O PROCESSO DE DESCOBERTA DO CONHECIMENTO COMO SUPORTE À ANÁLISE CRIMINAL: MINERANDO DADOS DA SEGURANÇA PÚBLICA DE SANTA CATARINA

Edson Rosa Gomes da Silva (Programa de Pós-graduação em Engenharia e Gestão do Conhecimento - Universidade Federal de Santa Catarina, Santa Catarina Brasil) - edson@egc.ufsc.br
Aires José Rover (Professor do Programa de Pós-graduação em Engenharia e Gestão do Conhecimento - Universidade Federal de Santa Catarina, Santa Catarina Brasil) - aires.rover@gmail.com

Resumo: Este artigo pretende apresentar os processos de mineração de textos e dados para dar suporte à análise criminal. Pretende-se vislumbrar a utilização pragmática dos processos no domínio da segurança pública. Como a análise criminal é um dos pilares da gestão integrada em segurança pública, e esta visa formular estratégias para um enfrentamento eficiente das ações criminosas, pode-se destacar que ela precisa de ferramentas capazes de auxiliar no processo de extração do conhecimento dos bancos de dados das instituições de segurança pública. A extração deste conhecimento ajudará na produção de análises criminais mais embasadas. A mineração de dados, por meio dos processos de *Knowledge Discovery in Database* (KDD) e *Knowledge Discovery Text* (KDT), podem ajudar as instituições a gerar análises criminais mais confiáveis e melhorar o entendimento do fenômeno da criminalidade. Estas análises podem também auxiliar as instituições na formulação de estratégias para atuar contra os criminosos de forma integrada e inteligente. A produção de análises criminais pelos analistas é cercada de vários passos, assim como os processos de KDD e KDT, mas o objetivo em comum em ambos é disponibilizar conhecimento para tomada de decisão. Desta forma, este artigo utiliza dados reais da secretaria de segurança pública, mais precisamente dos boletins de ocorrência registrados na polícia civil de Santa Catarina. Os dados são sobre algumas tipificações de crimes contra a vida. Para manipulação destes dados foi usada a ferramenta WEKA, empregando um algoritmo de associação. O algoritmo utilizado foi o apriori que associou os eventos dos dados das ocorrências registradas. O processo de KDT ajudou a refinar a amostra de dados e o KDD fez as inferências necessárias para colher alguns resultados. Com as associações do apriori se verificou qual a faixa etária entre as vítimas é mais suscetível aos crimes de homicídio, quais horários há maior número de homicídios dolosos e o estado civil das vítimas. Estas informações analisadas e disponibilizadas aos tomadores de decisão podem ajudar no planejamento das instituições que corroboram para a manutenção da segurança pública.

Palavras-Chave: Análise Criminal. Descoberta do Conhecimento. Mineração de dados. Segurança Pública.

The process of knowledge discovery as a support for crime analysis: data mining of Public Security of Santa Catarina

Abstract: This article presents the process of text mining and data to support crime analysis. It is intended to discern the pragmatic use of procedures in the field of public safety. As the crime analysis is a key component of the integrated public safety, and it aims to formulate strategies for effective confrontation of the criminal actions may be noted that it needs tools that can assist in the process of extracting knowledge from databases institutions of public security. The extraction of this knowledge will help in producing more informed crime analysis. Data mining, through the process of Knowledge Discovery in Database (KDD) and Knowledge Discovery Text (KDT) can help institutions generate crime analysis more reliable and improve our understanding of the phenomenon of crime. These assessments can also assist institutions in formulating strategies to act against criminals in an integrated and intelligent. The production of crime analysis by analysts is surrounded by several steps, as well as processes KDD and KDT, but the common goal of both is to provide knowledge for decision making. Thus, this paper uses actual data from the secretary of public safety, more specifically from police reports registered in the police of Santa Catarina. The data is on some of classifying crimes against life. For data manipulation WEKA tool was used with an association algorithm. The apriori algorithm was used that linked the events of the data of incidents recorded. The KDT process helped

to refine the sample data and the inferences made KDD necessary to reap some results. With the a priori associations were found between the age group which the victims are more susceptible to crimes of murder, which period has the greatest number of murders and marital status of victims. This information is analyzed and made available to decision makers can help in the planning of the institutions which support the maintenance of public safety.

Keywords: Criminal Analysis. Knowledge Discovery. Data mining. Public Safety.

1. INTRODUÇÃO

Para se atuar de forma efetiva em um problema é necessário entende-lo e analisar seus meandros. Na iniciativa privada é assim, as organizações procuram pesquisar para encontrar e desenvolver soluções pragmáticas para sanar as dificuldades suscitadas pelo ambiente das instituições, podendo ser no ambiente interno ou externo.

Na esfera publica é diferente, na grande maioria dos casos, ou seja, entender o problema não é o principal objetivo, mas sim atacá-lo de forma a mostrar que está tentando fazer alguma coisa para solucioná-lo. Mesmo que os ataques não surtam efeitos, o importante é mostrar boa vontade nas ações desenvolvidas e mitigar possíveis reclamações e críticas. Muito são os casos no qual esta incoerência acontece, pode-se destacar as áreas como a saúde, a educação e a segurança no Brasil, como casos típicos deste tipo de pensamento exposto. Para enfatizar se pega o caso da segurança pública como um bom exemplo a ser observado. As formas de atuar contra os criminosos são balizadas, na maioria dos casos, por soluções arcaicas de enfrentamento sem uma análise criteriosa.

O governo até se esforça para resolver, mas a cultura de atuar primeiro e só depois pensar se consolidou, e vem em curso por anos seguidos. Desta forma, buscando sanar estas incongruências empresas, universidades e instituições de pesquisa procuram dar maior atenção aos problemas sociais e alinham as teorias acadêmicas com técnicas de gestão aliadas as tecnologias para alcançar resultados mais satisfatórios no enfrentamento da criminalidade.

Como forma de gerar um diferencial nas ações surgiu a mencionada por muitos especialistas da área “Gestão Integrada da Segurança Pública”. A gestão integrada da segurança pública tem pressupostos que visam congregar os organismos de segurança pública em prol de um esforço coletivo para empreso dos ativos institucionais das corporações que atuam na segurança pública. A finalidade é aplacar a violência, a criminalidade e diminuir a sensação de insegurança. Para alcançar estes objetivos é necessário desenvolver estudos para área de segurança pública que visem verificar se os processos, as pessoas e as tecnologias estão sendo exploradas de forma satisfatórias para atingir bons resultados no controle, na prevenção e na repressão da criminalidade.

Para atingir estes objetivos precisa-se, primeiramente, conhecer o problema. Para isto é fundamental analisar os dados e as informações dos agregados criminais das instituições de segurança pública. Todos os estudiosos da área sabem que as atuações de segurança pública que almejem surtir os efeitos necessários precisam de estratégias bem definidas com atuações cirúrgicas. Para empreender estas ações cirúrgicas é necessário entender o funcionamento de uma gestão integrada em segurança pública.

O alicerce da gestão integrada da segurança pública são os conhecimentos produzidos por três áreas de extremo valor, mas que são negligenciadas por grande parte dos gestores nas instituições. Como de extrema necessidade destacam-se os setores de estatística policial, análise criminal e a inteligência de segurança pública que pode ser chamado de tripé do fundamento estratégico para ações de segurança pública. Estes setores são de fundamental importância dentro do contexto de atuação eficiente das instituições de segurança pública. Não se combate o que não se conhece, mas muitos tentam agir sem o conhecimento necessário e agem míopes e de forma simples em cima de um problema de extrema complexidade. Alias são poucas as instituições de segurança pública que utilizam este tripé estratégico de forma sábia. A grande maioria sequer domina os meandros destas três áreas de fundamental importância. Suas utilizações são superficiais e sem o devido foco de subsidiar a tomada de decisão dos agentes públicos. A maioria destes agentes públicos ainda acredita que para atuar eficientemente são necessários grandes efetivos de policiais, recursos volumosos e tecnologias caras e fantásticas.

Não é pretensão deste artigo adentrar em todas as três áreas do tripé de fundamentação estratégica. Até por que cada uma destas áreas necessita de uma boa atenção devido a suas especificidades. Pode-se mencionar que elas se complementam e fundem para um melhor resultado, ou seja, a estatística policial subsidia a análise criminal que por sua vez dá suporte a inteligência de segurança pública. Contudo, o domínio ou conhecimento aplicável de uma delas traz resultado quando bem empregado.

A engenharia do conhecimento pode auxiliar as três áreas do tripé de fundamentação estratégica das instituições de segurança pública. Seja na estatística policial, na análise criminal ou na inteligência de segurança pública. Neste paper, pretende-se ajudar o profissional de segurança pública a desenvolver as análises criminais com suporte de uma técnica da engenharia do conhecimento. Desta forma o foco deste artigo é apresentar o processo de *Knowledge Discovery in Database* (KDD) e *Knowledge Discovery Text* (KDT) para auxiliar a área de análise criminal com suporte de ferramentais computacionais (software open Source) da engenharia do conhecimento. Para alcançar este objetivo se trabalhará com os dados dos boletins de ocorrência da secretaria de segurança pública do Estado de Santa Catarina dos anos de 2008. Será uma pesquisa exploratória com aplicação de um algoritmo. Descrição e análise dos dados com suporte de especialista em segurança pública. O artigo está estruturado com esta introdução sobre o assunto tratado no artigo nessa primeira seção, e na segunda uma revisão sobre o processo de descoberta de conhecimento e banco de dados e textos. Na terceira seção fala-se da análise criminal. Na quarta seção aborda-se a técnica, as variáveis utilizadas e os passos da pesquisa. A quinta seção discute-se os resultados encontrados e na sexta seção as considerações finais. As seções são divididas em subseções para facilitar a abordagem dos assuntos tratados.

2. MINERAÇÃO DE DADOS

2.1. O Processo KDD

A mineração de dados pode ser considerada por muitos uma área de pesquisa multidisciplinar ou até interdisciplinar e nasceu por volta dos anos 80 a partir da necessidade de se analisar grandes bases de dados.

Usando-se uma definição pragmática para mineração de dados, entende-se que são as aplicações de técnicas estatísticas e de inteligência artificial em robustas quantidades de dados com a finalidade de descobrir relações e padrões relevantes entre estes dados (MORAES, 2008 apud SILVA, 2009).

Pode-se destacar que algumas áreas se envolveram nesta seara e se deve dar destaque a estatística, ciência da informação, ciência da computação, engenharia do conhecimento e engenharia de software, dentre outras disciplinas das mais várias esferas do conhecimento. A Figura 01 mostras alguns dos conhecimentos envolvidos na mineração de dados.

Embora o termo mineração de dados ou *data mining* tenha ganhado notoriedade e consequente visibilidade com essa nomenclatura, cabe aqui esclarecer que a mineração de dados é apenas um dos processos ou a ação central. Entretanto, há outros passos que são importantes no processo como um todo.

O termo utilizado por especialistas da área para identificar a técnica que tem como um dos passos a mineração de dados é Processo *Knowledge Discovery in Database*[1] (KDD).

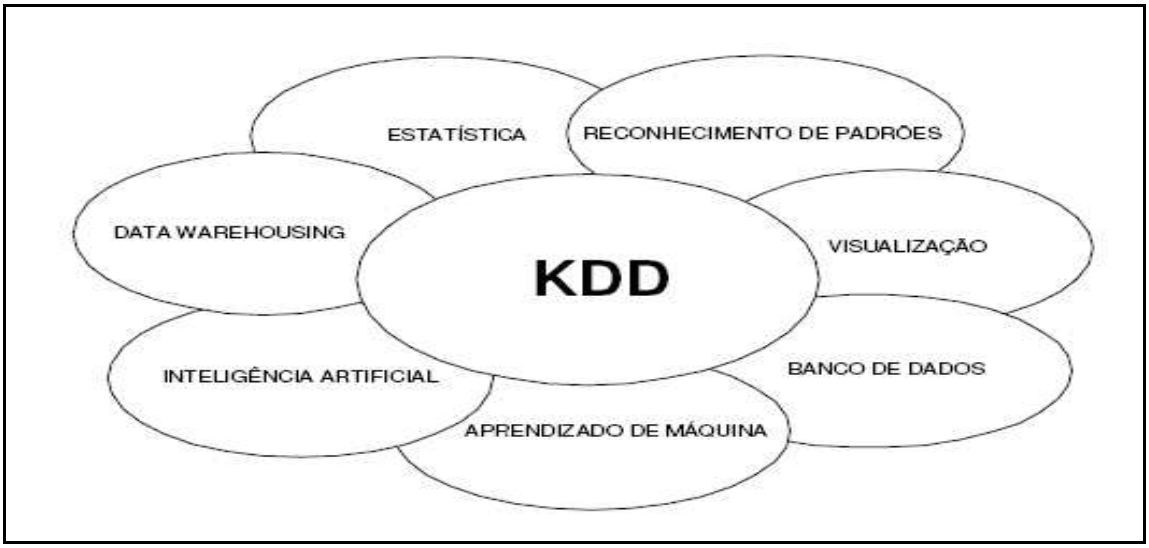

Figura 01 – Conhecimentos envolvidos na mineração de dados

O KDD é o termo utilizado para promover a descoberta do conhecimento em bases de dados, e assim identificar e descrever os relacionamentos implícitos entre as informações nos bancos de dados em sistemas de uma organização.

Segundo Ferro e Lee (2001) o termo processo implica em existir vários passos envolvidos no KDD, como preparação de dados, procura por modelos, avaliação de conhecimento e refinamento, todos estes repetidos em múltiplas iterações.

[1] Descoberta do conhecimento em banco de dados.

Fayyad et al. (1996 apud GOLDSCHMIDT, 2006) aponta o KDD como “um **processo**, de várias etapas, não trivial, **interativo** e **iterativo**, para **identificação** de **padrões compreensíveis**, **válidos**, **novos** e potencialmente **úteis** a partir de bases de dados”. Com a definição o autor apresenta algumas características que são inerentes ao processo de KDD que são expostas:

- **Interação** – pode ser entendido como a ação do homem com a máquina.
- **Iterativo** – são os refinamentos sucessivos que são necessários ao longo do processo.
- **Identificar** – é a procura pelos padrões que devem ser compreensíveis aos membros da organização.
- **Padrões** – é a busca pela forma adequada de representação do conhecimento.
- **Compreensão** – está ligada ao nível de representação de forma inteligível.
- **Validade** – está relacionada com a aplicação dos processos a um contexto determinado.
- **Novo** – refere-se à inovação que é proporcionada pelos novos conhecimentos que são extraídos que causam mudanças na atuação da organização.
- **Utilidade** – está ligada aos benefícios que o processo traz para instituição.

O processo de KDD segue algumas etapas e a literatura define uma seqüência lógica de ação para atingir um resultado satisfatório. As etapas são a limpeza dos dados, integração dos dados, seleção, transformação dos dados, mineração, avaliação e visualização dos resultados.

A Figura 02 ilustra as etapas do processo de KDD e abaixo dela se descreve os passos a serem seguidos segundo (BARIN; LAGO, 2008):

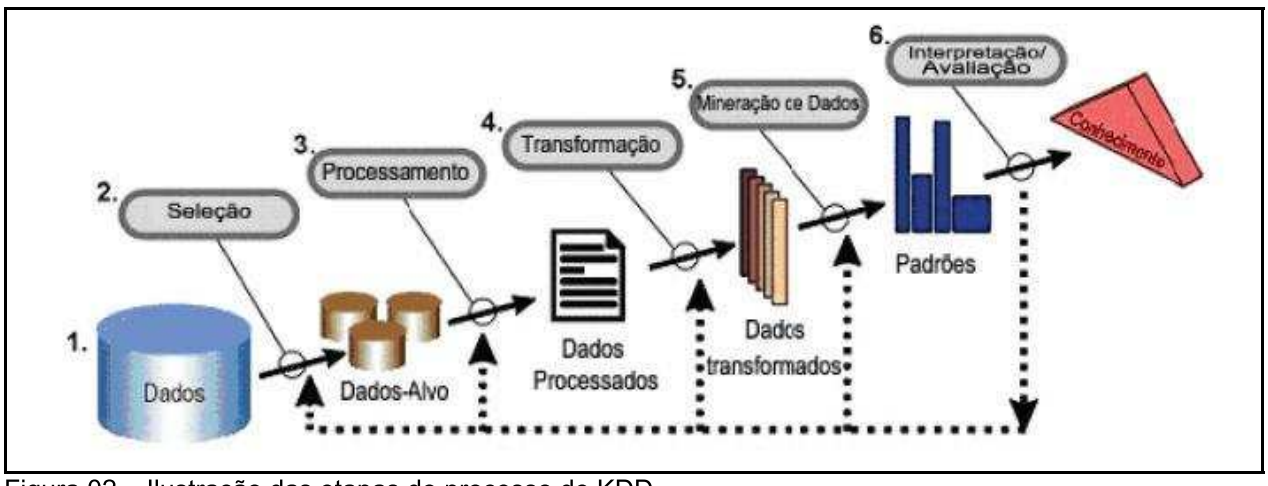

Figura 02 – Ilustração das etapas do processo de KDD
Fonte: Correia (2003 apud BARIN; LAGO, 2008)

1. Dados – Ver o tipo de conhecimento a ser extraído, procurando primeiro uma compreensão do domínio para posterior aplicação, buscando assim o conhecimento para contribuir com a tomada de decisão da organização.
2. Seleção – Criar um conjunto ou subconjunto de dados que será o foco da descoberta de novos conhecimentos.

3. Processamento – Também é conhecido como pré-processamento, nessa etapa realiza a limpeza de dados, incluindo as operações básicas de remoção de inconsistências, coleta das informações necessárias à modelagem, definição das estratégias para manusear campos ausentes, formatação dos dados à ferramenta de mineração.
4. Transformação – Nesta etapa os dados são armazenados de forma correta para facilitar o processamento da mineração dos dados, pois visa agregar valor semântico às informações e mitigar o número de variáveis a serem processadas na mineração.
5. Mineração de dados – seleção dos métodos a serem utilizados para localizar padrões através da descoberta do conhecimento sobre os dados. A maioria das técnicas de mineração de dados é baseada em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização e modelagem gráfica.
6. Interpretação e avaliação – fase na qual a descoberta do conhecimento resultante da mineração de dados é mostrada para os usuários, porém devem ser apresentadas com o pleno entendimento e interpretação dos resultados pelos usuários.

O processo KDD é também apresentado na literatura por fases que circundam os passos e estes podem ser observados nas seguintes fases:

- **Fase do pré-processamento** que visa à preparação dos dados disponíveis, que geralmente não estão dispostos em formato adequado, para a descoberta, análise e a extração de conhecimento.
- **Fase da extração de conhecimento** consiste da escolha da tarefa de mineração de dados a ser aplicada e na escolha da técnica e do algoritmo para extração de conhecimento.
  - Esta Fase pode ser dividida em duas formas:
    - Processo de verificação: usuário sugere uma hipótese acerca da relação entre os dados e testa a sua validade.
    - Processo de descoberta: os dados são vasculhados na procura de padrões frequentes, tendências e generalizações sobre os dados, sem intervenção ou ajuda do usuário.
      - O processo de descoberta pode ser subdividido em dois modelos:
        - Previsão: envolve a utilização de algumas variáveis (atributos da base de dados) para prever valores desconhecidos de outras variáveis de interesse.
        - Descrição: procurar por padrões que descrevam os dados e que sejam interpretáveis por seres humanos.
- **Fase de Mineração de Dados** tem a aplicação das tarefas a serem realizadas como forma de atingir os objetivos pretendidos:
  - Tarefa de Classificação: consiste em determinar em que categorias, já classificadas anteriormente, um determinado atributo em questão apresenta mais semelhança e pode ser considerado daquela classe. Processo pelo qual são examinadas as propriedades (aspectos, estrutura) de um objeto (dados) e atribuí-las a uma das classes predefinidas.

- Tarefa de Agrupamento: Agrupar é simplesmente classificar uma massa de dados em classes desconhecidas a princípio, em número ou forma. Assim, a diferença da classificação para o agrupamento é que a primeira já está determinada com um grupo de entidades e no segundo não há um grupo de entidades definido previamente e o agrupamento se dá por similaridade.
- Tarefa de Associação: A tarefa de associação consiste em descobrir associações importantes entre os itens, tal que, a presença de um item em uma determinada transação irá implicar na presença de outro item na mesma transação. Uma regra de associação é uma implicação na forma X → Y, e possui dois parâmetros básicos: um suporte e uma confiança; O suporte (frequência) é caracterizado pelo número mínimo de ocorrências, enquanto que a confiança (força) é um percentual das transações na base de dados que satisfazem o antecedente da regra (X) e também satisfazem o consequente da regra (Y).

- **Fase do pós-processamento** que consiste na avaliação das descobertas, isto é, se faz uma análise dos resultados obtidos e sua relevância. Nessa etapa ocorre a interpretação, visualização e validação das tendências descobertas, dos padrões extraídos. Quando não satisfeita a necessidade, possivelmente se retorna aos passos anteriores.

Depois de todo processo a organização realizará a implantação do conhecimento descoberto e irá incorporar esse conhecimento na estratégia de negócio e na atuação da instituição.

O KDD não é um processo muito complexo, mas demanda trabalho e bons profissionais para sua implementação, pois há um robusto trabalho de modelagem de dados e extração do conhecimento.

Segundo Kendal e Creen (2007) o processo KDD consiste na construção de um modelo numa situação em que você já procurou e sabe a resposta e, em seguida, aplicando-se a outra situação no qual se pretende responder a um problema similar. Uma vez que o modelo é construído, pode então ser usada em situações semelhantes, para responder às perguntas compatíveis.

A evolução da mineração de dados se deu ainda mais quando aumenta a necessidade de armazenar dados de negócios em computados, continuou com a melhoria do acesso aos dados e mais recentemente com os avanços das tecnologias da informação e a necessidade de gestão do conhecimento, que vem permitindo aos usuários navegar através das informações contidas em seus bancos de dados em tempo real (KENDAL e CREEN, 2007).

O processo de KDD envolve de forma objetiva a mineração de dados e aproveita o processo evolutivo da sociedade do conhecimento para além de proporcionar o acesso aos dados de maneiras retrospectivas e prospectivas com a navegação. Avançando proativamente para disponibilização do conhecimento para as várias áreas das organizações. A Figura 03 mostra os passos que são seguidos em uma organização que procura investir no conhecimento como fator de diferenciação e atuação estratégica visando ser competitiva no mercado.

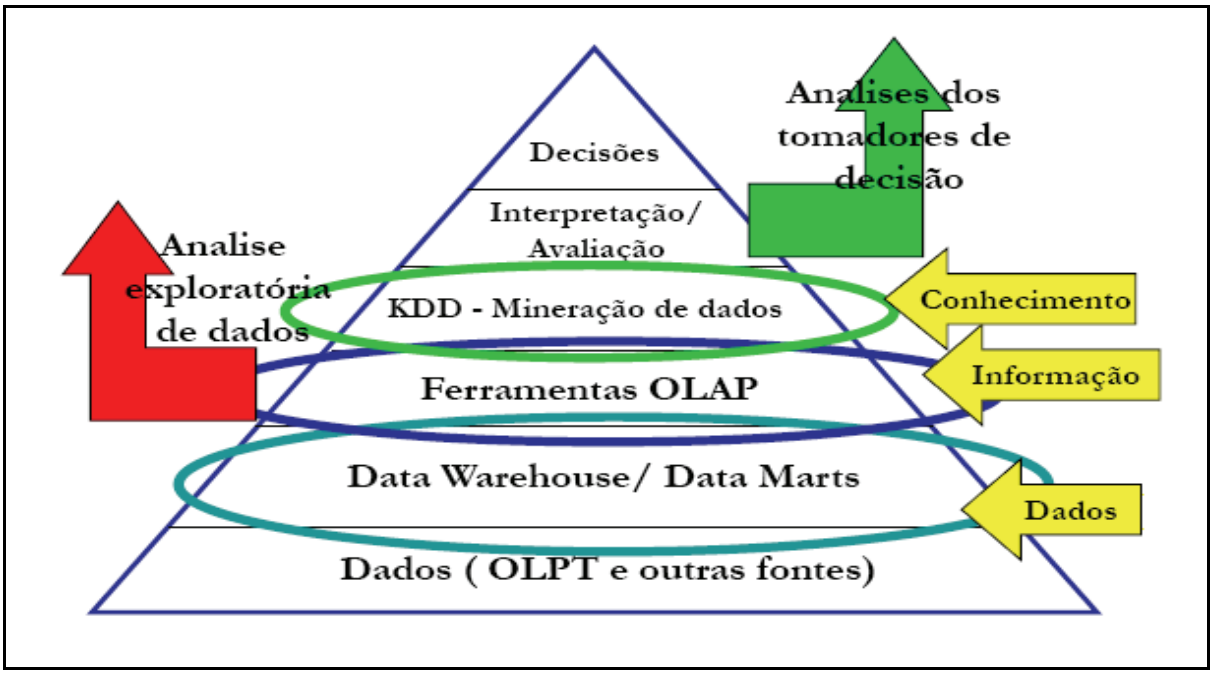

Figura 03 - Passos seguidos em organizações que procuram investir no conhecimento
Fonte: MORAES (2008 apud SILVA, 2009)

Pode-se verificar que o caminho consiste no tratamento dos dados e das informações disponíveis na instituição. Isto muitas vezes é difícil, pois configura um investimento de médio ou longo prazo. Entretanto, o processo KDD tem mais chances de ter resultados se for acompanhado da utilização de outras técnicas para estruturar o conhecimento organizacional. Outro ponto importante é que a adoção de outras técnicas pode diminuir os custos e o tempo de desenvolvimento da solução.

2.2. O Processo KDT

Na primeira década do século XXI o Gerenciamento Eletrônico de Documentos (GED) representa uma importante fonte de informação. Esta afirmativa é tão evidente que os sistemas de GED estão sendo adotados como forma de auxiliar no gerenciamento do conhecimento organizacional das empresas públicas e privadas. Segundo Felix (2009, p. 40) “o Gerenciamento Eletrônico de Documentos (GED) é uma tecnologia que oferece ferramentas para o gerenciamento e acesso às informações disponíveis tanto em papel como em meio eletrônico”. As instituições que tem um passivo em papel estão pegando estes documentos e os digitalizando com técnicas apropriadas. Os documentos que já se encontram armazenado em bancos de dados textuais, podem receber um nível de extração de conhecimento mais apurado com a realização de um processo comumente chamado de *text mining* (mineração de texto). Entretanto, a forma correta é denominada de *Knowledge Discovery Text* (KDT) (Descoberta de conhecimento em textos).

A Descoberta de Conhecimento em Textos (mineração de textos), segundo Passos (2006 apud BARIN; LAGO, 2008, p. 125), é “um campo multidisciplinar que inclui conhecimento de áreas como Informática, Estatística, Lingüística e Ciência Cognitiva”. Com base nos conhecimentos extraídos destas áreas, a mineração de texto define técnicas de extração em padrões ou tendências sobre grandes volumes de textos em linguagem natural para objetivos específicos.

Entre o KDD e o KDT há diferenças que podem ser destacadas como de suma importância, pois no primeiro realiza-se a extração do conhecimento a partir dos dados úteis armazenados em um banco de dados e no segundo a extração do conhecimento é feita nas informações potencialmente úteis contidas em textos. Em outras palavras, a diferença principal apresentada entre KDT e KDD, é que a primeira considera todo tipo de informação (base de dados, textos, e-mails, ou seja, informação não estruturada ou semiestruturada), enquanto que a segunda desenvolve as teorias e os modelos baseados nos dados de uma base estruturada.

No processo de KDT encontram-se resumidamente apresentados nas três fases:

- **Fase da preparação dos dados** que consiste na transformação do documento textual para sua forma vetorial; por este motivo se diz que o processo vetorial é a evolução dos Sistemas de Recuperação da Informação (SRI).
- **Fase da extração de conhecimento** que utiliza algoritmos de mineração para obter uma representação adequada dos dados por meio de atividades preditivas (classificação) ou descritivas (agrupamentos ou regras de associação).
- **Fase do pós-processamento** que consiste na avaliação das descobertas, isto é, analisa os resultados obtidos e sua relevância. Nesta etapa ocorre a interpretação, visualização e validade das tendências descobertas, dos padrões extraídos e possivelmente retorna aos passos anteriores, caso ocorra necessidade de melhorias do resultado.

Embora estas fases apresentem resumidamente o processo KDT, pretende-se utilizar neste artigo as etapas que foram propostas por Aranha e Passo (2007). A figura 04 apresenta de forma mais dissecada as fases do processo KDT.

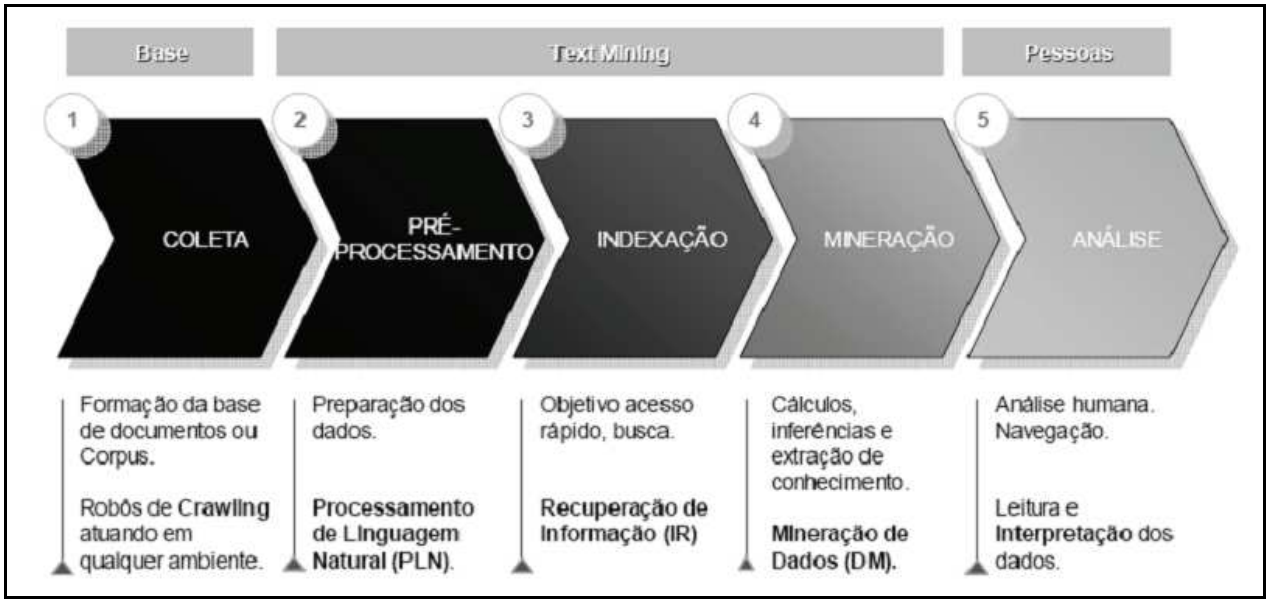

Figura 04 – Etapas do processo de mineração de textos
Fonte: Aranha e Passos, 2007.

Na explicação das fases se utiliza as descrições empregadas por Silva Filho et al. (2010, p. 3651-3652):

- **Coleta:** Nesta etapa ocorre o processo de busca e recuperação de texto com o proposito de formar a base textual da qual pretende-se extrair algum tipo de conhecimento. A coleta de texto é uma tarefa trabalhosa na qual se tem muitos desafios, pode-se destacar a localização da fonte de dados.
- **Pré-processamento:** É a etapa posterior a coleta que visa formatar corretamente os textos, ou seja, deixar homogênio o volume de texto coletado. É custosa esta fase, pois é necessário a aplicação, na maioria das vezes, de diversos algorítimos, que consomi grande parte do processo de extração do conhecimento. No pré-processamento há:
  - Análise léxica ou Tokenização: Finalidade de extrair unidades mínimas do texto. É a eliminação de uma palavra ou mais de texto livre que poder ser associada a simbolo, caracter ou pontuação.
  - Eliminação termos irregulares ou *stopwords*: A finalidade é retirar as palavras que não fazem diferença quando são indexadas, pois somente aumentam o tamanho do arquivo a ser utilizado. Exemplos são os artigos, conjunções, preposições, alguns verbos.
  - Normalização morfológica de termos ou *stemming*: As palavas em um texto podem assumir variadas formas, como exemplo pode-se destacar plural, gerúndio e sufixos temporais. O *stemming* constiste na remoção destas variações deixando apenas a raiz da palavra (ou *stem*).
- **Indexação:** Processo que visa organizar todos os termos que resultaram das fases anteriores a partir de fontes de dados que facilita o acesso e recuperação. Uma boa estrutura de índice garante rapidez e agilidade ao processo de mineração. O modelo booleano é uma das representações mais clássicas utilizadas na mineração de textos. Essa abordagem análisa a presença e ausencia do termo no documento, sendo binário, ou seja, {0,1} os pesos atribuidos aos termos. A vantagem nesta abordagem é a simplicidade e necessidade de pouco espaço de armazenamento.
- **Mineração de dados:** Depois de realizar a estruturação adequada dos textos e criar uma forma rápida de acesso a etapa de mineração é resposável pelo desenvolvimento de cálculos e inferências dos algoritmos para extração do conhecimento com a descoberta de padões e comportamento que possam ser úteis na próxima etapa.
- **Análise dos resultados:** Última etapa que deve ser executada por pessoa que tenha conhecimento do domínio e que esteja interessada no conhecimento extraído para apoiar a tomada de uma decisão com o processo de mineração de texto.

Silva Filho et al. (2010, p. 3652-3653) também apresenta as tarefas de mineração de texto, tal como na mineração de dados, com a classificação, agrupamento e associação, pois são comuns em ambos os processos de KDD e KDT e tem muita similaridade. Tem uma expansão de tarefas que são chamadas de sumarização e extração de informação no processo de KDT. Entretanto as tarefas de mineração de textos são mostradas para verificar sua aplicação de forma efetiva na mineração de textos como segue:

- **Classificação:** A classificação denominada de categorização de textos na mineração de textos é a tarefa de visa identificar os tópicos princípais em documentos textuais e associá-los a uma ou mais categorias predefinidas. Os textos categorizados ficam representados como de uma classe, de forma mais organizada, permitindo, assim, que determinado conteúdo seja acessado de forma facilitada e sem muito esforço. A literatura aponta que a árvore de decisão é utilizada na classificação de documentos por meio de indução da categorização de texto (MOULINIER et al., 1996 e JOACHIMS, 1998 apud SILVA FILHO et al., 2010).
- **Agrupamento:** A tarefa de agrupamento na mineração de textos é fundamentalmente dividida em dois tipos:
  - Termos do texto: Neste os grupos de termos similares são identificados com proposito de construir um discionário de termos que definam assuntos similares.
  - Documentos: O agrupamento por documento pretende identificar os documentos de assuntos similares e alocá-los em um grupo. Este método é muito útil na hipótese de não ser tem ideia das classes (assuntos) tratados nos documentos e se deseja separá-los por assuntos.
- **Associação:** A mineração de textos com regras de associação consintem na decoberta de relacionamentos existentes entre os termos presentes nos documentos. Este procedimento visa identificar os termos que mais se relacionam nos documentos analisados. Desta forma, tem-se pelas regras, quando há presensa, um termo no conjunto dos documentos, implica a presença de algum outro termo distinto no mesmo conjunto de documentos analisados com base nos valores de confiança e suporte definido pelo analista. A presença de um termo se correlaciona com a presença de outro no mesmo documento.
- **Sumarização:** Esta tarefa é utilizada para reduzir a massa de texto e obter ganhos significativos de desempenho quando ocorrer busca de informação necessárias. Também conhecida como criação automática de resumos tem como abjetivo a eliminação de dados (texto) dos documentos, quanto possível. Mas com a permanência do significado textual, ou seja, extrair resumos de textos apresentando como resultado palavras ou frases mais importantes para a compreensão dos textos analisados na íntegra, não necessitando da leitura completa do conjunto para entender os assunto tratados. Normalmente esta tarefa é desenvolvida acompanhada das tarefas de classificação e agrupamento. Para decidir se determinada sentença ou parágrafo será incluído no resumo ele pode ser mapeado por meio de classificação de sentenças (treinamento de algoritmo com as redes neurais artificiais ou classificadores bayesianos), outra aborgagem é a identificação de sentenças e parágrafos por conjunto de agrupamento.
- **Extração de Informação:** A tarefa de extração de informação também é conhecida como extração de caracteristicas, ela foca na obtenção automática de dados estruturados a partir de dados não estruturados. Tem como objetivo final o preenchimento de tabelas (*templates*). O preenchimento destas *templates* permite a vantagem dos dados tornarense estruturados e com isso possa ser manipulados com algoritmos clássicos de mineração de dados.

Nesta subseção apresentou-se o processo KDT com suas particularidades e procurou-se dar um entendimento dos meandos da mineração de textos. Tanto o processo KDD, quanto o processo KDT podem auxiliar na mineração de dados e textos em organizações públicas e privadas. Vários artigos apresentam as vantagens de se utilizar estas técnicas, mas há muitos domínios de conhecimento para explorar seu uso. Neste artigo está se focando na esfera pública, mas precisamente na segurança pública, uma área governamental e como apoio a análise criminal, que é abordada na próxima seção.

3. IMPORTANCIA DA ANÁLISE CRIMINAL

A Análise Criminal (AC) é tida como um conjunto de técnicas e procedimentos cuja finalidade é processar informações relevantes para a prevenção ou a repressão da criminalidade. Estas técnicas e procedimentos estão voltados para a determinação de padrões. Sua utilidade principal, mas não a única, são as áreas de prevenção, controle e combate ao crime pelo policiamento orinetado.

Alguns especialistas em análise criminal apontam que a análise criminal é:

> [...] o conjunto de procedimentos sistemáticos [...] [...] direcionados para o provimento de informações oportunas e pertinentes sobre os padrões do crime e suas correlações de tendências, de modo a apoiar as áreas operacionais e administrativas no planejamento e distribuição de recursos para prevenção e supressão de atividades criminais, auxiliando o processo investigatório e aumentando o número de prisões e esclarecimento de casos (GOTTLIEB, 1994, p.13).

A análise criminal, quando bem desenvolvida, pode prover os gestores da área de segurança pública, informações sobre questões sociais, geográficas, econômicas, entre outras que tenham relevância para o enfrentamento do fenómeno da criminalidade.

A AC é geralmente utilizada no apoio à investigação criminal, a inteligência de segurança pública e para apoio a tomada de decisão da autoridade. Na análise criminal grande números de informações são trabalhadas sistematicamente em prol de uma determinada necessidade. Estas necessidades são na maioria para planejar as ações a serem desenvolvidas nas várias esferas de comando das instituições.

3.1. A análise criminal e o Planejamento

A análise criminal está imbricada com o planejamento das ações, ou seja, a primeira serve, e deve subsidiar a criação do segundo. O problema, na maioria dos casos, é que o planejamento é feito sem se conhecer a realidade do que se pretende desenvolver. Na segurança pública, por exemplo, a estatística policial, a análise criminal e a inteligência policial ou de segurança pública tem que subsidiar a criação do planejamento das instituições de segurança pública. Estas ações visam direcionar a esfera estratégica, tática e operacional da instituição.

3.2. A Análise Criminal e o Planejamento Operacional

Nas organizações públicas o planejamento se configura como uma forte arma para antever os problema e traçar as metas de atuação a fim de trilhar um caminho controlado e bem sedimentado. Para isto as organizações desenvolvem

estudos com a finalidade de dar embasamento às ações que irão desenvolver. Na área de segurança pública não deve ser diferente, pois as ações devem ser bem formuladas e a análise criminal dá subsídios para o desenvolvimento do planejamento e das ações de cunho estratégico, tático e operacional. Magalhães apresenta uma leitura consistente sobre estas ações mencionando que:

> As Ações Operacionais (curto prazo), as Ações Táticas (médio prazo) e as Ações Estratégicas (longo prazo) orientadas, seqüenciadas, articuladas e formalizadas, compõem o conjunto de medidas que estruturam o planejamento organizacional. Com base nesse planejamento é que o Gestor deve avaliar, promover e orientar suas decisões de curto, médio e longo prazo. (MAGALHÃES, 2008, p.5).

Desta forma, Magalhães destaca que:

> O analista criminal, nas suas atividades de produção de conhecimento, deve buscar padrões e tendências criminais que, depois de identificados, constarão em seus relatórios de análise. Esses documentos, por sua vez, devem periodicamente ser difundidos para seus respectivos clientes. (MAGALHÃES, 2008, p.6).

Contudo, o que se entende por padrões criminais? Bem, os padrões criminais são as características identificáveis que se repetem em dois, ou mais, eventos criminais, em uma determinada série histórica, e que vincule, em tese, diversos eventos criminais entre si. Sabe-se que:

> Ao tratarmos do estudo dos padrões devemos ter em conta que o analista criminal não deve utilizar puramente o raciocínio jurídico para definição da sua tipologia criminal. Para o Analista Criminal o foco do comportamento humano é mais importante do que o enquadramento jurídico do fato. (MAGALHÃES, 2008, p.7)

Segundo Boba (2005), ocorreu nos Estados Unidos durante a década de 90, conforme pesquisas nacionais nos estados americanos, um número representativo de agências de polícias que investiram em tecnologia para análise criminal e mapeamento do crime.

Dantas e Souza (2008, p. 9) solidificam este pensamento e citam como exemplo o sistema norte-americano de análise criminal:

> A moderna prática da Análise Criminal está hoje fundamentada no uso intensivo da Tecnologia da Informação (TI), nela incluídos os chamados aplicativos de estatística computadorizada e de sistemas de informação geográfica, tendo como objeto de análise coleções de dados organizados em bases nacionais agregadas. Através da análise das bases nacionais de dados agregados é possível estabelecer relações entre várias categorias de dados e informações criminais, determinando padrões e tendências humanamente impossíveis de serem detectados manualmente. No caso norte americano, o FBI detém hoje as duas mais importantes bases nacionais de dados agregados para as atividades de análise criminal realizadas naquele país: Centro Nacional de Informação Criminal (NCIC) e Sistema de Relatórios Padronizados de Criminalidade (UCR).

Diante destas colocações se verifica a necessidade de aliar as tecnologias, por meio das técnicas disponíveis, sobretudo de engenharia do conhecimento, para realizar a extração de conhecimento das informações depositadas nos bancos de dados das instituições de segurança pública. Este é uma boa forma de auxiliar no processamento dos dados e disponibilizar, de forma mais útil, as informações aos analistas criminais.

4. ABORDAGEM METODOLÓGICA E TRATAMENTO DE DADOS

A metodologia utilizada neste artigo procura racionalizar a pesquisa sobre a análise criminal por meio da utilização de uma das ténicas de *datamine, que é* aplicada aos dados do boletin de ocorrência sobre homicídios dolosos (no qual o agente que comete o fato tem a intenção de causar o resultado, ou seja, assume o risco de causar a morte da vítima) da Secretaria de Segurança Pública do Estado de Santa Catarina. Este delito foi escolhido devido a percepção da sociedade em espelhar o nível de agressividade dos criminosos e pelo choque causado nas pessoas ao saberem os índices de homicidios nos seus Estados. Outro ponto é que este índice é monitorado pelos Estados, pela União (destaque para Secretaria Nacional de Segurança Pública – SENASP e pelo Ministério da Saúde) e por várias orgnizações internacionais.

Para alcançar este objetivo, utiliza-se apoio da pesquisa exploratória com extrapolação analítica nos dados. A pesquisa exploratória mescla a pesquisa documental com a pesquisa bibliográfica. A extrapolação se dá por meio da aplicação da técnica de *datamine* em uma amostra de dados do ano de 2008. Procurou-se suporte na bibliografia correspondente à área de interesse, ou seja, artigos e livros, que dessem sustentação às proposições de aplicação do processo de KDD e KDT.

Foram focalizadas, também, para complementação do estudo, a legislação que versa sobre a tipificação criminal, as normativas e diretrizes vigentes para área de segurança pública.

Como suporte empírico, foi utilizada a experiência do profissional envolvido na produção do artigo, pois este tem ativa experiência com a segurança pública, assim como contatos diversos com vários profissionais da área. Este profissional entende a necessidade e procura apresentar um olhar crítico, mas ponderado, sobre as dificuldades encontradas na condução da análise criminal pelos profissionais de segurança pública. Procura-se mostrar uma alternativa para analisar os dados criminais de forma efetiva e produzir relatórios que subsidiem as ações de segurança pública. Com este suporte, o artigo procura sugerir, com base na pesquisa exploratória e evidência empírica encontrada na mineração dos dados, as nuances por meio de uma exposição ponderada dos entendimentos a cerca do uso dos processos KDD e KDT nos dados de segurança pública.

No decorrer desta seção mostra-se os tratamentos empregados aos dados, a técnica utilizada, as ferramentas de mineração e o tratamento dos dados para que a análise alcaça-se os objetivos almejados.

4.1. Dados gerados para fase de pré-processamento

Seleção dos dados: OS dados da pesquisa foram selecionados por meio de consultas SQL (Structured Query Language, ou Linguagem de Consulta Estruturada). Os dados utilizados são do ano de 2008, que foram registrados no banco de dados dos boletins de ocorrências da Polícia Civil do Estado de Santa Catarina. As informações constantes nos campos são apresentadas no quadro 01 com a denominação correspondente que é utilizada para dar carga dos dados nas ferramentas durante o processo de KDD e KDT.

Ordem	Descrição dos Campos de Dados	Descrição dos Campos Utilizados	Tipo de Campo
01	Número de Identificação do Registro	registro	Numérico
02	Data do Registro da Ocorrência	data_registro	Data
03	Hora do Registro da Ocorrência	hora_registro	Horário
04	Data da Ocorrência	data_fato	Data
05	Hora da Ocorrência	hora_fato	Horário
06	Precisão do Momento da Ocorrência	precisão_momento	Texto
07	Endereço da Ocorrência	loca_fato	Texto
08	Bairro da Ocorrência	bairro_fato	Escolha
09	Cidade da Ocorrência	cidade_fato	Escolha
10	Nome da Vítima	nome_vitima	Texto
11	Natural	natural	Escolha
12	Data de Nascimento	data_nasc	Data
13	Sexo	sexo	Escolha
14	Estado Civil	Est_civil	Escolha
15	Natureza da Ocorrência	natureza_ocorrência	Escolha
16	Relato do fato	relato	texto

Quadro 01 – Descrição dos dados das ocorrências
Fonte: Secretaria de Segurança Pública. Elaborado pelo autor

O número de identificação do registro (é dado automaticamente pelo sistema de registro no momento da ocorrência), data do registro da ocorrência, hora do registro da ocorrência, data da ocorrência, hora da ocorrência, precisão do momento da ocorrência, endereço da ocorrência, bairro onde ocorreu, cidade onde houve o fato, nome da vítima, local que é natural (onde nasceu a vítima), data de nascimento, sexo, estado civil, natureza da ocorrência (tipo de ocorrência atendida) e relato do fato que é dado pela pessoa que fez o registro da ocorrência no distrito policial. Os dados colhidos tinham 16 (dezesseis) itens, mas houve um tratamento para retirada de alguns que não seriam utilizados na análise.

No quadro 02 são apresentadas as naturezas que fazem parte dos dados gerados para o estudo. É importante relatar que foi gerado um total de 3.100 dados, mas ao longo do tratamento os dados tiveram a amostra reduzida.

As naturezas das ocorrências foram retiradas da codificação utilizada no sistema de boletim de ocorrência. Foram gerados dados que tivessem o termo homicídio. Alguns foram descartados em decorrência de não atender o objetivo do estudo, mas poderão fazer parte de estudos futuros. Estes códigos foram:

- Homicídio Acidente Trânsito – Contra Homem, Mulher e Menor (593)
- Homicídio Acidente trabalho homem (38)
- Homicídio Culposo contra homem, Mulher e Menor (52)

Com a exclusão foram retiradas 683 e ficaram 2.417 registros. É importante ressaltar que alguns códigos embora pareçam não atender nosso objetivo, eles são necessários para fazer uma averiguação, pois nos casos das tentativas, muitas vezes, as vítimas são levadas para hospital e entram em óbito posteriormente. Nestes casos os históricos recebem alterações, mas o código do fato nem sempre sobre mudança em sua tipificação ou natureza da ocorrência.

Descrição da Natureza da Ocorrência	Número de registros
Homicídio Doloso contra homem	615
Homicídio Doloso contra Mulher	86
Homicídio Doloso contra Menor	55
Tentativa de Homicídio contra Homem	1155
Tentativa de Homicídio contra Mulher	144
Tentativa de Homicídio contra Menor	261
Violência doméstica - Homicídio doloso contra adolescente	7
Violência doméstica - Homicídio doloso contra mulher	17
Violência doméstica - Tentativa de Homicídio contra Adolescente	9
Violência doméstica - Tentativa de homicídio contra mulher	68
Total de Registro	**2.417**

Quadro 02 – Amostra de dados geradas por natureza da ocorrência em 2008
Fonte: Secretaria de Segurança Pública. Elaborado pelo autor

4.2. Ferramentas utilizadas

4.2.1. Ferramenta WEKA

O pacote de software WEKA (Waikato Environment for Knowledge Analysis) (Waikato Ambiente para Análise do Conhecimento) começou a ser escrito em 1993, usando Java, na Universidade de Wakato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O Weka encontra-se licenciado ao abrigo da General Public License sendo portanto possível estudar e alterar o respectivo código fonte. Neste artigo é utilizada a versão 3.7.1. A interface do WEKA é apresentada na figura 05.

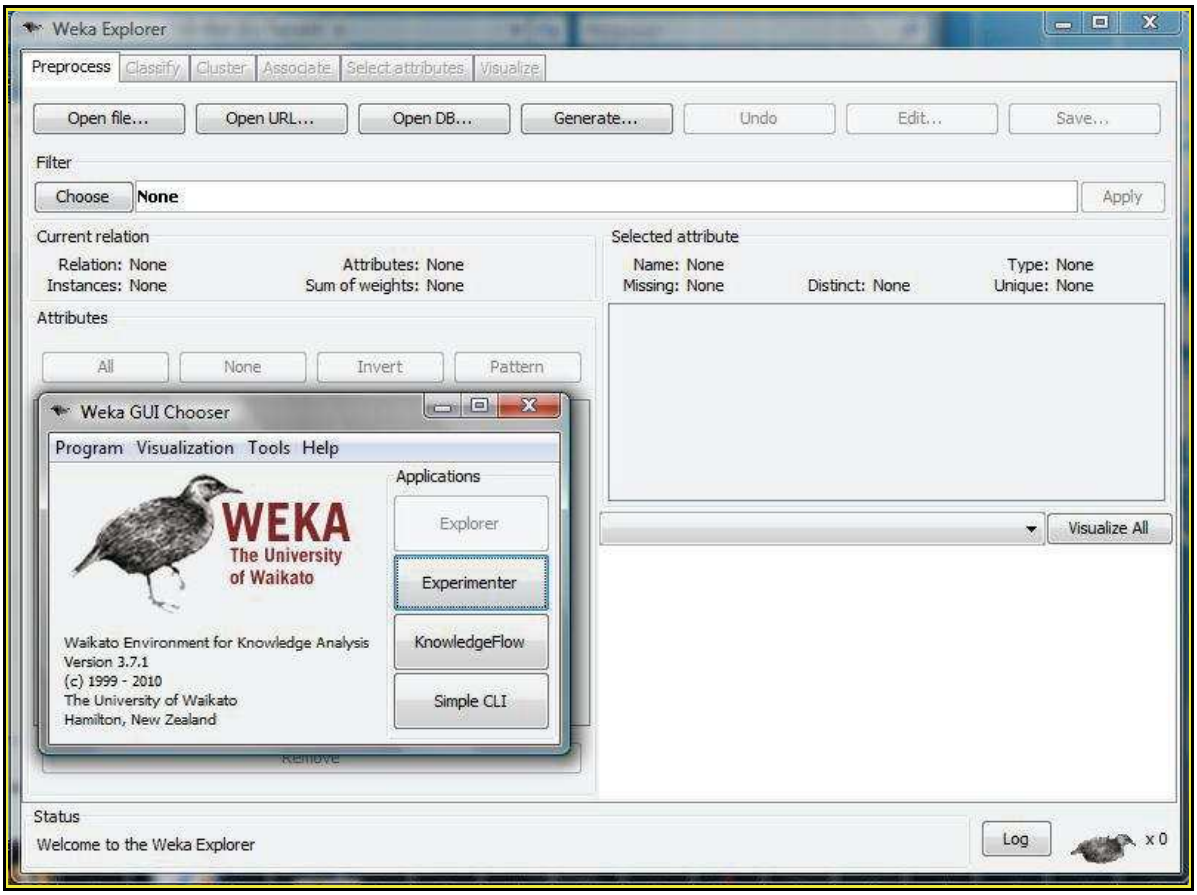

Figura 05 – Interface da ferramenta WEKA
Fonte: Imagem Produzida pelo Autor

De acordo com o site do WEKA, Universidade Waikato, a ferramenta trabalha com uma série de algoritmos, abaixo temos alguns métodos implementados no WEKA:

Métodos de classificação	Métodos para predição numérica
• Árvore de decisão induzida • Regras de aprendizagem • Naive Bayes • Tabelas de decisão • Regressão local de pesos • Aprendizado baseado em instância • Regressão lógica • Perceptron • Perceptron multicamada • Comitê de perceptrons • SVM	• Regressão linear • Geradores de árvores modelo • Regressão local de pesos • Aprendizado baseado em instância • Tabelas de decisão • Perceptron multicamadas

Quadro 03 – Algoritmos Suportados pelo WEKA
FONTE: Universidade de Waikato

O WEKA apresenta-se como uma boa ferramenta e com um bom suporte para importação de dados, podendo ser arquivo (neste caso o formato específico ARFF), URL (Uniform Resource Locator, em português Localizador Padrão de Recursos) ou de alguns bancos de dados.

4.2.2. API LUCENE

Existem várias ferramentas para auxiliar no processo de KDD e KDT, a ferramenta LUCENE é uma API - Application Programming Interface (interface entre aplicativo e programação). Esta ferramenta foi utilizada na fase de pré-processamento e indexação para a mineração de texto.

Segundo Guerra (2007, p. 1):

> A Apache desenvolveu uma API de nome Lucene que tem como utilidade recuperar informações em aplicações de arquivos. Esta funcionalidade se dá através de um engine (motor) de pesquisa, que permite a indexação de textos com alta performance. Isso torna possível executar buscas de qualquer dado que possa ser transformado em texto. O recurso pode ser aplicado para, por exemplo, localizar palavras inclusive em documentos em PDF, que anteriormente foram transformados em textos e indexados pelo Lucene.

Guerra destaca que a API Lucene trabalha indexando da seguinte forma:

> A indexação passa por um processo de análise do documento e, automaticamente, o converte para um texto simples. A extração do texto é feita a partir de um Analyser, classe que contém as regras para a realização desse trabalho de retirada do conteúdo. No entanto, é preciso saber que existem diversas implementações da classe Analyser que realizam essa mesma função (GUERRA, 2007, p. 2).

Na próxima subseção mostra-se como a API Lucene foi empregada no tratamento dos dados no processo de mineração de texto.

4.3. Tratamento dos dados

4.3.1. KDD

Pré-processamento: Nesta etapa da pesquisa se procurou retirar a acentuação das palavras, assim como eliminar os caracteres especiais que poderia gerar problemas para serem manipulados pelos algoritmos das ferramentas escolhida para desenvolver a mineração dos dados. Esta etapa é importante para deixar os dados padronizados na base e mitigar possíveis inconsistências no processo de mineração.

Transformação: Para obter um melhor resultado na aplicação da técnica de mineração de dados, alguns dados sofreram pequenas transformações. Os dados foram manipulados em formato XLS do software Windows Excel. Como exemplo, pode-se citar a data que estava expressa em dia, mês e ano. Esta foi separada e eliminou-se o dia do fato e passou a preservar o dia da semana (segunda-feira, terça-feira, quarta-feira...), mês (separado do dia e ano) e ano (separado do dia e mês). Outro caso é a faixa de horário que ficou expressa em manhã, tarde, noite e madrugada (08:00hs as 12:59hs, 13:00hs as 18:59, 19:00hs as 23:59hs e 24:00hs a 07:59hs respectivamente). Nos campos em que não havia informação foi incluído o

símbolo “?” (interrogação), esse símbolo é interpretado pela ferramenta WEKA como informação ausente. A figura 06 apresenta como os dados ficaram depois da transformação.

	A	B	C	D	E	F	G	H	I	J	K
1	semana_reg	mes_reg	ano_reg	hora_regis	semana_fa	mes_fato	ano_fato	hora_fato	bairro_fa	cidade_fato	natura
2	sabado	janeiro	2008	madrugada	sabado	janeiro	2008	madrugada	zona sul	balneario arroio	araran
3	domingo	abril	2008	madrugada	domingo	abril	2008	madrugada	zona rural	itaiopolis	itaiopo
4	domingo	abril	2008	madrugada	domingo	abril	2008	madrugada	zona rural	itaiopolis	itaiopo
5	domingo	abril	2008	madrugada	domingo	abril	2008	madrugada	zona rural	itaiopolis	itaiopo
6	domingo	abril	2008	madrugada	domingo	abril	2008	madrugada	zona rural	itaiopolis	itaiopo
7	quinta-feira	abril	2008	manha	sexta-feira	abril	2008	madrugada	zona rural	monte castelo	santa
8	domingo	novembro	2008	noite	domingo	novembro	2008	tarde	zona rural	santa terezinha	campo
9	quarta-feira	janeiro	2008	madrugada	quarta-feira	janeiro	2008	madrugada	zona rural	pinhalzinho	maravi
10	segunda-feira	janeiro	2008	manha	domingo	janeiro	2008	madrugada	zona rural	sao joao do oest	tenent
11	sabado	maio	2008	madrugada	sexta-feira	maio	2008	tarde	zona rural	itaiopolis	itaiopo
12	segunda-feira	novembro	2008	tarde	domingo	novembro	2008	noite	zona rural	sao bernardino	sao lo
13	sexta-feira	janeiro	2008	tarde	sexta-feira	janeiro	2008	manha	zona rural	ipira	piratub
14	terca-feira	outubro	2008	tarde	segunda-fei	outubro	2008	tarde	zona rural	riqueza	monda
15	sexta-feira	fevereiro	2008	tarde	quinta-feira	fevereiro	2008	madrugada	zano sul	balneario arroio	turvo
16	quinta-feira	fevereiro	2008	tarde	quinta-feira	janeiro	2008	noite	worspadt	blumenau	blume
17	sabado	marco	2008	noite	sabado	março	2008	noite	vostardt	blumenau	blume
18	quarta-feira	setembro	2008	noite	quarta-feira	setembro	2008	tarde	vostard	blumenau	blume

Figura 06 – Tratamento empregados nos dados
Fonte: Secretaria de Segurança Pública. Elaborado pelo autor

Depois das transformações necessárias os dados foram salvos no formato CSV e convertidos posteriormente no formato ARFF, que é específico para ser carregado na ferramenta WEKA. Entretanto antes de passar para o formato ARFF foi transformado o ponto e vírgula que separavam os dados e vírgula. Foi realizada uma anotação (cabeçalho) para que fosse carregado de forma correta o arquivo na ferramenta WEKA. O cabeçalho do arquivo ARFF é mostrado na figura 08. Temos então na primeira linha o nome do conjunto de dados atribuído pelo comando @relation nome_do_conjuto_de_dados, em seguida temos a relação dos atributos, no qual se coloca o nome do atributo e posteriormente o tipo ou seus possíveis valores, definido por @attribute nome_do_atributo tipo ou {valores}, na seção dos dados coloca-se o comando @data e nas próximas linhas coloca-se os registros, onde cada linha representa um registro. Desta forma, se salva o arquivo e fecha-se com as alterações no formato CSV. Depois é apenas necessário renomear com a extensão .arff (figura 07) e o arquivo ficará no formato reconhecido pela ferramenta WEKA.

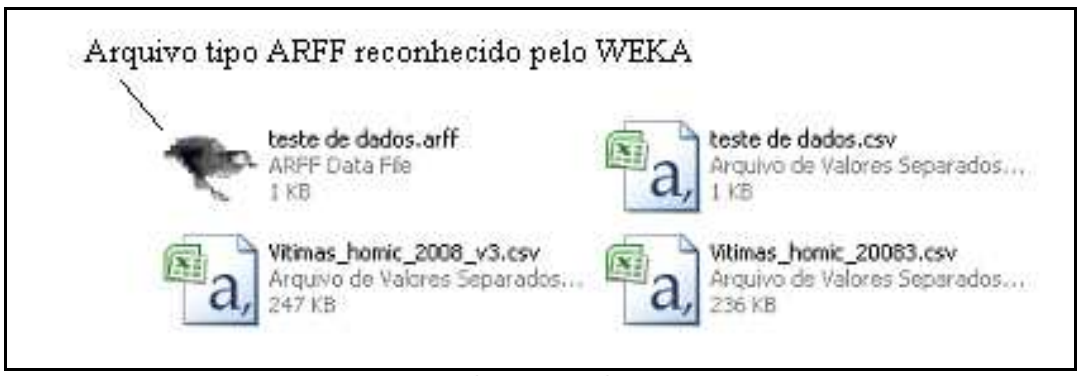

Figura 07 – Formato do arquivo WEKA formato arrf.
Fonte: Elaborado pelo Autor

Depois disso é necessário fazer a importação do arquivo para o WEKA. A ferramenta avisa dos erros que precisam ser sanados quando são carregados os dados para o sistema.

Figura 08 – Anotações para carga no WEKA
Fonte: Elaborado pelo autor

Depois da manipulação nos dados para que possam ser carregados, no caso aqui à ferramenta WEKA, é necessário traçar os parâmetros no sistema que irá realizar a mineração de acordo com o algoritmo empregado. Na quinta seção será mostrado os passos seguidos.

4.3.2. KDT

O processo de mineração de texto foi muito importante para este artigo, ele se mostrou fundamental para as etapas posteriores, ou seja, com o suporte da mineração de texto foi possível excluir os registros que não apresentavam natureza da ocorrência compatível com o objetivo deste artigo. Em outras palavras, as ocorrências que foram registradas como crimes de homicídios tentados, por exemplo, mas que tiveram seus relatos alterados pela autoridade policial em decorrência de falecimento, mas não se procedeu à mudança na tipificação, puderam ser encontradas com o processo de mineração de texto. Nesta subseção apresentam-se os passos desenvolvidos com o KDT.

Seleção: A seleção de dados já foi apresentada, mas nesta subseção fala-se exclusivamente dos dados do processo KDT.

Os relatos disponibilizados nos boletins de ocorrência e os números dos registros (atribuído automaticamente pelo sistema para cada ocorrência e que representa o nome do arquivo no formato “.txt”) foram utilizados para depurar as informações dos boletins e diminuir os números de dados para o processo de KDD. Para este processo foram usados os 1.661 registros, ou seja, os que fazem parte das tentativas, pois para os outros já há confirmação da ocorrência de homicídio. A figura 09 apresenta o exemplo de relato.

Algumas informações foram omitidas dos relatos para preservar a identidade das vítimas. Os arquivos de texto possuem a extensão “txt”.

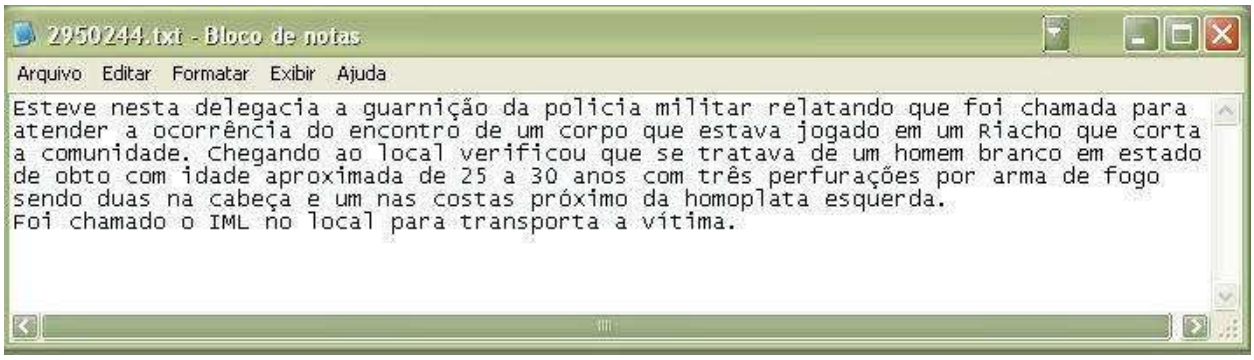

Figura 09 – Exemplo de relado de ocorrência
Fonte: Boletim de Ocorrência SSP

Pré-processamento: Nesta etapa é utilizado o Lucene, isto acontece por meio de suas classes, que estão disponíveis em JAVA. Realizaram-se, assim, apenas algumas rotinas para chamá-las e executá-las. A classe usada foi a brazilianAnalyzer, específica para trabalhar com o idioma português como menciona Guerra (2007).

Com auxílio desta classe foi realizados:

- Geração de *tokens* (termos) que temos em cada documento;
- Retirada de pontuações e caracteres especiais;
- Conversão das palavras em minúsculas; e
- Remoção das *stopwords*.
-

A figura 10 exibe o texto depois de aplicar a classe acima mencionada.

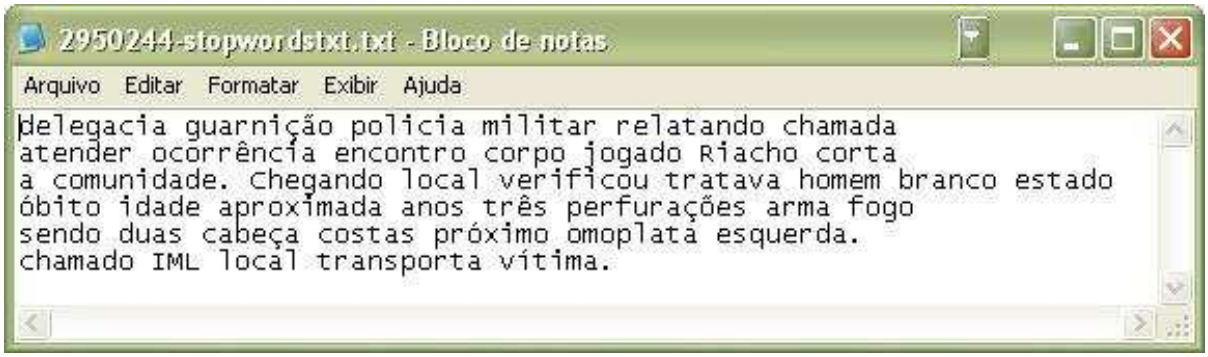

Figura 10 – Texto sem as stopwords
Fonte: Elaborado pelo Autor

Após de realizada a exclusão da *stopwords* aplicou-se o *stemming*, pois o Lucene tem duas classes para facilitar este processo. São elas:

- *BrazilianStemFilter*
- *BrazilianStemmer*

Estas classes são aplicadas apenas a textos em língua portuguesa. Depois da aplicação do stemming as palavras reduziram-se a seu radical. Desta forma foram removidas as letras finais dos termos que possuíam as mesmas variações morfológicas e as flexões. Este processo pode ver visto na figura 11.

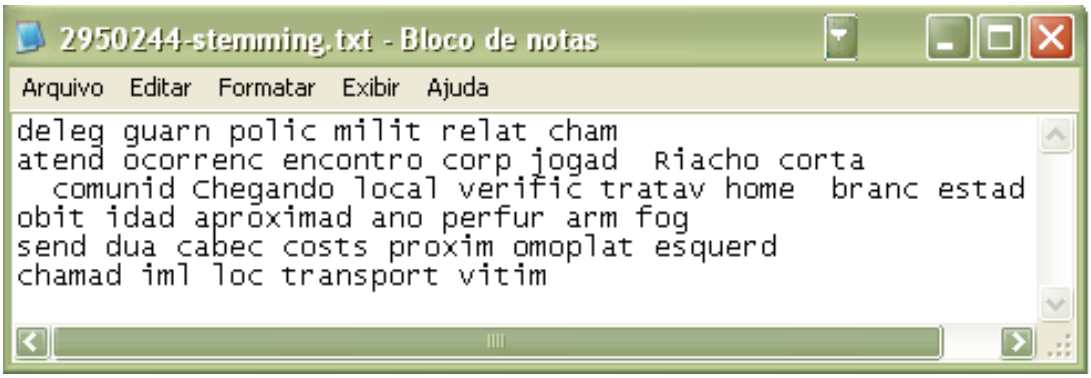

Figura 11 – Texto depois do processo de stemming
Fonte: Elaborado pelo Autor

Com estes procedimentos o número de termos caiu consideravelmente. Ele passou de 69.762 na etapa de processamento, sofrendo uma redução de aproximadamente 36% como pode ser vista na figura 12.

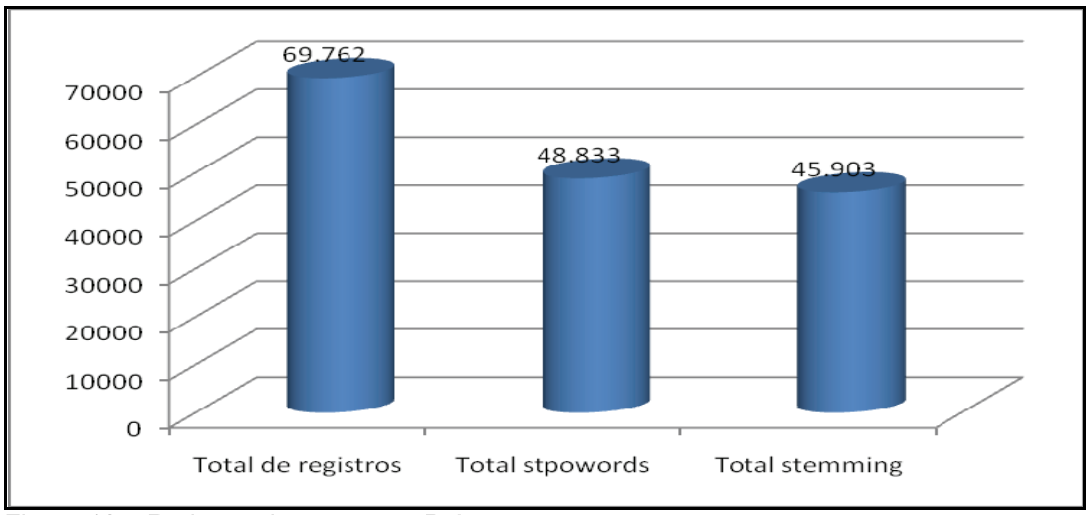

Figura 12 – Redução de termos no Pré-processamento
Fonte: Elaborado pelo Autor

Este procedimento gerou uma redução no número de termos e facilitou de certa forma o próximo passo, ou seja, a indexação.

Indexação: Chegando-se a fase de indexação dos termos restantes, para verificar não a frequência que são encontrados nos documentos, mas para verificar se os termos estão presentes nos documentos, pois a mineração dos textos aqui pretende apontar quais serão somados aos 756 registros para que se possa proceder ao processo de KDD.

Novamente algumas classes do Lucene foram utilizadas para realização desta atividade. O *document*, *directory*, *Analyzer* e *Indexwriter*. Estas classe são usadas para identificar as palavras que existem nos textos e guardá-los em um índice. Guerra (2007, p.3) destaca que:

> Em definição às nomenclaturas do Lucene, a classe Document é uma unidade de indexação e pesquisa que permite armazenar campos (Fields). Sobre a classe Field pode-se dizer que um field só pode ser armazenado em um *Document*, pois possui um nome e um valor. Não

é possível armazenar dois Fields com o mesmo nome em um documento. Mas um documento pode conter um ou mais *Fields*. A classe *Directory* é responsável por endereçar o índice. O armazenamento dos *Documents* é feito no *Directory*.

Na figura 13 o fluxo do processo de indexação do Lucene.

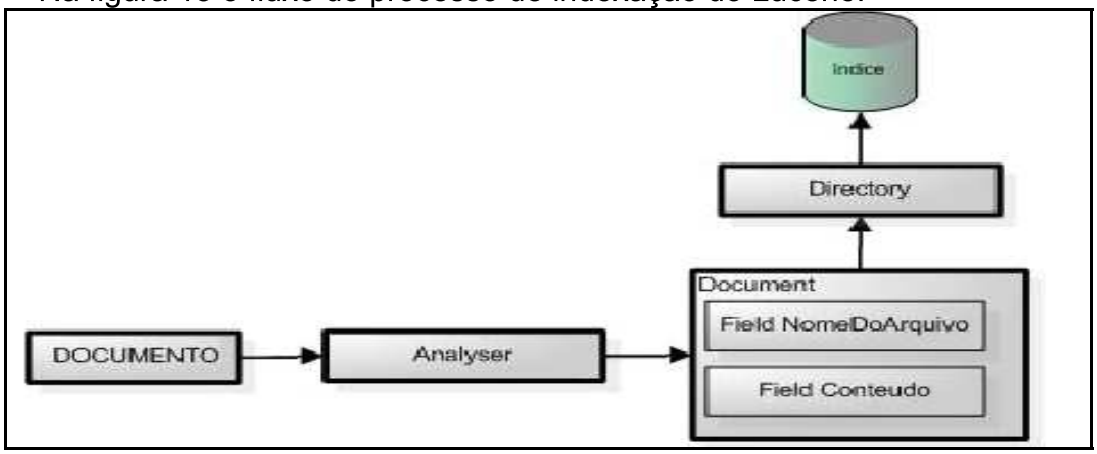

Figura 13 – Fluxo do processo de indexação do Lucene
Fonte: Guerra (2007, p.4)

Depois do processo de indexação dos documentos os termos foram preparados para serem transformados em formato para ARFF para ser importado para a ferramenta WEKA.

O mesmo tratamento foi dado com a criação de um arquivo CSV e posteriormente inclusão de um cabeçalho com os termos que se pretendiam encontrar.

As palavras mais recorrentes na indexação e que foram utilizadas na verificação dos relatos das ocorrências foram: faleceu, morto, sem vida e óbito. Estas 6 (seis) palavras foram utilizadas. As palavras sem e vida foram empregada separadamente e observado quando estas apareciam juntas nos textos. O quadro 04 mostra a relação das palavras e o formato que foram mineradas.

Palavras	Formato Mineração
faleceu	falec
morto	mort
sem vida	sem / vid
óbito	obit

Quadro 04 – Palavras utilizadas no KDT
Elaborado pelo Autor

Na próxima seção serão apresentados os processamentos dos dados e os resultados nos processos KDT e KDD.

5. ANALISE E DISCUSSÃO DOS RESULTADOS

Nesta seção trabalha-se com a apresentação dos passos finais dos processos de KDT e KDD. Destaca-se que o processo KDD teve seus passos mais explicitados devido a necessidade de focar mais em um dos dois processo. Ressalta-se também que não será discutida todas as associações geradas, mas alguns atributos para ser ter ideia de como se pode encontrar os resultados por meio dos processos.

A discussão vai mostrar que neste estudo o processo de KDT deu suporte a execução do KDD. Ele ajudou a realizar o refinamento dos dados de forma mais rápida e eficiente fazendo com que não fosse necessário ler todos os 1.661 registros para chegar aos 54 relevantes para o processo de KDD.

5.1. Aplicando o Processo KDT

Após as etapas de seleção, pré-processamento e indexação, que foram verificadas na subseção anterior, se procedeu a etapa de mineração de textos.

Foi desenvolvido um arquivo com os relatos das ocorrências, no qual cada documento compreendia uma transação. Para conseguir realizar a mineração gerou-se uma tabela, na qual a linha representa os documentos e as colunas os termos dos registros, depois do pré-processamento (após os processos de stopwords e stemming).

```
@relation relato

@attribute falec {1}
@attribute mort {1}
@attribute sem {1}
@attribute vid {1}
@attribute obit {1}


@data
```

Figura 14 – Arquivo de exemplo para mineração de texto
Fonte: Elaborado pelo Autor

A figura 14 mostra o exemplo do cabeçalho feito para o processo de KDT e a figura 15 apresenta a utilização do WEKA para mineração de texto nos relatos das ocorrências.

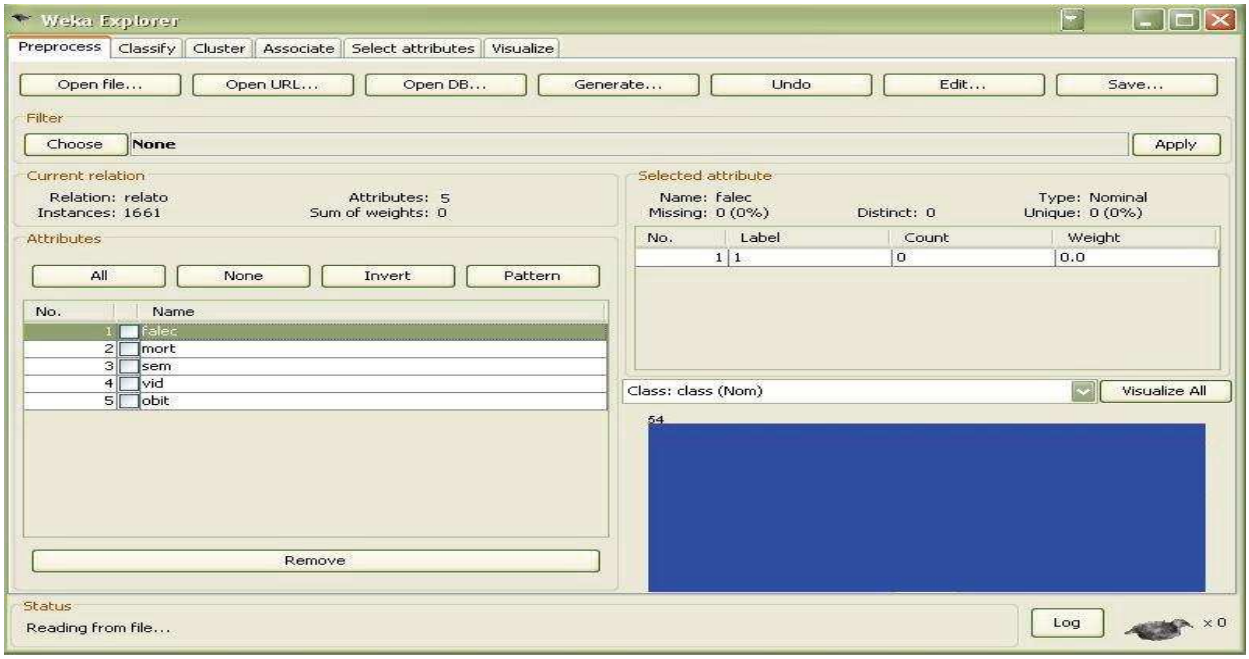

Figura 15 – Mineração dos relatos das ocorrências
Fonte: Elaborado pelo Autor

Com as 1.661instancias (registros) foram minerados 43.903 termos procurando encontra dentro das ocorrências, com a denominação de tentadas, as que teriam sofrido alteração no registro (corpo do texto), mas não em sua tipificação (natureza da ocorrência). A ferramenta WEKA realizou o procedimento de mineração de texto e encontro 54 registros que foram submetidos a leitura para verificar a consistência da mineração.

Na leitura para verificação, percebeu-se que dois registros apresentavam o relato não alterado e a pessoa não havia falecido. Estes foram recuperados no processo de mineração de texto, devido ao fato dos registros terem sido utilizadas as expressões “levado quase em estado de óbito ao hospital” e “conduzido quase sem vida ao pronto socorro”. Estas expressões referem-se aos registros realizados pelos comunicantes dos fatos ou pelos profissionais que lavraram os registros. Desta forma, foram descobertos 52 registros para serem acrescentados aos 756, já confirmados como sendo de homicídio doloso. Isto totalizou 808 registros que serão utilizados no processo de KDD. Na próxima subseção serão apresentados os passos usados na mineração dos dados com mais ênfase nos passos seguidos.

5.2. Aplicando o Processo KDD

O processo de mineração dos dados é uma fase de sua importância no processo de KDD. Nesta subseção mostra-se a os meandros que cercaram está fase e os passos envolvidos na mineração dos dados. Desta forma, depois das fases de seleção, processamento e transformação dos dados do processo KDD foi aplicado a mineração dos dados transformados e já com as anotações necessárias para salvar em formato ARFF que pode ser lido pelo WEKA. A figura 16 dá uma ideia de como os dados se encontravam para esta fase.

	A	B	C	D	E	F	G	H	I	J	K	L
1	@relation ocorrencias_homicidios_2008											
2	@attribute semana_reg {sabado, domingo, segunda-feira, terca-feira, quarta-feira, quinta-feira, sexta-feira}											
3	@attribute mes_reg {janeiro, fevereiro, marco, abril, maio, junho, julho, agosto,setembro, outubro, novembro, dezembro}											
4	@attribute ano_reg {2008, 2009}											
5	@attribute hora_registro {manha, tarde, noite, madrugada}											
6	@attribute semana_fato {sabado, domingo, segunda-feira, terca-feira, quarta-feira, quinta-feira, sexta-feira}											
7	@attribute mes_fato {janeiro, fevereiro, marco, abril, maio, junho, julho, agosto,setembro, outubro, novembro, dezembro}											
8	@attribute ano_fato {2008}											
9	@attribute hora_fato {manha, tarde, noite, madrugada}											
10	@attribute data_nasc {decada_20, decada_30, decada_40, decada_50, decada_60, decada_70, decada_80, decada_90, nao_informado}											
11	@attribute sexo {masculino, feminino, nao_informado}											
12	@attribute est_civil {casado, solteiro, viúvo, uniao_estavel, nao_informado, divorciado}											
13	@attribute natureza_ocorrencia {homicidio doloso contra menor, homicidio doloso contra mulher, homicidio doloso contra homem}											
14												
15	@data											
16	segunda-feira	janeiro	2008	manha	domingo	janeiro	2008	madrugada	decada_80	masculino	uniao_estavel	homicidio doloso contra homem
17	sabado	maio	2008	madrugada	sexta-feira	maio	2008	tarde	decada_80	masculino	uniao_estavel	homicidio doloso contra homem
18	terca-feira	outubro	2008	tarde	segunda-feira	outubro	2008	tarde	decada_70	masculino	casado	homicidio doloso contra homem
19	sabado	marco	2008	noite	sabado	marco	2008	noite	decada_70	masculino	solteiro	homicidio doloso contra homem
20	sexta-feira	janeiro	2008	noite	sexta-feira	janeiro	2008	noite	decada_80	masculino	nao_informado	homicidio doloso contra homem
21	domingo	fevereiro	2008	noite	domingo	fevereiro	2008	noite	decada_60	masculino	casado	homicidio doloso contra homem
22	sexta-feira	abril	2008	madrugada	quinta-feira	abril	2008	?	decada_70	masculino	solteiro	homicidio doloso contra homem
23	sabado	dezembro	2008	madrugada	sabado	dezembro	2008	madrugada	decada_80	masculino	casado	homicidio doloso contra homem
24	quinta-feira	maio	2008	noite	quinta-feira	maio	2008	noite	decada_30	masculino	casado	homicidio doloso contra homem
25	quarta-feira	marco	2008	madrugada	terca-feira	marco	2008	noite	decada_60	masculino	nao_informado	homicidio doloso contra homem
26	domingo	julho	2008	manha	domingo	julho	2008	madrugada	decada_70	masculino	separado	homicidio doloso contra homem
27	quinta-feira	janeiro	2008	noite	quinta-feira	janeiro	2008	?	decada_80	masculino	uniao_estavel	homicidio doloso contra homem
28	sabado	fevereiro	2008	noite	sabado	fevereiro	2008	?	decada_80	masculino	uniao_estavel	homicidio doloso contra homem
29	domingo	julho	2008	madrugada	sabado	julho	2008	noite	decada_80	masculino	uniao_estavel	homicidio doloso contra homem

Figura 16 – Registros transformados e prontos para carga na ferramenta WEKA
Fonte: Elaborado pelo Autor

Os dados foram transformados e carregados na ferramenta WEKA dentro do formato específico. A importação dos dados ocorreu e a figura 17, que mostra como ficaram na ferramenta.

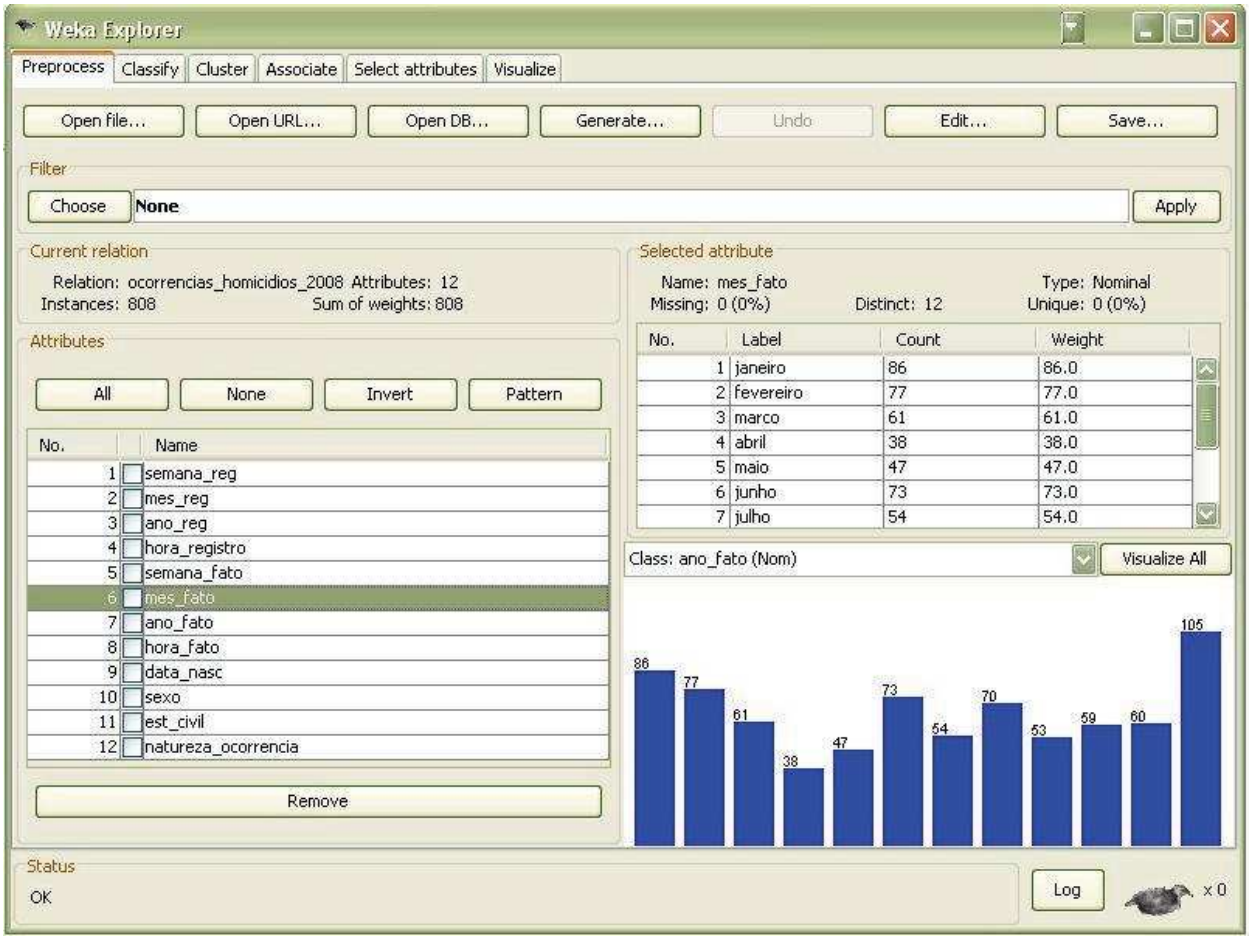

Figura 17 – Dados importados para ferramenta WEKA
Fonte: Elaborado pelo Autor

Após a importação dos dados a técnica de mineração escolhida pode ser aplicada. Como já mencionado se utiliza a técnica de associação por meio do algoritmo *apriori*. É importante destacar que para se aplicar o algoritmo *apriori* pela ferramenta WEKA é necessário configurar alguns parâmetros, embora o WEKA já venha *default* (padrão) para aplicar as regras de associação. O quadro 05 mostra os parâmetros a serem configurados, caso se verifique não atender a necessidade na mineração.

Sigla	Função
-T	Apontar o nome do arquivo de treinamento.
-N	Apontar o número máximo de regras a serem descobertas pelo algoritmo *Apriori*.
-C	Apontar a confiança mínima das regras descobertas.
-D	Apontar a variação para decréscimo do suporte mínimo, do limite superior até o limite inferior.
-U	Apostar limite superior para suporte mínimo das regras descobertas.
-M	Apostar limite inferior para suporte mínimo das regras descobertas.

Quadro 05 – Configuração dos parâmetros do algoritmo *Apriori* na ferramenta WEKA
Fonte: Adaptado de SILVA FILHO et al (2010)

O quadro 06 apresenta os valores que foram utilizados para os parâmetros neste estudo. Optou-se em usar na maioria dos parâmetros o valor *default* do WEKA. Os três parâmetros que não foram utilizados são o número máximo de regras (N), que no *default* do WEKA é 10, o valor do suporte mínimo (M), que no *default* do WEKA apresenta-se muito alto 0,1 (10%) e o valor da confiança (C), que no *default* do WEKA é 0,9 (90%) de confiança mínima. Assim os valores utilizados para estes três parâmetros (N, M, e C) foram alterados para satisfazer os anseios deste estudo e utilizou-se respectivamente 100 (N), 0,01 (M) e 0,75 (C), como se pode ver no quadro XX.

Sigla	Função
-T	0 (não é um arquivo de treinamento.
-N	100 (colocou-se um valor mais algo para receber mais regras geradas.
-C	0.75 (confiança mínima de 75%).
-D	0.05 (5% de variação para decréscimo do suporte mínimo).
-U	1.0 (100% como limite superior para suporte mínimo).
-M	0.01 (1% como limite inferior para suporte mínimo).

Quadro 06– Parâmetros configurados para utilizar no estudo na ferramenta WEKA
Fonte: Adaptado de SILVA FILHO et al (2010)

A figura 18 apresenta as configurações (circulo vermelho) e as associações geradas com os dados utilizados.

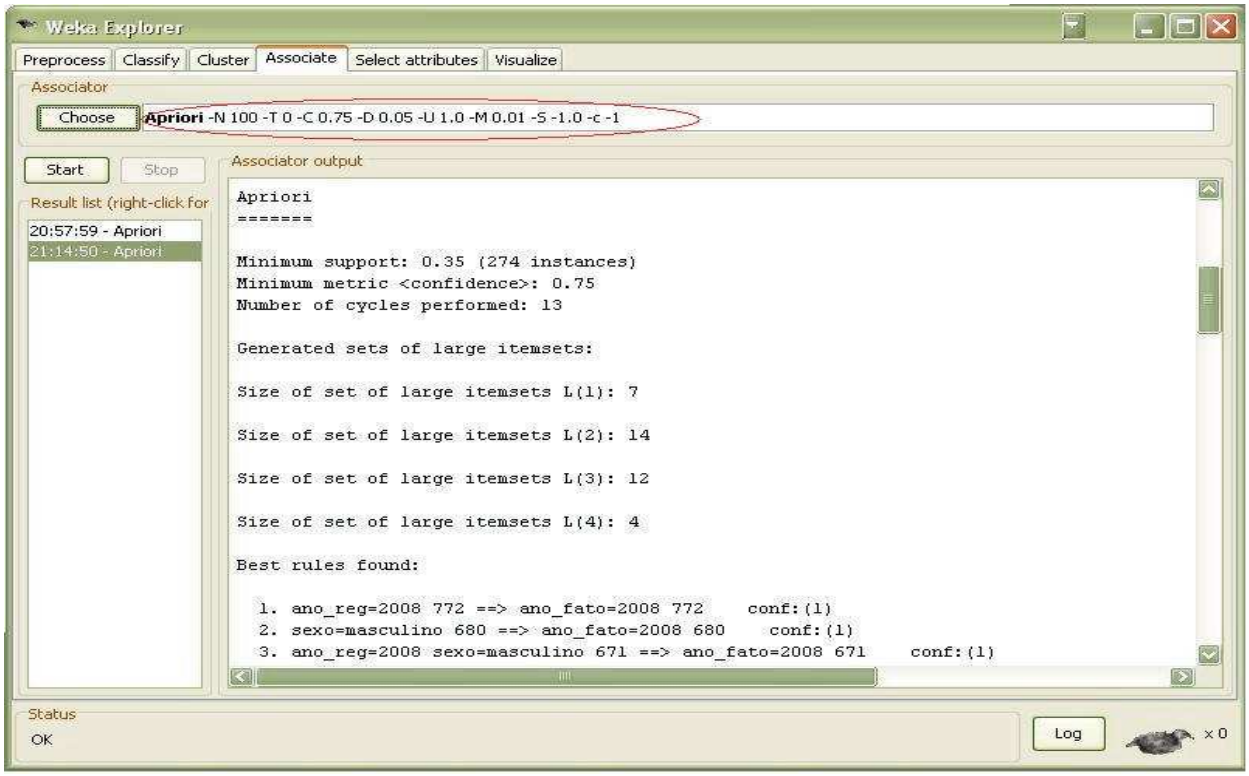

Figura 18 – Parâmetros de configuração e regras de associação geradas na ferramenta WEKA
Fonte: Elaborado pelo Autor

Como foi configurado para serem geradas 100 regras de associação (N) no WEKA é importante apresentar como estas regras podem ser lidas e interpretadas.

Analisando-se a regra (hora_fato=madrugada data_nasc=decada_80 natureza_ocorrencia=homicidio_doloso_contra_homem 104 ==> sexo=masculino 96 conf:(0.92)), pode-se observar que dentro dos 104 registros de ocorrências no qual a hora do fato foi a madrugada (entre 23:59 e 7:59) a década de nascimento da vítima era de 80, e visto que 96 dos caso o sexo da vítima era masculino, ou seja 92% (0.92) de confiança. A confiança é calculada dividindo 96/104 que é 0,92 ou 92%. Já o suporte é calculado 96/808 e chega-se a 0,11 ou 11%, nos quais 808 são o total de instâncias (registros) deste estudo.

Na próxima subseção serão tratados os resultados da mineração de dados, pois a mineração de texto teve seu fim com a seleção dos registros que foram acrescentados para mineração dos dados.

5.3. Interpretação e Discussão dos Resultados

Nesta subseção estão apresentados os resultado das associações geradas com o algoritmo *apriori* na ferramenta WEKA. Como forma de exemplificar os resultados das regras geradas é apresentada apenas algumas, pois como foi solicitado um número mínimo de 100, não é prudente nem interessante discutir todas, mas sim as mais relevantes. É importante destacar que mesmo com um confiança alta é com bons dados é necessário um especialista para analisar corretamente e entender as associações geradas e tirar as análises corretas.

Como suporte mínimo, apresentado a subseção anterior, foi escolhido 0,01 (mostra que a instancia mínima de cada atributo selecionado deve possuir no mínimo oito ocorrências, ou seja, 1% do total de instâncias (registros) utilizadas neste estudo), desta forma, os que tinha número de ocorrências (registros) menores que estes não figuraram em nenhuma das regras geradas. Da mesma forma a confiança mínima foi de 0,75 (75%).

Para um melhor entendimento da análise realizada se utilizou alguns atributos:

O primeiro é a hora do fato que compreendeu-se em manhã (8 as 12:59), tarde (13 as 17:59), noite (18 as 23:59) e madrugada (00 as 07:59).

A faixa de horário na qual mais aconteceu homicídio foram noite e madrugada e estas apareceram em grande parte das associações geradas. Os homicídios que ocorreram a noite perfizeram aproximadamente 33% das ocorrências registradas, já na madrugada ocorreram aproximadamente 32% dos homicídios. No período da tarde ocorreram 13% e de manhã 10%. Percebeu-se que 12% dos casos não foi precisado a hora do fato no registro da ocorrência.

Nos casos de maior incidência de homicídios as vítimas com estado civil solteiro são 46%, mas os homicídios a noite e madrugada a grande maioria, 66%, são solteiros. Os casados e com união estável representa 13% cada um deles e nestes dois casos a noite e madrugada correspondem um percentual grande na faixa horária de homicídios. Importante destacar que 20% dos registros, não foram identificados o estado civil da vítima. Isto desperta para possibilidade de utilizar o policiamento preventivo (rondas) nos períodos da noite e madrugada com viaturas, isso pode ajudar na prevenção de homicídios nas proximidades em que as morte ocorrem.

A data de nascimento foi tratada como década de nascimento, sendo utilizadas as décadas dos anos 20 a 90. A década de nascimento revela que boa parte das vítimas, 38% aproximadamente, tinha entre 19 e 28 anos de idade, pois elas nasceram na década de 80 e a data de falecimentos foi no ano de 2008. Juntando as análises se verifica que os jovens (19 a 28 anos) que mais morrem são solteiros.

No que diz respeito sobre o mês de maior incidência de homicídios, pode-se destacar os meses de dezembro e janeiro, que são os com maior número, ou seja, 13% e 10% concomitantemente. Calculando uma média mensal de 65 homicídios aproximadamente, percebe-se que janeiro e dezembro são os meses que tiveram um acréscimo de 25% e 39% respectivamente.

Apenas com estas modestas análises se verifica que focando em uma política de segurança pública destinada aos jovens e por meio de ações que venham a ser implantadas ao longo dos meses e intensificadas em janeiro e dezembro para o controle dos homicídios e reduções destes índices.

Assim, as análises criminais, com suporte de técnicas na mineração de dados, podem subsidiar o planejamento das instituições de segurança pública. Desta forma, utilizando como base nestas e outras análises, se pode criar as metas para atuar nos pontos mencionados. Claro que não é objetivos deste artigo propor formas de atuação, mas apresentar o processo de mineração de dados e texto para ajudar a desenvolver uma análise criminal mais elaborada.

6. CONSIDERAÇÕES FINAIS

A crescente criação e armazenamento de informação nos bancos de dados das instituições, sejam elas públicas ou privadas, está ocorrendo a uma velocidade enorme. Grande volumes de dados são gerados, armazenado e analisados para melhorar a competitividade das instituições e a prestação de serviço aos clientes. Esta dinâmica vem ocorrendo sobretudo nas últimas décadas com o surgimento e consolidação de modernas técnicas computacionais adivindas das tecnologias da informação e comunicação (TIC).

A utilização das TICs por muitas organização vem possibilitando que estas desenvolvam suas funções de forma mais eficiente e efetiva. Isto porque os sistemas geram as informações necessárias para acompanhar a evolução da instituição ao longo do tempo e assim aprender com as transações no seu dia a dia.

As empresas privadas estão investindo consideravelmente recursos em tecnologias para se manterem competitivas no mercado. As organizações públicas estão entrando neste caminho, aos poucos, buscando usar as tecnologias para dar mais eficiência na prestação de serviço aos cidadãos. Entretanto, isto vem acontecendo ainda de forma muito tímida.

Este artigo procuro vislumbrar a utilização das técnicas de mineração de textos e dados para subsidiar a análise criminal na segurança pública. Procurou mostrar os processos de KDD e KDT empregadas na descoberta de conhecimento em textos e dados da secretaria de segurança pública. As informações utilizadas foram dos boletins de ocorrência registrados na Polícia Civil em 2008 no Estado de Santa Catarina. O delito de homicídio foi o alvo do estudo e para isso foram usadas tipificações que envolvem estes crime.

Empregado os processos de KDT e KDD nos dados da segurança pública, o primeiro processo utilizado na mineração de textos (KDT) ajudou a refinar as informações para serem agregadas ao segundo processo de mineração de dados (KDD). Com o KDT foi possível retirar das tipificações de delitos tentados as que se concretivaram em morte das vítimas. Estas informações foram somadas as amostras especificas de homicídios dolosos para o processo de mineração de dados.

Os processos foram implementados e foi possível retirar resultados para subsidiar os analistas a criarem relatórios para apoiar a tomada de decisão das autoridades e dar embasamento ao planejamento das instituições de segurança pública.

Os resultados encotrados foram interessantes no que diz respeito a faixa de horária que ocorre os homicídios, sendo observado que a noite e na madrugada ocorrem o maior número. As pessoas solteiras são as maiores vítimas. E os meses de dezembro e janeiro de 2008 foram os que tiveram maior incidência de mortes.

Este trabalho foi desenvolvido com dados do ano de 2008, mas seria mais inriquecedor utilizar uma série histórica para retirar conclusões mais robustas. Entretanto este artigo procura mostrar que há possibilidade de se utilizar as técnicas apresentadas para auxiliar a entender a criminalidade e fomentar um enfrentamento melhor orientado pelo conhecimento produzido.

7. BIBLIOGRAFIA

ARANHA, Christian; PASSOS, Emmanuel. Automatic NLP for Competitive Intelligence. In: Emerging Technologies of Text Mining: Techniques and Applications. IGI Global Information Science Reference. 2007.

BARIN, Eliana C. Nogueira; LAGO, Decio. Mineração de Textos. Revista das Ciências Exatas e Tecnologia. v. III, n. 3, 2008.

BOBA, Rachel. Crime Analysis and Crime Mapping. Thousand Oaks, CA: Sage Publications. 2005.

CARVALHO, J. V.; Sampaio, M. C.; Mongiovi, G. Utilização de Técnicas de Datamining para o Reconhecimento de Caracteres Manuscritos. In: Simpósio Brasileiro de Banco de Dados, 14, 1999, Florianópolis. Anais. Florianópolis: UFSC, 1999, p. 235-249.

FAYYAD, U. M.; Platestsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In: American Association for Artificial Intelligence, p. 1–30, 1996.

DANTAS, G. F. L., SOUZA, N. G., As bases introdutórias da análise criminal na inteligência policial. Disponível em <www.mj.gov.br/main.asp?Team=%7B21F842C5-A1C3-4460-8A48-83F441C4808C%7D>. Acesso em: 22/09/2010.

FELIX, Aliny. Gestão da Informação através do Gerenciamento Eletrônico de Documentos (GED): suporte a gestão documental na Secretaria de Estado da Saúde de Santa Catarina (SES/SC). Florianópolis, 2009.

FERRO, Mariza; LEE, Huei Diana. O Processo de KDD – Knowledge Discovery in Database para Aplicações de Medicina. Semana de Informática de Cascavel (CEMINC) 2001.

GUERRA, Glaucio. Possibilitando alta performance na indexação com o Apache Lucene - Parte I. 2007. Disponível em: < http://www.devmedia.com.br/articles/viewcomp_forprint.asp?comp=4681>. Acesso em: 18/09/2010.

GOLDSCHMIDT, Ronaldo R.. KDD E Mineração de Dados - O Processo de KDD: Visão Geral. 2006. Disponível em: http://br.geocities.com/ronaldo_goldschmidt/downloads/KDD/060_KDD_Processo.pdf . Acesso em: 24/07/2009.

GOTTLIEB, Steven; ARENBERG, Sheldon; SINGH, Raj. Crime Analysis: from first report to final arrest. Alpha Publishing. California. 1994.

KENDAL, S. L.; CREEN, M. An Introduction to Knowledge Engineering. Springer-Verlag London Limited, 2007.

SILVA, Edson R. Gomes da. Governo Eletrônico na Segurança Pública: construção de um sistema nacional de conhecimento. Dissertação de Mestrado. Programa de Pós-graduação em Engenharia e Gestão do Conhecimento. Universidade Federal de Santa Catarina. 2009.

SILVA FILHO, Luiz Alberto da; FAVERO, Eloi Luiz;Cynthia Karina Lima de . Mining Association Rules In Data and Text – An Appplication In Public Sucurity. CONTECSI. 2010.

MAGALHÃES SILVA, E. Descoberta de Conhecimento com o uso de Text Mining: Cruzando o Abismo de Moore. Dissertação de Mestrado. Programa de Pós-Graduação em Informática - Universidade Católica de Brasília. 2002.

MAGALHÃES, Luiz C. Análise criminal e mapeamento da criminalidade – GIS. 2008. Disponível em: < http://jusvi.com/artigos/32343>. Acesso em: 10/09/2010.

The University of Waikato. Disponível em: < http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 25/09/2010.