УДК 004.912

Н.Ф. Хайрова, Аджит Пратап Сингх Гаутам

Национальный технический университет «Харьковский политехнический институт», Харьков

ОСОБЕННОСТИ ЭКСТРАКЦИИ И ИДЕНТИФИКАЦИИ ЗНАНИЙ WEB–КОНТЕНТА

В статье рассматриваются особенности идентификации знаний веб-страниц. Предлагается технология Web Content Mining, включающая выделение основного контента страницы, извлечение смысловых понятий и контентный анализ, использующий регулярные выражения. Разработанные регулярные выражения позволяют выделить отношения таксономии и репрезентативности между концептами веб-страницы.

Ключевые слова: *идентификация знаний, Web Content Mining, регулярные выражения, таксономия, отношение репрезентативности.*

Введение

Одной из центральных задач обработки информационных потоков Веба становится извлечение и идентификация знаний, выраженных текстами HTML. Эта задача является основной для направления Web Content Mining – добыча, извлечение и интеграция используемых данных, информации и знаний из контента Веб страницы. Web Content Mining, являющийся аналогом интеллектуального анализа данных для реляционных баз данных, представляет собой одну из трех составляющих Web Mining, включающего так же Web Structure Mining и Web Usage Mining.

Главная задача Web Content Mining, заключающаяся в извлечении и добыче полезных знаний из Веб-страницы, достаточно сложна, так как структура Web-данных неоднородна и нестандартна [1]. Веб документ может содержать несколько типов данных, таких как текст, изображение, аудио, видео, метаданные и гиперссылки. Некоторые из них частично структурированные, такие как HTML-документы; некоторые более структурированные, такие как данные таблиц и баз данных, но большая часть информации хранится в неструктурированных текстовых данных [2].

Интерес к этой теме, рассматриваемой как решение задачи Веба и как задача Natural Language Processing (NLP), в настоящее время фокусируется в двух направлениях: извлечение фактографической информации и определение тональности текстов

1. Постановка задачи

Представление текстовой информация на веб-страницах имеет ряд особенностей, к которым, прежде всего следует отнести с одной стороны структурированность представленного контента, с другой стороны наличие всевозможной побочной информации, своего рода «информационного шума». Такие особенности организации текстовой информации, требуют наряду с использованием общих технологий идентификации знаний из текстов, разработки особых этапов обработки текстовой информации, представленной на веб-страницах.

2. Технология выделения основного контента

Первый предлингвистический этап лингвистического процессора задачи Web Content Mining предназначен для выделения основного текстового контента из содержимого веб-страницы. Вся текстовая информация, представленная на такого рода страницах может быть разделена на три вида:

- основной контент, подлежащий дальнейшей лингвистической обработке;
- маловажная информация (например, наиболее часто просматриваемые ссылки по теме);
- информационный мусор (например, реклама).

Задача выделения основного контента заключается в разделении веб-страницы на информационные блоки (каждый блок представляет собой микро-информационный элемент [3]) и определение принадлежности данных блоков к вышеперечисленным классам информационной

нагруженности. Но так как до настоящего времени не существует единой стандартной схемы построения HTML-страницы, проблема извлечения основного контента веб-страницы не решена.

В основу предлагаемой технологии извлечения основного контента Веб-страницы положены следующие идеи:

• Контент html-страницы имеет некоторые особенности оформления текста (среднюю длину предложений, количество слов, количество стоп-слов (процент стоп-слов), определенную плотность включаемых гипертекстовых ссылок, количество картинок, насыщенность html- тегами и др.);

• Блок с наибольшим количеством слов считается блоком основного контента. Если имеется больше одного блока с одним и тем же количеством слов, выбирается первый блок.

• Основной контент веб-страницы обычно включен в html-блоки, имеющие специальные характеристики, которые могут быть проанализированы (такие как имя CSS-класса). Например, блок, имеющий имя , скорее всего будет содержать основной контент веб-страницы).

• Основной блок текстовой информации может использовать специальные стили: расположения, выравнивания, позиции, фона, цвета, размера и др.

• Шаблонная часть сжатой структуры дерева Compressed Structure Tree CST группы общих страниц сайтов, с учетом содержания текста в узлах, позволяет определить блок главного контента [4].

• Основной блок информационного контента наиболее релевантен к заголовку (определенного тегом <title>) и мета данным, включенным в заголовок веб-страницы тегами <meta keyword> и <meta description>.

3. Этапы идентификации знаний Веб-страницы

Предлагаемая технология Web Content Mining экстракции и идентификации знаний из текстов Веб-страницы включает:

1) метод извлечения смысловых понятий из текстовой информации HTML-страницы, применяемый для выделения концептов, и установления их принадлежности к определенному семантическому полю, т.е. их семантическую классификацию[5];

2) метод выделения репрезентативных элементов семантических полей лингвистических единиц:

a. формирование правил обработки имеющейся структурированности Веб-страниц;

b. контекстный анализ, включающий использование регулярных выражений, представляющих синтаксические шаблоны — фразы, позволяющие определить иерархические семантические связи между концептами.

Концепты, представленные на Веб-странице, уже имеют определенную организованность (рис. 1), часто связанную с их значением. Основные возникающие при этом проблемы: выделение ключевых концептов данной предметной области (ПО), и установление иерархических и синонимических связей между ними.

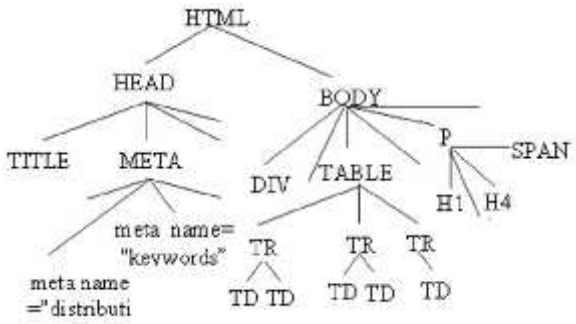

Рис. 1. Типичная структурная организация тегов веб-страницы

В предлагаемой технологии структурная организация текста на веб-странице используется в качестве дополнительной методики выявления отношения репрезентативности между понятиями данной предметной области. Для чего из всего множество тегов веб-страницы используется множество тегов, структурно организующих информацию веб-страницы:

- различные уровни заголовков, выражаемые тегами HTML: <H1>, …, <H6>;
- выделенные слова и фразы, обозначенные различными тегами HTML форматирования: <A>, <I>,…, <B>, <STRONG>, <BIG>, <EM>, <U>, <Li>, <DT> и т.д.

Проведенный анализ показывает, что определенные синтаксические конструкции, часто представленные на страницах Веб, выражают отношения таксономии и репрезентативности между концептами данной ПО. Например, наличие слов и словосочетаний: "Such as", "For example", "e.g.", "Including", "В качестве примера", "В частности", "Например", "Включает в себя", "Такие как". Были использованы следующие шаблоны регулярных выражений:

<concept>s such as <subconcept>*
such<concept>s as <subconcept>*
<subconcept> is a <concept>
<concept >[.,] например , <subconcept >
*<subconcept> (and|or) other <concept> s**

<concept >[.,] такие как , <subconcept > s+
такие<concept >[.,] как , <subconcept > s+

Для корректной работы шаблонов регулярных выражений для флективных языков на этапе морфологической обработки используются элементы стемминга.

Для имеющихся шаблонов разработаны правила определения существительных (NN) или групп существительных (NP), соответствующих элементам шаблона <*concept*>, <*concept_r*> и <*supconcept*>, которые позволяют определить семантические отношения таксономии и репрезентативности между элементами классов концептов ПО, выделенными на этапе экстракции и идентификации глубинных знаний из потоков текстовой информации.

Например, существительное или группа существительного шаблона <*subconcept* > являются представителями класса <NN> или <NP> шаблона <concept>, а соответственно NP <*concept*> представляет собой таксон множества элементов <NN> или <NP>, соответствующих правой части <*subconcept* > данного шаблона:

$$такие< concept >[.,]как,< subconcept> s +$$

$$< concept > s * such as < subconcept >$$

Первое регулярное выражение позволяет во фразе: "…*включает такие математические дисциплины как линейную алгебру, математический анализ, дифференциальные уравнения и…* " определить именные группы "*линейную алгебру*", "*математический анализ*" и "*дифференциальные уравнения*", в качестве концептов, находящиеся в таксономических отношениях с концептом "*математическая дисциплина*".

Шаблоны:

$$< concept \_ r > s * [.,]e.g.,(< concept > s+)+,$$

$$< concept \_ r > [.,]например[,](< concept > s+) +$$

позволяют выделять отношения репрезентативности. В данных шаблонах подстрока <*concept_r*> представляет собой существительное <NN> или группу существительного <NP>, являющиеся репрезентативными элементами для существительного или группы существительного правой части <*concept*> данных регулярных выражений.

Например, в предложении "*There are many Data Mining techniques, e.g. Clustering, Classification, Data Warehouses, Web Mining etc*." именная группа "*Data Mining techniques*", выражающая знак концепта, представляющего репрезентативных элемент подмножества концептов {"*Clustering*", "*Classification*", "*Data Warehouses* ","*Web Mining*"}.

Извлеченные таким образом ключевые концепты предметной области и идентифицированные иерархические (таксономические) связи, а так же выбор репрезентативного элемента семантического поля, позволяют формально представить знания данной ПО в виде таксономии.

Выводы

Проанализированные особенности экстракции и идентификации знаний из текста html-страниц позволяют разрабатывать технологию Web Content Mining в двух направлениях: решение задачи интеллектуального анализа данных и разработка технологии NLP, учитывающей особенности структурирования веб-контента. Предлагаемая технология включает предлингвистический этап выделения основного информационного блока, основанный на определении принадлежности текста к наиболее информационно-нагруженному элементу, и использование регулярных выражений для определения отношений репрезентативности и таксономии между концептами веб-страницы. При этом для выделения значащих концептов предполагается использование разработанных ранее методов экстракции и идентификации глубинных знаний из потоков текстовой информации

Список литературы

1. *Morinaga K. Mining product reputations on the Web / K.Morinaga, K. Yamanishi , Tateishi and T. Fukushima // Procs. of KDD. – 2002. – P. 341–349.*

2. *Wang Y. Web Mining and Knowledge Discovery of Usage Patterns / Y. Wang– 2000. Электрон. ресурс. – Режим доступа: https://cs.uwaterloo.ca/~tozsu/courses/cs748t/surveys/wang.pdf.*

3. *Broder A. Robust classification of rare queries using web knowledge / Broder A., Fontoura M., Gabrilovich E. et al. // Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval (SIGIR'07), 2007. – P. 231–238.*

4. *Lan Yi. Eliminating Noisy Information in Web Pages for Data Mining / Yi Lan, Liu Bing, Li Xiaoli // Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2003), Washington, 2003. – P. 296 – 305.*

5. *Хайрова Н. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення/ Н. Хайрова // Зб. наук. пр. Військового ін-ту Київ. нац. ун-ту. – К.: ВІКНУ, 2012. – Вип. № 38. – С. 239 – 245.*

Поступила в редколлегию 12.11.2014

Рецензент: д-р техн. наук, с.н.с. Е.С. Козелкова, Государственный университет телекоммуникаций, Киев.

ОСОБЛИВОСТІ ЕКСТРАКЦІЇ ТА ІДЕНТИФІКАЦІЇ ЗНАНЬ WEB-КОНТЕНТУ

Н.Ф Хайрова, Аджит Пратап Сингх Гаутам

Проаналізовано особливості екстракції та ідентифікації знань веб-сторінок. Запропоновано технологію Web Content Mining, що включає етапи виділення основного контенту сторінки витяг змістових понять і контентний аналіз. На етапі контентного аналізу використовуються регулярні вирази, які дозволяють виділяти явним чином відносини таксономії та репрезентативності між концептами веб-сторінки..

Ключові слова: *ідентифікація знань, Web Content Mining, регулярні вирази, таксономія, відношення репрезентативності.*

FEATURES EXTRACTION AND IDENTIFICATION OF KNOWLEDGE WEB-CONTENT

N.F. Hayrova, Ajit Pratap Singh Gautam

In the article the features of knowledge mining and knowledge identification of web-pages have been considered. The new kind technology of Web Content Mining has been elaborated. The technology is based on the method of extraction of semantic concepts from textual information and includes the steps: exarticulation of the main page-content, extraction of the semantic concepts and the content analysis. At the stage of content analysis regular expressions have been used. The regular expressions allow to manifestly distinguish relationships of the representation and taxonomy between concepts of the webpage. As elements of regular expressions were used nouns, nouns groups and special lexical constructs.

Keywords: *the identification of knowledge, Web Content Mining, regular expressions, taxonomy, attitude of representativeness.*