УДК 004.912

Светлов Н.Н.

студент

Научный руководитель: Соловьев Н.А., д.т.н

Оренбургский государственный университет

АЛГОРИТМ ПОИСКА КЛЮЧЕВЫХ ТЕРМИНОВ ВЕБ-СТРАНИЦЫ

Аннотация: Предложен алгоритм извлечения основного смыслового контента веб-страницы на основе её ключевых слов(keywords) и описания(description). Приведен пример использования математической модели русскоязычного текста для определения соответствия текста и ключевых слов страницы.

Ключевые слова: поиск терминов, веб, автоматическое извлечение терминов, математическая модель текстового документа.

Svetlov N.N.

student

Research supervisor: Soloviov N.A., d.t.s

Orenburg State University

Search algorithm for web-page keywords

Abstract: An algorithm for extracting the main semantic content of a web page based on its keywords and description is proposed. Shows an example of using mathematic model for Russian texts to determine the correspondence between the text and the keywords of the given page .

Keywords: terms search, web, automated terms extracting, text document mathematic model.

В рамках разработки системы адаптации контента веб-сайта для мобильных устройств перед нами встает задача выделения контента из html-разметки страницы для вставки текста в новую, подготовленную html-структуру.

Оптимизация страницы проходит в несколько этапов. Сначала система получает html разметку страницы. Затем из полученной разметки выбираются все текста из определенных тегов, выбираются подключаемые скрипты, стили и изображения. Основная сложность оптимизации – это вычленение из всего объема текстовой информации страницы той, которая содержит основной тезис, ключевые слова и описание страницы.

Данная математическая модель может быть использована для анализа существующих методов выделения ключевых слов из текста, а также для разработки новых методов.

Для выполнения этапа извлечения терминологических кандидатов данные методы не используют словари, онтологии или какие-либо другие семантические ресурсы. Кратко рассмотрим каждый из них [1]. Метод C-value базируется на использовании такой статистической метрики, как частота встречаемости строки в тексте. По сравнению с ней метрика C-value учитывает длину и вложенность терминологического кандидата . Вложенные термины (nested terms) – это понятия, содержащиеся в исходном тексте как по-отдельности, так и в составе других понятий [2]. Метрика, используемая методом C-value, подсчитывается согласно следующей формуле.

$$C - value(a) = \begin{cases} \log_2|a| \cdot \int(a) \\ a - \text{не вложен} \\ \log_2|a| \left(\int(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} \int(b) \right) \\ a - \text{вложен} \end{cases} \quad (1)$$

где a – терминологический кандидат;

$|a|$ - длина a – выраженная в количестве слов;

$\int(\cdot)$ – частота встречаемости кандидата;

T_a – множество извлеченных кандидатов, содержащих a;

$P(T_a)$ – количество кандидатов в T_a;

$\sum \int(b)$- сумма частот встречаемости кандидатов $b \in T_a$, содержащих a. То есть a является вложенным кандидатом по отношению к b;

Из вышеописанной формулы можно сделать вывод, что чем длиннее строка a, тем больше значение ее метрики. Это сделано для учета следующей закономерности. Более длинные строки встречаются в исходном тексте реже коротких. Следовательно, вероятность появления строки b в количестве f упоминаний меньше, чем вероятность появление строки a в количестве f раз, при условии, что $|a| < |b|$. По этой причине можно сделать вывод, что словосочетание b с большей вероятностью является термином по сравнению с a. Кроме этого, данный метод создан с предположением, заключающемся в том, что чем выше количество T_a – строк, содержащих a, тем больше степень независимости a [1].

Для адаптации данного метода к задаче анализа контента веб-страниц необходимо дополнительно проводить этап разбора html-тегов страницы и вырезать элементы с javascript и теги, в которых отсутствует наполнение (пустые).

Зачастую контент страницы заключается в теги <p> или <article>. Следую этому правилу на втором этапе необходимо извлечь весь текст из этих тегов. В результате мы получаем несколько разных по длине абзацев.

На следующем этапе разборе страницы выделяем ключевые слова из метатега <keywords> [3]. Производим выделение основных терминов из массива текстов, полученных на предыдущем этапе. Предварительно сделав выборку из первых 60% текстов с наибольшим количеством слов. По полученным массивам с ключевыми словами сравниваем их с ключевыми словами страницы и отбираем текст с наибольшими совпадениями

Использованные источники:

1. Петров А.С. Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста / Шульга Т.Э // Вестник ВГУ. 2017г №3. с 195-203.

2. Baroni M. BootCaT: Bootstraping Corpora and Terms from the Web / M. Baroni, S. Bernardini // Proceedings of LREC 2004. -2004. – Т.4. – С. 1313-1316.

3. Прохорова А.М. Seo-оптимизация // Евразийский союз ученых (ЕСУ) #30, 2016 | Экономические науки, 2016. С 79-82.