3
Preparação de DL para integração

O objetivo deste capítulo é, inicialmente, apresentar a extensão da arquitetura do ambiente de DL, para torná-la similar à arquitetura do Ambiente de Aprendizagem, de maneira a facilitar a integração dos mesmos. Em seguida, apresenta-se a extração de informação dos repositórios das DLs, para que se tenha o enfoque de reuso ou reutilização da informação compatível com os Ambientes de Aprendizagem, preparando, assim, o ambiente de DL para a integração.

3.1. Extensão da arquitetura do ambiente de DL

Na arquitetura do ambiente de DL, vista no Capítulo 2, o DLMS gerencia tanto os serviços, como os conteúdos dos documentos. Entretanto, como um dos objetivos deste trabalho é a integração dos repositórios do Ambiente de Aprendizagem e o de DL, será feita uma extensão da arquitetura do ambiente de DL, para que esta seja mais eficaz.

Fazendo uma analogia com a arquitetura do Ambiente de Aprendizagem, apresentada no Capítulo 2, foi observado que os DDs deveriam ser gerenciados por algum *software* específico, que garantisse as mesmas características de um Sistema Gerenciador de Banco de Dados, ou de um LCMS. Além disto, observando a necessidade de reuso e compartilhamento de material, seria interessante desenvolver uma estratégia similar em ambientes de DLs que contribuíssem para uma maior utilização dos conteúdos disponibilizados em meio digital. Deste modo, observou-se a necessidade de criar um "Sistema de Gerência de Objetos de Bibliotecas Digitais" (DLOMS), onde será possível extrair dos DDs, que são monolíticos, as informações para que os mesmos se tornem reutilizáveis, segundo o enfoque de aprendizagem. Com base neste conceito, define-se também, nesta tese, "Objetos de Bibliotecas Digitais" (DLOs), que serão os conteúdos reutilizáveis que compõem a DL. A Figura 9 (**B1.1**) apresenta a nova arquitetura modificada definida para a DL.

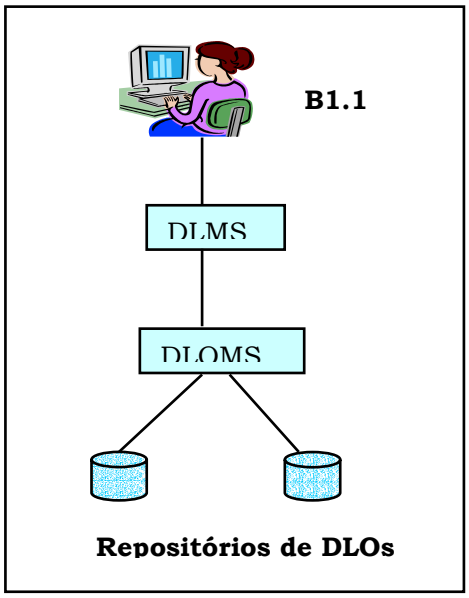

Figura 6- Arquitetura Modificada do Ambiente de DL

Onde o DLMS gerencia todos os serviços da DL, enquanto o DLOMS gerencia e trata todos os DLOs.

Os DLOs são extraídos dos DDs e esta operação é descrita na próxima seção.

3.2. Extração de informação dos DDs

A transformação dos DDs em DLOs implica em analisar estes documentos e extrair as informações de interesse, tais como conceitos e definições. A estratégia de como extrair estas informações poderia ser guiada pela organização destes documentos. Com base em experiência de trabalho na área de sistemas de bibliotecas, estima-se que cerca de 70% das DLs são compostas principalmente por teses e dissertações digitais, de tal forma que os DDs podem ser divididos em elementos pré-textuais, textuais e pós-textuais. Neste trabalho, considera-se apenas o elemento textual dos documentos. Este poderá ser subdividido, por exemplo, em suas seções. Entretanto, observando-se que em geral as DLs são utilizadas como um meio de impulsionar a aprendizagem, seria interessante adotar uma estratégia baseada na própria forma de se organizar os conteúdos de aprendizagem.

Este trabalho baseia-se na metodologia proposta pela CISCO (Cisco, 2001; Pereira, 2003) para extrair as informações destes documentos, pois as pesquisas realizadas por Merril (1996), Clark (1989), Barritt (2000), Cisco (2001) e Pereira (2003) mostraram que se trata da metodologia mais aceita e adotada pelas instituições. A seção a seguir apresenta a proposta da CISCO, para uma melhor compreensão da extração da informação do DD, ou seja, de sua transformação em um DLO.

3.2.1.Proposta da CISCO

A CISCO reconheceu a necessidade de abandonar a criação e disponibilização de conteúdos de aprendizagem compostos por blocos de treinamento grandes e inflexíveis, adotando a abordagem de LOs armazenáveis em banco de dados que podem ser reutilizados, pesquisados e modificados. A esta pesquisa a CISCO deu o nome de *Reusable Learning Object* (RLO) *Strategy* (Estratégia de Objeto de Aprendizagem Reutilizável). Essa, por sua vez, permite que a CISCO desenvolva, gerencie e entregue todo o tipo de informação e treinamento via Internet.

A CISCO adotou uma hierarquia de dois níveis para o desenvolvimento dos conteúdos reutilizáveis: RLOs e RIOs (Objetos de Informação Reutilizáveis), que foram classificados segundo os cinco tipos de informação propostos por Merril e Clark (Cisco, 2001), a saber: conceito, fato, procedimento, processo e princípio. O tamanho dos objetos depende, então, do nível de hierarquia: do menor elemento de mídia pesquisável até coleções desses elementos, que formam RIOs que, por sua vez, podem ser combinados para formarem uma lição (ou um RLO).

Com base nestes tipos de informação, observa-se que conceitos são elementos fundamentais das dissertações e teses digitais. Os elementos que estruturam os conceitos são: Introdução, Fatos, Definições, Exemplos, Contra-Exemplos, Analogias e Notas de Professor. Mais detalhes podem ser encontrados em (Cisco, 2001). Dentre estes elementos, o de maior importância é Definição.

Segundo Cisco (2001), definição é o elemento que identifica características claramente relacionadas ao conceito, enfatiza o termo que está sendo definido e identifica o conceito.

Assim, neste trabalho a palavra "definição" é utilizada neste sentido. As definições são enfatizadas porque elas representam os elementos mais importantes dos conceitos contidos nos DDs.

Neste trabalho, propõe-se a extração de definições automaticamente, sem que o professor ou aluno tenha que ler todo o documento para encontrá-la. Com isto, a busca e reutilização da informação são facilitadas e agilizadas.

A próxima seção apresenta explicações sobra as tecnologias utilizadas para extrair as definições dos documentos dos repositórios dos DLMSs e a experimentação realizada para demonstrar a viabilidade da proposta

3.2.2. Extração de Definição

Como foi visto no Capítulo 2, mineração de texto é o nome dado às técnicas de análise e extração de dados com base em textos ou frases. Como já mencionado, na etapa de processamento de textos 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

Como nesta tese a questão é extrair definições de DDs, trabalha-se com a técnica de extração de informação. Segundo Loh (1999), a estratégia mais utilizada para extrair informações no texto é analisar marcas ("*tags*") que possam indicar a presença de um certo dado. Por exemplo, o termo "anos" pode indicar que o numeral que o precede é a idade de algo ou alguém.

Grishman (1997) defende tal idéia, comentando que não adianta ter volumes grandes de textos se não há como saber que documentos específicos têm a informação necessária e onde ela se encontra dentro deles. O mesmo autor considera esta abordagem como um tipo de recuperação conceitual, em busca de informações para responder questões, ao invés de retornar documentos que contenham a informação.

Para este tipo de descoberta, utilizando estudo exploratório, o usuário pode utilizar regras gerais (formas de busca previamente definidas e comuns a vários textos) ou ele mesmo definir suas regras. Por exemplo, para encontrar o objetivo de um texto, o usuário pode procurar frases onde apareça a palavra "objetivo" (ou

seus sinônimos) ou então procurar passagens que contenham outros termos, os quais possam corresponder a um mesmo objetivo ("definiu-se", "apresenta-se", "será discutido", etc.).

Tendo como referência estas teorias, foi feito um estudo exploratório de alguns textos (Portugal, 2004; Ochi, 2004; Silva, 2004a; 2004b; Lopes, 2004; Penha, 2004; Wedemann, 2004 Barbosa, 2004), onde foram estabelecidas regras para se extrair automaticamente as definições contidas nos DDs dos DLMSs, com base em definições mais genéricas, de maneira a atender a qualquer contexto.

As regras geralmente são definidas através de uma representação comum na forma situação/ação. Um exemplo deste tipo de representação são regras do tipo “Se [condição] Então [ação]”.

As regras, na realidade, representam conhecimentos empíricos gerados com base no conhecimento do usuário. Estes conhecimentos referem-se à utilização de um grupo de características léxicas e sintáticas exploradas para a seleção das informações de entrada. Desta forma, o sistema processa a análise dos textos levando em conta o conhecimento envolvido na leitura de um texto por um ser humano.

Observa-se que, para se obter as definições encontradas nos DDs, os textos têm que ser pré-processados antes de se aplicar as regras definidas para a extração das definições, ou seja, neste trabalho foram extraídas as palavras mais importantes dos textos, denominadas VIO (Very Important Objects), em paralelo aos RIOs, pois a extração tem mais aplicabilidade e sentido quando as regras definidas são baseadas nestas palavras.

Em um primeiro momento, foi feita uma análise de vários DDs e, através de um estudo exploratório, foram estabelecidas as regras para a extração de definições.

Em seguida, as regras foram implementadas em uma aplicação Java. Como na aplicação destas regras há necessidade de se encontrar as VIOs, elas foram implementadas de maneira a selecionar as palavras cuja ocorrência no texto fosse maior que o resultado de uma função calculada sobre o número de páginas do documento.

Após a execução da aplicação Java observou-se que foram recuperadas algumas frases, as quais não eram definições. Com este resultado, foi observada a

importância de explorar a contribuição do pré-processamento dos textos para que se pudesse recuperar VIOs de maior relevância e não apenas de maior ocorrência.

A seguir são especificadas as regras para extrair as definições dos DDs, com a seguinte notação:

P1 = Primeira ocorrência de uma VIO numa frase;

Pn (n>1) = palavra consecutiva a P (n-1);

lexem = *exact match* (interação, combinação ou "casamento"perfeito);

VIO = *Very Important Object* (Palavra muito importante);

Semantic = significado da palavra, por exemplo, semântica do verbo ser: foi, são ...;

art = artigos definidos e indefindos;

pos = part-of-speech, que é parecido com a classe gramatical;

conj = conjunção;

prep = preposição.

Regras

Seja LV a lista de VIOs obtida no pré-processamento

Seja T o texto analisado contendo uma lista de frases

```
Para cada frase F do texto T faça
    Para cada VIO da lista LV faça
      SE P1 (lexem=VIO) + P2 (semantic=?ser?) + P3 (pos=?art?)
        Então
         P1 <-> {P1...P (pos="period")} * associa a palavra P1 com a seqüência de
         P1 até o próximo ponto (fim da frase F) *
         Senão
           SE P1 (lexem=VIO) + P2 (semantic=?ser?) + P3 (semantic=?definir?)
         Então
 P1 <-> {P3...P (pos="period")}
Senão
      SE P1 (lexem=VIO) + P2 (semantic=?ser?) + P3 (semantic=?explicar?)
 Então
   P1 <-> {P3...P (pos="period")}
 Senão
      SE P1 (lexem=VIO) + P2 (semantic=?ser?) +P3 (semantic=?especificar?)
   Então
   P1 <-> {P3...P (pos="period")}
      Senão
      SE P1 (lexem=VIO) + P2(semantic=?ser?) +P3 (semantic=?entender?)
   Então
    P1 <-> {P3...P (pos="period")}
      Senão
```

```
        SE P1 (lexem=VIO) + P2 (semantic=?consistir?) + P3 (pos=”prep”)
      Então
        P1 <-> {P3...P (pos="period")}
        Senão
        SE P1 (lexem=VIO) + P2 (semantic=?ser?) + P3 (semantic=?entender?)
      Então
        P1 <-> {P3...P (pos="period")}
        Senão
        SE P1 (lexem=VIO) + P2 (semantic=?ter?) + P3 (semantic=?ser?) + P4
        (semantic=?reconhecer?) + P5 (pos=”coj”)
      Então
        P1 <-> {P5...P (pos="period")}
        Senão
        SE P1(lexem=VIO) + P2 (pos=”conj”) + P3 (pos=”art”)
      Então
        P1 <-> {P3...P(pos="period")}
        Senão
        SE P1(lexem=VIO) + P2 (semantic=?consistir?) + P3 (pos=”prep”)
      Então
        P1 <-> {P3...P (pos="period")}
                        Fim-se
                      Fim-se
                    Fim-se
                  Fim-se
                Fim-se
              Fim-se
            Fim-se
          Fim-se
        Fim-se
      Fim-se
    Fim-se
  Proximo VIO de LV
Proxima F de T.
```

Posteriormente, foi utilizado o sistema Córtex, outra tese de doutorado em desenvolvimento na PUC-Rio, para pré-processamento dos textos, ou seja, para recuperar as VIOs de maior relevância e depois executar a aplicação Java com as regras que foram definidas.

O Córtex é um sistema de Reconhecimento e Classificação de Entidades (REM) baseado em modelos cognitivos, lingüísticos e estatísticos. O Córtex utiliza técnicas de Aprendizagem Automática a partir de regras primitivas para reconhecer padrões conhecidos e sugerir novos padrões, mas também inclui módulos dependentes da língua e regras adicionais para adaptar o sistema ao contexto específico.

Observa-se que com a execução do Córtex para recuperar VIOs de maior relevância e depois executarmos a aplicação em Java desenvolvida com as regras, poder-se-ia ter uma maior eficiência no resultado do experimento.

Foi realizada, então, duas experiências sobre uma amostra de 08 (oito) artigos, sendo eles heterogêneos e, portanto, complexos no sentido de possibilitar um experimento mais confiável.

No processo 1, foi executada a aplicação desenvolvida nesta tese, com as e regras sem o pré-processamento do Córtex. No processo 2, utilizou-se o Córtex para pré-processar o texto e depois foram aplicadas as regras definidas nesta tese.

Os resultados estão apresentados nas tabelas 1 e 2, onde, na Tabela 1, tem-se o resultado da análise do processo 1 e, na Tabela 2, tem-se o resultado da análise do processo 2. Nestas tabelas tem-se:

Nº – número seqüencial colocado no texto;

Referência – citação do texto;

Definições reais – definições encontradas no texto na análise humana;

Definições extraídas – definições extraídas na execução da aplicação;

Lixos – lixos extraídos na execução da aplicação, ou seja, frases que não são definições;

Definições corretas – definições corretas extraídas da aplicação dos processos.

$$P = \text{Pr}\,ecisão = \frac{N^{o}\ de\ \text{itens relevantes recuperados}}{\text{Nº total de itens recuperados pelo sistema}}$$

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

$$\text{F} = Medida\ padronizada\ compondo\ precisão\ e\ abrangência = 2 * P * {}^{A}\!/_{(P+A)}$$

Nº	Referência	Definições reais	Definições extraídas	Lixos	Definições corretas	P	A	F
1	Portugal (2004)	5	16	11	4	0,25	0,80	0,38
2	Ochi (2004)	8	16	12	4	0,25	0,50	0,33
3	Silva (2004a)	6	2	1	1	0,50	0,16	0,25
4	Lopes (2004)	0	0	0	0	0	0	0
5	Silva (2004b)	9	5	0	5	1,00	0,55	0,71
6	Penha (2004)	8	13	8	5	0,38	0,62	0,47
7	Wedemann (2004)	1	3	1	2	0,66	2,00	0,71
8	Barbosa (2004)	10	2	1	1	0	0,10	0
			Média			0,38	0,59	0,35

Tabela 1- Resultados de extração de definições do Processo 1

Nº	Referência	Definições reais	Definições extraídas	Lixos	Definições corretas	P	A	F
1	Portugal (2004)	5	7	3	4	0,57	0,80	0,66
2	Ochi (2004)	8	14	9	5	0,35	0,62	0,45
3	Silva (2004a)	6	4	2	2	0,50	0,33	0,40
4	Lopes (2004)	0	0	0	0	0	0	0
5	Silva (2004b)	9	5	0	5	1,00	0,55	0,71
6	Penha (2004)	8	9	3	5	0,62	0,58	0,62
7	Wedemann (2004)	1	1	0	1	1,00	1,00	1,00
8	Barbosa (2004)	10	0	0	0	0	0	0
			Média			0,49	0,49	0,47

Tabela 2 - Resultados de extração de definições do Processo 2

Da análise das tabelas, verifica-se que existem documentos nos quais o número de definições encontradas é muito baixo. Isto ocorre porque estes documentos contêm definições matemáticas que não foram contempladas nas regras. Também acontece que, em alguns destes documentos, a ocorrência de algumas palavras importantes é muito baixa, não permitindo a sua identificação como VIOs e, portanto, não sendo possível a extração das correspondentes definições.

Após a análise estatística apresentada nas tabelas 1 e 2, verificou-se que, como esperado, o valor da média da Precisão (P) foi maior no processo 2, ao passo que o valor da média da Abrangência (A) foi maior no processo 1. No entanto, a média da medida padrão F mostra-se mais elevada no caso do processo 2, comprovando que o pré-processamento do texto contribui para um melhor resultado na aplicação das regras de extração de definições.

Com este experimento, pôde-se mostrar a viabilidade de extrair DLOs dos DDs, fazendo com que a integração proposta neste trabalho se torne mais eficaz.