Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематического состава потока новостей

Шмулевич М. М., Киселев М. В.
mark.shmulevich@gmail.com, mkiselev@megaputer.com
Москва, Московский Физико-Технический Институт,
компания «Megaputer Intelligence»

Данная работа посвящена автоматической смысловой кластеризации текстов. Рассмотрено её применение к анализу тематического состава потока новостей. Предложен новый метод, названный *островной кластеризацией*, который основан на статистической мере корреляции встречаемости в текстах термов, характеризующихся значимым превышением их частот над средним уровнем. Показано, что он обладает набором качеств, необходимых для успешного решения задачи кластеризации текстов, и может быть применен для анализа тематической структуры новостного потока.

В настоящее время объем массивов текстовых документов в научной сфере, бизнесе, и других областях человеческой деятельности неуклонно растёт. Этим обусловлен растущий интерес к методам автоматической текстовой кластеризации. Наиболее часто используемым подходом к представлению текстов при кластеризации является подход, в котором текст рассматривается как неупорядоченный набор начальных форм входящих в него слов (Bag of Words).

При формировании деревьев решений для кластеризации текстов возможно учитывать агрегированные правила, построенные на основе статистических и семантических характеристик термов, выделяемых из текстовых документов. Один из таких методов, названный *островной кластеризацией*, рассмотрен в данной работе. Алгоритм основан на использовании статистической меры корреляции встречаемости в текстах термов, характеризующихся значимым превышением их частот над средним уровнем.

Первая часть алгоритма состоит в построении так называемого графа корреляций термов. Этот граф задается матрицей парных корреляций булевых переменных a_{ip}, отражающих наличие терма i в документе p, так что связь между термами i и j считается существующей при достаточно сильной (большей, чем пороговое значение) корреляции между переменными a_i и a_p. Степень корреляции между термами i и j определяется следующим образом. Пусть n — общее количество термов во всех документах, n_i — количество термов в документах, в которых встречается терм i. Обозначим общее число термов j во всех текстах как N_j,

а количество термов j в документах, содержащих терм i — как N_{ij}. Если принять гипотезу, что термы i и j распределены в документах независимо друг от друга, то вероятность того, что в документах, содержащих терм i, окажется N_{ij} или более термов j — это вероятность получения не менее N_{ij} успехов в серии из N_j испытаний при вероятности успеха одного испытания, равной $\frac{n_i}{n}$. Эта вероятность есть $p_{ij} = P_B(N_{ij}, N_j, \frac{n_i}{n})$, где $P_B(n, N, p) = \sum_{i=n}^{N} b(i; N, p)$ — биномиальное распределение. Вероятность p_{ij} может быть принята в качестве основы для расчета меры корреляции между термами i и j — чем она меньше, тем более коррелированы эти термы. Однако величина p_{ij} все же не совсем подходит для описания силы связи термов i и j, в частности, потому что она, как легко видеть, не симметрична: $p_{ij} \neq p_{ji}$. Поэтому в качестве меры корреляции термов берется $\tilde{p}_{ij} = \max(p_{ij}, p_{ji})$.

Одним из важных применений данного метода может быть анализ динамики тематической структуры потока новостей. Показано, что рассматриваемый метод удовлетворяет основным свойствам, которыми должна обладать процедура кластерного анализа для того, чтобы быть практически применимой к кластеризации больших массивов текстов вообще, и к анализу динамики тематической структуры потока новостей в частности:

— интерпретируемость найденных кластеров в терминах смысла содержания относящихся к ним документов;
— статистическая значимость группирования текстов в кластеры;
— возможность отнесения документа более, чем к одному кластеру;
— не более чем логлинейный рост времени работы кластеризатора с увеличением количества текстов;
— минимальная (или вообще отсутствующая) необходимость настройки со стороны пользователя.

Применение процедуры островной кластеризации было проиллюстрировано с использованием публично доступного массива новостей Reuters–21578. Было показано, что метод островной кластеризации может успешно решать задачу тематической кластеризации потока новостей, давая описание полученных результатов в понятных человеку терминах.

В работе были использованы средства пакета для анализа данных PolyAnalyst.

Работа выполнена при поддержке компании Яндекс, грант № 102903.

Литература

[1] *Fellbaum C.* WordNet: An Electronic Lexical Database. — MIT Press, 2005.

[2] *Hofmann T.* Probabilistic Latent Semantic Indexing // 22-nd Ann. ACM Conf. on Research and Development in Information Retrieval, 1999. — Pp. 50–57.

[3] *Apte C. Weiss S.* Data Mining with Decision Trees and Decision Rules // Corpus Linguistics: Investigating Language Structure and Use. — 1997. — № 13. — Pp. 197–210.

[4] PolyAnalyst data/text mining system. User manual — `www.megaputer.com`.