FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA – UNIVEM
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

LETICIA TONON

DATA MINING COMO FERRAMENTA PARA A RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS INSTITUCIONAIS

MARÍLIA
2014

LETICIA TONON

DATA MINING COMO FERRAMENTA PARA A RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS INSTITUCIONAIS

Trabalho de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília – UNIVEM, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador
Profº: Dr. Elvis Fusco

MARÍLIA
2014

TONON, Leticia

Data Mining como ferramenta para a Recuperação da Informação em Repositórios Digitais Institucionais / Leticia Tonon; orientador: Profº. Dr. Elvis Fusco. Marília, SP: [s.n.], 2014.

50 folhas

Monografia (Bacharelado em Sistemas de Informação): Centro Universitário Eurípides de Marília.

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Leticia Tonon

DATA MINING COMO FERRAMENTA PARA A RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS INSTITUCIONAIS

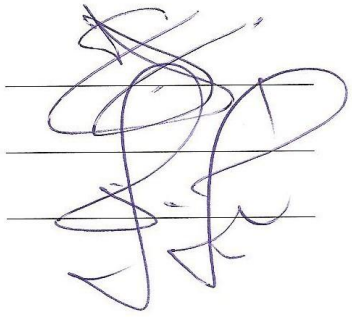
Banca examinadora da monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Sistemas de Informação.

Nota: 7,0 (Sete)

Orientador: Elvis Fusco

1°. Examinador: Paulo Augusto Nardi

2°. Examinador: Geraldo Pereira Junior



Marília, 01 de dezembro de 2014.

Dedico este trabalho aos meus pais, Antonio Tonon e Dalva Tonon, a meu irmão Lucas Tonon e a meu namorado José Miguel Moreno Calvo, pelo apoio e incentivo durante esta jornada da minha vida.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela sabedoria para superar os obstáculos que surgiram.

Aos meus pais, Antonio Tonon e Dalva Tonon, pelo apoio e incentivo nas horas em que precisava de orientação.

A meu irmão Lucas Tonon, pelos momentos descontraídos que foram essenciais nas horas de desânimo.

A meu namorado José Miguel Moreno Calvo, pela paciência e parceria com que me acompanhou durante esta etapa da minha vida.

A meu orientador Elvis Fusco, pela orientação e suporte. Agradeço também aos meus amigos de sala, nunca me esquecerei de vocês.

A Instituição, aos meus professores e a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

“Do or do not. There is no try.” Yoda

RESUMO

No contexto atual há um grande volume de dados armazenados nas bases informacionais dos repositórios digitais com isso a adversidade em encontrar informação útil em sistemas de recuperação da informação se intensificou, fazendo com que fossem exigidos processos de recuperação cada vez mais sofisticados. Este estudo pretende fazer uso de técnicas de mineração de dados para melhorar a recuperação de informação em repositórios institucionais digitais.

Palavras-Chave: Recuperação da Informação, Repositórios Digitais, Data Mining.

ABSTRACT

Currently there is a large volume of data stored in informational bases of digital repositories therewith the problem of finding useful data in information retrieval systems has intensified, making the processes of recovery increasingly sophisticated. This study aims to make use of data mining techniques to improve information retrieval in digital institutional repositories.

Keywords: Information Retrieval, Digital Repositories, Data Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RI	Recuperação da Informação
RD	Repositórios Digitais
DM	Data Mining
SRI	Sistemas de Recuperação da Informação
DC	Dublin Core
BD	Banco de Dados
XML	eXtensible Markup Language
ARFF	Attribute-Relation File Format
JDBC	Java Database Connectivity
SGBD	Sistema de Gerenciamento de Banco de Dados
CSV	Comma-separated values
JAR	Java ARchive
KDD	Knowledge Descovery in Databases
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Atualmente os repositórios digitais são utilizados cada vez mais como forma comum de busca de informações tanto para fins pessoais quanto acadêmicos buscando facilitar as pesquisas por sua enorme facilidade e agilidade na recuperação de informações.

Com o crescimento do volume de publicações e também das necessidades de informações dos usuários, sejam elas em papel ou em formato eletrônico, é importante que as bibliotecas possuam sistemas de informações capazes de armazenar e indexar informações bibliográficas de forma a facilitar a recuperação e disseminação aos usuários (Cardoso, 2000).

Dito isso, nota-se uma crescente necessidade por repositórios mais robustos, que facilitem a forma de armazenamento e recuperação destas informações que se perdem em meio a exponencial quantidade de material que é criado a cada dia.

Os repositórios digitais se caracterizam como ambientes facilitadores de acesso às informações, sem limitação de espaço e tempo, uma vez que nesses o tratamento dado ao recurso informacional requer uma descrição de forma e de conteúdo legível por máquinas com resultados compreensíveis aos humanos. Desse modo, destaca-se a necessidade de um tratamento de forma e conteúdo adequado para a representação e para a apresentação de informações, visando uma recuperação mais eficiente (Castro; Santos, 2008).

A busca por informações tem aumentado consideravelmente em ambientes acadêmicos brasileiros, especialmente de nível superior. Grande parte dos alunos tem acesso direto à rede Internet, ocasionando uma constante troca de informações e de conhecimento (Santarem, 2010).

O acesso aos grandes sistemas de recuperação de informação, também denominados de bancos de dados e, consequentemente, às suas bases de dados veio ampliar significativamente a qualidade das buscas bibliográficas, visto que essas bases proporcionam diversificados pontos de acesso à informação (Lopes, 2002).

O objetivo da recuperação da informação é armazenar e recuperar documentos existentes em grandes bases de dados conforme o conteúdo buscado pelo usuário, feito isso, os resultados são classificados e listados de acordo com sua relevância na busca.

Um dos grandes desafios encontrados na recuperação de informação é como atender às necessidades de informação do usuário de forma rápida e precisa. Várias pesquisas foram e

continuam sendo realizadas com o propósito de aumentar a precisão dos resultados de forma que o usuário possa encontrar todos os documentos que atendem às suas necessidades de informação (Kuramoto, 2002).

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, a recuperação da informação apresenta constantemente novos desafios e se configura como uma área de extrema relevância (Cardoso, 2000).

Desta forma, pensando em um sistema de recuperação que traga informações úteis ao usuário, destaca-se um processo do KDD (*Knowledge Discovery in Databases*): o Data Mining.

O Data Mining é um processo na descoberta do conhecimento em bancos de dados que consiste em analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não são visíveis. Para isto são utilizadas técnicas que envolvem métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades entre os dados pesquisados (Brusso, 1998).

A quantidade de informações geradas atualmente torna imprescindível a utilização de técnicas de Data Mining visando cruzar as informações buscadas de modo a trazer resultados mais relevantes.

1.1 Motivação e Justificativa

Tão importante quanto o armazenamento e manipulação das informações em sistemas, são os processos de recuperação da informação, que fazem com que usuários tenham acesso aos dados compilados para a tomada de decisão e geração de novos conhecimentos.

Neste sentido, a mineração de dados e seus algoritmos de inteligência artificial colaboram no processo de recuperação da informação.

1.2 Objetivos Gerais

Por meio do embasamento teórico foi possível definir o objetivo desta pesquisa que é estender o processo de Recuperação da Informação em Ambientes Informacionais Digitais de Repositórios Institucionais Digitais por meio de técnicas de Data Mining visando ampliar a relevância e a eficácia dos resultados de busca da informação. Além disso, a pesquisa tem como objetivos específicos:

- Apresentar técnicas de Data Mining que possam ser usadas como ferramentas em Sistema de Recuperação da Informação de Repositórios Institucionais Digitais;
- Criar um protótipo que implemente uma camada de Recuperação da Informação utilizando técnicas de Data Mining em Repositórios Digitais Institucionais que utilizam o DSPACE como ferramenta.

1.3 Metodologia

O projeto foi dividido nas seguintes fases:

- Levantamento de fontes bibliográficas para a definição do embasamento teórico do tema;
- Pesquisa de trabalhos correlatos sobre Recuperação da Informação, Data Mining e Repositórios Digitais;
- Estudo das técnicas de Data Mining e sua aplicação no conceito de Recuperação da Informação;
- Estudo sobre a utilização da WEKA;
- Implementar uma camada de Recuperação da Informação utilizando a API WEKA para a aplicação das técnicas de Data Mining em Repositórios Digitais que utilizam o DSPACE.

1.4 Trabalhos Correlatos

O projeto “Data Mining Aplicado à Identificação do Perfil dos Usuários de uma Biblioteca para a Personalização de Sistemas Web de Recuperação e Disseminação de Informações” utiliza as técnicas de Data Mining como ferramenta para a recuperação da informação personalizada em um sistema web desenvolvido para a Biblioteca Central da FURB (Jesus, 2014).

O projeto “O processo de bibliomineração: repositório de dados e mineração de dados para tomada de decisão em bibliotecas” tem como objetivo apresentar uma metodologia que utilize as técnicas de Data Mining em repositórios digitais a fim de compreender o comportamento de comunidades de usuários (Nicholson, 2004).

1.5 Organização do Trabalho

Esta monografia está organizada da seguinte forma:

No primeiro capítulo aborda a visão geral do trabalho, a motivação e justificativa, seus objetivos gerais, metodologia, trabalhos correlatos e a organização do trabalho.

O segundo capítulo apresenta a fundamentação teórica sobre Recuperação da Informação com seus modelos clássicos.

O terceiro capítulo traz a fundamentação teórica sobre Data Mining com suas técnicas.

O quarto capítulo expõe as etapas do desenvolvimento de um protótipo.

Por fim são apresentadas as conclusões e referências.

2 REPOSITÓRIOS DIGITAIS

Repositórios digitais são conjuntos de obras em vários formatos diferentes, mas digitais, disponíveis para serem acessadas através de meios computacionais, ou seja, a partir de redes de computadores locais ou pela Internet. O conteúdo dos repositórios digitais não possui limitações, podendo existir repositórios institucionais que agrupam projetos de pesquisas, teses de pós-graduação, trabalhos diversos desenvolvidos pelos membros da instituição, monografias, entre outros, e também repositórios para qualquer tipo de arquivos em formato digital que se deseje (Baeza-Yates; Ribeiro-Neto, 1999).

Tratando-se de repositórios institucionais de universidades, Lynch (2013) os define como sendo um conjunto de serviços que uma instituição de ensino oferece aos membros de sua comunidade objetivando o gerenciamento e disseminação eficiente dos materiais digitais gerados pela instituição e pelos membros de sua comunidade.

Compreende-se então que os repositórios institucionais de universidades têm como objetivo armazenar e propagar teses, documentos desenvolvidos pelos membros da instituição, monografias, periódicos científicos, entre outros. Além do repositório institucional há outro tipo de repositório, o repositório temático.

De acordo com Kuramoto (2006) um repositório temático é um conjunto de serviços oferecidos por uma sociedade, associação ou organização, para gestão e disseminação da produção técnico-científica em meio digital, de uma área ou subárea específica do conhecimento.

O repositório institucional é a reunião de repositórios temáticos, sob a responsabilidade técnica e administrativa de uma instituição ou organismo. Por consequência, este tipo de repositório é multidisciplinar e possui uma gama de tipos de documentos, em alguns casos, maior que um repositório temático. Além de agregar o conjunto de informações relativas e/ou de interesse para a instituição, dispõem de serviços referentes à organização, disseminação e acesso ao conteúdo digital (Café, 2003).

Em outras palavras o repositório temático, como o próprio nome diz, é um repositório que abrange somente temas específicos. Enquanto um repositório institucional é o agrupamento de vários repositórios temáticos. Ou seja, em um repositório institucional pode-se encontrar uma quantidade maior de documentos e temas a serem pesquisados.

Para a implantação de um repositório institucional, que é o tipo de repositório que será abordado no presente trabalho, há várias ferramentas disponíveis, algumas são softwares livres e outras softwares proprietários. Como o Dspace, o Fedora, entre outros.

Tais ferramentas apresentam características semelhantes quanto à forma com que armazenam seus dados. Todas elas estão amparadas por uma estrutura que define um banco de dados, relacional em grande parte das vezes, para armazenar as informações que são postadas pelos mais variados tipos de usuários (Santarem, 2010). O software que implantou o repositório a ser utilizado neste trabalho é o Dspace.

2.1 Dspace

O Dspace é um software livre muito utilizado para o desenvolvimento de repositórios institucionais. Foi desenvolvido pelo Massachusetts Institute of Technology (MIT) e Laboratórios Hewlett-Packard.

Sua estrutura provê um modelo de informação organizacional baseado em "comunidades" e coleções, o qual pode ser configurado de modo a refletir todo o conjunto de unidades administrativas de uma instituição. Permite a configuração do processo editorial nos moldes dos periódicos tradicionais, incluindo a possibilidade de revisão pelos pares. Suporta os mais variados tipos de formatos de arquivos digitais, incluindo textos, som e imagem (Lewis, 2013).

O DSpace é formado por diversos componentes, distribuídos por três camadas distintas (Romani; Fusco, 2010):

- A *Storage Layer* que é responsável pelo armazenamento físico dos metadados e dos conteúdos;
- A *Business Logic Layer* que trata da gestão dos conteúdos do arquivo, dos utilizadores, das políticas de autorização e do *workflow* e;
- A *Application Layer* que contém os componentes que comunicam com o mundo exterior, como por exemplo a interface *web* do utilizador e o serviço de suporte ao protocolo de coleta de metadados da OAI.

A camada *Storage Layer* é responsável pelos metadados, que são a descrição do documento que está armazenado no repositório, sendo eles de grande importância para esta

pesquisa e serão mais bem abordados na próxima seção.

O esquema de metadados utilizado pelo repositório é o Dublin Core, isto significa que a descrição do metadado segue os padrões desse esquema.

2.2 Metadados e o Padrão Dublin Core

Os metadados são "dados sobre dados", o que quer dizer que metadados referem-se à estrutura descritiva da informação sobre outro dado, o qual é usado para ajudar na identificação, descrição, localização e gerenciamento de recursos da *web*. Entretanto, eles podem ser aplicados em qualquer meio (Dziekaniak; Kirinus, 2004).

Na maioria dos sistemas de informação os metadados estão presentes, devido a isso há a necessidade da padronização desses metadados para que o compartilhamento seja possível, mesmo que entre diferentes instituições.

Com a utilização de padrões, o armazenamento, o acesso e a exibição aos documentos de um determinado repositório se tornam mais fácil. Pensando nisso, surgiram iniciativas voltadas a melhorar a descrição dos documentos na *Web*.

O Dublin Core é um dos padrões disponíveis para a descrição de documentos eletrônicos, e será o padrão utilizado no presente trabalho.

Dublin Core é uma iniciativa voltada à organização das informações na *web* de forma padronizada, ou seja, seu objetivo é determinar padrões de classificação e catalogação das informações. Esse esquema contêm 15 elementos que são apresentados na Tabela 1.

Tabela 1. Elementos do Dublin Core

Elementos	Descrição
Title (Título)	Nome dado a um documento
Creator (Autor)	Responsável pela elaboração do documento
Subject (Assunto)	Tema do documento
Description (Descrição)	Texto breve sobre o conteúdo do documento
Publisher (Editor)	Responsável por disponibilizar o documento
Contributor (Colaborador)	Faz contribuições no conteúdo do documento
Date (Data)	Data associada a alguma modificação no ciclo de vida do documento

Type (Tipo)	Natureza do conteúdo do documento
Format (Formato)	Tipo de mídia ou dimensão de um documento
Identifier (Identificador)	Faz referência a um documento
Source (Fonte)	Faz referência ao presente documento derivado de outro documento
Language (Língua)	Língua utilizada no documento
Relation (Relação)	Referência para um documento relacionado
Coverage (Cobertura)	A abrangência do conteúdo do documento
Rights (Direitos Autorais)	Informações sobre os direitos autorais do documento

Para coletar os metadados de um repositório que utiliza o padrão Dublin Core como esquema de metadados pode ser utilizado o protocolo OAI-PMH.

2.2.1 Protocolo OAI-PMH

O protocolo OAI-PMH é um mecanismo de coleta de metadados em repositórios. Esse protocolo é formado por seis verbos de requisição, cada verbo trás um tipo específico de resposta em XML e podem ser visualizados na Tabela 2.

Tabela 2. Verbos do Protocolo OAI-PMH

Verbo	Descrição
Identify	Utilizado para coletar informações que descrevem o repositório
ListMetadataFormats	Utilizado para listar os padrões de metadados do repositório
ListRecords	Utilizado para coletar os metadados do repositório
ListIdentifiers	Utilizado para coletar apenas o cabeçalho dos objetos do repositório
GetRecord	Utilizado para coletar os metadados de um único objeto do repositório
ListSets	Utilizado para listar os conjuntos do repositório

3 RECUPERAÇÃO DA INFORMAÇÃO

O termo recuperação da informação foi concebido por Mooers (1951), definindo que a área é responsável pelos aspectos intelectuais da descrição de informação e especificações para a busca, incluindo sistemas, técnicas ou máquinas empregadas para este fim e pode ser definida, no contexto da Ciência da Informação, como uma operação na qual são selecionados documentos a partir da busca de um usuário.

Alguns autores conceituam a Recuperação da Informação de forma menos complexa, enquanto outros de forma muito mais complexa. Quando conceituada de forma complexa a mesma fica também responsável pela catalogação, indexação e classificação das informações.

Há diferentes definições porém de forma geral todas semelhantes, para Baeza-Yates e Ribeiro-Neto (1999) a recuperação da informação é responsável pela representação, armazenamento, organização e acesso a itens informativos. Para isso é necessária à aplicação conjunta de técnicas de Processamento da Linguagem Natural e Inteligência Artificial (Scholtes, 1993).

O processamento da Linguagem Natural tem como objetivo simular o processamento da Linguagem Humana, ou seja, o intuito é “entender” a busca para que a recuperação da informação ocorra de forma mais eficiente.

As técnicas do processamento podem ser classificadas por meio do nível da interpretação linguística, que são (Liddy, 1998):

- Fonológico: comum em sistemas de recuperação da informação que lidam com a língua falada, ou seja, onde a busca pode ser realizada por meio da voz.
- Morfológico: está relacionado com a análise da estrutura e da classificação das palavras.
- Lexical: lida com o significado das palavras, duas palavras podem ter o mesmo sentido (partir e ir: mesmo significado) assim como uma palavra pode ter mais de um sentido (Manga: fruta ou manga de camisa).
- Sintático: determina a função de cada palavra em um conjunto de palavras assim podendo reconhecer se uma sequencia de palavras constitui uma frase ou não.
- Semântico: visa interpretar uma palavra ou conjunto de palavras. Uma frase

pode estar perfeitamente correta sintaticamente e não ser de fácil interpretação. O conhecimento do significado da palavra auxilia na formação de uma frase, que se aplicada conjuntamente com o nível sintático resulta em frases bem elaboradas.

- Discursivo: é analise da estrutura do documento, ou seja, como o mesmo é divido e apresentado (Exemplo: Introdução, Desenvolvimento, Conclusão).
- Pragmático: utiliza conhecimentos que não fazem parte do sistema, ou seja, analisa a frase conforme seu contexto levando em consideração a necessidade do usuário, suas preferências e seu objetivo quando formulou uma determinada frase para a realização da busca.

Assim como em outros tipos de técnicas, não há como eleger uma como a melhor técnica. Para cada situação irá existir uma técnica que melhor se encaixa.

Os estudos em recuperação da informação podem envolver vários aspectos, como comportamental, necessidades informações dos usuários, algoritmos de recuperação, diversidade sociolinguística, entre outros.

Atualmente grande parte das informações armazenadas pelo homem se encontra em formato digital. Uma das preocupações dos pesquisadores da área da Recuperação da Informação está em como organizar e representar as inserções de novos documentos neste acervo, bem como facilitar a busca aos documentos, em uma posterior recuperação (Miorelle, 2001).

Uma das ferramentas mais importante para auxiliar esse processo é denominada índice, que se trata de uma coleção de termos que indicam o local onde a informação desejada pode ser localizada (Cardoso, 2000). O índice permite que um termo seja encontrado mais rapidamente quando efetuada uma consulta no sistema de recuperação da informação, ou seja, seu objetivo é facilitar a busca de informações em uma base de dados.

A maioria dos sistemas de recuperação da informação dos repositórios digitais institucionais permite que o usuário expresse quais são as informações que ele precisa e a partir disso o sistema retorna os documentos considerados relevantes. Esses sistemas podem ser estruturados conforme a Figura 1.

Figura 1. Componentes de um Sistema de Recuperação da Informação

Primeiramente há a necessidade do usuário e os documentos que serão consultados, nos documentos é feito o processo de indexação, esse processo é uma representação na forma de índices dos documentos. Do outro lado a necessidade do usuário gera o processo de especificação de consulta, que é o momento que o usuário elabora a pergunta a ser consultada para o sistema. Esses componentes e processos se encontram no processo de recuperação, onde é feita a consulta nos documentos que foram indexados. O resultado do processo é uma lista que contêm os documentos recuperados que tentam chegar o mais próximo do que o usuário procura.

De forma geral todo sistema de informação é um sistema que recupera informações. Um dos tipos de instituição que adotaram tais sistemas foram as bibliotecas, num primeiro momento esses sistemas apenas automatizavam o processo de busca. Com o aumento das produções científicas muita informação tem sido gerada e isso intensifica a necessidade do desenvolvimento de técnicas capazes de melhor atender ás necessidades dos usuários.

Pensando nisso destaca-se o processo de Data Mining, que segundo Diniz e Neto (2000) é “o processo de extração de informações, sem o conhecimento prévio, de um grande banco de dados, e seu uso para tomada de decisões”.

O objetivo da personalização de conteúdo é garantir que a pessoa certa receba a informação certa no momento certo (Aranha, 2000). Com os métodos de Data Mining é

possível gerar perfis específicos para cada grupo de usuários, para desta forma tornar possível a personalização dos processos de recuperação da informação.

3.1 Modelos Clássicos

Nesta seção serão abordados os modelos clássicos de Recuperação da Informação. Antes de conhecer os modelos é necessário compreender o significado do termo de indexação, que de modo simplificado trata-se um conjunto de palavras-chave, cada documento possui um conjunto, extraídos manualmente ou automaticamente.

Os termos de indexação podem ser extraídos diretamente do texto dos documentos ou podem ser especificados por um humano, como é frequentemente feito pelos bibliotecários e pelos cientistas da informação. Independente de como os termos foram gerados eles são responsáveis por fornecer uma visão lógica dos documentos (Baeza-Yates; Ribeiro-Neto, 2011).

Dito isso, primeiramente será apresentado o Modelo Booleano, onde os termos de indexação não tem peso na associação. Após isso será apresentado o Modelo Vetorial e em seguida o Probabilístico, onde os termos de indexação são essenciais no processo de ordenação dos documentos.

É importante ressaltar que existem vários outros modelos de recuperação da informação, o importante é sempre utilizar o que mais se adéqua a situação.

3.1.1 Modelo Booleano

O modelo Booleano é um modelo de recuperação simples baseado na teoria de conjuntos e na álgebra Booleana. Como consequência, o modelo é bastante intuitivo e possui uma semântica precisa. Pela sua inerente simplicidade e formalismo elegante, o modelo recebeu uma atenção considerável no passado e foi adotado por muitos dos primeiros sistemas bibliográficos comerciais (Baeza-Yates; Ribeiro-Neto, 2011).

Neste modelo uma consulta pode ser considerada como expressão booleana quando são compostas pelos conectivos lógicos *AND*, *OR* e *NOT*. Os pesos dos termos de índice

assumem valores binários, pelo fato dos mesmos estarem presentes ou não em um documento.

As vantagens do modelo são sua simplicidade e o formalismo claro implícito no modelo. As desvantagens de sua utilização é a falta de ordem nas respostas, além disso, às vezes as mesmas se apresentam nulas ou muito grandes.

3.1.2 Modelo Vetorial

O modelo vetorial também é chamado de modelo espaço vetorial e representa cada documento como um vetor de termos, e cada termo possui um valor associado que indica seu grau de importância (peso – *weight*) para o documento, ou seja, cada consulta possui um vetor resultado construído através do cálculo da similaridade baseado no ângulo (co-seno) entre o vetor que representa o documento e o vetor que representa a consulta (Baeza-Yates; Ribeiro-Neto, 2011).

Este modelo propõe a representação de consultas e documentos como vetores de termos, isto é realizado por meio da atribuição de pesos binários aos mesmos. Por meio de um cálculo de similaridade é gerado o vetor que contêm o resultado de uma consulta.

Os pesos quantificam a relevância de cada termo para as consultas (W_{iq}) e para os documentos (W_{id}) no espaço vetorial. Para o cálculo dos pesos W_{iq} e W_{id} é utilizada uma técnica que faz o balanceamento entre as características do documento utilizando a frequência de um termo no documento (Cardoso, 2000).

Portanto, com uma quantidade de documento N e uma quantidade de termos T pode-se definir o inverso da frequência do termo.

3.1.3 Modelo Probabilístico

Na recuperação de informação, a modelagem probabilística é utilizada para classificar documentos em ordem decrescente de probabilidade de relevância de acordo com uma solicitação do usuário (Crestani; Lalmas; Van-Rijsbergen, Campbell, 1998).

Os modelos probabilísticos trabalham com um conjunto Q de consultas e um conjunto D de documentos de uma coleção (Fuhr; Pfeifer, 1994). Que como dito

anteriormente podem ser representados pelos termos de indexação (palavras-chave).

Segundo Baeza e Ribeiro (2011), no modelo probabilístico o peso do termo de indexação de uma consulta pode ser representado por $w_{i,q}$, e o peso do termo de indexação para um documento pode ser representado por $w_{i,j}$, todos são binários, $w_{i,q} \in \{0,1\}$, $w_{i,j} \in \{0,1\}$. A consulta, como dito anteriormente, é representa por q. O acréscimo do $+R_q$ (q. $+R_q$) indica que o documento é relevante à consulta, já o acréscimo do $–R_q$ (q. -Rq) indica que o documento não é relevante para a consulta. q. $P(+R_q|d_j)$ é a probabilidade de que um documento d_j seja relevante para a consulta q, e $P(–R_q |dj)$ é a probabilidade de que um documento d_j seja não-relevante para a consulta q.

Uma de suas vantagens é sua maior precisão se comparado com os demais modelos clássicos, porém a precisão da estimativa probabilística afeta diretamente o desempenho do modelo.

4 MINERAÇÃO DE DADOS

A Mineração de dados, também conhecida como Data Mining, faz parte de um contexto maior chamado KDD, sendo assim é necessário compreender as etapas desse processo antes de se aprofundar no conceito de DM.

A descoberta de conhecimento em bancos de dados (KDD) fundamenta-se no fato de que as grandes bases de dados podem ser uma fonte de conhecimento útil, porém não explicitamente representados, e cujo objetivo é desenvolver e validar técnicas, metodologias e ferramentas capazes de extrair o conhecimento implícito nesses dados e representá-lo de forma acessível aos usuários (Feldens, 1996).

Muitas vezes os termos Data Mining e KDD são confundidos como sinônimos. Porém o termo KDD é utilizado para descrever todo o processo de extração de conhecimento em um conjunto de dados. Já o termo Data Mining refere-se a uma das etapas deste processo (Carvalho, 2002).

Segundo Fayyad (1996), o KDD é dividido em cinco etapas: seleção, limpeza ou pré-processamento, transformação, Data Mining, interpretação e avaliação. Conforme pode ser visto na Figura 2.

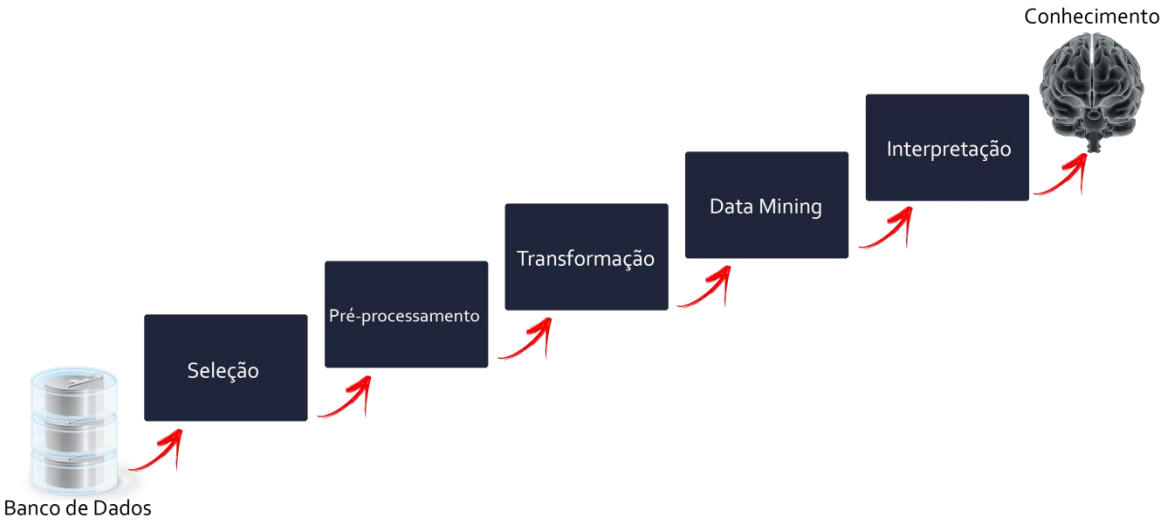


Figura 2. Etapas do processo de KDD.

O processo de descoberta se inicia a partir da seleção, nessa etapa os dados

relevantes de um BD são selecionados e agrupados. Depois que os dados foram agrupados vem a etapa de limpeza, onde dados irrelevantes são desconsiderados. Esses dados não são excluídos do banco de dados, são apenas desconsiderados no processo.

Em seguida vem a etapa de transformação que representa os dados de acordo com a tarefa de data mining escolhida previamente. Com o dado transformado é possível aplicar as técnicas de data mining, a etapa de DM que será abordada nesse capitulo.

E por fim a interpretação, nesta etapa os resultados obtidos a partir do processo de data mining são interpretados gerando o conhecimento. Após essa breve explicação sobre o processo de KDD será abordada a etapa mais importante para esta pesquisa, o Data Mining.

4.1 Data Mining

A curiosidade e a observação são características normais entre os seres humanos que estão sempre fazendo questionamentos e se arriscando para descobrir regras. Com a tecnologia o volume dados aumenta a cada dia e processo de encontrar padrões dentro de um conjunto de dados fica cada vez mais demorado. A mineração tem o intuito de suprir essa necessidade, podendo ser aplicado tanto em empresas como em projetos científicos, por esse motivo pode ser considerada como uma área multidisciplinar.

A definição mais importante de Data Mining tenha sido elaborada por Fayyad (1996): "o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Seu surgimento está ligado à associação de três áreas, a primeira delas e mais antiga é a estatística clássica. Tem papel fundamental nas ferramentas e técnicas que utilizam DM, em outras palavras, é a base para a construção das mesmas.

A segunda área é a Inteligência Artificial, com o objetivo de imitar as características humanas visando solucionar problemas estatísticos.

A terceira área é nomeada por *Machine Learning* podendo ser definida como uma intersecção entre a primeira e a segunda área, por fazer combinações entre heurística e estatística. Seu objetivo é automatizar o reconhecimento de padrões complexos para que possa chegar a uma conclusão inteligente baseada em dados.

Segundo Santarem Segundo (2010), DM é a busca de informações valiosas em grandes bases de dados. É um esforço de cooperação entre homens e computadores. Os

homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam os dados e mineram neles padrões que casem com as metas estabelecidas pelos homens.

Em outras palavras, a mineração consiste no uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu (Carvalho, 2002).

Antes de conhecer o processo de Data Mining é necessário saber a definição de dados, informação e conhecimento em nível de inter-relação, Figura 3.

Figura 3. Inter-relação entre Dados, Informação e Conhecimento.

Dados são elementos em sua forma bruta, ou seja, em sua maioria não são compreensíveis; Informação é o resultado do processo de tratamento dos dados, ou seja, aqueles dados não compreensíveis são transformados em algo que possa levar à compreensão; Conhecimento é o resultado da analise da informação, ou seja, é quando a informação passa a ser compreendida e utilizada para otimizar algum processo existente no sistema.

De modo simplificado, o processo de mineração se inicia a partir da limpeza de uma

fonte de dados onde, por exemplo, são retirados os ruídos e as redundâncias. Logo após, esses dados vão para os Data Marts e/ou Data Warehouses. Finalmente, usando os repositórios é possível selecionar colunas para entrarem no processo de Data Mining. A informação gerada desse processo (em forma de gráficos, planilhas, etc.) é analisada, e a partir dessa analise é possível transformar aquela informação em conhecimento.

Deve-se destacar que cada técnica de data mining ou cada implementação específica de algoritmos que são utilizados para conduzir as operações data mining adapta-se a alguns problemas que para outros, o que impossibilita a existência de um método de data mining universalmente melhor. Para cada particular problema tem-se um particular algoritmo. Portanto, o sucesso de uma tarefa de data mining está diretamente ligada à experiência e intuição do analista (Diniz, 2000).

4.2 Tarefas de Data Mining

As tarefas de DM podem ser divididas em dois grupos: descriptive data mining e predictive data mining. O primeiro se trata de uma mineração conduzida por objetivo, visando identificar padrões a partir de um determinado campo. O segundo grupo identifica padrões nos dados sem visar um determinado campo.

Dentro desses grupos, segundo Harrison (1998), existem seis tipos de tarefas. São elas: classificação, previsão, agrupamento por afinidade, estimação, segmentação e descrição. Nas próximas seções serão abordados os seis tipos.

4.2.1 Classificação

Segundo Berry e Linoff (1997) a classificação é a tarefa mais comum do DM. Ela consiste em examinar os aspectos de um objeto e ligá-lo a uma classe pré-definida.

Em outras palavras, essa tarefa analisa um conjunto de dados e fundamentado em suas características atribui o mesmo a uma classe. Abaixo estão listados alguns exemplos de classificação.

- Classificar tipos de doenças: classificar doenças fundamentadas nos sintomas

dos pacientes;

- Prever futuro devedor: classificar clientes fundamentados nos pagamentos anteriores;
- Prever futuro consumidor de um produto: classificar clientes fundamentados no perfil do mesmo.

4.2.2 Previsão

A tarefa de previsão é semelhante à tarefa de classificação, nessa tarefa os conjuntos de dados são classificados conforme uma atividade prevista ou valor estimado.

Na previsão a única forma de checar a acuracidade da classificação é esperar e ver (Berry; Linoff, 1997). Abaixo estão listados alguns exemplos de previsão.

- Prever se as ações da bolsa de valores subirão ou desceram nas próximas 24 horas;
- Prever futuros devedores dentro dos próximos seis meses;
- Prever consumidor que não comprará um produto nos próximos seis meses.

4.2.3 Agrupamento por afinidade

O agrupamento tem o objetivo de identificar grupos fundamentados nas semelhantes dos atributos do conjunto de dados.

As cadeias de varejo usam esta técnica para planejar a disposição dos produtos nas prateleiras das lojas ou em um catálogo, de modo que os itens geralmente adquiridos na mesma compra sejam vistos próximos entre si (Harrison, 1998).

Abaixo serão apresentados alguns exemplos de tarefa por agrupamento.

- Prever quais produtos poderiam ser colocados próximos para aumentar as vendas de um mercado;
- Identificar grupos de domicilio utilizando atributos como: escolaridade, idade, sexo, número de pessoas da casa;

- Auxiliar na organização dos produtos de um e-commerce.

4.2.4 Estimação

Segundo Harrison (1998), a estimação trabalha com resultados contínuos. Dado algum dado de entrada, nós usamos a estimativa para estipular um valor a uma variável contínua desconhecida, tal como renda, altura ou limite de cartão de crédito.

De maneira simplificada podemos dizer que essa tarefa é conduzida por um padrão já existente que visa predizer algum valor. Abaixo serão listados três exemplos da utilização dessa tarefa.

- Estimar a probabilidade de um consumidor não finalizar uma compra;
- Estimar o número de pessoas de uma casa;
- Estimar a renda de uma família.

4.2.5 Segmentação

Segmentação ou *clustering* é a tarefa de segmentar uma população heterogênea em um número maior de subgrupos homogêneos ou *clusters*. No *clustering* não há classes predefinidas (Berry; Linoff, 1997).

A diferença entre a tarefa de segmentação e a tarefa de classificação é que na primeira não existem classes predefinidas, já na segunda existem. Abaixo estão listados dois exemplos desta tarefa.

- Segmentação de animais com base em características semelhantes;
- Segmentação de funcionários em um projeto de acordo com suas habilidades.

4.2.6 Descrição

O intuito da tarefa de descrição de forma sucinta é descrever o que está acontecendo em uma base de dados. Essa descrição facilita a analise para a explicação de um comportamento por exemplo.

Conforme Harrison (1998) a divergência de gênero na política americana é um exemplo de como uma simples descrição “o número de mulheres que apoiam os democratas é maior do que o de homens” pode provocar grande interesse e estudos por parte de jornalistas, sociólogos, economistas e cientistas políticos, sem contar os próprios candidatos.

4.3 Técnicas de Data Mining

Berry e Linoff (1997) salientam que nenhuma técnica resolve todos os problemas de mineração de dados. A familiaridade com uma variedade de técnicas é necessária para encontrar o melhor caminho para resolver estes problemas.

Harrison (1998) indica que não há uma técnica que resolva todos os problemas de mineração de dados. A escolha dependerá da tarefa específica a ser executada e dos dados disponíveis para análise.

Nas próximas seções são apresentadas algumas técnicas de DM.

4.3.1 Redes neurais artificiais

De acordo com Harrison (1998) as redes neurais são provavelmente a técnica de mineração de dados mais comum, talvez sinônimo de mineração de dados para algumas pessoas.

Segundo Almeida (1995) as redes neurais têm sua origem em pesquisas neurológicas, e seu modelo de base é o cérebro humano. Como no cérebro humano, as redes neurais possuem neurônios interconectados de modo que os dados os percorram. Esses neurônios transmitem informação através de sinapses ou conexões.

O conceito-chave das redes neurais é a utilização de dados na criação de bases de conhecimentos. As redes neurais, ao contrário dos sistemas especialistas não precisam de um especialista para a criação da sua base de conhecimentos. Não trabalha com regras, sua aquisição é feita automaticamente a partir de exemplos coletados em bancos de dados (Almeida,1995).

Conforme Kotler (1998) o software de redes neurais, projetado conforme os padrões das células do cérebro humano pode, realmente, "aprender" a partir de grandes conjuntos de dados. Ao examinar repetidamente milhares de registros de dados, o software pode desenvolver um modelo estatístico poderoso descrevendo os relacionamentos e os padrões de dados importantes - nada que um pesquisador humano tenha tempo (ou capacidade visual) de fazer de maneira rigorosa e consistente.

Uma das principais vantagens das redes neurais é a sua variedade de aplicações. Devido a sua utilidade, as ferramentas que suportam redes neurais são fornecidas por várias empresas para uma variedade de plataformas. As redes neurais são interessantes também porque detectam padrões nos dados de forma analógica ao pensamento humano – um fundamento interessante para uma ferramenta de data mining (Pereira, 1998).

As redes neurais apresentam duas desvantagens: a dificuldade de compreender os modelos produzidos por elas e a particular sensibilidade ao formato dos dados que as alimentam. Representações de dados diferentes podem produzir resultados diversos, e o ajuste dos dados é uma parte significativa do esforço para utilizá-las (Bartolomeu, 2002).

A técnica de redes neurais artificiais é apropriada para as seguintes tarefas: classificação, previsão e estimação.

4.3.2 Árvores de decisão e indução de regras

As árvores de decisão são usadas para a mineração de dados direta (Harrison, 1998), particularmente para a classificação.

Berry e Linoff (1997) dizem que as árvores de decisão são um modelo poderoso produzido por uma classe de técnicas que inclui árvores de regressão e de classificação.

Segundo Kimball (1998), uma das principais vantagens das árvores de decisão é que o modelo é bem explicável, uma vez que tem a forma de regras explícitas. Isso permite às pessoas avaliarem os resultados, identificando atributos-chave no processo. Isso também é útil

quando os dados que entram possuem qualidade incerta. As próprias regras podem ser expressas facilmente como declarações lógicas em uma linguagem como *Structured Query Language* - SQL, de modo que possam ser aplicados diretamente em novos registros.

Harrison (1998) identificou como uma das principais vantagens das árvores de decisão a facilidade de explicação de seu modelo, devido a sua forma de regras explícitas.

Árvores de decisão é uma técnica apropriada para as seguintes tarefas: classificação e previsão.

4.3.3 Análise de seleção estatística

É uma forma de agrupamento usada para encontrar grupos de itens que tendem a ocorrer em conjunto em uma transação ou seleção estatística. Como técnica de agrupamento, é útil quando desejamos saber quais itens ocorrem ao mesmo tempo ou em uma sequencia particular. A informação resultante pode ser usada para vários objetivos, como planejar a arrumação de lojas, criar “pacotes” de produtos, entre outros (Harrison, 1998).

Souza (2000) faz referência a um exemplo típico de resultados da análise da seleção estatística que é o seguinte: compradores de ferramentas adquirem martelo e pregos ao mesmo tempo, assim como compradores de tinta adquirem também pincéis, mas não vice-versa.

Essa técnica utiliza a tarefa de agrupamento, é útil quando é necessário identificar ocorrências em um mesmo período de tempo ou em sequencia particular.

4.3.4 Algoritmos genéticos

Os algoritmos genéticos aplicam mecanismos de seleção genéticos e naturais para uma busca usada para encontrar conjuntos de parâmetros ótimos que descreve uma função preditiva. É usado para mineração de dados direta (Berry, 1997).

Os algoritmos genéticos usam operadores seleção, cruzamento e mutação para desenvolverem sucessivas gerações de soluções. Com a evolução do algoritmo, somente os mais previsíveis sobrevivem, até as funções convergirem em uma solução ideal. O algoritmo genético tem sido muito usado para aprimorar a técnica de redes neurais (Harrison 1998).

A técnica é apropriada para as tarefas de estimação e segmentação.

4.4 Ferramentas de Data Mining

Existem várias ferramentas que realizam a aplicação das técnicas de data mining. Alguns exemplos de ferramentas disponíveis são: Clementine, ODM (Oracle Data Mining), IBM Intelligent Miner e a Weka. Para a realização da mineração foi utilizada nesta pesquisa a ferramenta Weka.

A Weka é uma das ferramentas livres mais utilizadas para fins de pesquisa, seus algoritmos podem ser aplicados por meio da ferramenta ou usados em aplicações Java.

Para utilizar os algoritmos através da ferramenta usa-se sua interface gráfica, a Weka Explorer, apresentada na Figura 4. Com ela é possível aplicar as técnicas de data mining de forma simples e prática.

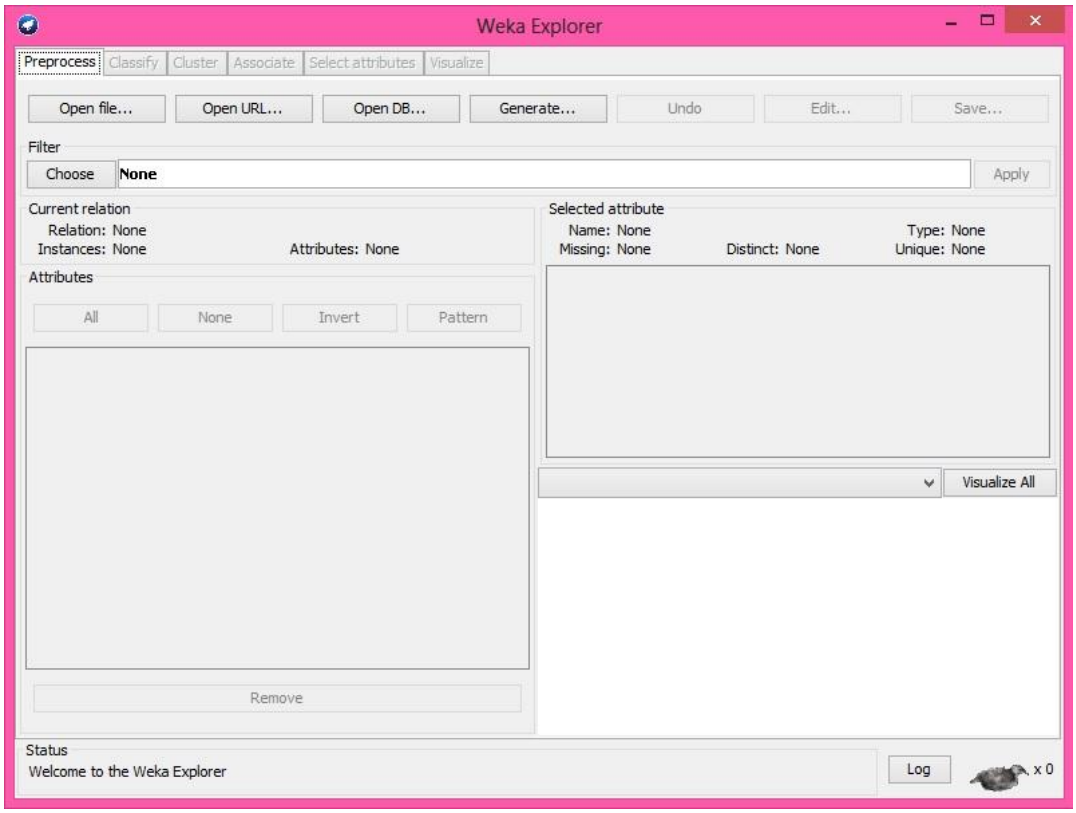


Figura 4. Weka Explorer

5 MODELO PROPOSTO DE DATA MINING EM REPOSITÓRIOS DIGITAIS

Com o estudo sobre Data Mining foi possível verificar a existência de diversas tarefas, onde para cada uma delas há técnicas específicas a serem utilizadas. E após o levantamento teórico foi possível concluir que tais técnicas poderiam ser utilizadas no processo de recuperação da informação nos repositórios digitais, conforme arquitetura apresentada na Figura 5.

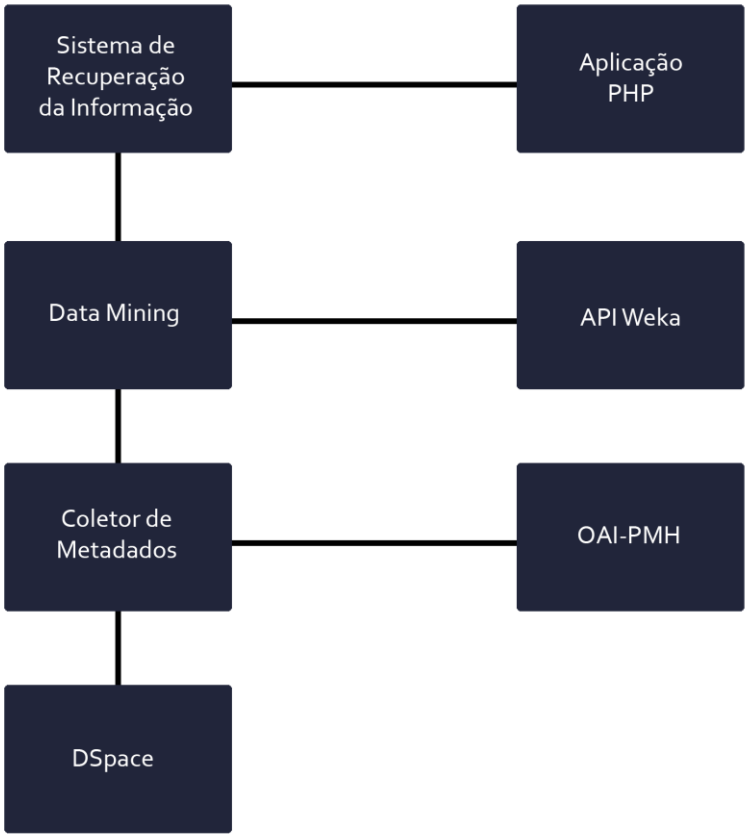


Figura 5. Arquitetura Proposta de Recuperação da Informação.

A arquitetura propõe a implementação de uma camada de Recuperação da Informação utilizando técnicas de Data Mining nos Repositórios Digitais. Essa camada trata-se de um aplicativo coletando os metadados, para que posteriormente possam passar pelas técnicas de Data Mining.

5.1 Coleta dos metadados

Diante da arquitetura proposta, foi desenvolvida em linguagem PHP na IDE Eclipe, uma aplicação responsável pela extração de metadados do servidor DSpace.

Para realizar a extração desses metadados foi necessário utilizar o comando ListRecord do protocolo OAI-PMH. Na Figura 6 é apresentada a *url* utilizada para retornar os metadados e uma breve explicação da mesma.

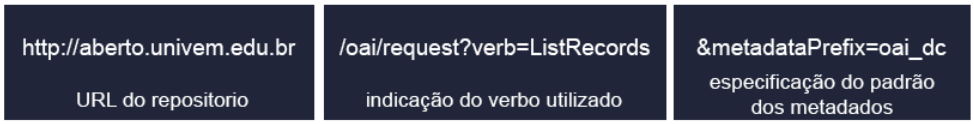


Figura 6. URL de requisição

Essa requisição retorna em XML todos os objetos presentes no repositório. É possível também realizar esta mesma requisição para um determinado período de tempo, com a utilização de argumentos disponíveis para este verbo.

5.2 Transformação e armazenamento dos metadados

Para que seja possível realizar a mineração nos metadados requisitados é necessário transformá-los e armazená-los de maneira que seja possível sua manipulação posteriormente. Diante disso, a aplicação foi desenvolvida para receber o XML resultante da requisição realizada e transformá-lo em array, para que sejam realizadas as operações necessárias com esses dados.

Com a transformação, esses dados são armazenados em um banco de dados que utiliza a linguagem SQL e o MySql como SGBD (Sistema de Gerenciamento de Banco de Dados). Na Figura 7 é apresentado o MER (Modelo Entidade Relacionamento) do banco, que se trata de uma forma de descrever os aspectos da informação que ali será armazenado.

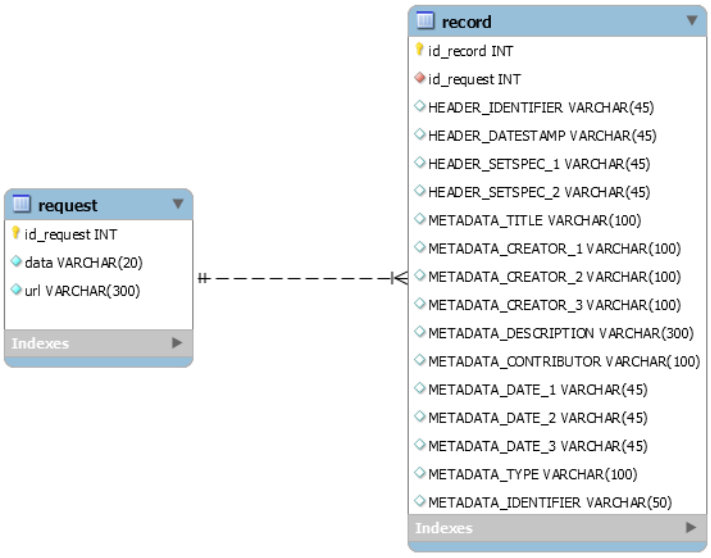


Figura 7. Modelo Entidade Relacionamento da aplicação

A tabela *request* é responsável por armazenar os dados da requisição, ela é formada pelos atributos id, data e *url*. O atributo id é o identificador de uma requisição, o data guarda a data em que foi feita a requisição e o atributo *url* é responsável por armazenar a *url* que foi gerada na requisição.

A tabela *record* contem atributos para armazenar os dados que foram coletados do servidor DSpace. Esses atributos foram criados com base nos elementos do Dublin Core, que foi abordado anteriormente nessa pesquisa.

5.3 Conexão entre o SGBD e a Weka

A Weka originalmente trabalha com bases de dados nos formatos “.ARFF” e “.CSV”, ou seja, com formato texto. Porém é possível também fazer a mineração a partir de tabelas de SGBDs relacionais.

Pensando em praticidade, essa é a melhor forma de se utilizar tal ferramenta já que a maioria das aplicações tem seus dados armazenados em bancos relacionais.

Para que a mineração nessas bases de dados seja realizada é necessário estabelecer uma conexão entre a ferramenta e o banco de dados. Sem a conexão seria necessário exportar os dados do banco de dados para texto e posteriormente transformá-los em “.ARFF” ou “.CSV”, o que levaria um tempo maior para realizar.

A Weka foi desenvolvida em Java, por esse motivo para realizar a conexão entre a mesma e o banco de dados fez-se uso do driver JDBC. Esse driver permite que uma aplicação Java possa interagir com um banco de dados.

Com a configuração correta basta executar a Weka a partir de uma janela de prompt permitindo a opção de conexão com o banco MySql, como pode ser visto na Figura 8.

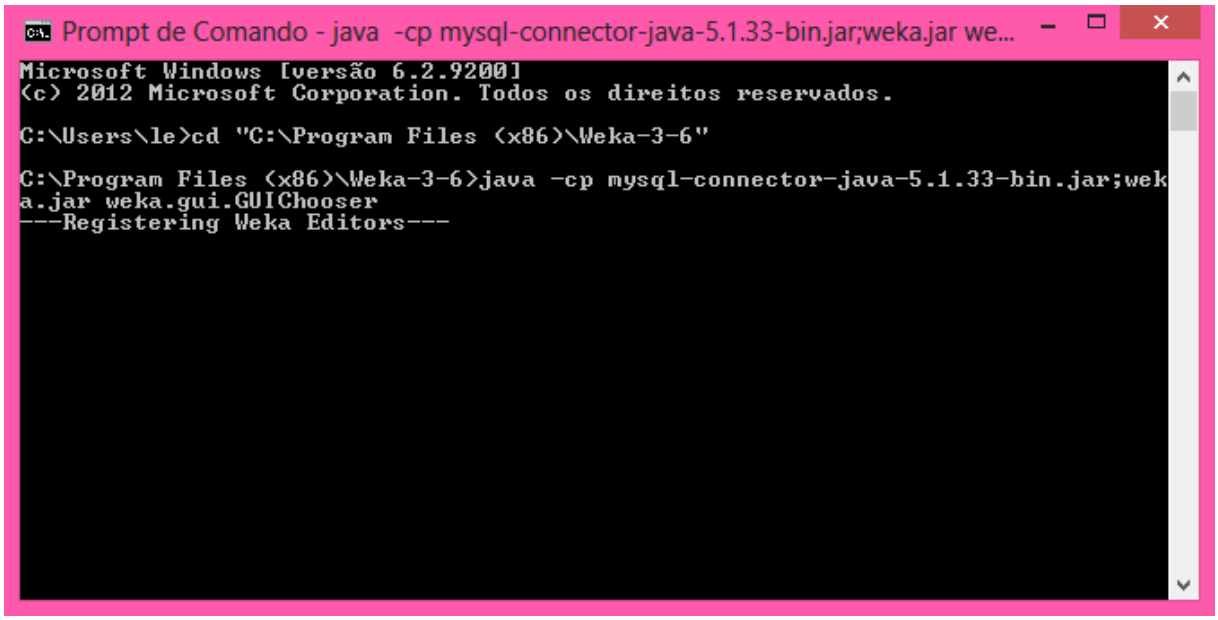


Figura 8. Executar Weka por linha de comando

O comando "-cp" indica onde estão as bibliotecas que são necessárias para rodar o programa desejado, depois é necessário especificar uma lista de arquivos JAR separando por ";" caso o sistema operacional seja Windows ou ":" no Linux. Nesse caso, o primeiro arquivo é o JAR do *Driver* JDBC e o segundo é o JAR da ferramenta Weka. E por fim deve-se apontar para o programa que se deseja rodar, que no caso é a interface gráfica da Weka.

Com a interface aberta é necessário inserir os parâmetros da conexão para conectar a ferramenta ao banco de dados. Após esse passo deve-se inserir uma *query* para obter os dados que o usuário deseja minerar. Esses dados são retornados e exibidos na ferramenta.

Como pode ser visto na Figura 9, foi executada uma *query* que retorna os dados que foram coletados do servidor DSpace.

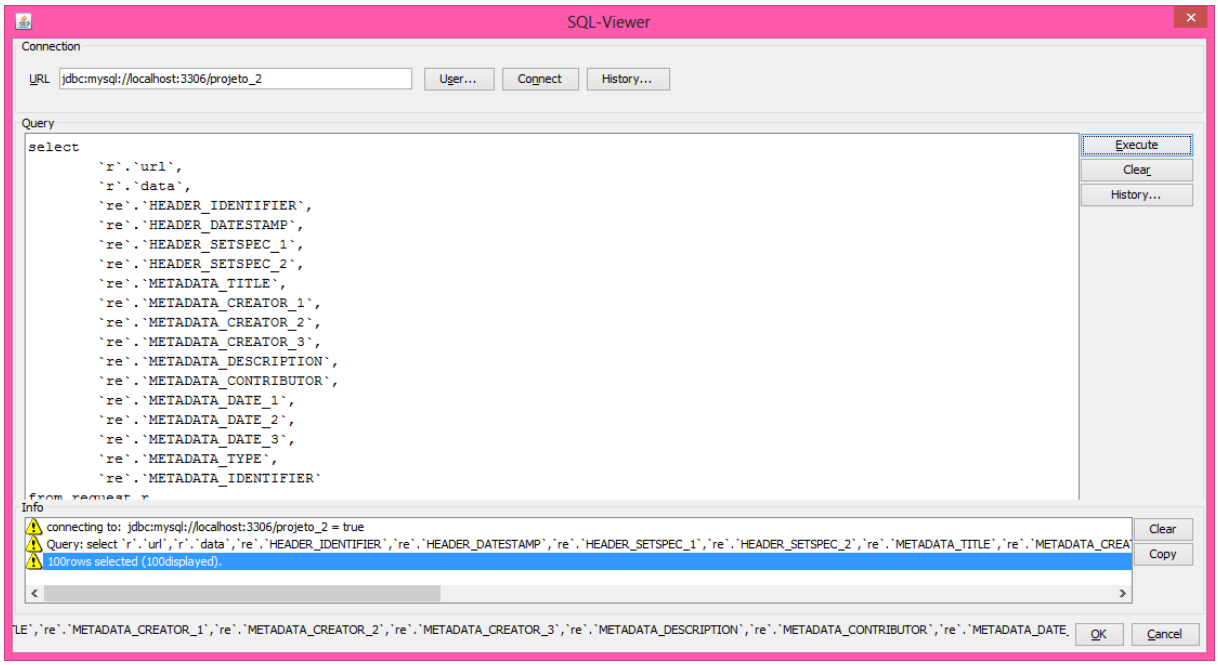


Figura 9. Execução de Query na Weka

Com os dados retornados é possível aplicar vários tipos de técnicas de mineração de dados, como será exemplificado na próxima seção.

5.4 Minerando dados na Weka

Essa seção aborda a aplicação de uma técnica nos dados que foram recuperados no servidor DSpace. Com a execução da *query* as informações do banco de dados poderão ser visualizadas na Figura 10.

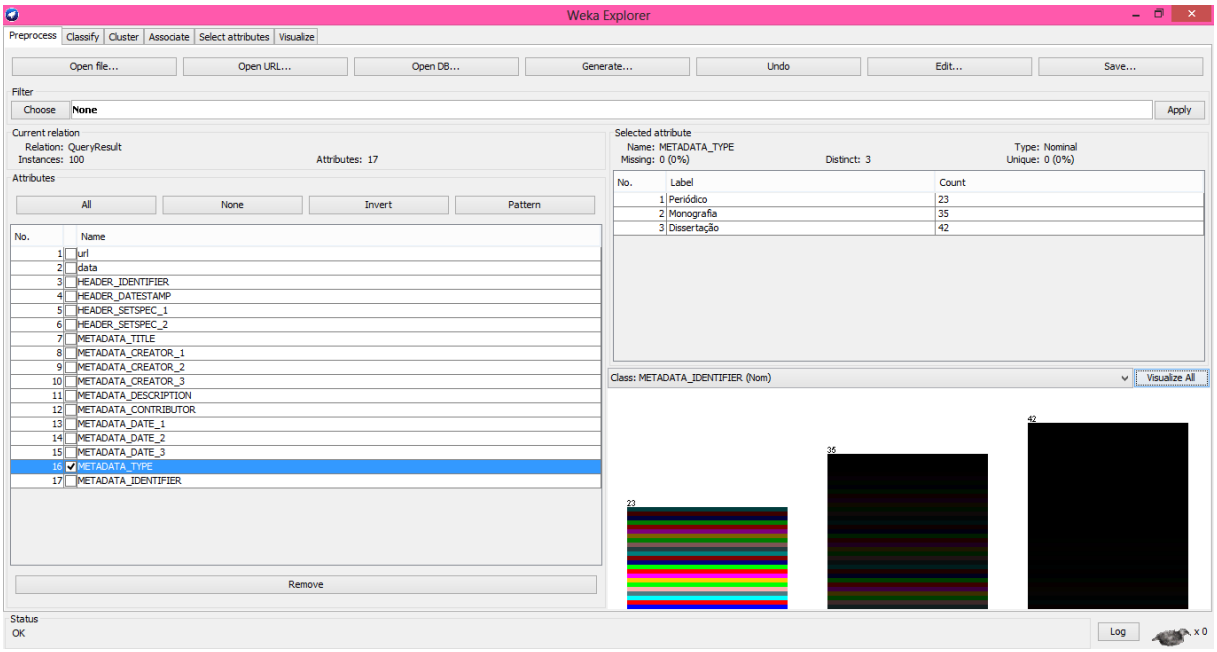


Figura 10. Mineração de dados com Weka

Na Figura 10 o atributo selecionado é o *Type*, esse atributo descreve qual o tipo do documento. Nesse caso existem três tipos: Periódico, Monografia e Dissertação, que podem ser vistos na coluna *Label*. A coluna Count apresenta a quantidade de ocorrências de cada *Label* que no total contabilizam 100 instâncias.

Para exemplificar foi utilizado um algoritmo que realiza a tarefa de classificação, mais especificamente baseado em arvores de decisão, o J48.

O algoritmo C4.5, desenvolvido na liguagem C, necessitava ser recodificado para a liguagem Java dessa forma surgiu o algoritmo J48. Atualmente é um dos algoritmos de mineração de dados mais utilizados.

Após escolher o algoritmo desejado, o usuário tem a opção de escolher o tipo de teste que deseja fazer, foi utilizada a opção *Percentage split* inserindo o valor de 70%. Isso quer dizer que o algoritmo utilizará 70% das 100 instâncias para aprender, os outros 30% serão utilizados para a realização dos testes.

Na Figura 11 é possível visualizar o resultado de aplicação do algoritmo J48 com as configurações de teste selecionadas.

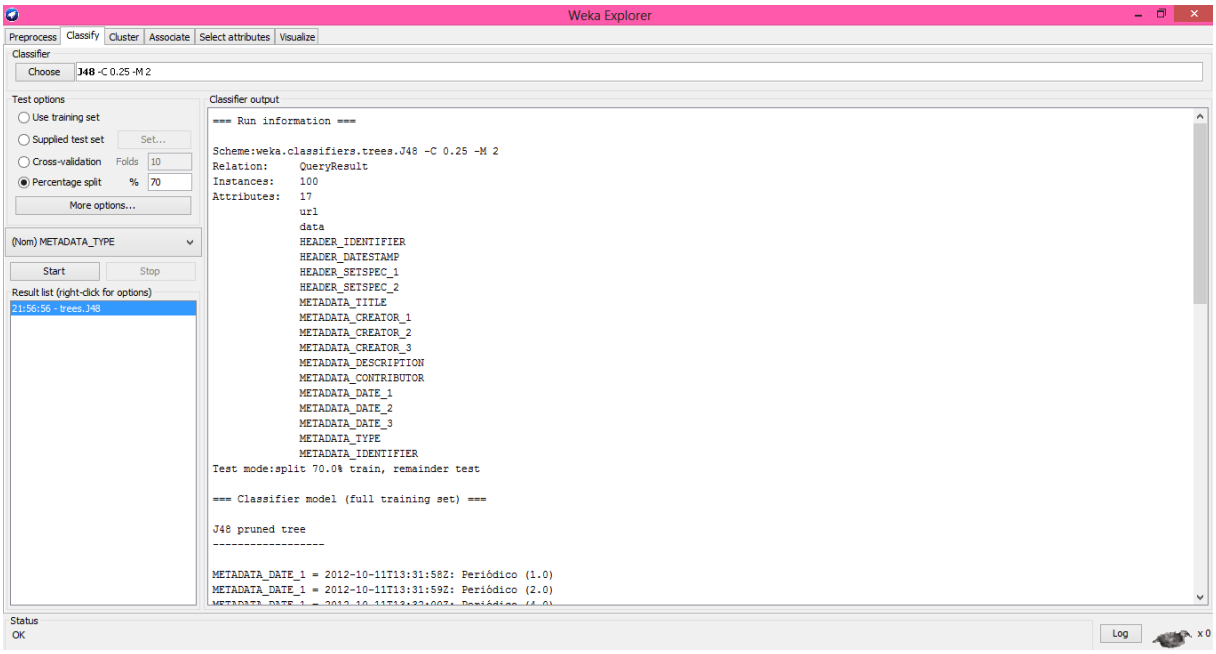


Figura 11. Resultado da Mineração de dados

Esse foi um exemplo de como uma técnica de mineração de dados utilizando a ferramenta Weka pode ser aplicada em uma base de dados. Porém, existem várias outras técnicas que podem ser utilizadas.

CONCLUSÃO

O intuito da presente pesquisa foi propor o uso da mineração de dados como ferramenta no processo de recuperação da informação dentro dos repositórios digitais institucionais visando melhorar esses processos trazendo melhor experiência ao usuário.

O levantamento bibliográfico possibilitou conhecer os conceitos que envolvem os repositórios digitais, assim como os modelos clássicos da recuperação da informação e por fim conceitos, tarefas e técnicas de data mining que podem ser aplicadas a este contexto.

Por meio desta pesquisa pôde-se verificar que a utilização de técnicas de mineração de dados em repositórios digitais institucionais ainda é pouco explorada, porém de grande valia quando utilizada adequadamente.

A união dos conceitos de RI e DM permite uma melhor utilização e exploração do conhecimento resguardado nestes repositórios. O resultado é a maximização da disseminação de conhecimentos que muitas vezes não são explorados por mera falta de um método que facilite e entregue de forma mais objetiva as informações sobre determinado assunto buscado pelo usuário.

A partir da hipótese levantada o primeiro passo foi a elaboração de uma arquitetura que apresenta a proposta de um modelo diferenciado para a recuperação da informação nos repositórios.

A arquitetura propõe uma camada de recuperação para os metadados por meio de um aplicativo desenvolvido em PHP. Posteriormente ocorre a aplicação das técnicas de data mining por meio da ferramenta Weka. Para enfim, a requisição a ser entregue ao usuário de maneira mais clara e objetiva, sem a necessidade de uma busca mais complexa por parte do mesmo.

Com a implementação da camada de recuperação da informação foi aplicado o algoritmo J48 nos metadados que foram coletados, exemplificando que o que foi proposto realmente pode ser feito.

Portanto, conclui-se que os objetivos do trabalho foram atingidos. Uma vez que:

- Foram apresentadas as técnicas de Data Mining que podem ser usadas como ferramenta no processo de recuperação da informação;
- Foi estudado o padrão de metadados para que os mesmos pudessem ser armazenados posteriormente;

- Foi realizado o armazenamento dos metadados coletados do servidor DSpace;
- O protótipo que implementa a camada de Recuperação da Informação utilizando técnicas de DM mostrou-se eficiente para o seu propósito.

Por fim compreende-se que a aplicação do algoritmo foi apenas um exemplo de utilização da ferramenta Weka em banco de dados relacionais. Porém existem vários outros algoritmos que podem ser utilizados.

Lembrando que não existe a melhor técnica, é necessário identificar qual técnica se encaixa melhor para a necessidade que deseja contemplar.

REFERÊNCIAS

ALMEIDA, Fernando C. Desvendando o uso de redes neurais em problemas de administração de empresas. RAE, São Paulo, 1995.

ARANHA, Francisco. Análise de Redes em Procedimentos de Cooperação Indireta: Utilização no Sistema de Recomendações da Biblioteca Karl A. Boedecker. São Paulo: EAESP/FGV/NPP, 2000.

BAEZA-YATES, R. e RIBEIRO-NETO, B. Modern information retrieval, ACM Press, 1999.

BAEZA-YATES, R. e RIBEIRO-NETO, B. Modern information retrieval: The Concepts and Technology Behind Scarch, 2ad Edition, Arrangement with Pearson Education Limited, 2011.

BARTOLOMEU, Tereza Angélica. Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento. 2002. 302f. Tese (Doutorado em Engenharia de Produção) - EPS. Universidade Federal de Santa Catarina, Florianópolis, 2002.

BARTON, M. R. Creating an institutional repository: LEADIRS workbook. Cambridge-MIT Institute, 2005.

BERRY, Michael J. A., LINOFF, Gordon. Data Mining techniques: for marketing, sales and customer support. USA: Wiley Computer Publishing, 1997.

BRUSSO, Marcos José. O paralelismo na mineração de regras de associação. Porto Alegre: UFRGS, 1998.

CAFÉ, L. et al. Repositórios institucionais: nova estratégia para publicação científica na Rede. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 26. 2003, Belo Horizonte. Anais... Belo Horizonte: INTERCOM, 2003.

CARDOSO, Olinda Nogueira Paes; Recuperação da Informação. Universidade Federal de Lavras, 2000.

CARVALHO, D. R. Um método hibrido árvore de decisão/algoritmo genético para data mining. Curitiba, 2002.

CASTRO, F. F. de; SANTOS, P. L. V. A. C. MarcOnt Initiative: representação e descrição de recursos informacionais na web. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO- ENANCIB, 9., 2008, Anais eletrônicos... São Paulo: ANCIB, 2008.

CRESTANI, F., LALMAS, M., VAN RIJSBERGEN, C. J., CAMPBELL I. "Is This Document Relevant?...Probably": A Survey of Probabilistic Models in Information Retrieval. ACM Computing Surveys, 1998.

DINIZ, Carlos Alberto R., NETO, Francisco Louzada. Data Mining: uma introdução. São Paulo: Associação Brasileira de Estatística, 2000.

DZIEKANIAK, G. V.; KIRINUS J. B. Web Semântica. Revista Eletrônica de Biblioteconomia e Ciência da Informação. 2004.

FAYYAD, U. M. et al. From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining. California: AAAI/The MIT, 1996. p.1-34.

FELDENS, M. A.; Descoberta de conhecimento aplicada à detecção de anomalias em bases de dados. Porto Alegre, 1996.

FRAKES, W. B. & BAEZA-YATES, R. Information Retrieval Data Structures & Algorithms, Prentice Hall, 1992.

FUHR, N, PFEIFER, U. Probebilistic Information Retrieval as a Combination of Abstraction, Inductive, Learning, and Probabilistic Assumptions. ACM Transactions on Information Systems, Vol. 12, No, 1, pp. 92-115, 1994

GEY, F. "Models in Information Retrieval". Folders of Tutorial Presented at the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR), 1992.

HARRISON, Thomas H. Intranet data warehouse. São Paulo: Bekerley Brasil, 1998.

JESUS, Alberto Pereira de. Data Mining aplicado à identificação do perfil dos usuários de uma biblioteca para a personalização de sistemas Web de recuperação e disseminação de informações. Florianópolis, 2014.

KIMBALL, R.; REEVES, L.; ROSS, M.; THORNTHWAITE, W. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. New York, John Wiley & Sons, 1998.

KOTLER, Philip. Administração de marketing: análise, planejamento, implementação e controle. 5ª ed. São Paulo: Atlas, 1998.

KURAMOTO, H. Informação científica: proposta de um novo modelo para o Brasil. Ciência da Informação, Brasília, 2006.

LEWIS, S.; YATES, C. The DSpace Course – Introduction to Dspace. CADAIR, 2008.

LIDDY, E.D. Enhanced text retrieval using Natural Language Processing. Bulletin of the American Society for Information Science, 1998.

LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. Brasília, 2002.

LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. Association of Research Libraries, n. 226, 2003.

MIORELLI, Sandra Teresinha. ED-CER: Extração do Sintagma Nominal em Sentenças em Português. 2001. 98f. Dissertação de Mestrado em Ciência da Computação - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul.

MOOERS, C."Zatocoding applied to mechanical organization of knowledge. In American Documentation, 1951.

NICHOLSON, Scott. The bibliomining process: data warehousing and data mining for library decision-making, 2004.

OLIVEIRA, Djalma de Pinho Rebouças. Sistemas de informações gerenciais: estratégicas, táticas e operacionais. 4ª ed. São Paulo: Atlas, 1997.

PEREIRA, W. A. L. Data warehouse – trabalho individual II. Porto Alegre, PUCRS, 1998.

ROMANI, Lucas S.; FUSCO, Elvis; SANTOS, Plácida L. V. Amorim da Costa Santos. Análise e Implantação de Repositório Digital Utilizando Software Livre DSPACE.

Programa de Pós-Graduação em Ciência da Informação. Universidade Estadual Paulista, 2010.

ROMANI, Lucas Salviano. Análise e Implantação de Repositório Digital utilizando Software Livre DSpace. 2009. Trabalho de Conclusão de Curso de Bacharelado em Ciência da Computação. Centro Universitário Eurípides de Marília, 2009.

SANTAREM SEGUNDO, José Eduardo. Representação Iterativa: um modelo para Repositórios Digitais / José Eduardo Santarem Segundo. – Marília, 2010.

SCHOLTES, J. C. Neural Networks in Natural Language Processing and Information Retrieval. PhD thesis, Institute for Logic, Language and Computation (ILLC).University of Amsterdam, 1993.

SHOLOM, M. Weis, Nitim Indurkhya; "Predict Data Mining"; Morgan Kaufmann Publishers, Inc, 1999.

SOUZA, O. R. M. Mineração de dados de um plano de saúde para obter regras de associação. Florianópolis, 2000. Dissertação (Mestrado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina.