

Средства построения сайтов с семантически структурированным контентом

Кот И. В.

Кафедра интеллектуальных информационных технологий
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
e-mail: ivancat@gmail.com

Аннотация—В работе рассматриваются средства построения сайтов с семантически структурированным контентом.

Ключевые слова: интеллектуальная система; база знаний; семантически структурированный контент; средства построения.

I. ВВЕДЕНИЕ

За последние годы объем данных, доступный для пользователей в сети Интернет значительно увеличился, и продолжает расти. Естественным следствием данного процесса является то, что появляются все новые и новые способы структурирования данных, основными причинами для таких тенденций является желание пользователей удобно перемещаться по статьям, размещенным в сети, а также легко находить именно то, что нужно. Наиболее распространенными подходами на данном этапе являются следующие: подход, основанный на использовании предварительно заданных ключевых слов статьи (т.н. тэгов), и подход, основанный на выделении значимых слов (наиболее распространен в поисковых системах). Тем не менее, ни один из данных подходов не решает поставленную задачу полностью, т.к. в первом случае ключевые слова могут быть не заданы или заданы неверно, а во втором могут быть неверно определены в силу несовершенства используемых алгоритмов.

Таким образом, если предположить, что существует система, которая позволила бы автоматически структурировать контент, вносимый в нее, то это не только бы решило поставленную задачу, но и позволило бы значительно улучшить качество предоставляемого контента. Такой системой могла бы стать веб-система, построенная с использованием базы знаний, т.к. в этом случае достаточно легко можно было бы не только хранить, но и обрабатывать данные (контент веб-системы), при этом система могла осуществлять т.н. интеллектуальный поиск, т.е. отвечать на поставленный пользователем вопрос, а не просто искать вхождение каких-либо ключевых слов [1].

II. СРЕДСТВА ПОСТРОЕНИЯ САЙТОВ С СЕМАНТИЧЕСКИ СТРУКТУРИРОВАННЫМ КОНТЕНТОМ

В предлагаемом подходе к построению сайтов с семантически структурированным контентом рассматривается решение, построенное на уже существующих разработках в данной области. В качестве основы можно рассматривать RDF-

хранилище, где все статьи, предоставляемые ресурсом, размечены определенным образом. Основным отличием в данном случае будет способ хранения информации: в отличие от классических способов все статьи ресурса будут храниться в базе знаний или т.н. sc-хранилище, представляющем из себя графодинамическому sc-память.

Поскольку в данном случае в качестве хранилища используется sc-память, то для нахождения той или иной статьи возможно использования интеллектуального решателя задач.

Аналогично семантической паутине, для доступа к ресурсам будут использоваться унифицированные идентификаторы ресурсов (URI), а также семантические сети и онтологии.

Традиционно в Интернете URI-идентификаторы используются для установки ссылок на адресуемый объект. В семантической паутине URI используются также для именования объектов, то есть каждый URI однозначно называет некоторый объект. Свои URI есть не только у статей, но и у объектов реального мира, и даже у абстрактных понятий. Поскольку URI глобально уникальны, они позволяют называть одни и те же предметы в разных местах в сети. Что касается второго пункта, использования онтологий, утверждения, кодируемые с помощью специального языка разметки, в дальнейшем можно интерпретировать с помощью онтологий, созданных по стандартам (например, RDF Schema и OWL), чтобы получать из них логические заключения.

Таким образом, одним из средств построения сайтов с семантически структурированным контентом, может стать использование специальной разметки для каждой из статей, хранение их в базе знаний с использованием онтологии и использование решателя задач для осуществления поиска[3].

III. ПРИМЕР САЙТА С СЕМАНТИЧЕСКИ СТРУКТУРИРОВАННЫМ КОНТЕНТОМ

В качестве примера сайта, контент которого структурирован семантически, можно рассмотреть уже существующий веб-ресурс проекта OSTIS, <http://ostis.net>.

Все статьи на сайте разделены на несколько уровней, т.н. разделов, т.е. каждый из уровней имеет связь со своим надразделом и подразделом соответственно, что значительно облегчает машинную обработку [2].

Помимо иерархических связей, на сайте также используются перекрестные ссылки между статьями каждого из разделов, что автоматически трансформирует изначально-иерархическую структуру

в сеть, узлами которой являются статьи, а связями - ссылки.

Поскольку любую сеть можно представить в виде графа, то для машинной обработки можно использовать алгоритмы, применяемые для обхода любой графовой структуры: обход в глубину и обход по уровням.

Еще одним достоинством такой организации информации является возможность использования интеллектуального решателя, разрабатываемого в рамках проекта OSTIS, для нахождения необходимой информации. А поскольку контент уже семантически структурирован, то осуществить его перенос в базу знаний достаточно легко.

Графически структуру сайта можно представить следующим образом:

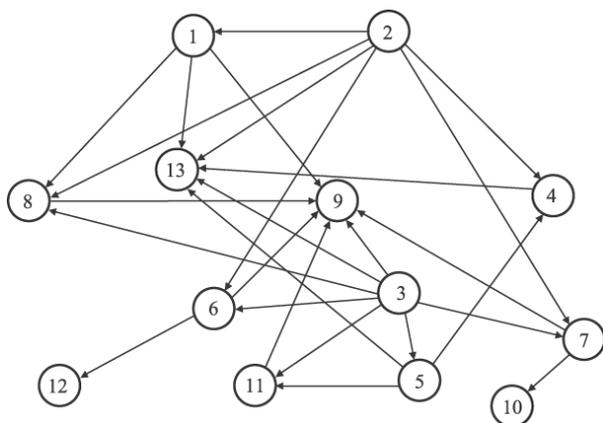


Рисунок 1. Представление структуры сайта в виде графа

Нужно отметить, что одного только наличия ссылок на подраздел, подраздел, а также смежные статьи недостаточно, поскольку семантика таких ссылок не определена соответствующим образом, другими словами, при машинной обработке нельзя определить, какой тип имеет данная ссылка. С этой целью на сайте введен специальный язык wiki-разметки, позволяющий явно указывать, к какому типу принадлежит тот или иной элемент, а также указывать связи между различными элементами базы знаний (в данном случае, статьями).

В дальнейшем процессор интерпретирует разметку превращая ее в набор тегов с определенной семантикой и заданным форматом отображения, что значительно упрощает последующую машинную обработку текста статьи, равно как и выделения связей с другими статьями.

Вид сайта для конечного пользователя указан на рисунках 2 и 3. Как видно, в данном случае пользователю, как и машине, достаточно легко ориентироваться в содержимом, поскольку ссылки на соответствующие направления четко обозначены и оформлены графически. Таким образом, можно говорить не только о том, такой подход к структурированию контента полезен не только для автоматизации процесса обработки данных, но и значительно облегчает работу пользователя с системой. В дальнейшем это потенциально позволит перемещаться по уровням и статьям при помощи голосового управления, т.к. система сможет

определить, что нужно делать, например, если необходимо спуститься на уровень ниже или перейти к следующей статье.

IV. ЗАКЛЮЧЕНИЕ

В работе рассмотрены средства построения сайтов с семантически структурированным контентом, основанные на использовании базы знаний, а также использовании интеллектуального решателя задач для нахождения тех или иных данных.

Результаты, приведенные в работе, апробируются в рамках открытого проекта OSTIS [2].

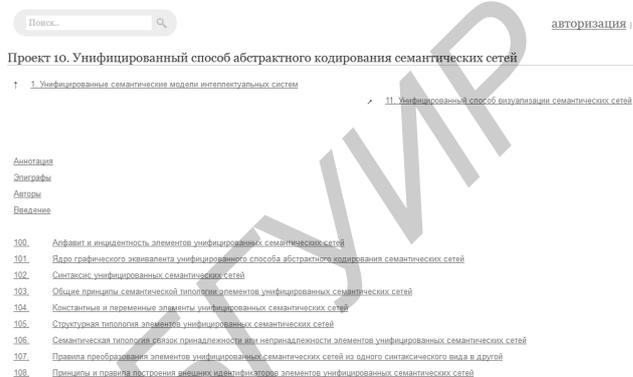


Рисунок 2. Вид заглавной страницы раздела документации сайта проекта OSTIS



Рисунок 3. Вид статьи, размещенной на сайте проекта OSTIS

[1] Голенков, В. В. Представление и обработка знаний в графодинамических ассоциативных машинах / Голенков В. В. [и др.]; под ред. В. В. Голенкова – Минск, БГУИР, 2001. – 412с.
 [2] Проект OSTIS [Электронный ресурс]. Минск, 2012. – Режим доступа: <http://ostis.net/>. – Дата доступа: 05.10.2012
 [3] Википедия [Электронный ресурс]. Минск, 2012. – Режим доступа: <http://wikipedia.org>. – Дата доступа: 05.10.2011