ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 25 МАТЕМАТИКИ Выпуск 2
2018

А. К. Мельников (Москва, НТЦ ЗАО «ИнформИнвестГрупп»). **Применение точных распределений в процедуре двухэтапной обработки текстов.**

Некоторые задачи, связанные с обработкой текстов [1], могут быть сведены к двухэтапной процедуре обработки, первый этап которой сводится к отбору текстов и основан на применении критериев согласия с тем или иным распределением, а второй этап состоит из углубленной обработки отобранных текстов.

1. Построение процедуры обработки. Пусть на первом этапе из массива, состоящего из M текстов длины n, содержащих знаки алфавита $A_N = \{a_1, \ldots, a_N\}$ мощности N,

$$T_{n,N}(j) = \{t_1(j), \ldots, t_n(j)\}, \qquad j = 1, 2, \ldots, M,$$

нам необходимо отобрать подмассив текстов, являющихся реализациями случайных выборок длины n из равновероятного распределения на алфавите мощности N. На втором этапе каждый из отобранных текстов необходимо подвергнуть дальнейшей углубленной обработке с получением положительного или отрицательного результата. При этом в подмассив нам необходимо отобрать не более $\overline{M} \ll M$ текстов, так как из-за ограничений, накладываемых на производительность имеющихся для проведения дальнейшей углубленной обработки вычислительных средств большего чем $\overline{M}$ количества текстов мы обработать не сможем.

На втором этапе — этапе углубленной обработки — результат обработки каждого из $\overline{M}$ текстов может быть положительным, а может быть и отрицательным. Будем предполагать, что положительный результат углубленной обработки будет получен в том и только в том случае, когда отобранный текст содержит равновероятное распределение входящих в него знаков, т. е. отобран в подмассив правильно. Определим число положительных результатов обработки $\overline{M}$ текстов отобранного подмассива через R^+. Из определения R^+ видно, что $R^+ \leqslant \overline{M}$. Блок-схема процедуры двухэтапной обработки текстов представлена на рис. 1.

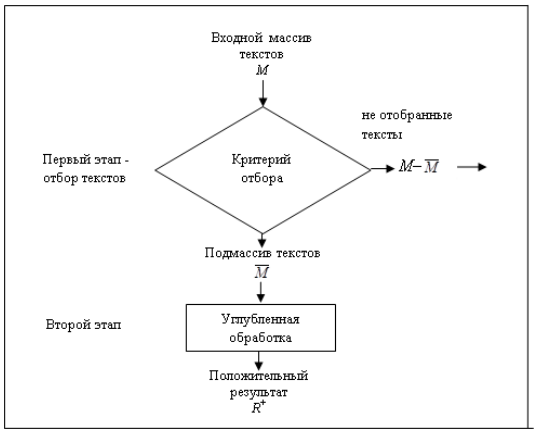

Рис. 1. Блок-схема двухэтапной процедуры обработки текстов

2. Эффективность обработки текстов. Под эффективностью обработки будем понимать величину $\omega_{\overline{M}}$, равную отношению числа положительных результатов углубленной обработки отобранных текстов R^+ к общему числу отобранных текстов $\overline{M}$:.

$$\omega_{\overline{M}} = \frac{R^+}{\omega_{\overline{M}}}.$$

Из определения эффективности видно что, $0 \leqslant \omega_{\overline{M}} \leqslant 1$, а $\max \omega_{\overline{M}} = 1$ достигается при $R^+ = \overline{M}$.

Простейшим способом решения поставленной задачи является сортировка всех M текстов $\{T_{n,N}(j) \,|\, j = \overline{1, M}\}$ по признаку равновероятности и выбор из отсортированного массива для обработки первых $\overline{M}$ текстов и их последующая углубленная обработка.

Рассмотрим ситуацию, когда M текстов даны нам не все сразу, а поступают последовательно, в течение определенного периода времени. Необходимо, обрабатывая последовательно каждый поступающий текст, принимать решение о его принадлежности к подмассиву для углубленной обработки, или отвергать. Данный подход к построению процедуры последовательной обработки поступающих текстов был впервые сформулирован академиком А. А. Боровковым еще в 60-х годах XX века на конференции по методам прикладной статистики и частично представлен им в [2] и [3].

Отбор текстов с равновероятным распределением знаков производится с помощью применения к каждому из M текстов критерия согласия с равновероятным распределением [4], использующего некоторую статистику S_n текста длины n, являющуюся функцией от h_i частот встречаемости знаков (исходов) текста a_i из алфавита A_N мощности $N - S_n = f(n, N)$ и распределение вероятностей значений используемой статистики (распределение) — $\mathbf{P}\{S_n \geqslant c\}$.

Также сделаем предположение, что на первом этапе при применении критерия отбора к каждому из M текстов мы сможем отобрать в подмассив $\overline{M}$ текстов. Отобранные $\overline{M}$ текстов будут содержать как $\overline{M}'$ текстов отобранных правильно, с равновероятным распределением входящих в них знаков, так и $\overline{M}''$ ошибочно отобранных текстов.

$$\overline{M} = \overline{M}' + \overline{M}''.$$

Заметим, что по принятым предположениям дальнейшая углубленная обработка текстов из числа ложно отобранных, не даст положительного результата, а значит $R^+ = \overline{M}'$. На рис. 2 представлена детализированная схема второго этапа обработки текстов.

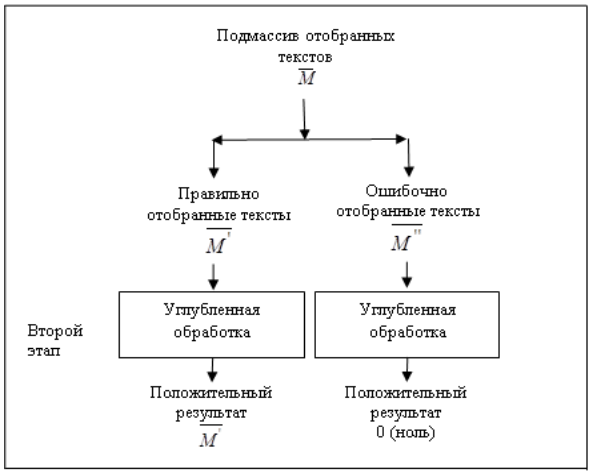

Рис. 2. Детализированная блок-схема второго этапа процедуры обработки текстов

На рис. 2 углубленная обработка показана двумя одинаковыми блоками для наглядности получения итоговых результатов в соответствии с принятыми предположениями. Тогда эффективность обработки принимает вид

$$\omega_{\overline{M}} = \frac{\overline{M}'}{\overline{M}} = \frac{\overline{M}'}{\overline{M}' + \overline{M}''}.$$

Величину $\overline{M}''$ определяет размер применяемого критерия — α [5]. Размер критерия α связан с вероятностью распределения значений применяемой в критерии статистики $S_n - \mathbf{P}\{S_n \geqslant c\}$ соотношением [6]

$$\mathbf{P}\{S_n \geqslant c\} = \alpha.$$

Число текстов $\overline{M}''$, ошибочно отобранных как тексты с равновероятным распределением знаков оценивается как

$$\overline{M}'' \cong \alpha\,\overline{M}$$

а тогда применяя тождественные преобразования имеем, что

$$\omega_{\overline{M}} = \frac{\overline{M}'}{\overline{M}} = \frac{\overline{M} - \overline{M}''}{\overline{M}} = \frac{\overline{M} - \alpha\,\overline{M}}{\overline{M}} = \frac{(1-\alpha)\,\overline{M}}{\overline{M}} = 1 - \alpha.$$

Подробное построение критерия согласия для принятия решения о равновероятном распределении знаков текста проведено в работе [7]. В зависимости от параметров выборки (n, N) применяются либо точные распределения, либо предельные [7]. Область параметров применения точных распределений определяется возможностями по производительности вычислительных средств, применяемых для их расчетов [8]. Область применения предельных распределений определяется из результатов, полученных Фишером в [10], Крамером в [4] и Кендаллом в [11]. В [9] показано, что существует область параметров (n, N), для которой не могут быть рассчитаны точные распределения, а предельные распределения применяться не могут — так называемая область неопределенности. В условиях невозможности расчета точных распределений для параметров из области неопределенности до настоящего момента использовались предельные распределения. Предложенный в [7] обобщенный статистический метод анализа текстов дает возможность использовать для параметров из области неопределенности распределения сколь угодно близко приближенные к их точным значениям и позволяет строить критерии с наименьшим уровнем значимости α, что дает при их использовании в процедуре обработки текстов наибольшую эффективность, позволяющую экономить дорогостоящий вычислительный ресурс.

Заключение. Предложенная двухэтапная процедура анализа текстов, использующая для построения критерия отбора текстов точные или близкие к точным распределения статистик, обладает высокой эффективностью при значительной экономии вычислительного ресурса. Рассмотрение относительной эффективности процедур отбора текстов, построенных с использованием предельных распределений и распределений, близких к точным, является предметом дальнейших исследований автора.

СПИСОК ЛИТЕРАТУРЫ

1. *Чеповский А. М.* Информационные модели в задачах обработки текстов на естественных языках. М.: Национальный открытый университет «ИНТУИТ», 2015, 228 с.
2. *Боровков А. А.* Вероятностные процессы в теории массового обслуживания. М.: Наука, 1972. 367 с. Stochastic processes in queueing theory. Springer, 1976, 280 p.

3. *Боровков А. А.* Математическая статистика. Новосибирск: Изд-во ИМ СОРАН, Наука, 1997, 772 с.
4. *Крамер Г.* Математические методы статистики. Пер. с англ. А. С. Монина, А. А. Петрова под ред. А. Н. Колмогорова. М.: Мир, 1975, 648 с.
5. *Ивченко Г. И.*, *Медведев Ю. И.* Введение в математическую статистику. М.: ЛЕНАРД, 2017, 608 с.
6. *Ивченко Г. И.*, *Медведев Ю. И.* Математическая статистика. М.: Книжный дом «ЛИБРОКОМ», 2014, 352 с.
7. *Мельников А. К.*, *Ронжин А. Ф.* Обобщенный статистический метод анализа текстов, основанный на расчете распределений вероятности значений статистик. — Информатика и ее примен., 2016, т. 10, в. 4, с. 89–95.
8. *Мельников А. К.* Сложность расчета точных распределений вероятности симметричных аддитивно разделяемых статистик и область применения предельных распределений. — Доклады ТУСУРа. 2017, т. 20, № 4, с. 126–130.
9. *Зелюкин Н. Б.*, *Мельников А. К.* Сложность расчета точных распределений вероятности значений статистик и область применения предельных распределений. В сб.: Электронные средства и системы управления: Материалы докладов XIII Международной научно-практической конференции. (29 ноября – 1 декабря 2017 г.) Томск: В-Спектр, 2017, Ч. 2, с. 84–90.
10. *Фишер Р. А.* Статистические методы для исследователей. М.: Госстатиздат, 1958, 257 с.
11. *Кендалл М. Г.*, *Стьюарт А.* Теория распределений. Т. 1. М.: Наука, 1966, 302 с.