УДК 004.832.2

МЕТОД КЛАССИФИКАЦИИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Ф.И. ТРЕТЬЯКОВ, Л.В. СЕРЕБРЯНАЯ

Белорусский государственный университет информатики и радиоэлектроники
ул. П. Бровки, 6, г. Минск, 220013, Республика Беларусь
Fiodor.Tretyakov@gmail.com, l_silver@mail.ru

Метод классификации текстов напрямую зависит от языка, на котором написан текст. Разнообразие словоформ русского языка усложняет процедуру классификации. Необходимы способы, приводящие к единому виду все многообразие словоформ. Разработан алгоритм, определяющий корни лексем и выполняющий автоматическую классификацию текстов на их основе.

Ключевые слова: классификация, разделяющая функция, стеммер.

Огромные коллекции электронных документов требуют создания эффективных методов обработки данных. Одному из них — частотной классификации текстов посвящена данная работа.

Существует множество способов решить названную задачу. Прежде всего выбор подхода зависит от количества исходных данных. Если имеется набор текстов-образцов и категорий, то речь идет о контролируемом обучении и классификации. Затем необходимо определить решающее правило и разделяющую функцию, с помощью которых будет выполняться классификация «незнакомых» текстов. Они подставляются в качестве параметра в разделяющую функцию, в результате чего определяется их принадлежность к одному из классов. Одним из существенных факторов, влияющих на выбор класса, является язык, на котором написан текст [1]. Настоящая работа посвящена русскоязычным текстам.

В русском языке лексемы имеют сложные и разнообразные структуры, что существенно затрудняет процедуру классификации текстов. Однокоренные слова могут иметь различные окончания, суффиксы и приставки, которые не должны влиять на результат классификации. Однако при проверке формального совпадения однокоренных лексем и получении отрицательного результата сравнения классификация текстов, построенная на основе неточных результатов сравнения, оказывается неверной. Поэтому для анализа русского языка в качестве разделяющей необходимо выбрать функцию, оперирующую только частью слова и выдающую ответ на его основе.

Одним из способов выделения определенной части слова в русском языке является стемминг. Это процесс нахождения основы слова заданной лексемы. Основа не всегда совпадает с морфологическим корнем слова [2]. Задача нахождения основы слова представляет собой давнюю проблему в области компьютерных наук. Первая публикация на заданную тему датируется 1968 годом. Стемминг применяется в поисковых системах для расширения поискового запроса пользователя и является частью процесса нормализации текста. На сегодняшний день созданы различные реализации алгоритмов стемминга. Они применяются для решения различных задач интеллектуальной обработки текстовой информацией.

Для решения задачи классификации используется специальный алгоритм стемминга под названием — стеммер [2]. Он может выделять значимую часть слова (стем). Для русского языка это основа и (или) корень. Основа определяется проще, однако в последствии с ее помощью получается менее качественная классификация. Например,

корень одного слова может быть эквивалентен корню и суффиксу другого. Поэтому алгоритмы, опирающиеся на корень слова, считаются более точными.

Для качественного выделения корня слова одного стеммера оказывается недостаточно. Для работы с русским языком можно использовать два дополнительных модуля грамматического словаря: лемматизатор и флексер (склонение и спряжение). С помощью лемматизатора слова приводят к базовой форме, что выполняется после обработки лексемы стемом. Флексер умеет выдавать все грамматические формы слова на основе базовой. Это позволяет улучшить результат, проверяя найденные фрагменты по набору форм ключевого слова.

Среди всех реализаций стеммеров можно выделить два типа:

- использующие словарь для выделения части слова;
- использующие эвристическую модель [2].

Для выделения корня слова был разработан программный модуль, включающий в себя стеммер, флексер и лемматизатор. Стеммер использует эвристическую модель. Рассмотрим алгоритм классификации на основе созданного модуля.

1. Обработать название всех категорий с помощью модуля, выделив корни слов и поместив результаты в соответствующий словарь. Каждая строка в нем имеет ключ, которым является корень слова, а значение в строке — количество всех словоформ по ключу из названия категории.

2. Выполнить шаг 1 для всех текстов, применив его не к названиям текстов, а к ним самим.

3. Найти для каждого текста наиболее подходящую категорию. Ее номер определяется значения переменной T, вычисленной по следующей формуле:

$$T = \sum_{\substack{i=0,\\ j=0}}^{\substack{i<n,\\ j<m}} a_i * b_j ,$$

где n — размер словаря категории,
m — размер словаря текста,
a_i — слово из словаря категории,
b_k — о слово из словаря текста.

4. Выбрать для текста категорию, где T максимально.

С помощью стеммера, флексера и лемматизатора можно классифицировать тексты с высокой точностью. Минусом является сложность архитектуры модуля.

Список литературы

1. TextMining. Глубинный анализ текста. Из цикла лекций «Современные Internet-технологии» для студентов 5-го курса кафедры Компьютерных технологий физического факультета Донецкого национального университета. ДонНУ, кафедра КТ, проф. В. К. Толстых.

2. *Третьяков, Ф.И.* Методы обработки текстовой информации [Текст] / Ф.И. Третьяков, Л. В. Серебряная // VI международная научно-практическая конференция «Актуальные вопросы методики преподавания математики и информатики».— Биробиджан, 2011.— С. 175–181.