Метод иерархической кластеризации массива текстовых материалов

Method of the hierarchical clustering of the array of text materials

Павленко / Pavlenko A.
Алексей Владимирович
(alxpavlenko@yandex.ru)
ФГКВОУ ВО «Череповецкое высшее военное инженерное училище радиоэлектроники» МО РФ,
младший научный сотрудник.
г. Череповец

Гатилов / Gatilov I.
Игорь Леонидович
(igor.gatilov@mail.ru)
доктор технических наук,
старший научный сотрудник.
ФГУП «18 Центральный научно-исследовательский институт» МО РФ,
главный научный сотрудник.
г. Москва

Ключевые слова: кластеризация текстов – clustering of texts; функционально-ролевая интерпретация – functional and role interpretation; нейросетевая классификация – neural network classification.

Предложен метод иерархической структуризации неупорядоченного массива текстовых материалов, основанный на определении степени тематической близости текстовых сообщений, составляющих данный массив.
Применение метода в перспективных комплексах обработки информации позволит снизить размерность задачи поиска текстовых сообщений за счет ее ограничения отдельными ветвями сформированной иерархии тематических кластеров. Кроме того, данный метод может быть использован в задачах формализации предметной области в целях подготовки данных при формировании онтологии.

The method of hierarchical structuration of the unregulated array of text materials based on determination of a level of subject closeness of the text messages making this array is offered.

Application of a method in perspective complexes of information processing will allow reducing dimensionality of the search of text messages due to its restriction with separate branches of the created hierarchy of subject clusters. Besides, this method can be used in tasks of data domain formalization for the purpose of data preparation when forming ontology.

В задачах информационно-аналитической обработки больших массивов текстовых сообщений нередко возникает потребность в их тематической структуризации, например в интересах поддержки деятельности аналитика по формализации предметной области. Целью разработки метода, описанного в данной статье, является снижение размерности задачи поиска текстового сообщения (ТС) по запросу аналитика в накопленных текстовых массивах комплексов информационной обработки. Снижение размерности задачи поиска ТС происходит за счет ограничения поисковых процедур отдельными ветвями иерархии тематических кластеров ТС. Структура метода представлена на рис. 1.

На первом этапе производится составление формального описания (ФО) случайно выбранного ТС на основе функционально-ролевой интерпретации [1]. Функционально-ролевая интерпретация предполагает описание ТС в виде множества функционально-ролевых структур (ФРС) – словарных конструкций в составе клаузы, описывающей некоторое событие. Клауза, как правило, представляет собой простое предложение, а ФРС является совокупностью трех основных элементов – действия, фигурирующего в описываемом событии, а также субъекта и объекта данного действия. На втором этапе для оценки попарной близости ТС применяется метод нейросетевой классификации, описанный в [2]. Результатом применения указанного метода является значение вероятности соответствия i-го формального описания ТС T_i j-му формальному описанию тематического класса $T_j^{\text{ЭТ}}$ (рис. 2). В данном случае $T_j^{\text{ЭТ}}$ будет являться формальным описанием ТС, выбранного на первом этапе.

Вышеуказанные значения вероятности определяются нейронной сетью прямого распространения с SOFTMAX-нормализацией выходов и при превышении какого-либо значения вероятности порогового коэффициента P текстовое сообщение считается принадлежащим соответствующему тематическому классу (в данном случае – кластеру). Первые три этапа повторяются до тех пор, пока каждое из ТС исходного массива не будет отнесено к какому-либо кластеру. Таким образом формируется первый (низший) уровень иерархии. Каждый кластер содержит информацию о включенных в него ТС и представляется одним формальным описанием, состоящим из наиболее характерных (частотных) для данного кластера элементов. По завершении формиро-

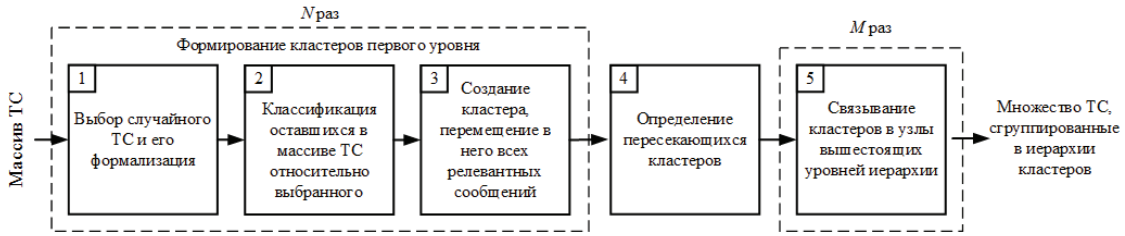


Рис. 1. Структура метода структуризации массива ТС

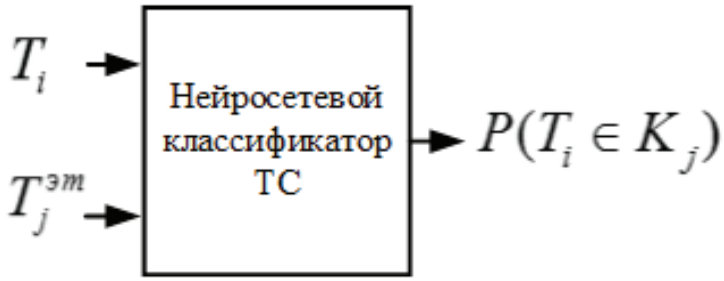


Рис. 2. Функциональная схема метода нейросетевой классификации

вания первого уровня кластеры проверяются на пересечение, в результате чего определяются ТС, соответствующие одновременно нескольким кластерам.

Формирование вышестоящих уровней иерархии осуществляется путем уменьшения на шаг ΔP порогового коэффициента P, на основании которого классификатором принимается решение о соответствии двух формальных описаний. С уменьшением P увеличивается количество кластеров, чьи формальные описания определяются как релевантные относительно друг друга. Коэффициент P уменьшается до тех пор, пока его значение не достигнет $P_{\min}$ (рис. 3), соответственно, количество уровней L сформированной иерархии выражением 1.

$$L = \frac{P_0 - P_{\min}}{\Delta P}, \qquad (1)$$

где:

P_0 – начальное значение порогового коэффициента;

$P_{\min}$ – минимальное значение порогового коэффициента;

ΔP – шаг уменьшения порогового коэффициента.

Формальное описание каждого кластера уровня 2 и выше представляется пересечением формальных описаний всех включенных в него кластеров более низкого уровня. Следует отметить, что иерархия нестрогая и не все кластеры могут быть связаны кластером высшего уровня, что отражается через формирование не одного, а нескольких деревьев и набора отдельных кластеров.

Временные затраты на структуризацию массива ТС определяются выражением 2.

$$t^{clust} = F * \sum_{i=0}^{N} W_i + R * M_0 * (\sum_{i=0}^{M_0} (N - 1 - \sum_{j=0}^{i} N_j^{clust}) + R * \sum_{i=0}^{M_0 - 1} \sum_{j=0}^{i} N_j^{clust} + R * \sum_{i=1}^{L} (M_i)^2 \oplus, \qquad (2)$$

где:

F – временные затраты на формализацию;

N – количество ТС;

W_i – количество слов в i-м ТС;

R – временные затраты на распознавание ФО;

M_i – количество кластеров на i-м уровне иерархии;

N_j^{clust} – количество ТС в j-м кластере.

Очевидно, что среднее значение временных затрат на проведение структуризации невозможно вычислить аналитически до её завершения, так как переменные M_i и N_j^{clust} априорно неизвестны. Кроме того, значительное количество переменных в выражении (2) не позволяет определить среднее значение требуемого времени на основании малого количества опытов. В связи с этим было осуществлено программное моделирование процедуры, при котором переменные M_i и N_j^{clust} генерировались случайным образом во время выполнения данной процедуры, что позволило имитировать формирование значительного разнообразия иерархических структур и рассчитывать временные затраты на формирование каждой из них. Проведение серии из 100 000 опытов при разном количестве ТС N в исходном массиве показало, что распределение временных затрат описывается нормальным законом (рис. 3). Это позволяет утверждать, что временные затраты на структуризацию

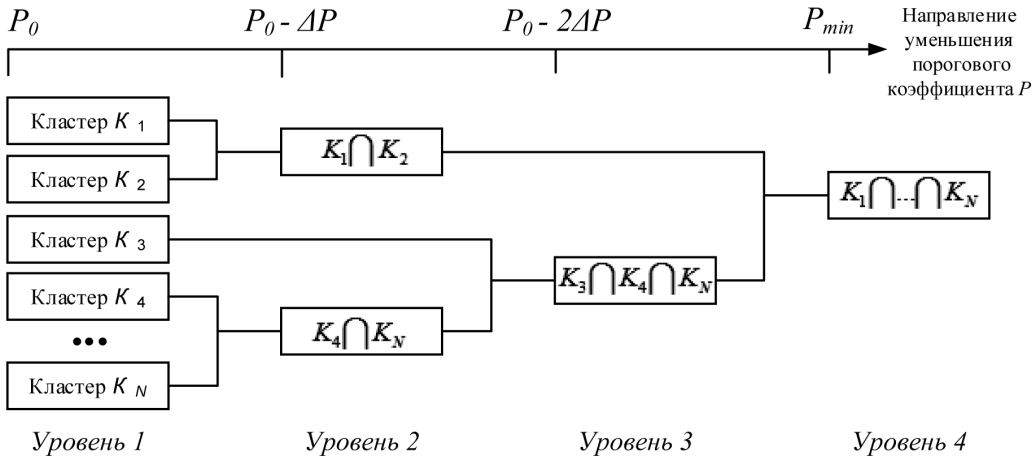
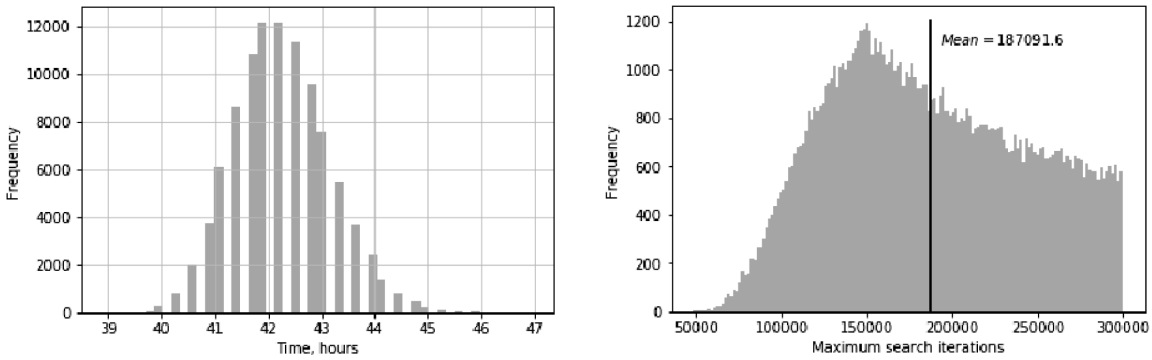


Рис. 3. Дерево кластеров ТС глубиной 4



а) *б)*

Рис. 4. Распределение временных затрат
(а) при структуризации массива объемом 300000 текстовых сообщений и распределение количества поисковых переборов ТС;
(б) по сформированной иерархии

массива из 300 тысяч ТС при среднем объеме 3500 слов с вероятностью 0,997 не превысят 45 часов (для сравнения, Национальный корпус русского языка содержит 335 076 текстов). Указанные временные затраты будут единовременными, так как построение иерархии тематических кластеров строится один раз и в дальнейшем лишь поддерживается в актуальном состоянии путем настройки классификатора (или нескольких классификаторов) на формальное описание каждого кластера.

Количество переборных процедур, осуществляемых при поиске по иерархии проиндексированных кластеров, в среднем меньше на 37,5 % по сравнению с количеством переборов, осуществляемых при поиске по неструктурированному массиву текстовых сообщений. Это подтверждается полученными распределениями количества поисковых переборов ТС по иерархиям кластеров общим объемом от 300 тысяч до 1 одного миллиона ТС (рис. 4).

Разработанный метод структуризации обладает одновременно следующими важными свойствами, необходимыми для корректной структуризации текстового массива:

1. Иерархичность.
2. Пересекаемость кластеров.
3. Инкрементность (т.е. отсутствие необходимости повторять процедуру структуризации заново при появлении новых ТС).
4. Отсутствие ограничений по числу возможных кластеров.

Одновременное наличие вышеуказанных свойств позволяет не только осуществить интерпретируемую структуризацию исходного текстового массива, но и

поддерживать актуальность содержания каждого узла иерархии путем настройки множества классификаторов на формальные описания кластеров и отбора из входного потока релевантных для них ТС. Кроме того, индексация содержания каждого кластера позволяет ускорить существующий поиск по ключевым словам за счет ограничения его отдельными ветвями сформированной иерархии вместо перебора всего массива текстовых материалов.

Литература

1. Столяров, М. Г. Способ определения информационной ценности текстового документа при полнотекстовом поиске, учитывающий отношения между понятиями предметной области / М.Г. Столяров, А.Ю. Новиков // Наукоёмкие технологии. – 2012. – № 8. – С. 87–90.

2. Павленко, А. В. Подход к нейросетевой классификации текстовых документов с использованием семантических признаков / А.В. Павленко, А.Ю. Новиков // Наукоёмкие технологии. – 2015. – № 12. – С. 67–70.