

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
КАФЕДРА ОТЕЧЕСТВЕННОЙ ИСТОРИИ

---

# **ЧЕЛОВЕК – ТЕКСТ – ЭПОХА**

**Выпуск 4**

**Аналитические практики и перспективы  
современного источниковедения**



**ИЗДАТЕЛЬСТВО ТОМСКОГО УНИВЕРСИТЕТА  
2011**

# **I. ИСТОЧНИКОВЕДЕНИЕ XXI в.: АНАЛИТИЧЕСКИЕ ПРАКТИКИ И ИССЛЕДОВАТЕЛЬСКИЕ ПЕРСПЕКТИВЫ**

**А.В. Бочаров  
АВТОМАТИЗАЦИЯ ОБРАБОТКИ  
НЕСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ  
И ПЕРСПЕКТИВЫ ГУМАНИТАРНЫХ НАУК В XXI в.<sup>1</sup>**

Одной из глобальных проблем XXI в. является так называемый «информационный взрыв», или, иными словами, рост диспропорции между объёмом информации, произведённой человечеством, и объёмом информации, которую люди способны потребить и усвоить.

Осмысление и решение данной проблемы чрезвычайно актуально для исторической науки, равно как и для всей гуманитаристики. Достаточно указать на то, что в наши дни почти каждая вузовская кафедра имеет возможность регулярно издавать один или несколько объёмных сборников статей или коллективных монографий каждый год. В итоге в узких предметно-тематических областях лавинообразно накапливаются никем не освоенные «завалы» информации. С одной стороны, эта лавина информации вызвана не столько расцветом науки, сколько конъюнктурными потребностями: количество и статус публикаций важнее их качества, оригинальности и практической актуальности. С другой стороны, необходимы специальные стимулы и поводы к прочтению или хотя бы к ознакомлению с огромным количеством отдельных текстов для оценки содержательных свойств. Такие поводы и стимулы может создать только автоматизированная организация баз знаний, в которых всё содержание текстов будет структурироваться и ранжироваться по множеству всех возможных тематических рубрик.

---

<sup>1</sup> Статья подготовлена в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» на 2002–2013 гг.

Информационный взрыв порождает избыток неиспользуемой информации и хаос неструктурированной информации. Поэтому один из главных путей выхода из сложившейся ситуации – превращение неструктурированной информации в структурированную. На этом пути будут изменяться способы организации и цели исследовательской деятельности во всех гуманитарных науках.

Исторических дисциплин это должно коснуться в наибольшей степени. Во-первых, потому что весь массив накапливаемой и сохраняемой в электронном виде текстовой информации – это потенциальный архив исторических источников; во-вторых, потому что историческая наука способна аккумулировать в себе концепции и методы всех остальных гуманитарных дисциплин.

Превращение неструктурированной информации в структурированную – главная цель контент-анализа текстов, являющегося особым способом сокращения пути читателя от языковых средств выражения к идеям. Субъектами структурирования информации могут быть как люди (ручной контент-анализ), так и компьютерные программы (автоматизированный контент-анализ). Ручной контент-анализ в современных условиях в любом случае связан с использованием компьютера и возможностей частичной автоматизации поиска единиц текста. Автоматизация контент-анализа – это нечто большее. По сути, это замена человеческого интеллекта искусственным. В случае полной автоматизации анализа экспертной компьютерной программой без участия человека производится толкование текстов, сведение его содержания к названиям и обозначениям тематик, фактов или смыслов для отображения их в виде: 1) свойств и атрибутов файлов, 2) иерархического дерева папок, 3) полей базы данных. Исследователю остаётся только указать, какие тексты нужно изучить, и нажать на одну, максимум на несколько кнопок.

В табл. 1. приведена авторская система определений понятия «Неструктурированная текстовая информация». При автоматическом анализе содержания текстовых документов посредством контент-анализа происходит перевод всех свойств текста из левого столбца таблицы в правый столбец.

Система определений понятия  
«Неструктурированная текстовая информация»

Отличительные признаки	Текстовая информация	
	Неструктурированная	Структурированная
Форма текста	Нестандартизированный и неформализованный текст, состоящий из предложений на естественном языке	Стандартизированный или формализованный список из символов, слов или словосочетаний
Содержание текста	Полнотекстовое изложение идей, смыслов и сюжетов (свободный текст)	Только краткие обозначения и название тематик, смыслов и сюжетов (строго лимитированный текст)
Образ предметной области (сферы реальности)	Описания реальности не разделяется явно и обязательно на части, которые напрямую сводятся к триаде « <i>сущность – признак – связь</i> »	Описание реальности явно и обязательно разделяется на части, которые напрямую сводятся к триаде « <i>сущность – признак – связь</i> »
Виды практических реализаций текстов	Разножанровые авторские тексты, не имеющие статуса документов, или тексты документов (отчёты, стенограммы, проекты, характеристики, заявления)	Таблицы и списки со значениями текстологических признаков
Уровень единообразия	Единообразие содержания в разных текстах из одного массива сводится к минимуму	Единообразие содержания таблиц и списков сведено к максимуму



Рис 1. Принципы автоматического контент-анализа текстов

Теоретические и методические основы автоматического анализа содержания текстовых документов были разработаны ещё на начальном периоде истории компьютеров в 1960-х гг., и с тех пор концептуальных изменений в данной области не наблюдалось. Одним из ведущих разработчиков этого направления был американский учёный Г. Сэлтон (Gerard Salton). В своей книге «Автоматическая обработка, хранение и поиск информации» (1968 г.) он перечисляет несколько принципов<sup>1</sup>, которые в упрощённом виде представлены на рис. 1.

Современным компьютерным системам доступно использование разного рода многомерных методов глубинного извлечения

<sup>1</sup> См.: Сэлтон Г. Автоматическая обработка, хранение и поиск информации. М., 1973. С. 23–24.

информации (data mining), таких, как, например, кластерный анализ или нейронные сети. Тем не менее в основе всех современных поисково-аналитических систем всё равно лежат те же принципы, что были обозначены Г. Сэлтоном.

Все поисковые системы используют контент-анализ для индексирования файлов с текстовой информацией. Индекс представляет собой таблицу, в которой каждому слову сопоставляется место хранения файла, в котором это слово встречается, и место в этом файле, где стоит это слово. Индексные файлы необходимы для того, чтобы при поисковом запросе пользователя не перебирать каждый раз заново слово за словом всё содержание текстов. Ведь этих текстов могут быть многие миллиарды (например, в Глобальной сети). При поисковом запросе или при построении классификаций документов обрабатываются уже не файлы с текстами, а только эта таблица-указатель, что многократно ускоряет поиски. Разные поисковые и экспертные системы отличаются оптимальностью алгоритмов построения таких таблиц-индексов и эффективностью управления ими.

В то же время поиск информации, основанный на словесной индексации, связан с множеством недостатков (рис. 2).

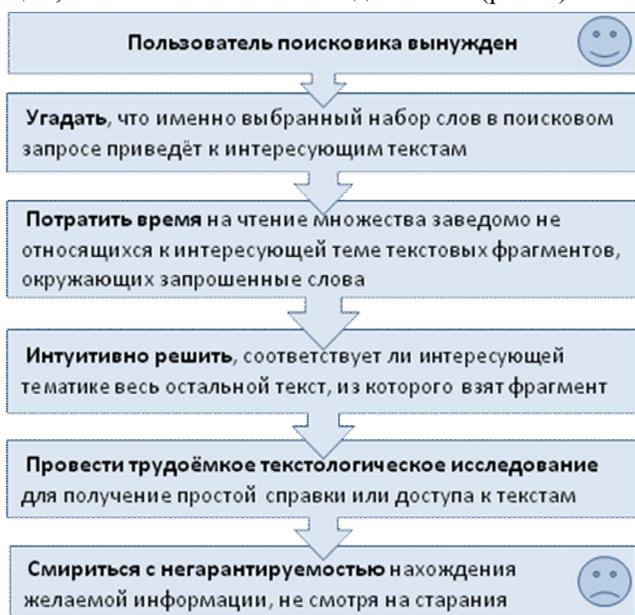


Рис. 2. Недостатки поиска информации, основанного на словесной индексации

От этих недостатков можно было бы избавиться, если бы индексация представляла собой некую тематико-смысловую предопределённую пользователем фильтрацию. Такая фильтрация давала бы возможность упорядоченного выбора из массива текстов, равно как и из набора фрагментов крупных текстов, структурированных по содержательно-смысловым критериям.

Попытка реализовать такую задачу на практике неизбежно ставит вопрос: как сделать, чтобы результаты поискового запроса не только соответствовали (были релевантными) наличию в текстах заданного набора слов, но и соответствовали тематикам, интересующим человека? С этим вопросом связано понятие «*пертинентность*» (от англ. *pertinent* – относящийся к делу, подходящий по сути), используемое в теориях поисково-информационных методов. Количественно пертинентность измеряется как отношение объема полезной информации к общему объему полученной информации.

В отечественной науке соотношение пертинентности и релевантности одним из первых начал изучать выдающийся библиотеквед и социолог А.В. Соколов<sup>1</sup>. Из современных русскоязычных авторов, наиболее активно касающихся этой проблематики, стоит назвать также специалиста в области семантического обеспечения интернет-сервисов Д.В. Ландэ<sup>2</sup>.

Если приблизить модель контент-аналитических подсчётов к модели тематического толкования текстов человеком, то организация и результаты информационного поиска станут более пертинентными конкретным информационным потребностям пользователей, а не просто релевантными запросу на встречаемость слов и словосочетаний. В настоящей статье предлагается способ повышения пертинентности информационных запросов эффективного поиска путём использования автоматизированного построения индекса тематического рубрицирования. До сих пор корректное тематическое рубрицирование текста было под силу проводить только человеку-эксперту.

В тематическом индексе с каждым целостным фрагментом текста соотносится перечень тематических рубрик (контекстов), для каждой из которых указаны её уровень или степень (вес, доля, процент) доминирования по отношению к другим контекстам в данном

---

<sup>1</sup> См.: Соколов А.В. Метатеория социальной коммуникации. СПб., 2001; *Он же*. Общая теория социальной коммуникации. СПб., 2002.

<sup>2</sup> См.: Ландэ Д.В. Поиск знаний в Internet. М., 2005; *Он же*. Основы интеграции информационных потоков. Киев, 2006.

фрагменте текста. Уровень доминирования может принимать значения в диапазоне от 0 до 100 или от 0 до 1. В качестве фрагмента текста может выступать как весь относительно небольшой текст (например, сообщение СМИ), так и явно выделяемая часть крупного текста (абзац, глава или раздел книги или статьи). Например, сообщение в прессе о ремонте в больнице может быть на 40% посвящено теме «ЖКХ», на 40% теме «Медицина и здравоохранение» и на 20% теме «Муниципальные власти».

В традиционных поисковых системах одно и то же слово ищется в разнотематических текстах. В отличие от них система с тематической индексацией позволила бы искать правильно понятые по смыслу синонимические и гипонимические ряды в документах по одной тематике и, что особенно важно, без выдачи информационного мусора.

Следует отметить, что в последние годы уже начали появляться интернет-сервисы, которые по ключевым словам вычисляют степень присутствия на web-странице некоторых наиболее общераспространённых тематик и жанров. Например, они могут отличать научный текст от остальных жанров. Из русскоязычных интернет-сервисов, способных на такое, можно указать сервис «Семантическое Зеркало» <http://www.ashmanov.com/tech/semantic/demo>. Тем не менее вычислить степень доминирования узких тематик, интересующих пользователя, такие сервисы не способны.

Попытка практической разработки тематической авторубрикации текстов неизбежно будет начинаться с концептуальной проблемы понимания отличий человеческого и компьютерного способов интерпретации смыслов текста. В чём суть этих различий?

Существующие компьютерные программы анализа текстов учитывают ключевые слова, исходя из предположения, что они тождественны по *вероятности* обозначения одного и того же смысла и для всех них эта вероятность равна 100%. Однако в реальности эта вероятность разная, так как в разных контекстах одно и то же слово может иметь разные смыслы. Именно поэтому ни одна поисково-экспертная программа не может самостоятельно и корректно определить, о какой узкой предметной области идёт речь.

Человек, в отличие от компьютера, заранее (*a priori*) знает, какие слова или выражения однозначно указывают на тематический контекст, какие не всегда, но часто – очень редко, какие – никогда, а какие могут указывать сразу на несколько рубрик. Затем (*a posteriori*) он соотносит это априорное знание с интерпретируемым текстом.

Отличия человеческого и формального компьютерного способов анализа текстов можно осмыслить посредством теории систем. В рамках теории систем тематическая рубрика текста (контекстный смысл) – это продукт конкретной текстовой системы, а взаимосвязь лексем-индикаторов тематической рубрики – это продуцент (взаимодействие частей системы). Отдельная лексема-индикатор (смысловой маркер) без связи с другими ещё не означает систему, т.е. не порождает смысла. Концепция формального перебора словарей индексов как раз и не учитывает продуцент – взаимодействие частей текстовой системы. Соответственно, эта концепция не в состоянии обеспечить правильное распознавание смыслов текста, вложенных в него человеком.

Различия подходов к автоматизации обработки неструктурированного текста стоит также попытаться представить в контексте парадигм научного познания. Работа авторубрикаторов и поисковиков с сочетаниями знаков в словесных индексах – это, по сути, учёт только инвариантных формализмов текстовых конструкций. Такая работа строится на методологии неопозитивизма и структуриализма. Однако с помощью обработки текстов по такой методологии компьютер не в состоянии, к примеру, различить, идёт ли в тексте речь о судах арбитражных или о судах морских. Для программы обработки данных в этом случае словоформа «судах» будет одним и тем же инвариантным формализмом. Математический анализ формализмов текста при таких вычислениях не выходит за рамки того, что в явной форме содержится в тексте. Поэтому обработка текстовых формализмов нуждается в дополнении деконструкцией вероятностных смыслов текста. В основе методологии такой деконструкции лежат постструктуралистский и постмодернистский проекты научного познания. Авторский взгляд на теоретические аспекты такой взаимодополняемости представлен в табл. 2.

Самым распространённым стандартом библиотечной тематической рубрикации по принципу онтологии является универсальный десятичный классификатор (УДК), представляющий собой примерно 143 тыс. концепций, организованных в таксономию. По каждой рубрике выбираются ключевые слова, определённые библиографом при библиографическом описании книги. Проблемы с автоматическим поиском нужного текста на базе УДК возникнут в тех случаях, когда одни и те же ключевые слова относятся к множеству разных рубрик, а такие случаи встречаются в библиографических указателях достаточно часто.

Подходы к обработке текстовой информации  
в разных парадигмах научного познания

Аспекты	Неопозитивистский и структуралистский подход	Постструктуралистский и постмодернистский подход
Таксономический	Иерархическая <b>древовидная</b> жёсткая классификация понятий	Неиерархическая <b>ризом</b> <sup>13</sup> вероятностных значений
Предметный	Отражает <b>внеязыковую</b> реальность предметной области	Отражает <b>языковую</b> реальность предметной области
Когнитивный	Структурированная <b>онтология</b> <sup>14</sup>	Неструктурированный <b>дискурс</b>
Языковой	Фиксирование <b>инвариантных формализмов</b> фразовых текстовых конструкций	Деконструкция <b>вероятностных</b> вариаций культурных <b>смыслов</b> текста

Практически неразрешимым в рамках УДК становится выявление тематической рубрики не целого текста, а его отдельных фраг-

<sup>13</sup> Ризома – утвердившееся в современных гуманитарных науках понятие постструктурализма и постмодернизма. Оно разработано в книге французских исследователей Ж. Делеза (Deleuze) и Ф. Гваттари (Guattari) «Ризома» (1974). Непосредственно французский термин rhizome заимствован из ботаники и означает специфическую форму корневища, не обладающего четко выраженным центральным подземным стеблем. Понятие введено в гуманитарные дисциплины в противовес понятию «структура» как четко систематизированному и иерархически упорядочиваемому принципу организации. Именно характер ризомы в наибольшей степени соответствует свойствам автоматизированной тематической индексации.

<sup>14</sup> Термин «онтология» в данном контексте используется в том смысле, в котором он распространился и употребляется специалистами в области семантических подходов к разработке баз данных. См., например: *Овдей О.М., Проскудина Г.Ю.* Обзор инструментов инженерии онтологий // Тр. 6-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции. М., 2004; *Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В.* Онтологии и тезаурусы: Учеб. пособие. Казань; Москва, 2006.

ментов. Для ускорения и автоматизации используемой всеми консективной компиляции на заданную тему решение таких затруднений было бы чрезвычайно важным. Решением может стать указание для каждого текста или его фрагментов не ключевых слов, а сопоставительных уровней доминирования разных тематических рубрик. Тогда при поиске станет ясно, в какой степени относится то или иное ключевое слово к интересующей нас рубрике. Однако проблема заключается в том, как адекватно вычислить уровень доминирования тематики.

Некоторые специалисты предлагают подсчитывать сумму отношений слов из тезауруса или из эталонных текстов по тематике к количеству слов в каждом предложении и отношение полученной величины к размеру текста<sup>15</sup>. Не исключено, что эта процедура может оказаться чрезмерно избыточной, сложной и ещё более объёмной, чем полная словесная индексация. Концепция эталонных текстов, в сущности, предлагает подогнать все новые и нестандартные тексты под один фиксированный по своей структуре и семантике текст (или фиксированную группу текстов, что по сути то же самое). Нет гарантии, что эталонный текст исчерпает всю семантику интересующей нас предметной области. Следовательно, тематическая индексация должна стать менее объёмной, более быстрой, более эвристичной, чем словесная индексация. Только тогда она будет коммерчески выгодной, а значит, и имеющей перспективы широкого распространения.

Для достижения этих целей, на наш взгляд, нужно учитывать расхождения человеческих и компьютерных принципов интерпретации тематики текста и попытаться приблизить компьютерные принципы к человеческим. Как было показано выше, при понимании текста человек восполняет информационную неполноту языковых формализмов априорной информацией о культуре и о языке в целом. Каким образом выразить эту информацию в числовой форме и включить её в формулу?

Ответ на этот вопрос предполагает обращение к математической концепции, учитывающей априорные вероятности. Концепту-

---

<sup>15</sup> См.: *Литинский Ю.В.* Средства информационного поиска и навигации в больших массивах неструктурированной информации // Науч.-практ. конф. «Проблемы обработки больших массивов неструктурированных текстовых документов». Фонд эффективной политики, 21–22 мая 2001 г. Режим доступа: <http://www.fep.ru/text/dataarrays04.html>

ально ответы на поставленные вопросы были даны ещё в конце XVIII в. в так называемой формуле Байеса.

Томас Байес (Thomas Bayes) (1702–1761) – английский священник и выдающийся математик, автор фундаментального исследования в области теории вероятностей «Эссе о решении проблем в теории случайных событий». Эта работа была обнаружена только после его смерти и в 1764 г. опубликована в «Трудах Лондонского королевского общества». Сам Байес, разработав концепцию априорной вероятности, ещё не смог найти нужного математического аппарата для неё. Через 10 лет после его смерти эту задачу решил П. Лаплас. Он представил теорему и формулу Байеса в её современном виде. В самом общем виде, суть формулы Байеса<sup>16</sup> – вычисление отношения условной вероятности, введённой априорно человеком, к полной вероятности всех возможных событий в изучаемой ситуации<sup>17</sup>.

Введение в математическую формулу неких «предпосылочных убеждений» может создать впечатление, что любой человек может прийти к любому выводу на основе одних и тех же данных. Это привело к тому, что теорему Байеса многие учёные XX в. посчитали субъективной и ненаучной. Вместо неё всю вторую половину XX в. приоритетными оставались статистические методы, требовавшие предоставить так называемую величину  $P$ , показывавшую вероятность ошибочности выводов для эмпирического апостериорного распределения статистических параметров, описывающих выборку данных. Ведущие научные журналы отказывались публиковать статьи, где статистические данные не сопровождалась величиной  $p$ -уровня. В исторической науке использование статистических методов для массовых исторических источников также неизменно опиралось и опирается на вычисление  $p$ -уровня.

Однако некоторые учёные всё же предупреждали, что величина  $p$  показывает достоверность выводов всего лишь при априорном условии, что имело место счастливое стечение обстоятельств, т.е.

---

<sup>16</sup> Формула Байеса:  $P(A|B) = P(B|A) / P(A)P(B)$ , где  $P(A)$  – априорная вероятность гипотезы  $A$  (смысл такой терминологии см. ниже);  $P(A|B)$  – вероятность гипотезы  $A$  при наступлении события  $B$  (апостериорная вероятность);  $P(B|A)$  – вероятность наступления события  $B$  при истинности гипотезы  $A$ ;  $P(B)$  – вероятность наступления события  $B$ .

<sup>17</sup> См.: Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. М., 2001. С. 68–69.

случайность, вводящая в заблуждение, а вовсе не искомая закономерность. В связи с этим во все области науки, в том числе гуманитарные, проникло и продолжает проникать множество абсурдных или сфальсифицированных результатов, получившихся случайно, без учёта всех математических тонкостей, при которых корректными будут оценки на основе  $p$ -критерия<sup>18</sup>.

Эти тенденции привели к тому, что сегодня популярность байесовской парадигмы постоянно растёт. Она задействуется, например, при так называемой «персонализации», когда системы информационных услуг автоматически учитывают в качестве априорной вероятности текущих интересов пользователя содержание его предыдущих запросов. Исходя именно из байесовского подхода, результаты поиска Google иерархически формируются из списков сайтов, которые выбирали другие люди, подавшие аналогичный запрос.

Самая знаменитая технология управления неструктурированной информацией, основанная на байесовском подходе, – это английская корпорация Autonomy [www.autonomy.com](http://www.autonomy.com). Её создатель Майкл Линч (Michael Lynch) утверждал, что с помощью формулы Байеса можно интерпретировать текст независимо от того, на каком языке он написан, т.е., по сути, без учёта дискурсивно-смысловых особенностей. Для простых производственных задач автоматизации документооборота и ускорения консалтинга клиентов этот подход оправдан. Однако такой подход не привёл и не сможет привести к созданию ожидаемого всем интернет-сообществом интеллектуального поисковика.

Решить задачу по созданию интеллектуального поисковика нового типа могут модели, более глубоко учитывающие вероятностную природу человеческого языка. О вероятностном восприятии языка одним из первых стал писать выдающийся отечественный учёный В.В. Налимов в ставшей уже культовой книге «Вероятностная модель языка. О соотношении естественных и искусственных языков» (1-е изд. 1974 г.). В.В. Налимов рассуждал о том, что смысл воспринимаемых слов расплывчатый, размытый, нечёткий, поэтому любое слово может в определённом контексте и ситуации обозначать любой смысл и любой предмет, а разные смыслы и

---

<sup>18</sup> См.: Мэтьюз Р. 25 великих идей. Научные открытия, изменившие мир / Пер. с англ. СПб., 2007. С. 138–141.

предметы могут с разной вероятностью обозначаться одним и тем же словом. В.В. Налимов предполагал, что, опираясь на теорию Байеса, в итоге можно получить некое «взвешенное произведение» смыслов текста<sup>19</sup>. Современный уровень развития вычислительной техники позволяет на практике реализовать теоретические прозрения В.В. Налимова.

Далее будет представлена авторская реализация байесовского подхода и теоретических прозрений В.В. Налимова. Теоретическим основанием предлагаемого подхода послужило предположение, что понимание смыслов текста зависит от отношения априорного знания к апостериорному, т.е. отношения предыдущего опыта к текущему. Вероятностное описание текущего опыта (в нашем случае текущего текста) может носить эмпирический статистический характер. Однако априорная вероятность получается совсем другим путём. Она скорее носит характер не эмпирический, а мировоззренческий. Она отражает уже имеющиеся знания о какой-либо предметной области. Она неизменна в ходе обработки данных. Априорная вероятность наличия в тексте какой-либо тематики воплощает некие когнитивные константы, отражающие константы культурные и языковые.

Аналитик, использующий байесовский подход, обладает численным выражением анализируемого неизвестного параметра ещё до начала сбора исходных статистических данных<sup>20</sup>. В нашем случае при вычислении уровня доминирования определённой тематики в тексте вышеупомянутым неизвестным параметром в формуле Байеса может считаться вероятность восприятия читателем данной тематики как доминирующей в конкретном тексте. Получая исходные статистические данные (в нашем случае путём контент-анализа текстов), мы присоединяем эти данные к ранее имевшейся информации о параметре (в нашем случае – к информации, полученной путём дискурсивного анализа предметной области, которой посвящены тексты). Дискурсивный анализ позволяет установить тезаурус лексических индикаторов для каждой интересующей нас тематики, которые могут указывать на её

---

<sup>19</sup> См.: *Налимов В.В.* Вероятностная модель языка: О соотношении естественных и искусственных языков. М., 1979. С. 74–95.

<sup>20</sup> См.: *Айвазян С.А., Мхитарян В.С.* Теория вероятностей и прикладная статистика. С. 269–272.

наличие в тексте. В системах авторубрики текстов тезаурусы выполняют роль тематических фильтров. Они отфильтровывают «сорные» слова и оставляют только тематически нагруженные.

Проверка наличия в тексте определённой тематики является для компьютерной программы статистической гипотезой. Как известно, её проверка в общем случае вычисляется как отношение произошедших событий, делающих истинность гипотезы более вероятной, ко всем возможным событиям в изучаемой выборке. В предлагаемой модели тематической индексации выборкой будет лексический состав анализируемого текста. «Всеми возможными событиями» при этом будут употребления в индексируемом тексте лексических индикаторов всех тематик из ограниченного списка тематик, которые нас интересуют. Это будет исходное предположение человека (аналитика-эксперта), что количество именно такое. В любом тексте может быть представлена какая-то доля от этой лексической полноты.

Какие тематики и в каком количестве должны быть в индексе? Чем руководствоваться при ограничении потенциально бесконечного смыслового поля? Будем исходить из того, что тема не должна быть слишком общей и определяться для подавляющего большинства текстов из большого массива (порядка нескольких тысяч) или их фрагментов (более 2/3 полного объёма отдельного текста). В то же время она не должна быть слишком узкой и редкой, встречаясь в слишком малом (менее 1%) количестве текстов или их фрагментов.

Каждое слово или выражение может потенциально выражать множество смыслов относительно разных контекстов и концептов конкретного текста. Необязательно да и невозможно учитывать все эти смыслы и контексты для решения конкретных поисково-прикладных задач. Поэтому величину представленности или доминирования тематики в тексте правильнее будет подсчитывать только относительно других тематик, заранее заданных пользователем в виде ограниченного списка интересующих его тематик. Это позволит учитывать смысловое взаимовлияние разных тематик в рамках одного текста для корректной тематической индексации. Следует подчеркнуть, что в современных поисковых системах такое взаимовлияние не учитывается.

Практическая реализация авторского подхода к тематической индексации выполнена на базе разработанной автором экспертной

компьютерной системы, предварительно названной «Матрица\_СМИ». Задача разработанной экспертной системы – тематическая авторубрикация сообщений СМИ или абзацев публицистической книги. Система генерирует матрицу данных, где одна строка данных соответствует описанию одного сообщения СМИ (либо одного абзаца книги). В столбцах (шкалах) матрицы показывается в процентах уровень доминирования определённой тематики по отношению к другим учитываемым тематикам (рис. 3).

Система реализована на базе языка VBA Excel. VBA Excel в данном случае играет для разработчика роль простого и доступного полигона для проверки алгоритмов. При необходимости коммерческой реализации система может быть преобразована в самостоятельный программный продукт на более мощных платформах. Главное ноу-хау в данном случае – это комплекс логико-математических и лингвистических методик, а не интерфейсные (в широком смысле) возможности.

Сообщения СМИ и публицистика выбраны в качестве источниковой базы не только из-за их актуальности, но и по методологическим соображениям. Дело в том, что язык СМИ отражает универсальную предметную область. В СМИ может быть написано о чём угодно с употреблением лексики от специально-научной до жаргонно-разговорной. Одновременно язык СМИ – это очень динамичное информационное поле, связанное с отслеживанием всех тенденций в жизни общества и в выразительных средствах языка. Поэтому наиболее продуктивной и показательной разработкой и отладкой системы автоматической тематической индексации текстов была бы именно для текстов СМИ.

В основу предлагаемой компьютерной системы была положена разработанная автором модель поисковой системы, ориентированной на пертинентность результатов запроса. На рис. 4 эта модель оформлена в виде когнитивной семантической карты и визуально обобщает взаимосвязь лингвистических, математических и поисково-сетевых аспектов, на которых может быть основана тематическая индексация. Далее эти аспекты будут раскрыты подробнее.

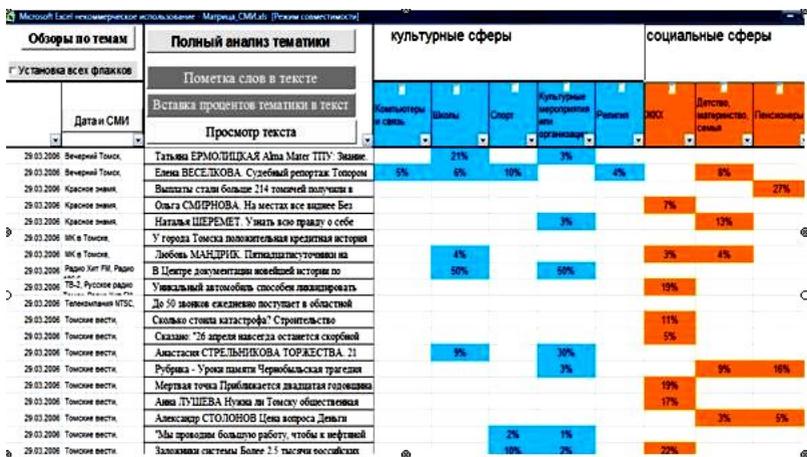


Рис. 3. Фрагмент тематической индексации сообщения в системе Матрица\_СМИ

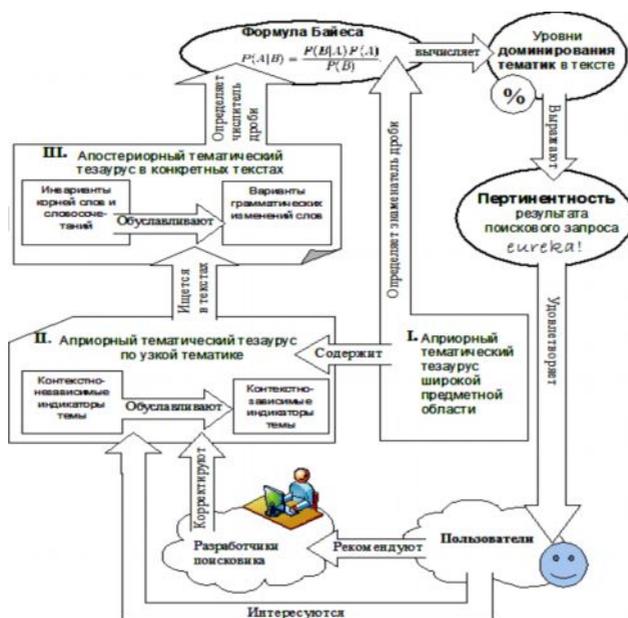


Рис. 4. Прикладная модель поисковой системы, основанной на пертинентности результатов запроса

Итак, в предлагаемой системе тематической индексации Матрица\_СМИ дискурс публицистики и региональных неспециализированных СМИ лёг в основу словарей-тезаурусов, составленных для вычисления доминирования разных тематик в сообщениях. Степень доминирования тематики представляет собой числовое выражение, которое складывается из трёх компонентов (рис. 4):

I. Априорный тематический тезаурус широкой предметной области. Именно он позволяет учесть соотношения в тексте нескольких тематик, поскольку суммирует представленность в анализируемом тексте тезаурусов для всех учитываемых тематических рубрик.

II. Априорный тематический тезаурус по узкой тематике. Он устанавливает количество *контекстнонезависимых* и *контекстнозависимых* индикаторов по отдельным тематическим рубрикам.

III. Апостериорный тематический тезаурус в конкретных текстах. Он определяет количество всех встречающихся в тексте лексических индикаторов.

Первые два компонента «делают вклад» в априорную вероятность доминирования темы, 3-й компонент – в апостериорную вероятность. Для первых двух компонентов состав и количество лексем-индикаторов заранее задан и предсказуем. В 3-м компоненте мы имеем дело со всем потенциально возможным лексическим составом, т.е. с языковым богатством, которое произвольно могло формировать смыслы в тексте. Согласно модели, 1-й компонент определяет знаменатель дроби в формуле Байеса, остальные компоненты – числитель дроби.

Рассмотрим вышеназванные три компонента подробнее.

И 1-й и 2-й компоненты содержат контекстнонезависимые индикаторы тематических рубрик. Эти индикаторы представляют собой лексемы или идиомы, которые всегда однозначно (чрезвычайно высокой вероятностью и ассоциативной устойчивостью) служат индикатором определённой тематики (контекста) в определённой дискурсивной практике.

Эксперименты автора на базе системы Матрица\_СМИ показали, что контекстнонезависимые индикаторы следует по-разному учитывать для тематик с объёмным («богатым») тезаурусом (например, тематики «Спорт» или «Торговля») и для тематик с небольшим («бедным») тезаурусом (например, тематика «Городская дума» или «Областная администрация»). Для тематики с богатым

тезаурусом в тематический фильтр полностью входят части словообразовательных гнезд, ассоциируемых с интересующей тематикой. При этом учитываются инвариантные словоформы без учёта изменяемости при склонении и спряжении. Для тематики с бедным тезаурусом учитываются все инвариантные словоформы с учётом изменяемости при склонении и спряжении. В обоих случаях учитываются все идиомы, устойчиво ассоциируемые с тематикой.

Следующий элемент расчётов – это контекстнозависимые лексемы. Они могут служить индикатором тематической рубрики только при наличии в одном текстовом фрагменте с этими лексемами хотя бы одного контекстн~~о~~зависимого индикатора данной рубрики. Например, слова «судья» или «суд» (в любых склонениях) могут служить индикатором тематики «Суды и расследования» или «Преступность» только при наличии рядом с ними других контекстн~~о~~зависимых индикаторов, например «прокурор», «преступн», «судебн» и т.п. Дело в том, что в переносном значении суды и судьи могут упоминаться в самых разных сферах жизни, а словоформа «суда» и вовсе может означать «морские суда».

Итак, для каждой тематической рубрики (предметной области) заранее создаются, а в процессе индексации текстов корректируются словари тезаурусы, содержащие контекстн~~о~~зависимые и контекстн~~о~~зависимые лексические индикаторы. Тематическая индексация в системе Матрица\_СМИ производится по следующим предметным областям, соответствующим разным сферам жизни:

**Властные сферы:** 1) Федеральные власти, правительство, Госдума РФ; 2) Областная администрация; 3) Областная дума; 4) Муниципалитет и мэрия; 5) Гордума; 6) Коррупция.

**Экономические сферы:** 7) Торговля, сфера обслуживания; 8) Производство, промышленность; 9) Банковский бизнес, акционирование, страхование; 10) Бюджетные финансы, налоги; 11) Нефть, газ, уголь, топливо, полезные ископаемые; 12) Атом; 13) Сельское хозяйство и село; 14) Трудоустройство и доходы населения; 15) Экология и природа; 16) Томские компании и предприятия (упоминаемые названия).

**Культурные сферы:** 17) Вузы; 18) Наука; 19) Компьютеры и связь; 20) Школы; 21) Спорт; 22) Культурные мероприятия или организации; 23) Религия.

**Социальные сферы:** 24) ЖКХ; 25) Детство, материнство, семья; 26) Пенсионеры; 27) Транспорт; 28) Здоровье и медицина.

**Силовые сферы:** 29) Преступность, судопроизводство, МВД, ФСБ, МЧС; 30) Армия.

Именно такая, а не иная специализация разных сфер жизни в виде отдельных тематик обусловлена одним из возможных представлений о тематической типологизации сообщений СМИ. В зависимости от потребностей пользователя экспертной системой возможна любая иная разбивка тематических категорий на специализированные подтемы. Например, тематику «Коррупция» также можно отнести в категорию «Силовые сферы и преступность» или в категорию «Экономические сферы».

Если какая-то из индексируемых тематик доминирует в конкретном тексте на уровне 100%, это значит, что в данном тексте никаких других тематик из тех, что учитываются, не присутствует. В остальных случаях несколько тематик присутствуют в текстах с разными уровнями доминирования, сумма которых равна 100%. Согласно предлагаемой модели процент доминирования тематик выражает для пользователя пертинентность результата его запроса при поиске текстов с интересующими тематиками (рис. 4).

При разработке математического аппарата тематической индексации встал важный вопрос: учитывать ли все случаи встречаемости в тексте однокоренных слов и выражений, относящихся к какой-либо тематике, или же учитывать только сам факт наличия в индексируемом тексте лексического инварианта? Иными словами, достаточно ли учитывать лексический индикатор однократно без вариантов его грамматических изменений в тексте?

Сплошной частотный анализ не соответствует *человеческому* восприятию текста и не всегда позволяет отделить по смыслу главную тему от косвенной второстепенной. Тематика тем глубже и шире раскрыта, чем больше по ней инвариантных слов и выражений из тезауруса по данной тематике. Причём некоторые слова и выражения будут относиться к определённой тематике только при наличии в окружающем тексте других, контекстно-независимых лексических индикаторов темы. Именно инвариантные слова и выражения из кратковременной памяти человека по ходу восприятия сообщения затем переходят в долговременные ассоциации и знания, которые порождаются прочитанным или услышанным сообщением.

Согласно предлагаемому здесь подходу достаточно учитывать частотность только единичного вхождения лексического инварианта, а не частоту всех случаев упоминаний леммы в тексте. В подтверждение рассмотрим пример из сообщения о ремонте концертного зала: 10 раз упоминается словосочетание «концертный зал» (тема культуры) и по одному разу упоминаются 10 разных терминов, напрямую связанных с техническими аспектами ремонта (тема ЖКХ). Никаких других терминов по теме «культура» в сообщении нет. Очевидно, что тема ЖКХ в этом сообщении представлена более многогранно и полно. Она доминирует над темой культуры, которая представлена лишь косвенно. Таким образом, абсолютная частота встречаемости одного и того же слова вовсе не связана напрямую со степенью доминирования темы. Человеческий интеллект не нуждается в учёте и подсчёте всех случаев употребления в тексте какого-либо термина, чтобы понять текст. Почему в таких подсчётах обязательно должен нуждаться искусственный интеллект, если они не только замедляют его работу, но и делают интерпретацию текста менее корректной? Искусственный интеллект<sup>21</sup> может выполнять разного рода стандартизированные задачи многократно быстрее и дольше человека. Практический смысл в использовании искусственного интеллекта появляется только тогда, когда он справляется с решением поставленных перед ним содержательных задач не хуже человека. Поэтому очень важно найти пути приближения компьютерных способов интерпретации текста к человеческим.

Эмпирическим экспериментальным путём на базе системы Матрица\_СМИ автор установил, что для пертинентного вычисления степени доминирования определённой тематики достаточно учесть, что та или иная лемма или идиома встречается в тексте в разных склонениях или спряжениях, не перебирая все случаи этих склонений и спряжений в рассматриваемом тексте. Учёт морфологических инвариантов на уровне пустого морфологического шаб-

---

<sup>21</sup> Наука пока понимает только некоторые механизмы человеческого интеллекта, в связи с этим под искусственным интеллектом понимается только логико-вычислительная часть разумной деятельности, реализованная в компьютерных системах. При этом не стоит путать интеллект и сознание. Насколько бы близкой не была имитация человеческого мышления в компьютере, это ещё не будет означать, что он обладает душой и сознанием. Однако эта проблематика лежит за пределами данной статьи.

лона позволяет, с одной стороны, ничего не пропустить, с другой – не делать избыточных подсчётов.

Помимо ускорения индексации, вероятностный учёт инвариантных вхождений приводит к тому, что так называемый спамдексинг (злоупотребление частотой ключевых слов с целью манипулирования поисковыми машинами) потеряет смысл. Нынешние «контент-аналитические войны» между разработчиками поисковых сервисов и разработчиками коммерческих сайтов также потеряют смысл.

В качестве демонстрационного примера вычисления априорной константы индексации для тематик с очень крупным тезаурусом в общих чертах рассмотрим априорный список для тематики «Культурные организации и мероприятия» в региональных СМИ. Всего по данной тематике определено 22 гиперонима (типа культурной деятельности): «Библиотеки», «Просвещение», «Театр», «Архивы», «Понятие сферы культуры», «Путешествия», «Зрители», «Художественность», «Деятели культуры», «Выставки и музеи», «Издательская деятельность», «Скульптура», «Поэзия», «Музыка», «Место проведения культурных мероприятий», «Концерты», «Медиа и кино», «Песня», «Танец», «Праздники местные», «Конкурсы», «Праздники официальные».

К каждому гиперониму привязан априорный список лексических инвариантов гипонимических рядов для определения апостериорного количества индикаторов темы в тексте. Далее, согласно принятому стандарту формализации в поисковом морфологическом шаблоне слов знак «?» означает наличие одного любого символа, а знак подчеркивания «\_» означает один пробел. Например, если в конце искомой словоформы стоит «??», это означает включение в поисковый тезаурус случая склонения или спряжения слова, выраженных окончанием из двух букв. Поскольку весь список гипонимов очень обширен (более 150), здесь приведён шаблон только для гиперонима «Место проведения культурных мероприятий»: культурн??\_центр, ультурн??\_центр, филармон, планетари, Хобби-центр, кинотеатр, кинозал, ТЮЗ, органн?? зал, органн?? зал.

В табл. 3 приведён список контекстнозависимых (второстепенных) индикаторов тематики «Культурные организации и мероприятия» с пояснениями того, почему конкретная лексема или идиома не может считаться первичным контекстнонезависимым индикатором.

Таблица 3

Характеристика некоторых контекстнезависимых индикаторов тематики «Культурные организации и мероприятия»

Характер переносных значений (причина невхождения в первичный список индикаторов темы)	Морфологические шаблоны (инварианты) индикатора	Пример употребления в переносном значении и не по теме «Культура»
не только в сфере культуры	творществ	« <i>законотворчество</i> »
	творческ??_человек	« <i>менеджер фирмы показал себя как творческий человек</i> »
	творческ??_человек	« <i>генеральному директору компании как творческому человеку...</i> »
не только на культурных мероприятиях	_зал_рукоплексал	« <i>зал рукоплексал депутату</i> »
	_оваци	« <i>выступление политика завершилось овациями</i> »
	экскурсия	« <i>экскурсия по квартире</i> »
	экскурсовод	« <i>экскурсовод на фабрике</i> »
	сценарист	« <i>сценарист политического переворота</i> »
метафорическое значение	_сцен?_	« <i>сцена действия</i> » и т.д.
	памятник	« <i>памятник бесхозяйственности</i> » и т.д.
	актер	« <i>он плохой актёр</i> » и т.д.
использование во фразеологизмах	_песн?_	« <i>из песни слова не выкинешь</i> » и т.д.
	театр_	« <i>театр абсурда</i> » и т.д.
	театр?_	« <i>театра абсурд</i> » и т.д.
	театр??_	« <i>театром абсурд</i> » и т.д.

Подводя итог и возвращаясь к теме статьи, обозначенной в её названии, рассмотрим вопрос о том, какие перспективы перед гуманитарными исследованиями открывает использование автоматизированных систем тематической индексации. Можно предположить, что в перспективе реализация идеи тематической индексации приведёт к специализации интернет-поисковиков. Она также позволит поисковым программам операционных систем автоматически создавать рубрицированный каталог всех электронных текстов на персональных компьютерах. Тематический поисковик сможет индексировать файлы и сайты по огра-

ниченному набору тематик из какой-либо узкой предметной области, скажем из истории России, но зато предельно корректно.

В интеллектуальных поисковиках будущего можно будет исключать из результатов поиска не просто нежелательные слова (их всех не предусмотреть), но и нежелательные тематики; регулировать ранжирование результатов поиска по степени доминирования тематики в текстах или в их фрагментах. Тривиальный поиск по словам, конечно, тоже останется, но он будет уже иметь вторичный характер.

Априорные величины для тематической индексации текстов по формуле Байеса могут сформировать только гуманитарии – специалисты в дискурсивных особенностях тех областей знаний, для которых происходит обработка текстов и документов. Речь идёт, в частности, о составлении дисциплинарных, субдисциплинарных и междисциплинарных идеографических словарей нового типа. В таких словарях по каждой предметно-тематической рубрике будет содержаться список и объяснение всех контекстно-инвариантных употреблений определённых словоформ, понятий или идиом. Для обозначения подобных «сборников контекстов» (с перечислением, но без анализа) в корпусной лингвистике используется термин «конкорданс». Однако существующие конкордансы носят исключительно языковедческо-справочный характер. Здесь же имеется в виду создание неких предметных баз знаний.

Эвристические возможности фильтрации, сортировки и компоновки информации в такой базе знаний как раз и откроют новые горизонты гуманитаристики. Они станут действенным поводом для новых форм рефлексии гуманитариев над методами и целями своих исследований, над своими языками и текстами и предоставят широкой аудитории новые поводы заинтересоваться этими текстами, будь то классические философские трактаты, электронные архивы сообщений СМИ за прошлые годы, базы данных с отчётной отраслевой документацией, протоколы заседаний или любые иные корпуса текстов. У гуманитариев, помимо задачи «выписывать из разных книжек в одну тетрадку», появятся новые, более творческие задачи, связанные с поиском способов автоматизации конспектирования, реферирования и аннотирования текстов. Автоматизация настраиваемых субдисциплинарных рубрикации, ранжирования и комбинаторики текстов создаёт новые стимулы их прочтения, новые направления интересов, новые способы интерпретации описанных в текстах событий и явлений. Соответственно, всё это может открыть и новые перспективы для гуманитарных наук XXI в.