2. Релевантность // WebEffector: [сайт]. URL: http://www.webeffector.ru/wiki/Релевантность (дата обращения: 09.04.2013).
3. Daniel Jurafsky, James H. Martin. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2009. 988 с.
4. Лифшиц Ю. Классификация текстов. Алгоритмы для Интернета. 2005. URL: http://yury.name/internet/ (дата обращения: 09.04.2013).
5. Померанцев А. Классификация. 2011. URL: http://rcs.chph.ras.ru/ Tutorials/classification.htm (дата обращения: 09.04.2013).
6. Ермакова Л.М. Методы классификации текстов и определения качества контента. // Вестник Пермского университета. Серия: Математика. Механика. Информатика. № 7. Пермь: Пермский государственный национальный исследовательский университет, 2011. С. 47-53.

А. В. Глазкова
Тюменский государственный университет
УДК 004.912

ОСНОВНЫЕ ПОДХОДЫ К РЕШЕНИЮ ЗАДАЧ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ДОКУМЕНТОВ

Аннотация. Рассматривается задача автоматической классификации текстовых документов. Описываются подходы к ее решению и случаи применения того или иного подхода.
Ключевые слова: автоматическая классификация текстов, автоматическая обработка документов, информационный поиск.

В настоящее время в связи с широчайшим использованием электронной почты и поисковых систем все более актуальным становится решение проблем организации эффективного доступа к информации. В связи с этим много внимания привлекает к себе задача автоматической классификации текстовых документов по определенным тематическим группам.

Целью классификации текстовых документов является разделение документов из имеющегося множества по классам. Для этого используется обучающая выборка документов и алгоритм обучения, при помощи которого можно получить классифика-

тор, или функцию классификации γ, отображающую документы в классы.

$$\gamma : X \to C .$$

Здесь X пространство документов, C фиксированное множество классов, $C = \{c_1, c_2, ..., c_{|C|}\}$. В процессе выполнения классификации необходимо обеспечить достижение высокой точности не только на обучающей выборке, но и на новых данных. В случае, когда обучающие выборки недоступны, речь идет о классификации без обучения, или кластеризации документов. В настоящей статье будут обсуждаться методы классификации текстов с обучением.

Для осуществления автоматической классификации документов часто используется наивный байесовский подход:

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c) ,$$

где $P\ t_k \mid c$ — условная вероятность того, что термин t_k появится в документе из класса c; $P(c)$ — априорная вероятность того, что документ принадлежит классу c.

При помощи наивного байесовского метода выбирается наиболее вероятный класс, или класс, имеющий максимальную апостериорную вероятность.

$$c_{map} = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k \mid c) ,$$

где $\arg\max_x f(x)$ — значение x, в котором функция f достигает максимума.

Существуют два способа построения наивной байесовской модели мультиномиальная модель, генерирующая по одному термину из лексикона в каждой координате внутри документа, и модель Бернулли, создающая индикатор для каждого термина лексикона:

1 для термина, присутствующего в документе, и 0 — для отсутствующего. Мультиномиальную и бернуллиевскую модели можно представить в виде равенств

$$P(d \mid c) = P(\langle t_1,...,t_k,...,t_{n_d} \rangle \mid c) \text{ мультиномиальная модель,}$$

$$P(d \mid c) = P(\langle e_1,...,e_i,...,e_m \rangle \mid c) \text{ модель Бернулли,}$$

где $\langle t_1,...,t_k,...,t_{n_d} \rangle$ — последовательность терминов, появляющихся в документе d; $\langle e_1,...,e_i,...,e_m \rangle$ — бинарный вектор, состоящий из M индикаторов присутствия каждого термина в документе d [1].

Особенностью наивной байесовской модели является то, что она рассматривает термины, входящие в состав документа, как независимые, что неестественно для модели естественного языка, где термины часто являются зависимыми друг от друга. Кроме того, мультиномиальная модель основана на предположении о позиционной независимости. Модель Бернулли игнорирует координаты термина в документе, поскольку в ней учитывается только присутствие или отсутствие термина. Несмотря на это, решения о классификации, принимаемые при помощи байесовской модели, довольно точны и результирующий класс в наивной байесовской модели обычно имеет большую вероятность, чем другие классы, хотя оценки сильно смещены от истинных вероятностей [2].

В наивной байесовской модели документ представляется в виде последовательности терминов, представленных в тексте, или бинарного вектора, содержащего информацию о терминах и их частотах. Другим возможным представлением документа может быть векторная модель. Каждый документ в такой модели рассматривается как вектор в едином векторном пространстве общем для всей коллекции документов. Вектор каждого документа состоит из действительных чисел, характеризующих вес термина в документе, зависящий, как правило, от отношения числа вхождений этого тер-

мина к общему числу терминов в тексте и частоты употребления термина в других текстах. Для векторной классификации документов необходимо определить границы между классами, поскольку именно они определяют результат классификации. Примерами реализации векторной модели могут послужить классификация Роккио и kNN (k nearest neighbors k ближайших соседей).

Классификация Роккио разделяет векторное пространство на области, окружающие центроиды центры масс всех документов в классе, по одному для каждого класса. Граница класса представляет собой множество точек, равноудаленных от центроидов двух соседних классов. Центроид класса вычисляется по формуле:

$$\vec{u}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d),$$

где D_c — множество документов из пространства D, принадлежащих классу c: $D_c = \{d : \langle d, c \rangle \in D\}$, $\vec{v}(d)$ — нормализованный вектор документа d.

Эта классификация проста в реализации и эффективна по скорости работы, однако неточна, если классы далеки от сфер с примерно одинаковыми радиусами.

Классификация kNN относит документ к классу, которому принадлежат k его ближайших соседей. Соседи выбираются из множества объектов, класс которых уже известен, документ относится к наиболее многочисленному классу. Существует вероятностный вариант метода kNN, согласно которому оценивается не только число ближайших соседей, относящихся к определенному классу, но и вероятность того, что документ принадлежит этому же классу. Метод kNN имеет большую временную сложность, чем другие методы классификации документов. При наличии обширного обучающего множества метод k ближайших соседей лучше справляется со сложными, например, несферическими, классами, чем метод Роккио [3].

На модели векторного пространства также основан подход к классификации документов, основанный на методе опорных векторов. Цель этого вида классификации состоит в том, чтобы найти разделяющие поверхности между классами, максимально удаленные от всех точек обучающего множества. Если обучающее множество при этом содержит два класса данных, то допускается линейное разделение между ними, и метод опорных векторов ищет разделяющую поверхность, находящуюся на наибольшем расстоянии от любых точек данных. Расстояние между поверхностью и ближайшей точкой данных называется при этом зазором классификации. При использовании опорных векторов подразумевается, что решающая функция полностью определяется подмножеством данных, влияющих на положение разделителя, эти точки называются опорными векторами [1].

Возможными расширениями данного метода могут служить метод классификации с мягким зазором для множеств, не являющихся линейно разделимыми, при работе с которыми допускается некоторое количество ошибок классификации в силу возможного наличия шумовых документов; метод опорных векторов с несколькими классами; нелинейный метод опорных векторов используемый в случаях, когда наборы данных не позволяют применять линейные классификаторы [4]. Также был предложен метод гибридного использования метода, основанного на векторной модели и метода опорных векторов [5].

Эффективность методов, применяемых для классификации, напрямую зависит от эффективности выбора признаков, по которым производится классификация текстов. Как правило, в качестве признаков используются термины, однако может быть эффективной разработка специальных признаков для решения узкоспециальных задач. Конструирование таких признаков не решается методами машинного обучения, а является задачей эксперта. Также многие термины, употребляемые в качестве классификационных признаков, могут быть сгруппированы при помощи тезауруса или вручную и считаться эквивалентными. Разработка новых и усовершенствова-

ние имеющихся подходов к выбору признаков классификации является одной из актуальных проблем автоматической обработки документов.

Разнообразие существующих классификационных подходов дает возможность выбора наиболее подходящего метода для решения конкретной поставленной задачи. Однако исследования, связанные с автоматической обработкой документов, нуждаются в разработке новых алгоритмов работы с текстами, что повысило бы качество традиционных методов классификации. До недавнего времени частота появления терминов и их близость в тексте являлись едва ли не единственными критериями оценки соответствия документа информационному запросу. Сейчас перспективным является применение сложных математических методов для решения задач классификации, а также разработка подходов к усовершенствованию первичной обработки исходных документов например, применению различных методов классификации для текстов, относящихся к разным тематическим категориям, или создание новых алгоритмов подбора классификационных признаков.

СПИСОК ЛИТЕРАТУРЫ

1. McCallum A., Nigam K. A Comparison of Event Models for Naive Bayes Text Classification // In AAAI/ICML-98 Workshop on Learning for Text Categorization. 1998.
2. Маннинг К., Кристофер Д., Рагхаван, Прабхакар, Шютце, Хайнрих. Введение в информационный поиск: пер. с англ. М.: ООО «И.Д. Вильямс», 2011. 528 с.
3. Hastie T., Tibshirani R., Friedman J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction Springer, 2009. 533 c.
4. Tsochantaridis I., Joachims T., Hofmann T., Altun Y. Large Margin Methods for Structured and Interdependent Output Variables // JMLP. 2005.
5. Chin Heng Wana, Lam Hong Lee b, Rajprasad Rajkumar, Dino Isa A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine // Expert Systems with Applications, 2012.