Recomendação Semântica de Plantas Medicinais

Manuel Cabeleira [1], Igor Cruz [1]

cabeleira@student.deu.uc.pt , igorcruz@student.dei.uc.pt,

[1] Universidade de Coimbra, Morada, 3000, Coimbra, Portugal

Resumo: Neste trabalho vamos implementar um web-site semântico para tratamento de informações relativas a plantas medicinais. Partindo de um web-site desorganizado sem qualquer tipo de ontologia associada, iremos percorrer um longo caminho desde o web-scrapping da informação à elaboração de mecanismos avançados de pesquisa semântica com a finalidade de construir um web-site em que a informação está devidamente acessível e tem associada uma camada semântica. Este tipo de abordagem permite aos utilizadores de plantas medicinais relacionar as plantas com as doenças e aumentar a velocidade das suas pesquisas. Permite também recomendar plantas semelhantes às pesquisadas por um dado utilizador, através de um processo de recomendação semântica. Serão utilizadas tecnologias como Ontologias, Protégé, Jena Triplestore, JSPs e SparQL.

Palavras-chave: web semântica; triplestore; plantas medicinais; semantic recomendation; semantic search

1. Introdução

A quantidade de informação existente na Internet aumenta de dia para dia, e a par com o aumento de informação, existe a necessidade de estudar também mecanismos que a permitam tratar, interpretar e filtrar. É importante que os utilizadores finais e as máquinas consigam encontrar o que procuram mais facilmente e mais objetivamente.

Com vista a aplicar os conceitos de web semântica aprendidos nas aulas, decidimos fazer um *web-site* onde implementámos técnicas semânticas sobre um *dataset* com informação sobre plantas medicinais utilizadas em medicina Ayurveda.

Este sistema medicinal, tem vindo a ser estudado na Índia há mais de 7000 anos tenso sido o precursor de muitos outros sistema. De entre estes sistemas são de destacar a medicina tradicional chinesa e a medicina moderna praticada em todos os hospitais hoje em dia. Este sistema faz uso de plantas medicinais como principal fonte de elementos ativos a utilizar em caso de doença.

Devido ao elevado número de plantas utilizadas neste sistema medicinal, é importante criar mecanismos de pesquisa de plantas que ajudem os especialistas a escolher as plantas que vão utilizar no tratamento.

Os especialistas em medicina Ayurveda estão interessados em relacionar as plantas com as doenças por estas tratadas, as partes das plantas que contêm os elementos ativos e as suas propriedades medicinais. Saber retirar partido destas informações permitirá levar a uma melhoria da performance destes especialistas, pois vão poder encontrar a planta a utilizar mais rapidamente ou encontrar alternativas a esta de uma forma mais eficaz.

As técnicas de web semântica são ideais para esta tarefa pois proporcionam modelos que imbuem a informação com o seu significado semântico.

Por forma a atingirmos um produto final (web-site semântico) congruente com as nossas intenções, o projeto foi subdividido em 6 fases:

- Extração do *dataset* usando técnicas de *Web Scrapping*,
- Criação do modelo da ontologia,
- Criação de Triplos usando o modelo criado como base,
- Implementação do *website* (Sub-divida em):
    - *Browsing*,
    - Recomendação Semântica,
    - Pesquisa Semântica.

2. Web Scrapping

Para recolher os dados utilizados neste trabalho tivemos de realizar *web-scrapping* do *website* [1].

Com o auxílio de expressões regulares foi feito o *parsing* do *website* por forma a recolher informações como nome científico, nome da família, descrição, partes comestíveis, partes utilizadas no tratamento de doenças e propriedades medicinais. Estas informações foram guardadas numa base de dados *SQL* bastante simples, para termos um suporte local com as nossas informações acessíveis e congruentes em caso do web-site original ser alterado ou ficar *offline*.

O *parsing* foi feito percorrendo a lista de plantas presentes no web-site uma a uma recolhendo as referências html dessa lista. As plantas com menos informações disponíveis foram automaticamente descartadas, visto que queríamos apenas ficar com plantas com informações associadas. No final do *parsing* obtemos 798 plantas que serão tratadas nos seguintes capítulos.

Algumas informações inconsistentes, como por exemplo nomes diferentes para a mesma doença, foram tratadas manualmente de forma a obtermos um *website* mais fiável. No entanto o tempo era escasso e não nos foi possível tratar toda a informação presente na base de dados, existindo assim alguns casos com informação redundante ou com várias designações para a mesma doença.

3. Ontologia

Para criar a ontologia, que não é mais que o modelo que representa o conhecimento utilizando um vocabulário partilhado para denominar tipos de conceitos juntamente com as relações e propriedades entre os mesmos utilizámos o Protégé.

A ontologia criada no contexto do projeto relaciona os conceitos de:

- Classificação científica de uma planta com os nomes científicos para:
  - Família
  - Género
  - Espécie
- Doenças tratáveis recorrendo a plantas,
- Propriedades médicas de plantas
- Partes de plantas.

Estes conceitos são, neste modelo, relacionados entre si através das seguintes relações:

- *HasClassification* - relaciona plantas com a sua classificação científica
- *HasTreatmentFor* - relaciona plantas com doenças que esta pode tratar
- *HasMedicalProperty*- relaciona plantas com propriedades medicinais
- *HasEdiblePart* - relaciona plantas com as partes que a constituem que são comestíveis
- *HasPlantPart* - relaciona plantas com as partes que a constituem e que são utilizadas no tratamento.
- *HasDescription* – relaciona plantas com um texto que descreve a sua aparência.

Todos os conceitos e todas as relações têm associadas *labels* que contêm termos de linguagem natural relacionados com cada conceito ou relação, por exemplo a relação *HasTreatmentFor* tem como *labels* termos como '*treats*', '*treat*' e '*diseases*' e o as entidades do tipo *Diseases* podem ter como *label* termos como '*Cancer*' ou '*Fever*'. Este tipo de *labels* permitirá mais tarde efectuar pesquisas semânticas mais avançadas.

A figura seguinte ilustra o diagrama gerado para a ontologia.

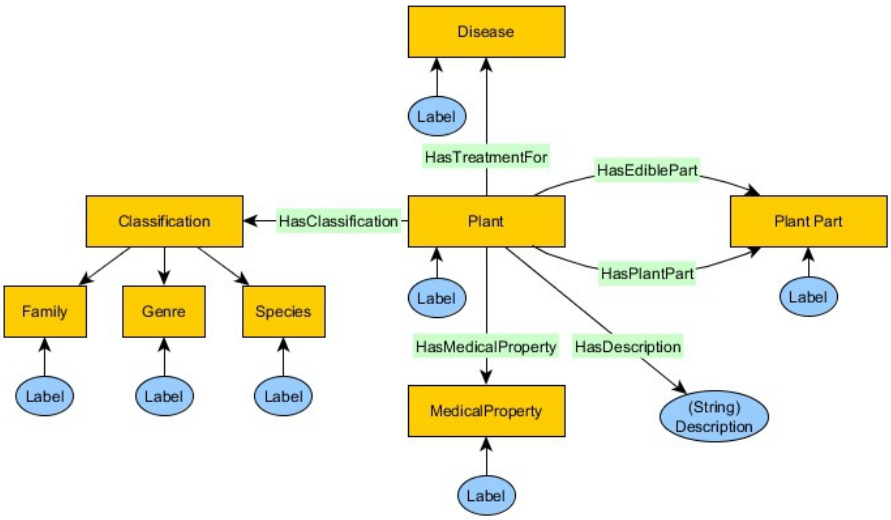

Figura 1 – Diagrama da ontologia utilizada para o projeto

4. Extração de triplos

Em web semântica o conhecimento é representado por triplos. Um triplo é uma entidade composta por sujeito-predicado-objecto e todo o tipo de relações são triplos. No presente projeto um exemplo de triplo é ”'plant52' hasTreatmentFor 'Disease54'“.

Para criar o nosso triplestore utilizámos a biblioteca Jena para a linguagem Java. Com esta biblioteca conseguimos criar o triplestore utilizando os dados previamente inseridos na base de dados SQL e o modelo criado no protege.

Para isto criamos uma rotina em Java que percorria os dados relativos a uma planta, presentes na base de dados e verificava através de uma query SparQL se as instâncias referentes àquela planta já tinham sido definidas ou não. Caso já tivessem sido definidas, o programa acrescentava a nova relação à instância, caso não tivessem sido definidas era criada uma nova instância com um novo identificador único (por exemplo Disease57, se a última inserida tivesse sido a Disease56) e era criada e a relação. Era também atríbuida a label correspondente à instância em causa (por exemplo Disease57 rdfs:label Nerves). Para cada classe definida no modelo foram definidas este tipo de instâncias.

No início do trabalho começámos por utilizar recursos para representar os indivíduos presentes na triplestore. No entanto, com o desenvolvimento do projeto verificámos que esta abordagem não era suficiente para representar o conhecimento.

Quando efetuávamos pesquisas por uma dada classe, por exemplo uma listagem de plantas de uma dada família não conseguíamos alcançar este tipo de conhecimento utilizando recursos. Foi aí que alterámos a nossa abordagem para a utilização de indivíduos.

Os indivíduos permitem referenciar as instâncias como sendo de uma determinada classe. Sendo assim já conseguíamos fazer pesquisas mais avançadas no *web-site* cingindo-nos por exemplo a uma família ou a um género de planta.

5. Browsing

Browsing é um tipo de estratégia para encontrar informação, normalmente utilizado num site, onde o utilizador navega através dos termos mais relevantes.

Para construção do web-site utilizado no nosso projeto utilizámos Java nomeadamente JSP para tratar da informação do lado do servidor. Para trabalhar o lado do cliente utilizámos diversas tecnologias, de entre elas destacam-se HTML, Jquery e Javascript.

No nosso projeto implementámos diversos tipos de *browsing*:

- Browsing por doença,
- Browsing por classificação
- Browsing por todos os termos relavantes contidos numa planta
- Browsing através de termos relacionados com a string utilizada na pesquisa semântica

Para fazer *browsing* por doenças utilizamos uma *query SparQL* que nos devolve as doenças existentes na *triplestore*. Como a quantidade de doenças é bastante elevada decidimos sub-dividi-las alfabeticamente para facilitar a interação com o utilizador.

O *browsing* por classificação é feito de uma forma um pouco diferente. O utilizador escolhe uma letra do alfabeto e é feita uma *query SparQL* que nos devolve as famílias de plantas que começam por essa letra. De seguida o utilizador escolhe uma família do menu e aparecem-nos os géneros dessa família. Escolhendo o utilizador um género são-lhe mostradas todas as plantas desse género.

Partido dos tipos de *browsing* implementados acima são mostradas as plantas em pequenas caixas que o utilizador pode clicar. Ao clicar numa planta as suas informações detalhadas são apresentadas, também estas informações são pesquisadas utilizando *queries* SparQL:

Figura 2 – Informações de uma dada planta

É a parir desde tipo de cenário que é possível ao utilizador efetuar pesquisa por termos relevantes contidos na planta. Na figura 2 seria possível navegar no web-site clicando em termos como *BASELLACEAE* (pesquisando plantas da mesma família), *Basella* (pesquisando plantas do mesmo género), *Abcesses* (pesquisando outras plantas que sirvam para tratar abcessos) ou *Diuretic* (pesquisando plantas com propriedades dietéticas).

Aquando da pesquisa semântica foi ainda implementado outro tipo de browsing, através de termos relacionados com a string utilizada na pesquisa semântica, por exemplo se o utilizador inserir na pesquisa semântica o termo *Diur* aparecerá como sugestão de *browsing* a palavra *Diuretic*.

6. Recomendação

Tendo toda a informação estruturada semanticamente é possível estruturar os web-sites de outra forma, uma das muitas vantagens que podemos extrair deste tipo de abordagem é a implementação de sistemas de recomendação baseados na informação semântica.

Para recomendar plantas similares às plantas visitadas pelo utilizador procedemos da seguinte forma:

Considerando que o utilizador está a ver uma dada planta, procuramos na base de dados de triplos todas as plantas com as mesmas propriedades medicinais ou com as mesmas doenças, ou da mesma família ou género do que a planta em vista. Estas plantas são colocadas numa lista e a cada uma destas propriedades é atribuída uma pontuação:

- Propriedades Medicinais : 2 pontos
- Doenças : 1 ponto
- Família : 2 pontos
- Género : 3 pontos

Da lista de recomendação criada são escolhidas as 6 plantas com pontuação mais alta e são mostradas ao utilizador as razões para a sua recomendação.

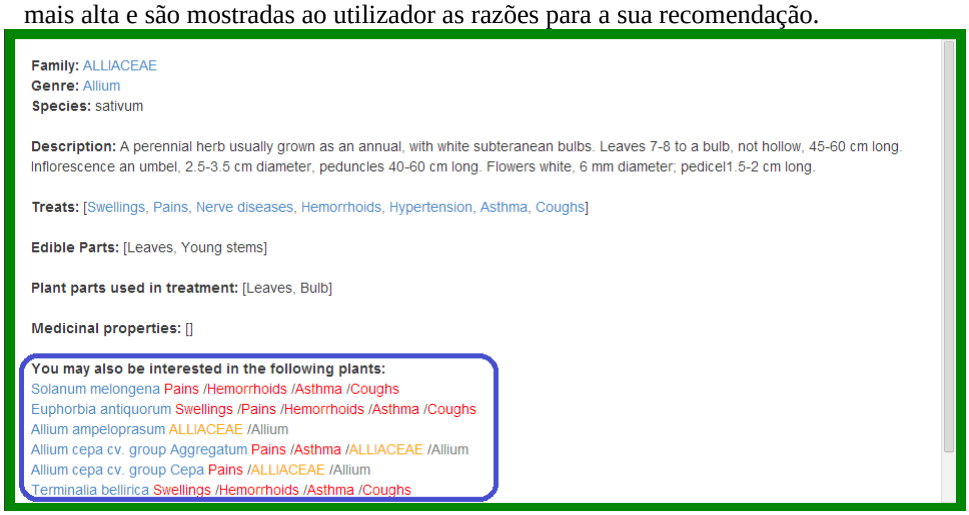

Figura 3 – Informações de uma dada planta

Na figura 3 podemos ver dentro da caixa azul, as plantas recomendadas. As diferentes cores representam as diferentes causas da recomendação. Vermelho é utilizado para doenças, Laranja para a família, cinzento para o Género e verde para propriedades medicinais.

7. Pesquisa Semântica

Pesquisa semântica é a estratégia de pesquisa de informação mais livre que foi implementado. Aqui o utilizador é livre de escrever uma query ao website que este processa e retorna plantas que estão relacionadas com a query efectuada.

A query que o utilizador insere na barra de pesquisa é composta por palavras-chave relacionadas com a informação que pretende obter. Por exemplo se um utilizador estiver interessado em obter informação sobre plantas cujas folhas tratem doenças de pele terá que inserir a query ‘leaves treat skin’, o website encarrega-se depois encontrar todas as plantas que o façam.

O algoritmo utilizado para esta pesquisa começa por isolar todas as palavras-chave que o utilizador inseriu e tenta descobrir qual a relação que estas palavras têm

com o nosso triplestore. Isto é feito através de uma query sparql cujo resultado é o tipo de relação.

Posteriormente o algoritmo constrói uma nova query sparql que pesquisa na base de triplos por plantas que estejam relacionadas com os conceitos contidos na query inserida. Esta query é composta por:

- Um header estático contendo os prefixos a utilizar na query
- Um comando select dinâmico que define todos os outputs que vão ser utilizados
- Um corpo de query composto por uma pilha de ‘templates’ de query

Cada um destes ‘templates’ é um excerto do corpo da mensagem que implementa a restrição imposta por uma palavra-chave. A estrutura de cada template depende unicamente do tipo de relação que cada palavra tem com o triplestore e é suficiente para, em conjunto com os restantes elementos da query, construir uma mensagem válida (usado sempre que se insere uma só palavra-chave). Estes templates foram também construídos de forma a ser possível integrar N (com N igual ao numero de palavras inseridas pelo utilizador) ‘templates’ numa query de forma a relacionar todas as palavras.

A query não só suporta palavras existentes na estrutura de triplos, mas também partes de palavras. Isto foi conseguido implementando REGEX, que está embebido na query sparql, a cada palavra-chave.

Como o resultado desta query pode não ser só plantas relacionadas estritamente com as palavras-chave, mas sim plantas relacionadas com palavras que contêm as palavras-chave, a cada planta recomendada está associada a razão pela recomendação na forma de conjuntos relação/palavra.

Na figura 4 está apresentado o resultado de uma pesquisa semântica para as palavras ‘leaves treat skin’. Daqui podemos verificar que as plantas não estão só relacionadas não só com a palavra ‘skin’ mas com outras palavras ou composições de palavras.

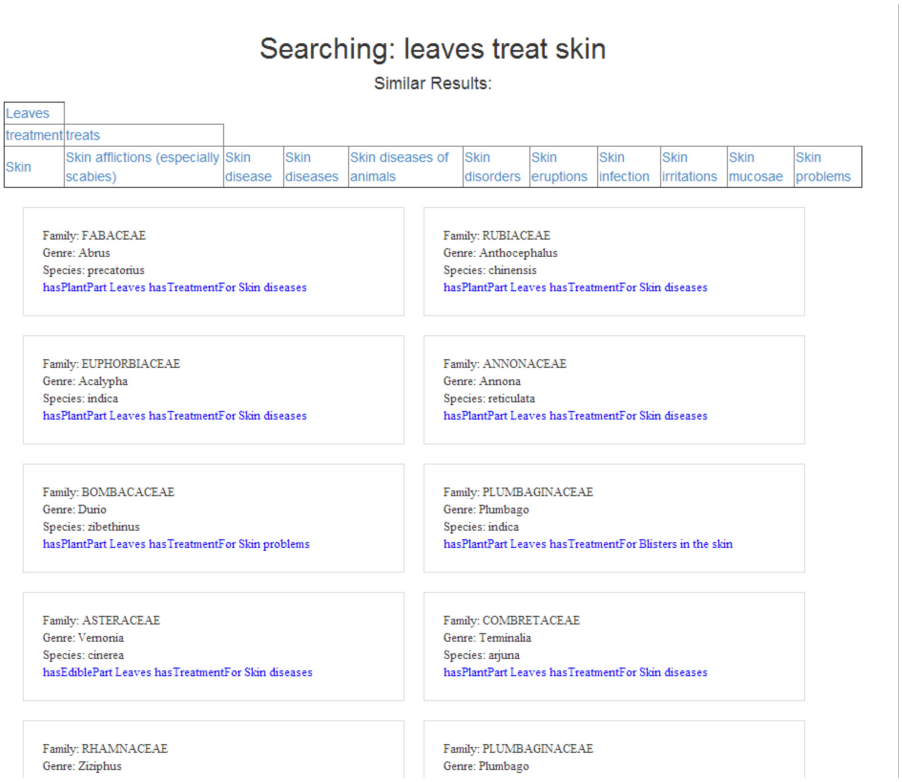

Figura 4 – Exemplo de pesquisa semântica

8. Conclusão

Com a realização deste projeto, foi-nos possível reconhecer a importância da Web Semântica na atualidade. Com a quantidade de informação que existe disponível na Internet existe a necessidade de pensar em aplicações que permitam, fazendo uso deste tipo de tecnologias para que responder de forma acertada a algumas pesquisas, ou otimizar aplicações já existentes melhorando a interface com o utilizador ou a forma como as informações são armazenadas.

Adquirir conhecimentos de web-semântica é assim uma mais valia para o mundo atual, visto que se podem criar aplicações bastante mais eficientes do que as já existentes.

Fazendo um balanço final acerca do projeto, cumprimos todas as metas propostas e todos os objetivos delineados e conseguimos também, de forma

autónoma, aprender todos os conceitos necessários para conseguir fazer o projeto com êxito.

As diferentes metas do trabalho foram de encontro aos objetivos estabelecidas. Para a realização do trabalho utilizámos ferramentas como o Protégé, SPARQL e TDB. Utilizámos também ferramentas e tecnologias auxiliares, tais como MySQL, técnicas de Screen Scrapping utilizando JAVA e expressões regulares e, no desenvolvimento WEB: HTML, JSPs, Jquery e Javascript.

Referências

Ayurvedic Medicinal Plants Of Sri Lanka:
"http://www.instituteofayurveda.org/plants/plants_list.php?s=Scientific_name". visitado em 19/01/2014.