UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RODRIGO ALEXANDRO FRANCISCHELLI

ESTUDO SOBRE A INFLUÊNCIA DA EXTRAÇÃO DE DADOS DA WEB NA DESCOBERTA DE CONHECIMENTO ESTRATÉGICO RELEVANTE

Profa. Dra. Helena G. Ribeiro
Orientadora

Caxias do Sul, Julho de 2013

“Se atribuirmos algum significado especial a um dado, este se transforma em uma informação (fato). Se os especialistas elaboraram uma norma (regra), a interpretação do confronto entre o fato e a regra constitui um conhecimento”.

Alberto Sulaiman Sade

AGRADECIMENTOS

Gostaria de registrar a minha gratidão a todas estas pessoas cuja colaboração foi determinante para a concretização deste trabalho. Agradeço primeiramente a Deus, que me deu forças e saúde para enfrentar mais esta etapa em minha vida. Em especial, agradeço aos meus pais e irmão, pela compreensão nas horas de ausência, pelo apoio nos momentos de dúvida e paciência nas horas de cansaço. O meu agradecimento a minha orientadora Prof. Dr. Helena Grazziotin Ribeiro, não só pela orientação do trabalho, mas pela colaboração e pelas valiosas sugestões durante esta etapa, pela disponibilidade, incentivo e empenho as quais foram determinantes para a execução e conclusão deste trabalho. Quero ainda agradecer a todas as pessoas que de uma forma ou outra estiveram presentes e contribuíram em algum momento para que eu atingisse aos meus objetivos.

Sumário

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
ASC II	*American Standard Code for Information Interchange*
CETEM	*Centro de Tecnologia Mineral*
CSV	*Comma Separated Values*
DATASUS	*Departamento de Informática do Sistema Único de Saúde*
DCBD	*Descoberta do Conhecimento em Banco de Dados*
IMPI	*Instituto Nacional da Propriedade Industrial*
KDD	*Knowledge Discovery in Databases*
PDF	*Portable Document Format*
PLN	*Processamento de Linguagem Natural*
SIC	*Serviço de Informações ao Cidadão*
STOPWORDS	*Lista de termos e palavras utilizadas para filtragem de texto*
TCC	*Trabalho de Conclusão de Curso*
TI	*Tecnologia da Informação*
TXT	*Arquivo de Texto*
WEKA	*Waikato Environment for Knowledge Analysis*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Este trabalho tem como objetivo investigar o impacto e a aplicabilidade da etapa de extração de dados de websites de classes heterogêneas no processo KDD (Knowledge Discovery in Databases), com o propósito de identificar o nível de obtenção dos dados e o grau de importância dos mesmos para a tomada de decisão estratégica de uma empresa. Ele descreve os conceitos relacionados a esta abordagem, as principais etapas e como funcionam os algoritmos, métodos e técnicas utilizadas para este modelo de processo. Um estudo de caso foi elaborado buscando identificar em base de dados pública o nível de informação e relevância adquirida das mesmas e o grau e resultados obtidos.

Palavras-chave: Mineração de dados, Extração do conhecimento, Descoberta de conhecimento em texto, Mineração de textos.

1. INTRODUÇÃO

Em mercados cada vez mais dinâmicos e competitivos, as empresas e as organizações com maior probabilidade de sucesso são aquelas que percebem as expectativas, necessidades e desejos dos clientes e se adequam de modo a satisfazê-los melhor que seus concorrentes. O sucesso somente ocorre se a empresa está corretamente orientada ao mercado, com compreensão e foco estratégico direcionado. Descobrir o que os clientes querem e surpreendê-los de modo a instigar a realização pessoal com a compra de um produto tornou-se um trabalho constante na vida das organizações. Para atender a esta demanda mercadológica recursos tecnológicos vem sendo aprimorados para que as necessidades das organizações possam ser supridas. A busca por informações que agregam valor aos negócios da empresa a fim de atingir seus objetivos requerem ferramentas mais adequadas para o processo de busca do mesmo. Como demonstra LEVITT (1986), a adoção de uma abordagem voltada ao mercado apresenta algumas questões básicas:

- Em que negócio estamos?
- Em que negócio poderíamos estar?
- Em que negócio gostaríamos de estar?
- O que precisamos fazer para entrar ou nos consolidar nesse negócio?

O entendimento destas perguntas pode sugerir um novo direcionamento e foco para a organização, afetando todo processo estratégico anteriormente definido pela mesma. A necessidade de melhorias constantes nos produtos e serviços de uma empresa torna a busca por informações um dos processos de maior importância, consequentemente um maior investimento em recursos é necessário para viabilizá-lo. Há alguns anos atrás, a descoberta do conhecimento era explorada apenas em bases de dados estruturados, devido a restrições de acesso a informação, onde as únicas fontes disponíveis eram a própria base de dados da empresa. Tornava-se difícil imaginar outra forma de aquisição de informação.

Porém, com a chegada da Internet e as novas formas de comunicação e relacionamentos vindas da mesma, o número de fontes de informações disponíveis chegou a um patamar estrondoso. Contudo, dificuldades também surgiram devido a dois fatores principais: o volume e a falta de estrutura das informações. Por isso, torna-se difícil agregar valor à informação disponível. Para se tirar proveito dessa enormidade de dados na Web é preciso ferramentas mais inteligentes, capazes de encontrar informações, rastrear e encontrar padrões de forma mais eficiente a fim de trazer benefícios para os negócios.

O objetivo deste trabalho é investigar o impacto e a aplicabilidade da etapa de extração de dados de websites de classes heterogêneas no processo KDD (Knowledge Discovery in Databases), com o propósito de identificar o nível de obtenção dos dados e o grau de importância dos mesmos para a tomada de decisão estratégica de uma empresa. A organização do texto está estruturada conforme descrito a seguir. O capítulo 2 apresenta o conceito, os tipos de descoberta do conhecimento, as etapas e as dificuldades encontradas neste processo. O capítulo 3 descreve o processo de extração de dados da web, explicando o modelo de mineração de textos com algumas de suas técnicas, métodos e ferramentas. O capítulo 4 apresentará o estudo de caso realizado na empresa H&TEC Soluções em Tecnologia no qual foram utilizados métodos de extração de textos da web e da ferramenta WEKA para processamento e mineração dos mesmos, conforme descritos na seção 3, buscando a identificação e relevância dos dados extraídos, a fim de obter um diferencial estratégico e competitivo no mercado. Por fim, será apresentada a análise dos resultados obtidos no estudo de caso desenvolvido.

1.1 MOTIVAÇÃO

Durante os últimos anos tem-se verificado um crescimento gigantesco na quantidade de documentos armazenados em meio eletrônico. Com esta grande massa de dados disponível, torna-se inviável efetuar um processo de análise estratégica em cima da mesma. Porém, com uma grande quantidade de

documentos disponíveis também se têm um maior potencial de informação que com as ferramentas ideais podem trazer os benefícios esperados pela organização.

A importância da etapa da extração de dados da web dentro do processo KDD influenciarão posteriormente no processo de mineração e análise dos mesmos levantando informações úteis para o suporte ás decisões, as estratégias de marketing e negócios, na busca e conquista de clientes, na descoberta de falhas em linhas de produção, entre outras. O aparecimento da Descoberta de Conhecimento em Texto vem se tornando cada vez mais importante para as organizações. Vários fatores contribuíram para o desenvolvimento de técnicas e ferramentas inteligentes para lidar com esta situação:

- A sobrecarga de informação textual disponível às pessoas e organizações.
- A grande dificuldade de extrair informação útil e relevante devido ao grande volume de dados existentes.
- O desenvolvimento rápido de técnicas e tecnologias de sistemas de administração da informação.
- Os avanços na tecnologia de computação, como computadores rápidos e arquiteturas de processamento paralelo.
- O desenvolvimento constante de técnicas de aprendizagem automática, e a inteligência artificial.
- A presença possível de ruídos, dados fora do padrão, informações perdidas.
- O descobrimento de conhecimento.

1.2 APLICABILIDADE

As tecnologias de mineração de dados podem ser aplicadas a uma grande variedade de contextos de tomada de decisão na área de negócios. A informação no ramo dos negócios vem representando um diferencial competitivo cada vez maior. Seu papel, na conjuntura atual, é bastante relevante para se alcançar os objetivos estratégicos do negócio. Um sistema deste tipo parte de uma tentativa de fornecimento pleno de informações, partindo dos dados disponíveis geradores de

informação construindo indicadores, identificação de tendências que apoiem a tomada de decisão.

As estratégias podem ser aplicadas de diversas formas em diversas áreas de negócio. Algumas delas citadas abaixo:

- Marketing – Incluem a análise do comportamento do consumidor com base em padrões de compra; determinar estratégias de publicidade e propaganda de marketing; segmentação de clientes, lojas e produtos.
- Finanças – Incluem a análise de avaliação de crédito; análise de desempenho de investimentos financeiros; avaliações financeiras e detecção de fraudes.
- Indústria – Incluem a otimização de recursos como equipamentos, força de trabalho, matéria-prima; melhoria de processo de produção;
- Saúde – Incluem a análise de eficácia de tratamentos; otimização de processos hospitalares;
- Outras Aplicações – Diversas outras áreas estão em busca da identificação de novos conhecimentos, como a área de seguros, bancos, comunicações, exploração de petróleo, etc.

2. DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (DCBD)

Nos últimos anos, o crescente aumento da disponibilidade de dados e informações providas pela internet vem tornando-se uma das principais ferramentas para a gestão estratégica de uma empresa, proporcionando conhecimento e consequentemente viabilizando alguma vantagem competitiva para os negócios. O processo de Descoberta de Conhecimento em Bases de Dados, ou KDD (*Knowledge Discovery in Databases*), nada mais é que um processo de busca e extração de conhecimento em bases de dados com o intuito de trazer algo novo, útil e pertinente, sendo que este processo ocorre de forma interativa e iterativa.

> O KDD se caracteriza por ser um processo não trivial, que busca gerar conhecimento que seja novo e potencialmente útil para aumentar os ganhos, reduzir os custos ou melhorar o desempenho do negócio, através da procura e da identificação de padrões a partir de dados armazenados em bases muitas vezes dispersas e inexploradas (THOMÉ, 2002: 11).

O processo de extração de conhecimento pode ser aplicado de forma interativa e iterativa, envolvendo diversas fases. Segundo ELMASRI, NAVATHE (2011), o KDD é composto por seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, mineração de dados e o relatório e exibição da informação descoberta. Estas etapas serão detalhadas na seção 2.2.

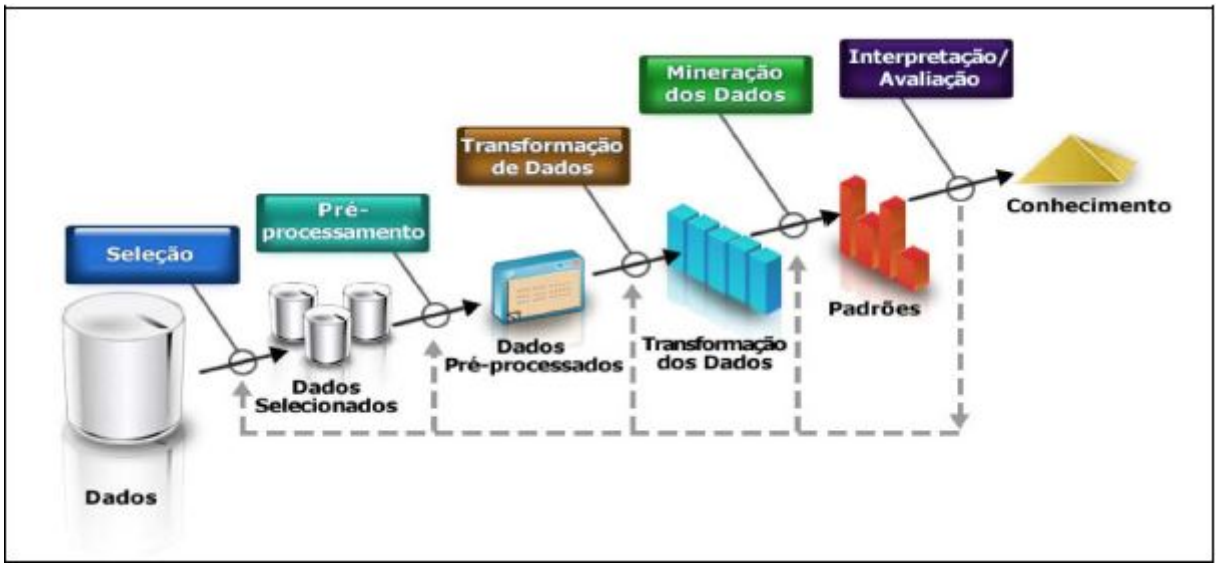


Figura 1: Etapas do processo de KDD (Adaptado de FAYYAD et al., 1996).

2.1 TIPOS DE DESCOBERTA DO CONHECIMENTO

A definição de escolha do modelo de descoberta do conhecimento tem papel fundamental no processo de captação de informações na organização. A quantidade de informação na Internet é tão grande que é impossível encontrar toda a informação que se necessita sem a correta definição da estratégia de busca da informação. Com a definição da estratégia a ser seguida, consegue-se estruturar as normas e procedimentos que ajudam a subsidiar a tomada de decisão.

2.1.1 Descoberta reativa e proativa

Segundo Choudhury e Sampler (1997) a descoberta de conhecimento ocorre de dois modos: o reativo e o proativo. No primeiro caso, o objetivo da descoberta de conhecimento é direcionado para uma solução de uma situação específica de um usuário. O usuário sabe o que quer e tem ideia de como solucionar o problema. Por outro lado, no modo proativo, o propósito de adquirir informação é exploratório sem a intervenção inicial do usuário, para utilização na detecção de problemas potenciais, oportunidades de negócio, analisar comportamentos de pessoas e mercado consumidor.

2.1.2 Inteligência competitiva

Quando se fala sobre Inteligência competitiva, em geral, busca-se encontrar a melhor definição de estratégias e ações que devem ser realizadas para beneficio da organização. Para que isso ocorra, as regras e objetivos aplicados na empresa devem estar bem definidos e de acordo com as necessidades impostas. A Inteligência competitiva visa explorar o ambiente externo da organização, utilizando-se principalmente da Internet como importante fonte de informação, buscando

avaliar o desempenho dos seus concorrentes dentro deste ambiente e obter informações que façam com que sua empresa supere seus concorrentes mesmo que as condições de mercado sejam as mesmas e buscando avaliar os pontos fortes e fracos de seus clientes e concorrentes e nas atividades de organizações que tenham produtos ou serviços similares dentro de um setor da economia.

2.2 O PROCESSO DE EXTRAÇÃO DE CONHECIMENTO

Ao iniciar este processo é fundamental que se saiba qual o propósito desta etapa, entender o problema e definir os objetivos a serem alcançados, mesmo não sabendo que tipo de resultado terá ao final. Porém é importante que ao final dessa etapa consiga-se identificar um padrão, uma métrica que possa orientar e ajudar na definição estratégica do negócio. A partir desse entendimento, podem-se seguir as etapas definidas no processo de DCBD.

2.2.1 Entendimento do negócio

A Descoberta de Conhecimento envolve a avaliação e a interpretação de padrões para que as decisões a serem tomadas acima disto possam refletir a expectativa de constituição do conhecimento esperado pela organização. A definição da estratégia se resume na identificação do problema, contudo é imprescindível nesta etapa o acompanhamento de um especialista da área para a correta identificação das necessidades e objetivos propostos pela visão estratégica empresarial.

Um entendimento claro e objetivo refletem diretamente aos objetivos a que se pretende atingir. Uma definição errada desta etapa pode tornar os resultados obtidos no processo de mineração de dados totalmente desprezíveis e ineficazes para o que foi proposto.

2.2.2 Seleção dos dados

A fase de seleção de dados envolve o processo de definição e especificação da(s) base(s) de dados a ser utilizada para o Processo de Descoberta do Conhecimento. Segundo (FAYYAD et al., 1996) nesta fase ocorre a escolha do conjunto de dados que será utilizada para análise.

Os dados que são utilizados para o processo de extração são oriundos de dois tipos principais de fontes de dados: (i) de uma Data Warehouse que segundo (INOMN, 1997) é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para auxiliar o processo gerencial de tomada de decisão; e (ii) base de dados externa, que são repositórios fora do domínio, publicados, por exemplo, em servidores de hospedagem na web, que estão legalmente de acordo com as politicas de segurança e privacidade de acesso aos mesmos, proporcionando a extração de dados (documentos de texto, artigos, jornais, revistas, textos livres, etc.) que posteriormente, na etapa de mineração de dados, trarão um determinado nível de informação para o negócio. Os dados extraídos são determinantes para a definição estratégica de negócio, a correta definição dos dados impacta diretamente na qualidade dos resultados a serem obtidos. Nessa etapa então, os dados serão organizados em uma base de dados, porém sem uma estruturação adequada para a correta análise e obtenção de conhecimento esperado.

2.2.3 Limpeza e pré-processamento dos dados

De acordo com (FAYYAD et al., 1996)**,** nesta etapa as operações básicas incluem a remoção de ruídos e desvios (se possível), a coleta da informação necessária para o modelo ou a adequação do ruído, decidindo sobre estratégias para lidar com campos incompletos, normalização ou indexação, etc...

Segundo NAVEGA (2002: 2), as bases de dados são dinâmicas, incompletas, redundantes, ruidosas e esparsas, necessitando de um pré-processamento para “limpá-las”. Métodos de redução ou transformação são utilizados para melhorar o desempenho do algoritmo de mineração de dados. É importante ressaltar que na etapa de limpeza de dados GOLDSCHMIDT & PASSOS (2005) identifica as seguintes funções que podem ser aplicadas para a limpeza de dados:

- Limpeza de informações ausentes: compreende a eliminação de valores ausentes em conjunto de dados;
- Limpeza de inconsistências: abrange a identificação e a eliminação de valores inconsistentes em conjunto de dados;
- Limpeza de valores não pertencentes ao domínio: compreende a identificação e a eliminação de valores que não pertençam ao domínio dos atributos do problema.

2.2.4 Transformação dos dados

Após o processo de seleção, limpeza e pré-processamento dos dados, algumas transformações adicionais dos dados podem ser necessárias. A ferramenta escolhida para o processo de mineração pode também ditar como se devem representar os dados. Novos dados podem ser agregados à base existente tornando o processo de identificação da informação mais rico e útil atendendo aos objetivos propostos. Nesta etapa é trabalhado o processo de redução da quantidade de dados, agrupando adequadamente os valores para não haver perda de dados.

2.2.5 Mineração de dados

A mineração de dados é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases ou fontes de dados, usando-as para efetuar decisões fundamentais. É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa.

Segundo THOMÉ (2002: 13) “Data Mining é a concepção de modelos computacionais capazes de identificar e revelar padrões desconhecidos, mas existentes entre dados pertencentes a uma ou mais bases de dados distintas”. “Data Mining se refere à mineração ou a descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados” ELMASRI, NAVATHE (2005: 620).

Para ELMASRI, NAVATHE (2011), o objetivo da mineração de dados se encontra em quatro classes distintas:

- Previsão: Consiste na identificação de comportamentos de certos atributos de dados em situações futuras.
- Identificação: Os padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade.
- Classificação: Os dados podem ser identificados em diferentes classes e categorias com base em combinações de parâmetros através do particionamento de dados.
- Otimização: Otimização do uso de recursos limitados para maximizar as variáveis de saída operando semelhantemente às técnicas de pesquisa operacional.

Conforme afirma (JESUS; MOSER; OGLIARI, 2011), cada técnica de mineração de dados ou cada implementação específica de algoritmos que são utilizados para conduzir as operações de mineração de dados adaptam-se melhor a alguns problemas que a outros, ou seja, os métodos são escolhidos e aplicados com base no problema e no objetivo estratégico que se busca. Para ELMASRI, NAVATHE (2011), as tarefas que descrevem os tipos de conhecimento descobertos

são: Regras de associação, hierarquias de classificação, padrões sequenciais, padrões dentro de séries temporais e agrupamento (*clustering).*

2.2.6 Avaliação

O processo de análise ocorre após a interpretação dos resultados minerados, onde regras, padrões possam ser validados através da utilização de técnicas e ferramentas apropriadas. Nesta etapa, além de ser possível visualizar os padrões extraídos é possível identificar ajustes podendo retornar o processo a uma das etapas anteriores, para que ao final da etapa, os dados interpretados possam se traduzir em informações úteis e em termos compreensíveis pelos usuários.

2.2.7 Consolidação do conhecimento

Segundo (FAYYAD et al., 1996), esta etapa do processo busca incorporar este conhecimento para melhorar o desempenho do sistema, incluindo a verificação e resolução de possíveis conflitos com conhecimentos previamente existentes. O conhecimento adquirido pode servir apenas para fins de documentação como também para complementação e correção dos modelos e processos existentes. Devido ao dinamismo dentro de uma base de dados, os dados acabam sofrendo grandes mudanças a qual requerem uma ação periódica de execução do processo de KDD para manter a integridade e veracidade da informação.

2.3 DIFICULDADES NA BUSCA DO CONHECIMENTO NA WEB

Apesar da grande potencialidade oferecida pela Mineração de Dados na Web, este recurso vem aos poucos ganhando espaço nas organizações. Porem, alguns fatores devem ser analisados para que a mineração tenha o sucesso esperado:

- As relações entre os atributos precisam ser muito bem definidas, caso contrário os resultados podem ser mal interpretados;
- Os testes devem ser exaustivos para que se consiga obter informações que possam levar à conclusões adequadas ao negócio;
- Uma interpretação falha pode disfarçar as falhas nos dados;
- Necessidade de usar um grande volume de dados e variáveis;
- O alto conhecimento exigido dos usuários sobre os métodos e técnicas de mineração para o correto funcionamento do processo;
- A escolha do repositório de dados para que a mineração e análise dos dados ocorram em prol do objetivo definido;
- A privacidade e a legislação em cima da obtenção e divulgação dos dados;
- Técnicas para lidar com bases de dados cada vez maiores, chegando à casa dos Terabytes;
- A velocidade com que os dados mudam faz com que os modelos gerem resultados inválidos;
- O problema da baixa qualidade dos dados;
- Complexidade dos relacionamentos entre os atributos;
- Os sistemas cada vez mais dependem de outros sistemas, gerando problemas de integração.

2.3.1 Tipos de dados e de bases

Atualmente, boa parte dos dados disponíveis para acesso eletrônico não estão mantidos em banco de dados. O principal exemplo disso são os dados da web, que geralmente possuem uma organização bastante heterogênea que mistura textos sem nenhuma informação aparente, com conjuntos de registros com

diferentes formatações, grande volume e muitos relacionamentos sem sentido. A facilidade de exploração dos dados é maior em fontes estruturadas uma vez que os dados já se encontram estruturados ajudando o processo de entendimento/organização do conteúdo dos dados. Isso implica na necessidade de se realizar buscas exaustivas nos dados ou buscas por palavras-chaves. Há diferentes fontes de dados a serem explorados, impondo problemas na sua utilização de acordo com a organização dos dados nessas fontes ou documentos.

- ***Documento livre/sem estruturação***: Texto livre é basicamente o texto onde não encontramos nenhuma forma de estrutura, e é o tipo mais encontrado. Não seguem necessariamente nenhum formato ou sequência, não seguem regras, não são previsíveis. Originalmente o objetivo de Extração de Informação era desenvolver sistemas capazes de extrair informações chave de textos em linguagem natural.
- ***Documentos semi-estruturados***: Não são textos totalmente livres de estrutura, mas também a estrutura existente não é tão severa, os textos semi-estruturados encontram-se no intermédio, ou seja, possuem uma estrutura parcial. Os dados são organizados como entidades semânticas, entidades similares são mantidas de forma agrupada, entidades em um mesmo grupo podem não ter os mesmos atributos, a ordem dos atributos não é necessariamente importante, nem todos os atributos podem ser obrigatórios, o tamanho e o tipo de um atributo podem variar dentro de um mesmo grupo.

- ***Documentos estruturados:*** Informações textuais contidas em banco de dados ou qualquer outro gênero de documento com uma estruturação rígida. Os dados são organizados em entidades de forma agrupada (relações ou classes), entidades de um mesmo grupo possuem as mesmas descrições (atributos) e as descrições para todas as entidades de um grupo (esquema) possuem o mesmo formato, o mesmo tamanho, estão todos presentes, e seguem a mesma ordem. São os dados mantidos em um SGBD.

Dados estruturados	Dados semi-estruturados
Esquema pré-definido	Nem sempre há um esquema pré-definido
Estrutura regular	Estrutura irregular (em nível tanto de atributos quanto de tipos)
Estruturada independente dos dados	Estrutura embutida nos dados
Estrutura reduzida	Estrutura extensa (para refletir as particularidades de cada dado, já que cada dado pode ter uma organização própria)
Estrutura fracamente evolutiva	Estrutura fortemente evolutiva (a estrutura dos dados modifica-se tão frequentemente quanto os seus valores)
Estrutura prescritiva (que permite definir esquemas fechados e restrições de integridades com relação à semântica dos atributos)	Estrutura descritiva
Distinção entre estrutura e dado é clara	Distinção entre estrutura e dado não é clara

Tabela 1: Diferenças entre dados estruturados e dados semi-estruturados.

2.3.2 Dados incompletos

Este problema é especialmente encontrado em bancos de dados comerciais. Atributos importantes podem ser perdidos se a base de dados não foi criada com as devidas restrições de integridade, tendo em vista a possível descoberta de conhecimento. Dados perdidos podem resultar de erros do operador, dados não preenchidos pelo vendedor durante uma transação e falhas de medidas, ou de uma revisão do processo de aquisição de dados ao longo do tempo, como por exemplo, novas variáveis são incluídas, mas eram consideradas sem importância a poucos meses atrás e ou tornaram-se obrigatórias por determinação de leis. Possíveis soluções incluem estratégias sofisticadas de estatística para identificar variáveis escondidas e dependências.

2.3.3 Redundância de dados

Um atributo pode ser redundante se ele puder ser derivado de outro armazenamento (tabela). Inconsistências em valores de atributos ou em nome de dimensões (salario, salario mensal, etc.) podem ser a causa de redundância em conjuntos de dados integrados, vindos de fontes diferentes. Uma técnica muito interessante para verificar redundância em conjuntos de dados é a utilização da análise de correlação. Esta análise mostra o grau de relacionamento entre as variáveis, fornecendo um número, indicando como as variáveis variam conjuntamente. A redundância a nível de atributo também pode ser identificada através da geração de registros idênticos gerados numa mesma entrada de dados.

2.4 ESTRATÉGIAS DE NEGÓCIO E MARKETING NA WEB

Não é possível construir um bom negócio sem uma estratégia adequada. Hoje, milhares de empresas e indivíduos de todo o mundo vendem produtos ou serviços através da Internet. A concorrência é muito grande e a persistência e a inovação são as peças chave para o sucesso. Diversos métodos existem e podem ajudar na estratégia da organização. “Empresas emergentes podem não ser gigantes globais, mas, devido à Internet, podem tornar-se competidores sérios em mercados globais. Muitas pequenas empresas de negócios estão começando a utilizar essa tecnologia para atingir clientes ao redor do mundo, comunicar-se com fornecedores, acelerar o fluxo de vendas e de pedidos, melhorar o serviço ao consumidor e tornar suas operações mais eficientes. Mais de 80% dos atuais donos de pequenos negócios contam com a Internet para auxiliar suas empresas a crescerem” (LAUDON & LAUDON, 2000).

O marketing e a comunicação são fundamentais para o sucesso de um empreendimento. Uma estratégia de marketing direcionada, um bom trabalho no site e redes sociais da empresa e a qualidade dos materiais de divulgação são fatores positivos para que o negócio ganhe diferencial competitivo no Mercado. A seguir

algumas das estratégias mais utilizadas:

- **Sites de Busca -** Os Sites de busca são de longe a forma mais popular de localização de informações e produtos na Web. Justamente por isso ocupa um lugar de grande destaque entre as estratégias de web marketing.

- **E-Mail Marketing -** As pesquisas de comportamento do usuário indicam que receber e enviar e-mails é disparado a atividade mais realizada pelos internautas, seguida de longe pela leitura de notícias e diversão. Quando se fala em mail Marketing, as pessoas pensam imediatamente em mala-direta eletrônica, não autorizada (spam). Na verdade, a grande força do e-mail Marketing é sua agilidade como canal de comunicação com o cliente.

- **Programa de Afiliados -** Um programa de afiliados é a montagem de rede de sites que divulgam o negócio e enviam potenciais clientes para o site do comerciante em troca de comissões sobre as vendas. Para implantar e gerenciar um programa de afiliados o comerciante precisa adquirir uma solução (software) que faça o rastreamento dos afiliados e possibilite o controle das vendas pelo comerciante e também pelo afiliado.

- **Banner, Links Patrocinados -** Os anúncios são peça fundamental na disputa pelo cliente e as empresas reservam parte expressiva de seus orçamentos em anúncios na TV, Rádio e mídia impressa. A grande novidade são os links patrocinados - anúncios pagos em sites de busca -- que estão ganhando fatias expressivas dos recursos destinados à publicidade on-line.

- **Compra Coletiva -** A compra coletiva é a mais nova forma de divulgação da Internet. Trata-se de uma parceria entre o comerciante e o site de compra coletiva para a divulgação de uma oferta com grande taxa de desconto (geralmente em torno de 60%). As ofertas duram um período curto de tempo, mas o suficiente para trazer uma grande quantidade de novos visitantes para o comércio.

Para um melhor aproveitamento destas estratégias é fundamental acompanhar e quantificar o resultado do investimento realizado, ou seja, quantas ações desejadas foram realizadas, ex: visitas, compras, downloads, etc. e comparar o retorno atingido com o investimento realizado. Isto poderia significar, por exemplo, responder a seguinte questão: “quanto me custa a venda do produto “X”, se eu utilizar a estratégia “A” ? ou a “B” ? ou a “C” ? ”. Isso pode ser feito com a utilização das ferramentas de mineração na web e representa uma das grandes vantagens das estratégias de Marketing na Web, possibilitando a mensuração com maior precisão e desempenho dos resultados obtidos.

3. EXTRAÇÃO DE DADOS DA WEB

Na última década a utilização da Internet no meio corporativo vem aumentando extraordinariamente, conforme pode ser visto na figura 2, que mostra o crescente aumento dos domínios de internet nos últimos anos. O grande potencial de negócio agregado ao uso da Internet, onde as empresas podem disponibilizar seus produtos, serviços e informações se tornou uma grande ferramenta de apoio estratégico para os negócios. Devido ao crescimento e grande volume de dados destas fontes de informações disponíveis na web, tornou-se necessário à utilização de ferramentas automatizadas para se extrair, filtrar e avaliar informações. A World *Wide Web* acabou por se tornar um grande repositório *on-line* de textos e páginas. Segundo CHEN, H. (2001), pesquisas realizadas neste período mostram que cerca de 80% do conteúdo *on-line* está em formato textual e dentro das empresas o mesmo percentual reflete a organização não estruturada dos dados.

3.1 INTRODUÇÃO A WEB MINING

Segundo ELMASRI, NAVATHE (2011), mineração de dados (*Data Mining)* refere-se à mineração ou descoberta de novas informações em termos de padrões ou regras com base em grandes quantidades de dados. A mineração de dados da web (*Web Mining*) se utiliza deste mesmo conceito seguindo os processos de descoberta de conhecimento em bases de dados (KDD), porém, a base de dados e as formas de extração destes dados necessitam de processos e ferramentas voltados ao ambiente Web.

Para OBERLE et al. (2003), mineração da Web é a aplicação de técnicas de Mineração de Dados ao conteúdo, estrutura e utilização dos recursos da Web, auxiliando na descoberta de estruturas tanto locais quanto globais (modelos ou padrões) entre as páginas Web.

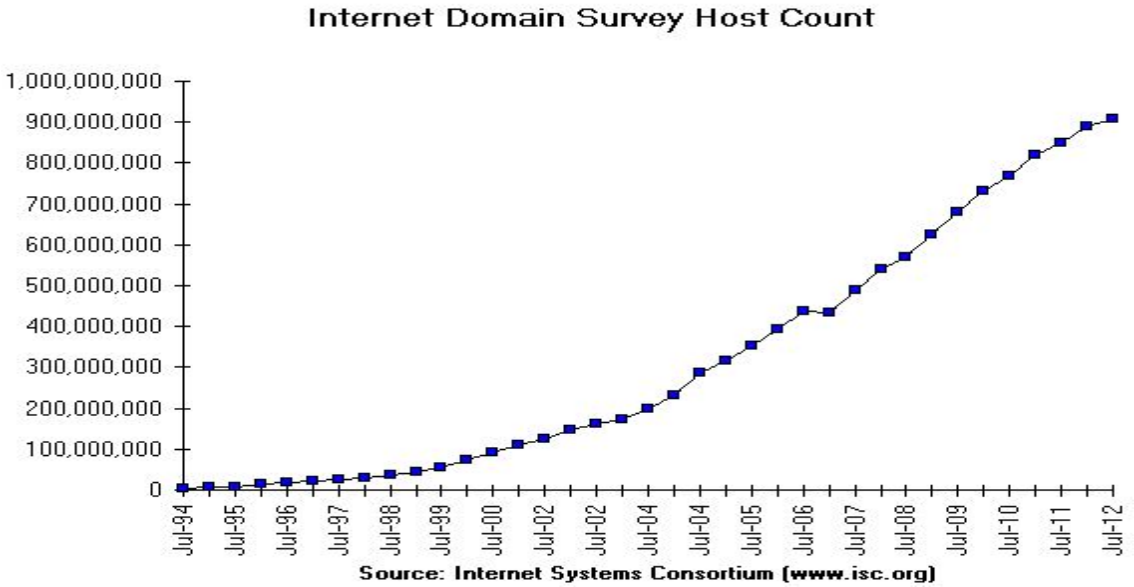


Figura 2: Internet Systems Consortium (www.isc.org). Relação Domínio x Período.

3.2 MINERAÇÃO DE TEXTOS

Devido ao grande volume de dados textuais armazenados no ambiente *Web*, torna-se difícil a análise e compreensão dos dados e informações nela contida. A mineração de textos permite a transformação desses dados não estruturados em conhecimento útil para as organizações. Este processo vem-se tornando bastante utilizado pelas empresas a nível estratégico proporcionando inovação e diferencial competitivo dentro do mercado. Sistematicamente isto leva a empresa a descobrir como agir e reagir ao mercado e seus potenciais concorrentes, utilizando-se da análise de seu ambiente interno com o ambiente externo.

Segundo ARANHA, PASSOS (2006) a mineração de textos, em geral, se refere ao processo de extração de informações de interesse e padrões não triviais ou descoberta de conhecimento em documentos de texto não estruturados. Pode ser visto como uma extensão da mineração de dados ou da descoberta de conhecimento em bases de dados estruturadas. Para EBECKEN, LOPES, ARAGÃO COSTA (2003), a Mineração de Textos pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais.

Segundo JUNIOR, RIBEIRO (2007), para o processo de extração de textos, há duas abordagens diferentes: a **Análise Semântica**, na qual o foco é na funcionalidade dos termos dos textos, tratando do significado das palavras e a **Análise Estatística**, que se preocupa com a frequência de aparição de cada termo. A Tabela 1 resume que áreas de conhecimento estão mais ligadas com os dois tipos de análise.

Análise Semântica	Análise Estatística
Ciência Cognitiva	Recuperação de Informação
Processamento de Linguagem Natural	Estatística
Mineração de Dados	Aprendizado de Máquina
Web Mining	Inteligência Computacional
	Mineração de Dados
	Web Mining

Tabela 2: Abordagens para a Análise de Textos e suas Áreas de Conhecimento.

3.2.1 Etapas do Processo de Mineração de Textos

Apesar da quantidade de informação extraída dos mecanismos de busca e recuperação, apenas uma parte da web é pesquisada, enquanto uma grande parte do conteúdo fica inacessível devido a restrições dos domínios e das ferramentas de busca. A dificuldade de se encontrar informações relevantes apesar da grande quantidade de informação na web e das diversas ferramentas disponíveis para buscá-las, torna a obtenção dos resultados retornados muitas vezes insatisfatórios. A busca por conjuntos de informações como livros, artigos, jornais, revistas, textos, etc..., em um repositório que agregue dos mais diversos assuntos, tratando de áreas relacionadas a negócios, informações do dia-a-dia, textos científicos, estatísticas, etc., possuem diversas restrições, onde o máximo que se consegue extrair é resumos e sumários dos textos. A grande maioria das bases públicas mundiais ditas de “acesso público“ como, por exemplo, a Proquest, GlobalResearch, Reuters, CETENFolha, são na verdade em sua maioria todas restritas quando o documento trata sobre dados qualitativos a um determinado ambiente ou grupo de domínio, e por muitas vezes a disponibilização do mesmo somente se dará através da compra do documento, gerando um determinado custo para quem pretende se arriscar a utilizar esta base de dados para análise e pesquisas em busca de um diferencial estratégico. As fontes de dados encontradas e que realmente possuem acesso público são as fontes que tratam de dados quantitativos como, por exemplo, a DATASUS, CETEM, IMPI e a SIC.

Conforme REZENDE, MARCACINI, MOURA (2011), o processo de extração e organização de conhecimento através da mineração de textos pode ser dividido em três fases principais: pré-processamento dos documentos, extração de padrões com agrupamentos de textos e avaliação do conhecimento, como mostra a Figura 3.

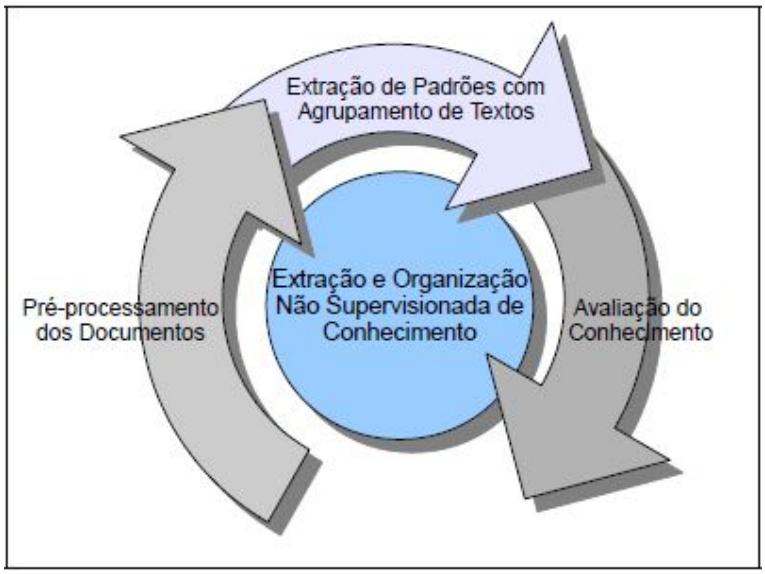


Figura 3: Fases da extração e organização não supervisionada de conhecimento.

3.2.1.1 Pré-processamento dos documentos

Nesta primeira fase é realizado um estudo do domínio da aplicação e a definição de objetivos e metas a serem atingidas no processo de Data Mining. Algumas questões importantes devem ser respondidas segundo Rezende (2003), nessa fase de identificação do problema, como:

- Quais são as principais metas do processo?
- Quais critérios de desempenho são importantes?
- O conhecimento extraído deve ser compreensível a seres humanos ou um modelo do tipo caixa-preta é apropriado?
- Qual deve ser a relação entre simplicidade e precisão do conhecimento extraído?

Nesta etapa, os dados disponíveis para análise, geralmente não estão em um formato adequado para a Extração de Conhecimento. Diversas transformações nos dados podem ser realizadas, entre elas:

- Extração e Integração: os dados disponíveis podem se apresentarem em diversos formatos, como arquivo-texto, arquivos no formato de planilhas, Banco de Dados ou Data Warehouse;

- Transformação: após a extração e integração dos dados, algumas transformações podem ser realizadas nos dados;

- Limpeza: os dados disponíveis para aplicação dos algoritmos de Extração de Padrões podem apresentar problemas advindos do processo de coleta. Esses erros podem ser de digitação ou leitura de dados pelos sensores;

- Seleção e Redução de Dados: em virtude das restrições de capacidade de memória ou desempenho (tempo de processamento), o número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de Extração de Padrões. Como solução, pode ser aplicado um método para redução dos dados antes de começar a busca de padrões.

3.2.1.2 Extração de padrões com agrupamentos de textos

Esta etapa esta direcionada ao cumprimento dos objetivos definidos na identificação do problema. Nessa fase é realizada a escolha, configuração e execução de um ou mais algoritmos para extração do conhecimento.

3.2.1.3 Avaliação do conhecimento

O conhecimento extraído pode ser utilizado na resolução de problemas da vida real, seja por meio de um Sistema Inteligente ou de um ser humano como apoio a algum processo de tomada de decisão. Um dos objetivos principais do Processo de Extração do Conhecimento é que o usuário possa compreender e utilizar o conhecimento descoberto. Após a análise do conhecimento, caso este não seja de interesse do usuário final ou não cumpra com os objetivos propostos, o processo de extração pode ser repetido ajustando-se os parâmetros ou melhorando o processo de escolha dos dados para a obtenção de resultados melhores numa próxima iteração.

3.2.2 Métodos de Descoberta do Conhecimento em Textos

Devido aos bancos de dados textuais obtidos da web serem desestruturados, a aplicação de técnicas para mineração dos dados torna-se mais difícil do que o processo de mineração convencional efetuado em dados estruturados. Os métodos apresentados a seguir nas seções 3.2.2.1 e 3.2.2.2 tratam de estruturar estes textos obtidos na web para que nos mesmos possam ser aplicadas as técnicas de mineração de dados para obtenção do conhecimento.

3.2.2.1 Recuperação de Informação

Este método atua sobre uma coleção de documentos onde filtros são apresentados para o processo de preparação dos dados apresentados a fim de chegar a uma conclusão plausível para o problema proposto. A Recuperação de Informação é feita através da entrada do usuário, gerando consultas para que os documentos relevantes sejam encontrados. Os principais modelos de recuperação

de informações são: Booleano, Espaço-vetorial, Probabilístico, Difuso, Busca Direta, Aglomerado e Contextual (SPARCK 1997).

3.2.2.2 Extração de Informação

A extração de informação é usado na área de Processamento de Linguagem Natural (PLN) com o objetivo de transformar dados semi-estruturados ou desestruturados (textos) em dados estruturados que serão armazenados em um banco de dados para obtenção do conhecimento. O processo extrai tipos específicos de informações contidas nos textos, havendo a possibilidade de construir regras de extração específicas do domínio. Posteriormente estas informações podem ser usadas como simples dados de entrada para o processo de mineração de dados. (FELDMAN, 2007).

3.2.3 Técnicas de Descoberta do Conhecimento

A partir dos dados já estruturados, técnicas podem ser aplicadas sobre a base de dados a fim de executar o processo de mineração dos dados. Nesta seção são apresentadas as principais técnicas utilizadas para o processo de descoberta do conhecimento.

3.2.3.1 Associação

O processo de extração por associação gera regras do tipo “Se X Então Y” a partir de um banco de dados de transações, onde X e Y são conjuntos de itens que concorrem em varias transações (WIVES, 2000). Esta técnica é bastante utilizada em mineração de textos, com o objetivo de descobrir as associações existentes entre termos e categorias de documentos. São extraídas informações de partes dos

textos, as quais são estruturadas em slots. As representações dos textos (em slots) são analisadas para encontrar similaridades e diferenças de informações. Por exemplo, similaridades podem ser detectadas em dados de um mesmo slot. As informações extraídas dos artigos são combinadas para exprimir contradição, adição, refinamento de informação, concordância, falta de informação, etc.

3.2.3.2 Sumarização

As técnicas de Sumarização são utilizadas para agrupar um conjunto de dados considerados similares em clusters ou grupos, ou seja, um resumo gerado a partir das palavras ou frases mais importantes.

3.2.3.3 Clusterização

É um método que identifica padrões homogêneos entre documentos, alocando-os em grupos conforme o seu grau de similaridade. A diferença fundamental entre clusterização e classificação é que no primeiro não existem classes predefinidas para classificar os registros em análise. Os registros são agrupados em função de suas similaridades conforme o conjunto de atributos selecionados. (MICHAEL, GORDON, 1997)

3.2.3.4 Classificação/Categorização

A classificação/categorização é uma técnica para identificar a classe ou categoria a qual pertence determinado documento baseando-se em um algoritmo pré-definido. O algoritmo analisa todos os documentos, aprende as regras e

armazena em uma base de conhecimento. Após os documentos são categorizados e estabelecidos a classe a qual pertence.

Conforme WIVES (2000), os métodos de classificação utilizam das seguintes técnicas:

a. Regras de inferência: baseiam-se em um conjunto de características para identificar a classe de um documento. São especificas para cada domínio;
b. Redes Neurais Artificiais: induzem um conjunto de regras a partir de arquivos de treinamento. São capazes de detectar mudanças nos dados. Seus resultados não são facilmente compreendidos;
c. Método de similaridade de vetores: as classes são representadas por vetores de palavras. O documento é comparado com cada vetor para ser agrupado;
d. Árvores de decisão: utilizam técnicas de aprendizado de máquina para induzir regras;
e. Classificadores de Bayes: tem como base a teoria da probabilidade. Informam a probabilidade de um documento pertencer a uma classe.

3.2.4 Tipos de Ferramentas de Extração

Com a utilização destes métodos de busca de dados que tem por objetivo identificar, indexar e devolver o(s) resultado(s) encontrado(s) da(s) consulta(s) em um mecanismo estruturado e organizado, o processo busca automatizar a obtenção de dados, proporcionando um considerável ganho de tempo e uma melhor definição dos dados para o processo de mineração de dados.

Há métodos e ferramentas que permitem extrair dados da web. São apresentados os dois principais métodos: motores de busca e os diretórios.

3.2.4.1 Motores de busca

Os motores de busca são ferramentas que auxiliam a busca por informações armazenadas em um sistema computacional a partir de palavras chave pré-definidas a fim de reduzir o tempo necessário para encontrar informações. O primeiro motor de busca baseado em robôs foi o WebCrawler lançado em abril de 1994. Para ARANHA, PASSOS (2006), WebCrawlers são programas especializados em navegar na internet, de forma autônoma com o objetivo de explorar e coletar documentos. Os crawlers fazem o trabalho automaticamente e recursivamente visitando páginas da Web, lendo, copiando e identificando os hiperlinks contidos nelas. O crawler é muito utilizado no processo de extração de dados onde as páginas visitadas e baixadas agilizam o processo de busca da informação.

Sendo os motores de busca o mecanismo mais utilizado para recuperação das informações na Web, seguindo o mesmo caminho do WebCrawlers diversas outras ferramentas surgiram posteriormente, tal como o Google, Northern Light Search, Snaket, Clusty, Kartoo, MetaCrawler, Grokker, Terrier e Copernic. Recentemente novas ferramentas surgiram com o crescimento das redes sociais. O Facebook e o Twitter, por exemplo, disponibilizam suas APIs aos desenvolvedores para que novos serviços possam ser desenvolvidos baseados nos dados e informações que estas redes sociais proporcionam. As APIs são mais adequadas para a coleta de dados de redes sociais online, pois oferecem os dados em formatos estruturados.

3.2.4.2 Diretórios

Os diretórios foram a primeira solução proposta para organizar e localizar os recursos. Eram utilizados quando o conteúdo da Web ainda era pequeno permitindo a busca não automatizada de informações na Web. Os serviços de diretório contêm apontadores para sites, organizados por categorias. Um webmaster regista um link para o seu site numa categoria do diretório fornecendo também uma descrição de

todo o conteúdo do site. Os diretórios permitem encontrar rapidamente links para sites sobre um determinado tema.

3.2.5 DCBD x Mineração de Textos

O campo de pesquisa da DCBD preocupa-se com o desenvolvimento de métodos e técnicas que buscam trazer sentido aos dados. Seu princípio básico é transformar dados sem nenhum valor estratégico de negócio em informações úteis para a organização. Ambas as formas de descoberta do conhecimento utilizam-se das mesmas técnicas tradicionais de transformação de dados em informação. A grande diferença destes é que na mineração de textos o processo de descoberta tem um trabalho maior na identificação e extração dos dados para análise, onde os dados estão todos eles em formato não estruturado. Enquanto que, na DCBD as bases de dados para o processo de identificação da informação encontra-se em formato estruturado tendo um foco maior no processo de mineração propriamente.

Após a etapa de pré-processamento, tanto a DCBD como a mineração de textos seguem os mesmos caminhos até a obtenção do resultado esperado, a informação.

4. PROPOSTA DE ESTUDO DE CASO

No mercado de trabalho, planejamento estratégico é requisito primário para um bom andamento dos negócios. A utilização dos recursos da web vem tornando-se cada vez mais presente na vida das organizações. Os dados e informações hoje presentes na Internet estão disponíveis a todos que buscam o saber, o conhecimento. Não existem mais limites para a busca do conhecimento na internet. Porém, da mesma forma que a informação está disponível a todas as pessoas, as mesmas impõem certas restrições de acesso para obtenção de dados que estrategicamente podem se tornar um grande diferencial para os negócios.

Neste capítulo será aprofundado o estudo da obtenção do conhecimento através da proposta de um estudo de caso de extração de dados em ambiente web em busca de informações mais significativas para o desenvolvimento estratégico do negócio. Após a execução do processo de descoberta do conhecimento, será feito uma avaliação sobre a importância estratégica das informações/conhecimento obtidos. O processo segue as etapas da Descoberta do Conhecimento em Textos que são: pré-processamento dos documentos, extração de padrões com agrupamentos de textos, avaliação do conhecimento (REZENDE, MARCACINI, MOURA, 2011).

4.1 H&TEC SOLUCOES EM TECNOLOGIA LTDA

A H&TEC Soluções em Tecnologia Ltda foi fundada em setembro de 2010 e surgiu com a proposta de tornar mais viável o acesso as tecnologias computacionais para seus clientes, proporcionando o crescimento dos negócios aliados ao que tem de melhor da tecnologia com um custo baixo. A H&TEC atua com foco direcionado na estratégia de negócio do cliente buscando soluções e alternativas tecnológicas que se encaixam as necessidades que visam o crescimento de seus clientes.

Desta forma, a H&TEC trabalha sempre em busca de inovação, novos processos, recursos que possam auxiliar na busca de soluções que atendam satisfatoriamente aos requisitos com rapidez, qualidade e baixo custo para os clientes que buscam ingressar ou aprimorar seu parque tecnológico.

4.2 DEFINIÇÃO DOS DOCUMENTOS PARA ANÁLISE

Para este trabalho, os dados que serão utilizados serão extraídos da base de dados da Internet Archive Text. A base de dados tem por característica ser aberta ao publico, ou seja, todos tem acesso as publicações existentes nela, tornando-se uma boa fonte de informação. A coleção a ser utilizada para os testes será composta por documentos da *Internet Archive Text* (http://archive.org/details/texts) a ser

apresentado na seção 4.2.2 a qual possui assuntos relacionados a negócios das mais diversas áreas mercadológicas. O processo de descoberta do conhecimento será desenvolvido seguindo o modelo proativo, conforme explicitado no capitulo 2.1.1. A seguir será apresentada a forma de extração, a ferramenta utilizada para mineração e a base de informações a serem extraídas.

4.2.1 Ferramentas de mineração de textos

Para o estudo foram selecionadas duas ferramentas de software de mineração de dados para trabalhar a descoberta do conhecimento. A primeira ferramenta avaliada foi o *WEKA* (http://www.cs.waikato.ac.nz/ml/weka/). Desenvolvida na linguagem *Java* pela Universidade de *Waikato* na Nova Zelândia, trabalha com diversas técnicas de *Data Mining*, além de ser um *software* livre e de fácil manuseio. A ferramenta foi selecionada considerando a confiabilidade dos algoritmos implementados e o fato da mesma ser de domínio público. A ferramenta Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Weka contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. Também é bem apropriado para o desenvolvimento de sistemas de aprendizagem de novas máquinas. A outra ferramenta selecionada foi o RapidMiner, um ambiente para a aprendizagem de máquina, mineração de dados, mineração de texto, a análise preditiva, análise e negócios. Ele é usado para pesquisa, educação, treinamento, prototipagem rápida, desenvolvimento, aplicação e aplicações industriais. Este também tem a característica de ser um software livre e vem sendo distribuído desde 2004. O RapidMiner possui uma gama maior de algoritmos implementados se comparado ao WEKA, porém o RapidMiner possui menos referências técnicas tornando-o mais difícil a sua utilização. A ferramenta WEKA foi escolhida para ser utilizada neste trabalho por ser uma ferramenta de software livre, simples, de fácil manuseio e com um bom referencial técnico. Esta ferramenta de mineração de dados é muito utilizada por profissionais que desejam aprender os conceitos básicos sobre mineração de dados. Através de sua interface gráfica

(conhecida como Weka Explorer) é possível conduzir processos de mineração de dados de forma simples, realizando a avaliação dos resultados obtidos e a comparação de algoritmos. Além disso, a ferramenta oferece recursos para a execução de tarefas relacionadas ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos. A ferramenta Weka trabalha preferencialmente com bases de dados no **formato texto**. Por esta razão, quase todos os tutoriais e apostilas sobre a ferramenta disponibilizados na Internet mostram como utilizar a Weka para minerar bases de dados estruturadas nos formatos ARFF ou CSV.

4.2.2 Base de dados da Internet Archive Text

A *The Brown Corpus of Standard American English* uma organização sem fins lucrativos, foi uma das primeiras a criar uma biblioteca digital com a proposta de processamento de textos. É constituído de textos de diversos gêneros, contando atualmente com aproximadamente três milhões e setecentos mil documentos. O acesso é livre para pesquisadores, historiadores, acadêmicos e público em geral. A Internet Archive Text, como é conhecido, contém uma ampla gama de documentos de ficção, livros populares, livros infantis, textos históricos e livros acadêmicos. Esta coleção é aberta à comunidade para a contribuição de qualquer tipo de texto. A base de dados conta com um acervo de 3.702.099 itens divididos em oito sub-coleções (American Libraries, Canadian Libraries, Universal Library, Community Texts, Project Gutenberg, Biodiversity Heritage Library, Children's Library, Additional Collections). A opção por essa base se deu em virtude da grande variedade de documentos que referenciam assuntos diversos, tanto a nível científico, acadêmico ou do cotidiano.

Os documentos serão extraídos em formato .TXT, que consiste geralmente em um arquivo com pouca formatação podendo ser facilmente lidos ou abertas por qualquer programa que lê texto. A base textual utilizada no presente estudo refere-se à área de estudo sobre estratégia de negócios, e foi obtida a partir do site http://archive.org/details/texts.

4.3 PRÉ-PROCESSAMENTO DOS DOCUMENTOS

. O objetivo do pré-processamento é extrair de textos escritos em língua natural, não estruturado em uma representação estruturada, concisa e manipulável por algoritmos de agrupamento de textos. Para tal, são executadas atividades de tratamento e padronização na coleção de textos, seleção dos termos (palavras) mais significativos e, por fim, representação da coleção textual em um formato estruturado que preserve as principais características dos dados. (FELDMAN, SANGER, 2006).

4.3.1 Seleção e limpeza dos dados

Na etapa de pré-processamento é necessário efetuar a seleção de dados considerada importante para a organização, ou seja, selecionar um conjunto de dados, pertencentes a um domínio, para que, a partir de um critério definido, ele possa ser analisado. O primeiro passo é a eliminação de *stopwords*, que são os termos que nada acrescentam à representatividade da coleção ou que sozinhas nada significam, como artigos, pronomes e advérbios. Essa eliminação reduz significativamente a quantidade de termos diminuindo o custo computacional das próximas etapas. Posteriormente, busca-se identificar as variações morfológicas e termos sinônimos. Para tal, pode-se, por exemplo, reduzir uma palavra à sua raiz por meio de processos de *stemming* a qual se tem como principal benefício a eliminação de sufixos que indicam variação na forma da palavra, como plural e tempos verbais.

Por exemplo, as palavras “*learning*” e “*learned*” são ambas convertidas para o *stem* “*learn*”. O algoritmo de *stemming* utilizado neste trabalho é o algoritmo de Porter [Porter 97], o qual varre uma string numa série de passos descritos a seguir:

- Primeiro passo: remove o plural, incluindo casos especiais tais como *“sses” “ies”*.
- Segundo passo: une padrões com alguns sufixos tais como: “fullness” -> “ful”, “ousness” -> “ous”.

Nessas transformações, removem-se os sufixos e os substituem por suas “raízes”.

- Terceiro passo: ocorre a manipulação das transformações necessárias para algumas palavras especiais conforme apresentadas a seguir: *“alize” -> “al”, “alise” -> “al”.*
- Quarto passo: a palavra analisada é checada perante mais sufixos, no caso da palavra ser composta incluindo: *“al”, “ance”, “ence”.*
- Quinto e último passo: checa se a palavra termina em vogal, fixando-a apropriadamente. Este algoritmo particular serve também para analisar e separar afixos e alguns prefixos simples, tais como: *”kilo”, “micro”, “milli”, “intra”, “ultra”, “mega”, “nano”, “pico”*.

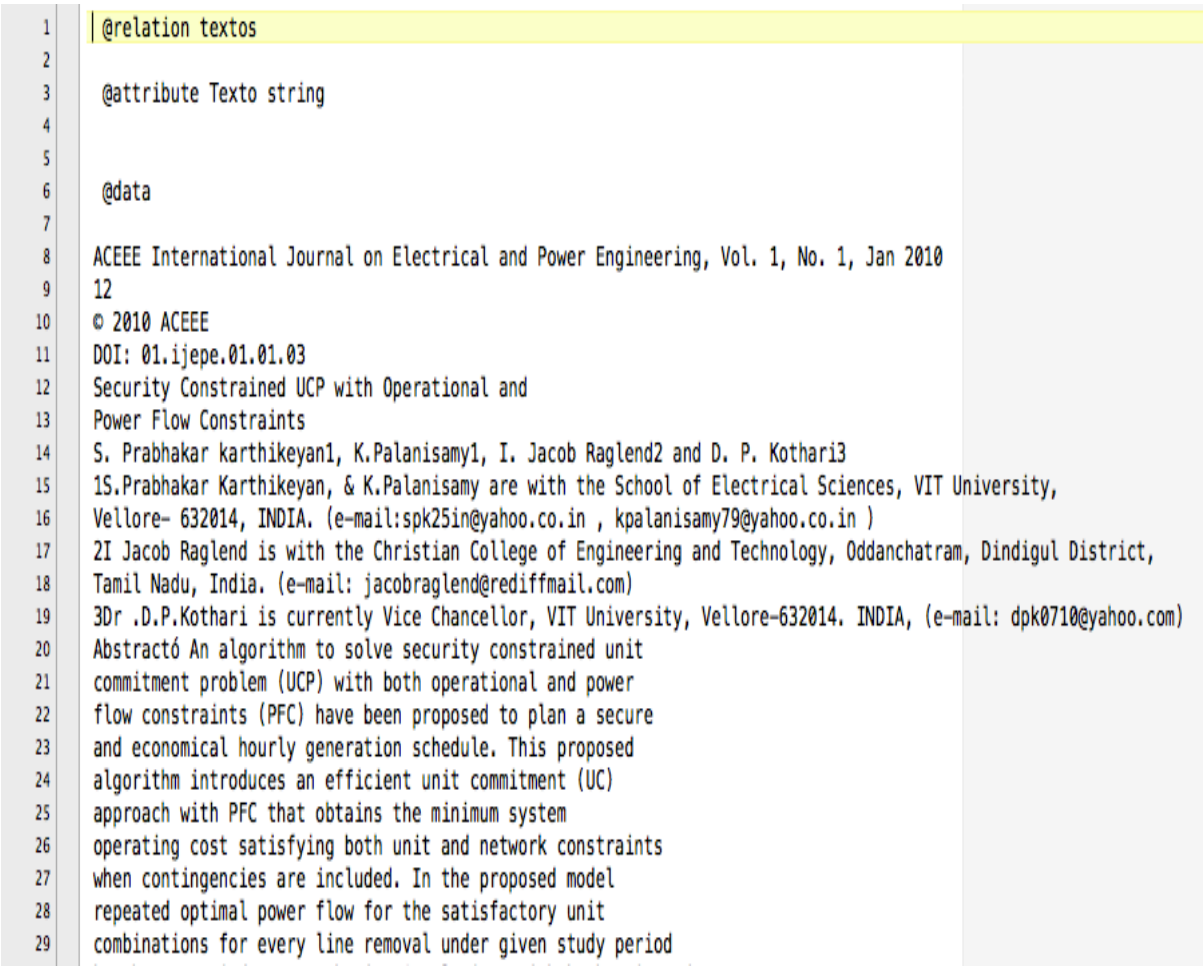
Os algoritmos em geral não se preocupam com o uso no qual a palavra se encontra (ARANHA, PASSOS, 2006), ou mesmo usar dicionários ou *thesaurus* que conforme EBECKEN, LOPES, ARAGÃO COSTA (2003) é utilizado para determinar as relações que são significativas entre termos próximos, sendo analisadas associações diretas, indiretas e categorias comuns de termos

Além disso, é possível buscar na coleção a formação de termos compostos, ou *n-gramas*, que são termos formados por mais de um elemento, porém com um único significado semântico (CONRADO, MARCACINI, MOURA e REZENDE, 2009). Após a seleção de dados, é necessário aplicar métodos de tratamento, pois na maioria das vezes os dados disponíveis para análise encontram-se em um formato inadequado para realização do processo, denominado de limpeza dos dados.

Para esta etapa, foram selecionados e extraídos os textos provenientes da base de dados conforme relatado na seção 4.2.2. Foram utilizadas palavras chave previamente selecionadas conforme o contexto de pesquisa abordado para o processo de seleção dos arquivos texto. As palavras utilizadas estão listadas a seguir: *business, knowledge, internet security, business strategy, competitive edge, social network, IT, technology.*

Os mesmos são extraídos em formato de arquivo pdf e posteriormente extraídos para texto puro. Percebe-se no arquivo texto que muito ”lixo“ (palavras, caracteres especiais, números, etc...) são encontrados dentro destes documentos,

conforme mostra a Figura 4 abaixo, sendo necessário efetuar uma limpeza prévia destes arquivos antes da execução do processo de mineração dos textos.



```
@relation textos

@attribute Texto string


@data

ACEEE International Journal on Electrical and Power Engineering, Vol. 1, No. 1, Jan 2010
12
© 2010 ACEEE
DOI: 01.ijepe.01.01.03
Security Constrained UCP with Operational and
Power Flow Constraints
S. Prabhakar karthikeyan1, K.Palanisamy1, I. Jacob Raglend2 and D. P. Kothari3
1S.Prabhakar Karthikeyan, & K.Palanisamy are with the School of Electrical Sciences, VIT University,
Vellore- 632014, INDIA. (e-mail:spk25in@yahoo.co.in , kpalanisamy79@yahoo.co.in )
2I Jacob Raglend is with the Christian College of Engineering and Technology, Oddanchatram, Dindigul District,
Tamil Nadu, India. (e-mail: jacobraglend@rediffmail.com)
3Dr .D.P.Kothari is currently Vice Chancellor, VIT University, Vellore-632014. INDIA, (e-mail: dpk0710@yahoo.com)
Abstractó An algorithm to solve security constrained unit
commitment problem (UCP) with both operational and power
flow constraints (PFC) have been proposed to plan a secure
and economical hourly generation schedule. This proposed
algorithm introduces an efficient unit commitment (UC)
approach with PFC that obtains the minimum system
operating cost satisfying both unit and network constraints
when contingencies are included. In the proposed model
repeated optimal power flow for the satisfactory unit
combinations for every line removal under given study period
```

Figura 4: Documento extraído da internet.

Depois de finalizado o processo de extração dos textos, todos os arquivos foram agrupados em um único arquivo texto para que o processo de limpeza inicial pudesse ser efetuado. Nesta primeira etapa apenas os caracteres especiais, como mostra a Figura 5 abaixo (utilizados conforme definição da tabela ASC II), foram removidos.

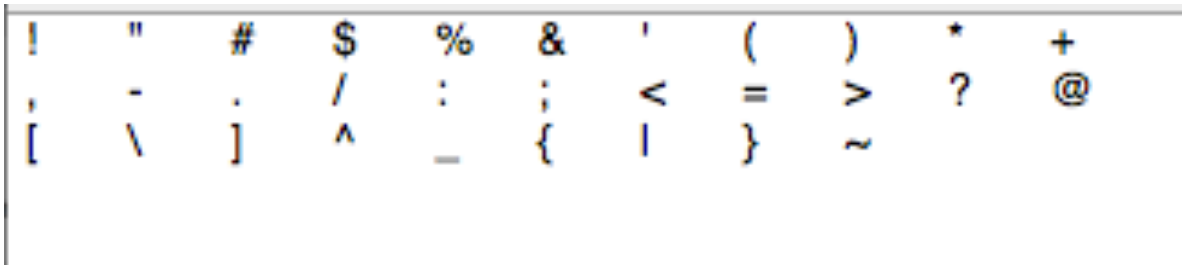


Figura 5: Tabela de caracteres especiais ASC II.

A remoção destes caracteres foi efetuada manualmente através de uma ferramenta de editor de texto. Este procedimento é necessário para que o arquivo possa ser importado para a ferramenta WEKA. Após a limpeza inicial, o arquivo texto é migrado para o formato de leitura padrão do WEKA, o arquivo ARFF. O WEKA suporta alguns formatos de arquivos, como por exemplo, o CSV e C4.5, além do formato próprio da ferramenta, o ARFF. Esta última possui um formato de arquivo em texto, contendo atributos e seus respectivos dados, para serem manipulados pela aplicação. O resultado obtido é demonstrado na Figura 6.

```
1  @relation textos
2
3  @attribute Texto string
4
5
6  @data
7
8  ACEEE International Journal on Electrical and Power Engineering Vol  1 No  1 Jan 2010
9  12
10   2010 ACEEE
11 DOI  01 ijepe 01 01 03
12 Security Constrained UCP with Operational and
13 Power Flow Constraints
14 S  Prabhakar karthikeyan1 K Palanisamy1 I  Jacob Raglend2 and D  P  Kothari3
15 1S Prabhakar Karthikeyan  K Palanisamy are with the School of Electrical Sciences VIT University
16 Vellore  632014 INDIA  e mail spk25in yahoo co in  kpalanisamy79 yahoo co in
17 2I Jacob Raglend is with the Christian College of Engineering and Technology Oddanchatram Dindigul District
18 Tamil Nadu India  e mail  jacobraglend rediffmail com
19 3Dr  D P Kothari is currently Vice Chancellor VIT University Vellore 632014  INDIA e mail  dpk0710 yahoo com
20 Abstract  An algorithm to solve security constrained unit
21 commitment problem UCP with both operational and power
22 flow constraints PFC have been proposed to plan a secure
23 and economical hourly generation schedule  This proposed
24 algorithm introduces an efficient unit commitment UC
25 approach with PFC that obtains the minimum system
26 operating cost satisfying both unit and network constraints
27 when contingencies are included  In the proposed model
28 repeated optimal power flow for the satisfactory unit
29 combinations for every line removal under given study period
```

Figura 6: Arquivo texto formatado para padrão ARFF do WEKA.

A Figura 6 apresenta o arquivo texto formatado em ARFF para ser processada na ferramenta WEKA. O formato ARFF tem os seguintes atributos que devem ser configurados: @RELATION: referente ao título do arquivo; @ATTRIBUTE: lista de termos e seu tipo de valor (INTEGER, REAL e NUMERIC para números, STRING para letras, entre outros); @ATTRIBUTE class: onde devem ser colocados os valores dos atributos entre chaves com as classificações dos textos; @DATA: lista de valores separados por vírgula e no final de cada linha deve constar a categorização referente aos valores. Para esta situação, o atributo "@ATTRIBUTE class:" não será utilizado neste projeto e o "@DATA" não será utilizada devido ao projeto se tratar de texto não estruturado.

Para iniciar o processamento dos textos é importante conhecer como funciona a ferramenta WEKA, iniciando pela tela inicial conforme mostra a Figura 7.

Figura 7: Tela inicial da ferramenta WEKA.

A Figura 7 apresenta a tela inicial da ferramenta de processamento de textos WEKA. O ambiente possui quatro opções de aplicações: 1) Explorer: que contem as funções de pré-processamento, análise, classificação e pré-visualização de resultados. 2) Experimenter: contém ferramentas de testes estatísticos entre esquemas de aprendizagem. 3) KnowledgeFlow: é semelhante ao Explorer mas com a vantagem de suportar a aprendizagem incremental. 4) Simple CLI: onde é possível incluir execuções diretamente na linha de comando pelo Sistema Operacional.

Para entender como é organizada a tela da aplicação Explorer pode-se observar a Figura 8.

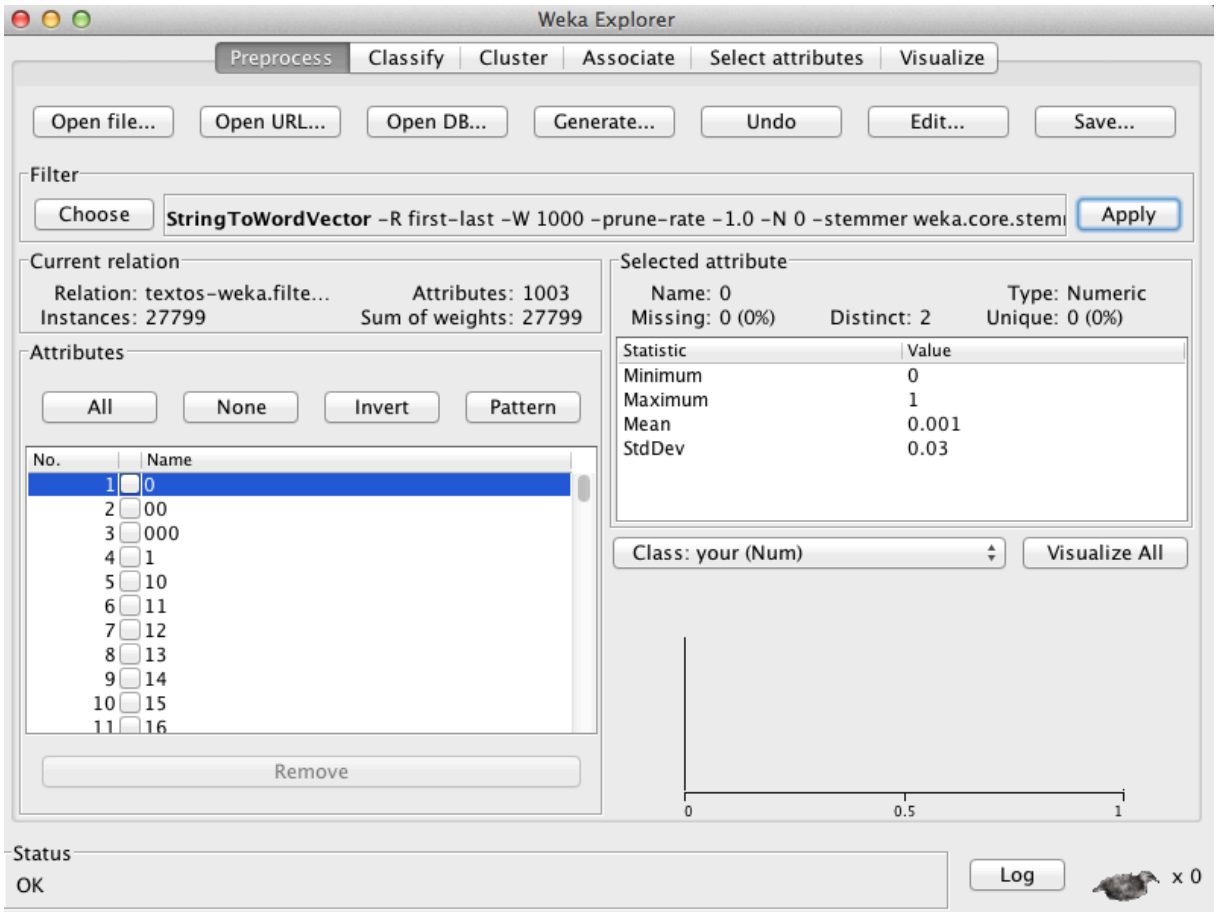


Figura 8: Tela Explorer da ferramenta WEKA.

A Figura 8 apresenta a tela Explorer da ferramenta WEKA onde no menu “Preprocess” é possível fazer o upload do arquivo ARFF cujo layout fora demonstrado na Figura 6. Para fazer o upload basta clicar em “Open File...”,

selecionar o arquivo e selecionar ok. Após, ainda na etapa de pré-processamento, foi selecionado o método de filtro por atributo não supervisionado “StringToWordVector”. Este filtro se faz necessário para que seja possível converter atributos do tipo string em um conjunto de atributos que representam uma determinada ocorrência da palavra a partir do texto contido na string. O conjunto de palavras (atributos) geradas após a execução do filtro pode ser visto na Figura 9.

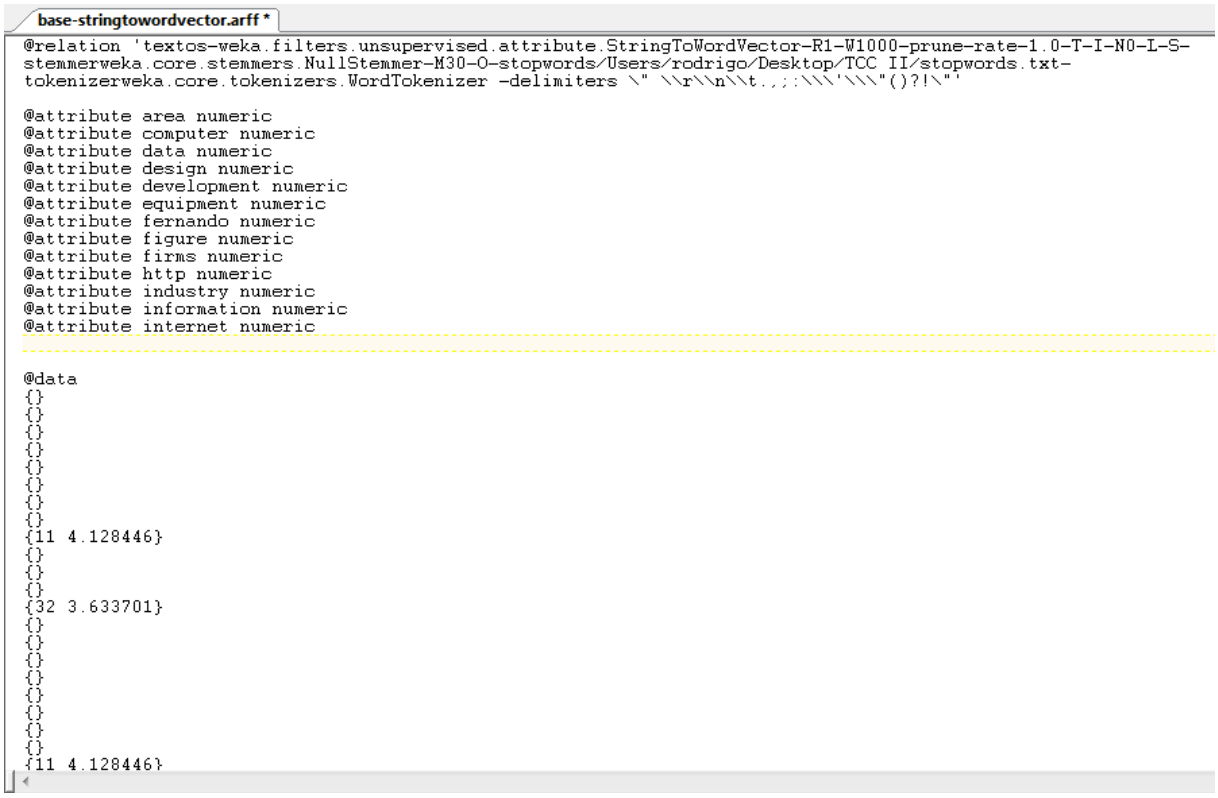


Figura 9: Resultado obtido pelo filtro “StringToWordVector”.

Durante o processo de transformação utilizando o filtro “StringToWordVector”, um outro elemento deve ser preparado para que com a utilização deste filtro se obtenha um resultado mais efetivo, o uso de stopwords. A definição da lista de stopword é providencial para que uma boa limpeza seja feita coerentemente e corretamente. Podemos verificar que antes da limpeza das stopwords o WEKA gerou 1003 atributos extraídos do arquivo importado (Figura 8). Após a execução do filtro com o auxílio das stopwords, os atributos diminuíram consideravelmente para 39 registros (Figura 11), tornando a busca por informação mais adequada e direcionada ao objetivo proposto.

Na Figura 10 é apresentada uma amostra de parte do arquivo de stopword utilizada no processo. O arquivo utilizado possui mais de 900 registros dentre elas letras, números, palavras e caracteres adicionados à lista de stopwords.

t	am	cannot	has
u	an	could	hasn't
v	and	couldn't	have
w	any	did	haven't
x	are	didn't	having
y	aren't	do	he
z	as	does	he'd
{	at	doesn't	he'll
|	be	doing	he's
}	because	don't	her
~	been	down	here
de	before	during	here's
about	being	each	hers
above	below	few	herself
after	between	for	him
again	both	from	himself
against	but	further	his
all	by	had	how
	can't	hadn't	

Figura 10: Lista parcial de stopwords.

Porém, ainda na etapa de pré-processamento, outros ajustes foram efetuados, removendo atributos considerados não pertinentes ao escopo do projeto, como por exemplo, nomes próprios como “Fernando, Silva, números, etc.…“ conforme demonstra a Figura 11.

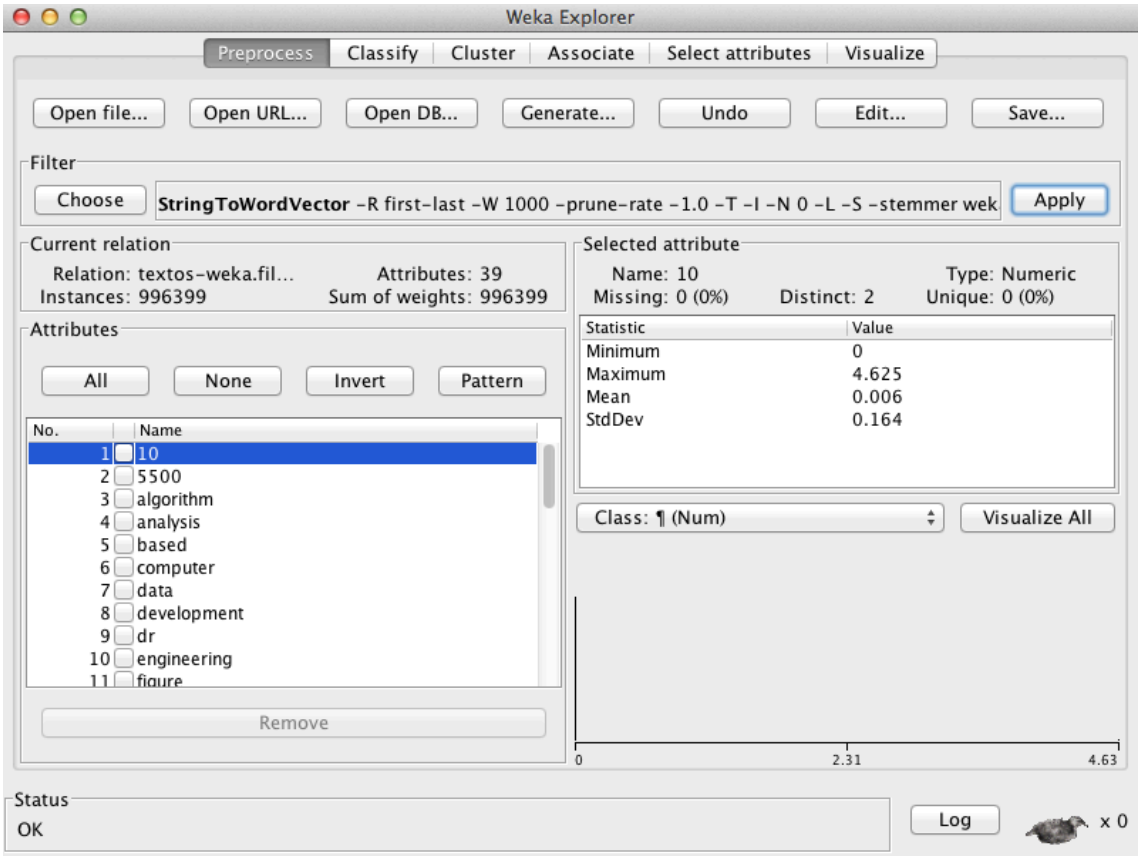


Figura 11: Limpeza dos atributos não pertinentes ao projeto.

É importante determinar nesta fase quais os atributos chaves que serão utilizados para os processos de clusterização e associação. Dessa forma, alguns outros atributos foram removidos até ficarem apenas os 10 atributos mais importantes. Atributos estes, selecionados baseando-se na proposta da busca do conhecimento para a elaboração, aprimoramento de um diferencial estratégico para os negócios na web. Assim, finalizando o processo de limpeza e organização do arquivo.

Esta limitação de atributos tem o intuito de propor uma melhor identificação dos mesmos tornando-os mais relevantes para o projeto, direcionando e fortalecendo o resultado tornando mais eficaz. Na medida em que menos atributos são selecionados os resultados gerados posteriormente na etapa de Associação serão menores e mais precisos, diminuindo assim, a ambiguidade e conflitos entre as regras geradas pela Associação e consequentemente da informação. Abaixo na Figura 12 consta a área de parametrização das variáveis do filtro “StringToWordVector”.

Figura 12: Parâmetros do filtro “StringToWordVector”.

O algoritmo foi configurado conforme os seguintes parâmetros:

IDFTransform / TFTransform – Métricas utilizadas para ponderar os termos. Os termos que aparecem pouco e não ajudam a distinguir uma palavra recebem um peso menor na classificação, os termos que aparecem bastante recebem um peso maior. Ambos foram parametrizados e definidos como verdadeiro.

lowerCaseTokens – Se definido, então toda a palavra é convertida para minúsculo antes de ser adicionado ao dicionário. Parâmetro foi definido como verdadeiro.

minTermFreq - Define a frequência mínima que uma palavra deve aparecer na base de dados consultada. Foi definida uma quantidade mínima de 1000 ocorrências do termo. Este valor foi determinado levando em consideração que a base de dados foi composta por 105 documentos diversos e entendendo que no mínimo, o termo teve que aparecer pelo menos 10 vezes dentro de cada documento.

stopwords – Palavras e termos q não tem validade para distinguir uma mensagem da outra. Foi montada uma lista com base no dicionário da língua inglesa contendo mais de 900 palavras, pronomes, etc...

useStoplist - Ignora todas as palavras que estão na lista de palavras irrelevantes, se definido como verdadeiro. Foi parametrizado para verdadeiro para que a lista de stopwords pudesse ser utilizada.

O resultado obtido após a execução deste método pode ser visto na Figura 13.

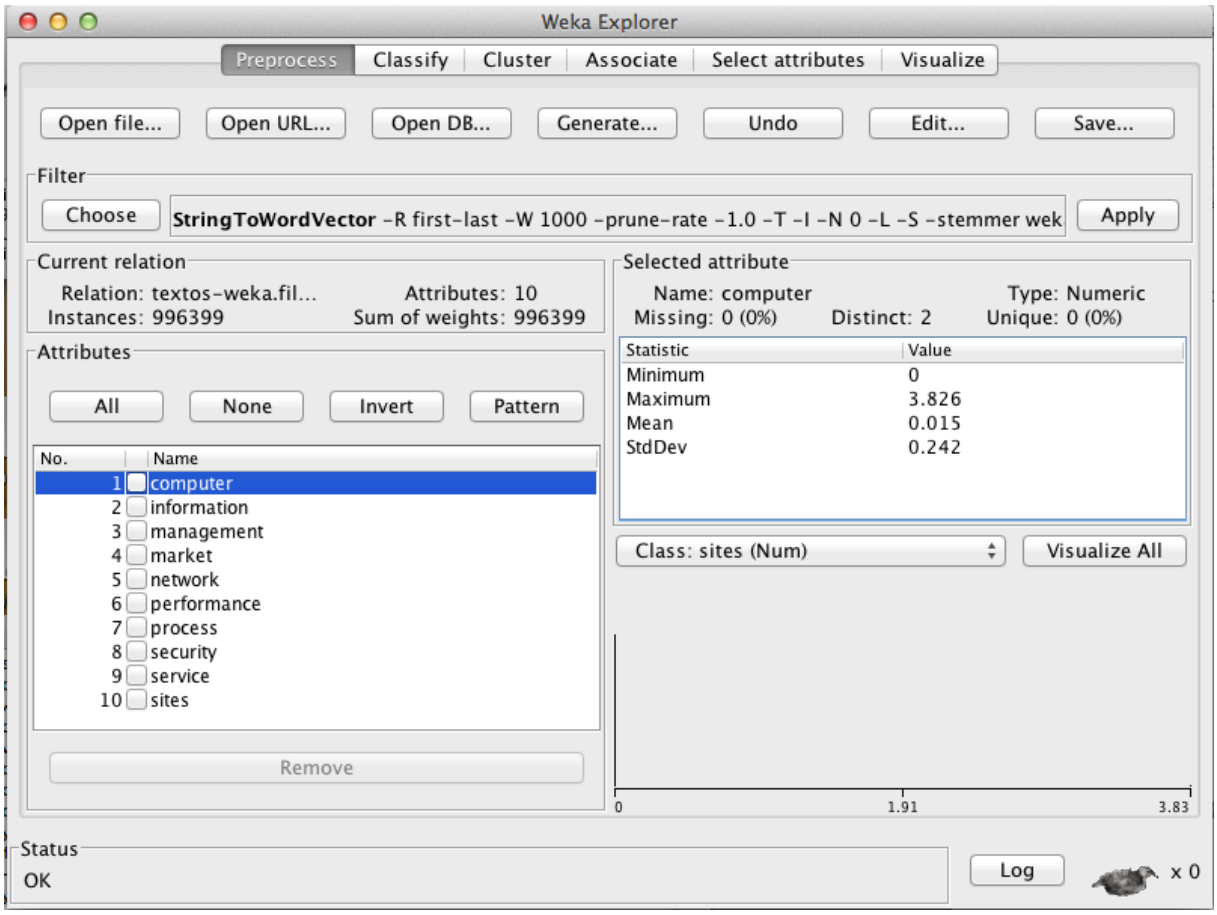


Figura 13: Execução do filtro “StringToWordVector”.

4.3.2 Transformação dos dados

Após o processo de limpeza dos dados é necessário realizar a codificação dos dados, com o intuito de que estejam na forma correta para serem usados como entrada dos algoritmos de mineração de dados, para posteriormente enriquecê-los de forma a agregar de alguma forma, mais informação para o processo de extração de conhecimento.

4.4 EXTRAÇÃO DE PADRÕES COM AGRUPAMENTOS DE TEXTOS

Para o processo de extração de padrões, após a execução da tarefa de agrupamento que tem por objetivo organizar um conjunto de objetos em grupos, baseado em uma medida de proximidade, na qual objetos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos objetos de outros grupos. Um dos algoritmos de agrupamento mais conhecidos, e que serve de base para inúmeros outros, é o algoritmo K-Means (ELSMARI, 2011). Este algoritmo iterativo usa como entrada um valor *K* correspondente ao número de grupos que deve ser formado; uma métrica de cálculo de distância entre dois registros, algumas condições de parada das iterações e os dados em si. O algoritmo cria *K* centróides com valores inicialmente randômicos, e itera primeiro marcando cada registro como pertencente ao centróide mais próximo e depois recalculando os centróides dos grupos de acordo com os registros pertencentes (ou mais próximos) a estes. A Figura 14 mostra seis passos da execução do algoritmo K-Means com $K = 3$ em um conjunto artificial de dados com dois atributos numéricos com valores entre 0 e 1, onde existem três grupos concentrados de pontos com uma quantidade considerável de ruído (pontos fora dos três grupos concentrados).

Os seis passos da execução do algoritmo K-Means mostrados na Figura 14 correspondem, respectivamente, à condição inicial (onde nenhuma iteração foi realizada, portanto os dados não são considerados pertencentes à nenhum dos grupos). A partir da primeira iteração os dados são marcados com tons de cinza diferentes, para facilitar a identificação dos grupos formados. Pode-se observar que os centróides dos grupos (indicados por pequenas cruzes) mudam sua posição, tentando se aproximar dos centros dos três grupos existentes. Nas primeiras iterações podemos observar claramente as mudanças das posições dos centróides, mas em iterações posteriores os mesmos quase não se movimentam, indicando que o algoritmo está convergindo.

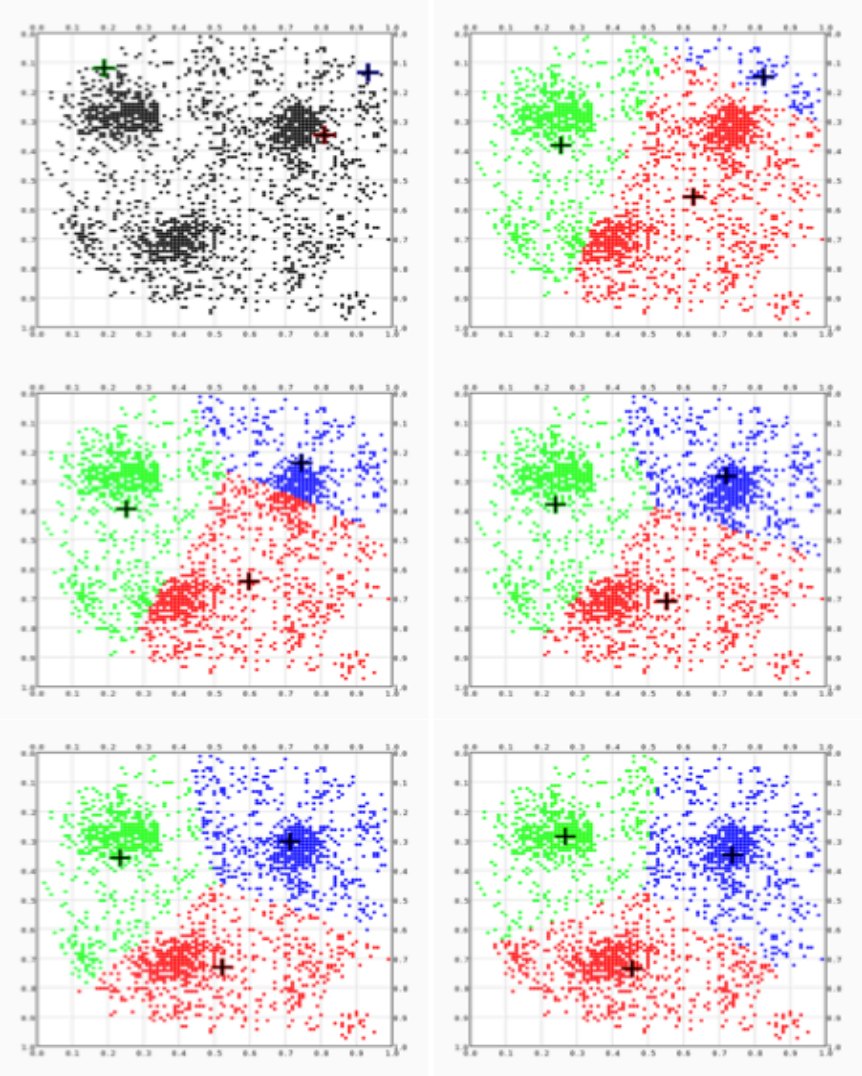


Figura 14: Etapas do algoritmo K-Means.

Depois de finalizado a etapa de pré-processamento, ocorre então o agrupamento dos atributos. Com o WEKA ainda aberto após o pré-processamento, foi selecionado o menu Cluster e após o algoritmo SimpleKMeans. Na Figura 15 são apresentadas as regras do algoritmo explicitado.

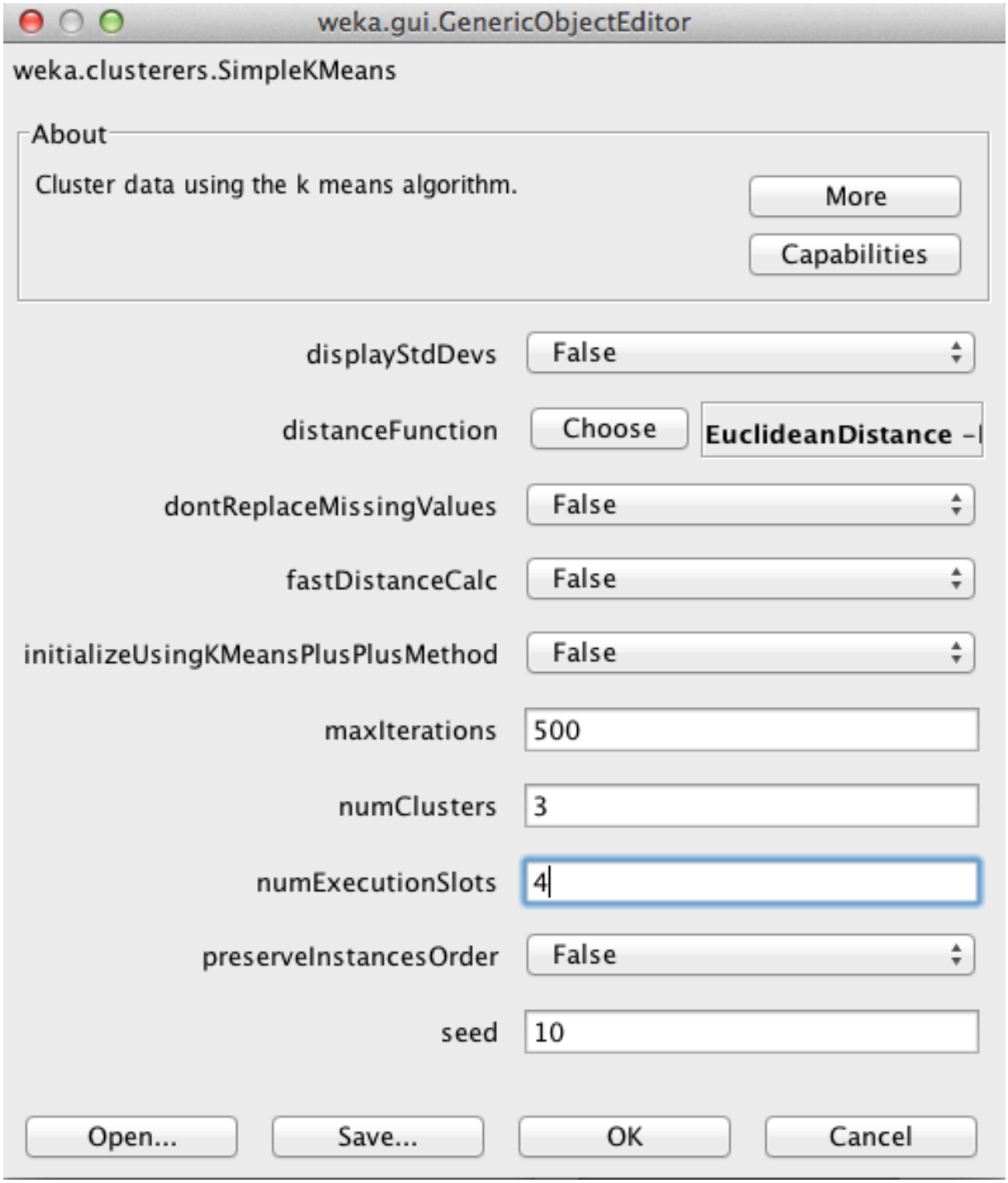


Figura 15: Regras do algoritmo SimpleKMeans.

Na tela de configuração do algoritmo foram parametrizadas as seguintes regras:

numClusters – Quantidade de conjunto de Clusters a serem gerados. Foram informados 3 conjuntos de clusters a ser gerado.

numExecutionSlots - Número de slots de execução (threads). Conjunto igual ao número disponível de CPU / núcleos.

Na Figura 16 abaixo é apresentado o resultado da execução do algoritmo de clusterização.

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 18127.07166360921
Missing values globally replaced with mean/mode

Cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                (996399)   (993780)     (1368)     (1251)
=============================================================
computer          0.0153     0.0154          0          0
information        0.011      0.011          0          0
management        0.0063          0     4.5684          0
market            0.0052     0.0052          0          0
network           0.0088     0.0088          0          0
performance       0.0056     0.0056          0          0
process           0.0058     0.0058          0          0
security          0.0149     0.0149          0          0
service           0.0058          0          0     4.6304
sites              0.009     0.0091          0          0



Time taken to build model (full training data) : 8.99 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      993780 (100%)
1        1368 (  0%)
2        1251 (  0%)
```

Figura 16: Execução do processo de clusterização.

Percebem-se na Figura 16 alguns pontos de interesse: Os centróides e desvios-padrão de cada *cluster* são mostrados nas linhas de 10 a 25. As instâncias aparecem subdivididas em 3 grandes clusters. Ao total foram geradas 996.399 instâncias, onde quase que sua totalidade está referenciada no cluster 0, onde o mesmo faz referência as instâncias computer, information, market, network, performance, process, security e sites. Nos clusters 1 e 2 respectivamente estão separados as instâncias management e service. Mostrando, dessa forma, que a área de gestão e serviços deve ser tratada separadamente e especificamente.

Na Figura 17 estão representados os clusters gerados pelo processo.

Figura 17: Visão dos Clusters gerados.

Após esta etapa, a execução foi salva em formato ARFF conforme Figura 18.

```
@relation 'textos-weka.filters.unsupervised.attribute.
@attribute Instance_number numeric
@attribute computer numeric
@attribute information numeric
@attribute management numeric
@attribute market numeric
@attribute network numeric
@attribute performance numeric
@attribute process numeric
@attribute security numeric
@attribute service numeric
@attribute sites numeric
@attribute Cluster {cluster0,cluster1,cluster2}

@data
0,0,0,0,0,0,0,0,0,0,0,cluster0
1,0,0,0,0,0,0,0,0,0,0,cluster0
2,0,0,0,0,0,0,0,0,0,0,cluster0
3,0,0,0,0,0,0,0,0,0,0,cluster0
4,0,0,0,0,0,0,0,0,0,0,cluster0
5,0,0,0,0,0,0,0,0,0,0,cluster0
6,0,0,0,0,0,0,0,0,0,0,cluster0
7,0,0,0,0,0,0,0,0,0,0,cluster0
8,0,0,0,0,0,0,0,0,0,0,cluster0
9,0,0,0,0,0,0,0,0,0,0,cluster0
10,0,0,0,0,0,0,0,0,0,0,cluster0
11,0,0,0,0,0,0,0,0,0,0,cluster0
12,0,0,0,0,0,0,0,3.850997,0,0,cluster0
13,0,0,0,0,0,0,0,0,0,0,cluster0
14,0,0,0,0,0,0,0,0,0,0,cluster0
15,0,0,0,0,0,0,0,0,0,0,cluster0
16,0,0,0,0,0,0,0,0,0,0,cluster0
17,0,0,0,0,0,0,0,0,0,0,cluster0
18,0,0,0,0,0,0,0,0,0,0,cluster0
19,0,0,0,0,0,0,0,0,0,0,cluster0
```

Figura 18: Arquivo gerado pelo processo de clusterização.

O arquivo (conforme mostra a Figura 18) é carregado novamente em modo de pré-processamento. Neste momento é selecionado o método de filtro por atributo não supervisionado “NumericToNominal” não sendo necessário alterar nenhum parâmetro deste filtro. Procedimento este necessário, pois o método de classificação não consegue efetuar a leitura de números, sendo preciso normalizar as informações do arquivo conforme demonstram as Figuras 19 e 20.

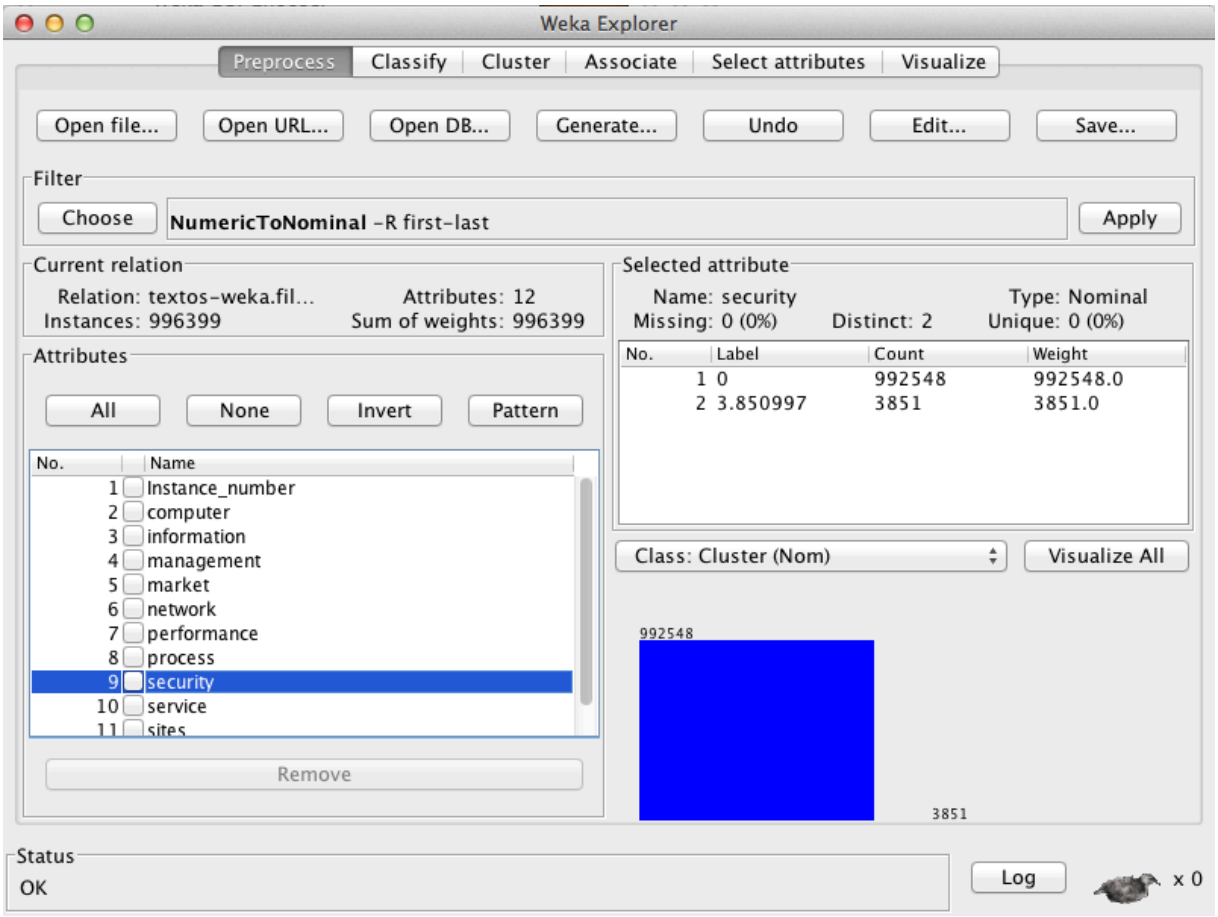


Figura 19: Processo de Normalização.

Na Figura 20 é apresentado o resultado da execução do algoritmo de normalização.

```
@relation 'textos-weka.filters.unsupervised.attribute.S

@attribute Instance_number {0,1,2,3,4,5,6,7,8,9,10,11,1
@attribute computer {0,3.825552}
@attribute information {0,4.107467}
@attribute management {0,4.568393}
@attribute market {0,4.717639}
@attribute network {0,4.293453}
@attribute performance {0,4.666765}
@attribute process {0,4.637046}
@attribute security {0,3.850997}
@attribute service {0,4.630365}
@attribute sites {0,4.268355}
@attribute Cluster {cluster0,cluster1,cluster2}

@data
0,0,0,0,0,0,0,0,0,0,0,0,cluster0
1,0,0,0,0,0,0,0,0,0,0,0,cluster0
2,0,0,0,0,0,0,0,0,0,0,0,cluster0
3,0,0,0,0,0,0,0,0,0,0,0,cluster0
4,0,0,0,0,0,0,0,0,0,0,0,cluster0
5,0,0,0,0,0,0,0,0,0,0,0,cluster0
6,0,0,0,0,0,0,0,0,0,0,0,cluster0
7,0,0,0,0,0,0,0,0,0,0,0,cluster0
8,0,0,0,0,0,0,0,0,0,0,0,cluster0
9,0,0,0,0,0,0,0,0,0,0,0,cluster0
10,0,0,0,0,0,0,0,0,0,0,0,cluster0
11,0,0,0,0,0,0,0,0,0,0,0,cluster0
12,0,0,0,0,0,0,0,3.850997,0,0,cluster0
13,0,0,0,0,0,0,0,0,0,0,0,cluster0
14,0,0,0,0,0,0,0,0,0,0,0,cluster0
15,0,0,0,0,0,0,0,0,0,0,0,cluster0
16,0,0,0,0,0,0,0,0,0,0,0,cluster0
```

Figura 20: Arquivo normalizado gerado.

Para a tarefa de Descoberta do Conhecimento, foi escolhido o modelo de descoberta por Associação. Este modelo "*tem por objetivo encontrar relacionamentos ou padrões frequentes que ocorrem em um conjunto de dados*" (BRUSSO, CERVO & GEYER, 2002); ou seja, é a busca por itens que ocorram de

forma simultânea frequentemente em transações do banco de dados. Por exemplo, “85% de todos os registros que contém os itens A, B, e C também contém D e E”. A porcentagem de ocorrência (85 no caso) representa a afinidade contida nestes itens. A associação trata-se de uma função endereçada à análise de mercado, onde o objetivo é encontrar tendências de associação dentro de um grande número de registros. Essas tendências de associação podem ajudar a entender e explorar padrões de compras, sugerir novas tendências de produtos e serviços e dar um direcionamento a estratégia de negócio da organização.

O algoritmo mais usado na implementação de regras de associação é o algoritmo Apriori. Este, procura identificar relações entre os itens de um conjunto de dados, que são descritas em forma de regras do tipo “Se X então Y”, ou “$X \rightarrow Y$”, onde X e Y são conjuntos de itens e $X \cap Y = \varnothing$ (AGRAWAL et al., 1994). Para a realização deste estudo, foi utilizado o algoritmo Apriori.

> O algoritmo Apriori usa uma abordagem de níveis para gerar regras de associação, onde cada nível corresponde ao número de itens que pertencem ao consequente da regra. Inicialmente, todas as regras de confiança alta que tenham apenas um item no consequente da regra são extraídas. Estas regras são então usadas para gerar novas regras candidatas. Por exemplo, se {acd} -> {b} e {abd} -> {c} forem regras de confiança alta, então a regra candidata {ad} -> {bc} é gerada pela fusão dos consequentes de ambas as regras. Suponha que a confiança para {bcd} -> {a} seja baixa. Todas as regras contendo item a no seu consequente, incluindo {cd} -> {ab}, {bd} -> {ac}, {bc} -> {ad} e {d} -> {abc} podem ser descartadas (TAN, 2009, p.417).

Para esta fase, o arquivo do WEKA já passou pelas etapas de pré-processamento, clusterização e normalização dos termos. A partir disso, selecionou-se o menu “Associate” e feito a busca do filtro, o algoritmo Apriori. Na Figura 21 são demonstrados os parâmetros do algoritmo.

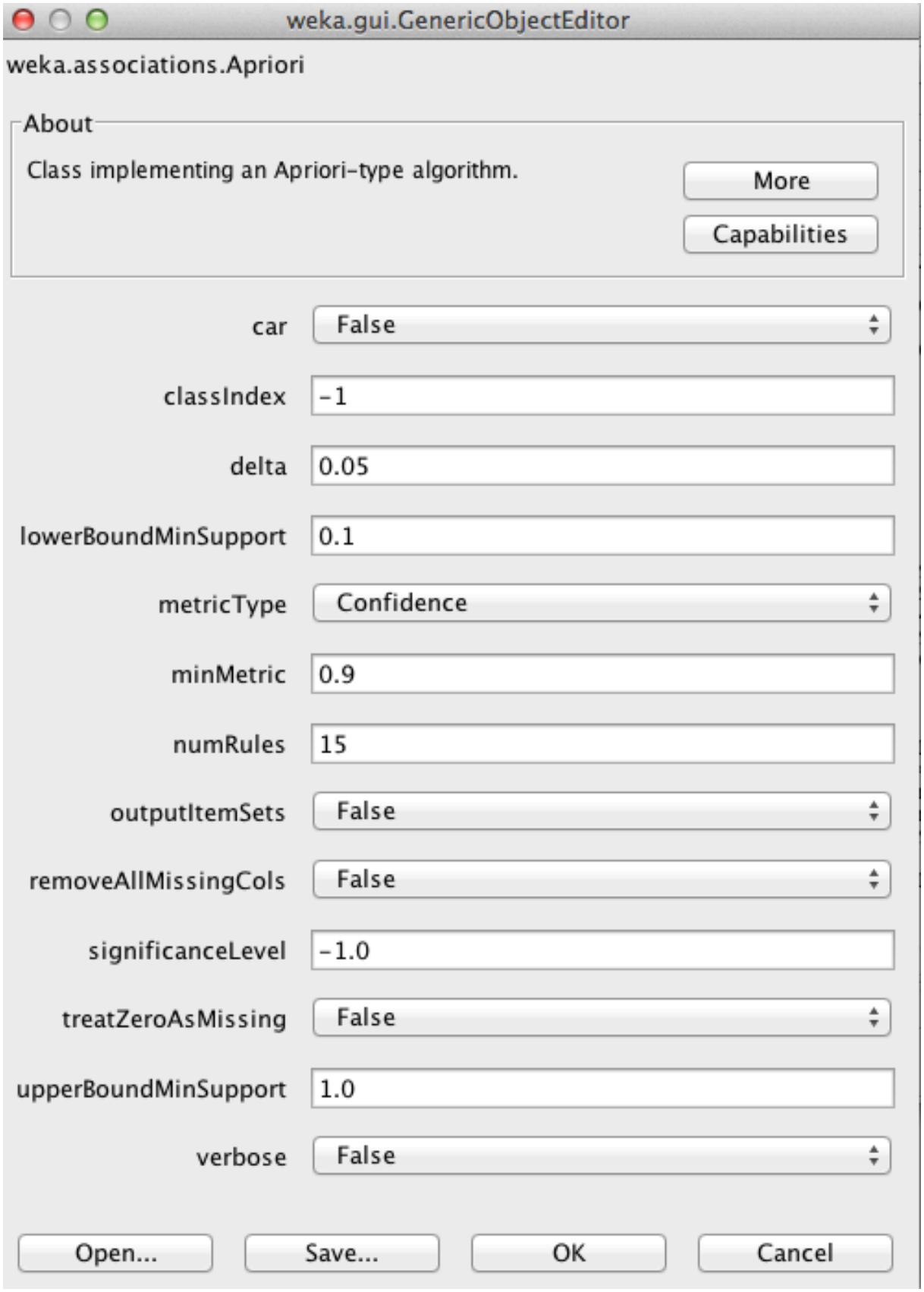


Figura 21: Parâmetros algoritmo Apriori.

Na tela de configuração do algoritmo foi alterado apenas o parâmetro numRules. Para este foi definido o valor 15, onde representa o numero de regras a encontrar através da relação de associação entre os atributos selecionados.

Percebe-se na Figura 22 que as 15 melhores regras de associação são mostradas. Os valores depois dos antecedentes e consequentes das regras são o número de instâncias ou ocorrências para as quais a associação ocorre. É interessante observar que o algoritmo usa tanto atributos positivos (“1”) como negativos (“0”), encontrando muitas regras significativas. O atributo referenciado está ao lado do valor da instancia conforme exemplo: management**=0** 995031 ==> market**=0** 993928. Neste exemplo a associação é: Se não tem “management” então não tem “market”.

```
Apriori
=======

Minimum support: 0.95 (946579 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 45

Size of set of large itemsets L(3): 120

Size of set of large itemsets L(4): 210

Size of set of large itemsets L(5): 252

Size of set of large itemsets L(6): 210

Size of set of large itemsets L(7): 120

Size of set of large itemsets L(8): 45

Size of set of large itemsets L(9): 10

Size of set of large itemsets L(10): 1

Best rules found:

 1. performance=0 995212 ==> market=0 994109    <conf:(1)> lift:(1) lev:(0) [-1] conv:(1)
 2. process=0 995160 ==> market=0 994057    <conf:(1)> lift:(1) lev:(0) [-1] conv:(1)
 3. service=0 995148 ==> market=0 994045    <conf:(1)> lift:(1) lev:(0) [-1] conv:(1)
 4. management=0 995031 ==> market=0 993928    <conf:(1)> lift:(1) lev:(0) [-1] conv:(1)
 5. network=0 994365 ==> market=0 993262    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
 6. sites=0 994290 ==> market=0 993187    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
 7. performance=0 process=0 993973 ==> market=0 992870    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
 8. performance=0 service=0 993961 ==> market=0 992858    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
 9. process=0 service=0 993909 ==> market=0 992806    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
10. management=0 performance=0 993844 ==> market=0 992741    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
11. management=0 process=0 993792 ==> market=0 992689    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
12. management=0 service=0 993780 ==> market=0 992677    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
13. information=0 993739 ==> market=0 992636    <conf:(1)> lift:(1) lev:(0) [-2] conv:(1)
14. network=0 performance=0 993178 ==> market=0 992075    <conf:(1)> lift:(1) lev:(0) [-3] conv:(1)
15. network=0 process=0 993126 ==> market=0 992023    <conf:(1)> lift:(1) lev:(0) [-3] conv:(1)
```

Figura 22: Resultado obtido pelo algoritmo Apriori.

Podemos tomar como exemplo da Tabela 3 a regra 4 onde se não tenho uma boa gestão nos negócios, não terei uma boa atuação no mercado. Complementando esta linha de pensamento com as regras 10 a 12 onde, se não tenho uma boa gestão nos processos de negócio e um serviço eficiente e de qualidade não atenderei satisfatoriamente as necessidades de mercado necessárias.

Regra	Instância antecedente	Então	Instância consequente
4	management=0 995031	==>	market=0 993928
10	management=0 performance=0 993844	==>	market=0 992741
11	management=0 process=0 993792	==>	market=0 992689
12	management=0 service=0 993780	==>	market=0 992677

Tabela 3: Regras extraídas do Apriori.

4.5 AVALIAÇÃO DO CONHECIMENTO

A etapa de Avaliação consiste na descoberta e validação dos resultados obtidos na fase anterior. Em geral, o principal objetivo dessa etapa é melhorar a compreensão do conhecimento descoberto pelo algoritmo minerador, validando-o através de medidas de qualidade. Depois desta tarefa é possível regressar aos passos anteriores que envolvem a repetição de processos, talvez com outros dados, outros algoritmos ou outras estratégias. São usadas técnicas de representação e visualização de conhecimento para apresentar o conhecimento de forma compreensível ao usuário. Esses conhecimentos serão consolidados em forma de gráficos, diagramas e outros tipos de relatórios demonstrativos.

Podemos perceber na Tabela 4 que todos os 10 atributos utilizados tem uma forte relação com a área mercadológica. Analisando as regras podemos tirar as seguintes conclusões:

1) Para atingir aos objetivos do negócio e ser uma empresa diferenciada no mercado de trabalho é necessário que os processos e serviços prestados pela empresa estejam alinhados e de acordo com a estratégia do negócio.

2) A gestão e a qualidade da informação que a empresa possui são imprescindíveis para se aventurar ao mercado.
3) Uma boa comunicação e o uso computacional fortalecem o crescimento dos negócios.
4) A utilização de recursos na internet através de websites agrega na busca da diferenciação mercadológica.

Regra	Instância antecedente	Então	Instância consequente
1	performance=0 995212	==>	market=0 994109
2	process=0 995160	==>	market=0 994057
3	service=0 995148	==>	market=0 994045
4	management=0 995031	==>	market=0 993928
5	network=0 994365	==>	market=0 993262
6	sites=0 994290	==>	market=0 993187
7	performance=0 process=0 993973	==>	market=0 992870
8	performance=0 service=0 993961	==>	market=0 992858
9	process=0 service=0 993909	==>	market=0 992806
10	management=0 performance=0 993844	==>	market=0 992741
11	management=0 process=0 993792	==>	market=0 992689
12	management=0 service=0 993780	==>	market=0 992677
13	information=0 993739	==>	market=0 992636
14	network=0 performance=0 993178	==>	market=0 992075
15	network=0 process=0 993126	==>	market=0 992023

Tabela 4: Relacionamento das regras extraídas do Apriori.

Baseado nos resultados obtidos da Tabela 4, podemos notar que os atributos resultantes do processo de extração e mineração de textos tem grande relevância, proporcionando a redução da incerteza no processo de tomada de decisão e o aumento da qualidade da decisão. As regras geradas indicam que para a obtenção do sucesso no mercado de trabalho os processos operacionais devem se tornar mais eficientes, e os processos gerenciais da empresa mais eficazes. Com essas melhorias nos processos empresarias a empresa pode reduzir custos, melhorar a qualidade e o atendimento ao cliente e criar novos produtos e serviços.

Com as melhorias oferecidas pela Tecnologia de Informação, as empresas podem ter novas oportunidades comerciais, permitindo a expansão para novos mercados ou novos segmentos de mercados existentes. Ainda que isso signifique

enfrentar muitas barreiras, principalmente no que tange ao custo elevado de investimento e complexidade da tecnologia de informação.

Os dados utilizados para o processo de extração proporcionaram a obtenção de informações (resultado das regras geradas pelo algoritmo Apriori) que ajudam a empresa a definir pontos de referencia estratégico, direcionar quais os pontos críticos que devem ser tratados internos e externamente, necessários para atingir ao objetivo do negócio. Os textos extraídos da web proporcionaram através dos resultados uma visão clara das necessidades e da constante mudança que as empresas de tecnologia têm de se adequar e melhorar para atender as expectativas do mercado consumidor.

O estudo foi direcionado para a obtenção do entendimento da capacidade de informação que um grupo de dados pode proporcionar, cabendo à alteração (por parte do especialista da empresa) das instancias inicialmente selecionados no processo de extração, a definição de que tipo de informação se deseja obter. A escolha e definição das instancias durante o processo de extração e limpeza dos dados é determinante para que novas regras e novos resultados possam ser gerados de acordo com a necessidade da empresa, sejam de novos produtos, serviço, etc...Estes resultados foram obtidos através da análise de um especialista, onde o mesmo, através das informações obtidas com o gestor da empresa, pode alinhar as regras extraídas com os objetivos estratégicos direcionados pelo gestor. Dentro do campo analisado (empresa de tecnologia atuando na prestação de serviços), o especialista identificou similaridades entre as regras.

Dentro do escopo do trabalho nas relações entre as regras 1, 7 e 8 percebe-se que o especialista entende que, para uma empresa atuar com prestação de serviços é indispensável “performance”, “process” e “service”, ou seja, os processos da empresa para que ocorra a efetivação do serviço devem ser enxutos, claros e efetuados com rapidez. O especialista identificou também que, nas regras 4, 10, 11 e 12 (“management”, “performance”, “process” e “service”) as relações entre as instancias puderam indicar que a gestão dos serviços, dos processos (tanto internos como externos da empresa) e a velocidade com que tudo isto ocorre influência diretamente na qualidade dos serviços prestados pela empresa. Com esta informação podemos identificar mais subregras subentendidas, como a questão da qualidade dos serviços, visto que, uma melhor gestão das áreas nos processos

dentro da empresa pode melhorar significativamente a qualidade dos serviços prestados aos clientes e fornecedores. Proporcionando uma melhor transparência no atendimento e na relação para com os mesmos.

CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos uma visão diferenciada de como o processo de extração de dados da web pode proporcionar uma melhor seleção de dados para serem utilizados na mineração dos textos, proporcionando um diferencial estratégico e competitivo para as empresas. Neste trabalho são apresentados os principais conceitos, os resultados que se deseja obter com o processo do KDD, bem como as técnicas e modelos utilizados para o processo.

As Tecnologias e os Sistemas de Informação são fundamentais na obtenção e extração de informações estratégicas de bases de dados que auxiliem na tomada de decisão, já que o papel que a informação desempenha na atualidade e dentro das organizações é de extrema importância e seus benefícios são evidentes.

O trabalho buscou a descoberta do conhecimento através da utilização de regras de associação. O algoritmo utilizado tem a vantagem de descobrir regras no formato **se - então**, fáceis de serem interpretadas pelo especialista encarregado de obter as informações. Porém algoritmos de regras de associação têm a desvantagem de exigirem um alto tempo de processamento, muito comum nos casos de *text mining*. Para reduzir esse problema, utilizou-se das técnicas básicas de pré-processamento de textos (remoção de *stop words* e *stemming*). O método de pré-processamento proposto mostrou-se eficaz na redução do tempo computacional e ainda assim permitir a descoberta de regras com alto fator de confiança, porém durante o processo de levantamento das informações um grande problema encontrado foi à definição da base de dados para o trabalho. A grande restrição de acesso a estas informações restringiu bastante o campo de obtenção dos dados. Os documentos utilizados da base de dados da Brown Corpus (Internet Archive Text) foram relevantes para o trabalho, mas insuficientes para se conseguir uma visão

mais realista dos negócios no mundo competitivo. Muitos destes documentos possuíam maior teor científico e menos informações relacionadas ao mercado dos negócios.

As dificuldades encontradas neste trabalho estiveram relacionadas mais propriamente com o conteúdo dos documentos e não com o processo de DCBD. Para isto, uma maior atenção foi dada aos processos de extração, limpeza e preparação dos dados que foram utilizadas para a descoberta do conhecimento.

Assim, algumas direções para pesquisa futura são sugeridas a seguir. Inicialmente, cabe ressaltar que é interessante a busca de novas bases de dados não estruturadas na medida em que está utilizada, não reflete as informações do mercado de trabalho em sua totalidade, dando um direcionamento mais voltado ao ambiente científico e acadêmico.

No futuro poder-se-á automatizar o processo de extração e limpeza dos dados permitindo uma maior agilidade na execução desta etapa proporcionando a realização de experimentos mais extensos com uma maior base de documentos extraídos. Além disso, outras medidas de interesse de regras poderão ser investigadas no futuro, talvez se utilizando de outras informações disponíveis na Internet.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R.; 1994. **Fast Algorithms for Mining Association Rules**. In PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES (1994: Santiago, Chile).

ARANHA C., PASSOS E. **Revista Eletrônica de Sistemas de Informação** (Nº2, 2006) - ISSN 1677-3071 doi:10.5329/RESI.

BOENTE, A. N. P. ; OLIVEIRA, F. S. G. ; ROSA, J. L. A. **Utilização de Ferramenta de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa**. Anais do Simpósio de Excelência em Gestão e Tecnologia, v. 1, p. 123-132, 2007.

BRUSSO, M. J.; CERVO, L. V.; GEYER, C. F. R. **Ferramenta para Descoberta de Regras de Associação em Banco de Dados Relacionadas a área da Saúde**.

Porto Alegre, 2002, 12 páginas. Monografia (Graduação). Universidade Federal do Rio Grande do Sul.

CAMPOS, R. N.T., 2005. **Agrupamento Automático de Páginas Web Utilizando Técnicas de Web Content Mining**. Dissertação de M.Sc, Universidade da Beira Interior, Covilhã, 2005.

CHEN, H. **Knowledge management systems**: a text mining perspective. University of Arizona (Knowledge Computing Corporation), Tucson, Arizona, 2001.

CHOUDHURY, Vivek & SAMPLER, Jeffrey L. **Information specificity and environmental scanning**: an economic perspective. MIS Quarterly, Março de 1997.

ELMASRI, Ramez.; NAVATHE, Shamkant**. Conceitos de Data Mining. In:** Sistemas de Banco de Dados. São Paulo: Pearson Addison Wesley, 2005, p. 624-645**.**

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados.** 6.ed. São Paulo: Pearson, 2011. xviii, 788 p. ISBN 9788579360855.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). **Advances in knowledge discovery and data mining**. [S.l.]: AAAI/MIT Press, 1996.

FELDMAN, Ronen; SANGER, James. **The text mining handbook**: advanced approaches in analyzing unstructured data. New York: Cambridge University Press, 2007.

FREITAS (H.), JANISSEK (R.) e MOSCAROLA (J.). **Dinâmica do processo de coleta e análise de dados via web**. CIBRAPEQ - Congresso Internacional de Pesquisa Qualitativa, 24 a 27 de março, Taubaté/SP, 2004. 12 p..

GOLDSCHIMIDT, R e PASSOS, E. **Data mining: Um guia prático**. Rio de Janeiro: Campus, 2005.

INMON, W. H.; HACKATHORN, R. D. **Como usar o Data Warehouse**. Rio de Janeiro: Infobook, 1997.

IADIS, Conferência Ibero-Americana WWW/Internet 2005. **Mineração da utilização de páginas web, identificando padrões através das requisições dos usuários.** ISBN: 972-8924-03-8 © 2005 IADIS.

JESUS, A. P. de; MOSER, E. M.; OGLIARI, P. J. **Data mining aplicado na identificação do perfil dos usuários de bibliotecas para a personalização de sistema web de recuperação e disseminação de informações**. http://inf.unisul.br/ ines/workcomp/cd/pdfs/2371.pdf, 2011.

JUNIOR C., RIBEIRO J. - (2007). **Desenvolvimento de uma metodologia para mineração de textos** / João Ribeiro Carrilho Junior ; orientador: Emmanuel Piseces Lopes Passos. – 2007. 96 f. ; 30 cm. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de

Janeiro, 2007.

LEVITT, T. “**Marketing myopia**”. *Harvard Business Review,* jul./ago. 1906, p.45-56. _. “Marketing myopia – Retrospective commentary”. Harvard Business Review, 53 (5), 1975, p. 177-181. _. The marketing imagination. Nova York: The Free Press, 1986.

LAUDON, K. C.; LAUDON, J. P. **The Internet: electronic commerce and electronic business**. In: _____; _____. *Management information systems*: organization and technology in the networked enterprise. 6. ed. Upper Saddle River, NJ: Prentice Hall, 2000. cap. 10, p. 290-329.

MICHAEL J. A. Berry; GORDON Linoff, “**Data Mining Techniques for Marketing, Sales, and Customer Support**”; John Wiley & Sons, Inc., 1997.

M. S. CONRADO, R. M. MARCACINI, M. F. MOURA, and S. O. REZENDE, “**O efeito do uso de diferentes formas de geração de termos na compreensibilidade e representatividade dos termos em coleções textuais na língua portuguesa**” in WTI’09: II International Workshop on Web and Text Intelligence, 2009, pp. 1–10.

NAVEGA, Sérgio. **Princípios essenciais do Data Mining**. São Paulo, SP: 2002. Publicada nos Anais do Infoimagem.

N. F. F. EBECKEN, M. C. S. LOPES, and M. C. de ARAGÃO COSTA, “Mineração de textos,” in **Sistemas Inteligentes: Fundamentos e Aplicações**, 1st ed., S. O. Rezende, Ed. Manole, 2003, ch. 13, pp. 337–370.

Oberle, Daniel; Berendt, Bettina, Hotho, Andreas et al, 2003. **International Atlantic Web Intelligence Conference**. *Conceptual User Tracking, Web Intelligence*. Madrid, Spain, pp. 155-164.

PORTER, M.F. **An algorithm for suffix stripping. In *Readings in Information Retrieval***. Morgan Kaufmann, 1997.

R. FELDMAN and J. SANGER, ***The Text Mining Handbook****: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

REZENDE, S. O., MARCACINI, R. M., MOURA, M. F. / **Revista de Sistemas de Informação da FSMA** n. 7 (2011) pp. 7-21.

SPARCK JONES, Karen; WILLET, Peter (Ed.). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997.

WIVES, Leandro Krug. **Tecnologias de descoberta de conhecimento em textos aplicadas a inteligência competitiva**. 2000. 100p. Exame de qualificação

(Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

XAVIER, P. M. **Mineração de Dados: Tecnologias e Ferramentas** – Trabalho de Conclusão do Curso de Informática da UCPel – 2000.

TAN, PANG-NING; STEINBACH, MICHAEL; KUMAR, VIPIN. **Introdução ao *Data Mining*: Mineração de Dados**. Rio de Janeiro: Ciência Moderna, 2009.

THOMÉ, Antônio C. G. Data Warehouse, Data Mining. In: **Redes Neurais – Uma ferramenta para KDD e Data Mining**. [s.i.]: 20--.