Использование оптимизационного метода роя частиц в стохастическом алгоритме категоризации текста

Пекарь Д.В.; Тихоненко С.Г.
Кафедра интеллектуальных систем, факультет радиофизики и компьютерных технологий
Белорусский Государственный Университет
Минск, Беларусь
e-mail: pekar.dima@gmail.com, siarhei.tsikhanenka@googlemail.com

Аннотация **– задачей категоризации текстов является автоматическое отнесение текстовой информации к одной из предопределенных категории. В представленной работе рассматривается новый вектор признаков для категоризации текстов, а также алгоритм на основе метода роя частиц для его оптимизации.**

Ключевые слова: категоризация текста; стохастический алгоритм; метод роя частиц

I. Введение

Текстовая информация является одним из наиболее распространенных типов обмена информацией. Для эффективного манипулирования текстами необходимы автоматические алгоритмы на основе анализа их содержимого.

В представленной работе рассматривается новый набор признаков для категоризации текстов, отличающийся от наиболее используемых наборов, основанных на различных методах [1-4]. Разработанный вектор представляет собой набор слов-индикаторов: релевантных слов, характерных для текстов определенной категории, и нерелевантных (не характерных) для данной категории слов.

II. Вектор признаков и его оптимизация

Пусть N – число категорий текстовой информации, V_i^{rel} – счетное множество слов, встречаемых в текстах i- ой категорий, V_i^{irrel} – счетное множество слов, которые не характерны для i- ой категорий, определяемое выражением (1):

$$V_i^{irrel} = \bigcup_{\substack{j=1 \\ j \neq i \\ V_i^{rel} \cap V_j^{rel} \neq \varnothing}}^{N} V_j^{rel} \setminus V_i^{rel}, \quad (1)$$

Перед нахождением множеств V_i^{rel} и V_i^{irrel} проводиться предварительная обработка текстов: удаление стоп слов [5] и стемматизация [6]. Множество V_i^{irrel} строиться в несколько этапов. На первом этапе извлекаются все слова из всех текстов принадлежащих i - ой категории. Затем, итеративно, перебираются все тексты из оставшихся категорий, и если отрывок содержит слова из V_i^{rel}, то из него выбираются все слова во множество V_i^{irrel}, которые не входят в V_i^{rel} так, что выполняется условие:

$$V_i^{rel} \bigcap V_i^{irrel} = \varnothing, \quad (2)$$

Цель работы алгоритма состоит в нахождении за определенное число итераций таких подмножеств $\widetilde{V}_c^{rel} \subseteq V_c^{rel}$ и $\widetilde{V}_c^{irrel} \subseteq V_c^{irrel}$ для некоторой заданной категории C, для которых целевая функция (3) принимает наибольшее значение:

$$f_{target}(\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}) = P_{precision}^c = \frac{\left| Text_{rlevant}^c \bigcap Text_{retrieved}^c \right|}{\left| Text_{retrieved}^c \right|}, \quad (3)$$

где $Text_{relevant}^c$ – множество текстов, относящихся к тестируемой категории C, $Text_{retrieved}^c$ – множество текстов, отнесенных к тестируемой категории C после применения алгоритма классификации, $|Text_{retrieved}^c|$ – мощность множества, определяющая число элементов в данном множестве. В представленном алгоритме значение целевой функции представляет собой точность классификации текста, вычисленную для категории C при заданных подмножествах $\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}$.

Для описания содержащихся слов в искомых подмножествах $\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}$ строиться вектор W_c с бинарными значениями для заданной категории C, так, что его длина равна сумме мощностей множеств V_c^{rel}, V_c^{irrel} и j-ый элемент w_{cj} вектора W_c принимает значение равное 1, если j-ый элемент множества V_c^{rel} присутствует во множестве $\widetilde{V}_c^{rel}$ при $j = \overline{1...\left|V_c^{rel}\right|}$ или при $j = \overline{\left|V_c^{rel}\right| + 1...\left|V_c^{irrel}\right|}$, элемент с индексом $j - \left|V_c^{rel}\right|$ множества V_c^{irrel} присутствует во множестве $\widetilde{V}_c^{irrel}$. Предложенный вектор имеет вдвое меньшую размерность, чем предложено в [7], что повышает скорость работы алгоритма и снижает размер требуемой памяти для размещения данных. В случаях, отличных от описанных, j-ый элемент w_{cj} вектора W_i принимает значение 0. Таким образом, присвоение значения равное 1 определенному элементу вектора W_c, означает добавление соответствующего элемента (слова) из множеств V_c^{rel}, V_c^{irrel} во множества $\widetilde{V}_c^{rel}$, $\widetilde{V}_c^{irrel}$ в зависимости от индекса модифицируемого элемента.

III. Оценка найденного решения

Оценка найденного решения осуществляется с помощью бинарного классификатора:

$$Cl(\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}) = (T \bigcap \widetilde{V}_c^{rel} \neq \varnothing) \wedge (T \bigcap \widetilde{V}_c^{irrel} = \varnothing), \quad (4)$$

где T – множество различных слов в анализируемом тексте. Оценка эффективности найденного решения осуществляется при фиксированных подмножествах $\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}$ согласно алгоритму, изображенному на рисунке 1:

Библиотека БГУИР

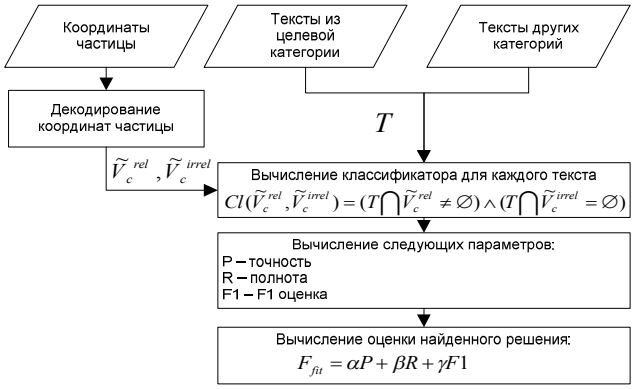

Рис. 1. Алгоритм вычисления оценочной функции

IV.МЕТОД РОЯ ЧАСТИЦ

Метод оптимизации с помощью роя частиц (Particle Swarm Optimization, далее PSO), базируется на моделировании поведения множества частиц в пространстве параметров задачи оптимизации [8], который использует для решения множество частиц, где каждая частица представляет собой возможное решение задачи оптимизации. Каждая $i-ая$ частица может быть представлена как объект с набором следующих параметров: $X_i = (x_{i1}, x_{i2}, x_{i3}, ..., x_{iS-1}, x_{iS})$ - положение частицы в S- мерном пространстве, $V_i = (v_{i1}, v_{i2}, v_{i3}, ..., v_{iS-1}, v_{iS})$ - скорость частицы, $P_i = (p_{i1}, p_{i2}, p_{i3}, ..., p_{iS-1}, p_{iS})$ - наилучшее положение частицы. Также важным параметром является наилучшее достигнутое положение в рамках всего роя частиц: $G_i = (g_1, g_2, g_3, ..., g_{S-1}, g_S)$. Тогда изменение скорости отдельной $i-ой$ частицы определяется следующим выражением:

$$v_{id} = w*v_{id} + c1*R*(p_{id} - x_{id}) + c2*R*(g_d - x_{id}), \quad (5)$$

где w - коэффициент инерции, $c1$ - ускорение для локального наилучшего положения, $c2$ - ускорение для глобального наилучшего положения, R - случное значение $R \in (0;1)$, $d = \overline{1...S}$ - индекс координаты.

Кодирование решения осуществляется следующим образом. Размерность пространства решений выбирается таким образом, что выполняется условие:

$$\left|V_C^{rel}\right| + \left|V_C^{irrel}\right| = S, \quad (6)$$

В конечном итоге каждая координата ассоциирована с определенным словом из исходного словаря. Если значение ассоциированной координаты равно 1, то соответствующее слово из исходного словаря добавляется в результирующий словарь, т. е. во множества $\widetilde{V}_c^{rel}, \widetilde{V}_c^{irrel}$. Для реализации описанного кодирования решения, осуществляется дискретизация положения частиц с помощью сигмоидной функции:

$$P(x) = \frac{1}{1 + e^{-x}}, \quad (7)$$

с учетом правила:

$$\widetilde{x} = \begin{cases} 1, rand() \le P(x) \\ 0, rand() > P(x) \end{cases}, \quad (8)$$

Структурная схема, описанного алгоритма категоризации текста, приведена на рисунке 2.

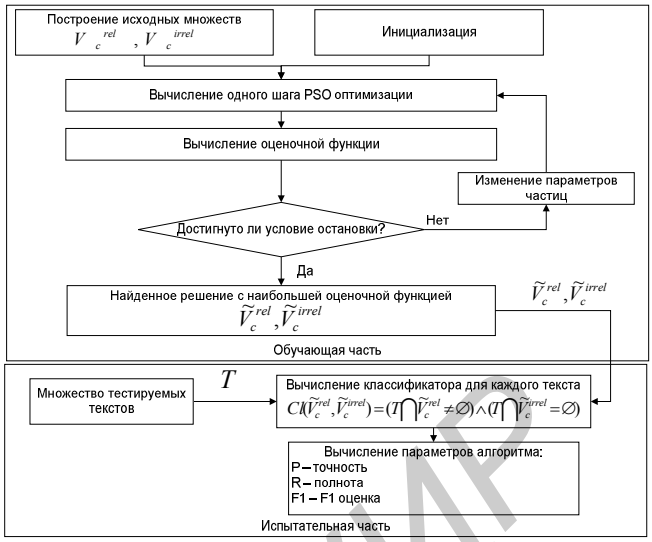

Рис. 2. Алгоритм категоризации текстовой информации

V.РЕЗУЛЬТАТЫ

Тестирование проводилось с использованием коллекции текстов Reuters-21578 [9], которая содержит 21578 текстов. Для оценки эффективности работы алгоритма были использованы 10 наиболее употребляемых категорий. В таблице 2 приведены результаты тестирования алгоритма.

Табл. 2. Результаты тестирования алгоритма

№	*Категория*	*Точность*	*Полнота*	*F1 оценка*
1	acq	0,75	0,69	0,72
2	earn	0,96	0,92	0,94
3	coffee	0,86	0,46	0,6
4	crude	0,94	0,79	0,86
5	gold	0,88	0,96	0,92
6	interest	0,66	0,75	0,56
7	money-fx	0,62	0,55	0,54
8	ship	0,64	0,47	0,55
9	sugar	0,91	0,94	0,93
10	trade	0,65	0,93	0,76

[1] Tao Li, Shenghuo Zhu, Mitsunori Ogihara, "Text categorization via generalized discriminant analysis", Information Processing and Management, vol. 44, 2008, pp. 1684–1697.

[2] Edda Leopold, Jorg Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?", Machine Learning, vol. 46, 2002, pp. 423–444.

[3] Abdellaif Rahmoun and Zakaria Elberrichi, "Experimenting N-Grams in Text Categorization", The International Arab Journal of Information Technology, vol. 4, no. 4, Oct. 2007, pp. 377-385.

[4] Nerijus Remeikis, Ignas Skucas, Vida Melninkaite, "Text Categorization Using Neural Networks Initialized with Decision Trees", Informatica, 2004, vol. 15, no. 4, 551–564

[5] Journal of Machine Learning Research [Electronic source] / Access mode: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

[6] M.F. Porter, "An algorithm for suffix stripping" Program, vol. 14, no. 3, 1980, pp.130-137.

[7] Pietramala Adriana, "A Genetic Algorithm for Text Classification Rule Induction", Lecture Notes in Computer Science, vol. 5212, 2008.

[8] J. Kennedy, R. Eberhart, "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks. IV. 1995. pp. 1942–1948. doi:10.1109/ICNN.1995.488968.

[9] Machine Learning and Intelligent Systems. — [Electronic source]. — Access mode: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection.

Библиотека БГУИР