SUPeRB

Sistema Uniformizado de Pesquisa de Referências Bibliográficas

Simpósio Doutoral da Linguateca

Luís Cabral <luis.m.cabral@sintef.no>

Outubro de 2006

Sumário

- Motivação
- Alguns conceitos
- Objectivos
- Métodos e recursos utilizados
- Sistemas já disponíveis
- SUPeRB
- Conclusões

Motivação

Motivação

- Facilitar o acesso a documentação científica disponível on-line
- Melhorar um recurso para a comunidade de PLN, o catálogo de publicações da Linguateca
  - Melhorar a usabilidade na manutenção
  - Melhorar a qualidade, providenciando meios de recolha de informação

Motivação

- Disponibilizar um conjunto de ferramentas capaz de executarem determinadas tarefas vitais para processamento bibliográfico
  - Pesquisa e extracção de referências e elementos bibliográficos
  - Armazenamento da informação recolhida
  - Gestão e actualização dos dados
  - Avaliação da relevância
- Melhorar o processamento de documentação escrita na língua portuguesa

Alguns conceitos

Publicação: Documento no formato em papel ou electrónico. Pode ser um livro, revista académica, dissertação, relatório, manual técnico ou parte destes, como um capítulo de um livro ou artigo publicado numa conferência (parte das actas da conferência)

Referência bibliográfica: ou *citação*, é uma breve nota que permite identificar uma publicação. É composta por vários *elementos bibliográficos* como o nome de autor, titulo, data da publicação, local jornal ou conferência onde foi publicado, etc.

[Marrafa & Ribeiro 2001]
Palmira Marrafa & António Ribeiro) "Quantitative Evaluation of Machine Translation Systems: Sentence Level" (In *Proceedings of the MT Summit VIII: Fourth ISLE workshop* (Santiago de Compostela, 22 de Setembro de 2001), pp. 39-43. http://www.eamt.org/summitVIII/papers/marrafa.pdf

- Referência Bibliográfica
- Elementos bibliográficos

Alguns conceitos
Estilos bibliográficos

- Referências bibliográficas podem ser representadas como texto simples, na forma de linguagem natural.
- Existem inumeros estilos bibliográficos, depedendo do domínio a que pertencem ou onde são publicados.

Alguns conceitos
Exemplos

IEEE (Institute of Electrical and Electronics Engineers)

[1] P. Marrafa and A. Ribeiro, ``Quantitative evaluation of machine translation systems: Sentence level," in *Proceedings of the MT Summit VIII: Fourth ISLE workshop*, pp. 39-43, 2001.

APA (American Psychological Association)

[**Marrafa and Ribeiro, 2001**] Marrafa, P. and Ribeiro, A. (2001). Quantitative evaluation of machine translation systems: Sentence level. In *Proceedings of the MT Summit VIII: Fourth ISLE workshop*, pages 39-43.

ACM (Association of Computer Machinery)

[1] MARRAFA, P., AND RIBEIRO, A. Quantitative evaluation of machine translation systems: Sentence level. In *Proceedings of the MT Summit VIII: Fourth ISLE workshop (2001), pp. 39-43.*

Alguns conceitos
Formatos bibliográficos

- Referências podem ser representadas ou armazenadas num formato estruturado, identificando devidamente cada elemento.
  - Próprio para ser processado por programas
- Uma vez mais temos vários formatos bibliográficos:
  - BibTeX, RIS, EndNote

Alguns Conceitos
Exemplos

BibTeX

```
@inproceedings{1141121239,
    author ={Palmira Marrafa and António Ribeiro},
        title ={Quantitative Evaluation of Machine Translation Systems: Sentence Level},
        year ={2001},
        booktitle ={Proceedings of the MT Summit VIII: Fourth ISLE workshop},
        pages ={39-43},
        location ={Santiago de Compostela},
        url ={\url{ http://www.eamt.org/summitVIII/papers/marrafa.pdf}}
}
```

EndNote

```
%0 Conference Proceedings
%A Palmira Marrafa
%A António Ribeiro
%T Quantitative Evaluation of Machine Translation Systems: Sentence Level
%D 2001
%B Proceedings of the MT Summit VIII: Fourth ISLE workshop
%P 39-43
%W Santiago de Compostela
%U http://www.eamt.org/summitVIII/papers/marrafa.pdf
```


Objectivos

- Desenvolver um sistema de capaz de extrair elementos bibliográficos a partir da Web, o SUPeRB
  - Ser funcionalmente integravel com um repositório local
    - Integrar com o catálogo da Linguateca
  - Sistema modularizado
    - Capaz de processar diversas tarefas individualmente

Sistemas já existentes

Sistemas já existentes

- Repositórios bibliográficos
  - CiteSeer, DBLP ou a The Collection of Computer Science Bibliographies
- Sistemas de pesquisa bibliográfica
  - Google Scholar ou o Microsoft Live Academic Search
- Programas dedicados à manutenção de referências bibliográficas
  - BibTeX, Endnote, ...
- Sítios dedicados à gestão bibliográfica pessoal
  - CiteUlike, eprints, Connotea

Sistemas já existentes

Repositórios bibliográficos

- Dedicados normalmente a um domínio específico, podendo abordar mais do que um
  - O arxiv.org possui artigos de física, matematica, Ciência de Computadores e Biologia quantitativa
- Recolha de informação é feita através do
  - acesso a BD de conferências
  - Acesso a Bibliotecas
  - Inserção manual
  - Extracção a partir de documentos disponívies online

Sistemas já existentes

Sistemas de pesquisa

- Não tem um domínio específico.
- Recolha de informação é feita através de:
  - pesquisa em páginas Web (crawling)
  - Bibliotecas académicas
  - Repositórios bibliográficos
- Ponte entre bibliotecas académicas e utilizadores. Parece ser esta a mensagem "comercial" que tentam passar.

Sistemas já existentes

Repositórios contra Motores de pesquisa

distinguem domínios	Apresentam todos os domínios sem distinção
Disponibilizam o acesso à colecção completa (descarregar) ou através de serviços Web	Só permitem a pesquisa Web.
Fornecem o número de publicações indexadas	Não dão a conhecer nem valores nem as fontes que usam (bibliotecas ou repositórios)

Sistemas já existentes

Gestores bibliográficos (on-line)

- Repositórios de referências bibliográficas pessoais
- Fracos na area de pesquisa Web. A introdução dos dados é feita de forma
  - manual, atráves de formulários
  - Semi-automática, onde os utilizadores fornecem o URL e este é processado com a ajuda de templates
- Multi-domínio,depende das preferências dos utilizadores ou do domínio a que estes pertencem.

Sistemas já existentes

Gestores bibliográficos (on-line)

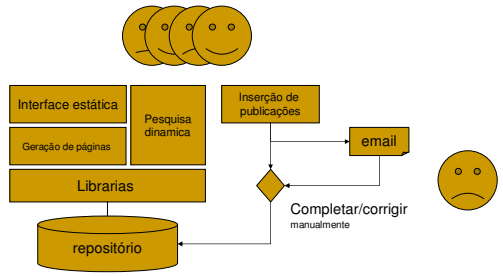
- A indexação é feita usando os elementos bibliográficos e Incluem normalmente apenas a referência bibliográfica (e o abstract)
- Indexação pode ser feita também através da classificação manual das referências
- Dado que estes gestores funcionam como uma comunidade, a classificação é partilhada gerando uma BibSonomia

SUPeRB

Sistema Uniformizado de Pesquisa de Referências Bibliográficas

SUPeRB

… O catálogo **sem** o SUPERB

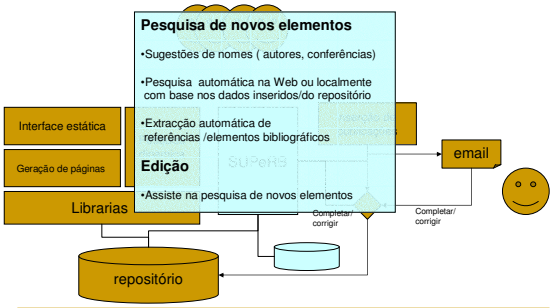


SUPeRB

Sistema Uniformizado de Pesquisa de Referências Bibliográficas

- A inserção de páginas não é fácil
  - Formulário extenso
  - Não fornece ajuda
- A verificação dos dados é ainda mais complicada
  - Não existem métodos para ajudar o gestor a verificar e completar os dados
- Não existem interfaces para a reedição dos dados
  - Necessário editar directamente para alteração dos dados

SUPeRB



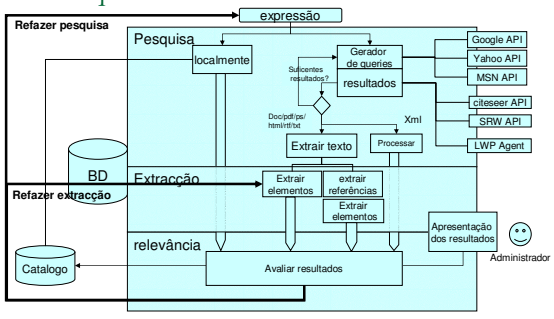
SUPeRB

Oferece

- Métodos de pesquisa na Web
- Métodos de extracção de informação bibliográficas
- Métodos de validação da informação
- Métodos e recursos para facilitar a interacção com o utilizador e o gestor
- Métodos para editar os recursos existentes

SUPeRB

...Arquitectura



SUPeRB
A pesquisa

SUPeRB
A pesquisa

- Parâmetros de entrada
  - expressão fornecida pelo utilizador
  - extraído de uma referência
    - Interface do catálogo permite consultar o SUPeRB, enviando elementos bibliográficos.
    - Posteriormente pode comparar com a referência completa
  - URL
    - Tarefa simples de descarregar e analisar o conteúdo de um URL, recolhendo

SUPeRB
A pesquisa - recursos

- A motores de pesquisa através de serviços Web
  - Baseada em motores de pesquisa bastante abrangentes
  - Uso de serviços Web
    - API avançadas
    - Informação estruturada
    - Não é necessário processar páginas Web
    - Reduz a probabilidade de erros

SUPeRB
A pesquisa

- Responsável por processar as respostas
  - Descarregar e extrair o texto dos documentos
  - Aceder a repositórios bibliográficos para descarregar meta-informação (através de WebServices)

SUPeRB
A pesquisa (resultados)

- São suficientes?
  - Pode gerar novas pesquisas para tentar encontrar resultados
- Após obter os resultados
  - Os documentos são descarregados e o texto é extraido dos seguintes formatos.
    - HTML, pdf, ps, rtf, doc e txt ☺
    - Capaz de processar ficheiros sem extensão(mime/ type)
  - URL que pertencam a repositórios bibliográficos acessíveis por serviços Web podem ser processados de forma diferente...

SUPeRB
A pesquisa (repositórios bibliográficos)

- Informação de URL que pertencam a repositórios bibliográficos pode obtida através de serviços Web
  - *Citeseer API*
  - *SRU / SRW*
- *Vantagem de ter a meta-informação devidamente estruturada*

SUPeRB
A extracção de Referências

SUPeRB
A extracção de referências

- Ferramentas como o ParaTools mostram-se incapazes de extrair correctamente as referências bibliográficas
  - Probabilidade de sucesso muito baixo
  - Só lida com documentos em Inglês
  - Limitado à extracção de documentos académicos
    - abstract, introduction,..., references

SUPeRB
A extracção de referências

- ::Citation.pm
  - Extracção das referências bibliográficas de documentos no formato académico
    - Apresenta melhores resultados que o modulo que susbtitui
  - Agrega blocos de texto que contenham possíveis elementos bibliográficos
    - Datas, nomes, páginas, acronimos e palavras que ocorram em referências bibliográficas (“in proceddings”, “editors”,)
    - Necessário garantir que os blocos encontrados são referências bibliográficas através de filtros bastante restritos
    - A extracção de texto (feita anteriormente) de documentos html identifica possíveis blocos de referências bibliográficas (<li> , <p>, etc)

SUPeRB
de referências a elementos bibliográficos

Paratools tem dois métodos para processar referências

- Pattern Matching
  - Uma colecção de mais de 400 padrões de referências
  - Cada elemento tem um determinado peso,
  - Uma referência pode fazer match com vários padrões mas só um é retornado, aquele cujo conjunto de elementos tem maior peso
  - referências incompletas não fazem match com o padrão correcto
  - Facilmente podem ser adicionados novos padrões

SUPeRB
de referências a elementos bibliográficos

- Um conjunto (wrapper) de várias funções para capturar elementos bibliográficos independentemente
  - Limitado a um conjunto de elementos bibliográficos
  - Melhor performance do que o método anterior
  - Pode encontrar alguns elementos correctos ou outros incorrectos
  - Podem ser adicionados novos elementos a pesquisar
    - Mas é hard-coded

SUPeRB
de referências a elementos bibliográficos

- Outros métodos auxiliares
  1. Determinar se um elemento é um nome
     - Recorrendo à base de dados ou repositorios de EM
  2. Expandir iniciais
     - Recorrendo à base de dados ou repositorios de EM
  3. Abreviaturas de conferências?
     - Recorrendo à base de dados
  4. Determinar a língua da publicação
     - Lingua::Identify
  5. Determinar o tipo de publicação
     - Verificar a existência ou ausência de determinados elementos

SUPeRB
A extracção personalizada

- Situações podem requerer uma extracção personalizada
  - O SUPeRB pode possuir templates para páginas ou domínios conhecidos.
    - Extracção dos dados (referência/elementos bibliográficos) estruturados uma vez que a estrutura do documento é conhecida
  - O utilizador introduz um padrão de captura simples (expressão regular) para extrair os dados
    - Como no caso de ter dados um URL como argumento ao sistema

SUPeRB
A extracção de elementos bibliográficos (dos documentos)

- A partir dos documentos académicos
  - Tal como existe um bloco de referências existe também um bloco descritivo do documento
    - titulo, autores, resumo
  - cabeçalhos ou rodapés nas páginas a indicar autores, jornais, datas, etc (difícil)

SUPeRB
A relevância

- A relevância dos elementos bibliográficos depende de:
  - Obtenção dos dados
    - Serviços Web ou processados de documentos
  - parâmetros usados na entrada
    - e a sua ocorrência no documento (de onde os dados foram extraídos)
    - e a sua ocorrência na referência (de ...)
    - Métodos de extracção

SUPeRB
Relevância /Validação humana

- Após calcular automáticamente os candidatos mais prováveis, é apresentado ao utilizador (ou gestor) um conjunto de referências bibliográficas mais prováveis
  - Uma referência exacta
  - Várias referências encontradas no âmbito da pesquisa
  - Assinaladas aquelas que já pertencem ao catálogo ou que são similares

SUPeRB
Relevância /Validação humana

- O utilizador pode então selecionar as referências para
  - Adicionar ao catálogo
  - Unificar/substituir referencias já existentes
  - Corrigir antes de inserir (elementos bibliográficos mal capturados)

Alguns módulos produzidos

- WebSearch.pm
  - Módulo para interagir com as API de pesquisa
- Utils.pm
  - Conjunto de métodos para extrair o texto
- Data.pm
  - Interace para o MySQL
- Citation.pm
  - Métodos para extrair referências de texto
- ReferenceParser.pm
  - Interface para chamar diversos métodos do Paratools
- Styles.pm
  - Módulo para converter entre formatos e estilos bibliográficos
- EvalRef.pm
  - Módulo para avaliar referências bibliográficas (e elementos). Compara e calcula a relevância entre referências bibliográficas

Alguns dos recursos produzidos

- O catálogo de publicações ☺
- Listas de referências bibliográficas que surgem no mesmo contexo da pesquisa
- Listas de nomes (autores) ou conferências utilizadas para aumentar a interacção com o utilizador
  - Usando AJAX é possível fornecer informação ao utilizador à medida que este preenche um formulário

Conclusões

Conclusões

- Ainda existem poucos resultados e testes para tirar conclusões mas:
- A fase de pesquisa não necessita de tanta interacção humana
  - Os resultados necessitam ser quantitativos não qualitativos
  - A qualidade depende dos argumentos dados para usar como expressão de consulta da sua disponibilidade na Web
- Enquanto que a fase de extracção/ relevância está mais dependente da supervisão humana
  - A lista de candidatos é extensa
  - Existem actualmente duas fases de validação
    - Selecção de uma (ou mais referências bibliográficas)
    - Confirmação da extracção dos elementos bibliográficos

Conclusões

- Os meios actuais para extrair elementos bibliográficos são limitados
  - Não são 100% seguros
  - O tipo de elementos bibliográficos que extrai é limitado, necessitando de ser extendido
- As técnicas de relevância podem não dispor do texto do documento, restringindo-se a usar a própria referência e o contexto onde foi encontrada (url, expressão de pesquisa)
  - Mas a supervisão humana pode colmatar este problema

Conclusões

- Do ponto de vista do utilizador/gestor
  - Aumentou-se a intereacção automática entre o utilizador e a interface
    - Sugerir nomes de utilizadores
    - Sugerir referências bibliográficas encontradas no contexto da pesquisa
      - Pode necessitar ainda de métodos auxiliares para reduzir os candidatos