УДК 004.057.3:004.423.4

ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ВЕБ-СТРАНИЦ

Капитонов О.А., Тютюнник В.М.

ФГБОУ ВПО «Тамбовский государственный технический университет», Тамбов, Россия e-mail: vmt@tmb.ru

Рассмотрены основные направления развития алгоритмов и методов поиска информации в интернете, а также предложено направление развития семантической разметки веб-документа, которое позволит сделать веб-документ более удобным для обработки поисковыми системами. Для этого вводятся микроформаты, главная задача которых – сделать веб-страницы максимально удобными для семантического анализа методами Text Mining. Технология микроформатов позволяет с помощью тегов-контейнеров разметить веб-страницу, выделив элементы текста и обозначив их параметрами: предложение, слово, рисунок, таблица, формула, заголовок, абзац, коллокация, новый термин, часть речи, причинно-следственные связи и т.д. Приведён пример разметки части предложения, которая значительно повышает надёжность поиска профессиональной информации в интернете.

Ключевые слова: поиск информации, микроформаты, теги-контейнеры, семантическая разметка, анализ текста

LOGICAL-LINGUISTIC MODEL OF WEB-DOCUMENTS' SEMANTIC MARKING

Kapitonov O.A., Tyutyunnik V.M.

TambovStateTechnologyUniversity, Tambov, e-mail: vmt@tmb.ru

The basic ways of development of algorithms and information retrieval methods on the Internet are considered, and also the direction of development of a semantic marking of the web-document which will help us to make the web-document more convenient for processing by search engines is offered. Micro-formats are introduce for this purposes and for solving the main problem, that is to make web-pages maximal convenient for a semantic analysis by Text Mining methods. A micro-formats technology allow us to mark web-page with the help of tag-container, that is to pick out text's elements and to mark its by parameters: sentence, word, figure, table, formula, title, paragraph, collocation, new term, part of speech, causal relationship, etc. The example of the part of sentence' mark, which is more raising the reliability of professional information retrieval on the Internet, is given.

Keywords: information retrieval, micro-formats, semantic marking, text mining

Использование информационно-поисковых машин (ИПМ) приобретает огромную значимость при нахождении сведений, необходимых для решения задач оптимизации, анализа, принятия решений и управления практически во всех сферах деятельности [1]. В литературе [1–3] специалистами и рядовыми пользователями отмечается неудовлетворительное качество работы даже известных и широко распространённых ИПМ.

Рассмотрим основные методы, применяемые при поиске сведений в интернете.

Самый первый, простой и до сих пор распространённый – булева модель поиска. Она строится на теории множеств и математической логике. Её популярность связана прежде всего с простотой реализации, которая позволяет индексировать и выполнять поиск в больших документальных массивах. Эта модель формирует результаты поиска на основе факта присутствия тех или термов в документах. В 1983 г. Г. Солтоном (G. Salton), Э.А. Фоксом (E. Fox) и Г. Ву (H. Wu) [4] предложена расширенная булева модель, которая предполагала вычисление и использование весовых коэффициентов для каждого терма.

Ранее, в 1975 г., Г. Солтон предложил векторно-пространственную модель описания данных (Vector Space Model), которая послужила основой создания системы SMART [5]. В этой модели документу ставится в соответствие вектор в евклидовом пространстве, где каждому терму, использующемуся в документе, сопоставляется его весовое значение, которое определяется с помощью статистических методов. Запрос также представляется вектором в евклидовом пространстве термов. Соответствие документа запросу вычисляется как скалярное произведение их векторов.

В 1977 г. С.Э. Робертсон (S.E. Robertson) и К. Спарк-Джонс (K. Sparck-Jones) предложили вероятностную модель поиска [6]. В данной модели поиска вероятность того, что документ релевантен запросу, основывается на предположении, что термы запроса по-разному распределены среди релевантных и нерелевантных документов. При этом используются формулы расчета вероятности, базирующиеся на теореме Байеса.

На наш взгляд, векторные и вероятностные модели поиска информации хорошо проработаны и подошли к своему естественному пределу, особенно в больших

динамических объёмах документов. Однако исторически сложилось так, что максимальную эффективность показали алгоритмы, уделяющие внимания не тексту, а самому документу. Эти алгоритмы работают не с информационным наполнением сайта, а с его свойствами как интернет-ресурса. Они могут быть основаны на двух правилах [7]:

1. Если документ А ссылается на документ Б, то можно считать, что автор А рекомендует Б. То есть ранг веб-страницы тем выше, чем больше других страниц ссылается на неё и чем эта страница популярнее.

2. Если документы А и Б связаны гиперссылками, то вероятнее тот факт, что их тематика родственна, чем обратное утверждение.

Другой подход, базирующийся на связях между веб-страницами, – представление текстовой коллекции как многосвязного графа, по которому осуществляются переходы. Описав вероятности переходов, можно прийти к Марковской вероятностной модели. Эта модель рассматривает граф как матрицу вероятностей переходов между документами, это позволяет определить наиболее значимые документы в корпусе с точки зрения их способности к посещаемости [7].

Одним из первых таких алгоритмов был PageRank, положенный в основу известнейшей сегодня ИПМ Google. Фактически PageRank является не алгоритмом поиска, а алгоритмом ранжирования результатов. Сам поиск можно осуществлять, например, классической булевой моделью, а порядок выдачи результатов определять с помощью PageRank.

Сегодня практически нет ни одной ИПМ, которая не пользовалась бы информацией, получаемой на основе анализа кода ресурса.

На заре веб-технологий появились и так называемые логические теги – инструменты, предназначенные для упрощения задачи появления сведений. Например, тег <abbr> указывает на то, что данное слово – аббревиатура, элемент <acronym> – на появление акронима, а тег <address> – на то, что в нём прописан адрес, и т.д. Для того чтобы подчеркнуть семантическую значимость тех или иных слов, можно объявить их ключевыми, а в тексте выделить тегом . Однако такие механизмы стали терять свои позиции. Можно выделить несколько причин этого:

– отсутствие системного подхода во внедрении логических тегов;

– огромные сложности, возникающие при внесении каких-либо изменений;

– злоупотребление логическими тегами для продвижения сайтов в рейтингах ИПМ.

Другая модель семантической разметки была предложена в 2004 году Ч. Тантеком, она базировалась на понятии микроформатов – протоколов разметки, использующих стандартные теги для придания объектам, в них содержащимся, некоторой семантической нагрузки. Этот подход позволяет решить проблему внедрения новых элементов в язык гипертекстовой разметки, позволяя всем пользователям, знакомым с концепцией микроформатов, создавать свои спецификации к семантической разметке, регистрируя их на сайте сообщества microformats.

По словам создателя технологии Ч. Тантека, микроформаты представляют собой тропу к семантическому вебу. Но существующие на сегодняшний день спецификации разметки позволяют лишь упростить обработку некоторой информации. Так, например, микроформат hCard, использующийся для публикации контактной информации людей, компаний, организаций и адресов и призванный на более высоком уровне заменить логический тег <address>, даёт возможность ИПМ в автоматическом режиме корректно обрабатывать «визитные карточки» людей и организаций. Но для комплексного семантического анализа произвольных текстовых сведений необходимо совершенно новое семейство микроформатов.

Решение задачи семантического поиска не может быть осуществлено без технологии глубинного анализа текстов их смысла и представления его в базах знаний на основе методов искусственного интеллекта.

Это семейство технологий объединено общим названием Text Mining. Разрабатываемые на основе статистического и лингвистического анализа, методов искусственного интеллекта, эти технологии предназначены для проведения смыслового анализа.

Основное назначение Text Mining – выбирать из текстов наиболее ключевую и значимую информацию для пользователей [8, 9].

Рассмотрим основные задачи, решаемые в рамках технологий Text Mining:

– извлечение из документа элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотаций;

– классификация и кластеризация документов;

– автоматическое выявление прежде неизвестных связей в уже имеющихся данных;

– извлечение знаний из текстовых сведений;

– автоматическое реферирование текстов;

– выявление феноменов – понятий и фактов;

– создание принципиально нового вида поиска, который в отличие известных под-

ходов устанавливает не формальную релевантность – соответствие результатов поиска запросу, а позволяет осуществлять поиск с помощью смысла текстов, то есть искать сведения на основе пертинентности – соответствия информационной потребности.

Для глубокого семантического анализа текстов вполне справедливо определение, данное для предшествующего Text Mining семейства технологий Data Mining Г. Пятецким-Шапиро из GTE Labs: «Процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1].

Text Mining – относительно молодые технологии, появившиеся в девяностых годах XX века как направление анализа неструктурированных текстов. Технологии Text Mining сразу же взяли на вооружение такие методы Data Mining, как классификация или кластеризация. Современные системы на основе Text Mining используются для управления знаниями, выявления шаблонов в текстах, автоматического «проталкивания» и распределения информации на основе информационной потребности пользователя.

Однако при организации семантического анализа текстов встаёт ряд проблем несемантического характера, но препятствующих глубинному анализу текстовых сведений.

С нашей точки зрения, *задача микроформатов* состоит в том, чтобы сделать вебстраницы максимально удобными для семантического анализа методами Text Mining.

Подойдём системно к вопросу выявления и решения этих проблем. Начнём с макроуровня. На этом этапе рассмотрения вебстраница в общем виде представляет собой множество объектов: заголовки, текстовые сведения, рисунки, таблицы, указание на источники, на классификаторов информации, например, УДК. Все эти элементы так или иначе являются содержимым теговконтейнеров. Необходимо понимать, что семантическое содержание текста зависит от данных нетекстовых объектов, поэтому важно предусмотреть систему именования объекта. Это некий параметр, который указывается в открывающем теге-контейнере именуемой информации, и ему присваивается имя объекта с тем, чтобы мы всегда смогли обратиться к данному объекту. Назовём этот параметр *smf_name*.

Тип элемента (рисунок, таблица, формула, абзац, заголовок или же просто выделенный фрагмент) легко определяется с помощью тега-контейнера и его параметров.

Более низкий уровень рассмотрения текстовых сведений – это множество предложений. Для того чтобы произвести семантический анализ, необходимо осуществить синтаксический разбор предложений и словосочетаний. Эта задача решена большинством систем, работающих с текстовой информацией. Однако встречаются случаи, когда данные алгоритмы не в состоянии автоматически определить взаимосвязь слов в предложении или словосочетании (программы Office в этом случае предполагают, что предложение не согласовано). Это может происходить из-за несовершенства алгоритмов и сложности русского языка, из-за появления неизвестных системе слов, терминов и символов, которые человек трактует как подлежащее, сказуемое или другие члены предложения, а машина это сделать не в состоянии. Для решения данной проблемы необходимо семейство микроформатов, которое позволит недвусмысленно задать роли слов и других элементов в предложении, а также связи между ними. Естественно, нет необходимости для каждого слова текста определять его роль и связи. Возникает необходимость разработки системы, которая позволит выявить предложения и словосочетания, нуждающиеся в дополнительной синтаксической разметке.

На этом этапе необходимо уделить внимание таким элементам, как коллокации [10–11]. Существует несколько определений понятия коллокации. В корпусной лингвистике под коллокацией чаще всего понимают последовательность слов, которые встречаются вместе чаще, чем можно было бы ожидать исходя из случайности распределения [10]. Другое определение говорит о том, что коллокации отражают ограничения совместного использования слов, например, указывают, какие предлоги употребляются с данным глаголом [10]. Коллокация – статистически устойчивое сочетание в тексте. По степени устойчивости можно выделить свободные сочетания, связанные сочетания, идиомы [11].

Необходимо с помощью технологии микроформатов дать пользователю возможность выделить коллокации в тексте. Естественно, сделать это самому крайне сложно. Однако существует множество разработанных методов выявления коллокаций. Если на этапе создания веб-публикации выявить пары слов с подозрением на коллокации и предложить пользователю определить данные словосочетания, задача становится решаемой.

Спустимся теперь на уровень слов, из которых состоят текстовые сведения. На этом этапе рассмотрения основными проблемами являются синонимия и омонимия.

Кроме того, неудобство при семантическом анализе представляют неизвестные слова. Предложим полную классификацию неизвестных слов. Неизвестное слово может быть либо термином, либо именем собственным. Под терминами будем понимать неизвестные слова, аббревиатуры и прочее. Для обработки терминов удобно использовать микроданные itemscope, которые позволяют дать определение термину и впоследствии извлечь из этого определения информацию. Если мы хотим подчеркнуть, что это аббревиатура, то в качестве контейнера можно использовать тег <abbr>. Для разметки имён собственных необходимо указать на термин, определяющий множество, к которому принадлежит объект, а также на некоторую именованную область, где описываются его отличительные черты.

В качестве классического примера использования данных микроформатов рассмотрим отрывок предложения: «…F позволяет найти значение…». Его разметка будет иметь вид «… <span class = smf smf_rol = ”podlegachie” smf_type = “name = «function» URL = «http//…» smf_name = ”form_1”> F позволяет найти значение…». Поясним пример. Задан класс семантической разметки «class = smf», параметр smf_rol указывает на роль объекта в предложении (в данном случае «подлежащее») и также при необходимости может указать на связь между словами в предложении или словосочетании. Параметр smf_type содержит необходимую информацию для определения класса элементов, к которому принадлежит размеченная сущность; smf_name – параметр, указывающий на имя объекта, который задаёт выделенный фрагмент. В данном случае этот объект – формулы, вставленные в текст картинкой. Задание формул как вычислимых элементов также является важной задачей семантической разметки веб-страниц.

Дополнительная разметка: причинно-следственные связи и спец-символы.

Предложенная разметка позволит сделать веб-страницы прозрачными для глубокого семантического анализа. Веб-разработчику будет достаточно воспользоваться открытым ПО, которое поможет ему найти элементы, подлежащие разметке. После этой процедуры программное обеспечение сможет без труда произвести синтаксический разбор и определить каждое слово единственным образом. С нашей точки зрения именно этот подход позволит микроформатам проложить тропу для семантических методов контент-анализа (Text Mining) и стать настоящей дорогой в долгожданный семантический веб.

Список литературы

1. Козлов Д.Д. Информационно-поисковые системы в Internet: текущее состояние и пути развития. – М.: МГУ, 2000. – 18 с.

2. Ландэ Д.В., Санарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: ЛИБРОКОМ, 2009. – 264 с.

3. Ландэ Д.В. Поиск знаний в Internet. – М.: Диалектика-Вильямс, 2005. URL: http://poiskbook.kiev.ua (дата обращения 15.03.2012).

4. Недошивина Е.В. Учёт синтаксических связей при поиске коллокаций // Natural Language Processing. – 2008. – 3 с.

5. Федотов А.М., Барахнин В.Б. Проблемы поиска информации: история и технологии. – Новосибирск: Изд-во НГУ, 2008. – 18 с.

6. Шарапов Р.В., Шарапова Е.В., Торохова Е.А. Учёт гипертекстовых ссылок между документами при ранжировании результатов поиска. – Воронеж: Изд-во ВГУ, 2001. – С.4.

7. Ягунова Е.В., Пивоварова Л.М. От коллокаций к конструкциям // Тр. Ин-та лингвист. исследований РАН. – 2011. – 43 с.

8. Berry M.W. Survey of Text Mining. Clustering, Classification, and Retrieval. – Berlin: Springer-Verlag, 2004. – 244 p.

9. Robertson S.E., Jones K.S. Simple, proven approaches to text retrieval // Cambridge Technical Report. – 1997.

10. Salton G., Fox E., Wu H. Extended Boolean information retrieval // Communications of the ACM. – 2001. – Vol. 26, № 4. – P. 35–43.

11. Salton G., Wong A., Yang C.A. Vector Space Model for Automatic Indexing // Communications of the ACM. – 1975. – Vol.18, № 11. – P. 613–620.

References

1. Kozlov D.D. *Informitsionnopoiskovyesistemy v Internet: tekuscheesostoyanieiputirazvitiya* [Information retrieval systems into Internet: the current state and the way of the development]. Moscow, MSU Publ., 2000. 18 p.

2. Lande D.V., Sanarskiy A.A., Bezsudnov I.V. *Internetika: navigatsiivslozhnykhsetaykh: vodeliialgoritmy* [INTERNETIKA: Navigation in the complex networks: model and the algorithms]. Moscow , LIBROKOM, 2009. 264 p.

3. Lande D.V. Search for knowledge into Internet. M.: Dialectician-Williams, 2005. URL: http://poiskbook.kiev.ua.

4. Nedoshivina E.V. *Ucheotsintakticheskikhsvyazejpripoiskekollokatsij* [Calculation of syntactic connections with the search for collocation] // Natural Language Processing, 2008. 3 p.

5. Fedotov A.M., Barakhnin V.B. *Problemypoiskainformatsii: istoriyaitekhnologii* [Problems of retrieval for the information: history and the technology]. Novosibirsk, NSU Publ., 2008. 18 p.

6. Sharapov R.V., Sharapova E.V., Torokhova E.A. *Ucheotgipertekstovykhssylokmezhdudokumentamipriranzhirovaniirezultatovpoiska* [Calculation of hyper-text references between the documents with the ranking of the results of the search]. Voronezh, VSU Publ., 2001. pp. 4.

7. Yagunova E.V., Pivovarova L.M. *Otkollokatsijkkonstruktsiyam* [From collocation to the constructions] // Tr. Institute Linguist. Studies RAS, 2011. 43 p.

8. Berry M.W. Survey of Text Mining. Clustering, Classification, and Retrieval. Berlin, Springer-Verlag, 2004. 244 p.

9. Robertson S.E., Jones K.S. Simple, proven approaches to text retrieval // Cambridge Technical Report, 1997.

10. Salton G., Fox E., Wu H. Extended Boolean information retrieval / G.Salton, // Communications of the ACM. 2001. Vol.26, no. 4. pp. 35–43.

11. Salton G., Wong A., Yang C.A Vector Space Model for Automatic Indexing // Communications of the ACM. 1975. Vol.18, no. 11. pp. 613–620.

Рецензенты:

Сумин В.И., д.т.н., . профессор кафедры управления и информационно-технического обеспечения, Воронежский институт ФСИН России, г. Воронеж;

Сысоев В.А., д.т.н., профессор кафедры прикладной информатики, Тамбовский филиал Московского государственного университета культуры и искусств; г. Тамбов.

Работа поступила в редакцию 11.12.2012.