УДК 681.3

К СОЗДАНИЮ ТИПОВОГО КОМПОНЕНТА ОПЕРАЦИОННОЙ ПЛАТФОРМЫ ЗАДАЧ КОНТЕНТ-МОНИТОРИНГА ТЕКСТОВ

В.А. Литвинов, С.Я. Майстренко, И.Н. Оксанич
Институт проблем математических машин и систем НАН Украины
e-mail: takam@post.com

1. В последние годы все большую актуальность приобретают задачи и системы компьютерного анализа текстовой информации.

Наряду с универсальными "высокоинтеллектуальными" системами анализа ("text mining"), основанными в частности, на построении и обработке семантических сетей, отображающих множество смысловых понятий текста, существует потребность в создании специализированных проблемно-ориентированных систем контент-мониторинга текстов [1], близких по функциям к традиционному контент-анализу. Эти задачи возникают при создании СППР на высших уровнях руководства (корпоративного, государственного). Примером таких задач является экспресс-мониторинг новостных электронных публикаций в соответствии с заданными лингвистическими условиями, определяющими наличие и распределение в тексте заданных ключевых слов. Для реализации таких задач полезно располагать готовым набором инструментов, реализующих типовые операции "нижнего уровня", не зависящие от целей и конкретного характера контент-мониторинга.

2. В основе операционной платформы (ОП) подавляющего большинства прикладных задач контент-анализа текстов лежит небольшой набор элементарных операций, включающих:

- поиск в тексте заданных лексических единиц (слов "образцов"), установление факта и показателя крайности вхождения слов в текст;
- определение "расстояния" между вхождениями слов;
- поиск словосочетаний, образующих заданные смысловые категории, и т.п.

Процедуры выполнения именно этих операций и составляют основу операционной платформы рассматриваемых задач анализа текстов (ОП АТ).

В связи с тем, что степень независимости перечисленных операций различна, представляется целесообразным двухуровневый подход к разработке ОП АТ, основанный на отделении операций количественного анализа требуемых характеристик текста от операций непосредственной обработки символов текста (рис. 1). При этом процесс разбивается на 2 этапа:

1) предварительная обработка текста с целью получения первичных данных, описывающих результаты "наложения" искомых слов на синтаксическую структуру текста;

2) целевая количественная обработка первичных данных и вычисление количественных характеристик текста, создающих основу для аналитической обработки и последующих выводов.

Такой подход дает возможность структурировать и упростить логику процессов обработки текста. Первый этап является стандартным, алгоритм его выполнения не

зависит от текстов, слов и искомых характеристик. Второй этап несет определенную содержательную нагрузку и зависит от целей анализа текстов.

Под линейной моделью разметки текста понимается последовательность линейных адресов вхождений слов в виде кортежей «<слово > <номер вхождения > <текущий номер символа текста >». Структурная модель разметки представляет собой совокупность относительных адресов вида <слово > <номер вхождения > <номер раздела (абзаца) > <номер предложения в разделе> <номер символа в предложении >. Структурная модель сдержит всю основную информацию, которую можно извлечь в результате «чтения» текста и сопоставления с искомыми словами.

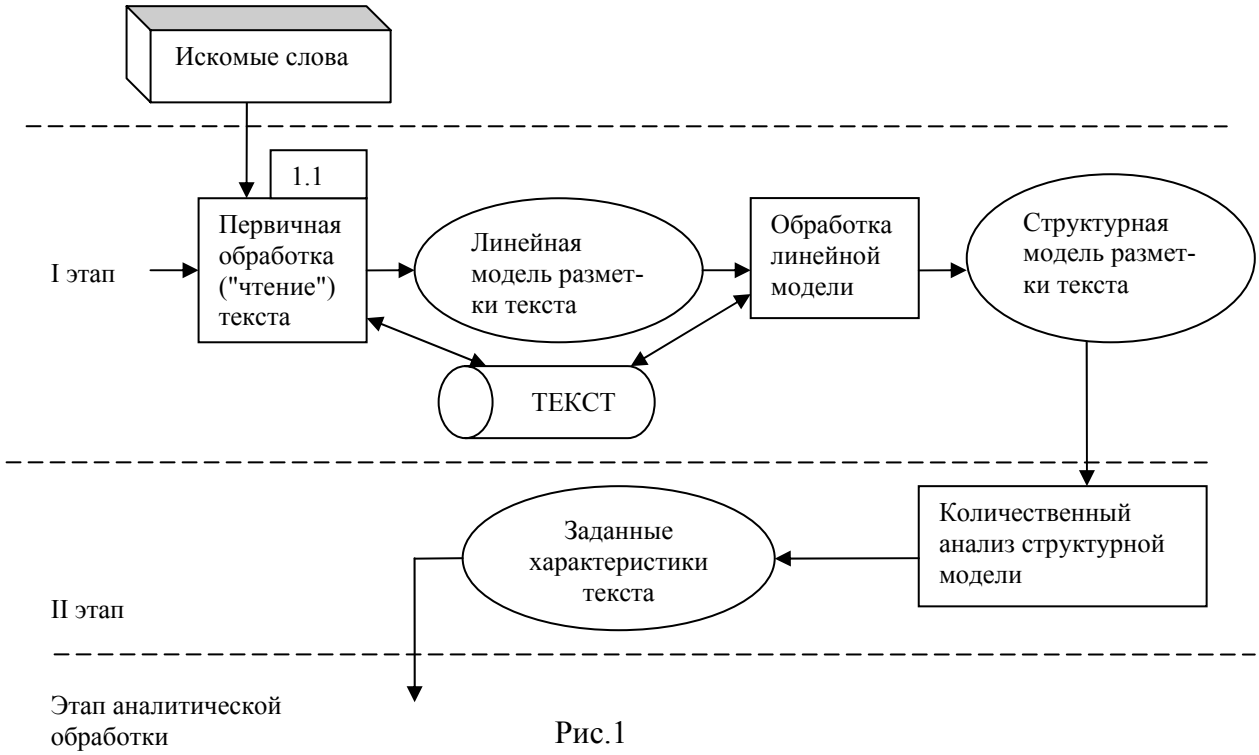

Рис.1

3. Реализация 1-го типового этапа и, в частности блока 1.1, основана на алгоритмах поиска вхождения подстроки (образца) в строку (текст). Существует большое количество алгоритмов выполнения этой операции, ориентированных на различные специфические условия задачи поиска. Аналитический обзор алгоритмов со ссылками на первоисточники приведен в [2], сводные результаты обзора отражены, например в [3].

Лучшим алгоритмом для поиска образца в обычных естественно-языковых (ЕЯ) текстах считается алгоритм Боуэра-Мура (первоисточник [4]). Алгоритм и его модификации теоретически исследованы и описаны во многих источниках (см. например, [2]) и не нуждаются в подробном изложении. Ответ на вопрос о реальных временных характеристиках обработки ЕЯ-текста алгоритмами Боуэра-Мура на современных ЭВМ дают результаты сравнительных экспериментальных исследований, проведенных для алгоритма «грубой силы» (ГС), простого алгоритма Боуэра-Мура (БМ1) и улучшенного алгоритма Боуэра-Мура (БМ2). В алгоритме БМ1 сдвиги образца в процессе поиска определяются вектором смещений, а в алгоритме БМ2 – матрицей смещений. Последнее обеспечивает возможность выполнять регистронезависимый поиск

и поиск по шаблону, содержащему «подстановочные» (неспецифицированные) символы, не учитываемые при анализе текста. Например, для образца «?оды» будут найдены слова «коды», «годы» и т.п. Исследования алгоритмов проводились на компьютере (P4, 2.6 Ггц, 1.0 ГБ ОЗУ) для следующих данных:

- длина образца m = 2, 4, 8, 16, 32, 64 символа;
- длина текста n = 70 МБ, 140 МБ, 280 МБ;
- алфавит – кодовая таблица ASCII.

В результате проведенных исследований были получены следующие зависимости максимального времени поиска от:

- длины текста для длины образца m = 8 БТ, примерно соответствующему средней длине содержательной единицы ЕЯ-текста русского языка (рис.2);
- длины образца для текста длиной n = 140 МБ (рис.3);
- количества символов шаблона k для образцов разной длины для улучшенного алгоритма Боуэра-Мура (БМ2)(табл.1-3).

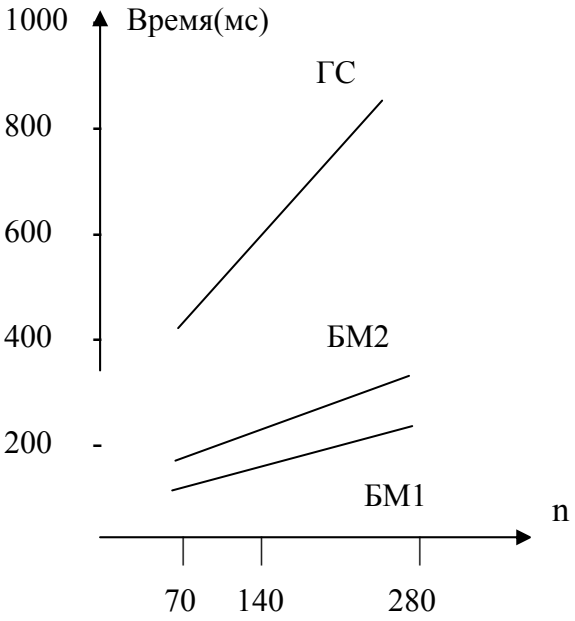

Рис.2 Зависимость времени поиска от длины текста

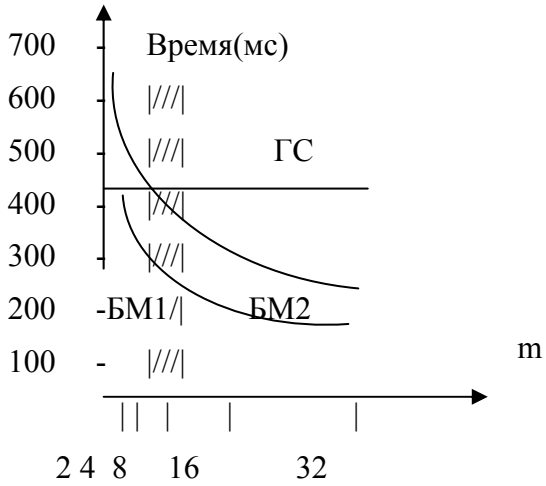

Рис.3 Зависимость времени поиска от длины образца

Таблица 1 Зависимость времени поиска от количества символов шаблона для m=8 БТ

k	0	2	3
k/m	0	0,25	0,37
Время (мс)	226,3	224,4	221,5

Таблица 2 Зависимость времени поиска от количества символов шаблона для m=16 БТ

k	0	2	4	6	8	10
k/m	0	0,12	0,25	0,37	0,50	0,62
Время (мс)	160	159	158,5	157,5	151,3	147,7

Таблица 3 Зависимость времени поиска от количества символов шаблона для m=32 БТ

k	0	2	4	6	8	10	12	14	16	18	20	22
k/m	0	0,6	0,12	0,18	0,24	0,30	0,36	0,42	0,48	0,54	0,60	0,66
Время (мс)	108,5	108,6	108,7	108,6	102,8	102,7	99,7	102,3	101	98	98	94,8

4. Результаты проведенных экспериментов совместно с имеющимися литературными данными дают основания для вывода о целесообразности использования для блока 1.1 типового этапа рис.1 комбинации алгоритмов БМ1-БМ2. При этом в качестве основного используется программа БМ1, а БМ2 "вступает в игру" при наличии подстановочных символов в образце и/или прямого задания регистронезависимого поиска.

Литература

1. Федорчук А.Г. Контент-мониторинг информационных потоков. http://www/nbuv/ua/articles/2005/05fagmip.html.

2. Christian Charras , Thierry Lecroq, Handbook of Exact String-Matching Algorithms // www-igm.univ-mlv.fr/~lecroq/string/index.html.

3. Точный поиск подстроки в строке. http://program.rin.ru/razdel/html/828.html.

4. BOYER R.S., MOORE J.S., A fast string searching algorithm, *Communications of the ACM,* 1977.