

ОТЗЫВ

на автореферат диссертации Черняк Екатерины Леонидовны,
**«Разработка вычислительных методов анализа неструктурированных текстов
с использованием аннотированных суффиксных деревьев»**
на соискание учёной степени кандидата технических наук
по специальности 05.13.18 – «Математическое моделирование, численные методы
и комплексы программ»

Работа Екатерины Черняк связана с актуальной задачей обработки текстов на естественном языке: определением релевантности некоторой «строки» (поискового запроса, ключевого слова и т. д.) «тексту». В работе предлагается алгоритм решения этой задачи с помощью аннотированных суффиксных деревьев (АСД) – структуры данных, позволяющей эффективно представлять языковые модели. Автором впервые предложена адаптация алгоритма Уокконена для построения АСД за линейное время.

Фактически АСД представляет из себя упакованные в суффиксное дерево n -граммные языковые модели с разными значениями n . Для оценки релевантности соответствия строки тексту предлагается метрика СУВСС (средняя условная вероятность символа в совпадении). Эта метрика фактически представляет из себя взвешенное скалярное произведение вектора «строки» и вектора «текста» в пространстве n -грамм. Оказывается, что она легко вычислима с помощью построенного АСД.

Полученная модель верифицируется на задаче рубрикации научных статей, задаче автоматизированного построения таксономий, а также на задаче поиска (фильтрации) по однословному ключу для фильтрации обценной лексики. Результаты сравниваются с другим классическими методами, например, использованием коэффициента Жаккара и поиску ближайшего совпадения по расстоянию Левенштейна в задаче фильтрации. Во всех случаях предлагаемые методы, основанные на АСД, показали либо лучшие результаты, либо результаты, сравнимые с лучшими.

При исследованиях были использованы программы WikiDP (для извлечения структурированных данных из Википедии) и EAST (для предварительной обработки текста и быстрого построения АСД), созданные при участии автора.

В качестве возможных направлений развития работы можно указать на преобразование АСД в конечный автомат, позволяющий в духе алгоритма Ахо-Корасик при переходе к "следующему" началу подстроки не искать очередную подстроку заново. Также было бы интересно провести сравнение со статьёй [Ionia Veritawati, Ito Wasito, T. Basaruddin. *Text Preprocessing using Annotated Suffix Tree with Matching Keyphrase* // *IJECE*, Vol. 5, No. 3, 2015, pp. 409-420].

Стоит также обратить внимание на то, что в информационном поиске (где «строка» – это поисковый запрос, а «текст» – документ) часто требуются более сложные модели, учитывающие расширения и переформулировки запроса, исправления опечаток, а также пользовательские данные (например, клики пользователей по похожим запросам на документы).

Замечания по автореферату:

1. Имеется ряд опечаток: например, вместо «каждая строка помечена непустой подстрокой строки S », очевидно, должно быть «каждое ребро помечено непустой подстрокой строки S » (стр. 7).
2. Для нестрогих неравенств следовало бы использовать символ " $\leq$ ", а не " $\leq$ " (стр. 4).
3. При сравнении разных методов решения задач обработки текста стоило бы сравнивать не только вычислительную сложность, но и затраты потребляемой памяти.
4. n -граммные языковые модели со сглаживанием названы «вычислительно неэффективными» (стр. 4), хотя на самом деле их эффективность зависит от используемых структур данных.

Считаю, что перечисленные недостатки не снижают общей ценности работы и не могут существенно повлиять на главные теоретические и практические результаты. Судя по представленным в автореферате сведениям, диссертация соответствует требованиям ВАК при Минобрнауки России, а соискатель Черняк Екатерина Леонидовна заслуживает присуждения учёной степени кандидата технических наук по специальности 05.13.18 – «Математические модели, численные методы и комплексы программ».

19.10.2016

А.Зобнин

Зобнин Алексей Игоревич,
кандидат физико-математических наук

Руководитель группы морфологии
ООО «Яндекс»
119021, Москва, ул. Льва Толстого, 16

подпись *Зобнина А.И.*
ЗАВЕРЯЮ
БЕЛОВА Л.Г.
РУКОВОДИТЕЛЬ СЛУЖБЫ ПО РАБОТЕ
СО ШТАТНЫМИ СОТРУДНИКАМИ
НА ОСНОВании ДОВЕРЕННОСТИ №57
ОТ 10 МАРТА 2016
19 ОКТ 2016

