М.Е. ГУСАРОВА, К.В. МАКАРОВ

Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени А.Г. и Н.Г. Столетовых»

Система компьютерного контент-анализа исторических документов

УДК 303.642.023, 004.912

В статье показывается возможность использования метода контент-анализа при работе с историческими документами с применением компьютерных технологий.

The paper deals with content analysis possibility in process of the work with historical document by means of computer technologies.

Одним из методов эмпирической социологии является контент-анализ, в ходе которого обрабатывается текстовая или графическая информация. Последняя, в свою очередь, переводится в количественные показатели и подвергается статистической обработке. Как правило, контент-анализ используется в социологических исследованиях либо как самостоятельный метод, либо в сочетании с другими эмпирическими методами – анкетным опросом, интервью и т.д. Процедура достаточно трудоёмкая, особенно при большом количестве текстов. Наша цель – показать возможность использования контент-анализа в исторических исследованиях с применением компьютерных технологий.

Остановимся сначала на методике проведения данного метода. После определения проблемы и постановки задач исследования следует отобрать эмпирический материал, содержание которого будет выступать в качестве объекта контент-анализа. Им могут быть любые источники коммуникации – журналы, газеты, официальные документы, фильмы и т.п. Следует также решить, все ли материалы из отобранных источников необходимо использовать и за какой период (всё это определяется исходя из задач исследования). Следующий этап – операционализация понятий. Обозначаются единицы анализа, которые классифицируются таким образом, чтобы каждая

единица могла быть отнесена только к одной категории. Единицей анализа может выступать как отдельное слово, так и законченный текст. Перевод содержания документов в категориальную схему – это, по сути, кодировка текста, которая осуществляется опытными, специально подготовленными кодировщиками. После кодировки данных можно анализировать полученные результаты, представлять их в виде таблиц и графиков [1]. Именно кодировка и представление результатов считаются самыми трудоёмкими, особенно при большом количестве исходного материала. Поэтому возникает необходимость применения на данных этапах компьютерных технологий [2].

Рассмотрим применение данного метода в исследовании системы образования как составной части социальной политики российского государства в XVIII веке. Для решения обозначенной цели были поставлены следующие задачи: сравнить отношение к системе образования в периоды правления Петра I, Елизаветы Петровны и Екатерины II; выявить, какие преобладали типы учебных заведений, включая и вновь открывающиеся, в обозначенные периоды; определить, как решался в законодательстве вопрос о финансировании образования. В качестве объекта исследования были выбраны материалы Полного Собрания законов Российской империи. Из них были отобраны статьи, относящиеся к периодам правления вышеперечисленных монархов и содержащие информацию по образованию.

Далее были обозначены смысловые единицы текста, объединённые по категориям, и выделены единицы счёта (представлены в табл. 1).

Таблица 1

Классификатор контент-анализа

№ п/п	Категория анализа	Единицы анализа (смысловые единицы)	Единица счета
1.	Открытие учебных заведений в XVIII веке	1) о распространении наук 2) учинить училища (заведение училищ) 3) о школах (школы) 4) о обучении (обучать) 5) школы для обучения	Количество законодательных актов
2.	Финансирование учебных заве-	1) сумма на содержание оных	Количество упоминаний в законодательстве

	дений в XVIII веке	2) на содержание казенных	
3.	Виды учебных заведений в XVIII веке	1) Воспитательное Общество благородных девиц 2) Гарнизонные школы 3) Воспитательное училище 4) Горное училище 5) Сухопутный кадетский корпус (кадетский корпус)	Количество упоминаний в законодательстве по каждому учебному заведению
4.	Изучаемые предметы в учебных заведениях	1) предметы учения 2) изучаемые науки 3) науки	Количество упоминаний в законодательстве
5.	Характер образования в XVIII веке	1) - избирать добровольно хотящих … иных со принуждением - положить штраф - без свидетельствованных писем жениться не допускать 2) - всякого чина людей - доказывать дворянство их	Количество упоминаний в законодательстве
6.	Отношение власти к образованию	- Петр I (1700-1725) - Елизавета Петровна (1741-1761) - Екатерина II (1762-1796)	Количество всех вышеперечисленных единиц анализа по каждому из императоров.

Проведенные работы по определению единиц счета являются необходимыми для разработки программы компьютерного анализа и её внедрения с целью обработки исторических документов.

Процесс построения системы компьютерного контент-анализа состоит из следующих основных этапов:

- анализ структуры документов для обработки;
- формализация процесса анализа документов;
- проектирование структуры системы;
- проектирование структуры хранимых системой данных;
- разработка алгоритмов обработки текстовых документов;
- реализация полнотекстового поиска;
- программная реализация системы;
- апробация системы.

Исходными данными для проектирования информационной системы анализа исторических документов являются: комплект исто-

рических текстов, перечень ключевых слов и фраз для анализа (категорий и единиц анализа), правила трактовки результатов анализа.

Работу системы контент-анализа можно представить следующими взаимосвязанными этапами:

Рис. 1. Этапы работы системы контент-анализа

Представленные этапы легли в основу спроектированных алгоритмов обработки текста, а также структуры самой системы (см. рис. 2). Ключевыми подсистемами являются:

- подсистема взаимодействия с пользователем;
- подсистема обработки исходного текста;
- подсистема поиска единиц анализа;
- подсистема интерпретации результатов.

Одно из ключевых мест в системе занимает база данных, обеспечивающая хранение исходных текстов, перечень параметров анализа текстов, в том числе единиц анализа, результаты анализа. Подобный подход позволяет в любой момент времени возвращаться к уже обработанным текстам и проводить новый анализ, либо поновому интерпретировать полученные ранее результаты.

Рис. 2. Структурная схема системы компьютерного контент-анализа

Проектирование структуры данных и методы дальнейшей работы полностью зависят от выбранной платформы (системы управления базами данных (СУБД)). Исходя из целей разработки, которые указывают в качестве основных операций – операции обработки текстовых массивов была выбрана СУБД PostgreSQL [3]. В пользу выбора именной этой СУБД можно привести следующие аргументы: наличие готовых средств полнотекстового поиска, богатая поддержка лингвистики, включая подключаемые словари. Данный функционал СУБД очень важен в рамках создания системы обработки текстовых документов.

Структура базы данных, используемой в рамках системы контент-анализа приведена на рис. 3.

Разработанная система представляет собой универсальный инструмент, позволяющий проводить анализ произвольных текстов. Это достигается благодаря реализованным алгоритмам контент-анализа, а также используемой для анализа методике.

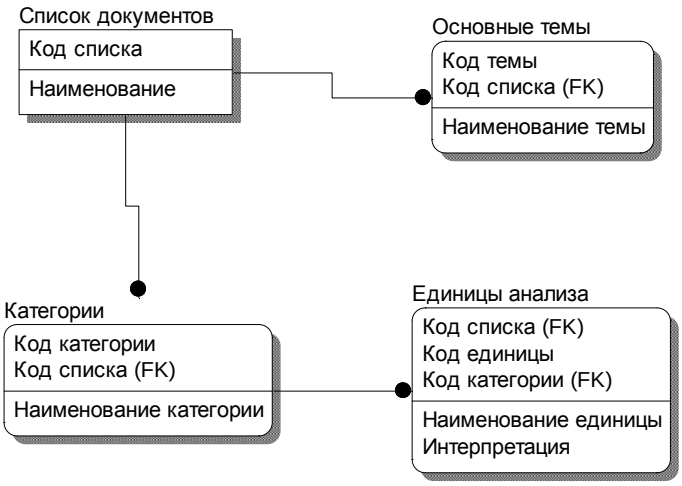


Рис. 3. Структура данных системы контент-анализа

Обобщенный алгоритм функционирования всей системы представлен на рис. 4.

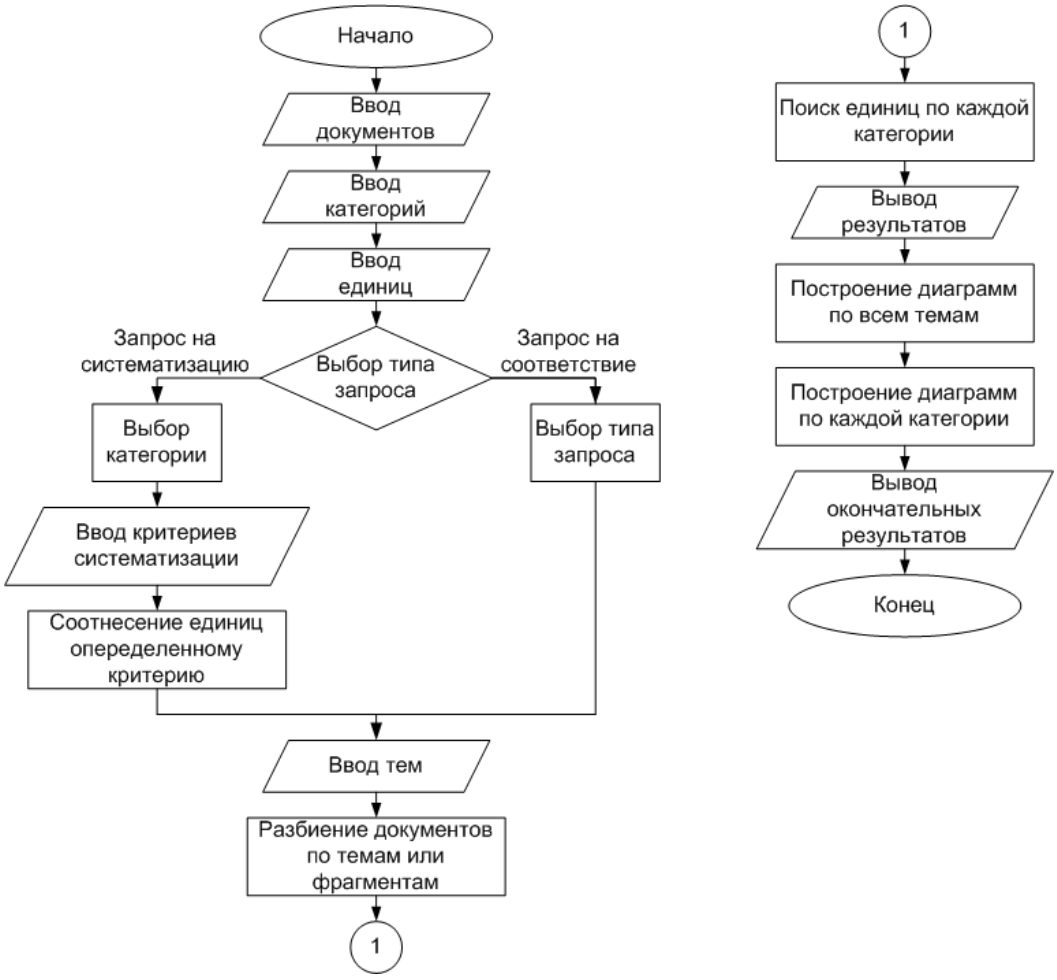


Рис. 4. Блок-схема алгоритма работы системы

Первоначальный ввод категорий и единиц анализа позволяет произвести настройку системы на анализ текстов конкретной тема-

тики. Далее осуществляется либо процесс систематизации данных в представленных документах, либо проводится проверка текстов на соответствие единицам анализа. На следующем этапе происходит разбиение текстов согласно заданным критериям на отдельные темы или произвольные фрагменты. Дальнейшая обработка сводится к поиску единиц анализа в полученных ранее фрагментах.

Полученные результаты анализа помещаются в базу данных. В системе предусмотрены разнообразные возможности по визуализации полученных результатов. Например, в виде диаграмм, показывающих количество найденных в тексте единиц счета (рис. 5).

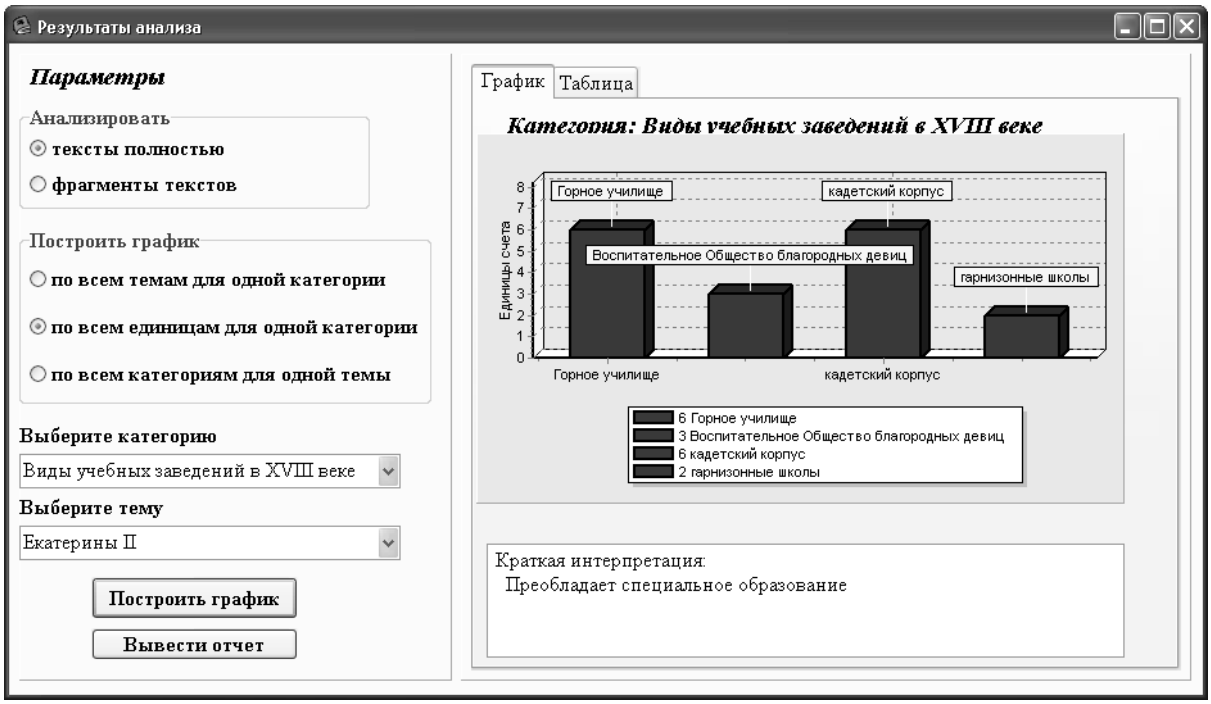
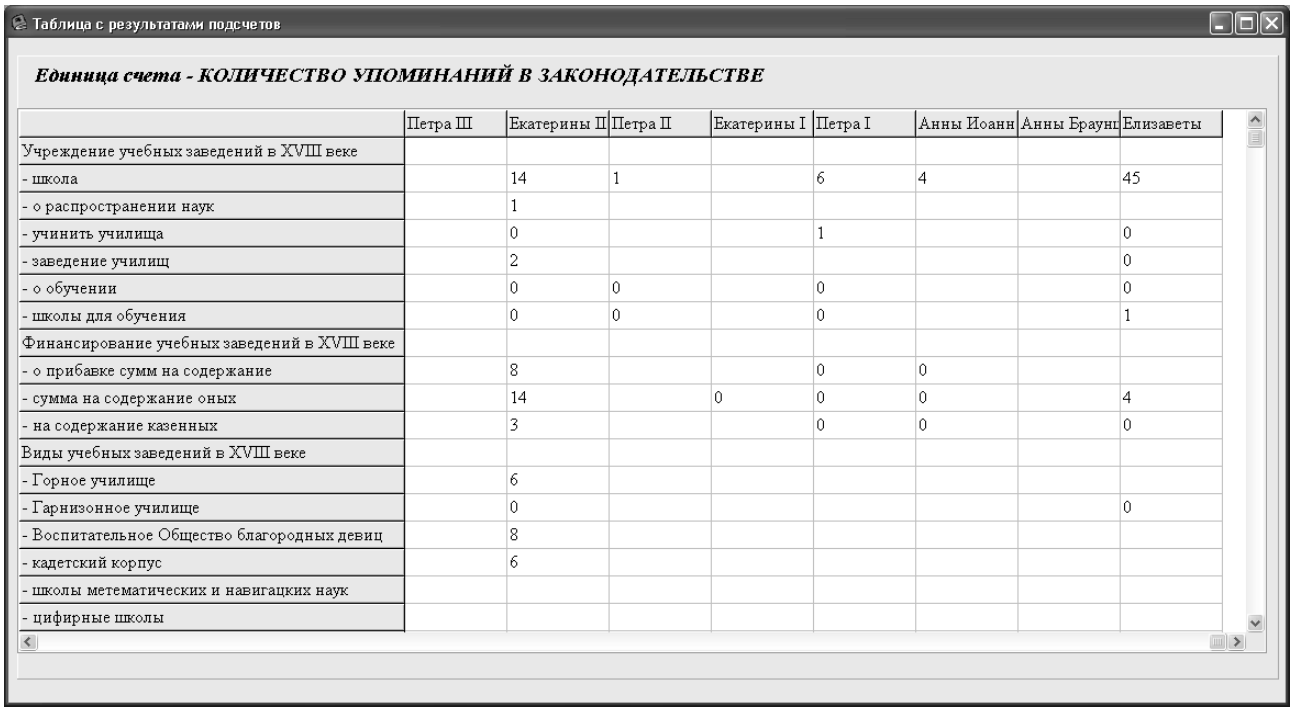


Рис. 5. Результаты анализа текстов: по всем единицам для одной категории

Результаты анализа могут быть представлены в привычной табличной форме (рис. 6). Хранение результатов получаемых в результате проведения каждого из этапов анализа позволяет при необходимости эксперту не только контролировать процесс анализа, но и использовать промежуточные результаты для повышения его гибкости.



	Петра III	Екатерины II	Петра II	Екатерины I	Петра I	Анны Иоанн	Анны Браунш	Елизаветы
Учреждение учебных заведений в XVIII веке								
- школа		14	1		6	4		45
- о распространении наук		1						
- учинить училища		0			1			0
- заведение училищ		2						0
- о обучении		0	0		0			0
- школы для обучения		0	0		0			1
Финансирование учебных заведений в XVIII веке								
- о прибавке сумм на содержание		8			0	0		
- сумма на содержание оных		14		0	0	0		4
- на содержание казенных		3			0	0		0
Виды учебных заведений в XVIII веке								
- Горное училище		6						
- Гарнизонное училище		0						0
- Воспитательное Общество благородных девиц		8						
- кадетский корпус		6						
- школы метематических и навигацких наук								
- цифирные школы								

Рис. 6. Оценка встречаемости единиц анализа по всем текстам

Окончательная обработка и интерпретация полученных результатов (см. табл. 2) проводится человеком-экспертом. В рамках исследования исторических текстов были выделены особенности социальной политики в области образования XVIII века.

Таблица 2

Фрагмент протокола контент-анализа

Единицы анализа	**Единицы счета**	**Петр I**	**Елизавета Петровна**	**Екатерина II**
О распространении наук	Количество законодательных актов	-	-	1
Учинить училища	Количество законодательных актов	-	-	2
О школах	Количество законодательных актов	3	10	9
О обучении	Количество законодательных актов	1	2	7
Школы для обучения	Количество законодательных актов	1	3	5
Общий показатель по категории «открываемые учебные заведения в XVIII веке»		5	15	24
Воспитательное общество благо-	Количество упоминаний в законода-	-	-	4

родных девиц	тельстве по каждому учебному заведению			
Гарнизонные школы	Количество упоминаний в законодательстве по каждому учебному заведению	-	5	1
Воспитательное училище	Количество упоминаний в законодательстве по каждому учебному заведению	-	-	4
Общий показатель по категории «виды учебных заведений в XVIII веке»		4	5	15
Сумма на содержание оных	Количество упоминаний в законодательстве	-	-	5
На содержание казенных	Количество упоминаний в законодательстве	-	-	2
Общий показатель по категории «финансирование учебных заведений в XVIII веке»		0	0	7

Общее суммарное значение по категории анализа «открытие учебных заведений в XVIII веке» составило: Петр I – 5, Елизавета Петровна – 15, Екатерина II – 24 упоминания. Можно предположить, что проблемам образования и просвещения больше внимания уделялось при Екатерине II, даже если учесть, что периоды их правления различались количественно (Петр I правил 29 лет, Елизавета Петровна – 20 лет, Екатерина II – 33 года). Правление Екатерины II в исторической науке носит название «просвещённого абсолютизма». Этим отчасти и можно объяснить такое повышенное внимание к сфере образования.

Значительное внимание в законодательной политике Екатерины II уделялось открытию воспитательных учебных заведений (4 упоминания в законодательстве), в которых осуществлялось и обучение воспитанников, и их полное содержание (в первую очередь это касалось детей-сирот).

Анализ законодательства позволил определить, что большая часть открываемых учебных заведений относилась к уровню специ-

ального образования. Особенно это проявилось в периоды правления Елизаветы и Екатерины II, где значительно выделилась такая смысловая единица, как «гарнизонные школы», что показывает широкую их распространенность на протяжении большей части XVIII века как специального учебного заведения. Такое внимание к гарнизонным школам можно объяснить необходимостью комплектования кадрами регулярной армии, которая при Петре ещё только начинала создаваться.

Подобным образом проводится подсчёт и анализ результатов по всем категориям и смысловым единицам.

Использование созданной программы позволило значительно ускорить процесс проведения контент-анализа и сократить его трудоёмкость. Реализованные в системе механизмы обработки текста обеспечивают возможность проведения анализа текстовых документов, посвященных произвольной тематике.

Литература

1. Социология / Отв.ред. П.Д. Павленок. – М., 2002. – 1036 с.

2. Компьютеризированный статистический анализ для историков / Под ред. Л.И. Бородкина, И.М. Гарсковой – М., 1999. – 187 с.

3. PostgreSQL наиболее продвинутая открытая СУБД в мире. Полнотекстовый поиск в PostgreSQL. [Электронный ресурс]. Режим доступа: www.postgresql.ru.net/docs/fullsearch.html, свободный.– Загл. с экрана.