В статті визначено, які семантичні технології слід використовувати для більш докладного опису контенту web-сайтів, що дозволяє пошуковим машинам видавати більш релевантну інформацію користувачеві. Особливу увагу приділено створенню методології впровадження семантики у web-сайт

Ключові слова: методологія, семантика, пошукові машини, web-сайт

В статье определено, какие семантические технологии следует использовать для более подробного описания контента web-сайтов, что позволяет поисковым машинам выдавать более релевантную информацию пользователю. Особое внимание уделено созданию методологии внедрения семантики в web-сайт

Ключевые слова: методология, семантика, поисковые машины, web-сайт

The article defines semantic technologies which should be used for a more detailed description of web-site content that allows search engines to return more relevant information to a user. Particular attention is paid to the creation of the methodology of embedding semantics into the web-site

Keywords: methodology, semantics, search engines, web-site

УДК 004.75:681.3

МЕТОДОЛОГИЯ ВНЕДРЕНИЯ СЕМАНТИКИ В WEB-САЙТ

А.В. Витько
Кандидат технических наук, доцент*
Контактный тел.: (057) 702-13-37
E-mail: alexandra_vitko@yahoo.com

Л.Р. Стрелец*
*Кафедра искусственного интеллекта
Контактный тел.: 093-979-99-01
E-mail: milastrelec@yahoo.com
Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166

1. Введение

Интернет – это крупнейший из когда-либо существовавших информационных репозиториев, его содержание постоянно растет и представлено на самых разнообразных языках и практически во всех областях знаний [1]. Но, в конечном счете, становится все труднее находить смысл во всем этом содержимом. Поисковые системы способны находить информацию, содержащую определенные слова, но эта информация не всегда оказывается именно той, что требуется. Поиск основан на содержании страниц, но не на семантическом значении этого содержания или информации о странице.

Неотъемлемой частью сегодняшней всемирной сети является SEO-продвижение и оптимизация сайтов для поисковых систем. Анализ и изучение статистических данных посещаемости и аудитории сайта, постоянное расширение списка ключевых запросов, увеличение контента и ссылочной массы на сайт позволяет существенно повысить посещаемость ресурса и финансовых доход сайта. Для более грамотного поискового продвижения сайтов используется семантическое ядро, которое позволяет более качественно описать контент с помощью ключевых слов и правильно оптимизировать сайт для индексации поисковыми системами.

Для того, чтобы сделать интернет более практичным, нужно структурировать данные, тем самым повысить их ценность. Благодаря постоянной структуре, они могут быть использованы во многих отношениях. Структурирование и правильная разметка страниц становится возможным с использованием семантических технологий.

Данная работа посвящена исследованию концепции SemanticWeb и его элементов, существующих поисковых систем и их алгоритмов индексации вебстраниц, а также разработке методологии внедрения семантики в web-сайт.

2. Постановка задачи

Для быстрого поиска информации в Интернет разработаны специальные программы, которые по заданным адресам и ссылкам мгновенно отыскивают нужную информацию. При этом число обработанных информационных ресурсов может достигать сотен тысяч.

Недостатки нынешнего представления и разметки информации на большинстве сайтов:

– HTML теги не несут семантической нагрузки;

– процент полезной информации меньше процента разметки;

– разметка, в том числе, из-за большой сложности и вложенности (например, проблема табличной верстки) содержит много ошибок;

– машины не понимают и, следовательно, не анализируют смысл информации;

– поиск неудобен и сложен, часто результаты неудовлетворительны и не релевантны;

– SEO-оптимизаторы умышленно используют разметку несоответствующую правилам для кратковременного эффекта высоких позиций в поисковых запросах [2].

Исправление этих недостатков и коррекция нынешней системы представления информации возможна за счет внедрения семантических технологий в web-сайт.

В работе необходимо разработать методологию внедрения семантических технологий в web-сайт. Спроектировать архитектуру web-сайта с внедренной семантикой и взаимодействия с ней. Рассмотреть семантические технологии и выделить некоторые из них для построения архитектуры web-сайта. Исследовать принципы работы поисковых систем и рассмотреть возможности более грамотного поискового продвижения сайтов с помощью семантического ядра.

3. Результаты исследования

SemanticWeb дает возможность пользователям получить качественный результат поиска и, в свою очередь, дает возможность владельцам сайтов получить больше целевого трафика, благодаря тому, что пользователи действительно находят полезную и нужную для них информацию [3].

Структурирование данных повышает ценность данных. Благодаря постоянной структуре, они могут быть использованы во многих отношениях.

Для того, чтобы внедрить семантику в веб-сайт, следует воспользоваться семантическими веб-технологиями. Одной из сложных частей является разработка онтологии, которая соответствует данным. Онтология, как правило, один из важнейших элементов успешной реализации семантических веб-проектов [4].

Предложенная методология внедрения семантики в web-сайт включает в себя такие семантические технологии, как микроформаты, семантическое описание страницы с помощью RDF, онтологический движок и описание семантического ядра сайта для поисковой оптимизации.

Каждая из этих технологий позволяет в структурированном виде представить данные и повысить уровень соответствия запросам пользователей.

На рис. 1 представлена архитектура web-сайта с внедренной семантикой. Условно её можно разделить на три блока:

- блок создания контента;
- база знаний;
- блок внешних анализаторов контента.

Блок базы знаний состоит из трех составляющих:

- онтологический движок;
- TripleStore;
- база данных.

Онтологический движок – это основной обработчик данных на сайте. На вход онтологического движка поступает новый информационный ресурс, из которого он извлекает данные (ключевые слова, метаданные). Эти данные отправляет в TripleStore в формате .rdf и в базу данных.

TripleStore – база данных, в которой информация хранится в виде триплета «субъект-предикат-объект».

База данных выполняет функцию хранения текстов и ключевых слов соответствующих им. Все эти компоненты вместе создают базу знаний сайта.

Блок создания контента включает в себя семантические технологии, которые используются при создании нового материала или редактировании старого материала на сайте, используя любую CMS. Этот блок включает в себя два блока: микроформаты и семантическое ядро.

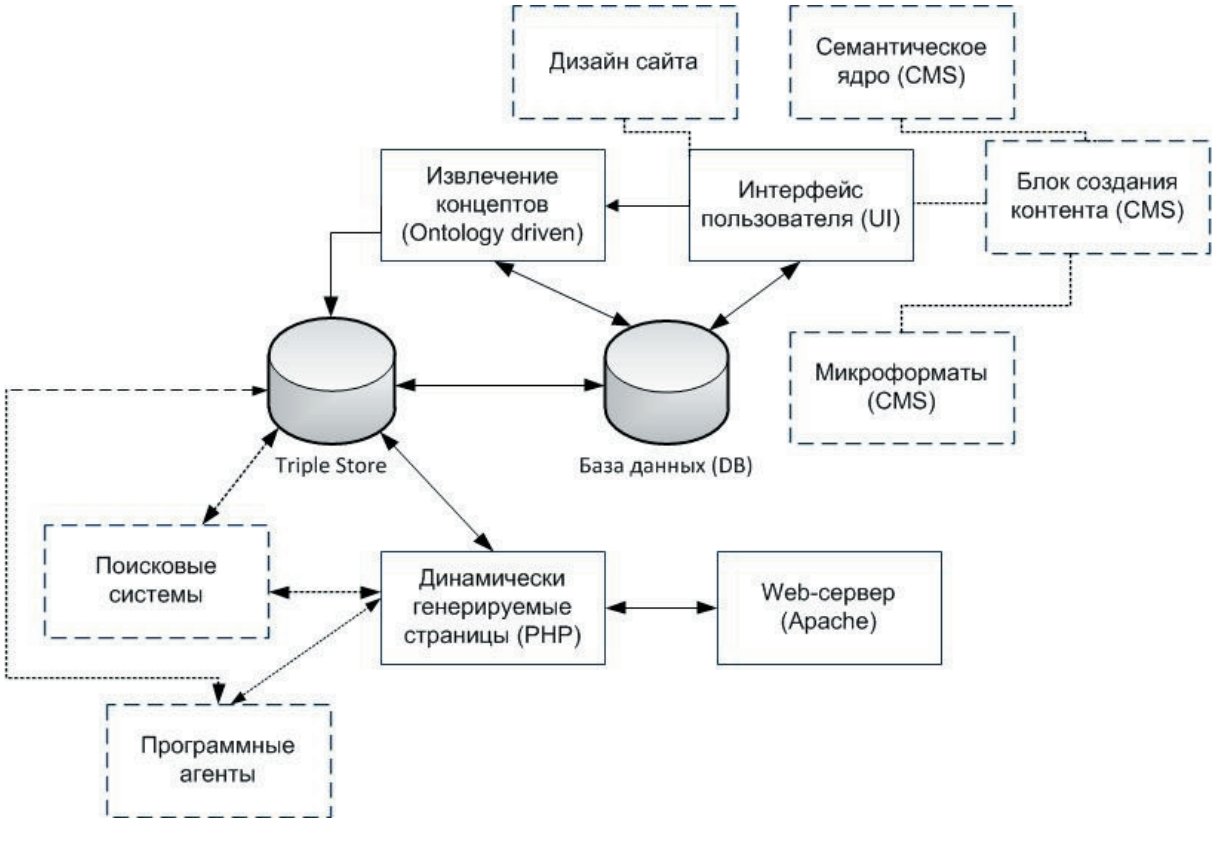

Рис. 1. Архитектура web-сайта с внедренной семантикой

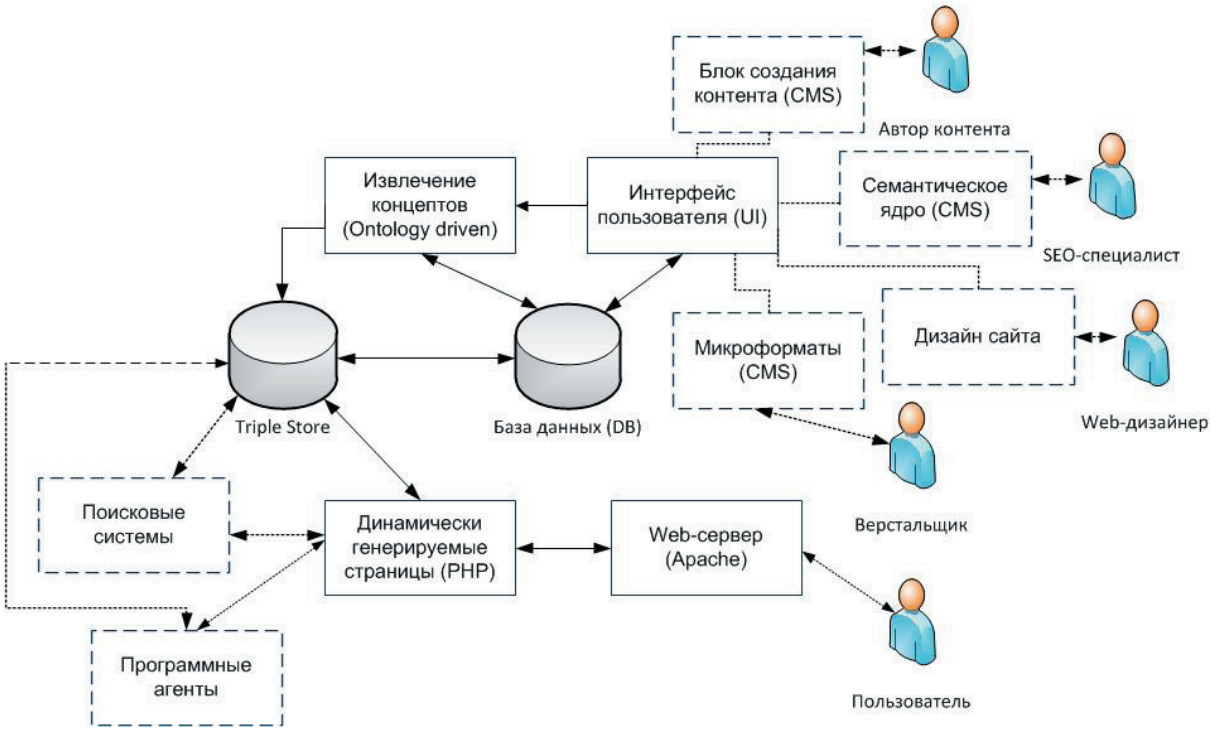

Рис. 2. Взаимодействие с семантическим сайтом

При создании статьи (либо другого материала) автор контента пишет текст. Для того, чтобы правильно разделить статью на информационные блоки, нужно использовать при верстке страницы микроформаты. Микроформаты позволяют представлять данные сайта в четко сформированном виде для поисковых машин. Это повышает вероятность правильной индексации контента.

Также на данном этапе создается семантическое ядро статьи. Оно подразумевает под собой описание статьи с помощью ключевых слов. Ключевые слова сайта определяются его тематикой. Для правильного составления семантического ядра сайта необходимо: определить необходимую словоформу; изъять из семантического ядра ключевые слова не дающие трафика.

С данной системой работают создатели контента, SEO-специалист, верстальщик, web-дизайнер. На рис. 2 представлено взаимодействие с семантическим сайтом.

Данная архитектура представлена в общем виде, так как при разработке нужно учитывать индивидуальные требования к сайту, которые выставляет заказчик. Но это архитектуру можно использовать как базовую и основную для различных типов web-сайтов.

4. Выводы

В статье был проведен анализ текущего состояния интернет. Выявлены недостатки представления данных на большинстве сайтов первого поколения web.

На основе теоретических исследований была разработана методология внедрения семантики в web-сайт. Она является общей и может быть использована как базовая методология внедрения семантики для различных типов web-сайтов. Также в статье спроектирована архитектура web-сайта исходя из предложенной методологии.

Данная архитектура позволяет за счет использования семантических технологий повысить уровень индексации сайта поисковыми системами, дать возможность использования данных другими сайтами и корпорациями, дает возможность динамического наполнения web-страниц.

Литература

1. WorldWideWebConsortium, [электронный ресурс] – Режим доступа: \www/ URL: http://www.w3.org/ – Загл. с экрана.
2. Рябов, В.А., Несвижский, А.И. Современные веб-технологии, [электронный ресурс] – Режим доступа: \www/ URL: http://www.intuit.ru/department/internet/mwebtech/1/ – Загл. с экрана.
3. Рябов, В.А., Несвижский, А.И. Семантические технологии и микроформаты, [электронный ресурс] – Режим доступа: \www/ URL: http://www.intuit.ru/department/internet/mwebtech/20/ – Загл. с экрана.
4. Рогушина, Ю.В. Технологии SemanticWeb и их использование при разработке интеллектуальных приложений [Текст] /Ю.В. Рогушина, А.Я. Гладун; Международный научно-учебный центр информационных технологий и систем НАН Украины и МОН;– К. : Университет, 2004. – 215с.