Математические структуры и моделирование
2009, вып. 20, с. 93–96

УДК 681.3.06

ТЕМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ДОКУМЕНТОВ ПО СТЕПЕНИ БЛИЗОСТИ ТЕРМОВ

А.С. Епрев

В статье рассматривается один из методов тематической классификации текстовых документов по близости термов, входящих в описания заданных тематик и поступающих в систему документов.

Введение

Классификация текстовых документов для электронных библиотек и баз знаний рассматривается как один из возможных вариантов решения проблемы эффективного доступа к информационным ресурсам этих систем. Проблема заключается в сложности ориентирования в этих массивах. Отсутствие возможности получить наиболее актуальную и полную информацию по конкретной теме делает бесполезной большую часть накопленных ресурсов. Использование классификаторов позволяет сократить трудозатраты на поиск нужной информации.

Тематическая классификация документов является задачей автоматического определения тематики документа по заданному множеству возможных тематик. Причем каждый документ соответствует какой-нибудь одной из заданных тематик.

Проблеме классификации текстовой информации уделено много внимания [1–3]. Большинство методов классификации основываются на использовании простой векторной модели описания документов. В рамках этой модели документ описывается вектором, в котором каждому используемому в документе терму сопоставляется его значимость в этом документе. Значимость терма рассчитывается на основе статистической информации о встречаемости термов в документе. Описание тематики также представляется вектором, и для оценки близости документа и тематики используется скалярное произведение векторов описания тематики и описания документа.

В качестве исходных данных используется матрица *термы-на-документы*. Столбцы этой матрицы – документы, а строки – термы. Элементами этой матрицы являются частоты использования данного терма в данном документе.

Омский государственный университет им. Ф.М. Достоевского.
E-mail: a.eprev@gmail.com

Таким образом, каждый терм, документ и тематику можно представить в виде векторов в общем пространстве. Размерность этого пространства – k. Близость между любой комбинацией термов и/или документов может быть вычислена при помощи скалярного произведения этих векторов. Таким образом, чтобы отнести документ к одной из возможных тематик, достаточно вычислить наиболее близкую из возможных тематику.

1. Метод классификации документов

Классическая задача классификации документов по заданному набору тематик [2] заключается в определении для каждого поступающего в систему документа одной (или нескольких) тематик, к которым этот документ относится.

Существует множество методов классификации, но в их основе лежит один и тот же обобщенный алгоритм [3]:

- построение описаний тематик;
- построение описания поступившего в систему документа;
- вычисление степени близости описания тематик и описания документа и выбор наиболее близкой тематики.

1.1. Описания тематик и документов

Будем считать, что тематика документа определяется его «словарем». Различные синтаксические формы одного и того же слова, в словаре представляются базовой словоформой (термом). Из словаря исключаются наиболее употребительные слова, такие как предлоги, местоимения и прочие.

Описанием документа является все множество встречающихся в документе термов.

Тематики также описываются наборами термов, только эти наборы содержат не все употребляющиеся в данной тематике слова, а лишь небольшое их подмножество.

1.2. Построение описаний тематик

Тематика задается небольшим множеством относящихся к ней документов. Чтобы построить описание тематики, необходимо выявить отличия этой тематики от остальных. Это позволит сформировать набор термов, наилучшим образом подчеркивающих особенности рассматриваемой тематики.

Выбор термов для описания тематик производится при помощи следующего алгоритма:

1. Построение общего словаря термов W. В него включаются все термы, которые используются в документах, задающих тематики.

2. Для каждого терма $\omega \in W$ вычисляется оценка вероятности его использования в документах d данной тематики C:

$$P(\omega|C) = \frac{|\{d : d \in C, d \supset \omega\}|}{|C|}.$$

3. Для каждой тематики C строится тематический словарь. В этот словарь включаются термы, вероятность использования которых в рассматриваемой тематике превосходит вероятность их использования в любой другой тематике $C_i \in \Omega$, т. е.

$$P(\omega|C) \geq \frac{\sum\limits_{C_i \in \Omega} P(\omega|C_i)}{|\Omega|}.$$

Для каждого терма из тематического словаря тематики C вычисляется его значимость по следующей эмпирической формуле:

$$Order(\omega, C) = \frac{P(\omega|C)^2}{\sum\limits_{C_i \in \Omega} P(\omega|C_i)^2}.$$

4. Значимость термов $Order$ задает отношение порядка на множестве термов каждого из тематических словарей. Используя это отношение, из тематического словаря выбирается несколько наиболее значимых термов, которые будут использоваться в качестве описания этой тематики.

Оптимальное количество термов для включения в описание необходимо определить опытным путем.

1.3. Вычисление степени близости

Как уже было сказано выше, описываемый подход основывается на предположении, что тематика документа определяется его словарем.

Если мы определим функцию $Proximity(\omega_1, \omega_2)$, которая сопоставляет каждой паре термов некоторую оценку их тематической близости, то оценка тематической близости документа и тематики определяется тематической близостью термов, входящих в их описания.

Одним из вариантов оценки близости документа и тематики является среднее арифметическое попарных оценок тематической близости термов из описаний документа $d \in D$ и тематики $C \in \Omega$:

$$Hit(d, C) = \frac{\sum\limits_{\omega_i^d \in d} \sum\limits_{\omega_j^C \in C} (Proximity(\omega_i^d, \omega_j^C))}{|C| \cdot |D|}.$$

После вычисления степени близости документа с возможными тематиками можно выбрать одну или несколько тематик с наиболее высокими оценками, т. о. классифицировать документ в одну или несколько тематик.

1.4. Тематическая близость термов

Степень тематической близости пары термов характеризует, насколько часто эти термы употребляются в документах одной и той же тематики, но не обязательно присутствуют в одних и тех же документах.

Построим матрицу *термы-на-документы* X, строки которой отражают распределение термов по документам из обучающего систему набора документов. В качестве оценки тематической близости пары термов используется скалярное произведение соответствующих строк этой матрицы. Следовательно, чтобы вычислить степень тематической близости между всеми парами термов, достаточно вычислить матрицу XX^T. Таким образом, мы только что задали функцию $Proximity$:

$$Proximity(\omega_1, \omega_2) = XX^T[\omega_1, \omega_2].$$

2. Заключение

Мы рассмотрели метод классификации документов по степени близости к заданным тематикам. Однако ему присущи следующие недостатки [4, стр. 203]:

- метод не обнаруживает зависимости между термами, которые часто используются в документах одной и той же тематики, но редко встречаются вместе;
- случайные зависимости и ошибки правописания оказывают существенное влияние на получаемые оценки и негативно сказываются на точности метода;
- размер матрицы *термы-на-документы* очень велик даже для небольшого числа документов и поэтому метод весьма ресурсоемкий.

Открытым остается вопрос об оптимальном количестве термов для включения в описание тематик. Необходимо провести исследование: в рамках компьютерного эксперимента выявить зависимости качества классификации документов от числа термов, определяющих тематику.

Литература

1. Koller, D. Hierarchically classifying documents using very few words / D. Koller, M. Sahami // Proceedings of the Fourteenth International Conference on Machine Learnin. – 1997. – P. 170–178.
2. Lewis, D. A comparison of two learning algorithms for text categorization / D. Lewis, M. Ringuette // Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieva. – 1994. – P. 81-93.
3. Кураленок, И.Е. Автоматическая классификация документов с использованием семантического анализа / И.Е. Кураленок, И.С. Некрестьянов // Программирование. – 2000. – С. 31-41.
4. Ландэ, Д.В. Основы интеграции информационных потоков / Д.В. Ландэ. – К.: Инжиниринг, 2006.