ISSN 2311-8814

«НАУКА | RASTUDENT.RU»

Электронный научно-практический журнал

График выхода: ежемесячно
Языки: русский, английский
ISSN: 2311-8814
Издатель: компания INFLASH
Учредитель: ИП Соколова А.С.
Место издания: г. Уфа, Российская Федерация
Прием статей по e-mail: rastudent@yandex.ru
Место издания: г. Уфа, Российская Федерация

Федотов Р.Г. Классификация текстовых документов. Уменьшение размерности задачи и повышение производительности // Наука-RASTUDENT.RU. – 2014. – No. 4(04-2014) / [Электронный ресурс] – Режим доступа. – URL: http://nauka-rastudent.ru/4/1310/

УДК 004

Федотов Руслан Геннадьевич,
Магистрант,
Факультет Электроники и Системотехники,
Московский Государственный Университет Леса,
Мытищи, Россия

Классификация текстовых документов. Уменьшение размерности задачи и повышение производительности

Аннотация: В данной статье рассматривается понятие классификации и основные направление, где она используется. Так же автор описывает основные способы предварительной обработки текстовых документов, для уменьшения размерности задачи классификации и повышение производительности систем, такие как стемминг, лемматизация, стоп-слова.

Ключевые слова: классификация, стемминг, лемматизация, обработка текста

Classification of text documents. Reducing the size of the problem and increase productivity

Fedotov Ruslan Gennadievich,
Magistrant,
Faculty of Computer Sciences,
Moscow State Forest University,
Mytischi, Russia

Abstract: This article discusses the concept of classification and the main area where it is used. So the author describes the basic pretreatment methods of text documents, in order to reduce the dimension of the problem of classification and performance improvement systems, such as stemming, lemmatization, stop words.

Keywords: classification, stemming, lemmatization, stop words

Классификация – одна из давних проблем, окончательно не решенная и по сегодняшний день. Под классификацией понимают группировку изучаемых объектов по видам, типам или другим признакам на основании содержания объекта для удобства дальнейшего их исследования. Следует различать классификацию и кластеризацию. Кластеризация так же выполняет группировку объектов по категориям, но здесь они заранее не известны.

Автоматическая классификация очень часто применяется в таких областях, как:

- Фильтрация спама;
- Сортировка новостей;
- Проверка авторства;
- Составление интернет-каталогов;
- Автоматическое аннотирование;

В наше время с появлением интернета и быстрым ростом информации в нем очень остро стоит проблема её классифицировать. Существует множество методов классификации, некоторые делают упор на качество классификации, некоторые на скорость. Но когда информации слишком много, необходимо делать предварительную обработку. Для этого существует несколько способов, которые помогают не только повысить производительность самой системы, но и уменьшить размерность информации, обрабатываемой классификатором.

Стемминг слов

Стемминг — это процесс нахождения основы слова для заданного исходного слова. Основа слова необязательно совпадает с морфологическим корнем слова.

Русский язык относится к группе флективных синтетических языков, то есть языков, в которых преобладает словообразование с использованием аффиксов, сочетающих сразу несколько грамматических

значений, поэтому данный язык допускает использование алгоритмов стемминга.

Русский язык имеет сложную морфологическую изменяемость слов, которая является источником ошибок при использовании стемминга. Обычно стеммером пользуются для поиска текста с имитацией учета морфологии. Под имитацией подразумевается неустранимо большое количество ошибок и нерелевантных результатов, которые возникают, если применять только стеммер. В русском языке источником ошибок при стемминге являются всевозможные изменения корня слова - беглые гласные, к примеру [1].

В качестве решения проблемы плохих результатов поиска со стеммером для русского языка можно использовать два дополнительных модуля грамматического словаря - лемматизатор и флексер (склонение и спряжение). С помощью лемматизатора можно приводить слова к базовой форме, поэтому после сопоставления слова со стемом можно уточнить результат с помощью лемматизации. Второй модуль - флексер, который умеет выдавать все грамматические формы слова на основе базовой. Это позволяет уточнять результаты поиска, проверяя найденные фрагменты по набору форм ключевого слова.

Самым распространенным алгоритмом стемминга является алгоритм Портера (Porter, 1980). Оригинальная версия этого алгоритма была только для английского языка, но впоследствии был создан проект «Snowball», в котором использовалась основная идея алгоритма, и реализованы стеммеры для большинства индоевропейских языков, включая русский.

Лемматизация

Это одна из прикладных дисциплин языкознания, она достаточно часто используется для морфологического анализа текстов, для чего все словоформы приводятся к их первоначальному виду [2]. В результате

которой удаляются только флективные окончания и возвращается основная, или словарная, форма слова, называемая леммой.

В русском языке словарной формой считается:

- Существительные – именительный падеж, единственное число (книгами – книга);
- Глаголы – инфинитивная форма (читали - читать);
- Прилагательные – единственное число, именительный падеж, мужской род (зарубежными - зарубежный);

Стоп-слова

Это слова, не несущие какой-либо самостоятельной смысловой нагрузки. В целях уменьшения базы данных системы не учитывают стоп-слова при индексировании, заменяя специальным маркером. К ним относятся:

- Союзы и союзные слова
- Местоимения
- Предлоги
- Частицы
- Междометия
- Указательные слова
- Цифры
- Знаки препинания
- Вводные слова
- А также ряд некоторых существительных, глаголов, наречий (например, сайт, давать, всегда, однако и др.)

В связи с постоянным развитием и усовершенствование существующих алгоритмов поиска, классификации, кластеризации и пр. базы данных стоп-слов обновляются и изменяются.

Рассмотренные способы повышения производительности и уменьшения размерности задачи далеко не все, которые существуют для обработки текстовой информации, но в большинстве случаев их хватает для подобных систем классификации.

Список литературы:

1. http://www.solarix.ru/for_developers/api/stemmer.shtml (дата обращения: 09.04.2014)
2. http://searchenginez.ru/lemmatizaciya-chto-eto/ (дата обращения: 09.04.2014)
3. http://delaem-krasivo.ru/programmirovanie/234-stemming-i-lemmatizaciya.html (дата обращения: 09.04.2014)
4. Губин М.В., Морозов А.Б. Влияние морфологического анализа на качество информационного поиска// Консорциум «Кодекс». – 2006г. - С. 1-6.

Дата публикации: 11.04.2014