ПРИМЕНЕНИЕ МНОГОПОТОЧНЫХ ВЫЧИСЛЕНИЙ ДЛЯ УСКОРЕНИЯ ПРОЦЕССА ПОИСКА ЭКВИВАЛЕНТНЫХ ТЕКСТОВ

Д. С. Бухаров
Служба автоматизированных систем диспетчерского управления, филиал АО «СО ЕЭС» «Региональное диспетчерское управление энергосистемы Иркутской области»
Иркутск, Российская Федерация
E-mail: bukharovds@gmail.com

В работе представлен модифицированный подход к поиску эквивалентных текстов, основанный на формировании маски заданного текста в виде множества слов-элементов. Сопоставление текстов с заранее подготовленной маской выполняется в многопоточном режиме, что позволяет существенно ускорить процесс решения задачи.

Введение

В настоящее время наиболее актуальными задачами поиска эквивалентных текстов являются «библиотечный поиск» и «поиск плагиата». Формализация таких задач – достаточно сложный процесс, поскольку сравнение текстов выполняется на естественном языке, который содержит в себе множество аспектов слабо поддающихся математико-алгоритмическому описанию.

Процесс поиска эквивалентного текста можно разделить на два этапа:

Этап 1. Формирование поискового множества (маски текста).

Этап 2. Сравнение поискового множества с некоторым набором текстов.

I. Построение поискового множества

Пусть M_i $(i = \overline{1,k})$ – некоторый текст, содержащий элементы $a_1, a_2, \ldots, a_n$, k – заданное количество текстов, N – текст, содержащий элементы $b_1, b_2, \ldots, b_m$, и сравниваемый с M_i. Необходимо найти такой текст M_i из k вариантов, при котором

$$Q = \max_i \left(\frac{h_i(M_i \cap V)}{h(V)} \right) \qquad (1)$$

где $h(V)$ – мощность множества V, образованного из элементов текста N, при этом каждый элемент N обработан согласно определенным правилам; $h_i(M_i \cap V)$ – мощность пересечения множеств M_i и V.

Под элементом текста понимается некоторое сочетание символов: слова, числа, цифробуквенные выражения.

При решении задачи (1) выполняется поиск такого текста M_i из k возможных вариантов, который содержит в себе наибольшее количество элементов из множества V, при этом, если отношение $Q = 0$, то i-ый текст не содержит в себе ни одного подобного элемента из V. Если $Q \to 0$, то тексты M_i и V являются эквивалентными.

Прежде чем начать сравнение текста N с некоторым набором текстов M_i $(i = \overline{1,k})$, необходимо специальным образом обработать элементы текста N и составить такое поисковое множество V, при котором решение задачи (1) будет максимально корректным. Для этого необходимо учесть ряд особенностей: частотность слов (из поискового множества удаляются малозначащие высокочастотные слова с целью снижения вероятности обнаружения текста, не являющегося эквивалентом, для которого $Q \to 1$), морфемная структура слова (отбрасываются малозначащие служебные морфемы, выделяются корни слов), регистр букв (сравнение выполняется с учетом заданного в тексте регистра букв, а также в режиме единого регистра), цифробуквенные выражения. Подробнее о предварительной обработке представлено в [1].

II. Сравнение текстов

При сравнении поискового множества V с текстами M_i $(i = \overline{1,k})$ вычисляется значения выражения (1), являющегося критерием эквивалентности тестов. После выполнения всех k вычислений выбирается текст, для которого значение Q является наибольшим. Данный текст отмечается как эквивалентный для множества V.

При проведении многократных вычислительных экспериментов выявлены два недостатка описанного подхода.

Недостаток первый. Корректность решения существенно снижается, если заменить в текстах M_i $(i = \overline{1,k})$ значащие слова на их синонимы. Как следствие, для таких текстов значение выражения (1) существенно уменьшается, снижая вероятность определения этих текстов как эквивалентов.

Для решения данной проблемы модифицировано ранее разработанное программное обеспечение: выполнена интеграция с базой данных синонимов слов.

На этапе формирования V для каждого элемента запоминаются все синонимы. В процессе сравнения выполняется последовательное сопоставление слов из M_i с каждым синонимом из множества V. Если определено соответствие

хотя бы по одному из синонимов, то результат сравнения считается положительным, при этом совокупность всех синонимом для каждого элемента из V определяется как единый элемент множества V, т. е. мощность $h(V) = const$ независимо от количества синонимов.

Недостаток второй. Существенные временнные затраты на сопоставление текстов: сравнение 287 «поисковых» текстов (используемых для формирования V) с 915 оригиналами в однопоточном режиме на компьютере с процессором AMD Athlon II Dual-Core M300 2.00 GHz под управлением операционной системы Windows 7 выполняется 3,5 часа. Результат корректного сопоставления составляет 98,25 % без учета синонимов слов. Пять из 287 текстов (оставшиеся 1,75 %) предварительно обработаны: большинство значащих слов заменено на их синонимы.

Добавление синонимов в поисковые множества увеличивает время решения задачи до 4 часов, но при этом улучшается результат корректного сопоставления до 100 %.

Для ускорения процесса сравнения текстов выполнена многопоточная реализация основного вычислителя. Поскольку операция сравнения текстов однотипная, то она выполняется в параллельном режиме каждым потоком. Общим ресурсом, доступ к которому выполняется последовательно, являются множество текстов M_i, масив с результатами вычисления выражения (1) и номер обрабатываемого в данный момент времени текста N. Каждый поток работает со своим поисковым множеством V, сформированным из одного из 287 текстов.

За обработку результатов вычислений отвечает отдельный поток, который после сопоставления всех текстов формирует файл с результатом решения задачи. Взаимодействие модулей разработанного программного обеспечения схематично представлено на рис. 1.

Сравнение 287 поисковых текстов с 915 оригиналами (с учетом синонимов при формировании множества V) при двух потоках выполняется 1,87 часа. Процесс сопоставления текстов ускорился приблизительно в 2,14 раза. Дальнейшее увеличение количества потоков не приводит к улучшению, поскольку процессор AMD Athlon II Dual-Core является двуядерным, и обработка третьего и последующего «излишне» созданного потока выполняется по мере высвобождения вычислительного времени процессора.

Решение данной задачи в однопоточном режиме на компьютере с четырехядерным процессором Intel(R) Core(TM) i7-5600 CPU 2.60 GHz под управлением операционной системы Windows 7 выполняется 2,8 часа. При двух потоках – 1,3 часа, при четырех потоках – 0,6 часа (скорость сопоставления текстов увеличилась приблизительно в 4,67 раза).

III. Заключение

Итак, вычислительный эксперимент при значительном количестве сравниваемых текстов на компьютерах с различным количеством ядер процессоров показал, что применение многопоточных вычеслений для решения трудоемких задач обеспечивает хорошое ускорение.

Наилучший результат достигается при равном количестве потоков программы и ядер процессора компьютера, на котором выполняются расчеты. Поскольку задача сравнения текстов является достаточно трудоемкой, то чрезмерное увеличение количества потоков не только не приводит к ускорению процесса решения задачи, но и замедляет его.

Список литературы

1. Бухаров, Д. С. О поиске эквивалентных текстов / Д. С. Бухаров // Прикладная информатика. – 2016. – № 3. – С. 46–52.

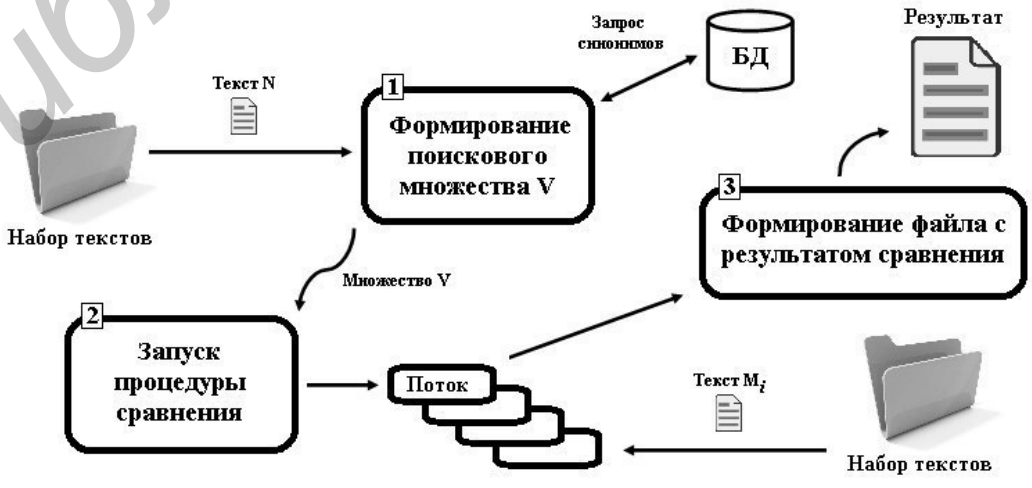

Рис. 1 – Схема взаимодействия модулей программы