УДК 002
ББК 70
Д 63

Д 63 **Документ как социокультурный феномен:** Сборник материалов IV Всероссийской научно-практической конференции с международным участием / Под общ. ред. Н.С. Ларькова. – Томск: Томский государственный университет, 2010. – 624 с.

ISBN 5-94621-295-8

Сборник содержит доклады и выступления, подготовленные к IV Всероссийской научно-практической конференции с международным участием «Документ как социокультурный феномен» (Томск, 29–30 октября 2009 г.).

УДК 002
ББК 70

Редакционная коллегия:
д.и.н., профессор Н.С. Ларьков (отв. ред.),
д.и.н., профессор Ю.В. Куперт,
д.и.н., профессор О.А. Харусь,
д.и.н., профессор С.Ф. Фоминых,
д.и.н., профессор А.С. Шевляков

Техническая работа с материалами:
И.Е. Яцковец

Издание осуществлено при финансовой поддержке
РГНФ, проект № 09-01-64180 г/Т.

ISBN 5-94621-295-8 © Томский государственный университет, 2010

А.В. Бочаров

РАСХОЖДЕНИЯ ЧЕЛОВЕЧЕСКИХ И КОМПЬЮТЕРНЫХ ПРИНЦИПОВ ИНТЕРПРЕТАЦИИ НЕСТРУКТУРИРОВАННОГО ТЕКСТА В КОНТЕКСТЕ АВТОМАТИЗАЦИИ ТЕМАТИЧЕСКОЙ ИНДЕКСАЦИИ

Все компьютерные поисково-экспертные системы основаны на автоматическом индексировании текстовой информации посредством учёта встречаемых в текстах слов. Однако обработка результатов такой индексации при поисковых запросах или авторубрикации документов даёт много информационного мусора, что в ситуации современного информационного взрыва ведёт к накоплению непродуктивных затрат времени и средств.

До сих пор эту проблему пытаются решить (например, в поисковых интернет-сервисах), используя семантические алгоритмы уже имеющихся словарных индексных баз данных, в которых указано, какое слово сколько раз и по каким сетевым адресам встречается. При этом с избавлением от одного информационного мусора в информационных запросах нередко только заменяют его новым.

В общем виде задача, рассматриваемая здесь автором, такова: для более интеллектуального и эффективного поиска необходимо также построение индексов тематического авторубрицирования (доминирования контекстов, соотношения предметных областей, степени представленности сфер описываемой реальности). Для решения этой задачи необходимо разработать алгоритм, в котором будет реализована своеобразная аппроксимация (приближение) компьютерной числовой оценки доминирования темы в тексте к субъективному человеческому восприятию большего или меньшего доминирования темы. Прежде чем понять возможную логику и, может даже, философию такого алгоритма, рассмотрим сначала конкретные примеры, показывающие действительную практическую актуальность заявленной проблематики.

Рассмотрим некоторые показательные примеры неизбежности хаоса избыточной информации при традиционной словесной индексации даже в наиболее продвинутых системах семантического анализа. Ни одна поисковая или экспертная программа (по крайней мере, для русского языка) не может корректно определить, о какой предметной области (сфере жизни, тематическом контексте) идёт речь, к примеру, в сообщениях о каких-либо мероприятиях (конференциях, заседаниях, выставках, ярмарках, круглых

столах и т.д.). Программы могут только выдавать термины, которые встречаются рядом с интересующим понятием, но эти термины во многих случаях могут не иметь отношения к тому, что нужно для действительно полезной и функциональной тематической рубрикации. По такому принципу анализируют и выдают информацию современные поисковые метамашины Teoma, Quintura, Clusty и Nigma и подобные им, автоматически структурирующие по определённым понятийно-терминологическим разделам результаты запросов, сделанных на платформе других поисковых машин. Окончательное структурирование полезной информации всё равно придётся делать вручную. То же самое можно сказать о лучшей на сегодняшний день экспертной системе по обработке неструктурированной текстовой информации на русском языке RCO (официальный сайт http://www.rco.ru), которая ниже будет рассмотрена подробнее.

Самые впечатляющие возможности в системе RCO имеет модуль фактографического анализа текстов (RCO Fact Extractor). Это модуль автоматического извлечения из текста описаний ситуаций – событий и фактов определенного типа, с определенными участниками (покупатель, продавец, товар и т.п.). Общее количество семантических шаблонов в библиотеке модуля составляет около 700. Примеры шаблонов по тематике «Кадры»: «Должности» (Должностное лицо – персона; Должность – занимаемая или освобождаемая должность; Организация – место работы); «Забастовки персонала» (Организация – название; Участник – участники акции протеста, сотрудники организации; Акция – акция протеста). «Договоры» (Участник 1 – первая сторона (организация, персона); Участник 2 – вторая сторона (организация, персона).

Таким образом, описание событий и ситуаций из неструктурированного текста преобразовывается в стандартизированный структурированный вид. При автоматическом наложении шаблонов на неструктурированный текст теоретически предполагается соответствие смыслового контекста интересующей тематике, но практически соответствие тематике предварительно не учитывается. Поэтому наложение шаблона может нередко давать некорректные или абсурдные результаты.

Для проверки этого утверждения демонстрационная версия экстрактора фактов RCO Fact Extractor была использована при анализе массива текстов всех сообщений СМИ Томска за март 2006 г., представленного в электронном архиве региональной администрации. Это почти 5 тыс. инвариантых полнотекстовых сообщений томских газет и региональных теле- и радиоканалов. RCO Fact Extractor было дано задание выделить

в этом массиве текстов не менее 10 тыс. отдельных фактов и составить на основе этого тематически рубрицированное досье.

Ниже приведены только несколько из множества показательных примеров некорректного рубрицирования в фактографическом досье вследствие опоры только на словесную, а не на тематическую индексацию.

К рубрике «Суды и расследования» были, в частности, отнесены следующие сообщения:

– о том, что томские спектакли вынесены на суд театральной общественности Москвы;

– о том, что спикер Томской Думы был подвергнуть «допросу» журналистов;

– о том, что заместитель мэра представил на суд депутатов целевую программу по коммунальной сфере.

К рубрике «Боевые действия» были отнесены сообщения:

– о том, что установка связи Томск-Колпашево позволит «убить несколько зайцев»;

– о том, что гости из сибирских регионов будут соревноваться в беге на лыжах и стрельбе;

– о том, что местные экологи волновались по поводу отстрела перелётных птиц;

– о том, что в Томске происходит большое сражение, связанное с подписанием договоров между коммунальными компаниями.

Таких курьёзов множество, и связаны они с тем, что компьютер пока не понимает смысла слов, он рассматривает только сочетания знаков. Поэтому, когда он встречает в двух разных текстах одинаковое сочетание знаков с разным смыслом, он не сможет различить эти смыслы простым перебором вариантов словарных соответствий. Именно поэтому в приведённом примере программа не может различать, о каких «судах» идёт речь: о тех, которые плавают по морям; о тех, в которых рассматриваются преступления, или о тех, от которых зависит популярность актёров театра. Ищутся все грамматические варианты словоформы «суд», независимо от того, в каком контексте и в каком смысле эти грамматические варианты употреблены. Аналогичные причины ошибок наблюдаются также при рубрицировании путём сравнения с эталонными документами-шаблонами.

В целом система RCO очень эффективна. Именно поэтому она и взята для демонстрации актуальности проблемы тематической индексации. Из тех фактов и рубрик, которые были определены, большинство определе-

но правильно. Однако, во-первых, для значительной части фактов в исследуемой выборке сообщений СМИ тематичекие рубрики определены абсолютно неверно или не совсем корректно. Во-вторых, большой массив предполагаемой целями программы информации по рубрицированию и фактографии вообще будет незамечен в рамках применяемой концепции поиска и извлечения знаний из текста. И это относится к лучшей на сегодняшний день экспертной системе анализа неструктурированной информации на русском языке, уже задействованной на практике и получившей признание в серьёзных консалтинговых компаниях и в информационных отделах силовых органов.

С фразеологизмами можно справиться путём фильтрации и перебора. Можно создать фильтр, который будет отсекать все фразеологизмы со словом «убить», как не относящиеся к тематике боевых действий. Такой способ замедлит работу программы, потребует много ручной работы квалифицированных филологов и программистов, но, тем не менее, он будет справляться с задачей в рамках традиционной парадигмы перебора вариантов. Однако с выражениями, использованными в тексте в переносном смысле, принцип перебора вариантов уже не справляется. «Сражения» и «битвы» могут происходить в самых разных сферах: от битвы за урожай до сражений за подписание договора между коммунальными организациями. Программа, автоматические относящая эти понятия к тематике «боевые действия», отнимет много времени на отсечение информационного мусора.

Эксперт может создавать новые фильтры и менять их содержание, но любое слово всё равно будет иметь разные смыслы в разных контекстах. Если тематический фильтр никак не учитывает возможные контекстно-семантические зависимости лексем, а основан только на переборе их грамматико-морфологических форм, то неизбежно накопление ошибок. Всё это в конечном итоге «пожирает» время аналитиков и документоведов, снижая эффективность их работы.

Многие понятия и слова могут иметь отношение к любым сферам жизни, то есть их конкретный смысл может зависеть от бесконечного множества контекстов. Смысл всего текста вырастает из связанных с этим текстом по тематике других текстов, то есть из культурного контекста. Причём человек может один и тот же контекст наделять разными смыслами, то есть назначение одного из возможных смыслов будет носить вероятностный характер. Вывод из этих примеров таков: когда речь идёт о потенциальной бесконечности, простым перебором словарных фильтров уже не обойтись.

Сравнительная характеристика человеческих и компьютерных принципов интерпретации неструктурированного текста

Критерий сравнения	Определение значений и смыслов слов и выражений	
	человеком	компьютером в имеющихся программах
Учёт окружающего тематического контекста	При определении смысла слов и выражений всегда учитывается тематический контекст окружающего текста (рубрикация по сферам жизни и по жанрам), исходя из всего багажа общекультурных индивидуальных знаний	Компьютер не способен самостоятельно определить смысла слов и выражений, так как в него закладывают только знания о структуре языка, без соотнесения их с общекультурными знаниями
Учёт априорной вероятности значения слова	Человек заранее (a priori) знает, какие слова или выражения однозначно указывают на тематический контекст, какие могут не всегда, но часто, какие – очень редко, какие – никогда, а какие могут указывать сразу на несколько рубрик. Затем (a posteriori) он соотносит это априорное знание с интерпретируемым текстом	В программах a priori предполагается, что отдельные слова и выражения могут только однозначно указывать или не указывать на один из тематических контекстов. То есть априорная вероятность значения слов игнорируется, так как всегда равна единице (100%). В результате a posteriori компьютер не может соотнести вероятность обозначения словом разных смыслов с тематикой окружающего текста
Учёт частот встречаемости слова	Если эксперт накапливает опыт интерпретации понятий в текстах по определённой тематике, то он не делает каждый раз сравнения частот встречаемости всех слов во всех прочитанных по теме текстах	Вероятность принадлежности слов к тематической рубрике вычисляется статистически по средней частоте встречаемости одних слов относительно частот всех встречающихся слов в ограниченной выборке текстов
Необходимость специальной формализованной разметки текста	Человек не нуждается для чтения и понимания текста в дополнительной структурированной и формализованной разметке. Он оценивает тематику и контексты исходя только из непосредственного содержания	Многие программные разработки извлечения информации из текстов и авторубрикации связывают это с дополнительной разметкой текстов. Динамичность web-пространства и вообще любых редактируемых электронных документов делает любую семантическую разметку неактуальной и бесперспективной

Вопрос в том, как научить компьютер вычислять вероятность присутствия в контексте разных смыслов, и, уже исходя из этих вероятностей, определять то, каким смыслом обладает интересующее слово.

В первую очередь разберёмся, чем отличаются человеческие и компьютерные принципы интерпретации неструктурированного текста. Эти различия представлены в таблице.

Обобщая, можно констатировать, что все расхождения человеческих и компьютерных принципов интерпретации неструктурированного текста исходят из одного базового отличия: обладание человеком априорной информацией о вероятных смыслах, которые могут быть в конкретном тексте в конкретной ситуации. Это означает, что для выполнения поставленной задачи нужно найти математическую концепцию и соответствующий ей математический аппарат, работающие с априорными вероятностями. Теоретические и прикладные аспекты решения данной задачи выходят за рамки данной публикации и требуют специального, более подробного рассмотрения.