На правах рукописи

Цыганов Илья Германович

УДК: 681.5

МНОГОАГЕНТНАЯ АВТОМАТИЗИРОВАННАЯ СИСТЕМА АДАПТИВНОЙ ФИЛЬТРАЦИИ ПОТОКОВ ТЕКСТОВОЙ ИНФОРМАЦИИ

Специальность 05.13.01 –
Системный анализ, управление и обработка информации

А в т о р е ф е р а т

диссертации на соискание ученой степени кандидата технических наук

Москва - 2005

Работа выполнена в Московском Государственном Техническом Университете им. Н.Э. Баумана

Научный руководитель: заслуженный деятель науки РФ,
доктор технических наук, профессор В.А. Шахнов

Официальные оппоненты: доктор технических наук, профессор В.М. Черненький
кандидат технических наук, Н.Б. Лиханов

Ведущее предприятие: ОАО "Концерн "РТИ-Системы"

Защита диссертации состоится «13» октября 2005 г. на заседании диссертационного совета Д 212.141.02 в Московском Государственном Техническом Университете им. Н.Э. Баумана.

Ваши отзывы в двух экземплярах, заверенные гербовой печатью, просьба высылать по адресу: 107005, 2-ая Бауманская ул., д. 5.

Автореферат разослан «25» сентября 2005 г.

Ученый секретарь
диссертационного совета
к.т.н., доцент Иванов В.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Важнейшей задачей, стоящей в современных гетерогенных системах обработки информации, таких как система электронной почты (ЭП) глобальной сети Интернет, является сокращение объемов незапрашиваемой информации (т.н. "спама" – от англ. *spam*), обработка которой приводит к значительным потерям. В системах ЭП для решения этой задачи применяются системы специализированного класса – автоматизированные системы фильтрации незапрашиваемой рассылки (АСФ НР), обеспечивающие отделение потоков важной информации от потоков незапрашиваемых сообщений.

АСФ НР выявляют незапрашиваемую рассылку (НР) на основании автоматического анализа совокупности признаков, содержащихся в сообщении. Однако, распространители НР, с целью обхода фильтров, постоянно изменяют внешние признаки сообщений.

Поэтому в современных условиях АСФ НР должна, во-первых, обладать способностью быстро реагировать на факт изменения поведения распространителей НР, во-вторых, обеспечивать быструю и качественную самонастройку в соответствии с новым набором признаков НР.

В настоящее время разработано большое число АСФ НР различного типа, однако, ни одна из них не обеспечивает в полной мере учета всей совокупности перечисленных требований, что приводит к низким показателям качества их работы. Современные системы при уровне ложного блокирования 2-3% пользовательских сообщений обеспечивают пропуск до 30% сообщений НР. Снижение показателей качества в различных системах обусловлено различными факторами.

Так в современных многопользовательских системах (Brightmail, Kaspersky Antispam, SpamAssassin и др.) снижение качества фильтрации обусловлено малой скоростью реакции системы на ошибки фильтрации. Это связано с тем, что пользователи этих систем, выявляющие ошибки фильтрации, фактически исключены из контура настройки системы, поскольку не считаются надежными источниками информации (пользователь может ошибаться, иметь субъективное мнение и пр.). А в существующих индивидуальных АСФ НР (SpamBully, ifile и др.), качество фильтрации достаточно мало, поскольку фильтры обучаются на ограниченном числе сообщений, поступающих только конкретному пользователю.

Все это показывает, что качество фильтрации существующих АСФ НР может быть увеличено за счет применения комплексных многопользовательских систем фильтрации, обеспечивающих полномасштабное участие пользователей в процессе выявления ошибок фильтрации и соответствующей настройке фильтров. Это делает задачу разработки научных основ (математического аппарата и функциональной структуры) построения многоагентных автоматизированных систем адаптивной фильтрации (АСАФ) актуальной.

Цель работы. Исследование моделей, методов и алгоритмов адаптивной фильтрации потоков информации в группах пользователей и разработка многопользовательской многоагентной обучаемой пользователями автоматизированной системы адаптивной фильтрации, обеспечивающей автоматическое выявление и блокирование незапрашиваемой рассылки в потоках сообщений систем электронной почты глобальной информационной сети Интернет.

Решаемые задачи:

1) исследование и анализ функциональных возможностей и математического аппарата современных АСФ НР и разработка требований к архитектуре и математическому аппарату многопользовательских адаптивных АСФ НР;
2) исследование и разработка многоагентной архитектуры АСАФ НР, позволяющей пользователям участвовать в процессах выявления ошибок, фильтрации и настройки фильтров АСФ НР;
3) исследование и разработка математических моделей, методов и алгоритмов, обеспечивающих адаптивную фильтрацию потоков информации в группах пользователей: методов формирования обучающей и тестовой выборки, методов построения пространства признаков, методов классификации текстовых сообщений по выявленным признакам, методов коллективной фильтрации.
4) исследование и выбор методов программной и аппаратной реализации средств фильтрации НР и разработка реализации АСАФ НР в виде аппаратно-программного комплекса, обеспечивающего фильтрацию НР на основе разработанных принципов;
5) экспериментальное исследование эффективности предложенных моделей, методов и алгоритмов, определение оптимальных параметров, обеспечивающих наивысшие показатели качества фильтрации НР, разработка рекомендаций по настройке системы.

Методы исследования. При решении поставленных задач использована теория информационных систем, теория экспертных систем и обработки знаний, нейроматематика, теория нейронных сетей, теория оптимизации, математический аппарат теории автоматического управления, теория вероятностей и математическая статистика, теория Марковских случайных полей.

Научная новизна работы состоит в следующем:

1) Проведено исследование, классификация и систематизация существующих многопользовательских АСФ НР с точки зрения функциональной структуры, особенностей реализации основных функций и применяющегося в них математического аппарата.
2) Исследована и разработана архитектура многоагентной автоматизированной системы адаптивной фильтрации НР, обеспечивающая эффективное взаимодействие пользователей системы при выявлении ошибок фильтрации и настройке фильтров.
3) Исследованы и разработаны математические модели и методы формирования пространства признаков в задаче анализа содержания сообщений электронной почты, что позволило повысить точность анализа за счет учета значимых словосочетаний.
4) Исследован и разработан метод синтеза нейронной сети с переменной структурой, входным сигналом которой являются разряженные векторы большой размерности (до десяти тысяч).

Достоверность полученных научных результатов, выводов и рекомендаций диссертационной работы подтверждена:

1) результатами экспериментальных исследований;
2) результатами внедрения разработанной многопользовательской АСАФ НР в корпоративной системе электронной почты Международного Института Экономики и Права (4000 пользователей);
3) результатами внедрения разработанных в работе моделей, методов и алгоритмов, а также программного комплекса экспериментального

исследования алгоритмов фильтрации текстовой информации в учебный процесс МГТУ им. Н.Э. Баумана.

Полученные в работе результаты наглядно демонстрируют эффективность использования разработанных моделей, методов и алгоритмов для решения задач автоматической фильтрации незапрашиваемой рассылки.

Положения, выносимые на защиту:

1) архитектура, функциональный состав и интерфейсы АСАФ НР;
2) метод формирования пространства признаков в задаче фильтрации текстовых сообщений, обеспечивающий учет, как значимых слов, так и значимых словосочетаний;
3) методы и алгоритмы обучения многослойной нейронной сети с переменной структурой, входным сигналом которой являются разряженные векторы большой размерности;
4) аппаратно-программная реализация многоагентной АСАФ НР;
5) результаты экспериментальных исследований разработанных методов и алгоритмов решения задачи фильтрации незапрашиваемой рассылки.

Практическая ценность работы. Разработанные в диссертации методы формирования пространства признаков, методы учета указаний группы пользователей о выявленных ошибках фильтрации и алгоритмы настройки нейронных сетей, а также аппаратно-программный комплекс, реализующий многоагентную АСАФ НР, построенный на основе разработанных принципов позволяют:

1) повысить эффективность фильтрации потоков незапрашиваемой информации;
2) автоматизировать и упростить контур настройки современных АСФ НР;
3) сократить время, требуемое на адаптацию АСФ НР к выявленным ошибкам;
4) автоматизировать и сократить ручной труд при наладке системы;
5) предоставить пользователям гибкий механизм управления процессом фильтрации сообщений НР.

Разработанные алгоритмы и программы могут быть использованы для дальнейшего развития и совершенствования систем интеллектуальной фильтрации и управления потоками текстовой информации.

Реализация результатов. Разработанная в работе аппаратно-программная реализация АСАФ НР внедрена в корпоративную систему электронной почты Международного Института Экономики и Права и обеспечивает фильтрацию потоков сообщений, поступающих нескольким тысячам пользователей этой системы.

Полученные в работе математические модели, методы и алгоритмы, а также разработанный комплекс экспериментальных исследований алгоритмов адаптивной фильтрации потоков текстовой информации, внедрен в учебный процесс МГТУ им. Н.Э. Баумана.

Апробация работы. Результаты работы были представлены на Международной молодежной научно-технической конференции "Наукоемкие технологии и интеллектуальные системы", (Москва, 2003, 2004), Международной молодежной научной конференции "Информатика и системы управления в XXI веке", (Москва, 2003 г.), студенческой научной конференции "Студенческая научная весна – 2002", (Москва, 2002).

Публикации. По материалам и основному содержанию работы имеется 15 публикаций в научно-технических журналах и трудах конференций.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения, списка литературы. Общий объем диссертации 208 страниц, 84 рисунка, список использованных источников из 237 наименований.

СОДЕРЖАНИЕ И РЕЗУЛЬТАТЫ ДИССЕРТАЦИИ

Во введении обоснована актуальность решения задач борьбы с НР, сформулирована цель и задачи исследования, обоснована научная новизна и изложена структура диссертации, показано место АСФ НР рассматриваемого класса в современных глобальных системах обработки информации.

В первой главе рассматриваются вопросы, связанные с исследованием современного состояния проблемы незапрашиваемой рассылки, классификацией и систематизацией технических средств борьбы с НР, и разработке на их основе требований к АСФ НР.

Указываются предпосылки НР, виды НР, типы распространителей НР, способы распространения НР.

Разработана обобщенная архитектура АСФ НР, в составе которой выделены агенты двух типов: агенты пользователей (АП) и агенты фильтрации (АФ) (рис. 1). Выделены основные функции АСФ НР: фильтрация, выявление ошибок и настройка.

Это позволило осуществить классификацию и систематизацию АСФ НР по архитектуре, способам реализации основных функций, и применяющемуся математическому аппарату.

Сформулированы ключевые недостатки существующих АСФ НР, основными из которых являются следующие:

1) недостаточный уровень автоматизации контура настройки АСФ;

2) слабое участие пользователей в настройке АСФ НР.

Результаты исследования показали необходимость разработки АСФ НР нового класса, в которых фильтрация, выявление ошибок и настройка фильтров осуществляется группой пользователей системы.

В работе разрабатываются требования к системам нового класса – автоматизированным системам адаптивной фильтрации незапрашиваемой рассылки (АСАФ НР).

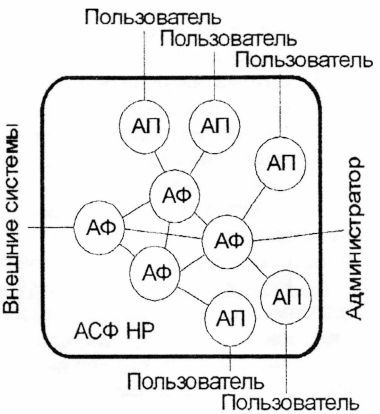

Рис. 1. Обобщенная архитектура АСФ НР

Во второй главе производится разработка функциональной модели, математических методов и алгоритмов АСАФ НР.

Определяются принципы построения системы из групп взаимодействующих агентов, описываются потоки информации в группах агентов и между отдельными агентами, функциональный состав отдельных агентов.

Обмен информацией о результатах обработки сообщений в каждом из агентов осуществляется с помощью специализированной структуры – статус сообщения:

$$S = < S^T, S^V >,$$

где S^T - двухсимвольный мнемонический код ($S^T \in \{HE, HT, HC, SC, ST, SE\}$), S^V - целое число в интервале $S^V \in [-100,100]$. Числовое значение определяет степень уверенности агента: большие положительные значения характерны для большей уверенности в том, что сообщение является ЛПС (легитимным пользовательским сообщением), меньшие отрицательные значения характерны для большей уверенности в том, что сообщения является СНР (сообщением незапрашиваемой рассылки). Мнемонический код разбивает числовой диапазон [-100,100] на ряд поддиапазонов и интерпретируется следующим образом: НЕ – точно ЛПС; НТ – ЛПС с уверенностью, большей заданной; НС – ЛПС с уверенностью, меньшей заданной, аналогично для СНР. Величины уверенности определяются с помощью порогов в каждом из агентов самостоятельно.

В главе подробному исследованию и разработке подвергаются математические модели, методы и алгоритмы, использующиеся в агентах АСАФ НР для решения задачи адаптивной фильтрации потоков текстовых данных.

Фильтрация обеспечивается с помощью последовательного применения трех фильтров: формального, контекстного и коллективного (рис. 2.).

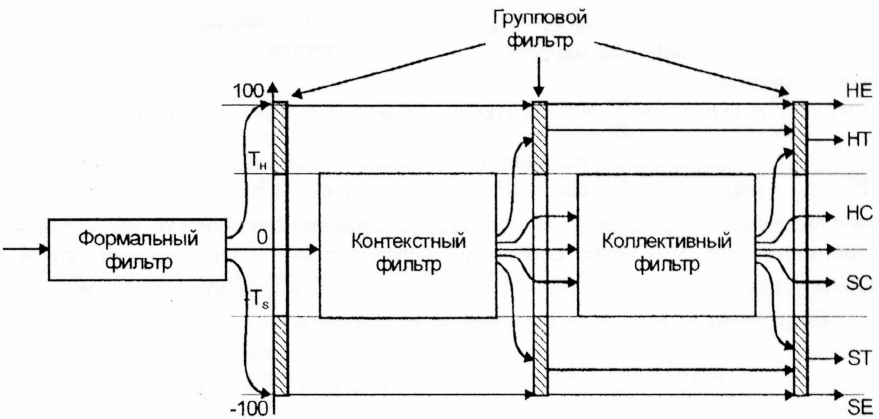

Рис. 2. Фильтрация в агентах АСФ НР

Формальный фильтр представляет собой алгоритм фильтрации на основе черных и белых списков, а также методов сигнатурного анализа сообщений.

Контекстный фильтр обеспечивает фильтрацию по признакам, извлекаемым из текстовой части сообщений (отдельные слова и словосочетания).

Работа коллективного фильтра основывается на обработке результатов фильтрации сообщений в других агентах АСАФ НР на основе применения методов корреляции.

Обучение контекстного и коллективного фильтра требует использования обучающей (настройка фильтра) и тестовой (оценка качества настройки) выборок. Формирование выборок осуществляется автоматически по алгоритму, представленному на рис. 3.

Анализ текста сообщений в контекстном фильтре осуществляется с использованием признаков двух типов: отдельных слов и словосочетаний.

Вначале осуществляется выбор подмножества слов, которые будут использоваться для представления текстов сообщений. На основании выбранных слов осуществляется выбор комплексных признаков, состоящих из группы слов.

Выбор значимых слов осуществляется из слов сообщений обучающей выборки, отбираемых согласно одного из трех исследуемых в работе критериев:

1) по частоте;

2) максимальная информативность (Information Gain - IG);

3) критерий χ^2.

Критерий по частоте выбирает заданное число слов, имеющих максимальную частоту в массиве сообщений.

Критерий IG оценивает значимость каждого слова в соответствии с информативностью данного признака:

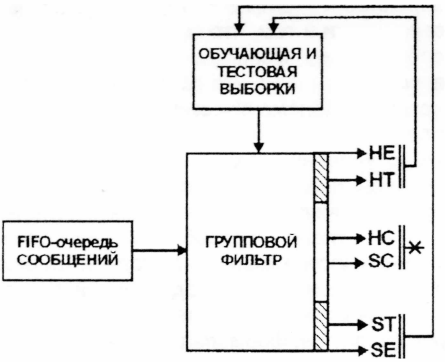

Рис. 3. Формирование обучающей и тестовой выборок

$$G(w) = -\sum_{e \in E} p_e \lg p_e + p(w)\sum_{e \in E} p(\varepsilon = e \mid w)\lg p(\varepsilon = e \mid w) + p(\overline{w})\sum_{\varepsilon_i \in E} p(\varepsilon = e \mid \overline{w})\lg p(\varepsilon = e \mid \overline{w}),$$

где $E=\{-1,1\}$ – пространство решений (-1 – соответствует классу СНР, 1 – классу ЛПС); p_{-1} - вероятность класса СНР, p_1 - вероятность ЛПС, $p(w)$ - вероятность сообщения со словом w, $p(\overline{w})$ - вероятность сообщения без слова w, $p(\varepsilon = e \mid w)$ - вероятность того, что класс сообщения соответствует e при условии, что в сообщении имеется слово w, $p(\varepsilon = e \mid \overline{w})$ - аналогично при отсутствии слова w.

В критерии χ^2 используется следующая формула:

$$\chi^2(w) = \max_{\varepsilon \in E} \chi^2(w,\varepsilon),\ \chi^2(w,\varepsilon) = \frac{|M|(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)},$$

где A - количество сообщений, в которых w и ε появились совместно; B - количество сообщений, при которых слово w встречается с другим классом; C - количество сообщений, относящихся к классу ε, но в которых не встречается w, D - количество сообщений, не относящихся к классу ε, в которых нет слова w; $|M|$ - общее число сообщений в обучающей выборке.

Из числа отобранных слов осуществляется построение словаря.

Формирование комплексных признаков осуществляется с помощью моделирования лингвистической структуры классов. Класс моделируется с помощью функции:

$$P(X = x) = p(m \mid \varepsilon = const) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n),$$

где $x(m) = (x_1(m), x_2(m), ..., x_n(m))$ - вектор сообщения ($x_i = 0$, если i-ое слово словаря присутствует в сообщении m и $x_i = 1$ если наоборот), n – число слов, отобранных на этапе выбора значимых слов.

Моделирование осуществляется в виде закона распределения Гиббса:

$$p(X=x)=\frac{1}{Z(\lambda_1,\lambda_2,...,\lambda_F)}e^{\sum_i^F \lambda_i f_i(x)}=\frac{1}{Z(\lambda)}e^{\lambda\cdot f(x)},$$

где $f(x)=\{f_i(x),\ i=1,...,F\}$ – множество бинарных функций признаков, λ - соответствующие им коэффициенты (параметры), $Z(\lambda)=\sum_x e^{\sum_i \lambda_i f_i(x)}$ - нормировочная постоянная.

Функции-признаки являются либо простыми, либо составными. Простые функции-признаки представляют отдельные слова, а их общее число равно числу слов в словаре. Простые функции-признаки обращаются в 1 только в том случае, если вектор-аргумент имеет в компоненте, соответствующей представляемому слову, единицу, а во всех остальных случаях обращается в 0. Составные признаки представляют несколько слов и обращаются в единицу на соответствующем наборе компонент, тогда как на всех прочих имеют значение 0.

Задача формирования пространства признаков решается с помощью выбора такого множества признаков $\{f_i\}$, чтобы результирующее распределение $p(X=x)$ было максимально близко к эмпирической функции распределения $p'(X=x)$ при заданном конечном числе признаков. Близость двух распределений в работе оценивается с помощью классических функционалов теории информации:

1. $I(p':p)=\sum_x p'(x)\log\frac{p'(x)}{p(x)}\to\min$,

и двойственного ему функционала:

2. $I(p:u)=\sum_x p(X=x)\log\frac{p(X=x)}{u(x)}=\sum_x p(X=x)\log p(X=x)\to\min$

при условии $\sum_x f_i(x)p(X=x\,|\,f)=\sum_x f_i(x)p'(X=x)=\theta_i$, $i=1,...,F$, где $\theta_i\in R, i=1,...,F$ - постоянные, $u(x)$ – равномерный закон распределения.

Решение задачи производится с помощью последовательного, многоитерационного алгоритма, на каждой итерации которого производится поиск очередного оптимального признака из числа кандидатов (структурный синтез) и добавление его в модель за счет оптимального выбора параметра λ_i (параметрический синтез). Алгоритм продолжается до тех пор, пока не будет добавлено заданное количество признаков.

Очередной признак выбирается из числа кандидатов. Изначально множество кандидатов содержит только простые признаки. После выбора и добавления в модель признак исключается из числа кандидатов, а множество кандидатов расширяется за счет комбинирования выбранного признака с признаками в текущем множестве кандидатов.

С целью выбора признака из числа кандидатов производится оценка его значимости в соответствии с текущей моделью. Для этого используется следующий функционал:

$$I(f^*)=\inf_\beta I(p':p(f^*,\beta)),$$

где f^* - оцениваемый кандидат, β - соответствующий ему коэффициент, $p(f^*,\beta)=\frac{1}{Z'(f^*,\beta)}e^{\beta f^*+\sum_i \lambda_i f_i}$, $Z'(f^*,\beta)=\sum_x e^{\sum_i \lambda_i f_i+\beta f^*}$. Выбирается признак, имеющий

наименьшее значение указанного функционала, что в аналитической форме может быть представлено следующим образом:

$$f^*_{опт} = \arg\min_{f^* \in F_c}\left[\log\frac{1-P^1(f^*)}{1-E^1(f^*)} + E^1(f^*)\log\left(\frac{P^1(f^*)}{1-P^1(f^*)}\frac{E^1(f^*)}{(1-E^1(f^*))}\right)\right],$$

где F_c – множество кандидатов, $P^1(f^*)$ - вероятность того, что признак f^* равен *1* относительно текущей модели *p(x)*, $E^1(f^*)$ - среднее значение признака f^* на эмпирической выборке.

Представленные алгоритмы позволяют сформировать пространство признаков, включающее в качестве признаков как отдельные слова, так и словосочетания. Пусть общая размерность пространства признаков N, Таким образом каждое сообщение можно представить в виде вектора: $x = (x_1, x_2, ..., x_N)$, компоненты которого соответствуют либо отдельным словам, либо словосочетаниям.

На основе сформированного пространства признаков осуществляется синтез многослойной нейронной сети. При этом осуществляется построение оптимальной гиперповерхности, разделяющей пространство векторов обучающей выборки *X* на ряд непересекающихся областей, каждая из которых относится к одному из двух классов (т.е. $X \to Y$, где Y – пространство решений нейросети).

Построение кусочно-линейной гиперповерхности, разделяющей два класса векторов, осуществляется с помощью последовательного алгоритма, который заключается в постепенном увеличении числа гиперповерхностей, до тех пор, пока не будет достигнуто условие останова.

Опишем применяющийся алгоритм. Вначале пространство признаков делится с помощью одного нейрона. Синтез нейрона осуществляется с помощью минимизации одного из рассмотренных ниже функционалов вторичной оптимизации. Затем на основании того же критерия полученные при делении области делятся еще раз и т.д. На рис. 4. представлена общая структурная схема алгоритма: *I* – блок определения параметров нейрона; *II* – блок разделения входной обучающей последовательности; *III* (пунктир) – алгоритм синтеза многослойной нейронной сети на первом шаге, аналогично которому строятся блоки *III*; *VI* - блок управления.

Выбор параметров отдельных нейронов первого слоя на каждом шаге алгоритма осуществляется в соответствии с минимизацией одного из четырех рассматриваемых в работе критериев вторичной оптимизации:

1. $a(n+1) = a(n) + K^* sign\left[\overline{x_a(n)}^{m_*}\right]\overline{x(n)}^{m_*}$;
2. $a(n+1) = a(n) + 2K^*\overline{x_a(n)x(n)}^{m_*}$;
3. $a(n+1) = a(n) + K^* sign[x_0(n)]sign(x_i(n))$;
4. $a(n+1) = a(n) + 2K^* x_a(n) sign(x_i(n))$,

где *a(n)* – вектор коэффициентов нейрона на текущей итерации, $x_a(n) = \varepsilon(n) - \sum_{i=0}^{N} a_i x_i(n)$ - величина аналоговой ошибки, $x_0(n) - \varepsilon(n) - sign\left[\sum_{i=0}^{N} a_i x_i(n)\right]$ - величина дискретной ошибки (предполагается $x_0(n)=1$ для всех n), $\varepsilon(n)$ - эталонные данные о классе сообщения, K* - матрица, определяющая скорость

настройки алгоритма, а применение усреднения $\overline{x(n)}^{m_n} = \frac{1}{m_n} \sum_{i=n-m_n}^{n} x(i)$ позволяет снизить уровень шумов в контуре настройки, m_n - константа.

Рассматриваются методы синтеза последующих слоев. Вначале осуществляется проверка возможности реализации последующих слоев в виде одного нейрона. Если это реализуемо, то на этом процесс синтеза прекращается. Если нет, то производят формирование последующих слоев в виде порогово-дизъюнктивной сети.

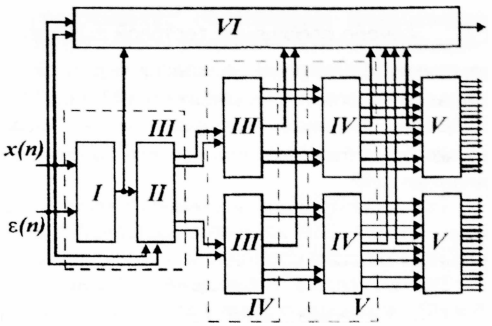

Рис. 4. Структурная схема алгоритма последовательного формирования кусочно-линейной разделяющей гиперповерхности

Рассмотрим расчет значения статуса сообщения в коллективном фильтре. Рассматривается Q дополнительных агентов, в которых значение статуса для сообщения m равно $S_j[m], j = 1,...,Q$. Это позволяет каждому сообщению поставить в соответствие следующий набор величин:

$$\vartheta_0^j(m) = \begin{cases} 1, S_j[m] \in \{HE, HT\} \\ -1, S_j[m] \in \{SE, ST\} \\ 0, по-другому \end{cases}.$$

Аналогичная величина рассчитывается для данного агента - $\vartheta[m]$. В результате для каждого сообщения обучающей выборки $m(n)$ рассматривается величина коэффициента корреляции:

$$r_{0j} = \frac{\sum_{n=1}^{|M|} \left(\vartheta[m(n)] - \overline{\vartheta}\right)\left(\vartheta_0^j[m(n)] - \overline{\vartheta_0^j}\right)}{\sqrt{\sum_{n=1}^{|M|} \left(\vartheta[m(n)] - \overline{\vartheta}\right)^2 \left(\vartheta_0^j[m(n)] - \overline{\vartheta_0^j}\right)^2}}, \text{ где } \overline{\vartheta} = \frac{\sum_{n=1}^{|M|} \vartheta[m(n)]}{|M|}, \ \overline{\vartheta_0^j} = \frac{\sum_{n=1}^{|M|} \vartheta_0^j[m(n)]}{|M|}.$$

Величина r_{0j} показывает насколько можно в данном агенте доверять результатам идентификации класса сообщения в j-ом агенте.

Величина статуса для любого сообщения m на основании результатов идентификации класса сообщения в других агентах рассчитывается следующим образом:

$$S_0^*[m] = \overline{\vartheta} + \frac{\sum_j \left(\vartheta_0^j[m] - \overline{\vartheta_0^j}\right) r_{0j}}{\sum_j |r_{0j}|}.$$

Оценка качества фильтрации в работе осуществляется с помощью тестовой выборки на основании оценок точности и ошибки классификации. Указанные параметры определены следующим образом:

точность: $\Lambda = \frac{\mu \cdot N_{1,1} + N_{-1,-1}}{\mu \cdot |Leg[M']| + |Spm[M']|}$,

ошибка: $E = \frac{N_{-1,1} + \mu \cdot N_{1,-1}}{\mu \cdot |Leg[M']| + |Spm[M']|}$,

где $N_{e_1 e_2}$ - число сообщений тестовой выборки, которые относятся к классу e_1, но отнесенные фильтрами к классу e_2, $e_1, e_2 \in E = \{-1,1\}$, *Spm[M]* – множество сообщений класса СНР в множестве М, *Leg[M]* – множество ЛПС, М' - множество сообщений тестовой выборки, μ - коэффициент, позволяющий учесть увеличенную стоимость ошибок при выявлении ЛПС относительно ошибок при выявлении СНР.

Также определены величины точности и ошибки при выявлении сообщений каждого класса. Эти величины используются для оценки работы алгоритмов по отдельным классам сообщений в экспериментальной части работы.

Третья глава посвящена созданию аппаратно-программного комплекса АСАФ НР, решающего задачу адаптивной фильтрации НР.

Среди основных наиболее крупных программных компонентов АСАФ НР выделены следующие (см. рис. 5):

1) компонент ядра агента АСАФ НР (ядро);
2) компонент взаимодействия (КВ);
3) компонент доступа к данным (КДД);
4) компонент ввода и отображения данных (КВОД).

КВ реализуется на базе стандартного программного обеспечения (ПО) системы Postfix, КДД – на основе Apache Web Server, КВОД – на базе Outlook Express.

В качестве серверной ОС выбрана ОС RedHat, в качестве клиентской ОС – MS Windows XP.

В качестве языка программирования и средств разработки приложений АСАФ НР выбраны соответственно язык C++ и KDevelop (для серверной части), и Microsoft Visual Studio .NET 2003 (для клиентской).

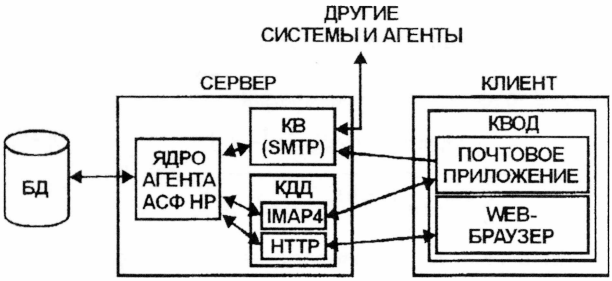

Рис. 5. Программные компоненты АСФ НР

В работе рассматриваются программные модули собственной разработки.

Центральным модулем агентов АСАФ НР является boa.bin, который отвечает за выполнение функций запуска, перезапуска, останова агентов АСАФ НР на данном сервере, а также функций их мониторинга. Управление этим модулем осуществляется в режиме командной строки, опции которой следующие:

```
boa.bin  [agent_name1, agent_name2, ...] [--start]
         [--restart] [--list] [--activelist] [--passivelist]
         [--eventlog K [N-]M] [--config file.conf]
```

В работе также описываются все библиотеки, используемые в процессе работы модуля, основные интерфейсы взаимодействия, конфигурационные файлы, а также файлы в виде интерпретируемого кода на языке PHP.

На основе разработанного программного обеспечения формулируется оценка потребной производительности системы и разрабатываются требования к аппаратному обеспечению АСАФ НР. Рассматриваются варианты реализации АСАФ НР в системах различного уровня: системы с малой пропускной способностью, системы со средней пропускной способностью, системы обработки в реальном времени и магистральные системы. Также рассматриваются системы мобильной связи.

Четвертая глава посвящена экспериментальному исследованию АСАФ НР.

Рассматриваются цели и задачи экспериментального исследования, среди которых основной является определение оптимальных параметров рассматриваемых в работе алгоритмов, а дополнительной – исследование эффективности рассматриваемых методов.

В работе приводится методика экспериментальных исследований. Согласно используемой методике все эксперименты повторяются 10 раз на различных комбинациях обучающей и тестовой выборов, результаты которых усредняются. Обучение алгоритмов осуществляется на обучающей выборке, а оценка результатов – на непересекающейся с ней тестовой.

Экспериментальные исследования проводятся с использованием большого числа наиболее типичных сообщений каждого класса, отобранных организацией SpamAssassin в результате собственных экспериментальных исследований. На основе массива SpamAssassin сформировано три обучающих выборки разной сложности: SA-E (обычная), SA-H (сложная) и SA-A (объединение первых двух).

Вначале исследованию подверглись методы формирования пространства признаков, включающие: исследование методов формирования множества слов, использующихся для обучения; исследование методов сокращения размерности пространства признаков; исследование методов выбора комплексных признаков.

Результаты исследования методов сокращения размерности пространства признаков для класса ЛПС массива SA-E представлены на рис. 6. Полученные результаты показали, что для всех массивов и всех классов сообщений наиболее эффективным методом выбора значимых признаков является критерий χ^2 при числе признаков около 2000.

Исследование методов выбора комплексных признаков осуществляется на основании отобранных в предыдущем разделе признаков. Исследование алгоритма, описанного в главе 2, приводит к экспоненциальному росту мощности множества кандидатов с ростом числа итераций. В работе рассматривается метод ограничения числа рассматриваемых признаков, среди которых рассматриваются те комплексные признаки, у которых отдельные составляющие слова имеют между собой высокий коэффициент корреляции. Это позволяет отбросить большую часть признаков, которые не представляют собой устойчивых лексических образований.

Из соображений производительности среди признаков с высоким коэффициентом корреляции должны выбираться наиболее частые, т.е. такие, которые встречаются в максимальном числе сообщений.

На рис. 7 представлена гистограмма, показывающая среднее число сообщений в массиве SA-E, содержащих признаки с соответствующим коэффициентом корреляции. (для SA-H – аналогично).

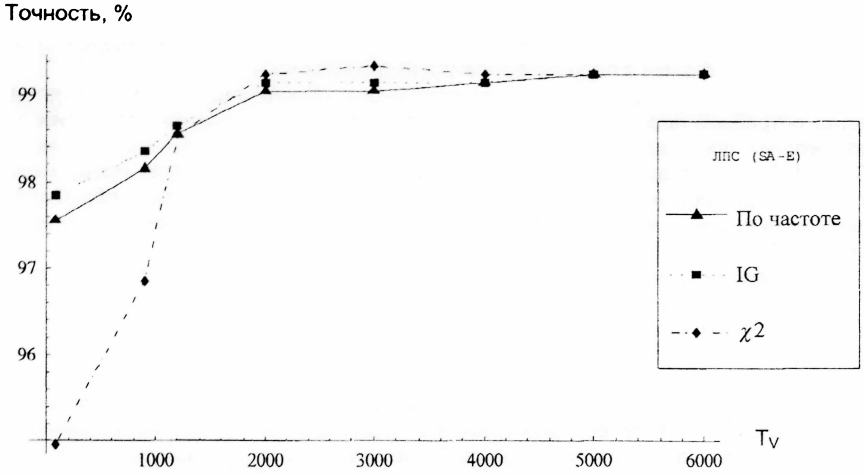

Рис. 6. Точность распознавания ЛПС в массиве SA-E в зависимости от числа признаков T_V, представляющих сообщения

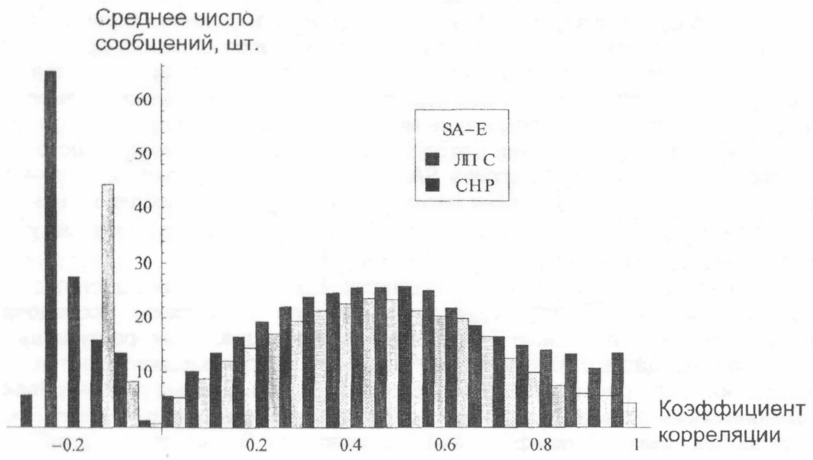

Рис. 7. Среднее число сообщений в массиве SA-E, содержащих признаки с заданным коэффициентом корреляции

Описанным требованиям к комбинированным признакам соответствуют признаки со значениями коэффициента корреляции, близкими к 0,4-0,5.

В следующем эксперименте оценивалось пространства признаков, содержащее фиксированное число признаков, выбор которых осуществлялся из

различных окрестностей точки r=0,5. Результаты данного эксперимента показали, что достаточно использовать признаки из окрестности [0,45; 0.55] коэффициента корреляции.

Исследования числа признаков, отбираемых из заданной окрестности точки r=0.5, показали, что достаточно отобрать порядка 4000 признаков, а результирующая размерность пространства признаков составила порядка 6000 признаков.

Увеличение максимального числа входящих в комплексные признаки слов более 3-х не дало увеличения качества классификации, а на практике достаточно ограничиться числом слов равным 2.

Исследование методов обучения нейронной сети включает следующие шаги: исследование настройки одного нейрона с использованием различных функционалов вторичной оптимизации; исследование числа итераций; исследование выбора шага; исследование выбора числа нейронов первого слоя.

Исследования критериев вторичной оптимизации показали, что рассматриваемые в работе критерии можно упорядочить по эффективности их применения в данной задаче следующим образом: "2а", "1а", "1д", "2д".

При исследовании числа итераций в работе рассматриваются три "стоимостных" варианта: 1) стоимость ошибки при распознавания ЛПС равна стоимости ошибки при распознавании СНР ($\mu = 1$); 2) стоимость ошибки в распознавании ЛПС в три раза выше стоимости ошибки при распознавании СНР ($\mu = 3$); 3) стоимость ошибки в распознавании ЛПС в десять раза выше стоимости ошибки при распознавании СНР ($\mu = 10$).

Результаты значений по обобщенной точности приведены на рис. 8 для массива SA-E (аналогично для массива SA-H).

Зависимость показывает, что пик распознавания достигается при числе итераций порядка 10^6.

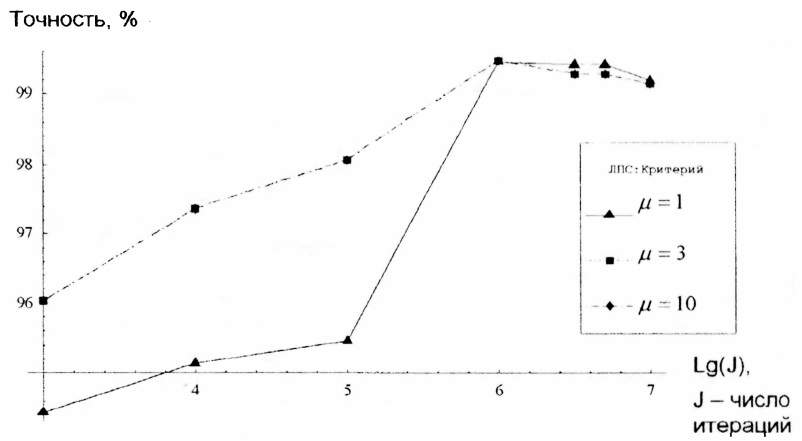

Рис. 8. Зависимость точности фильтрации НР в массиве SA-E в зависимости от числа итераций (J) при различных стоимостных вариантах.

На рис. 9 представлены зависимости, отображающие качество распознавания в зависимости от величины шага обучения нейросети для массива SA-E. Анализ этих кривых позволяет рекомендовать в качестве шага обучения выбрать величину 10^{-4}

Проведено исследование выбора числа нейронов первого слоя. Результаты этого эксперимента приведены на рис. 10 (для массива SA-E). Полученные зависимости имеют общий характер. Полученные данные позволяют

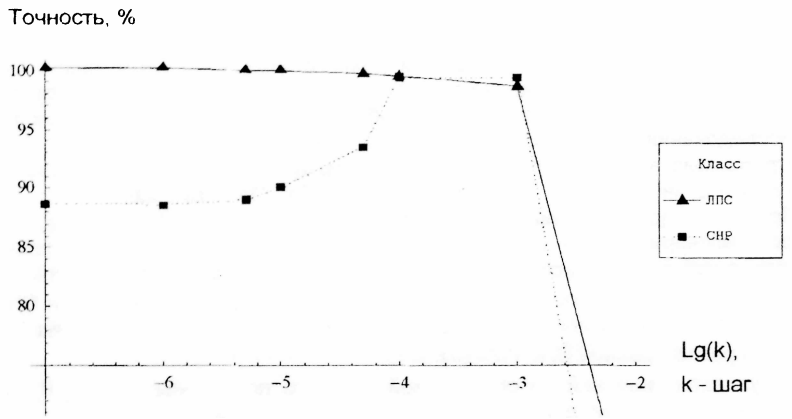

Рис. 9. Точность распознавания сообщений в массиве SA-E при различной величине шага обучения, k

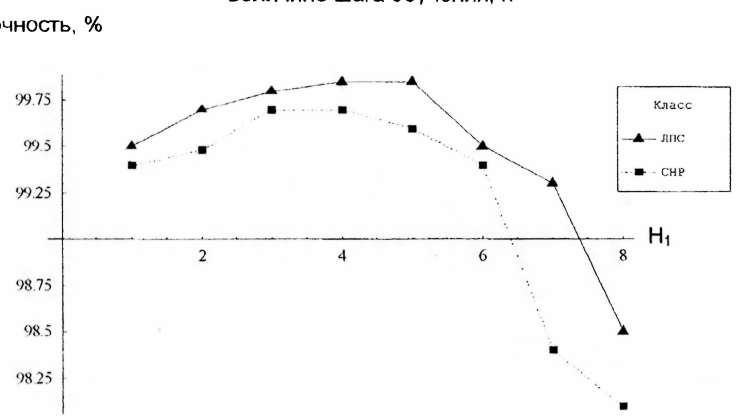

Рис. 10. Зависимость точности распознавания сообщений каждого класса от числа нейронов первого слоя в массиве SA-E, H_1

рекомендовать в качестве числа нейронов первого слоя для выборки SA-E 4-5 нейронов, тогда как для выборки SA-H – 3 нейрона (рисунок не приводится).

Исследование порогов позволяет определить величины порогов T_S и T_H. Пороги позволяют отделить диапазон выходных значений нейронной сети, в которых нейронная сеть может иметь ошибку распознавания (диапазон "U"), от диапазонов уверенного распознавания НР (диапазон "S", порог T_S) и ЛПС (диапазон "H", порог T_H). Диапазон S соответствуют мнемоническим кодам SE и ST, диапазон H – HE и HT, а диапазон U – кодам HC и SC.

В экспериментах данного раздела осуществляется выбор порогов T_S и T_H, обеспечивающих минимум числа сообщений, попадающих в поддиапазон "U". Результаты этих экспериментов представлены в табл. 1

Таблица 1.

Исследование порогов распознавания сообщений каждого класса

№ п/п	Массив	T_S	$P_{S->S}$	$P_{S->U}$	T_H	$P_{H->H}$
1	SA-E	-18	87,0%	13,0%	12	89,2%
2	SA-H	-22	83,7%	16,3%	29	78,4%

Из табл. 1 следует, что при 100% распознавании ЛПС для массива SA-E будет пропущено порядка 13% сообщений класса СНР, а для массива SA-H порядка 16,3%.

В экспериментальных исследованиях коллективной фильтрацию имитировалась совместная работа двух агентов, получающих одновременно два непересекающихся подмножества сообщений. Оценка результатов производилось по тестовой выборке. Результаты, усредненные по обоим агентам приведены в табл. 2.

Таблица 2.

Результаты исследования методов совместной фильтрации

№	Массив	$P_{S->S}$	$P_{H->H}$	$P_{S->U}$
1	До применения совместной фильтрации	86,0%	88,9%	14,0%
2	После применения совместной фильтрации	91,2%	94,1%	8,8%

Таким образом, результаты экспериментов демонстрируют, что применение методов совместной фильтрации позволило повысить качество фильтрации на 5,2%, сократив число пропущенных сообщений НР до 8,8%.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Проведена классификация и систематизация существующих автоматизированных систем фильтрации незапрашиваемой рассылки (АСФ НР) по архитектуре (функциональному строению), способам реализации их основных функций, а также использующемуся в них математическому аппарату.

2. Исследованы и разработаны математические методы и алгоритмы, позволяющие производить автоматическое формирование обучающей выборки с учетом указаний пользователей о выявленных ошибках фильтрации.

3. Исследованы и разработаны математические методы формирования пространства признаков в задаче анализа содержания сообщений, что позволило повысить точность анализа за счет учета значимых словосочетаний.

4. Исследованы и разработаны методы синтеза нейросетей с переменной структурой, показана эффективность применения нейросетей этого класса в задаче фильтрации НР, определены их оптимальные параметры.

5. Разработан аппаратно-программный комплекс фильтрации НР на основе предложенных в работе принципов.

6. Проведены экспериментальные исследования математических методов и алгоритмов, использующихся в АСАФ НР на реальном материале сообщений электронной почты Интернет, определены оптимальные параметры алгоритмов, предложенных в работе, а также методики их автоматического выбора. Результаты проведенных экспериментов показали, что при 0% ложной фильтрации пользовательских сообщений разработанная система обеспечивает пропуск порядка 8,8% сообщений НР, что превосходит аналогичные показатели существующих систем данного класса.

ОСНОВНЫЕ ПОЛОЖЕНИЯ ДИССЕРТАЦИИ ИЗЛОЖЕНЫ В СЛЕДУЮЩИХ РАБОТАХ:

1. Цыганов И.Г., Власов А.И. Архитектура корпоративной многоагентной автоматизированной системы фильтрации информационных потоков // Информационные технологии. – 2005. – №1. – С. 34-41.

2 Цыганов И.Г., Власов А.И. Адаптивная фильтрация информационных потоков в корпоративных системах на основе механизма голосования пользователей // Информационные технологии. – 2004. – №9. – С. 12-19.

3. Цыганов И.Г., Смирнова Е.Г. Исследование и анализ предпосылок распространения незапрашиваемой рассылки в глобальных гетерогенных сетях передачи информации // Научно-технический сборник. ВТУ при Спецстрое России. - 2004. – Вып. 8. – С. 114- 133.

4. Цыганов И.Г., Руденко М.И. Метрики текстов в автоматизированных системах обработки информации // Наукоемкие технологии и интеллектуальные системы: Сборник научных трудов VI Международной молодежной научно-технической конференции. – М., 2004. – С. 86-93.

5. Цыганов И.Г. Решение задачи автоматизированной контекстной классификации с помощью стохастического моделирования лингвистической структуры категорий // Наукоемкие технологии и интеллектуальные системы: Сборник научных трудов VI Международной молодежной научно-технической конференции. – М., 2004. – С. 178-185.

6. Цыганов И.Г. Оценка применимости нейросетевых парадигм при решении задачи сквозного семантического анализа текстовых сообщений // Наукоемкие технологии и интеллектуальные системы: Сборник научных трудов V Международной молодежной научно-технической конференции. – М., 2003. – С. 66-77.

7. Цыганов И.Г. Применение нейросетевых методов для фильтрации SPAM сообщений // Информатика и системы управления в XXI веке: Сборник научных трудов Международной молодежной научной конференции. – М., 2002. – С. 26-33.

8. Цыганов И.Г. Нейросетевые методы автоматизированного анализа информационных потоков в масштабе реального времени // Студенческая научная весна – 2002: Сборник докладов студенческой научной конференции с международным участием. – М., 2002. – С. 19-24.

Подписано к печати 24.05.2005. Заказ №165. Объем 1,0 п.л. Тираж 100 экз.
Типография МГТУ им. Н.Э. Баумана, 107005, г. Москва, 2-ая Бауманская ул., д. 5.