Recuperação de Padrões em Arquivos Postscript para um Sistema de Bibliotecas Digitais

Aluna: Lidiane Vogel Sander

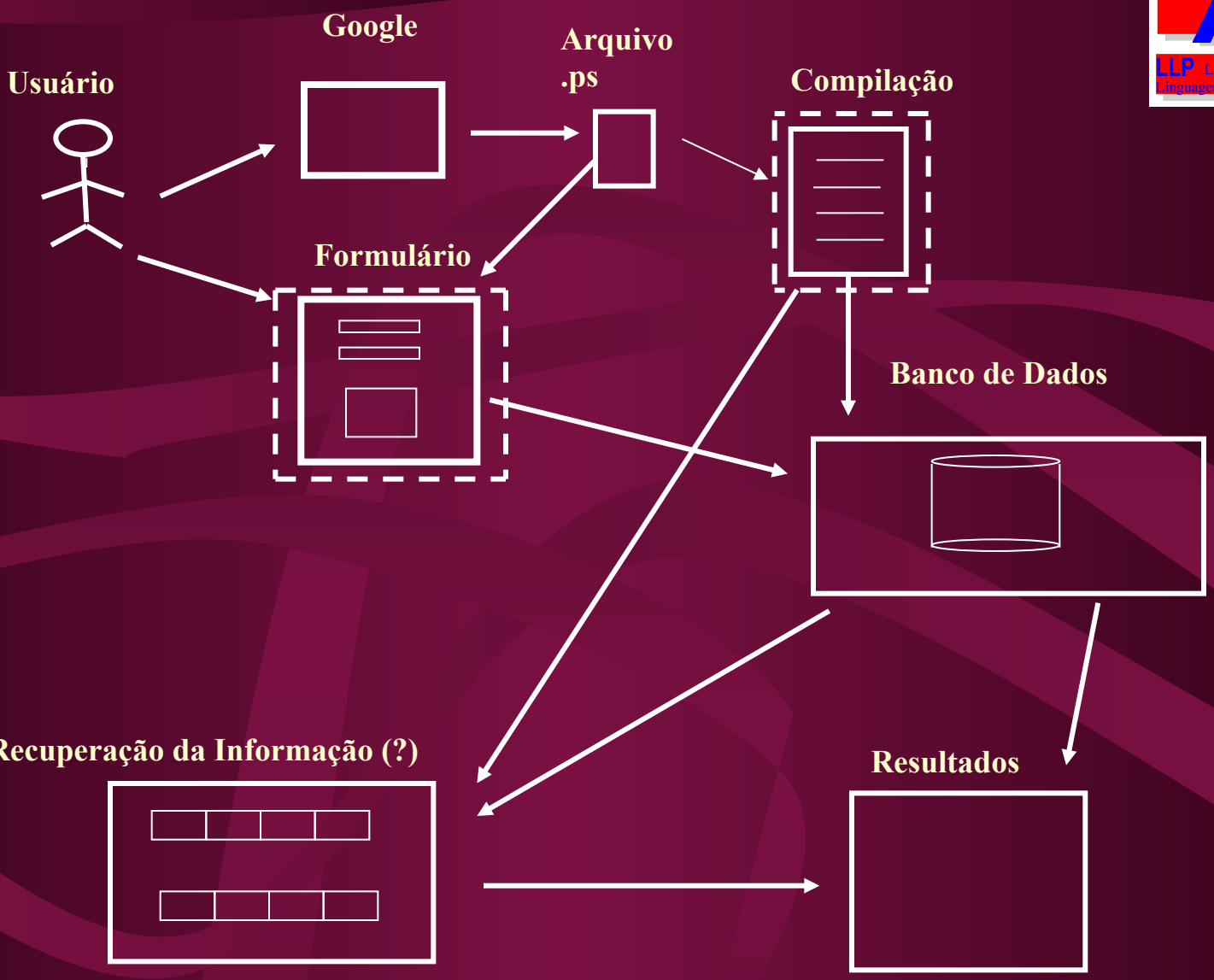
Orientadora: Mariza Andrade S. Bigonha

O Problema

- Aumento do volume de informações disponíveis → organização dessas informações.
- Tentativas existentes: bibliotecas digitais e servidores de publicações.

A Proposta

- Estudo da viabilidade de construção de um sistema web para atuar como uma biblioteca digital.
- Necessidade do laboratório SIAM – compartilhar informações comuns de maneira mais simples, organizada e eficiente.
- Possibilidade de ampliação para atender outras necessidades, em meios acadêmicos ou não, observadas as restrições da proposta.


Usuário
Google
Arquivo
.ps
Compilação
Formulário
Banco de Dados
Recuperação da Informação (?)
Resultados

Desenvolvimento

1. Comparação com sistemas já existentes:

   1.1 Koala Document Fingerprinting - Carnegie Mellon University – Sistema para identificar documentos relacionados textualmente.

   - Proposta: documento é passado ao servidor, convertido para .txt e comparado (fingerprinting) aos documentos do banco existente.
   - Ponto comum: submissão e conversão de documentos.

... Desenvolvimento

1.2 Co-Citer – Programa desenvolvido pela Cogitum para capturar textos e citações da Internet.

- Ponto comum: organização da informação – inserção de algumas informações pelo usuário e captação automática de outras.
- Diferenças: armazena informações pré-selecionadas pelo usuário; não envolve documentos como um todo (upload, download).

... Desenvolvimento

1.3 DBLP Computer Science Bibliography – University of Trier, Alemanha.

- Informações bibliográficas sobre grande parte dos periódicos e publicações em Ciência da Computação.
- 400000 artigos e milhares de links para home pages de cientistas da computação.
- Pesquisa por autor, título, assunto ou navegação manual.

... Desenvolvimento

- Diferenças:
  - Servidor de bibliografias, não é serviço de disponibilização de artigos.
  - Submissão de arquivos: via email, para o responsável pela manutenção do serviço.

... Desenvolvimento

- 1.4 BDBComp – Biblioteca Digital Brasileira de Computação
- Acesso à produção científica nacional na área de computação.
- Organização e aquisição de trabalhos científicos
- Três formas de incluir itens na biblioteca:
  - Relação de artigos publicados em eventos nacionais;
  - Cadastramento direto via formulário;
  - Integração com outras bibliotecas digitais ou repositórios.

... Desenvolvimento

- Diferenças:
  - Pesquisa por autor, título, ano, evento.
  - Voltada para produção nacional.
  - Várias formas de inclusão de documentos.
  - Disponibilização no site não-automática.

... Desenvolvimento

2. Estudo da Linguagem PostScript

   2.1 Gramática (definição BNF) PostScript para o desenvolvimento de um compilador → reconhecimento de padrões nos documentos.

– Dificuldade: encontrar definição disponível.

... Desenvolvimento

2.2 Testes com o código da linguagem: desenvolvimento de arquivos .ps "puros" e comparação com arquivos .ps "gerados" (ferramentas como o LaTeX) → complexidade/volume do código.

... Desenvolvimento

- Trecho de um arquivo .ps “gerado”

%%Page: 1 1
1 0 bop -40 -384 a Fi(Eighth)18
b(International)j(Conference)f(on)d
(Information)i(and)e(Kno)n(wledge)i(Mana
gement,)g(CIKM)e(99,)g(Kansas)g
(City)l(,)h(Missouri,)f(No)o(v)o(ember)h

... Desenvolvimento

- Trecho de um arquivo .ps “puro”

```
/Times-Roman findfont   % Get the basic font
20 scalefont            % Scale the font to 20 points
setfont                 % Make it the current font
newpath                 % Start a new path
72 72 moveto            % Lower left corner of text at
   (72, 72)
(Hello,world!) show     % Typeset "Hello, world!"
```

... Desenvolvimento

2.3 Conversão de arquivos .ps para .txt → difícil, susceptível a erros, cara.

- PStoText: aplicativo para extrair textos de arquivos PostScript e PDF.
  - Acoplado ao Ghostscript
  - Heurística para reconhecimento de palavras: strings separados por uma distância inferior à uma média específica pertencem à mesma palavra.
  - Trabalha com codificações segundo convenções da Adobe.

... Desenvolvimento

2.4 Construção do site:

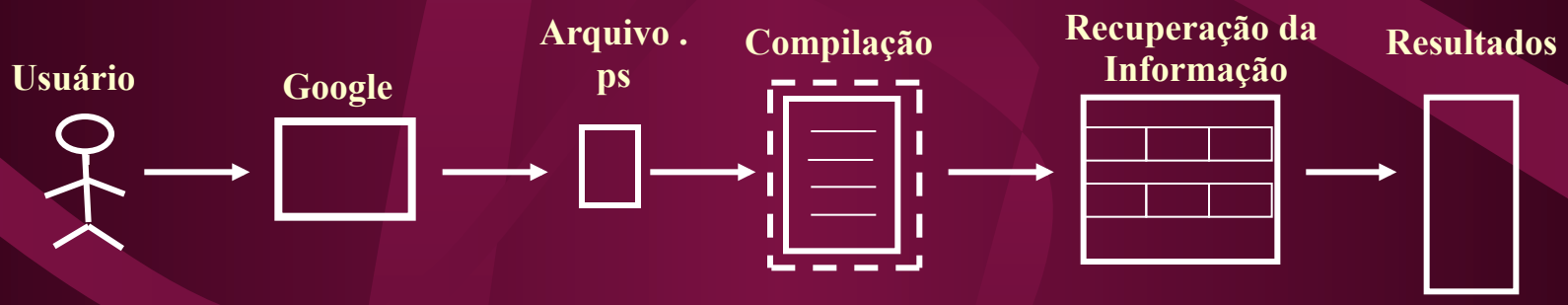
- Definição das linguagens a serem usadas: php e mySQL (Banco de dados).
- Construção dos mecanismos de inserção de arquivos.
- Download de arquivos.

... Desenvolvimento

- Utilização da ferramenta de busca FreeFind:
  - Ferramenta livre
  - Indexação (spidering) externa – percorre links e forma índice (base de dados)
  - htDig “versus” FreeFind
  - Problemas: tempo, anúncios, idioma.

Limitações da Versão

- Não utilização do Banco de Dados – informação redundante.
- Inclusão de arquivos PostScript apenas.
- Mecanismo de indexação e busca.
- Download de arquivos.



Trabalhos Futuros

- Mecanismo de busca específico
- Armazenamento em banco de dados
- Ampliação para inclusão de outros tipos de documentos.

Referências Bibliográficas

- Adobe Systems Inc. *PostScript Language Reference Manual*, Addison Wesley, 1995, 5th printing.
- Adobe Systems Inc. *PostScript Language Tutorial and Cookbook*, Addison Wesley, 1987.
- McGilton, Henry ; Mary Campione. *Postscript by Example*. Reading, MA : Addison-Wesley, c1992.
- http://www-2.cs.cmu.edu/afs/cs/user/nch/www/koala-info.html
- http://www.cogitum.com/co-tracker-text/more.shtml
- www.informatik.uni-trier.de/~ley/db
- www.cs.indiana.edu/docproject/programming/postscript/postscript.htm
- http://research.compaq.com/SRC/virtualpaper/pstotext.html