Luís Cabral

SUPERB
Sistema Uniformizado de PEsquisa de Referências Bibliográficas

FEUP

Proposta de dissertação para obtenção do grau de Mestre em Engenharia Informática

Orientador**:** Prof. Doutor Eugénio de Oliveira
Co-orientador: Eng. Luís Sarmento

Mestrado de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto

Junho de 2005

Resumo

A pesquisa e disponibilização de referências bibliográficas têm vindo a tornar-se um assunto com particular interesse à medida que a Internet tem aumentado o seu tamanho.
A Internet tem sido um dos principais meios de divulgação científica e a disponibilização dessa informação tem sido garantida, através de sítios. São estes que divulgam a informação científica desenvolvida pelas entidades académicas ou de investigação científica responsáveis por esses sítios. No entanto a pesquisa de documentos, e outra informação relevante, tais como referências, domínios ou assuntos relacionados, pode tornar-se bastante penosa quando o tamanho do sítio é consideravelmente grande. Isto só é possível com um sistema que pesquise a Internet, como um motor de pesquisa genérico. Porém, os motores de pesquisa não permitem avaliar os documentos, fornecer os meta-dados, ou dar-nos alternativas ao documento, isto é, documentos idênticos.
Que assunto trata? Onde posso encontrar outro documento semelhante? Onde foi publicado? São as perguntas que fazemos antes de pesquisarmos por entre dezenas de alternativas aparentemente todas elas validas e que parecem responder a uma ou outra das nossas perguntas.
É neste âmbito que surge assim a proposta de desenvolver um sistema de pesquisa de referências bibliográficas: Um sistema automático e fácil de interagir, que dado um mínimo de informação, tal como "autor", o "domínio" ou o próprio documento, é capaz de pesquisar documentação e simultaneamente ser capaz de fornecer informações relevantes, tais como referências bibliográficas, documentos semelhantes ou documentos do mesmo autor.
Esta tarefa não é trivial, sendo necessário aceder a repositórios de informação ou às páginas do autor, já disponível na Internet e que contém alguma da informação requerida. É depois necessário integrar a informação das diversas fontes e usar uma base de conhecimento interna de forma a acelerar o processo de pesquisa.

Índice

1. Introdução 4
2. SUPERB (Descrição da tese) 5
   2.1. Casos de Uso 6
   2.2. Motivação 7
3. Estado da arte 7
   3.1. Sistemas de Extracção de Informação da Internet 8
   3.1.1. Armadillo 8
   3.1.2. KnowItAll 9
   3.2. Técnicas fundamentais para a extracção 11
   3.2.1. Análise terminológica 11
   3.2.2. SIEMÊS 12
4. Objectivos 12
5. Metodologias 13
   5.1. Módulos a implementar 14
   5.2. Ferramentas a usar 15
   5.3. Arquitectura do SUPERB 16
6. Plano de actividades 17
7. Considerações finais 18
8. Referências 19

1. Introdução

As pesquisas bibliográficas são um ponto relevante na realidade académica e científica. É necessário ser-se capaz de extrair mais referências e de citar correctamente a partir dos conteúdos de documentos que estão disponíveis na Internet.
À medida que a Internet tem vindo a crescer, a capacidade de pesquisa de referências bibliográficas de forma eficiente tem vindo a tornar-se uma tarefa crítica. Dada a espontaneidade da Internet no que diz respeito à documentação científica, um Blog e um artigo científico podem ser considerados igualmente relevantes, se tiverem as mesmas palavras-chave.
Por outro lado, a tarefa de gerar bibliografia manualmente, num formato standard e de fornecer a sua disponibilização ao nível de uma universidade, centro de investigação ou empresa é uma tarefa demasiado penosa, mas mais ainda ao nível da Internet, pelo que muitos destes documentos acabam "perdidos" na Internet.
A automatização de processos de pesquisa sobre a Internet, ou parte dela, a extracção de informação de conteúdos de documentos, a capacidade de interligar e referenciar artigos ou que permitam o armazenamento dos dados extraídos no processo e consequente divulgação surge assim como um passo vital na divulgação de informação cientifica. Actualmente existem bases de dados on-line consideravelmente grandes que disponibilizam informação bibliográficas, tais como o CiteSeer (Bollacker et al. 1998) ou o DBLP (Ley, 2002), usando motores de pesquisa na Internet, como interfaces para pesquisar e aceder aos dados.
O DBLP, cujas iniciais significa *DataBase systems and Logic Programming,* é uma base de dados de informação bibliográfica com cerca de 650.000 documentos indexados a partir de *proceedings* e revistas cientificas.
O DBLP armazena a informação em *Table of Contents* (TOCs) no formato XML, para armazenar a informação de cada artigo, estando os registos disponíveis pela rede. O acesso à informação é feito através de um motor de pesquisa próprio(http://dblp.uni-trier.de), não existindo nenhuma API, pelo que pude averiguar, para extrair os dados directamente do repositório XML.
O CiteSeer é uma biblioteca digital que indexa artigos científicos *pdf* e *ps* e fornece algumas funcionalidades tais como estatísticas das referências e citações, indexação de documentos relacionados e extracção de meta-dados dos documentos indexados ou actualizações regulares da informação.
O CiteSeer possui actualmente 723.140 artigos indexados e, tal como o DBLP, este possui da mesma forma uma interface sobre a forma de motor de pesquisa para facilitar o acesso à informação.
O CiteSeer (http://citeseer.ist.psu.edu) já interligava as páginas dos documentos aos registos do ACM Portal (http://portal.acm.org/portal.cfm) e recentemente anunciou que passaria a ligar os seus documentos também às referências bibliográficas do DBLP, segundo a informação dada pelo próprio sistema. Com esta nova funcionalidade, o CiteSeer aproveita para seu beneficio a repetição de documentos e das suas referências pela Internet. Possui ainda uma API pública para o acesso as funcionalidades de que dispõe (Yves Petinot et al., 2004).
Estes tipos de repositórios são bastante fiáveis e são o primeiro passo para a obtenção de informação básica, servindo para iniciar módulos mais complexos como os *wrapper*, usados na extracção de informação de páginas de Internet.

É ainda importante referir que normalmente estes sistemas ordenam os documentos através de critérios probabilísticos baseados no número de citações, para avaliar a sua importância ou qualidade.

2. SUPERB (Descrição da tese)

A proposta que surge neste âmbito tem portanto como finalidade, um sistema capaz de pesquisar referências bibliográficas de forma automática, o SUPERB.
O SUPERB, Sistema Uniformizado de PEsquisa Referências Bibiliograficas, tem como objectivo tornar-se num assistente na pesquisa automática de bibliografia. A partir de um conjunto de sementes que podem incluir títulos de documentos, nome de autores, referências bibliográficas (ex. ISBN), palavras-chave ou o próprio documento, o SUPERB deverá ser capaz de recolher referências bibliográficas adicionais da Internet sobre os mesmos autores e/ou assuntos.
Deverá ser possível recorrer a sistemas de pesquisa Internet já existentes, tais como o Google, ou a repositórios de informação bibliográfica, como o CiteSeer, o DBLP ou outros directórios de publicações disponíveis pela Internet.
O SUPERB tem como objectivo manter uma filosofia semi-automática, permitindo receber retorno do utilizador acerca da informação que vai sendo encontrada de forma a refinar a pesquisa iterativamente. Idealmente pretende-se que o sistema possua algumas capacidades de aprendizagem de forma a poder facilitar futuras pesquisas.
Um outro requisito que merece consideração é a possibilidade de exportar os resultados obtidos nas pesquisas, para formatos habitualmente empregues por sistemas de gestão bibliográfica. A fim de proporcionar o uso do sistema aos utilizadores, deverá ser tomado em conta usabilidade do sistema, pelo que a criação de uma interface que facilite o uso do sistema será tida em conta.
Actualmente já existem soluções para aquilo a que nos propomos no entanto são ou repositórios que não podem ser considerados representativos de toda a Internet, não contem informação adicional sobre o domínio dos documentos, ou então são motores de pesquisa que não estão preparados fornecer informação adicional acerca do documento. No que diz respeito à área da língua portuguesa, pouco ou nada tem sido desenvolvido que possa ser considerado viável. Desta forma pretende-se que o SUPERB possua capacidades de processamento para extracção de referências bibliográficas especialmente adaptadas à língua portuguesa.
Um factor ainda importante e que será alvo de estudo, deverá abordar as defesas a ter em conta para a validação dos actuais mecanismos globais de referenciação. O CiteSeer e o DBLP poderão conter políticas de divulgação científica? O DBLP limita a extracção de informação a um conjunto limitado de fontes, podendo ter autores prolíferos por possuírem muitas publicações nas fontes extraídas, enquanto que outros autores com um grande impacto na sua área de desenvolvimento podem ter um número pequeno de publicações indexadas. A contagem de citações um métodos muito comuns pela qual os repositórios de bibliografia indexam os documentos.
A análise de motores de pesquisa como o Google (http://www.google.com) ou o Yahoo (http://www.yahoo.com) para determinar a viabilidade e facilidade na pesquisa e extracção de informação para o português será também alvo de estudo.

2.1. Casos de Uso

Nesta secção serão descritos alguns casos de uso do SUPERB sucintamente.

2.1.1. Pesquisa simples

O utilizador acabou encontrar uma referência num artigo em papel, pelo qual se mostrou bastante interessado. Acede ao SUPERB pelo Navegador e introduz a referência no campo de pesquisa. O sistema devolve uma página contendo os meta-dados obtidos, inclusive o resumo do documento e um URL onde o documento está disponível. Apresenta também uma lista de documentos semelhantes pela qual o utilizador pode facilmente navegar, escolhendo qualquer dos resultados, onde pode por novamente visualizar a informação desse documento e uma lista de ficheiros semelhantes.

2.1.2. Descarregar documento

Um utilizador acabou de ler um artigo no seu computador num formato comum. Uma vez que considerou esse artigo interessante, gostaria de poder avaliar outras perspectivas, para além da perspectiva do autor. O utilizador acede à interface do SUPERB para pesquisar mais informação relacionada com o assunto. Acede à interface que permite enviar um documento para ser analisado pelo SUPERB.

O sistema apresenta uma página com os meta-dados obtidos a partir do documento e da Internet. Exibe também uma lista de documentos que o sistema considerou idênticos ao submetido. Facilmente o utilizador pode repetir os passos para qualquer um dos documentos que obteve como resposta. Desta forma o sistema não só dará um conjunto de respostas relacionadas com a nova pesquisa mas também começará a especificar o domínio da pesquisa, tendo sempre em conta as pesquisas anteriores.

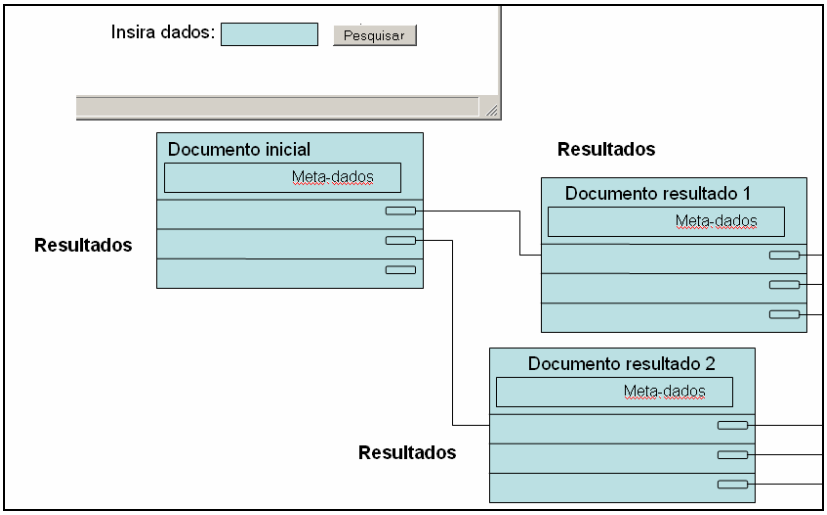

Figura 1 Esquema da àrvore de pesquisa possiveil com o SUPERB

2.1.3. Acesso a partir de um sítio exterior

Um caso de uso que pode ser considerado é ter este sistema inserido numa biblioteca on-line para poder complementar os utilizadores na busca por informação.

O sítio possui um motor de pesquisa interno para aceder ao catálogo de livros ou documentos disponíveis, inclusive publicações ou dissertações. Possui no entanto uma hiperligação para a API do SUPERB em cada dissertação que é listada.
Quando um utilizador desejar, pode carregar na hiperligação, que através da API SUPERB envia o título e o nome do autor, podendo ainda enviar o URL do documento, se este estiver disponível on-line. O SUPERB devolverá os meta-dados do documento, bem como uma lista de documentos que são considerados semelhantes.

2.2. Motivação

Esta proposta teve origem na ligação como Bolseiro da Linguateca pelo que se encontra no âmbito dos objectivos desta: a divulgação da língua portuguesa e disponibilização de recursos. Alguns dos projectos em que tenho estado envolvido, como o Corpógrafo (www.linguateca.pt/Corpografo) ou SIEMES, que recorre a um repositório local, o REPENTINO (http://www.linguateca.pt/REPENTINO) os quais serão discutidos mais à frente, estão relacionados com a extracção de informação a partir de textos da Internet.
Esta proposta permitirá alargar os conhecimentos nesta área, possibilitando o desenvolvimento de um sistema de pesquisa de bibliografia para a língua portuguesa. Pelo que pude aperceber-me ao longo deste processo de análise bibliográfica, esta área não tem sido muito abordada no que diz respeito ao tratamento da língua portuguesa em particular, notando-se uma ausência de ferramentas que produzam os resultados desejados. Outro factor que me motiva é a possibilidade de poder desenvolver um sistema que possa ser útil e que tenha uma aplicação real, tanto mais que, como já referido, não existem aplicações semelhantes à que aqui é proposta.
Os actuais sistemas de pesquisa de referências bibliográficas são repositórios que são na realidade bases de informação necessitando ser actualizados periodicamente ou então os dados têm que ser introduzidos manualmente. O seu conhecimento não vai além da informação armazenada no próprio repositório e de outros repositórios idênticos, desde que os documentos em causa façam parte do próprio conhecimento, não trazendo nada de novo. Um sistema que pesquise a Internet em tempo real põe um desafio interessante do qual se podem obter resultados interessantes.
Para concluir, ao longo desta pesquisa inicial, verifiquei que muitas instituições criam os seus próprios repositórios locais, existem repositórios que possuem domínios específicos, com algumas centenas de documentos. Como podemos chegar a todos eles e como podemos avalia-los para saber se é encontramos aquilo que estávamos à procura? Isto é aquilo que este sistema deve permitir fazer com o máximo de facilidade.

3. Estado da arte

O CiteSeer associa os documentos baseando-se mais no autor ou nas co-referências, ou seja, se dois documentos contêm referências que ocorram em ambos simultaneamente, então conclui-se que estes são idênticos. Logo, o conteúdo do texto é menosprezado, a terminologia de cada documento é ignorada.

Técnicas que permitam verificar a terminologia de cada documento e que permitam comparar documentos com base nesses dados é possível, podendo proporcionar resultados mais recompensadores. Esta é portanto uma metodologia que pode ser vantajosa.
A possibilidade de relacionar documentos do mesmo autor ou cujos temas são idênticos, é uma tarefa que pode ser conseguida através da pesquisa de dados a partir da extracção de informação de texto da Internet.
A tarefa que aqui é proposta tem por objectivo desenvolver um assistente que permita a pesquisa de documentos com base num conjunto de sementes extraídas de um documento ou dados pelo utilizador. Entende-se por sementes aquelas palavras ou termos que caracterizam um documento e que serão usados na descoberta de outros documentos semelhantes. O sistema deverá ser capaz de encontrar outros documentos relacionados de forma semi-automática. Para obter esta informação, é necessário processar um texto para dai extrair palavras-chave que caracterizem esse documento. As palavras-chave que se pretende encontrar nos textos são Entidades Mencionadas (i.e. uma entidade que é referenciada num determinado contexto), terminologia técnica relevante e as referências. Com base nas sementes obtidas, o assistente constrói *queries* que usa para invocar vários sistemas de pesquisa na Internet ou repositórios de forma a obter prováveis candidatos.
Estas metodologias não são novas, mas não dizem respeito à informação bibliográfica. A informação bibliográfica está normalmente armazenada em repositórios pela Internet, Estas metodologias estão de facto implementadas mas noutros sistemas que tem outras finalidades, como por exemplo para a obtenção de informação geográfica a partir de textos da Internet (Etzioni et al., 2005).
Por esse motivo, as seguintes secções dizem respeito a sistema de extracção de informação genérica ou com outros propósitos que não a bibliografia. Mas estes sistemas são bons exemplos da utilização de tecnologias relativamente simples e até aparentemente ingénuas. No entanto, funcionam na prática pela elevada redundância na Internet. O sistema proposto assenta nos mesmos princípios da utilização de técnicas simples.

3.1. Sistemas de Extracção de Informação da Internet

Existem algumas ferramentas que são vistas como o estado da arte no que diz respeito à extracção de informação a partir da Internet. Esta subsecção faz uma breve análise a algumas ferramentas de extracção de informação de forma não supervisionada ou levemente supervisionada.

3.1.1. Armadillo

O Armadillo (Ciravegna et.al., 2004) é um sistema que recorre a vários serviços para anotar e extrair informação, de forma automática e com o mínimo de intervenção humana, de um domínio específico para um repositório. Esta informação armazenada é depois ser aplicada na descoberta de novas instâncias. Por exemplo, este sistema pode extrair nomes de filmes de texto, sendo capaz de reconhecer e de relacionar títulos de filmes como:

"The big chill"

e

"Big chill, the".

Recorrendo à verificação de redundância (múltiplas provas em diversas fontes distintas) as anotações são validadas e usadas para extrair novas informações, permitindo assim uma expansão continua e automática da base de conhecimento.

O Armadillo extrai informação de vários serviços de Internet que por sua vez tem funções específicas e recorrem a outros sistemas. É através destes sistemas que o Armadillo obtém a informação necessária a extracção de informação de um determinado domínio. Por exemplo, um serviço de reconhecimento de entidades nomeadas de um sítio de uma universidade recorre a um sistema de reconhecimento de entidades nomeadas para identificar potenciais nomes. Outros serviços procurariam obter artigos, de um investigador identificado no serviço anterior, do CiteSeer ou do DBLP. Cada serviço produz resultados pouco fiáveis só por si, de pouca precisão mas a combinação dos diversos serviços produz resultados com uma precisão alta.

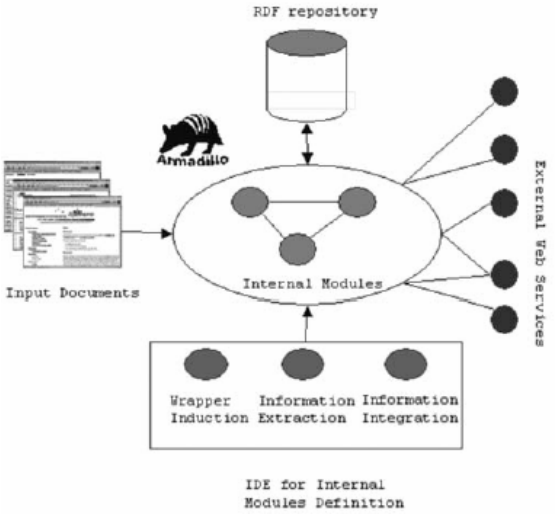

Figura 2 Arquitectura do Armadillo

A informação obtida pelos vários serviços é de seguida integrada, através de ontologias num repositório RDF, onde é armazenada.
Explorando a redundância da informação na Internet e posteriormente no repositório gerado, o Armadillo extrai informação com diferentes graus de confiança e expande a sua base de conhecimento inicial. Estas metodologias evitam a aquisição de informação espúria baseadas em informações erradas.
O Armadillo funciona com o mínimo de intervenção humana, o utilizador fornece um URL e alguma informação adicional não requerendo anotações manuais. Após a intervenção do utilizador, os dados que este alterou, apagou ou adicionou, podem ser usados novamente para reiniciar a aprendizagem de forma a obter mais informação e maior precisão.

3.1.2. KnowItAll

Outro sistema que lida com a pesquisa e extracção de informação de forma não supervisionada ou levemente supervisionada é o KnowItAll (Etzioni et al., 2005). Ao contrário do Armadillo, este permite a pesquisa independente de domínio, através de oito conjuntos de padrões que permitem a determinar candidatos a factos, como localizações geográficas, através da instanciação de uma classe.
Este sistema permite extrair informação como CIDADEDE("*Porto*","*Portugal*") a partir de texto comum como

"…a cidade do Porto, em Portugal …."

Que como pode se compreender facilmente, indica que Porto é uma cidade de Portugal.

O KnowItAll assenta essencialmente em 3 métodos distintos:

- Aprendizagem de padrões capazes de serem usados tanto com regras de extracção ou de validação das instancias extraídas
- Extracção de subclasses (ex. capaz de extrair subclasses de cientista (ex. "*physicists"*, "geologists", *etc.*)
- Capaz de extrair listas de classes, através da aprendizagem de um "wrapper"

Qualquer um destes métodos dispensa a marcação de textos para aprendizagem, dado que a informação extraída pelos padrões é carregada no *BootStrapping* de forma a gerar regras de extracção.

As regras geradas no *BootStrapping,* são usadas pelo extractor para gerar questões a motores de pesquisa que retornam páginas que são ser processadas, recorrendo novamente às regras, para extrair informação das páginas obtidas. Estas regras são construídas através de um predicado simples com por exemplo "*City"*. A partir deste predicado, o KnowItAll é capaz de produzir um padrão de extracção, restrições, *bindings* e palavras-chave.

Predicate:	City
Pattern:	NP1 "such as" NPList2
Constraints:	head(NP1)= "cities"
	properNoun(head(each(NPList2)))
Bindings:	City(head(each(NPList2)))
Keywords:	"cities such as"

Ilustração 1 Regra de extracção gerada substituindo a classe "*City*" numa regra de base

O extractor usa expressões regulares, para identificar nomes simples e listas de nomes simples.

"*The tour includes major cities such as New York, central Los Angeles and Dallas*".

O KnowItAll é capaz de identificar *New York*, *Los Angeles* e *Dallas* como nomes que pertencem à classe "*City*", na frase anterior.

As restrições servem para especificar a cabeça de cada frase ou de cada nomes simples e listas de nomes simples. As palavras-chave servem para fazer *queries* a motores de pesquisa. De facto é primeiro feita a pesquisa através da palavra-chave e só depois é que é possível aplicar a regra às páginas obtidas na pesquisa.

Os resultados obtidos na extracção de informação podem, por sua vez, gerar mais regras ao serem usadas. Este processo é também completamente automático, ao contrário de muitos outros que requerem a criação manual de sementes de treino. No início de cada ciclo, o KnowItAll favorece os predicados e regras que tenham sido mais produtivas.

O KnowItAll é ainda capaz de criar uma regra semelhante a "*City*", para a palavra "*Town*", se o predicado especificar "*town* as well as *city*".

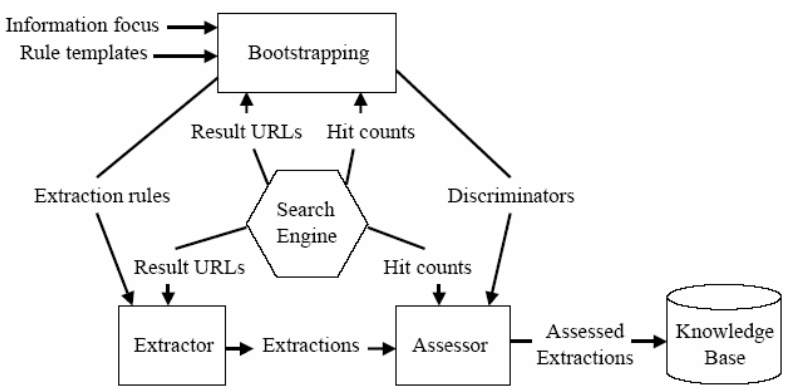

Figura 3 Grafo de fluxo do KnowItAll

O KnowItAll recorre também a sistemas Internet de pesquisa, tratando a Internet como um corpus de texto gigante. O Assessor usa estatísticas *pointwise mutual information* (PMI) entre palavras e frases obtidas pelos motores de pesquisa, para validar a qualidade dos factos extraídos.

3.2. Técnicas fundamentais para a extracção

Anteriormente falamos de sistemas completos, agora importante falar de alguns sistemas bem como de técnicas de extracção/classificação de informação/conhecimento que serão úteis no contexto deste projecto. Tratam-se de sistemas com os quais estou particularmente familiarizado dado que participei no seu desenvolvimento. Esta subsecção tem especial interesse para a avaliação de documentos na língua portuguesa, uma vez que todos os sistemas que aqui serão mencionados, foram desenvolvidos para a língua portuguesa em particular.

3.2.1. Análise terminológica

A extracção terminológica de documentos pode permitir extrair pistas importantes sobre os documentos em análise, com um nível de precisão consideravelmente alto, que o torna suficientemente fiável (Sarmento, 2005). Com base nisto, é possível que a terminologia sirva como um método alternativo na procura e na comparação de bibliografia.

O Corpógrafo (http://www.linguateca.pt/corpografo) possui um módulo, o FLEXT, para a extracção terminológica de termos, independente do domínio dos textos. Este módulo permite fazer uma estimativa probabilística da redundância de N-Gramas extraídos de texto, análise do contexto e análise morfológica das palavras. Este processo é usado para obter possíveis candidatos de termos de forma semi-automático, complementando uma metodologia que garante um grau de precisão razoável com a validação de um terminologista antes dos termos serem inseridos numa base de dados terminológica.

Este sistema, razoavelmente simples de implementar, permite assim facilitar o reconhecimento de termos através de um esforço mínimo que se limita à validação humana de um perito, ficando o resto do trabalho a cargo do sistema e das metodologias que implementa.

3.2.2. SIEMÊS

O SIEMES, Sistema de Identificação de Entidades Mencionadas e Estratégia Siamesa, é um sistema que integra o conhecimento do REPENTINO e uma metodologia que testa um conjunto de regras de forma a analisar o contexto. As duas estratégias complementam-se de forma poder classificar entidades mencionadas em texto extraído da Internet, de forma relativamente fiável.

Trata-se de um sistema que funciona de forma não supervisionada, ainda que durante as suas recentes aplicações, tenha sido reajustado, essencialmente pela inserção de novas regras, que efeito de forma fácil, ainda que no código, ou pela inserção de exemplos de entidades nomeadas considerados suficientemente representativos de parte de uma categoria. Por exemplo,

"Escola C+S Pires de Lima"

Tem uma cabeça forte (*"Escola C+S"* ou somente *"Escola"*), representativa de parte da categoria "Organização:: Ensino/I&D".

Este projecto superou as nossas expectativas quando, recentemente, participou no HAREM (http://www.linguateca.pt/HAREM), Avaliação de Reconhecimento de Entidades Mencionadas, tendo este ficado classificado numa posição consideravelmente favorável.

Ainda recentemente, o SIEMES foi aproveitado, por outros elementos da Linguateca, para a participação no CLEF, na avaliação de sistema de perguntas e respostas.

O REPENTINO, REpositório de ENTIdades NOmeadas, é um repositório com mais de 400.000 exemplos de entidades Nomeadas em português. Está categorizado em 11 categorias e cerca de 100 subcategorias (ex. "Luís Cabral" é um Ser::Humano, "Rio Douro" é um Local::Hidro). OS exemplos foram extraídos do WPT03, um recurso que contem textos da Internet portuguesa, e da própria Internet.

O REPENTINO possui uma interface para a Internet, sob a forma de motor de pesquisa que de forma simples e rápida, pesquisar os dados do repositório, não deixando no entanto de possibilitar pesquisas avançadas (http://www.linguateca.pt/repentino).

4. Objectivos

O sistema que se pretende desenvolver, assenta em duas partes distintas que devem ser analisadas:

- A parte teórica, que tem em consideração as técnicas a aplicar para se avaliar o grau de semelhança entre documentos, envolvendo uma análise das técnicas simples para a extracção de referências bibliográficas, nomeadamente através da extracção de terminologia. O sistema desenvolvido deverá ser capaz de se ajustar automaticamente, optando por termos mais significativos para a pesquisa. Estes termos poderão ser identificados com base nos termos já extraídos, aplicados e que fornecem resultados que chamam o interesse do utilizador.

  Para cumprir estes objectivos, será ainda necessário que estas metodologias sejam implementadas e devidamente testadas, de forma a provar a sua aplicação.

- Uma componente prática, que culminará na disponibilização de uma ferramenta prática e útil para a pesquisa de bibliografia na Internet. Pretende-se desenvolver um motor de pesquisa automático, capaz de pesquisar a Internet e que aceda e processe os resultados encontrados e os apresente ao utilizador de forma explicita. Este sistema deverá permitir ao utilizador aceder a um leque de funcionalidades que facilitem e automatizem a pesquisa de informação

  Deverá apresentar dados como autores, título, local de publicação, hiperligação para o documento e documentos alternativos que possam ser semelhantes ao original.

  Deverá permitir a extracção de informação sobre formatos comuns, como o bibtext ou o XML. E ainda deverá ser um sistema que possa facilmente ser aplicado a outros sistemas ou aplicações, através de uma API de Internet.

5. Metodologias

O desenvolvimento deste sistema deverá envolver várias metodologias distintas como a pesquisa de documentos e o processamento e extracção de informação de textos provenientes da Internet.

O sistema proposta deverá obter um ficheiro directamente da Internet, recorrendo a motores de busca (por exemplo, o Google, Yahoo, etc.) com base em parâmetros dados pelo utilizador (titulo ou nome dos autores). Alternativamente o sistema permitirá descarregar um ficheiro directamente do computador do utilizador ou de um URL. O documento poderá ser aceite em qualquer dos formatos habitualmente usados na publicação de documentos científicos na Internet.

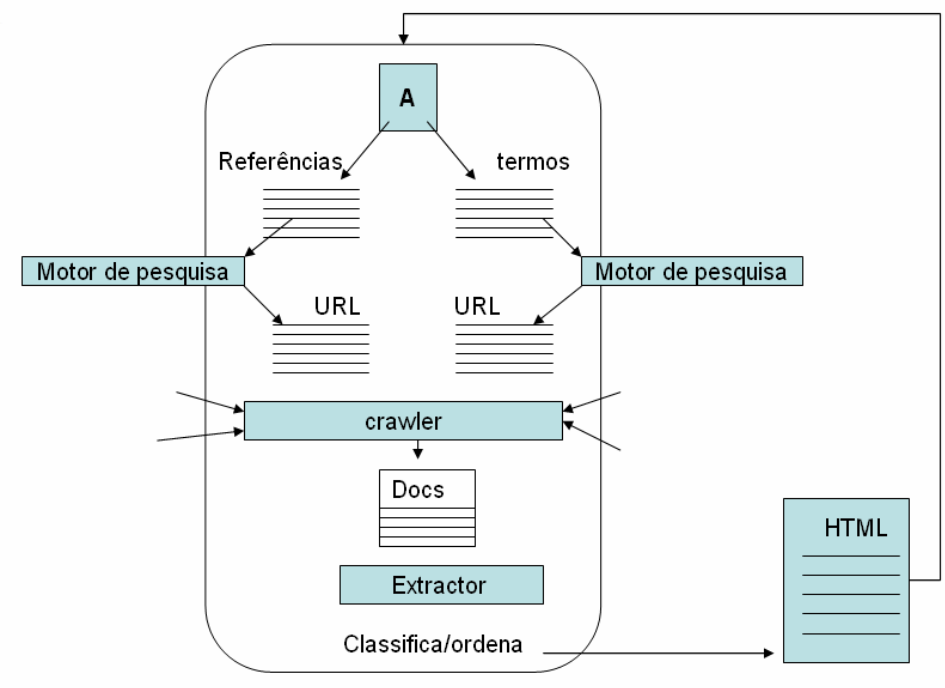

Figura 4 Workflow do SUPERB

Após ter obtido o documento, o SUPERB analisará o texto extraído do documento através de métodos probabilísticos de forma a obter elementos do texto que caracterizem esse mesmo documento, o seu domínio, inclusive todos os outros dados de um documento científico (referências, autores, titulo, etc.). O SUPERB deverá depois usar esta informação para tentar obter documentos similares a partir da Internet. Este passo não se limita a localizar os dados da Internet. Deverá ainda obtê-los e avaliar a sua semelhança com o fornecido pelo utilizador, usando para tal, medidas de comparação em relação ao texto que compõe cada um dos documentos.

5.1. Módulos a implementar

Para isto, o SUPERB deverá ser composto por diferentes módulos cada um com uma função específica:

- Descarregar documentos directamente para o sistema

O SUPERB será capaz de descarregar um ficheiro do computador do utilizador ou de um URL por ele indicado.

- Extracção de textos, de documentos em vários formatos, através da integração com vários programas próprios.

A Internet contém documentos em diversos formatos, havendo a necessidade de que o sistema disponibilize formas de converter, através de aplicações especializadas, todos os formatos que possa ser necessário tratar para um formato único, texto. Será assim possível que os módulos de extracção de informação possam processar todos os documentos que possam surgir.

- Acesso a motores de pesquisa ou repositórios de informação externo

Este passo consiste em consultar de forma automática vários serviços disponíveis na Internet, gerando os parâmetros da pesquisa e processando os dados que o serviço devolver (Daume & Brill, 2004). Isto pode ser auxiliado por API próprias e que já estão disponíveis, ou então através do processamento de documentos HTML retornados pelo serviço, criando-se para tal uma API própria.

- Extracção de informação relevante a partir de texto

Nesta fase, deverão ser aplicadas várias metodologias, como o uso reconhecimento de entidades ou a extracção de terminologia (Ciravegna et al, 2004, Etzioni et al.,2005 e Santos et al., 2003). A metodologia a ser implementada aqui, aborda métodos probabilísticos e linguísticos que garantam que os dados extraídos são de facto representativos do documento. Recorrendo a informação extraída da Internet em tempo real, pode-se pesquisar documentos em profundidade, isto é, pesquisar uma hierarquia de relações entre ficheiros até que sejam obtidos resultados suficientes. Estas metodologias deverão no entanto possuir uma rapidez considerável, mesmo em conjunto com os serviços externos. Para isto, a base de conhecimento deverá poder armazenar informação temporária de forma a acelerar o processo.

- Base de Conhecimento

A Base de conhecimentos tem como propósito permitir que o SUPERB possua uma base de dados temporária relativamente às pesquisas mais recentes e frequentes de forma a acelerar o processo de pesquisa.
Deverá conter ainda a informação necessária aos módulos de extracção, tais como padrões comuns ou regras a aplicar aos documentos (Ciravegna et al, 2004 e Etzioni et al.,2005). O armazenamento desta informação como dados, pode permitir a alteração destes dados de forma fácil e sem ter que alterar o código.
O esquema desta base de dados deverá ser pensado, atendendo à informação que aqui se pretende agrupar, não devendo armazenar informação desnecessária mas apenas informação que permita representar um documento., Não tem como propósito armazenar documentos completos ou manter a informação dos documentos permanentemente.

- Interface de Internet

Interface desenvolvida para facilitar o acesso a todas as funcionalidades disponibilizadas pelo sistema. Pesquisas simples ou compostas, visualização de dados e analise de dados, avaliação dos resultados e importação da informação para o computador do utilizador.

- API de rede

Deverá existir também uma interface para o acesso ao sistema através de um protocolo e de métodos simples para poder ser acedido por outros programas ou motores de pesquisa. O sistema deverá usar o HTTP para, através de XML ou métodos normais como o GET e o POST poderem obter dados do sistema, devidamente formatados. Esta API permitirá a fácil implementação noutros sistemas, como por exemplo, num outro sistema que poderia usar o SUPERB como um serviço externo de bibliografia.

5.2. Ferramentas a usar

O SUPERB deverá correr como um serviço disponibilizado por um servidor acessível pela Internet. Para alem deste factor, existe já muito trabalho elaborado sobre esta área em diversas linguagens mas uma das mais comuns, em parte devido à sua facilidade em lidar com expressões regulares, é o Perl. É também uma linguagem familiar (com a qual lido no dia a dia). Para alem disso, há ainda a ter em conta que se pretende desenvolver um sistema livre, facilmente distribuível, capaz de correr num servidor e necessita de ter recursos locais de armazenamento de dados. Colocadas estas questões, determinou-se que a melhor opção deveria ser um sistema LAMP (Linux + Apache + MySQL + Perl), composta por ferramentas livres.
O Linux é um sistema reputado no seu uso como servidor de Internet e com um leque de ferramentas já incluídas de tal forma que o acesso a todas as outras aplicações e respectiva instalação, se necessário, é relativamente fácil. Apesar disso, e dado que todas estas ferramentas (Apache, MySQL e Perl) estão disponíveis para Windows, não deixa de ser possível colocar-se a possibilidade de que o SUPERB possa correr a partir de um servidor Microsoft, não sendo no entanto uma prioridade.

O Apache é um servidor de Internet, livre, e disponibilizado com a maioria das distribuições em Linux. A sua integração com o Perl já existe através de módulos do próprio apache e do Perl.
O MySQL é uma base de dados relacional bastante conhecida, sendo também distribuída com a maioria das distribuições em Linux. Um dos servidores de base de dados mais conhecidos e muito comum em servidores de Internet. É bastante fiável, rapida e possui drivers que permitem a acesso a partir de outras aplicações ou linguagens, inclusivamente o Perl.
O Perl é uma linguagem de scripting. Contêm centenas de módulos que lhe permitem um variadíssimo leque de funcionalidades, inclusive a interacção com outras ferramentas, nomeadamente o Apache e o MySQL. Possui ainda muitas outras funcionalidades que a tornam uma linguagem onde é possível executar qualquer tarefa.
Adicionalmente a estas, é deverão ainda ser necessárias algumas aplicações para executar outras tarefas, como por exemplo, para a conversão diversos formatos de documentos para texto (pdf2txt, ps2txt, etc.).

5.3. Arquitectura do SUPERB

A arquitectura planeada para o SUPERB, é bastante simples, decompondo as várias especificidades em partes distintas.
A parte principal do sistema deverá ser composto pelos módulos que processam o texto, desde a sua conversão até à extracção de informação do mesmo. Nesta parte deverão ainda ser incluídos as metodologias para permitir o acesso à base de conhecimento. Já existe um módulo Perl para o propósito, o DBI. Pode no entanto ser mais fácil encapsular este módulo dentro de outro mais específico para o sistema, com o intuito de simplificar a inicialização e os acessos à informação.
O acesso a serviços externos e a repositórios deve está separada, em API distintas, uma vez que cada um desses recursos é diferente dos outros. Cada uma das API deve estar preparada para lidar com o respectivo serviço, podendo estar já a usar API próprias para aceder ao serviço, ou então a ter que processar os resultados gerados em HTML.

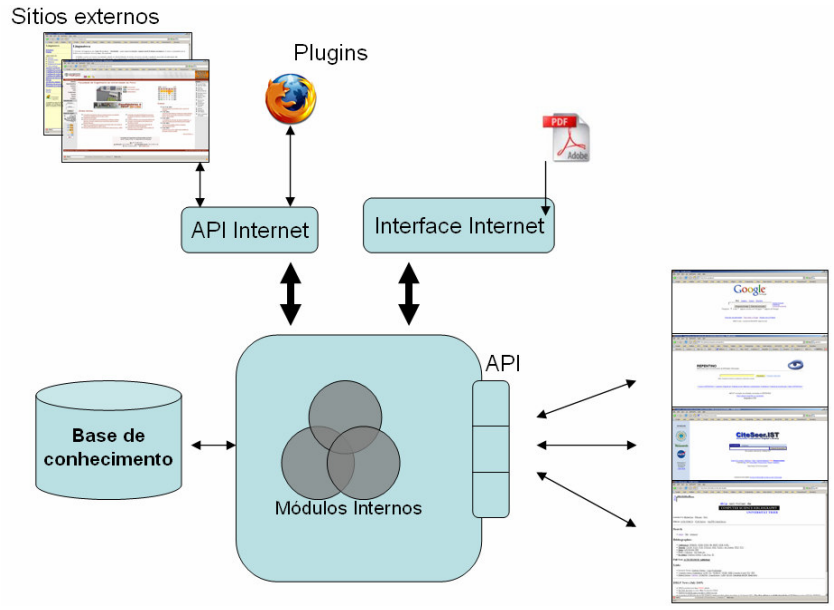

Figura 5Esquema da arquitectura do SUPERB

A interacção com o sistema deverá ainda ser possível através de duas interfaces distintas:

- Uma interface de Internet, acessível através de um navegador como o Internet Explorer ou o Mozilla Firefox via Browser onde será possível aceder ao máximo de funcionalidades que o sistema disponibilize.
- Uma API que receberá pedidos provenientes de sites ou de aplicações e que devolverá os resultados.

6. Plano de actividades

Este projecto pode ser decomposto em várias actividades que podem ser escalonadas:

a. Pesquisa bibliográfica. A pesquisa, que já está em curso, deverá fornecer a base para a escrita do estado da arte. (2 semanas)
b. Definição e estudo das entidades associadas ao problema e construção de uma ontologia respectiva (2 semanas)
c. Estudo e implementação de formas de acessos automática aos repositórios já existentes através de motores de pesquisa ou API próprias para o efeito.
   Esta fase encontra-se dividida, sendo que cada repositório será analisado independentemente, e em fases distintas do projecto. Esta divisão foi planeada para, em caso de atrasos se poder dispensar a implementação do acesso a um dos repositórios. (1 mês e 1 semana)
d. Desenvolvimento dos módulos de processamento de documentos. Nomeadamente para a extracção de terminologia e de referências bibliográficas que estes contenham (1 mês).
e. Desenvolvimento dos módulos de pesquisas de documentos associados, usando a informação extraída pelos módulos desenvolvidos na etapa *d.* e usando os mecanismos de acesso desenvolvidos na tarefa *c.* (por exemplo, o acesso a fontes externas) (1 mês)
f. Desenvolvimento da interface de pesquisa. Nesta fase será desenvolvida a interface do sistema. Deverá ser uma interface de Internet, com diversas janelas, para as funcionalidades proporcionadas (pesquisa, visualização de resultados, optimização da pesquisa, qualificar os resultados da pesquisa). (2 semanas)
g. Testes funcionais do sistema. Nesta fase, pretende-se testar o sistema de forma a detectar, resolver problemas ou optimizar o sistema. Esta fase de testes deverá inclusive afectar a usabilidade da interface. (2 semanas)
h. Escrita da dissertação. Fase final do projecto onde se procederá à descrição escrita do projecto, desde o estado da arte, metodologias aplicadas e dos resultados obtidos. (um mês e meio)

O planeamento das actividades foi elaborado de forma a estar compreendido num espaço temporal de seis meses, o tempo estimado para a elaboração da tese. A carga horária deverá variar entre as 20 e as 30h semanais. O sobre-posicionamento de actividades ocorre mas só em actividades que estão interligadas ou que são consideradas menos penosas.

	Setembro	Outubro	Novembro	Dezembro	Janeiro	Fevereiro
a.						
b.						
c1.						
c2.						
c3.						
d.						
e.						
f.						
g.						
h.						

Tabela 1 Diagrama temporal de actividades

No caso das actividades começarem a sofrer atrasos, estão previstos que se possa suprimir a actividade *c3*., dado que já deverão ter sido desenvolvidas as API para dois outros sistemas externos. O início de cada tarefa deverá ser cumprido, e caso a tarefa anterior ainda não tenha sido concluída como previsto, poderá maximizar-se carga horário semanal. Com estas precauções, estima-se uma recuperação de duas a três semanas.
Para a escrita da dissertação espera-se que um mês e meio seja suficiente, tendo em conta que o estado da arte deverá ser parcialmente redigido na fase inicial, nos meses de Setembro e Outubro, ao longo da pesquisa bibliográfica.

7. Considerações finais

O trabalho que aqui é proposto, vai para além da criação de um repositório estático ou que necessita ser actualizado periodicamente. Na Internet, a informação está dispersa por inúmeros sítios em vários formatos e onde a melhor forma de lá chegar, muitas vezes, é através de um motor de pesquisa comum. Porém, os motores de pesquisa não permitem avaliar os documentos, fornecer os meta-dados, ou dar-nos alternativas ao documento, documentos idênticos.
O conceito de procurar toda a Internet é assustador, mas os motores de pesquisa já possuem informação sobre a localização de milhares de documentos na Internet. O que eles não fazem, é aprofundar a pesquisa de forma a obter informação como "Quem o escreveu?", "Em que conferência/revista foi submetido?" ou "Que outros documentos há semelhantes a este?"
As metodologias que aqui foram propostas, deverão permitir uma nova abordagem na forma como é feita a pesquisa bibliográfica, principalmente na forma como os documentos são relacionados entre si.
Com a combinação destas metodologias será possível responder às perguntas feitas em cima. E estas perguntas poderão ser feitas sobre qualquer documento que esteja disponível na Internet, não apenas sobre um repositório.
Por outro lado existe ainda a compilação de informação de várias fontes, a informação não está disponível somente no documento. Existem dados que estão em repositórios,

páginas pessoais dos autores, páginas de instituições, páginas das conferências, ou de revistas cientificas on-line. Este tipo de informação pode ser facilmente integrada para fornecer ao utilizador o máximo de informação disponível na Internet, num único ponto. Para concluir, esperamos criar um sistema que de facto mostre ser útil e inovador. Para garantir isso, este sistema optou por adaptar-se à natureza da Internet , permitindo pesquisas em tempo real para, usando de seguida métodos de avaliação e de pesquisa mais aprofundados.

8. Referências

[1] Kurt D. Bollacker, Steve Lawrence, C. Lee Giles. "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications" in Proceedings of the Second International Conference on Autonomous Agent, 1998

[2] Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda2, Hui Han, Isaac Councill. "CiteSeer-API: Towards Seamless Resource Location and Interlinking for Digital Libraries" in ACM Press, 2004

[3] Michael Ley. "The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives" in SPIRE 2002

[4] Steve Lawrence, C. Lee Giles, Kurt Bollacker. "Digital Libraries and Autonomous Citation Indexing" in IEEE Computer Society Press 1999

[5] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu Tal Shaked, Stephen Soderland, Daniel S.Weld, and Alexander Yates. "Unsupervised Named-Entity Extraction from the Web: An Experimental Study". Department of Computer Science and Engineering University of Washington, in Artificial Intelligence Journal, 2005

[6] Fabio Ciravegna, Sam Chapman, Alexiei Dingli, Yorick Wilks. "Learning to Harvest Information for the Semantic Web" in Proceedings of the 1st European Semantic Web Symposium (ESWS-2004), Heraklion, Greece, May 10-12, 2004

[7] Fabio Ciravegna and Daniela Petrelli. "User Involvement in Adaptive Information Extraction: Position Paper", in Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, held in conjunction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001

[8] Luís Sarmento. "A Simple and Robust Algorithm for Extracting Terminology". In META Simposium - For a Proactive Translatology (Université de Montréal, Québec, Canadá, 7-9 April 2005)

[9] Santos, D., Maia, B. & Sarmento, L. "Gathering empirical data to evaluate MT from English to Portuguese". In Proceedings of LREC 2004 (Workshop on the Amazing Utility of Parallel and Comparable Corpora). Lisboa, Portugal, 25 May 2004

[10] H. Daume, E. Brill. “Web Search Intent Induction via Automatic Query Reformulation” in Proceedings of HLT 2004

[11] Agichtein, Eugene & Gravano, Luis."Snowball: Extracting Relations from Large Plain-Text Collections", in Proceedings of the Fifth ACM Conference on Digital Libraries (San Antonio, TX, USA, June 2-7, 2000), pp. 85-94.