УДК 004.912

СТАТИСТИЧЕСКИЕ МОДЕЛИ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ

© 2017 О. Ю. Лазарева, М. С. Боломутова

Московский политехнический университет (г. Москва, Россия)
Высшая школа печати и медиаиндустрии (г. Москва, Россия)

В работе рассматриваются основные модели статистического метода выделения ключевых слов из текстовых документов: модель на основе закона Ципфа и модель TF-IDF. Выделены основные преимущества и недостатки применения рассмотренных моделей выделения ключевых слов.

Ключевые слова: анализ текста, выделение ключевых слов, статистические методы выделения ключевых слов, модель на основе закона Ципфа, модель TF-IDF.

Развитие информационных технологий достаточно сильно затронуло процесс обучения, внедрив в него новые модели и концепции. Одной из таких концепций является применение электронных изданий в образовательном процессе. Некоторые электронные издания по сути своей являются в какой-то мере обучающей системой, которая оперирует некоторым массивом документов, навигацию по которым могут существенно облегчить ключевые слова.

Также выделение ключевых слов в электронных обучающих системах может помочь в определении дидактических единиц курса – логически самостоятельных частей учебного материала, по своему объему и структуре соответствующих таким компонентам содержания как понятие, теория, закон. Дидактические единицы могут использоваться в электронной обучающей системе для построения карты изучения материала, для более точного определения знаний учащихся во время тестирования, для помощи обучающемуся.

В настоящее время все острее становится проблема интеллектуальной обработки текстов, их структуризации и категоризации. Для того чтобы подвергнуть исходный документ интеллектуальной обработке необходимо его специальным образом подготовить, так как текст в своем исходном виде не подходит для интерпретации классификатором. Для решения этой проблемы к тексту применяется процедура индексации.

Индексация – это выделение некоторых признаков, которые содержит текст. В качестве таких признаков могут быть некоторые слова в тексте, несущие информацию о тематике документа. Такие слова называют терминами или ключевыми словами. Фактически задача индексации документа заключается в выделении ключевых слов из некоторого текста.

Существует несколько методов для выделения ключевых слов: статистические, гибридные и нейросетевые.

Статистический метод основывается на ранжировании всех слов текстового документа по частоте их встречаемости. В реализациях статистического метода выделения ключевых слов выделяют два подхода использование одного документа и некоторого массива документов с целью увеличения точности извлечения ключевых слов. Каждый из данных методов положен в основу моделей выделения ключевых слов: модель на основе закона Ципфа и модель TF-IDF. Принцип каждой из этих моделей построен на вычислении ранга ключевого слова путем применения специальной весовой функции.

Закон Ципфа основывается на понятии TF (англ. term frequency – частота слова, так же это отношение числа вхождения некоторого слова к общему количеству слов документа), так как в каждом тексте можно выделить статистические закономерности. Слова, встречающиеся в тексте можно разделить на слова, которые имеют высокую частоту употребления, но при этом не несут никакой смысловой значимости, и слова, которые имеют более низкую частоту встре-

Лазарева Ольга Юрьевна – Московский политехнический университет Высшая школа печати и медиаиндустрии, доцент кафедры информатики и информационных технологий, к. т. н., lazarevaoy@gmail.com.
Боломутова Марина Сергеевна – Московский политехнический университет Высшая школа печати и медиаиндустрии, студентка Института принтмедиа и информационных технологий, marinabolomutova@yandex.ru.

чаемости в тексте при этом обладая большей значимостью.

В процессе анализа текста измеряется частота встречаемости каждого слова, далее берутся эти частоты и располагаются по убыванию. Полученную последовательность пронумеровывают согласно позиции значения частоты, присвоенные каждой частоте номера, являются рангом слова. Таким образом, наиболее часто используемые слова имеют ранг 1. Тогда вероятность встретить произвольное, заранее выбранное слово будет равна отношению количества вхождений этого слова к общему числу слов в тексте:

$$TF_i = \frac{n_i}{T}, \quad (1)$$

где n_i – количество вхождений слова t_i,

T – количество слов в тексте,

TF_i – частота слова t_i.

Произведение частоты слова TF на порядковый номер частоты (ранг) будет постоянным для любого данного слова t_i вычисляется согласно формуле 2:

$$TF_i \times r_i = const, \quad (2)$$

где const – некоторая константа,

r_i – порядковый номер (ранг) частоты слова.

Согласно этой формуле, выведенной Ципфом, частота использования слова в тексте изменяется по гиперболе, в зависимости от количества вхождений. Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье – в три раза реже, чем первое, и так далее.

Модель на основе закона Ципфа используется крайне редко, так как ранг ключевого слова рассматривается в рамках одного документа, размер которого явно не соизмерим с современными объемами информации, где информационное пространство составляют множество документов.

Дальнейшим развитием применения величины TF стало появление метода TF-IDF. Данный метод эффективен для рассмотрения ранга ключевого слова среди множества документов, в которых оно упоминается. Метод TF-IDF заключается в извлечении ключевых слов и выделении словосочетаний путем сопоставления величины встречаемости слова в рамках одного документа с величиной встречаемости этого же слова в корпусе документов IDF (англ. inverse document frequency – обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в документе:

$$IDF = \log\frac{D}{DF}, \quad (3)$$

где D – общее количество документов (корпус),

DF – количество документов, в которых встретилось слово t_i.

Величина IDF уменьшает ранг широкоупотребительных слов. Для каждого уникального слова в рамках корпуса документов имеется только одно значение IDF.

Вычисление ранга ключевого слова в модели TF-IDF происходит по формуле (4):

$$W_i = TF_i \times IDF_i, \quad (4)$$

где W_i – ранг ключевого слова,

TF_i – частота слова в документе,

IDF_i – частота слова в корпусе документов.

Вес некоторого слова пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах корпуса. Таким образом, более высокий ранг будут иметь ключевые слова, имеющие большую частоту в рамках отдельного документа и меньшую частоту использования в других документах корпуса.

Ключевое слово в тексте может содержать в себе несколько слов, то есть ключевое слово является словосочетанием. Далеко не каждое словосочетание может быть ключевым. При индексации текстов проверяется сила связи слов в словосочетании, не случайно ли появление слов словосочетания рядом.

Для выделения ключевых словосочетаний в тексте используется мера MI, которая вычисляет частоту встречаемости слов словосочетания рядом друг с другом. Величина MI вычисляется по формуле (5).

$$MI = \log\frac{Fab \times n}{Fa \times Fb}, \quad (5)$$

где Fab – частота употребления ключевого словосочетания в документе,

Fa –частота употребления в документе только первого слова,

Fb –частота употребления в документе только второго слова,

N – количество упоминаний данного словосочетания.

Принцип выделения ключевых слов может быть весьма полезен для облегчения процесса построения плана обучения – ключевые слова, описывающие ёмко содержание каждой лекции, выделяются в отдельной лекции учебного курса. Таким образом, гораздо проще выстроить последовательность изучения дидактических единиц курса путем выстраивания связей между ключевыми словами, выделенных из

каждой лекции. Ключевые слова также являются базой для облегчения навигации по курсу дисциплины.

ЛИТЕРАТУРА

1. Астраханцев Н. А. Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии / Н. А. Астраханцев // Труды Института системного программирования РАН. – 2014. – Т. 26. – № 4. – С. 7-20.

2. Гринева М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов / М. Гринева, М. Гринев // Труды Института системного программирования РАН. — 2009. – Т. 16. – С. 155-165.

3. Захаров В. П. Автоматическое выявление терминологических словосочетаний / В. П. Захаров, М. В. Хохлова // Структурная и прикладная лингвистика. – 2014. – № 10. – С. 182-200.

4. Лазарева О. Ю. Архитектура интеллектуальной обучающей системы для оценки компетенций учащихся вузов / О. Ю. Лазарева // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. – М.: МГУП, 2014. – № 5. – С. 55-64.

5. Лазарева О. Ю. Методы выделения ключевых слов в контексте электронных обучающих систем вузов / О. Ю. Лазарева, М. С. Боломутова // Молодой ученый. – 2016. – № 26. – С. 143-146.

6. Попов Д. И. Нечеткая оверлейная модель учащегося в интеллектуальной обучающей системе / Д. И. Попов, О. Ю. Лазарева // Научный вестник Московского государственного технического университета гражданской авиации. – М.: МГТУ ГА, 2015. – № 213 (3). – С. 141-148.

7. Попов Д. И. Модель проверки знаний обучающихся на основе когнитивной карты учебного курса / Д. И. Попов, О. Ю. Лазарева // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. – М.: МГУП, 2015. – № 3. – С. 88-94.

8. Шереметьева С. О. Методы и модели автоматического извлечения ключевых слов / С. О. Шереметьева, П. Г. Осминин // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. – 2015. – Т. 12. – № 1. – С. 77-81.

9. Popov D. I. A Knowledge Testing Production Model Based on a Cognitive Map for SWI-Prolog Applications / D. I. Popov, O. Y. Lazareva // International Journal of Emerging Technologies in Learning. – 2015. – Vol. 10. – № 6. – P. 62-65.

STATISTICAL MODELS OF KEYWORDS ALLOCATION

© 2017 O. Yu. Lazareva, M. S. Bolomutova

Moscow Polytechnic University (Moscow, Russia)
Higher School of Print and Media Industry (Moscow, Russia)

The paper deals with main statistical models of selection of keywords from text documents: the model based on Zipf's law and the TF-IDF model. The basic advantages and disadvantages of the use of models for the allocation of keywords are listed.

Keywords: analysis of the text, keyword selection, statistical models of selection of keywords, model based on Zipf's law, TF-IDF model.