ОБРАБОТКА WEB-ИНФОРМАЦИИ НА ОСНОВЕ МОДЕЛИ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

А.А. Селиванова

Белорусский государственный университет информатики и радиоэлектроники,
кафедра программного обеспечения информационных технологий
220013, ул.П.Бровки,6, г.Минск, Республика Беларусь
+375 29 7973569; e-mail: asteria_87@mail.ru

Рассмотрены информационные технологии, применяемые для поиска и анализа web-информации различного профиля. Предложен метод обработки медицинской web-информации на основе модели представления знаний. В качестве модели представления знаний выбрана иерархическая семантическая сеть. Рассмотрены вопросы извлечения знаний из web-ресурсов.

Ключевые слова: web-информация, семантическая сеть, медицинская диагностика, база знаний.

1 ВОЗМОЖНОСТИ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ WEB-ИНФОРМАЦИИ

Internet является одним из основных источников информационных ресурсов, наряду с традиционными. Однако в нем содержится огромное количество ненужных и повторяющихся данных. Поэтому постоянное изменение и увеличение информации в Internet требует структурирования и систематизации для ее эффективного использования, как обычным пользователем всемирной сети, так и специалистами различных предметных областей, осуществляющих поиск и анализ необходимой информации. Становится более очевидным отсутствие эффективных методов извлечения и формализации знаний из электронных документов и web-страниц, для дальнейшего, с учетом смысла, анализа. Одной из наиболее актуальных проблем обработки web-информации является обеспечение эффективного и быстрого поиска практически полезных знаний, необходимых для принятия решений в различных сферах человеческой деятельности [1].

Наиболее распространенные поисковые системы используют в основном количественные методы и алгоритмы анализа web-ресурсов, не используя качественные методы, которые производят анализ метаданных с информационным наполнением страницы. Они зачастую не обеспечивают адекватного выбора информации по запросу пользователя.

Одним из решений этой проблемы является использование средств и методов искусственного интеллекта для представления и определения метаданных, описывающих ресурсы web. А именно, использование моделей представления знаний, таких как семантические сети и различные их вариации.

Такой подход обеспечит не только синтаксический, но семантический анализ web-информации. Также при использовании различных модулей для работы с семантической сетью возможно решение таких задач, как формализация, интеграция, обмен знаниями и их повторное использование [2].

2 ОБРАБОТКА МЕДИЦИНСКОЙ WEB-ИНФОРМАЦИИ НА ОСНОВЕ СЕМАНТИЧЕСКОЙ СЕТИ

Использование иерархической семантической сети для обработки web-информации, относящейся к конкретной предметной области, рассмотрим на примере модели семантической сети для медицинской диагностики. Такие online-системы сегодня весьма популярны среди пользователей Internet.

Знания любой предметной области отражаются в модели представления знаний. Семантическая сеть - это граф, вершины которого соответствуют объектам или понятиям, а дуги, связывающие вершины, определяют отношения между ними. В семантических сетях продукционные правила представлены наиболее приближенно к мыслительному процессу человека и позволяют [3]:

добавлять или удалять фрагменты сетей;
добавлять или удалять связи или вершины;
проверять, что некоторый элемент содержится в сети;
находить элементы, общие для двух и более сетей.

Структура семантической сети может быть различной. В предлагаемой модели вершинами графа являются такие объекты, как симптомы заболевания, синдромы (объединенные какой-либо характеристикой группы симптомов, определяющие возможное заболевание), заболевания (факт существования у больного патологического процесса), результаты обследований и диагнозы. Кроме того, в силу специфики предметной области, наличия классификации заболеваний и симптомов, каждая вершина может иметь иерархическую структуру. Классификации заболеваний или симптомов представляет собой дерево и помещается в иерархическую вершину семантической сети.

В сети имеются различные виды n-арных отношений, отношений, связывающих более двух понятий. Основной вид связи между вершинами графа для данной модели – связь «Является признаком». Например, симптом кашель

– является признаком – заболевания верхних дыхательных путей. Таким образом, связь «является признаком» подчиненная связь между симптомами и заболеваниями, симптомами и синдромами, синдромами и заболеваниями, заболеваниями и диагнозом. Для обозначения обратных отношений используется связь «имеет признак». Например, заболевание астма – имеет признак – затрудненное дыхание. Логическое отношение «И» – связь между вершинами одного вида. А также связи внутри иерархии именуются как «элемент класса» («a kind of»).

В иерархической семантической сети, таким образом, есть возможность делить сеть на подсети и устанавливать связи не только между вершинами, но и между фрагментами сети. Различные фрагменты сети упорядочены в виде дерева, где вершины – это отдельные подсети, а дуги – это отношения «видимости». Это позволяет значительно сократить время поиска решения по сети, не включая в поиск подсети, не связанные отношениями видимости.

Семантическая сеть работает с помощью правил, которые управляют переходами между вершинами по дугам сети, связывая условия и действия. Правила объединяют элементы сети, используя типы связей между ними, так, они строят логические цепочки утверждений. Например, правило звучит «Если у пациента имеется сильный кашель, то возможно у него бронхит». Правила состоят из связок «если-то», из условия, или нескольких условий, из вывода (нескольких выводов) и их вероятности. Ниже описан способ представления правил в предлагаемой модели базы знаний.

<правило>::=(ЕСЛИ<условие>ТО<вероятность> <действие>)

<условие>::=<утверждение>

<утверждение>::=<логическое утверждение>|

<описательное утверждение>

<логическое утверждение>::=(И{<утверждение>}*)|

({ИЛИ<утверждение>}*)|(НЕТ{<утверждение>}*)

<действие>::=<описательное утверждение>

Для учета неопределенности используется коэффициент определенности утверждений, который целесообразно вычислять по правилам нечеткой логики. При работе с сетью он может указываться как для исходных данных, так и для вывода утверждений. Например, «Если у пациента слабые хрипы в груди, то возможно заболевание легкая форма астмы». С помощью нечетких множеств наиболее полно можно определить степень проявления симптомов, что существенно влияет на выходные данные, а именно, постановку диагноза, уточняя его и отбрасывая неподходящие варианты, полученные при поиске.

Такая семантическая сеть позволяет провести эффективный и быстрый поиск и смысловой анализ даже больших объемов медицинской информации. По аналогии с областью медицинской диагностики, такую семантическую сеть можно использовать для обработки web-информации из других предметных областей не менее эффективно.

3 ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ WEB-РЕСУРСОВ

Одним из основных аспектов извлечения знаний из web-информации являются онтологии, как средство построения распределенных и неоднородных систем баз знаний на основе Интернет. Онтология представляет собой формальное, явное описание понятий предметной области и отношений между ними, а также правила для составления новых понятий и отношений. Формально записанные знания в онтологии составляют семантическую основу – базу знаний, для компьютерного анализа информации [1].

Четко определенный семантический базис предметной области, позволяет организовать более "осмысленный" анализ информации в электронных документах. Любые естественно-языковые конструкции, с помощью которых может выражаться та или иная информация, содержат в явном или неявном виде предмет обсуждения, семантическую идентификацию которого можно осуществить благодаря наличию онтологии предметной области. Описание онтологий и метаданных web-страниц и содержится в иерархической семантической сети.

Предложенная модель представления знаний реализована в работе в виде распределенной экспертной системы. В подобной системе процесс обработки web-информации может состоять из следующих этапов: поиск информации по предметной области при помощи семантической сети, анализ связей в сети с другими элементами, выявление знаний, поиск похожих документов в Internet, что может быть полезно при различных видах деятельности пользователей.

Основными достоинствами семантических сетей при обработке web-информации являются: качественный, эффективный и соответственно более точный анализ web-страниц, хорошая структурированность данных и возможность поиска документов из конкретной предметной области по смыслу.

ЛИТЕРАТУРА

[1] Применение онтологий при создании электронных образовательных ресурсов./ В.И. Аверченков [и др.] // Известие Орел ГТУ. – 2006. № 1. – С. 6–11.

[2] Средства информационного поиска и навигации в интернет: опыт развития языковых технологий./ А.Е. Ермаков, В.В. Плешко// Казань: Отечество – 2003.

[3] Попов Э.В. Статические и динамические экспертные системы / Э.В.Попов [и др.]// – М.: Финансы и статистика. – 1996. – 320с.