Государственное образовательное учреждение
Высшего профессионального образования
«Оренбургский государственный университет»

Рыбина Алена Игоревна

Автоматизация процесса определения релевантности текста информационному запросу методом латентно-семантического анализа

230100.68 – Информатика и вычислительная техника

Научный руководитель
кандидат технических наук,
Цыганков А.С.

Оренбург 2015

ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЙ

Объект - информационное и программное обеспечение поисковой системы**.**

Предмет - методы, модели и средства определение релевантности текста поисковому запросу.

Границы исследования - осуществление процесса поиска текстовой информации**.**

Цель: Разработка автоматизированной поисковой системы с повышенной точностью поиска соответствия информационному запросу.

Задачи :

1. Проведение анализа предметной области, определение существующих и разрабатываемых подходов поисковых механизмов.
2. Определение критериев качественного функционирования системы поиска.
3. Разработка поискового алгоритма на основе латентно-семантического анализа.
4. Создание эффективного поискового механизма.
5. Прототип автоматизированной системы использующей предложенный метод определения релевантности текстов.
6. Результаты экспериментального исследования разработанного прототипа и оценки его эффективности.

СХЕМА ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ

СИСТЕМНЫЙ АНАЛИЗ ПРОЦЕССА ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА

- 1.1 Анализ проблем процесса определения релевантности текста
- 1.2 Анализ аналогов поисковых алгоритмов
- 1.3 Концептуальная постановка задачи исследований и её формализация

МЕТОДЫ И МОДЕЛИ СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ТЕКСТА

- 2.1 Исследование моделей описания текстового контента
- 2.2 Развитие модели текстового контента для задачи поиска
- 2.3 Разработка алгоритма семантического представления текстов

РАЗРАБОТКА СРЕДСТВ ПРОЦЕССА ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА

- 3.1 Разработка алгоритма системы определения релевантности текста
- 3.2 Разработка алгоритма определения оптимальных параметров
- 3.3 Разработка алгоритма выявления латентных связей

ИССЛЕДОВАНИЯ ЭФФЕКТИНОСТИ СИСТЕМЫ ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА ИНФОРМАЦИОННОМУ ЗАПРОСУ МЕТОДОМ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА

- 4.1 Методика оценки эффективности поиска информации
- 4.2 Сравнительная оценка эффективности поиска информации
- 4.3 Направления дальнейших исследований

ОСОБЕННОСТИ ЭКСПЛУАТАЦИИ ПОИСКОВЫХ СИСТЕМ

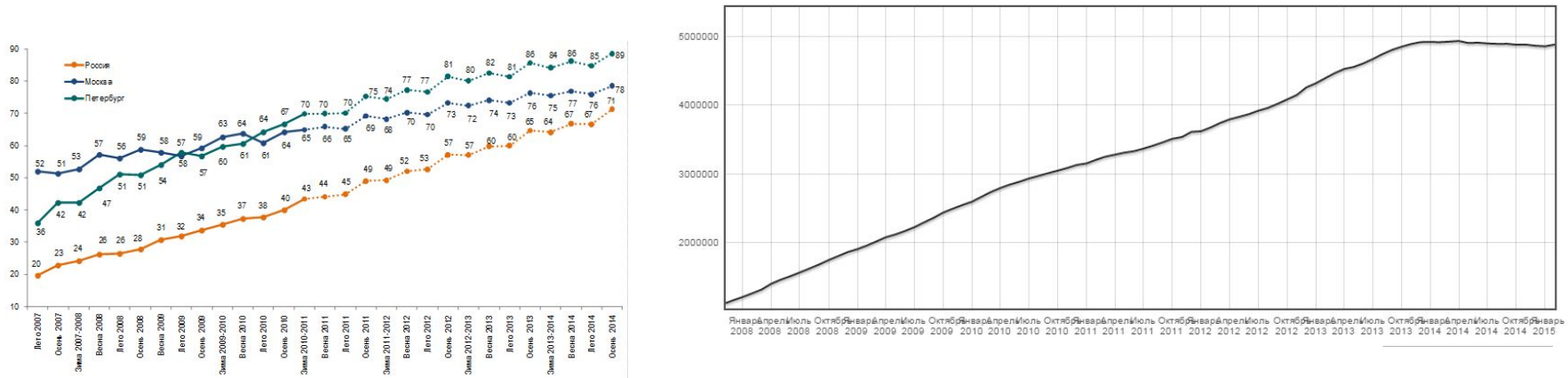


Рисунок 1 – Динамика роста интернет аудитории и количества доменов

Рисунок 2– Динамика изменения возрастного состава интернет-аудитории

По данным TNS Web Index, доля пользователей старшей возрастной группы растёт год от года.

Объект исследования: $OI = \{ Mt \{ Mob \{ S \}\}$ (1.1)

где ***Mt*** – метод поиска релевантной информации;

Mob – модель объекта исследования;

S – средства поиска информации.

ПРОТИВОРЕЧИЯ ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА ИНФОРМАЦИОННОМУ ЗАПРОСУ

Проблемы практики

- Увеличение количества пользователей сети Internet
- Для построения хорошего запроса необходимо уметь использовать специфичный язык запросов поисковых систем.
  Обычно пользователь не обладает достаточной квалификацией.
- Увеличение количества сайтов и web-документов

Проблемы теории

- Методы поиска информации базируются на поиске прямых вхождений слов из запроса в текст и не в полной мере учитывают их семантическое содержание web-документов
- Существующие алгоритмы требуют существенных ресурсов, что снижает производительность поисковых систем

Противоречие между существенно возросшим количеством web-Документов в совокупности с низким уровнем квалификации пользователей и методами поиска, не учитывающими семантическое содержание документа и чувствительными к использованию специфического языка запросов.

Предмет исследования

$$PI = \{ Mt, Mpr, I \}, \quad (1.2)$$

где ***Mt*** – методы поиска информации;
Mpr – модель описания текста;
I – объем информации для определения релевантности.

АНАЛИЗ АНАЛОГОВ ПОИСКОВЫХ АЛГОРИТМОВ

КОНЦЕПТУАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЙ И ЕЁ ФОРМАЛИЗАЦИЯ

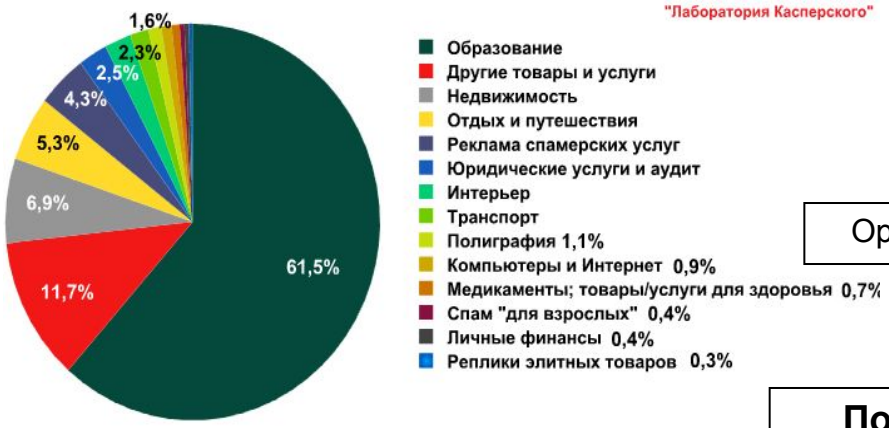


Рис. X – Тематика служебной переписки

Рис.X – Методы классификации

Рис.X – Подходы к задаче фильтрации

Рис.X – Методы борьбы с НЭС

Целевая функция

где R – ошибки поиска;

$L \in \{Let_i\}$ – множество web-документов;

$P=(p_1,p_2,p_3,\ldots p_i)$ пространство признаков, характеризующих L;

A – алгоритм классификации к одному из классов $K \in \{k_1,k_2\}$.

ИССЛЕДОВАНИЯ МОДЕЛЕЙ ОПИСАНИЯ ТЕКСТОВОГО КОНТЕНТА

Векторная модель

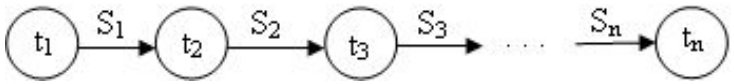
t_n – терм (смысловая единица) в n-ом документе D
(слово, понятие, предложение и т.д.)

- множество термов документа D,

- множество свойств термов t_i в *D.*

Модель на основе графа $\mathbf{D = (t,S)}$

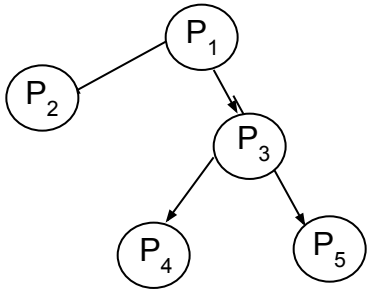
Синтаксическое представление
(дерево зависимостей)



где S1..Sn – расстояние между словами

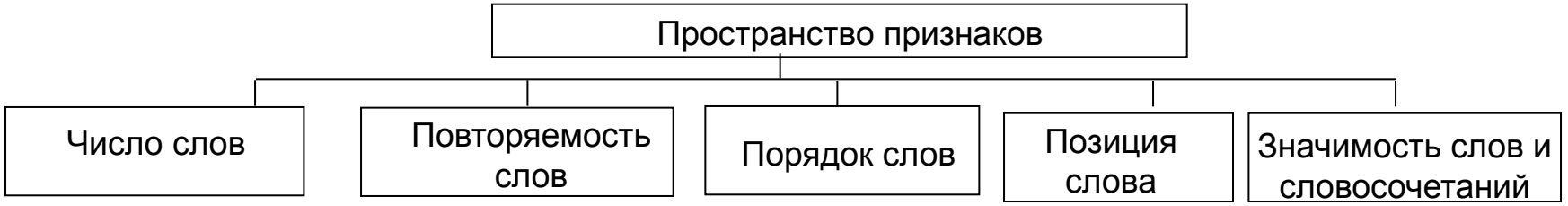
Семантическое представление
(семантические сети, семантический граф)

$\mathbf{D = (P,O)}$



P_n – понятия в тексте
O_n – отношение между понятиями

МОДЕЛЬ WEB-ДОКУМЕНТОВ



Модель web-документа

$$S(p_i)=<t_i, w(t_i)>$$

где t – i-ый терм в документе;

p_i – пространство признаков, определяющих сообщение;

$w(t_i)$ – вес терма в документе после удаления стоп-слов.

где S_j – j-ое сообщение электронной корреспонденции;

ω_{ij} – вес терма i в сообщении j;

N – число термов в сообщении.

$$\mathrm{tf}(w,d) = \frac{n_{wd}}{n_d}$$

где n_{wd} — число вхождений слова в документ, n_d — общее число слов в данном тексте;

$$\mathrm{idf}(w,D) = \log \frac{|D|}{|(d \supset w)|},$$

где $|D|$ — число текстов в корпусе, а $|(d \supset w)|$ — число текстов, в которых встречается w. Теперь

$$\mathrm{tfidf}(w,d,D) = \mathrm{tf}(w,d) \times \mathrm{idf}(w,D).$$

РАЗРАБОТКА АЛГОРИТМА СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ТЕКСТОВ

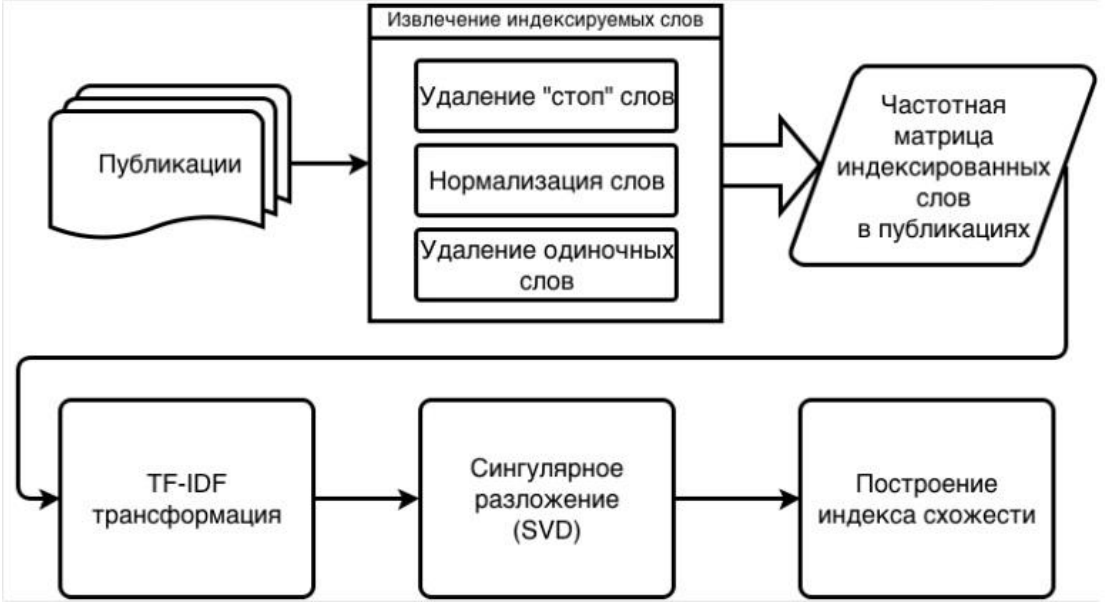
Матрица признаков базы документов *Lk*

$$L_k = <T_k, w(t_j)>$$

где T_k – k-ый терм сообщения;
$w(t_j)$ – вес терма в документе j;

где L_k – база документов k;
ω_{ij} – вес терма i в документе j;
N – число термов в базе;
M – число документов в базе.

РАЗРАБОТКА АЛГОРИТМА СИСТЕМЫ ОПРЕДЕЛЕНИЯ РЕЛЕВАНТНОСТИ ТЕКСТА



Сингулярное разложение матриц

$$A=U\,S\,V^{T},$$

где U и V^T — ортогональные матрицы размером $n*n$ и $m*m$, соответственно,
а S — диагональная матрица с сингулярными числами матрицы A на диагонали.

Диагональные элементы матрицы S имеют вид:

$$S_1>S_2>\ldots>S_n>0$$

РАЗРАБОТКА АЛГОРИТМА ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНЫХ ПАРАМЕТРОВ

- определить влияние параметра *K* на количество шумов в результативной матрице корреляций.

- Нахождение оптимального параметра K, при котором количество шумов будет минимально. $f(x) \to \min_{x \in U}$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1								
c2	0.19	1							
c3	0.00	0.00	1						
c4	0.00	0.00	0.47	1					
c5	-0.33	0.58	0.00	-0.31	1				
m1	-0.17	-0.30	-0.21	-0.16	-0.17	1			
m2	-0.26	-0.45	-0.32	-0.24	-0.26	0.67	1		
m3	-0.33	-0.58	-0.41	-0.31	-0.33	0.52	0.77	1	
m4	-0.33	-0.19	-0.41	-0.31	-0.33	-0.17	0.26	0.56	1

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1								
c2	0.91	1							
c3	1.00	0.91	1						
c4	1.00	0.88	1.00	1					
c5	0.85	0.99	0.85	0.81	1				
m1	-0.85	-0.56	-0.85	-0.88	-0.45	1			
m2	-0.85	-0.56	-0.85	-0.88	-0.44	1.00	1		
m3	-0.85	-0.56	-0.85	-0.88	-0.44	1.00	1.00	1	
m4	-0.81	-0.50	-0.81	-0.84	-0.37	1.00	1.00	1.00	1

Рисунок 5 – Корреляция в исходной матрице и в преобразованной

$$P(t_k|d_i, w_j) = \frac{P(w_j|t_k)P(t_k|d_i)}{\sum_{l=1}^{|T|} P(w_j|t_l)P(t_l|d_i)+\gamma P_{noise}(w_j|d_i)+\epsilon P_{background}(w_j)}.$$

$$P_{noise}(w_j|d_i) = \frac{\gamma P_{noise}(w_j|d_i)}{\sum_{l=1}^{|T|} P(w_j|t_l)P(t_l|d_i)+\gamma P_{noise}(w_j|d_i)+\epsilon P_{background}(w_j)}.$$

$$P_{background}(w_j) = \frac{\epsilon P_{background}(w_j)}{\sum_{l=1}^{|T|} P(w_j|t_l)P(t_l|d_i)+\gamma P_{noise}(w_j|d_i)+\epsilon P_{background}(w_j)}.$$

РАЗРАБОТКА АЛГОРИТМА ВЫЯВЛЕНИЯ ЛАТЕНТНЫХ СВЯЗЕЙ

ПРОГРАММНЫЙ ПРОЕКТ ПРОТОТИПА СИСТЕМЫ КОНТЕНТНОЙ ФИЛЬТРАЦИИ ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ

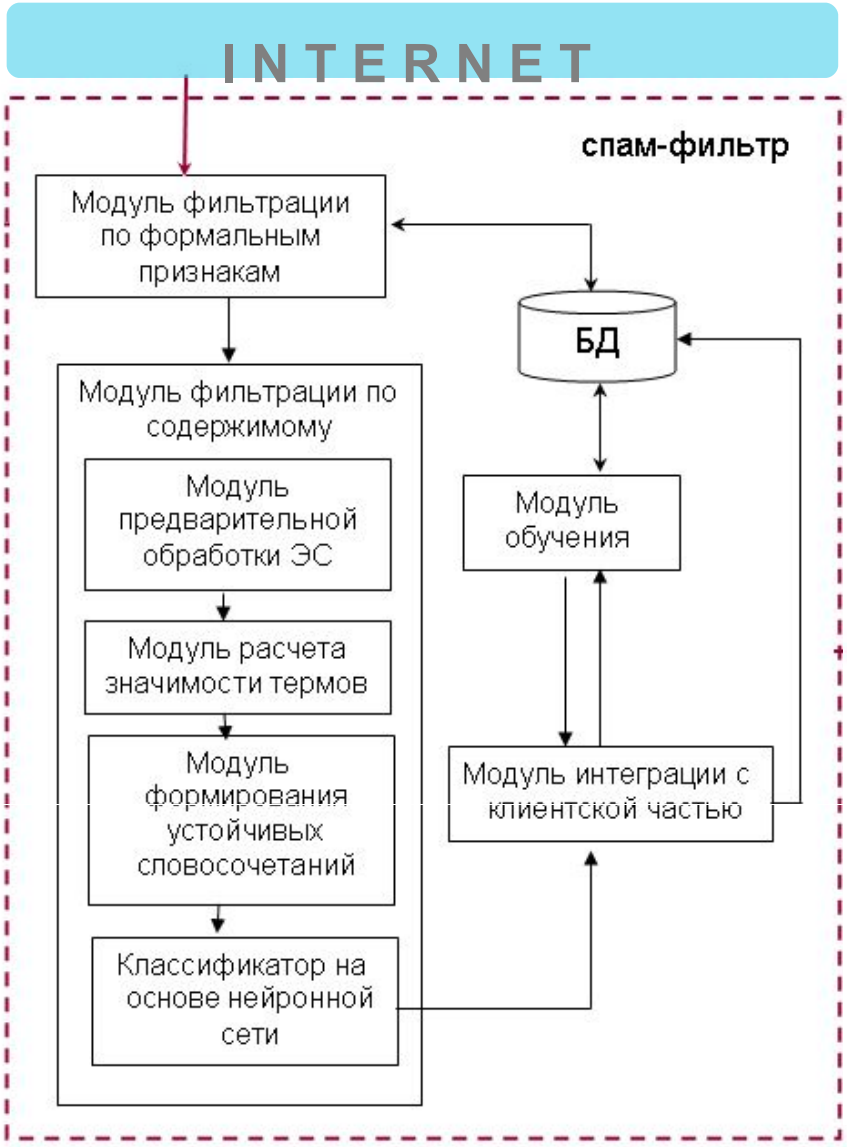


Рисунок X – Архитектура системы контентной фильтрации

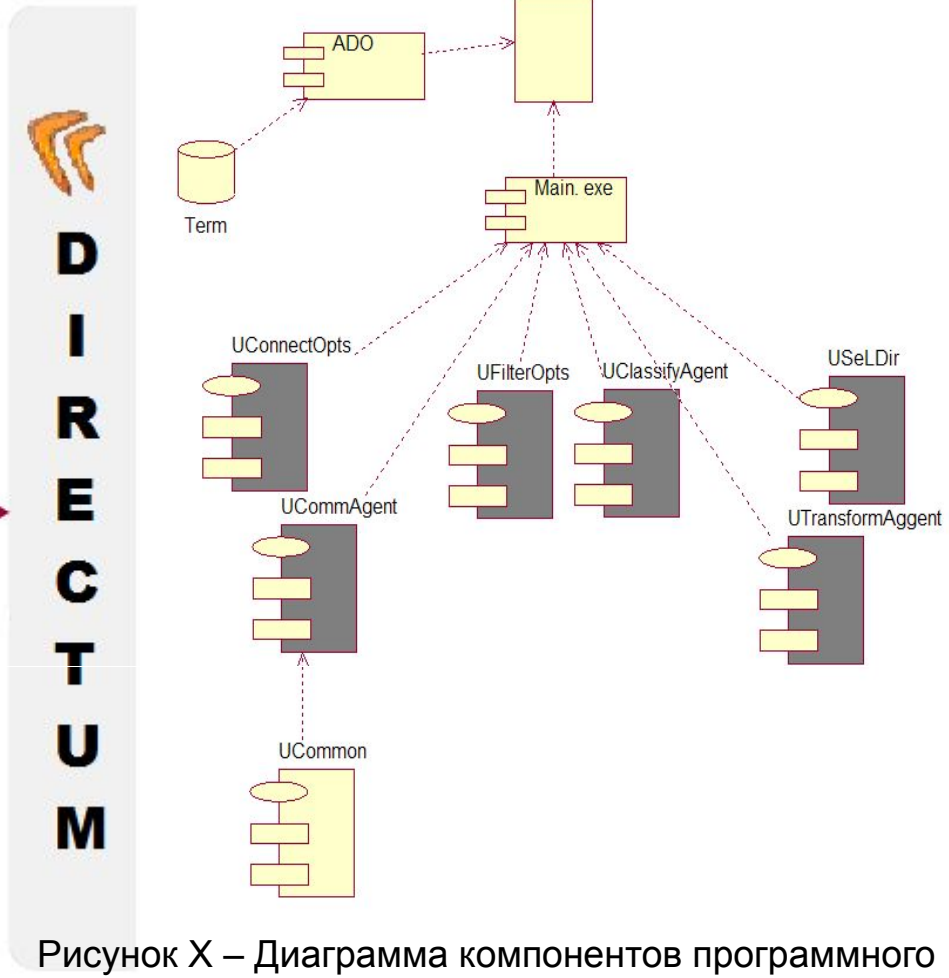


Рисунок X – Диаграмма компонентов программного проекта системы контентной фильтрации

ПРОЕКТ БАЗЫ ДАННЫХ И ИНТЕРФЕЙС СИСТЕМЫ КОНТЕНТНОЙ ФИЛЬТРАЦИИ

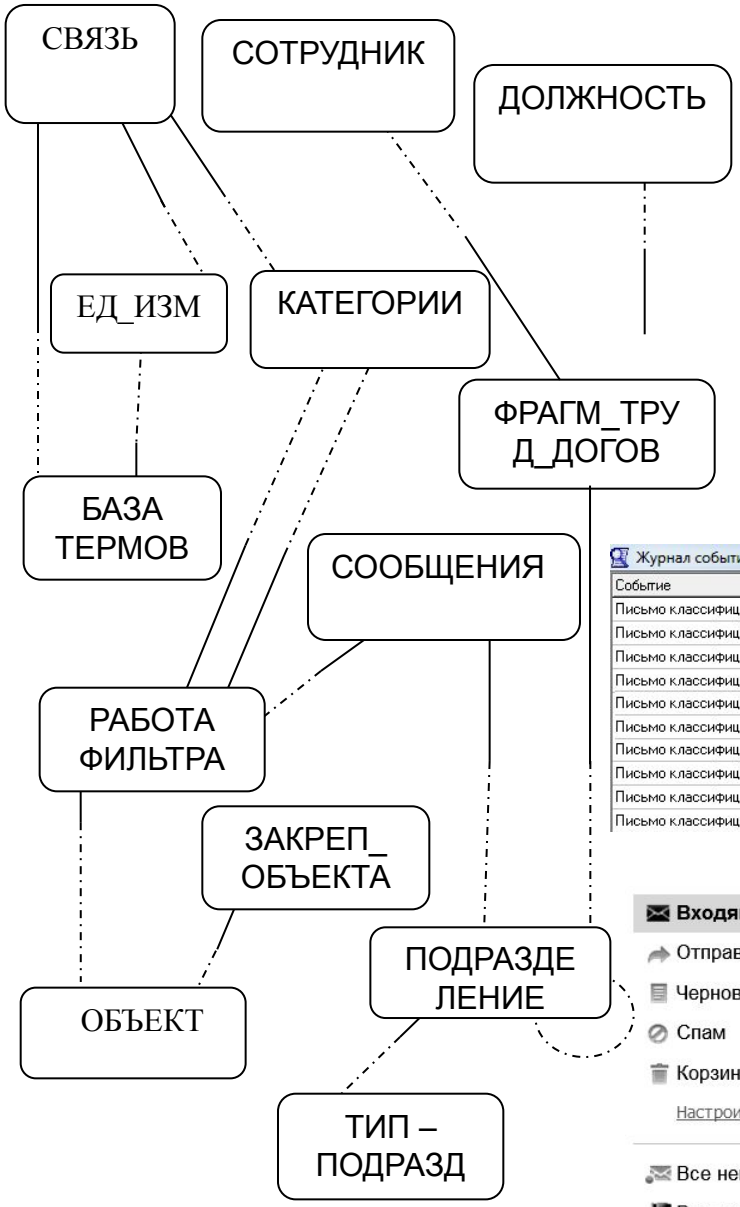


Рис. X – Инфологическая модель предметной области

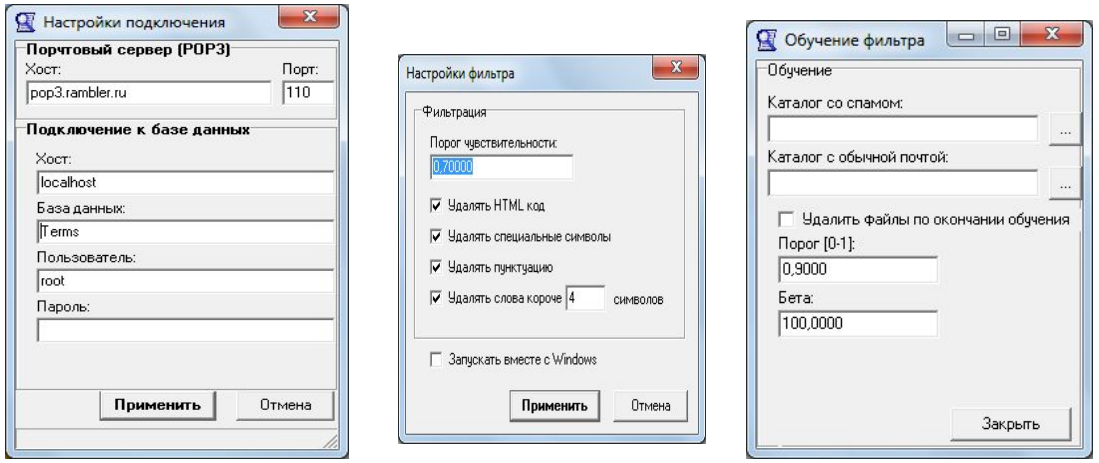
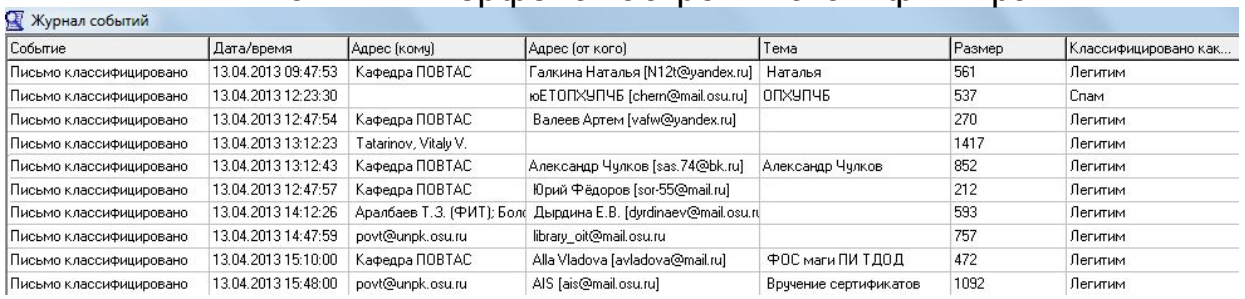


Рис. X – Интерфейс настройки спам-фильтра


Журнал событий

Событие	Дата/время	Адрес (кому)	Адрес (от кого)	Тема	Размер	Классифицировано как...
Письмо классифицировано	13.04.2013 09:47:53	Кафедра ПОВТАС	Галкина Наталья [N12t@yandex.ru]	Наталья	561	Легитим
Письмо классифицировано	13.04.2013 12:23:30		юЕТОПХУПЧБ [chem@mail.osu.ru]	ОПХУПЧБ	537	Спам
Письмо классифицировано	13.04.2013 12:47:54	Кафедра ПОВТАС	Валеев Артем [vafw@yandex.ru]		270	Легитим
Письмо классифицировано	13.04.2013 13:12:23	Tatarinov, Vitaly V.			1417	Легитим
Письмо классифицировано	13.04.2013 13:12:43	Кафедра ПОВТАС	Александр Чулков [sas.74@bk.ru]	Александр Чулков	852	Легитим
Письмо классифицировано	13.04.2013 12:47:57	Кафедра ПОВТАС	Юрий Фёдоров [sor-55@mail.ru]		212	Легитим
Письмо классифицировано	13.04.2013 14:12:26	Аралбаев Т.З. (ФИТ); Бол	Дырдина Е.В. [dyrdinaev@mail.osu.r		593	Легитим
Письмо классифицировано	13.04.2013 14:47:59	povt@unpk.osu.ru	library_oit@mail.osu.ru		757	Легитим
Письмо классифицировано	13.04.2013 15:10:00	Кафедра ПОВТАС	Alla Vladova [avladova@mail.ru]	ФОС маги ПИ ТДОД	472	Легитим
Письмо классифицировано	13.04.2013 15:48:00	povt@unpk.osu.ru	AIS [ais@mail.osu.ru]	Вручение сертификатов	1092	Легитим

Рис. X – Журнал событий

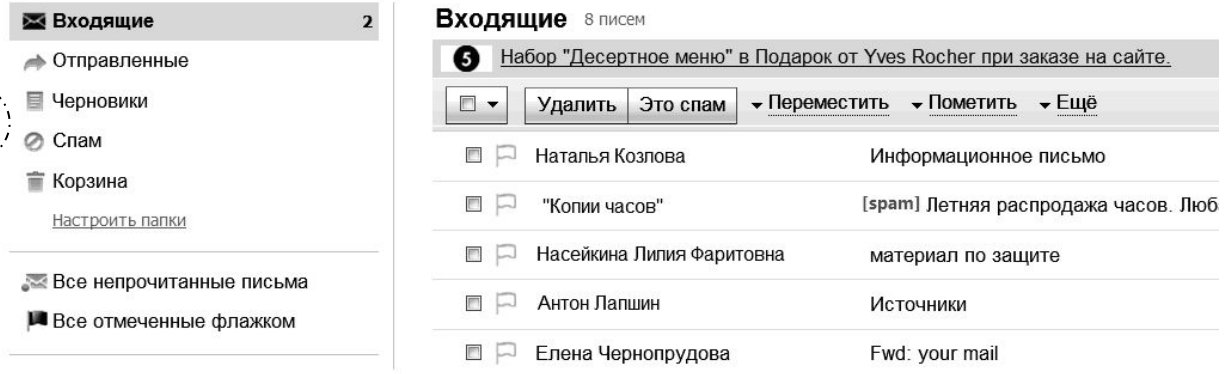


Рис. X – Интерфейс классификации спам-фильтра

МЕТОДИКА ОЦЕНКИ ЭФФЕКТИВНОСТИ

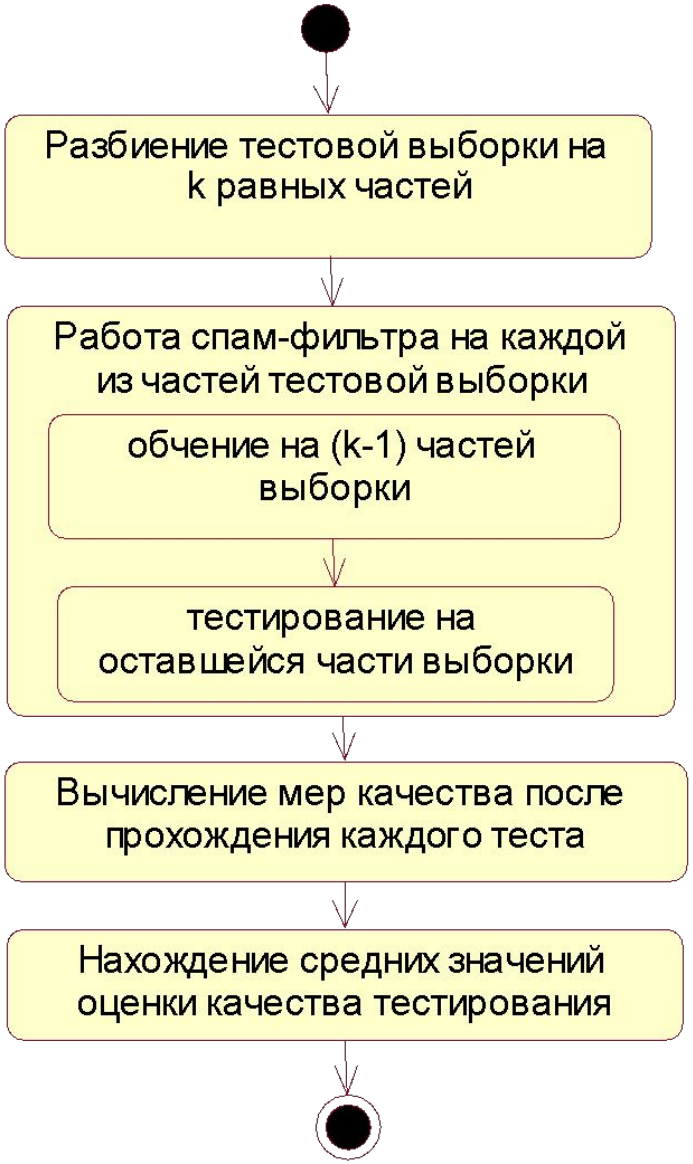


Рис. X – Методика проведения эксперимента методом k–подмножеств (k-foldes)

Ошибка 1 рода
(принятие решения о легитимности сообщения, когда оно является спамом)

$$\alpha = FNsp / Nsp , \quad (8)$$

Ошибка 2 рода
(принятие решения о спамности сообщения когда оно является легитимным)

$$\beta = FP_l / N_l , \quad (9)$$

Мера полноты (*precision*) (12)
(оценивает долю верного распознавания относительно всех объектов определенного класса)

Мера точности (13)
(оценивает долю верных обнаружений относительно всех объектов)

F мера (15)
(сводная оценка качества классификации)

N_{sp} – число объектов, относящихся к классу спам;

N_l – число объектов, относящихся к классу легитимных сообщений;

FN_{sp} – число спам-рассылок, классифицированных как легитимное письмо;

FP_l – число легитимных писем, классифицированных как спам-рассылка.

TP_l – число правильно классифицированных легитимных ЭС ($TP_l = N_l - FP_l$)

МЕТОДИКА ПРОВЕДЕНИЯ ИМИТАЦИОННОГО ЭКСПЕРИМЕНТА

Рисунок 4.5 – Схема имитационного эксперимента

Таблица 4.1 – Перечень тестовых сообщений

№	Вид сообщения	Тематика сообщений
1	спам	Пустые сообщения, содержащие только ссылки или вложения.
2	спам	Реклама товаров ()
3	спам	Реклама услуг (юридических, бухгалтерских, строительных, образовательных, туристических, медицинских и проч.)
4	спам	Приглашения на курсы, предложения схем «отмывания» денег, и т.п.
5	легитим	Деловая переписка (со знакомыми пользователями) свободная форма
6	легитим	Деловая переписка (со знакомыми пользователями) приказы, распоряжения и т.п.
7	легитим	Приглашения на участие в грантах, конференциях, выставках и т.п.

Название	Модель ЭС	Вес	Метод сокращения признакового пространства	Выделение устойчивых словосочетаний	Алгоритм классификации
Met1	векторная	Tf-idf	RF	+	нейрон. сеть Art
Met2	векторная	Ltc	RF	+	нейрон. сеть Art
Met3	векторная	Ltc	RF	-	нейрон. сеть Art
Met4	векторная	Ltc	IG	+	нейрон. сеть Art
Met5	векторная	Tf-idf	IG	+	нейрон. сеть Art

Методика оценки результатов имитационного эксперимента

Результатом ИЭ являются определение средних значений двх вероятностных характеристик - вероятности принять решение о легитимности сообщения, когда оно спам (***α – ошибка 1 рода***) и вероятность отвергнуть решение о легитимности сообщения, когда оно легитимно (***β - ошибка 2 рода***), сводной оценки качества классификации (F-мера), полноты и точности.

ОЦЕНКА ЭФФЕКТИВНОСТИ ПРОТОТИПА СИСТЕМЫ СПАМ-ФИЛЬТРАЦИИ

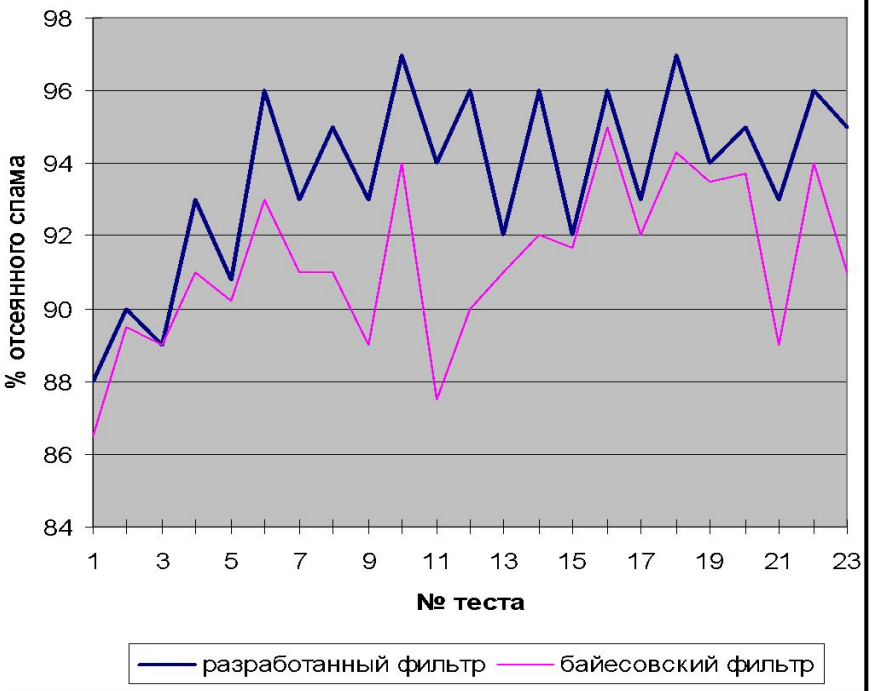


Рис. X – Сравнительная оценка эффективности предложенного спам-фильтра и фильтра на основе байесовского классификатора

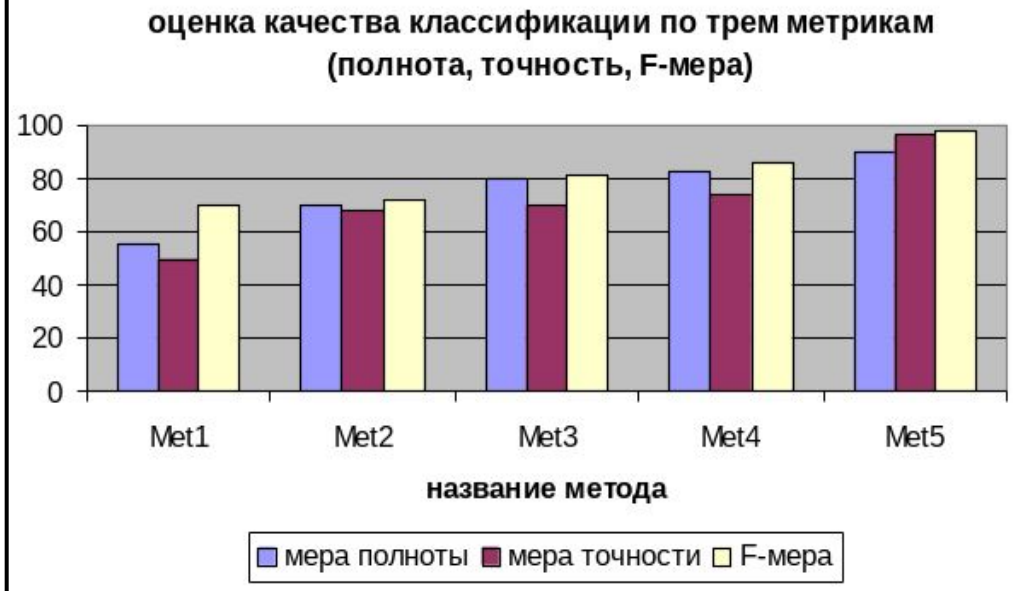


Рис. X – Результаты имитационного эксперимента

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИОННОГО ИССЛЕДОВАНИЯ

1 Научная новизна модели ЭС заключается в применении меры значимости для определения веса признаков в ЭС (термов) позволяющей сократить характерный разброс в частотах различных термов

Во первых, предложен комбинированный метод сокращения признакового пространства, основанный на том, что для каждого терма в сообщениях определенного класса вычисляется величина , характеризующая значимость терма для определенного класса (spam\legitim)

Во вторых, предложенная методика выделения устойчивых словосочетаний позволяет без потери смыслового содержания выделить термы характеризующие данное сообщение(класс), тем самым выделить признаки легитимности сообщения в отличии от существующих фильтров учитывающих только признаки спама.

2 Новизна методики и алгоритмов фильтрации НЭС заключается в развитии нейросетевых методов классификации и новом практическом применении нейронной сети ART для осуществления идентификации несанкционированных рассылок электронной почты.

АПРОБАЦИЯ, ПУБЛИКАЦИИ

Научные и практические результаты диссертационных исследований

обсуждались и получили одобрение на 5-ти всероссийских научно-практических конференциях с международным участием (ОГУ 2003- 2008 гг.; СПГТУ 2008 г.) и 3-х региональных научных семинарах «Актуальные вопросы информационных технологий теории управления» (ВУ ВПВО 2006 -2008 гг.);

опубликованы в 10-ти печатных работах, одна из которых – в издании, определенном ВАК России для опубликования научных результатов диссертаций на соискание ученых степеней, в 2-х свидетельствах о государственной регистрации программ, а также в четырех отчетах о НИР на спецтемы.

НАПРАВЛЕНИЯ ДАЛЬНЕЙШИХ ИССЛЕДОВАНИЙ

Анализ среды Internet как предпосылки НСР	Исследования механизмов спам-рассылок	Разработка методов и средств спам-фильтрации
Анализ системного и прикладного ПО. Исследование сетевого оборудования и анализ протоколов. Анализ современных типовых технологии получения информации о спам-рассылках.	Систематизация и моделирование механизмов спам-рассылок и других аномальных событий. Модели спам-рассылок, интегрированных с информационными атаками. Создание механизов адаптивной защиты	Обнаружение и предотвращение спам-рассылок. Обнаружение и защита от сетевых вирусов Активное противоборство спам-воздействиям Анализ рисков возникновения спам-атак, их последствий и определение фактической степени необходимой защиты.