destinado ao processamento do algoritmo AnaphoraNET .

O programa DataAssociations calcula co-ocorrências entre entidades extraídas de textos e desenvolve estruturas matriciais para modelagens de Mapas de Inteligência e cenários, através do tratamento executado pelos algoritmos:

- Chen & Lynch - Coseno e Cluster
- AnaphoraNET
- Hauck

Os algoritmos Chen & Lynch e Hauck serviram de base para desenvolvimento do algoritmo AnaphoraNET , sendo incorporados como opções

de testes ao sistema DataAssociations e servindo de inspiração para adaptação do Algoritmo AnaphoraNET .

O sistema DataAssociations extrai entidades úteis de históricos policiais, processando o algoritmo AnaphoraNET que é orientado por dicionários temáticos. Como resultante do processamento é desenvolvida uma **Matriz de Relacionamentos Criminais**, cuja estrutura representa uma rede semântica de relacionamentos vértice-arco de um grafo direcionado.

4.7.1
Organização dos dados para tratamento das associações

A organização dos dados para indexação das entidades extraídas foi estruturada em duas matrizes:

Matriz temporária 1 - Relacionamento entre entidade e documento

Matriz temporária 2 - Relacionamento entre entidades (para cada entidade)

Uma estrutura para consolidação dos resultados foi gerada, onde são computados somatórios das freqüências entre entidades e documentos e freqüências entre entidades.

Os resultados permitem indexar:

- As freqüência das entidades em cada documento;
- Os nomes das entidades extraídas;
- As freqüências entre pares de entidades em cada documento;
- As freqüências consolidadas entre pares de entidades e total das freqüências em cada documento da coleção.

As co-ocorrências são tratada em pares distintos. Cada entidade recebe uma verificação recorrente, pesquisando-se a existência simultânea de pares de entidades em cada documento. Para armazenamento das freqüências, indexadas por tipo de entidade, são utilizadas matrizes temporárias residentes em memória. O resultado é armazenado em uma estrutura, cuja representação é uma rede semântica contendo arcos direcionados, ponderados pelo algoritmo AnaphoraNET.

Devido ao intenso trabalho computacional requerido para tratamento das co-ocorrências e à intensa interação entre entidades, a estrutura temporária de matrizes permanece em memória durante toda execução do processamento. Esta arquitetura é requerida para otimização do tempo de processamento, exigindo uma configuração robusta de hardware para tratamento de algoritmos de extração (Chen & Lynch, 1992; Hauck et al., 2002).

Chen & Lynch (1992) citam o problema da "fragmentação de informações" que é causado pelo excesso de conexões entre entidades relevantes extraídas pela mineração em grandes massas de documentos. Como forma de prevenir o problema da "fragmentação de informações" é recomendável a extração através de um conjunto limitado de palavras-chave e categorias previamente classificadas segundo suas importâncias para os objetivos esperados.

4.7.2
Visão funcional do sistema DataAssociations

Primeira etapa para produção de resultados - o sistema DataAssociations organiza uma fila de documentos selecionados para uma pesquisa, alimentados em quantidade variável e originados de múltiplos arquivos. O Sistema DataAssociations suporta o tratamento de arquivos texto mais comuns tipo .TXT, .RTF ou .DOC, cujos formatos podem ser combinados livremente em qualquer pesquisa. A Figura 4.10 apresenta um exemplo de uma seleção de arquivos organizados para tratamento pelo sistema DataAssociations.

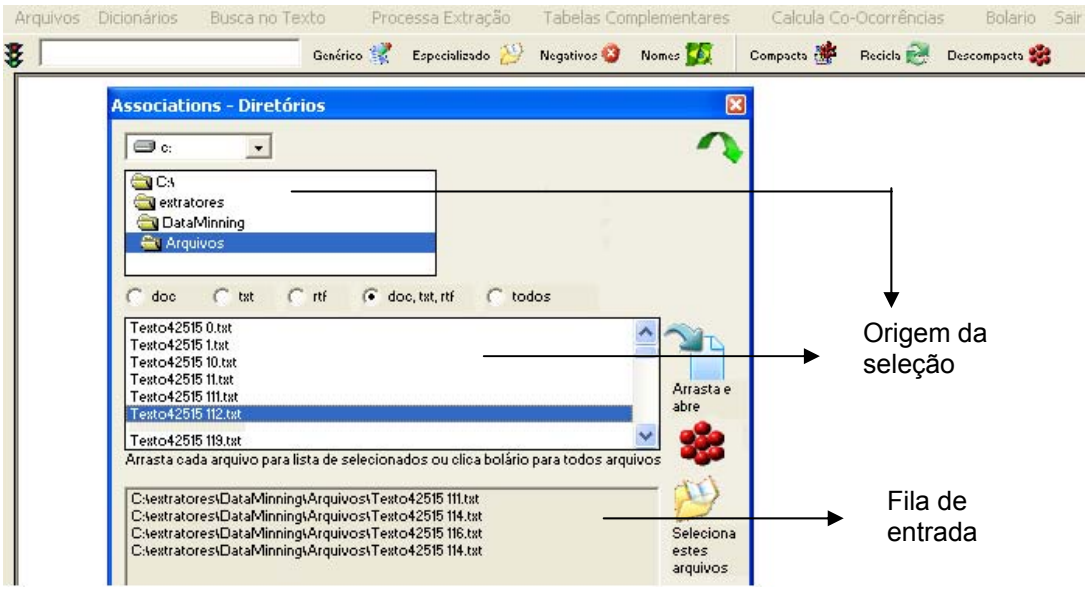

Figura 4.10 – Caixa de seleção de arquivos para pesquisa - sistema DataAssociations

A Figura 4.11 apresenta um painel para seleção de dicionários pelo sistema DataAssociations, destacando um filtro para inclusões que permite configurar o limite mínimo de freqüências das entidades na matriz de relacionamentos.

Figura 4.11 – Caixa de diálogo para ativação de dicionários Sistema DataAssociations

Segunda etapa para produção de resultados - o sistema DataAssociations gera uma estrutura temporária consolidando os totais das freqüências das entidades extraídas nos documentos selecionados.

As entidades são extraídas segundo orientação dos dicionários temáticos selecionados processados pelo algoritmo AnaphoraNET. Um exemplo de uma

estrutura inicial gerada pelo algoritmo AnaphoraNET pode ser observada na Figura 4.12.

Palavra -Chave	Freqüência
dp	56
vulgo	47
inq	44
favela	36
traficante	28
polícia	25
dinamica	22
dre	22
mandado	21
traficantes	20
cv	19
prisão	16
comando vermelho	15
preso	15
policiais	15

Figura 4.12 – Extração consolidada de entidades extraídas pelo sistema DataAsoociations

Terceira etapa para produção de resultados - nesta etapa o sistema DataAssociations processa a matriz de relacionamentos, seguindo os passos do algoritmo ativado. A seleção do algoritmo é executado através de um painel de opções que exibe alternativas de algoritmos suportados pelo sistema DataAssociations (Figura 4.13).

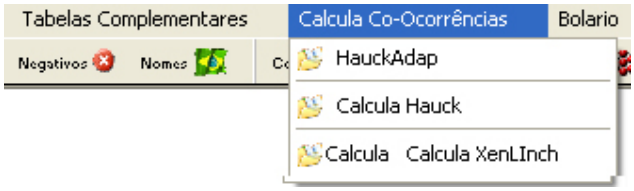

Figura 4.13 – Painel de opções de algoritmos de extração para o sistema DataAssociations

4.8
Resultados do Processamento do Sistema DataAssociations

Como decorrência da extração, o sistema DataAssociations produz os seguintes resultados:

4.8.1
Resultado 1

Geração de uma estrutura temporária contendo totalizações de freqüências das entidades e variáveis temporárias como WeightFactor para cada entidade extraída.

4.8.2
Resultado 2

Geração de uma matriz temporária contendo o número de documentos onde cada par de entidades foi encontrado em conjunto (Figura 4.14)

Documentos onde a Entidade está presente

	luiz fernando da co	comando vermelh	marcos marinho d	marcos antonio pe	ederson jose gonc	amigos dc
luiz fernando da cost		5	5	5	2	
comando vermelho	5		3	2	1	
marcos marinho dos	5	3		5		
marcos antonio pere	5	2	5			
ederson jose goncal	2	1				

Figura 4.14 – Matriz de freqüência das entidades nos documentos - sistema DataAssociations

4.8.3
Resultado 3

Geração de uma matriz temporária contendo freqüências consolidadas para cada par de entidades extraídas, calculadas processada pelo sistema **DataAssociations** (Figura 4.15)

TFunção ADAPT Entidade - Entidade em todos os Documentos

	luiz fernando da co	comando vermelh	marcos marinho d	marcos antonio pe	ederson jose gonc	amigos dc
luiz fernando da cost		559,878	454,901	244,947	74,720	
comando vermelho	486,263		156,951	43,264	14,400	
marcos marinho dos	369,551	146,806		198,989		
marcos antonio pere	174,805	35,549	174,805			
ederson jose goncal	44,977	9,980				
amigos dos amigos	123,827	123,827	73,142	29,257	9,738	

Figura 4.15 – Matriz de freqüências consolidadas entre entidades - sistema DataAssociations

4.8.4
Resultado 4

Geração de uma matriz de resultados finais normalizados, contendo freqüências consolidadas para cada par de entidades. (Figura 4.16)

Função Normalizada ADAPT Entidade - Entidade

	luiz fernando da co	comando vermelh	marcos marinho d	marcos antonio pe	ederson jose gonc	amigos do
luiz fernando da cost		1,000	0,813	0,438	0,133	
comando vermelho	0,869		0,280	0,077	0,026	
marcos marinho dos	0,660	0,262		0,355		
marcos antonio pere	0,312	0,063	0,312			
ederson jose goncal	0,080	0,018				

Figura 4.16 – Dados normalizados das freqüências entre entidades - sistema DataAssociations

4.8.5
Saída dos dados para análise

A estrutura resultante da matriz normalizada corresponde a um grafo direcionado, cujos vértices estão representados pelas entidades nominais extraídas e seus arcos estão representados pelos resultados dos relacionamentos entidade-entidade.

A estrutura é armazenada em um arquivo auxiliar (arquivo cenário) gerado para posterior análise, concluindo o ciclo desenvolvido pelo algoritmo AnaphoraNET .

O arquivo para análise compõe-se de três informações, correspondentes à cada par de entidades associadas:

- Código numérico dos vértices conectados;
- Valor da associação calculado pelo algoritmo AnaphoraNET;
- Nome de referência das entidades conectadas.

4.9
Funções complementares do Sistema DataAssociations

Representações de estruturas resultantes das associações entre entidades podem ser obtidas como subproduto do processamento do Algoritmo

AnaphoraNET . As representações podem ser geradas em dois formatos: formato tabela (Figura 4.17) ou formato gráfico estrela (Figura 4.18). Ambas as representações (gráfico ou tabela) indicam os valores bidirecionais dos arcos entre os pares de entidades vinculadas, que traduzem os valores das co-ocorrências entre as associações.

Qualquer entidade pode ser selecionada para representação de relacionamentos ativando a linha correspondente à entidade desejada, descrita na coluna zero da linha da matriz resultante do processamento do Algoritmo AnaphoraNET .

O formato tabela destaca o nome da entidade selecionada como centralizadora complementada por uma lista contendo todas as entidades vinculadas, o numero de documentos em que são encontradas juntas e valores correspondentes aos relacionamentos bidirecionais com a respectiva entidade.


luiz fernando da costa

Titulo Associado ao Nó

Ordem	Nome	Documentos	Relação A->B	Relação B->A
3	comando vermelho	7	0,033	0,041
4	marcos marinho dos santos	5	0,037	0,070
5	cidade de deus	2	0,096	0,326

Figura 4.17 – Seleção de relacionamentos entre entidades sistema Associations - Formato tabela.

O formato gráfico (Figura 4.18) apresenta uma representação centralizada pela entidade selecionada que conecta-se em formato de estrela com todas as entidades vinculadas. Os valores e nomes das associações encontram-se registrados nas linhas de cada respectivo vínculo.

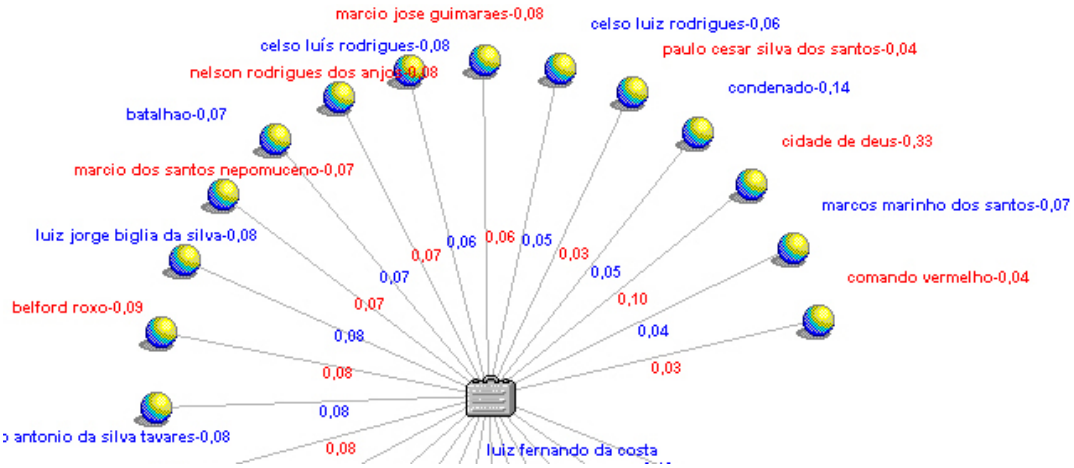

Figura 4.18 – Seleção de relacionamentos entre entidades sistema Associations - Formato estrela

A programação do sistema DataAssociations, destinado ao processamento do algoritmo **AnaphoraNET** foi desenvolvida pelo autor em 2008, para os laboratórios requeridos pela Tese, utilizando linguagem de programação Microsoft Visual Basic 6, que contemplou rotinas para tratamento dos algoritmos e rotinas para representação gráfica das funções de relacionamento e matrizes contendo estruturas temporárias de totalizações das freqüências entre entidades.

5
Conclusões deste capítulo

Este capítulo apresentou a extração de entidades a partir de históricos policiais, usando como suporte da extração dicionários temáticos. As entidades extraídas foram tratadas como vértices e o relacionamento entre as entidades, como arcos de um grafo direcionado. O valor dos arcos foi computado pela freqüência, segundo a qual as entidades foram identificadas nos documentos pesquisados e a freqüência, segundo a qual cada par de entidades foi identificada em conjunto nos documentos da coleção pesquisada.

A estrutura em formato de grafo, resultante da extração, foi aplicada na modelagem de um mapa de relacionamentos, representando um cenário criminal que posteriormente servirá para preparação de um Mapa de Inteligência que será utilizado para análise e descoberta de conhecimento em uma investigação criminal.