2
Fundamentação

Neste capítulo são abordados conceitos importantes para uma melhor compreensão desta tese. Na primeira seção são apresentados conceitos de Ambiente de Aprendizagem e Biblioteca Digital, ambientes de *softwares*, cujos repositórios serão integrados. A segunda seção contempla a descrição das principais arquiteturas para integração de dados e tratamento de heterogeneidade. Finalmente, na terceira seção são apresentadas as técnicas de *Text Mining* (mineração de texto), técnica esta utilizada para extrair objetos de aprendizagem dos DDs.

2.1.Ambientes de Aprendizagem e Bibliotecas Digitais

Este trabalho visa integrar repositórios de Ambientes de Aprendizagem e Bibliotecas Digitais, que serão brevemente descritos a seguir.

2.1.1.Ambientes de Aprendizagem

E-learning vem despertando interesse de grandes e pequenas empresas, dispostas a desenvolver tecnologias ou oferecer serviços variados. Das tecnologias disponíveis, as mais relevantes para o *E-learning* estão relacionadas à autoria de cursos e LMS. A maioria dos LMS disponíveis no mercado possibilita acessar cursos *on-line* e outras atividades de aprendizagem, matricular alunos, coletar e armazenar dados sobre a atuação dos estudantes e administrar usuários e cursos.

Segundo Guest e Juday (2002), LMS são a espinha dorsal de um projeto de *E-learning*, pois centralizam as funções e os processos de aprendizagem, registrando os passos de cada usuário, estatísticas de desempenho e cursos oferecidos.

Segundo Raghavan (2001), um LMS, do ponto de vista do usuário final, fornece uma maneira eficaz de atender suas necessidades individuais e meios de localizar e registrar facilmente as atividades de aprendizagem mais relevantes para melhorar os níveis de habilidade do aprendiz. Um LMS não focaliza a criação, a reutilização, a gerência ou a melhoria do conteúdo de aprendizagem. Para isso surgiram os LCMS (Learning Content Management Systems), que gerenciam o conteúdo de aprendizagem e respectivos metadados.

O material ou conteúdo de aprendizagem é referenciado na literatura especializada por LOs. Não será discutido acerca do termo utilizado e sua definição mais formal, mas o importante é ressaltar que há um enfoque na reutilização destes conteúdos de aprendizagem. A arquitetura de um ambiente de aprendizagem (**A1**) está representada na Figura 1, onde o usuário acessa o LMS. Este, por sua vez, acessa o LCMS, que gerencia os LOs.

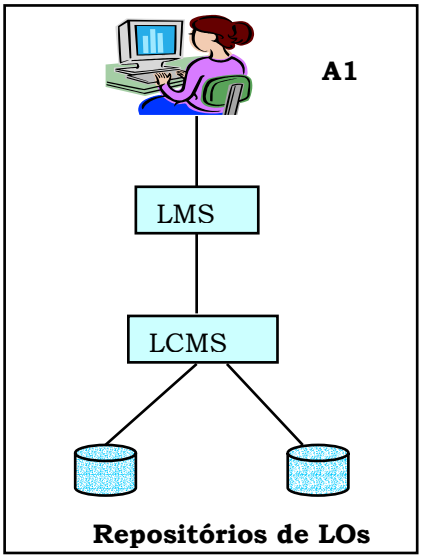

Figura 1 – Arquitetura do Ambiente de Aprendizagem

Um LCMS pode pesquisar e recuperar um LO para o usuário final como uma unidade individual para satisfazer uma necessidade específica ou para recuperar o objeto como parte de um curso, um currículo ou uma atividade de aprendizagem definida em um LMS.

Portanto, o LMS e o LCMS representam duas categorias de produtos distintas e complementares, geralmente presentes em um Ambiente de Aprendizagem.

2.1.2. Bibliotecas Digitais

O surgimento da Internet a partir dos anos 90 vem mudando de maneira radical o papel das bibliotecas no ciclo de intermediação e acesso a documentos. Também vem tendo um impacto significativo nas formas de comunicação científica e, conseqüentemente, nos sistemas de informação em Ciência e Tecnologia. Diferentes processos sociais, econômicos e tecnológicos convergem para configurar a situação atual das formas de comunicação científica.

Com o crescimento da informação digital e do número de documentos eletrônicos, as bibliotecas tradicionais estão incorporando gradualmente os serviços das bibliotecas digitais. No entanto, a designação de "biblioteca digital" não é um termo único para definir uma mesma idéia. Segundo Barker (1994) e Marchori (1997) existem quatro definições de bibliotecas: convencional (polimídia), eletrônica, digital e virtual. A seguir, essas definições serão descritas de forma objetiva, buscando fornecer subsídios para os leitores desse trabalho.

- **Biblioteca Convencional ou Polimídia** - atualmente diversas mídias são utilizadas como meios independentes para armazenamento da informação. As bibliotecas convencionais possuem livros e periódicos que convivem com fitas, vídeos, CD-ROMs, microfilmes, *software* de controle e armazenamento etc., dando-lhes também o nome de bibliotecas polimídias. Os processos de gerenciamento e organização nestas bibliotecas são praticamente manuais e, apesar de os computadores estarem disponíveis para os usuários, esta tecnologia não é utilizada para a realização de qualquer forma de automação das bibliotecas**.**

- **Bibliotecas Eletrônicas** – a biblioteca eletrônica se refere aos sistemas de bibliotecas nos quais os processos básicos da biblioteca são de natureza eletrônica, o que implica ampla utilização de computadores e de suas facilidades na construção de índices *on-line*, busca de textos completos e na recuperação e armazenamento de registros.

- **Biblioteca Digital** - a biblioteca digital contém apenas informação na forma digital, podendo residir em meios diferentes de armazenamento, como memórias eletrônicas, tais como os discos magnéticos e óticos. Desta forma, a biblioteca digital não contém livros na forma convencional. A informação pode ser acessada em locais específicos e/ou remotamente, por meio de redes de computadores.

- **Biblioteca Virtual** - a biblioteca virtual é conceituada como um tipo de biblioteca que, para existir, depende da tecnologia da realidade virtual. Neste caso, um *software* próprio acoplado a um computador sofisticado reproduz o ambiente de uma biblioteca em duas ou três dimensões, criando um ambiente de total imersão e interação. É então possível, ao entrar em uma biblioteca virtual, circular entre as salas, selecionar um livro nas estantes, "tocá-lo", "abri-lo" e "lê-lo". Obviamente, o único "lugar" onde o livro realmente existe é no computador e na imaginação do leitor.

Os sistemas de bibliotecas têm crescido muito e todas as bibliotecas polimídias estão se adequando às novas tecnologias**.** Contudo, dependendo da perspectiva, a designação de "biblioteca digital" poderá conter significados diferentes: significar a simples automatização das bibliotecas tradicionais, tendo funções de biblioteca de uma nova forma ser um sistema de informação baseado em um repositório de informação *on-line* ser uma coleção de serviços de informação distribuídos; ou ser um espaço distribuído de informação interligada ou um sistema neural de informação multimídia. Todas as definições mencionadas contribuem para a conceituação de bibliotecas digitais.

Com o crescimento das bibliotecas digitais e a grande tendência de compartilhar os conteúdos das mesmas, visando a facilitar esta disseminação dos conteúdos de forma eficiente, surgiu a OAI (*Open Archives Initiave*), que desenvolve e promove padrões de interoperabilidade entre repositórios digitais. Ela criou o protocolo OAI-PMH (*The Open Archives Initiative Protocol for Metadata Harvesting*), um meio simples de se compartilhar metadados entre servidores distribuídos.

Pode-se considerar, então, que uma Biblioteca Digital é um agrupamento de tecnologias de armazenamento e de comunicação, conjuntamente com o conteúdo e *software* necessários para estender os serviços fornecidos pelas bibliotecas polimídias baseadas em papel e em outros meios de catalogação, pesquisa e disseminação da informação.

A arquitetura de uma biblioteca digital envolve o seu sistema de gerência (DLMS), *software* que gerencia todos os serviços da biblioteca digital, tais como validação de usuários, consulta aos documentos armazenados em seus repositórios (DDs) e catalogação de DDs. Na Figura 2 está representada a arquitetura (**B1**) de uma biblioteca digital, onde um usuário acessa o DLMS, que gerencia todos os serviços da DL que, por sua vez, acessa os repositórios.

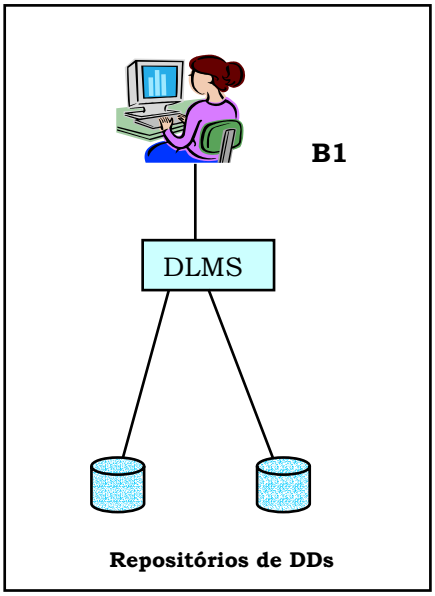

Figura 2 - Arquitetura do Ambiente de DL

2.2. Integração de Dados Heterogêneos

Com a evolução dos Sistemas de Gerência de Banco de Dados (SGBDs), as grandes empresas utilizam diferentes tecnologias de Banco de Dados, convivendo, assim, com um alto grau de heterogeneidade de *softwares*, bem como heterogeneidade de modelos de dados, tendo, então, que enfrentar complexos problemas de integração.

Para resolver estes problemas de integração de dados distribuídos e heterogêneos, existem hoje várias propostas de solução, tais como Banco de Dados Heterogêneos (Sheth & Larson, 1999) (Barbosa, 2001a), sistemas baseados em mediadores (Wiederhold, 1995) e tecnologia de agentes (Genesereth, 1994). As diferenças entre estes sistemas estão principalmente na forma de interação deste com os componentes locais

Na seção a seguir, será apresentada a tecnologia de sistemas baseados em mediadores para resolver o problema de heterogeneidade, pois será a tecnologia utilizada neste trabalho.

2.2.1.Mediadores

Os mediadores são definidos como módulos de software que exploram o conhecimento representado em um conjunto ou subconjunto de dados para criar informações para aplicações residentes em uma camada superior. Os mediadores são utilizados como uma camada intermediária entre a camada das aplicações e a camada das fontes de dados, tendo como função a aplicação do conhecimento especializado a um domínio específico para agregar valor (Wiederhold, 1992).

A arquitetura de mediadores efetua acesso aos dados distribuídos em múltiplas fontes de informação através de consultas que são submetidas ao sistema, via mediador, e este as transforma em subconsultas para serem enviadas às fontes de dados. As subconsultas geradas pelo mediador devem ser traduzidas para linguagens de consulta de cada fonte de dados componente. Ao final, os resultados das consultas são traduzidos e a resposta é devolvida ao usuário.

A arquitetura definida por Wiederhold utiliza hierarquia de camadas para separar as tarefas dos usuários, das bases de dados e dos mediadores, configurando, portanto, uma arquitetura de três camadas. As bases de dados podem ser autônomas e heterogêneas, uma vez que na maioria das vezes não foram projetadas visando o compartilhamento. Na Figura 6, vê-se que as interfaces, para serem suportadas, fornecem um corte onde serviços de rede de comunicação são necessários.

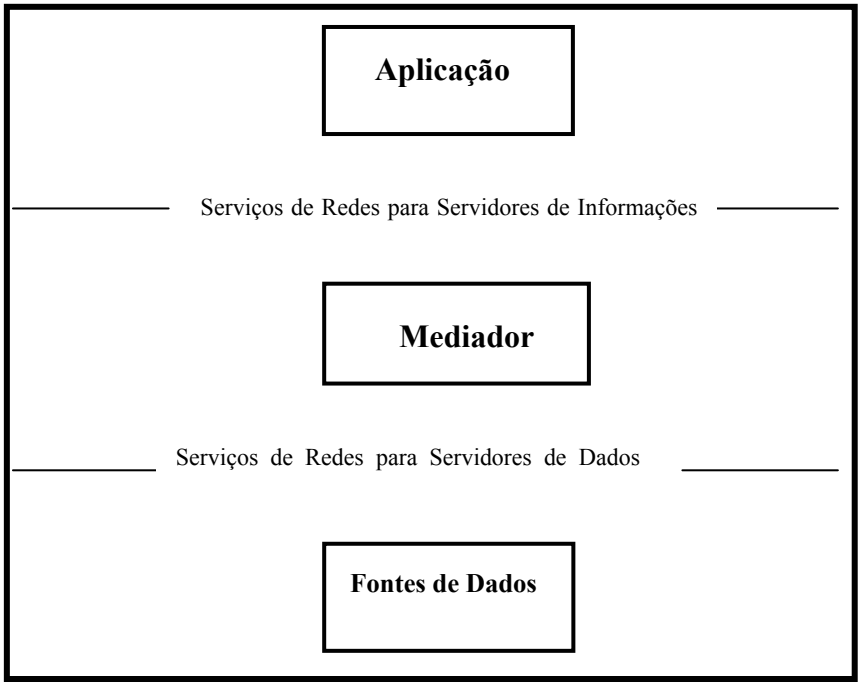

Figura 3 – Camadas da Arquitetura de Mediadores

Na camada da Aplicação os usuários acessam aplicações independentes. Na camada do Mediador múltiplos mediadores são gerenciados pelos especialistas de domínios. Na camada da Fontes de Dados têm-se múltiplas fontes de dados, gerenciadas pelos administradores de Banco de Dados.

A interface entre as estações de trabalho dos usuários e os mediadores necessita de uma linguagem de alto nível que forneça flexibilidade, repetição e avaliação. A estrutura básica da linguagem deve ser adaptável para que novas funções possam ser suportadas pelo mediador, junto à rede, para desenvolver novas funcionalidades. É importante observar que é necessário uma máquina e uma interface de comunicação amigável.

A interface entre o mediador e as fontes de dados pode utilizar padrões de banco de dados, como SQL e protocolo Remote Data Access (RDA) que fornecem acessos a bases de dados pelo mediador. Um mediador que combina informações de diversos bancos de dados pode usar seu conhecimento para controlar o processo de integração, especificando operações relacionais diretamente.

As três camadas da Figura 6 fazem um intercâmbio explícito, favorecendo

a flexibilidade sobre a integração, já que permitem a reorganização das estruturas dos dados e redistribuição dos dados sobre os nós de processamento das redes de comunicação, devido à separação entre as aplicações dos usuários e as fontes dos dados.

Os módulos mediadores serão mais eficientes se puderem servir a uma variedade de aplicações. As aplicações formarão as suas tarefas através da aquisição de informações do conjunto de mediadores disponíveis. O mediador, por sua vez, poderá usar diferentes visões sobre um ou muitos bancos de dados. Novos mediadores serão criados pela não disponibilidade da informação.

A distribuição dos mediadores, replicados ou não, através da rede de comunicação, seria feita pela melhor localização, a fim de facilitar a sua manutenção, o uso de seu conhecimento pelos bancos de dados e a sua autonomia. O uso de mediadores, conforme a Figura 6, é colocado por Wiederhold como uma evolução lógica da arquitetura cliente-servidor, pois na mediação uma camada extra de *software* é inserida entre o cliente e o servidor, rompendo esse acoplamento.

Os sistemas baseados em Mediadores possuem uma arquitetura bastante flexível, onde a integração dos dados é realizada no momento em que as consultas são submetidas.

Por causa da enorme quantidade de informações e fontes de dados disponíveis, tem-se a necessidade de utilizar metadados nesta arquitetura, para a heterogeneidade semântica destas fontes seja tratada.

2.2.2.Heterogeneidade Semântica

A heterogeneidade semântica é reconhecida como um dos mais importantes e difíceis problemas encontrados ao se lidar com interoperabilidade entre múltiplas fontes de banco de dados. Com o advento da Internet este problema tem se agravado muito por causa da enorme quantidade de informações e fontes disponíveis.. Para a sua solução, metadados e ontologias possuem um importante papel, uma vez que agem como descritores, podendo representar os conflitos de representação, os mapeamentos, a localização, assim como o tipo de cada fonte de dados.

2.2.2.1.Metadados

Segundo Sumpter (1994), "Metadado é a informação sobre o dado que permite o acesso e gerenciamento deste dado de maneira eficiente e inteligente."

Metadados se aplicam a uma grande variedade de acervos de dados convencionais que podem, ou não, estar disponíveis em redes eletrônicas de computadores, tais como acervos de dados bancários, de bibliotecas tradicionais ou acervos de dados não convencionais, como os de sistemas de informações geográficas, de bibliotecas digitais e de documentos multimídia.

Nos acervos de dados considerados convencionais, onde a informação se encontra estruturada, cada objeto do mundo real tem um identificador único. Nestes acervos, os problemas de gerenciamento e localização dos dados podem ser resolvidos utilizando SGBDs, que utilizam metadados para descrever e gerenciar os dados. O problema maior reside nos acervos não convencionais que envolvem dados não estruturados e que podem ser representados textualmente ou não.

A especificação e utilização de padrões metadados garantem a existência de um conjunto de informações comuns sobre um determinado tema ou área, com regras claramente estabelecidas e aceitas pela comunidade envolvida. Padrões facilitam a compreensão, a integração e o uso compartilhado de informações entre usuários de diferentes formações, com diferentes níveis de experiência e diferentes propósitos. O estabelecimento de padrões implica em um compromisso entre usuários e provedores de informações, que devem mutuamente aceitar, colaborar e usar as terminologias e definições estabelecidas.

Um padrão de metadados é formado por um conjunto de elementos descritores que podem estar relacionados. Geralmente são padronizados nomes, informações ou grupos de dados utilizados para descrever um determinado tipo de acervo. Caso existam relacionamentos, estes também devem constar do modelo de padronização. Geralmente, a definição de padrões de metadados é feita por um grupo de pessoas onde, entre seus componentes, se encontram usuários que detêm o conhecimento sobre um determinado tipo de material, como bibliotecários ou profissionais de ciência da informação junto com profissionais de informática.

O grau de complexidade de um padrão de metadados pode ser elevado quando no ambiente a ser descrito existe uma grande diversidade de informações

manipuladas, cada qual com características diferentes e que devem ser integradas de forma a se obter um modelo de padronização coerente.

Apesar da complexidade de alguns padrões de metadados, o conjunto de descritores deve conter apenas informações apropriadas e suficientes para descrever o dado de forma que a informação nele contida além de ser compreendida por qualquer pessoa, possa também ser compilada/interpretada pelo computador, pois pode servir de subsídio a sistemas de busca e recuperação de informações (Yearger e Mcgrath, 1996).

Atualmente, existem vários padrões de metadados utilizados nos LMS. Dentre os padrões destacam-se o LOM (*Learning Object Metadata*), do *Institute of Electrical and Electronics Engineers* (*IEEE*), *Learning Technology Standards Committee* (LTSC), o IMS (*Information Management Systems*), do *IMS Global Learning Consortium*, e o SCORM (*Sharable Content Object Reference Model*), da *Advanced Distributed Learning* (ADL) (Pereira, 2003a). Além destes padrões, encontra-se na literatura o Dublin Core (DC, 2001) e o ARIADNE (Alliance of Remote Instructional Authoring and Distribution Networks for Europe) (Ariadne, 2005).

Nos Sistemas de Bibliotecas Digitais, também existem vários padrões de metadados, tais como o MARC – *Machine Readable Cataloging* (LC, 2005), desenvolvido pela LC (*Library of Congress* – Biblioteca do Congresso Americano) e utilizado na maioria dos Sistemas de Bibliotecas, pois é um padrão de registros bibliográficos, o METS (*Metadata Encoding and Transmission Standard)*, que é uma iniciativa da DLF (*Digital Library Federation*), e o Dublin Core, utilizado também nos LMS.

A heterogeneidade destes padrões dificulta a interoperabilidade de recursos e, conseqüentemente, o compartilhamento de informações. Por isto, na arquitetura que propomos terá que ser tratada esta heterogeneidade para integrar os repositórios de DLs e LMSs. Para solucionar este problema, primeiro é necessário entender os conceitos associados a cada um dos elementos representados nos padrões de metadados e, a partir disso, estabelecer as possíveis relações de equivalência entre os mesmos.

Neste trabalho serão enfocados três padrões de metadados que serão utilizados no estudo de caso: o Dublin Core, que é utilizado nas duas áreas e por isto é adotado para definição dos metadados dos Objetos Digitais, o MARC, que é

utilizado no Sistema de Bibliotecas da PUC-Rio (Pergamum), onde se tem todos os metadados dos acervos da instituição, e o LOM, que é utilizado no projeto PGL (*Parthership in Global Learning*), um projeto de cooperação internacional entre instituições de ensino e pesquisa para promover a educação baseada na *Web*, do qual a PUC-Rio, através do Laboratório de Banco de Dados (TecBD) do Departamento de Informática, faz parte. Mais detalhes destes padrões são apresentados nos apêndices A, B e C.

2.2.2.2. Ontologia

Atualmente, há um grande interesse no desenvolvimento de ontologias para facilitar o compartilhamento de informações.

Segundo Gruber (1995), "Uma ontologia é uma especificação de uma conceituação, isto é, uma descrição de conceitos e relações que existem em um domínio de interesse". Basicamente, uma ontologia consiste desses conceitos e relações, enquanto suas definições, propriedades e restrições são descritas na forma de axiomas. Esta definição serve apenas para especificar um conjunto de conceitos. Do ponto de vista adotado neste trabalho, o importante em uma ontologia é sua aplicação. A construção de ontologias comuns tem sido proposta como abordagem promissora para a interoperabilidade de sistemas (Guarino, 1998).

Ontologias podem ser planejadas para facilitar a reutilização de bibliotecas de objetos e para modelagem de problemas e domínios. A última meta desta técnica é a construção de uma biblioteca de ontologias que podem ser reutilizadas e adaptadas para diferentes classes comuns de problemas e ambientes.

Em uma arquitetura que permite processamento inteligente de consultas num Sistema de Banco de Dados Distribuído, a informação disponível nos diferentes repositórios deveria ser descrita por visões semânticas. De fato, cada repositório de dados deveria ser descrito por pelo menos uma visão semântica. Assim, a heterogeneidade e distribuição existente entre os repositórios de dados no Sistema de Informações Integrador estariam escondidas dos usuários, que poderiam trabalhar com um número pequeno de visões semânticas.

Ontologias têm sido aceitas como poderosas ferramentas de descrição de conceitos e, por esta razão, são muito apropriadas para representar o papel de visões semânticas.

Ontologias podem constituir um modelo semântico sobre os repositórios de dados nelas descritos. Termos na ontologia podem representar qualquer informação, existente ou não no repositório em um momento específico, isto é, uma ontologia pode descrever, por exemplo, que há livros em um determinado repositório, embora seja possível que livros não tenham sido armazenados no repositório. Como ontologias são abstrações, elas também podem descrever qualquer tipo de formato de dados, desde textual até objetos multimídia.

Segundo Nieto (1998), é possível obter uma integração de ontologias através de uma ontologia global, onde se representa todos os repositórios que serão integrados. Pode-se também ter o procedimento com múltiplas ontologias porque o gerenciamento de ontologias globais integradas envolve problemas de administração, manutenção, eficiência e consistência dos dados, muito difíceis de se resolver. Uma ontologia muito grande também pode ser muito difícil para um usuário navegar e compreender. Também, não é real assumir que uma ontologia global simples possa descrever todos os dados disponíveis na *Web*, por exemplo. Além do mais, uma ontologia integrada global força todos os usuários a utilizar seu vocabulário. Assim, diferentes ontologias descritas utilizando diferentes vocabulários podem satisfazer as necessidades do usuário de uma maneira melhor, reduzindo problemas de consistência e eficiência.

As ontologias são ligadas por relacionamentos inter-ontologias. Esses relacionamentos podem ser utilizados para dois propósitos: o primeiro, para traduzir consultas de usuários de uma ontologia para outra, e o segundo, para indiretamente apoiar o processamento de consultas que acessariam dados descritos por múltiplas ontologias.

2.3. Mineração de Texto

Mineração de texto (*Text Mining ou Knowledge Discovery from Texts - KDT*) é o nome dado às técnicas de análise e extração de dados a partir de textos ou frases. Estas técnicas se baseiam em regularidades, padrões ou tendências dos textos de linguagem natural. Das análises, é possível descobrir o significado de palavras, o tipo da frase que está sendo analisada (afirmação, interrogação ou exclamação) e, além disso, o assunto que está sendo tratado em determinadas frases. Segundo (Tan, 1999), pode-se definir *KDT* ou *Text Mining* como sendo o processo de extrair padrões ou conhecimento, interessantes e não triviais, com base em documentos textuais.

Mineração de textos surgiu inspirada em mineração de dados (*Data Mining*), que é uma forma de descobrir padrões em bases de dados altamente estruturadas. Em um mundo onde há uma busca incessante por conhecimento, tais técnicas tornam-se uma das formas de agilizar e facilitar a seleção do que pode ser produtivo e o que pode ser descartado em textos não estruturados.

Através da análise de textos é possível a descoberta de conceitos, classificações automatizadas e sumarizações para documentos não estruturados. Trata-se de um campo multidisciplinar que envolve várias técnicas, tais como recuperação da informação, análise de texto e categorização de texto, extração da informação.

O processo de mineração de texto pode ser dividido em três etapas principais: preparação de dados textuais, processamento dos textos e pós-processamento da mineração. Na Figura 7, é importante notar que esse processo é totalmente iterativo, possibilitando o retorno a uma etapa anterior ou o recomeço de todo o processo.

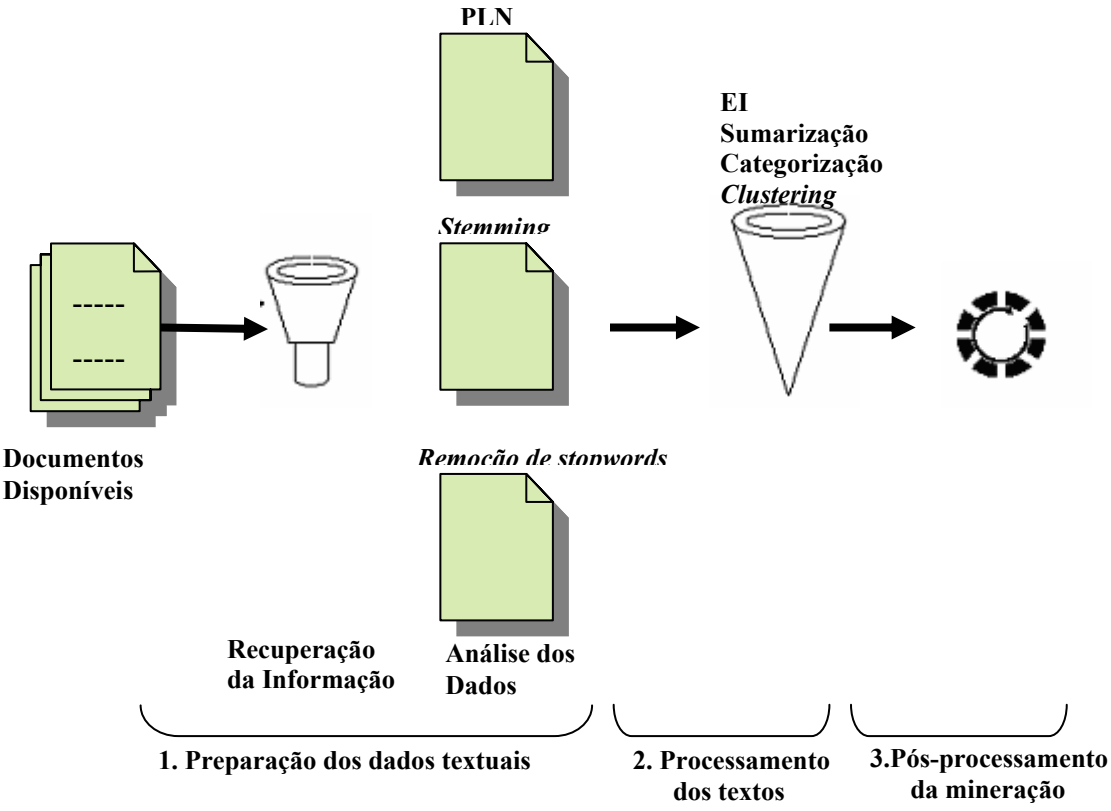

Figura 4 - Etapas do Processo de Mineração de Texto

Nas seções subseqüentes serão descritas as etapas desse processo, assim como algumas das técnicas utilizadas.

2.3.1. Preparação dos dados textuais

A preparação dos textos é a primeira etapa do processo de descoberta de conhecimento em textos ou mineração de texto. Esta etapa envolve a seleção das bases de textos que constituirão os dados de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo dos textos, ou seja, toda a informação que não refletir alguma idéia considerada importante poderá ser desprezada.

Além de promover uma redução dimensional, esta etapa tenta identificar similaridades em função da morfologia ou do significado dos termos, de modo a aglomerar suas contribuições.

Na preparação dos dados textuais a primerira técnica utilizada é a recuperação da informação e em seguida será feita a análise dos dados que serão processados. Nas seções a seguir, serão detalhadas algumas técnicas destas fases.

2.3.1.1. Recuperação da Informação

A Recuperação da Informação consiste em identificar, no conjunto de documentos de um sistema, quais atendem à necessidade de informação do usuário. O usuário de um sistema de recuperação da informação está interessado em recuperar a "informação" sobre um determinado assunto.

Segundo (Baeza,1995), Recuperação de Informação (IR – *Information Retrieval)* é um campo vasto e complexo, que engloba técnicas para construções de sistemas que vão de formas de como representar informações até como acessá-las. Essas técnicas não devem se concentrar nos dados por eles manipulados, mas sim nas informações contidas nestes dados. O grande desafio destes sistemas está em responder, da melhor forma, consultas de informações feitas por seus usuários.

As bibliotecas foram os primeiros usuários destes sistemas e, no começo, sistemas de IR não passavam de uma evolução dos catálogos de livros. Com o surgimento da *Web* e de bibliotecas digitais, este campo acabou tendo novos desafios já que, diferentemente de sistemas de bibliotecas, as informações contidas na *Web* têm um caráter dinâmico, descentralizado e não uniforme.

Basicamente, um sistema de IR é formado por uma coleção de documentos previamente catalogados, um módulo de busca e uma interface com o usuário. Na Figura 8 é apresentado um diagrama descrevendo as partes de um sistema de IR.

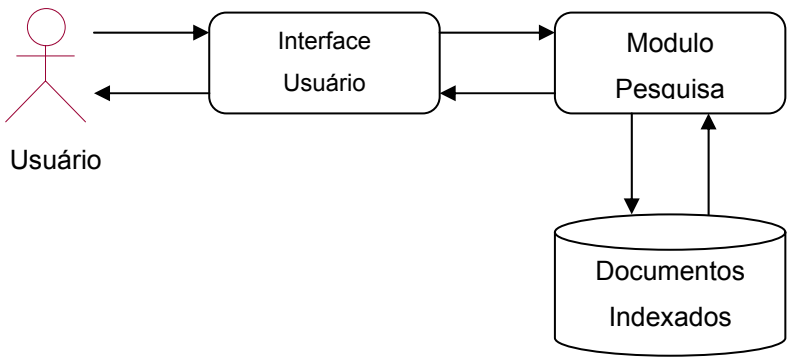

Figura 5 - Arquitetura Sistema IR

O Usuário, através de uma interface, faz a pesquisa passando como argumento o que necessita. O modulo de pesquisa acessa os documentos indexados e retorna o resultado para a interface do usuário.

O IR pode ser considerado o primeiro passo de um processo de mineração de texto.

Segundo (Burke,1999), no ambiente digital que vem se configurando nas últimas décadas, os acervos de objetos digitais se multiplicam tanto no que se refere à sua tipologia como na sua complexidade. Neste novo cenário, textos, imagens, sons, vídeos, páginas Web e diversos outros objetos digitais requerem diferentes tipos de tratamento e representação para que a recuperação da informação seja eficaz.

O processo de representação busca descrever ou identificar cada documento através de seu conteúdo. Tal representação geralmente é realizada através do processo de indexação. Durante a indexação são extraídos conceitos do documento através da análise de seu conteúdo e traduzidos em termos de uma linguagem de indexação, tais como cabeçalho de assunto, thesaurus, etc. Esta representação identifica o documento e define seus pontos de acesso para a busca e pode também ser utilizada como seu substituto.

A eficiência de um sistema de IR está diretamente ligada ao modelo que o mesmo utiliza. Um modelo, por sua vez, influencia diretamente no modo de operação do sistema.

Esta técnica, em alguns casos não é utilizada, pois dependendo do problema utilizará todo o conteúdo para fazer a mineração.

2.3.1.2. Análise dos dados

A atividade de análise dos dados visa reduzir a complexidade computacional da próxima etapa do processo (processamento dos textos) observando separadamente as palavras num documento. Isso pode ser alcançado pela eliminação de palavras consideradas irrelevantes (*stopwords*), pela identificação

de variações morfológicas (*stemming*) e pelo conhecimento do significado das palavras PLN (Processamento de Linguagem Natural).

A seguir descreveremos cada técnica da análise dos dados.

Stemming

Freqüentemente, o usuário especifica uma palavra em uma consulta, mas somente uma variante desta palavra é apresentada em um documento relevante. Plurais, formas de gerúndio e sufixos de tempo passado são exemplos de variações sintáticas que impedem uma perfeita combinação entre uma palavra da consulta e uma palavra do documento respectivo. Este problema pode ser parcialmente solucionado com a substituição de palavras pelos respectivos stems delas. Segundo (Strzalkowski,1999), Stem é o conjunto de caracteres resultante de um procedimento de stemming. Ele não necessariamente é igual à raiz lingüística, mas servirá como uma denotação mínima não ambígua do termo. O s*temming* realiza a remoção de sufixos e prefixos dos termos, escolhendo a qual radical deve ser relacionada a ocorrência. Além de reduzir o tamanho da coleção de termos do sistema, esta técnica permite que consultas realizadas por um usuário não se restrinjam apenas à forma do nome contida na consulta, sendo possível o retorno de documentos com suas variações sintáticas.

A conversão dos termos em radicais consiste na remoção do sufixo dos termos para gerar apenas os radicais de acordo com as regras gramaticais da linguagem. Isto é feito para agrupar termos que possuem o mesmo significado conceitual, como classificar, classificado, classificação e classificando.

A conversão dos termos em radicais auxilia muito o processo de filtragem e classificação de documentos, tanto na redução do número de termos diferentes existentes no vetor que representa um documento (minimizando um dos maiores problemas da classificação que é o de trabalhar com uma enorme escala de termos), assim como também na melhor definição de peso para os termos, pois se um termo aparece *n* vezes em um documento e o plural do mesmo termo ocorre mais *m* vezes, sem a radicalização existiriam duas entradas no vetor, mas utilizando a radicalização o termo só apareceria uma vez no vetor e com *n*+*m* ocorrências.

Um dos algoritmos mais citados na literatura é o *Porter Stemming Algorithm* (Porter, 1980). Este é um algoritmo simples e muito eficiente para a radicalização de termos. O algoritmo é executado em cinco passos, sendo que cada passo realiza uma transformação sobre o termo alvo. Cada passo é formado por um conjunto de regras do tipo: Se um termo *t* possui mais do que *s* sílabas e termina com o sufixo *SUFIX*, o sufixo *SUFIX* é substituído por *SUF*.

O *Porter Stemming Algorithm* foi definido para a formação de radicais da língua inglesa. Para a língua portuguesa é necessário considerar as devidas adaptações.

Após o pré-processamento, o texto estará transformado em um conjunto de termos que são os termos que realmente podem identificar a categoria do texto.

Remoção de *Stop Words*

A remoção de *Stop Words* é o processo que elimina palavras consideradas irrelevantes, tais como artigos, pronomes, interjeições, advérbios, preposições, etc. Essas palavras, normalmente, são eliminadas porque não traduzem a essência do texto e, por isso têm baixo valor semântico, sendo conhecidas também por palavras negativas. O benefício da eliminação de *stopwords* num documento está na redução do seu tamanho. Segundo estudos de Yates-Baeza e Ribeiro-Neto (1999), eles mostram que com a eliminação de *stopwords*, os documentos podem diminuir em até 40% do seu tamanho.

Segundo Frakes e Beaza-Yates (1992) é possível encontrar uma lista contendo 425 *stopwords*. Embora, na literatura disponível, o maior número de *stoplist* (lista de *stopwords*) é no idioma inglês, em http://www.cin.ufpe.br/~compint/aulas-IAS/ias/stoplist/portugues.txt é possível obter uma lista com mais de 400 *stopwords* no idioma português.

2.3.2. Processamento dos textos

Após o pré-processamento, é feito o processamento do texto, nesta etapa destacam-se extração da informação, *clustering*, sumarização e categorização de texto. Cada uma dessas tarefas extrai um tipo diferente de informação dos textos e é indicada para resolução de problemas distintos. Por isto, os objetivos do

processo de mineração de texto devem ser definidos para que seja escolhida a tarefa mais adequada. A seguir, será apresentada uma breve explicação da extração da informação, pois será a técnica utilizada neste trabalho.

2.3.2.1. Extração da Informação

A busca do conhecimento a partir de bases de dados textuais não estruturadas exige a compreensão de dados armazenados. Um leitor adquire o conhecimento do texto, fácil e naturalmente, identificando a informação relevante e memorizando-a. Automatizar esta atividade é tão complexo como construir um sistema para compreender a linguagem natural (Moulin, 1992). Para evitar este complexo processamento, o processo de Extração de Informação (EI) simula o trabalho do leitor, compreendendo o conteúdo de um texto sem a manipulação profunda das características semânticas do texto.

Tipicamente, a EI envolve a identificação de padrões que representam um contexto chave dentro do texto. Além disso, a EI utiliza um conjunto de filtros que, junto com os padrões, irão representar, de forma estruturada, a informação contida em cada texto, possibilitando a atualização de uma base de dados ou a melhora de uma recuperação de informações posterior (Jacobs e Rau, 1993).

Do ponto de vista das técnicas utilizadas, a EI pode ser vista como qualquer método que filtre informações de um grande volume de texto. Grishman (1997) estreita a definição para "a identificação de instâncias de uma classe particular de eventos ou relacionamentos num texto em linguagem natural, e a extração de argumentos relevantes do evento ou do relacionamento". Então, Grishman conclui que a EI envolve a criação de uma representação estruturada da informação selecionada e extraída do texto.

EI visa à redução do tempo de processamento e à obtenção de informações de melhor qualidade. Esta tarefa isola os fragmentos relevantes do texto, extrai informação relevante dos fragmentos e une então esta informação em uma estrutura coerente (base de dados resultante). Os sistemas de EI não tentam interpretar o texto, mas analisam as partes destes que contêm a informação relevante (Lehnert, 1996). A relevância é determinada por um conjunto de regras pré-definidas que devem especificar, tão exatamente quanto possível, que tipo da informação o sistema espera encontrar. Estas regras podem ser definidas com base

em estudo exploratório do texto ou através de algoritmos de aprendizagem de máquina.

2.3.3. Pós-processamento da Mineração

O pós-processamento dos dados consiste da fase de validação das descobertas efetuadas pela etapa de processamento dos dados e da visualização dos resultados encontrados.

Algumas métricas de avaliação de resultados, ferramentas de visualização e conhecimento de especialistas ajudam a consolidar os resultados. Uma delas é a precisão (*precision*), que avalia o quanto o sistema acerta. A expressão de cálculo da precisão é a seguinte:

$$\text{Pr}\,ecisão\,(\text{P}) = \frac{\text{Número de itens relevantes recuperados}}{\text{Número total de itens recuperado pelo sistema}}$$

Onde o número de itens relevantes recuperados equivale ao número de itens que o sistema classificou de forma consistente; enquanto o número total de itens recuperados corresponde ao número total de itens fornecidos pelo sistema.

A métrica da precisão é baseada na noção de itens classificados corretamente. Na literatura ainda não foi definido um valor mínimo que determine se um sistema obteve uma alta ou baixa precisão. Essa avaliação ainda é realizada de forma subjetiva e, geralmente, junto com o(s) especialista(s) do domínio, pois a participação deste(s) nesta etapa é também muito importante, pois ajuda a verificar o conhecimento extraído do processo e resolver situações de conflito. Nesse caso, o sistema necessita atingir uma taxa de precisão aceitável pelos especialistas.

A outra é a abrangência, que avalia o quanto o sistema traz de lixo em seu resultado. O cálculo é o seguinte:

$$\text{Abrangência}\,(\text{A}) = \frac{\text{Número de itens certos recuperados}}{\text{Número total de itens certos}}$$

Onde o "Número de itens certos recuperados" equivale ao número de itens que o sistema recuperou de forma consistente; e o "Número total de itens certos" corresponde ao número de itens certos existentes no texto.

A medida final que será utilizada para verificar qual o melhor sistema é F, que é uma medida padronizada para calcular a composição da Precisão e Abrangência, dada pela seguinte expressão:

$$F = \frac{2 * P * A}{(P + A)}$$

O sistema que obtiver o maior F é o melhor sistema.

Neste capítulo foram apresentadas explicações de algumas técnicas e alguns conceitos utilizados na tese, para facilitar um melhor entendimento da mesma. No próximo capítulo será apresentada a comparação entre os ambientes e as operações que devem ser feitas na Biblioteca Digital para alcançar uma integração mais útil.