BDTC - Uma Biblioteca Digital de Trabalhos Científicos com Serviços Integrados

Cristiano R. Cervi [1]
Edimar Manica [2]
Carina F. Dorneles [3]
Renata Galante [2]

Resumo: Este artigo apresenta a BDTC, uma biblioteca digital de trabalhos científicos. A BDTC é composta por um conjunto de serviços que busca atender a algumas necessidades dos usuários. O artigo foca a arquitetura da BDTC, mostrando especificamente os serviços de autoarquivamento de documentos, extração de metadados de arquivos PDF e uso de índices que fornecem suporte à busca por similaridade.

Palavras-chave: Bibliotecas digitais. Autoarquivamento. Extração de dados. Indexação por similaridade.

Abstract: *This paper presents BDTC, a digital library for scientific work. BDTC is composed of a set of services that try to attend some user needs. This paper focuses on the BDTC architecture, showing specifically the following services: self-archiving of documents, metadata extraction from PDF files and the use of indexes that provide support to similarity search.*

Keywords: *Digital libraries. Self-archiving. Data extraction. Similarity indexing.*

1 Introdução

Atualmente, pesquisadores e acadêmicos têm se beneficiado muito com o crescimento acelerado das tecnologias web, pois os resultados de pesquisa podem ser publicados e acessados eletronicamente tão logo tenham sido realizados. Esta possibilidade é vantajosa na medida em que minimiza as barreiras de tempo e espaço associadas à publicação tradicional. Nesse contexto surgem as bibliotecas digitais como repositórios de dados que, além dos documentos digitais propriamente ditos, ou de apontadores para esses documentos, armazenam os metadados associados, muito úteis em processos de indexação e sistemas de busca.

Uma biblioteca digital pode ser definida como uma coleção organizada de objetos digitais associada a um conjunto de serviços acessíveis em ambientes distribuídos, com o objetivo de atender às necessidades de comunidades de usuários [10]. Os objetos digitais podem ser criados digitalmente ou, ainda, convertidos para o formato digital. Os serviços oferecidos podem incluir desde mecanismos simples de navegação e consulta até complexos sistemas de recomendação e personalização. As comunidades de usuários é que possibilitam delinear os objetivos da biblioteca, uma vez que deve estar alinhada às necessidades dos usuários que a utilizam.

As características que diferenciam uma biblioteca digital de outra podem estar associadas aos serviços que são disponibilizados por tais sistemas [17]. Esses serviços são implementados de acordo com os objetivos da biblioteca e das tecnologias utilizadas para seu desenvolvimento. Serviços de busca são oferecidos para facilitar a consulta do usuário ao conteúdo da biblioteca digital, ao passo que interfaces amigáveis de navegação oportunizam uma fácil manipulação dos objetos digitais, permitindo ao usuário localizar, visualizar e acessar conteúdos.

[1]Curso de Ciência da Computação, UPF, Campus 1 - BR 285 - Passo Fundo (RS) - Brasil
{cervi@upf.br}
[2]Instituto de Informática - II, UFRGS - Porto Alegre(RS) - Brasil
{edimar.manica, galante@inf.ufrgs.br}
[3]Departamento de Informática e Estatística - INE, UFSC - Florianópolis (SC) - Brasil
{dorneles@gmail.com}

doi: 10.5335/rbca.2009.007

Outros tipos de serviços podem estar associados a bibliotecas digitais, como, por exemplo, autoarquivamento de conteúdo, extração de metadados e busca por similaridade. O auto-arquivamento refere-se à possibilidade de o próprio autor de um objeto digital, ou um mecanismo de sistema, realizar o trabalho de inserção dos objetos na biblioteca digital. Isso agiliza o processo de cadastramento e auxilia, pelo menos em grande parte, o trabalho de um administrador do sistema. Extração de metadados refere-se ao processo de extrair dados importantes a respeito dos documentos armazenados nas bibliotecas digitais e que, posteriormente, podem ser usados para diferentes fins, tais como indexação de termos e busca de objetos digitais. Consultas a bases de dados de bibliotecas digitais podem se tornar extremamente massantes e cansativas ao usuário, pois ele deveria conhecer todas as diferentes representações ligadas ao nome de um autor. Por exemplo, o Google Scholar retorna diferentes documentos para as consultas "Willian W. Cohen" e "Willian Cohen". Neste caso, um módulo que forneça suporte à consulta por similaridade pode ser de grande importância num sistema de biblioteca digital.

Este artigo apresenta uma proposta de uma biblioteca digital de trabalhos científicos, a BDTC, que faz uso de serviços que têm o objetivo de prover suporte aos três pontos considerados anteriormente: autoarquivamento de conteúdo, extração de metadados e busca por similaridade. Os serviços são utilizados para que os conteúdos possam ser disponibilizados pelos próprios autores dos documentos digitais e a indexação destes objetos seja realizada por um mecanismo com suporte à similaridade.

O artigo está organizado como segue. Na Seção 2 são apresentadas a arquitetura da BDTC e uma visão geral dos serviços oferecidos; na Seção 3, os serviços de autoarquivamento, extração de metadados, indexação por similaridade e a interface de busca a objetos digitais são detalhados. A Seção 4 apresenta brevemente a interface de consulta utilizada na BDTC. Na Seção 5, são descritos os trabalhos relacionados. Finalmente, a Seção 6 apresenta as conclusões, bem como são esboçados trabalhos futuros.

2 Arquitetura da BDTC

A Figura 1 apresenta a arquitetura atual da BDTC. De forma geral, além da disponibilização de uma interface amigável ao usuário, a BDTC dispõe de quatro serviços:

- autoarquivamento: serviço que possibilita ao próprio autor enviar seu trabalho para publicação, sem intermédio de terceiros;
- consulta de objetos digitais: torna possível a consulta de documentos armazenados na BDTC;
- extração de metadados: executa a extração de metadados a partir de documentos não estruturados, especificamente documentos PDF (Portable Document Format); e
- indexação por similaridade: serviço que utiliza índices que fornecem suporte à busca por similaridade a termos da base de dados.

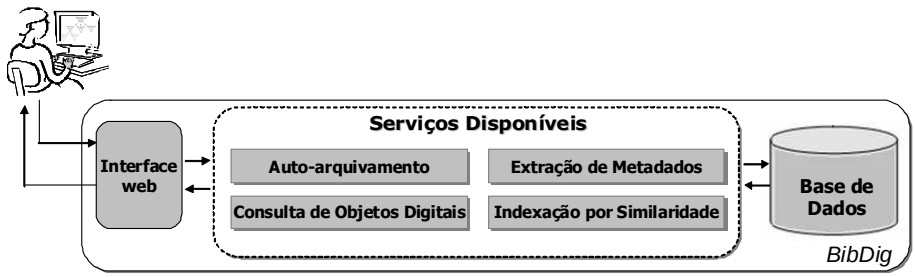


Figura 1. Arquitetura da BDTC com os serviços atualmente disponíveis

Basicamente, a arquitetura da BDTC refere-se à interface de apresentação ao usuário, aos serviços disponibilizados e à base de dados com os conteúdos digitais. As interfaces envolvem dois ambientes: o ambiente público, acessado por qualquer usuário, e o ambiente de administração, ao qual apenas os administradores da biblioteca têm acesso. É no ambiente do administrador que são realizadas operações de correções, quando algum conteúdo apresentar qualquer tipo de problema.

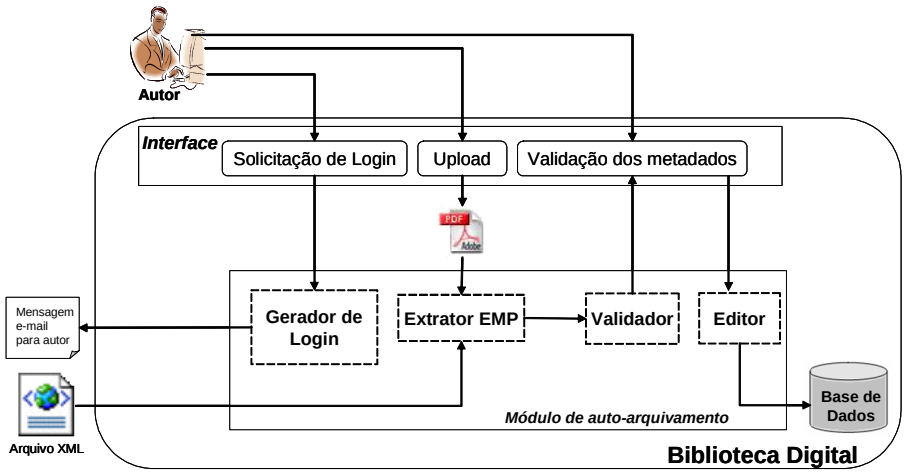


Figura 2. Seqüência do processo de submissão, validação e publicação

Os objetos digitais da biblioteca são armazenados no banco de dados PostgreSQL[4]. A base de dados foi projetada para ser flexível com relação aos tipos de dados suportados. Atualmente somente arquivos texto são passíveis de armazenamento, mas nada impede que outros formatos, tais como áudio, vídeo ou imagens possam ser utilizados. Para isso, apenas a camada de aplicação deve sofrer adaptações, uma vez que o banco de dados já está preparado para receber qualquer tipo de dado multimídia.

3 Serviços disponíveis

Nesta seção, é apresentado o funcionamento dos serviços disponibilizados pela BDTC, pela implementação dos principais módulos: o autoarquivamento, a extração de metadados e a indexação por similaridade.

3.1 Autoarquivamento

A metodologia de autoarquivamento é dividida em quatro etapas: (i) autenticação do autor e submissão do trabalho; (ii) extração de metadados; (iii) validação dos metadados; (iv) publicação dos documentos. De forma geral, a submissão dos trabalhos é feita pela criação de uma conta temporária de usuário, que permite apenas a submissão do trabalho pelo autor, sendo posteriormente excluída do sistema. O processo de extração e validação dos metadados faz com que o documento submetido pelo autor seja processado de forma a extrair os metadados automaticamente, sem a necessidade de serem digitados por eles, ou por terceiros. A tarefa de validação dos metadados é opcional e consta da intervenção do autor na confirmação da veracidade dos metadados extraídos. A publicação dos metadados e do documento é o processo de armazenar as informações extraídas. A Figura 2 apresenta uma visão geral da metodologia de autoarquivamento proposta. A interface possibilita três formas de interação com o usuário durante o processo de autoarquivamento: (i) solicitação de login; (ii) upload do documento; (iii) validação dos metadados extraídos do documento.

3.1.1 Autenticação do autor e submissão

O início do processo se dá pela autenticação do autor como usuário do sistema. Para isso, o autor deve acessar a biblioteca digital e solicitar uma senha através da interface de solicitação de login, informando um nome de login e seu nome. Com essas informações é possível que o sistema gere automaticamente o e-mail do usuário, através da concatenação do login informado, seguido de "@" e do domínio da instituição que possui a biblioteca digital. Em seguida, é enviada uma mensagem para o endereço de e-mail do autor, contendo o login e a senha gerados para a utilização da biblioteca digital. Caso o endereço de e-mail não seja válido, é retornada ao sistema uma mensagem informando que o destinatário não existe. Com isso, a transação que havia sido iniciada é cancelada

[4]Banco de dados objeto-relacional de código aberto. Disponível em http://www.postgresql.org.br/

e todas as operações, descartadas. Essa validação é importante porque evita o cadastro de usuários que não sejam vinculados a uma determinada instituição, limitando o conjunto de usuários que podem acessar o sistema[5].

Tal procedimento é efetuado a fim de que o sistema gere os logins automaticamente, sem a necessidade de um administrador ter de realizar o cadastro dos autores, e se evite que pessoas mal intencionadas incluam arquivos indevidos. A utilização do domínio da instituição para a composição do e-mail é proposital, a fim de que o sistema não seja utilizado por qualquer usuário. Assim, mesmo que qualquer indivíduo informe um nome de login válido, de algum membro da instituição, não poderá acessar a conta de e-mail para adquirir a senha de acesso à biblioteca digital. De posse do login de usuário e senha, o autor pode efetuar a submissão dos trabalhos através do upload dos arquivos PDF para a biblioteca digital.

3.2 Extração de metadados

A extração dos metadados dos documentos submetidos pelo autor, especificamente documentos PDF, é feita pelo extrator EMP (Extracting Metadata from PDF file), especialmente desenvolvido para que o processo de extração possa ser realizado. O processo faz uso de um arquivo XML, que funciona como um template, onde são especificados os campos de metadados que devem ser extraídos do documento. O processo de extração é automático, não sendo necessária a intervenção do usuário em nenhum momento.

```
<!ELEMENT structure (coverSheet*, otherPages*)>
<!ELEMENT coverSheet (metadata+)>
 <!ATTLIST coverSheet page CDATA #REQUIRED>
<!ELEMENT metadata (position?, prefix?, suffix?, separator?)>
 <!ATTLIST metadata id CDATA #REQUIRED>
<!ELEMENT position  (#PCDATA)>
 <!ATTLIST position  type (general|prefix) #REQUIRED>
<!ELEMENT prefix (#PCDATA)>
 <!ATTLIST prefix type (start|all_line) #REQUIRED>
<!ELEMENT suffix (#PCDATA)>
 <!ATTLIST suffix type (new_line|same_line) #REQUIRED>
<!ELEMENT separator (#PCDATA)>
<!ELEMENT otherPages (metadata+)>
 <!ATTLIST otherPages startPage CDATA #REQUIRED
                      endPage   CDATA #REQUIRED>
```

Figura 3. Estrutura do arquivo XML

3.2.1 Template XML

Como os documentos PDF não possuem uma estrutura semântica, é necessário o uso de um documento XML, que funciona como um template, tornando possível que o módulo de extração identifique os metadados que devem ser extraídos do documento. Este documento XML torna possível que o processo de extração possa ser utilizado para qualquer documento PDF, independentemente de seu layout de apresentação. O documento XML possui a estrutura DTD (Document Type Definition) apresentada na Figura 3.

A raiz do documento XML é representada por um elemento chamado `structure`, que é composto pelos elementos `coverSheet` e `otherPages`. O elemento `coverSheet` representa a página que contém metadados que são sempre encontrados em uma única página e são representados através do elemento `metadata`. A identificação do número da página onde esses metadados se encontram é feita poe meio do atributo `page`. O elemento `otherPages` descreve um intervalo de páginas em que a ocorrência dos metadados não possui uma página específica, sendo possível sua distribuição dentro deste intervalo, definido pelos atributos `startPage` e `endPage`. No caso do elemento `otherPages`, os metadados também são representados dentro do elemento metadata.

Cada elemento `metadata` deve possuir um atributo `id`, que é usado para identificá-lo, cujo valor é o campo do metadado. Para que o metadado seja reconhecido no texto, é necessário identificar os limites de início e fim do string do metadado, que podem ser representados pela combinação dos seguintes elementos: (i) `prefix`,

[5]Esta opção pode ser adaptada durante a implementação da metodologia.

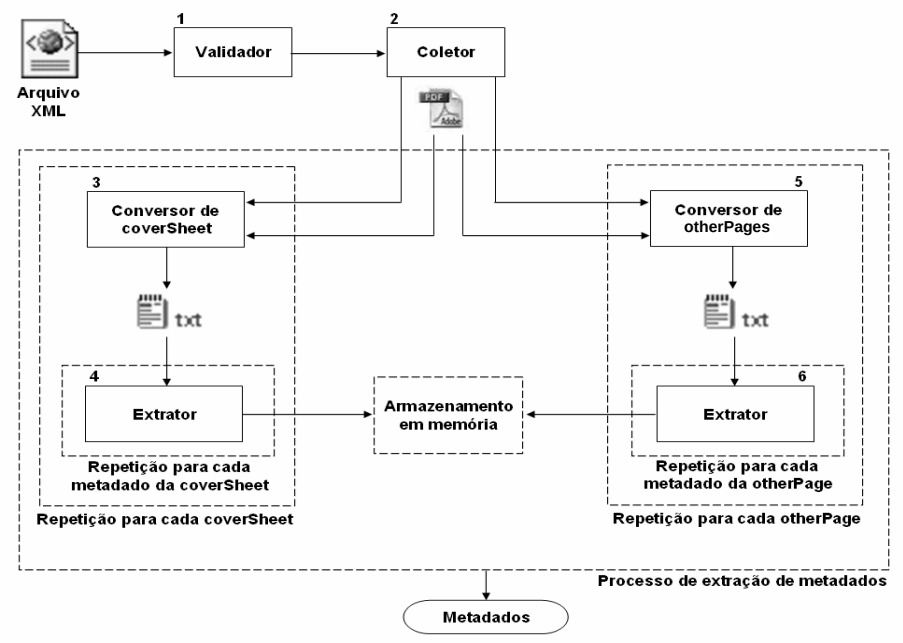


Figura 4. Processo de extração dos metadados

que define um string que sempre antecede o metadado e possui o atributo `type`, o qual identifica se ele se encontra no início da mesma linha do metadado ou se corresponde a toda a linha anterior; (ii) `position`, conforme o valor de seu atributo `type`, pode indicar a linha onde o metadado se encontra, ou qual o número de ocorrência do string que antecede o metadado, para o caso de este string se repetir na página, ou no intervalo de páginas onde se encontra o metadado em questão; (iii) `suffix`, que define uma cadeia de caracteres que sempre se encontra ao fim do metadado, e possui o atributo `type` para informar se esta cadeia de caracteres se encontra na mesma linha do metadado ou no início da linha seguinte. Quando existe um metadado composto por um conjunto de valores, tal como as palavras-chave que descrevem um documento, utiliza-se o elemento `separator` para indicar o caractere usado na separação dos valores deste conjunto.

3.2.2 Processo de extração

O processo de extração dos metadados inicia com a interpretação do arquivo XML, como apresentado na Figura 4. Inicialmente, o arquivo XML é lido (1), a fim de validar a existência de elementos que, para serem corretamente usados, necessitam de outros. Por exemplo, um elemento `position`, tendo um atributo `type` com valor `prefix`, necessita da ocorrência do elemento `prefix`, cujo valor indica o prefixo de um metadado. Em seguida, o validador envia o documento XML válido ao coletor de informações (2), que deve separar os campos de metadados em duas partes: aqueles encontrados em uma única página e aqueles encontrados num intervalo de páginas. As informações coletadas e o documento PDF servem de entrada para o extrator EMP, que extrai os metadados descritos nos elementos `coverSheet` e `otherPages`.

Para cada elemento `coverSheet` e para cada elemento `otherPages`, especificados no arquivo XML, os seguintes procedimentos são necessários: (i) transformação da página referente ao elemento `coverSheet` em um arquivo TXT (3) e conversão do intervalo de páginas referente ao elemento `otherPages` em um arquivo TXT (5); (ii) extração de metadados, a partir do arquivo TXT usado como entrada para o extrator (4 e 6), através da leitura do arquivo TXT e extração dos metadados. Em seguida, o conteúdo extraído é associado ao metadado correspondente. Depois de concluída a etapa de extração de todos os metadados contidos em todos os elementos `coverSheet` e `otherPages`, os resultados são integrados a fim de montar a resposta final que contém todos os metadados extraídos.

3.2.3 Validação dos metadados extraídos e publicação dos documentos

A fim de verificar se a extração dos metadados foi executada de forma correta, pode ser necessário que o usuário valide os metadados extraídos, indicando se os dados correspondem a valores válidos. Como exemplos dessa validação podem ser citados: (i) se data extraída corresponde a uma data válida; (ii) se ano extraído é menor ou igual ao ano atual; (iii) se o endereço eletrônico especificado para o autor é válido.

O processo de submissão, validação e publicação do documento só deve ser finalizado quando todas as operações iniciadas forem executadas. Dessa forma, se alguma delas não obtiver êxito, a transação deve ser abortada, e todas as operações são descartadas; ao contrário, constatando-se o sucesso dessas operações, a transação é efetivada.

3.3 Indexação

Como não é possível assegurar com exatidão que duas representações se referem ao mesmo objeto do mundo real, a solução é prover um mecanismo que forneça uma noção de proximidade entre os valores. Isso torna possível a construção de busca por similaridade aos objetos digitais armazenados na biblioteca digital. Além disso, para que uma busca utilizando uma função de similaridade torne-se eficiente, é necessário que exista um mecanismo de indexação dos dados armazenados na base, com suporte à similaridade. A utilização de uma tabela de índices tradicional, indexando a base apenas pelos termos existentes nas instâncias, utilizando o modelo vetorial [2] por exemplo, não é satisfatória, pois não se deseja uma comparação exata. A Figura 5 apresenta um exemplo[6] de uma consulta que não retorna nenhum resultado, pois o mecanismo de indexação utilizado faz uso dos termos exatos contidos nos documentos. O exemplo trata de um caso, extremo, em que os valores utilizados na busca não equivalem a nenhum dos termos existentes nos documentos da base de dados.

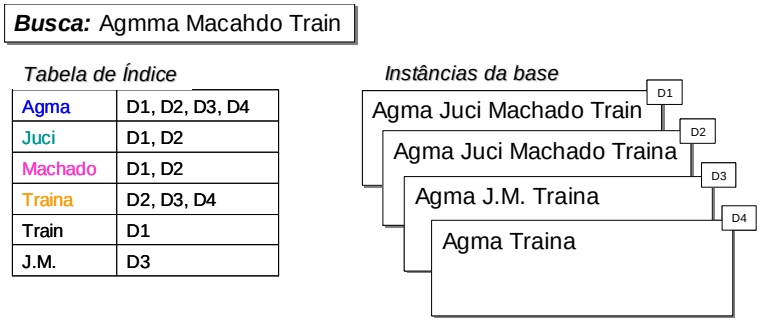


Figura 5. Exemplo de uma consulta sem retorno

3.3.1 Criação de índices

Antes da descrição da proposta, alguns conceitos e definições são apresentados a fim de tornar mais fáceis o entendimento e a descrição da abordagem.

Assim, seja d_i um termo qualquer, $D = \{d_1, ..., d_n\}$ representa o conjunto de termos de uma instância, onde $i \geq 1 \leq n$. Isso significa que os termos de uma instância são representados através de um conjunto D, onde cada termo é representado por d_i. Por exemplo, supondo a instância D_1 apresentada na Figura 5, o conjunto D_1 = {Agma, Juci, Machado, Train}. Da mesma forma, os termos de uma consulta são representados por um conjunto $C = \{c_1, ..., c_n\}$, onde c_i é um termo qualquer e $i \geq 1 \leq n$. Por exemplo, considerando uma consulta "`Agma Traina`", o conjunto C_1 = {Agma, Traina}.

Um termo em uma instância, ou em uma consulta, pode ser dividido em n-grams, que é usado para indicar um conjunto de n caracteres que ocorre sequencialmente no termo. Por exemplo, supondo o termo d_i = Juci, pode ser dividido nos seguintes 3-grams {juc, uci}. Assim, seja g_i um 3-gram qualquer, $G = \{g_1, ..., g_n\}$ representa o conjunto de 3-grams de um termo, sendo $i \geq 1 \leq n$. Por exemplo, considerando o termo d_3 = `Machado`,

[6]Os dados apresentados nos documentos são valores reais retirados do servidor da BDBComp

pertencente ao conjunto D_1, tem-se para D_1 o conjunto $G_{d3}^{D1} = \{\text{mac, ach, cha, had, ado}\}$. Da mesma forma, tendo a consulta "`Agma Traina`", o conjunto $G_{c2}^{C1} = \{\text{tra, rai, ain, ina}\}$.

Como mencionando inicialmente, a tabela de índices deve ser criada para os dados utilizados como campos de pesquisa em um sistema de biblioteca digital. Para fins de exemplificação, no decorrer do artigo são utilizados dados da biblioteca digital BDBComp[7].

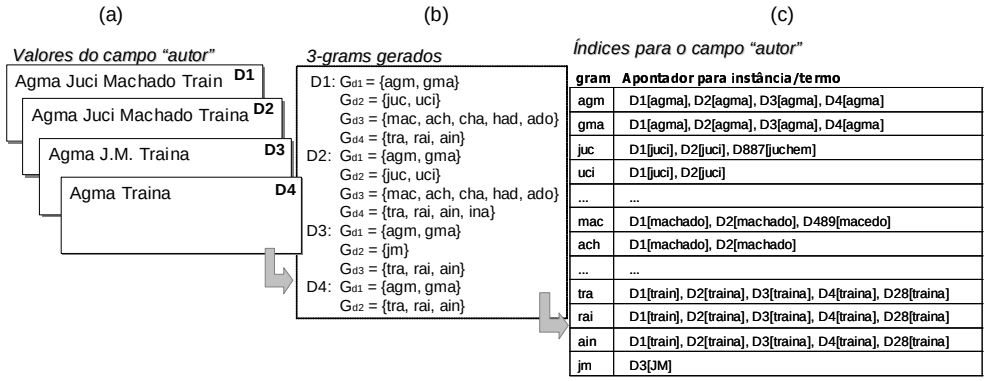


Figura 6. Índices n-grams

Para a criação da tabela de índices, devem ser executados os passos descritos a seguir.

- Pré-processamento nos dados utilizados como campos de pesquisa. Por exemplo, para que se gere uma tabela de índices para Autor, é necessário efetuar um pré-processamento de todos os dados deste campo. Para cada termo encontrado em cada instância, cria-se um conjunto G de q-grams, sendo $q = 3$. Por exemplo, considerando a Figura 6, para cada nome de autor apresentado nas instâncias descritas em (a) foram criados os conjuntos de 3-grams mostrados em (b).
- Criação da tabela de índices, com os q-grams. Os q-grams criados durante o pré-processamento das instâncias são armazenados em uma tabela juntamente com um vetor que indica quais as instâncias, e o termo em cada instância, que possui o 3-gram. Continuando no exemplo da Figura 6, para cada 3-gram criado em (b) é gerada a tabela apresentada em (c). A tabela de índices gerada possui para cada 3-gram um vetor com apontadores, que indicam quais as instâncias que possuem o 3-gram, bem como o termo, na instância, que o possui. Por exemplo, o 3-gram mac possui o vetor [D1[machado], D2[machado], D3[macedo]] associado a ele, indicando que nas instâncias D1 e D2 o termo associado a ele é machado, e que na instância D3 o termo associado a ele é macedo.

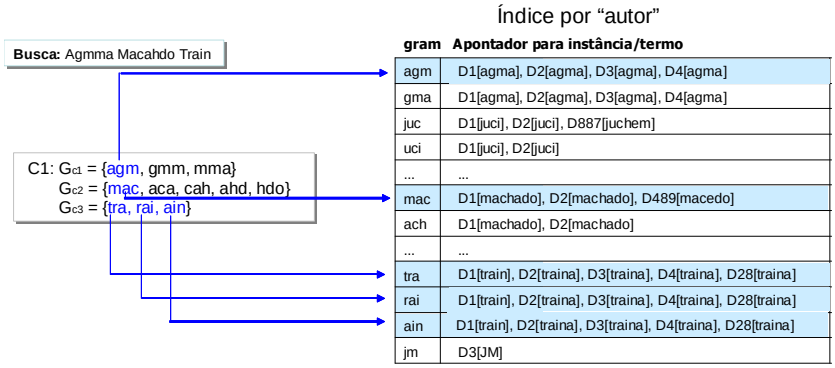


Figura 7. Instâncias utilizadas na comparação com a consulta

[7]Os dados foram obtidos através de consultas utilizando a página www.lbd.dcc.ufmg.br/bdbcomp/

As consultas também deve ter conjuntos de 3-grams criados a partir de seus termos. Considerando o exemplo da Figura 7, em que a consulta é "Agmma Macahdo Train", os conjuntos gerados para cada um dos termos são: $G_{c1} = \{\text{agm, gmm, mma}\}$, $G_{c2} = \{\text{mac, aca, cah, ahd, hdo}\}$ e $G_{c3} = \{\text{tra, rai, ain}\}$. Obviamente, o objetivo de criar os 3-grams para a consulta é que como os índices criados para as instâncias são em 3-grams a busca pelos dados deve ser feita com 3-grams também. Dessa forma, as instâncias que devem ser utilizadas para comparação são aquelas indicadas pelos 3-grams gerados para a consulta. Utilizando o exemplo apresentado na Figura 7, apenas as instâncias descritas nas linhas em destaque são usadas na comparação.

3.3.2 Processo de comparação

Instancia/termo	***C1[c1]***	***C1[c2]***	...	***C1[cn]***	***Similaridade Final***
D1[d1], D1[d2], ...,D1[dn]	*Sim(d1, c1)*	*Sim(d2, c2)*	...	*Sim(d3, cn)*	[0,1]
...	...	...	...	...	...
D28[], D28[], D28[d4]	*Sim(d1, c1)*	*Sim(d2, c2)*	...	*Sim(d3, cn)*	[0,1]

Figura 8. Estrutura auxiliar usada no cálculo da similaridade

Já que o arquivo de índices, descrito na seção anterior, possui a indicação da instância e do termo na instância em que um determinado 3-gram está localizado, não é necessário acessar os dados no banco de dados a fim de efetuar a comparação. Assim, o processo de comparação dos termos da consulta com os termos das instâncias da base de dados é efetuado com o auxílio de uma estrutura temporária. O principal objetivo desta estrutura é evitar acesso a disco cada vez que se fizer necessário efetuar o cálculo de similaridade entre os termos da consulta e os termos das instâncias. Informalmente falando, uma estrutura temporária possui a lista de termos que cada instância da base possui (identificados através da tabela de índices, pelos apontadores), o valor de similaridade entre cada termo da consulta e os termos da instância, e o valor de similaridade total entre a consulta e a instância (Figura 8). A estrutura temporária é utilizada em tempo de execução da consulta e é construída da forma descrita a seguir.

- Utilizando os valores dos apontadores da tabela de índices, cria-se uma linha para cada instância que possui os mesmos 3-grams da consulta.
- Para cada termo da consulta é criada uma coluna, cujo valor é o escore de similaridade entre o termo na consulta e o termo na instância que possui o mesmo 3-gram.
- Cada termo da consulta é comparado com o termo da instância que contém o mesmo 3-gram. A função de similaridade usada para comparar os termos é a q-Grams.
- Os valores de similaridade são utilizados para cálculo do valor final de similaridade entre a consulta e cada instância. O cálculo é feito por meio da média aritmética dos valores de similaridade entre os termos (média aritmética entre os valores, da mesma linha, das colunas dos termos da consulta). Este valor final serve como parâmetro para corte dos resultados. Somente aquelas instâncias com um valor de similaridade maior do que um limiar (threshold) são retornadas como resposta[8].

A Figura 9 apresenta um exemplo de uma estrutura temporária utilizada. A estrutura é mostrada em forma de uma tabela, onde cada tupla da tabela é composta por instância/termo, os termos da consulta e o valor de similaridade entre consulta e instância.

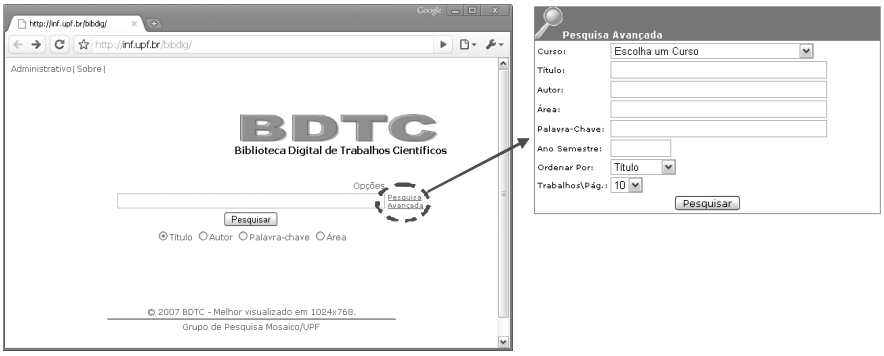
4 Interface de consulta aos objetos digitais

A interface de consulta da BDTC funciona da forma descrita a seguir. Ao entrar na página principal da biblioteca o usuário tem acesso à interface criada para realizar busca aos objetos digitais (trabalhos científicos)

[8] A definição do que seria um valor de limiar apropriado não é foco do presente trabalho. Trabalhos na literatura têm despendido muito estudo nesta questão [19]

	Termos da consulta			
Instancia	***Agmma***	***Macahdo***	***Train***	***Similaridade (soma/nr. Tokens maior string)***
D1[agma], D1[machado], D1[train]	1	0,93	0.98	0,73
D2[agma], D2[machado], D2[traina]	1	0,93	1	0,74
...	...	...	...	...
D28[traina]	0	0	1	0,33

Figura 9. Estrutura auxiliar usada no cálculo da similaridade



(a) (b)

Figura 10. Interface de acesso e pesquisa avançada

disponibilizados (Figura 10(a)). Esta página possibilita a escolha de qual atributo será usado para realizar a consulta aos arquivos. Os possíveis atributos a serem utilizados na busca são: título, autor, palavras-chave ou área. As pesquisas realizadas desta forma permitem consulta a um atributo por vez. Para que o usuário possa consulta mais de um atributo ao mesmo tempo, é disponibilizado um módulo para consulta avançada (Figura 10(b).

Figura 11. Resultados de uma pesquisa e recomendações do sistema

Os resultados das buscas efetuadas sobre a BDTC são retornados ao usuário conforme apresentado na

Figura 11. Nesta figura é possível identificar duas principais áreas: (i) dos resultados diretamente relacionados aos termos de busca usados pelo usuário; (ii) resultados que mostram trabalhos recomendados pelo sistema. Na área destinada ao resultados da busca, o usuário pode fazer o download do arquivo ou ler o resumo do trabalho. Ambas as funcionalidades estão acessíveis através de ícones. Já na área destinada à recomendação, o usuário visualiza alguns trabalhos que podem ter relação com o tema desejado. A recomendação feita pela BDTC é bastante simples e é baseada apenas nas palavras-chaves utilizadas para cadastrar o objeto digital e no nome do orientador do trabalho.

5 Trabalhos relacionados

Baseado no conceito de autoarquivamento, alguns trabalhos abordam os procedimentos para a implementação de mecanismos que possibilitam a disponibilização de documentos em bibliotecas digitais. Em [18] é apresentado um serviço que permite a diversos usuários submeterem os metadados de seus trabalhos científicos ao repositório da BDBComp [11]. Este serviço é compatível com a Open Archives Initiative [16], fazendo com que os metadados relacionados ao documento sejam coletados através do Protocol Metadata Harvesting [16]. O serviço de autoarquivamento da BDBComp não permite o armazenamento do documento digital. Isso faz com que o autor, ou até mesmo uma pessoa autorizada por ele, deva informar ao sistema um endereço eletrônico onde se encontram os objetos digitais a serem visualizados.

A submissão de documentos à Biblioteca Digital de Teses e Dissertações da USP[9] é realizada pelos próprios alunos, através do site, onde existem instruções para submissão de seus trabalhos, sendo guiados pelo sistema durante o processo de submissão [12]. O processo inclui desde a digitação dos metadados referentes ao documento, por parte dos alunos, até a complementação e verificação do cadastro a cargo dos funcionários da instituição.

A Biblioteca Digital de Teses e Dissertações da UFSCar[10], após diversas experiências no processo de submissão e publicação de documentos, adotou um sistema no qual os agentes envolvidos são apenas a biblioteca e o aluno. No processo o aluno retira uma ficha catalográfica na Biblioteca Comunitária da UFSCar para anexar à sua tese ou dissertação. Após, o bibliotecário e o aluno, conjuntamente, preenchem no sistema todos os metadados necessários. Finalmente, o documento é publicado e o processo encerrado [1].

O serviço de autoarquivamento do Eprints [11] é baseado na inserção de metadados por meio de uma interface Web, onde são definidos pelo administrador do sistema. Os usuários podem submeter o texto completo dos documentos digitais, ou, ainda, especificar um endereço eletrônico indicando sua localização. O processo de submissão só é permitido mediante o fornecimento de uma conta para o usuário [15].

A DSpace [20] realiza o processo de autoarquivamento fornecendo uma interface web aos usuários [15]. O serviço é compatível com OAI [16] e o usuário necessita de um login e uma senha para realizar a publicação dos documentos. Quando um trabalho é submetido, inicia-se um processo de revisão, realizado por revisores devidamente habilitados pelo sistema. O processo se encerra quando os revisores autorizam sua publicação. Havendo rejeição, o autor do documento é comunicado, recebendo a relação de alterações/sugestões que devem ser realizadas antes de uma nova submissão. A DSpace não permite que qualquer usuário realize a submissão de documentos, mas aqueles que estão habilitados podem visualizar os metadados, alterá-los, removê-los e até mesmo inserir novos.

O processo de autoarquivamento desenvolvido na BDTC difere dos demais por envolver como agente principal, apenas o autor do documento e sua interação com o sistema. A tarefa do usuário é simplesmente a de submeter o arquivo no formato PDF e revisar e confirmar os metadados que foram extraídos. Todos os procedimentos de extração, bem como os de publicação, são realizados de forma automática pela sistema. Isso simplifica o processo, agiliza a publicação e minimiza a ocorrência de erros, uma vez que não é o usuário quem tem a tarefa de informar os metadados referentes ao documento que será publicado.

No que diz respeito à indexação dos dados, um método que poderia ser utilizado para indexar um banco de dados de uma biblioteca digital é o espaço vetorial [9], utilizado em aplicações multidimensionais para acesso a dados multimídia. Cada característica de um objeto é mapeado para um número, que indica quanto desta ca-

[9] http://www.teses.usp.br/
[10] http://www.bdtd.ufscar.br/
[11] http://www.eprints.org

racterística está presente no objeto. A partir daí, se existirem *n* características relevantes, haverá um vetor de tamanho *n* para cada objeto. A distância entre dois objetos é dada pela distância entre os seus vetores. Muitos métodos de acesso métrico têm sido propostos com o intuito de indexar objetos embutidos em um espaço métrico [6, 21, 8, 4, 22, 5]. Entretanto, esta abordagem não pode ser usada para campos textuais, pois não é possível mapear texto para números significativos.

O trabalho desenvolvido em [13] propõe a aplicação de uma função "torcida" (twisting function) sobre o espaço métrico para gerar um novo espaço, o "espaço torcido" (twisted space). Neste novo espaço, os objetos são arranjados de forma diferente, possibilitando um agrupamento mais apropriado para valores textuais. No entanto, este novo espaço não pode ser utilizado diretamente em consultas, sendo necessária a combinação do novo espaço com o espaço métrico.

Um proposta que se assemelha à descrita neste artigo é a apresentada em [7], onde é proposto o "índice tolerante a erro" (Error Tolerant Index (ETI)). A proposta é descrita no contexto de um data warehouse, e o principal objetivo é criar uma tabela de índices com 3-grams apontando para todos as relações que possuem aquele gram. Outros métodos de indexação [3, 14] também efetuam o pré-processamento da base de documentos para construir tabelas com q-grams indicando a localização dos documentos que possuem os termos com aqueles *grams*. O diferencial do trabalho proposto aqui é a utilização de uma estrutura auxiliar que minimiza o número de comparações e o acesso a disco.

6 Conclusões

Este artigo apresentou uma biblioteca digital de trabalhos científicos, a BDTC, que faz uso de serviços como autoarquivamento de conteúdo, extração de metadados dos objetos digitais e busca por similaridade. As bibliotecas digitais que se utilizam de um mecanismo de autoarquivamento têm reduzido o tempo de submissão e publicação dos documentos, uma vez que os únicos agentes envolvidos no processo são o próprio autor do documento e o sistema.

Com relação ao processo de extração automática dos metadados dos objetos digitais, a ocorrência de erros é minimizada, ficando a critério do autor realizar apenas a confirmação dos metadados extraídos. Assim, automatiza-se o processo de extração, pois todo o trabalho fica a cargo de um mecanismo automático, que segue apenas um template no formato XML.

O mecanismo de indexação por similaridade proposto é simples, útil e facilmente aplicável em qualquer sistema de busca por similaridade. A principal contribuição do trabalho está na definição de uma estrutura auxiliar que facilita a comparação da consulta com as instâncias da base, diminuindo o tempo de execução, pois não necessita acesso a base de dados.

Como trabalhos futuros pretende-se implementar três serviços: (i) um mecanismo que forneça interoperabilidade entre bibliotecas digitais, desenvolvido obedecendo ao padrão OAI (Open Archive Initiative) através de um protocolo denominado PMH (Protocol Metadata Harvesting); (ii) um sistema de personalização, onde usuários seriam identificados através de um perfil pessoal; (iii) um mecanismo de recomendação mais avançado que o atualmente implementado, onde, baseado no perfil do usuário, o sistema recomende trabalhos científicos baseado em seu interesse.

Referências

[1] ALMEIDA, S. M. Auto-arquivamento em Bibliotecas Digitais de Teses e Dissertações. In: SEMINáRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 2006, Salvador, Bahia. **Anais...** [S.l.: s.n.], 2006.

[2] BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval**. [S.l.]: ACM Press / Addison-Wesley, 1999.

[3] BAEZA-YATES, R.; NAVARRO, G. A practical index for text retrieval allowing errors. In: LATIN AMERICAN CONFERENCE ON INFORMATICS (CLEI), 1997, Valparaiso, Chile. **Proceedings...** [S.l.: s.n.].

[4] BOZKAYA, T.; OZSOYOGLU, M. Distance-based indexing for high-dimensional metric spaces. In: SIGMOD RECORD (ACM SPECIAL INTEREST GROUP ON MANAGEMENT OF DATA), 1997. **Anais...** [S.l.: s.n.], 1997. p. 357–368.

[5] BRIN, S. Near neighbor search in large metric spaces. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 1995, Zurich, Switzerland. **Proceedings...** [S.l.: s.n.], 1995. p. 574–584.

[6] BURKHARD, W. A.; KELLER, R. M. Some approaches to best-match file searching. **Communications of the ACM**, [S.l.], v. 16, n. 4, p. 230–236, 1973.

[7] CHAUDHURI, S. et al. Robust and efficient fuzzy match for online data cleaning. In: SIGMOD, 2003. **Proceedings...** ACM, 2003. p. 313–324.

[8] CHIUEH, T. Content-Based Image Indexing. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 20, 1994. Santiago, Chile. **Proceedings...** [S.l.: s.n.], 1994. p. 582–593.

[9] GAEDE, V.; GÜNTHER, O. Multidimensional access methods. **ACM Computing Surveys**, [S.l.], v. 30, n. 2, p. 170–231, 1998.

[10] GONÇALVES, M. A. **Streams, Structures, Spaces, Scenarios, and Societies (5S)**: a formal digital library framework and its applications. 2004. Tese (Doutorado em Ciência da Computação) - Virginia Polytechnic Institute and State University.

[11] LAENDER, A. H. F.; GONÇALVES, M. A.; ROBERTO, P. BDBComp: building a digital library for the brazilian computer science community. In: ACM/IEEE JCDL, 4., 2004, Tucson, Arizona. **Proceedings...** [S.l.: s.n.], 2004. p. 23-24.

[12] MASIERO, P. C. et al. Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo. **Ci. Inf.**, [S.l.], v. 30, n. 3, p. 34-41, 2001.

[13] NADVORNY, C. F.; HEUSER, C. A. Twisting the Metric Space to Achieve Better Metric Trees. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2004, Brasília, DF. **Anais...** [S.l.: s.n.], 2004.

[14] NAVARRO, G. et al. Indexing text with approximate q-grams. In: ANNUAL SYMPOSIUM ON COMBINATORIAL PATTERN MATCHING, 2000. **Proceedings...** LNCS 1848, 2000.

[15] NIXON, W. J. D. **Initial Experiences with EPrints and DSpace at the University of Glasgow**. [S.l.]: Ariadne, 2007. Disponível em: <http://www.ariadne.ac.uk/issue37/nixon>. Acesso em: 31 jul. 2009.

[16] OPEN Archives Initiative (OAI). Disponível em: <http://www.openarchives.org>. Acesso em: 31 jul. 2009.

[17] PEDRONETTE, D.; SILVA TORRES, R. da. Uma Plataforma de Serviços de Recomendação para Bibliotecas Digitais. In: XXIII SIMPÓSIO BRASILEIRO DE BANCO DE DADOS(SBBD), 2008, Campinas, São Paulo. **Proceedings...** [S.l.: s.n.], 2008.

[18] SILVA, L. V. **Um Serviço de Auto-arquivamento de Publicações Científicas Compatível com o Padrão OAI**. 2004. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Minas Gerais, Belo Horizonte.

[19] STASIU, R. K.; HEUSER, C. A.; SILVA, R. Estimating recall and precision for vague queries in databases. In: CONFERENCE ON ADVANCED INFORMATION SYSTEMS ENGINEERING, CAISE, 17., 2005, Porto, Portugal. **Proceedings...** Berlin:Springer-Verlag, 2005. (Lecture Notes in Computer Science).

[20] TANSLEY, R. et al. The DSpace institutional digital repository system: current functionality. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 3., 2003, Houston, Texas, Washington, DC, USA. **Anais...** IEEE Computer Society, 2003. p. 87-97.

[21] UHLMANN. Satisfying general proximity / similarity queries with metric trees. **IPL: Information Processing Letters**, [S.l.], v. 40, 1991.

[22] YIANILOS, P. N. Excluded middle vantage point forests for nearest neighbor search. In: IMPLEMENTATION CHALLENGE WORKSHOP: NEAR NEIGHBOR SEARCHES, 1999, Baltimore, Maryland. **Proceedings...** [S.l.: s.n.], 1999.