К.Г. Паутов
Центр новых информационных технологий БТИ АлтГТУ
Научный руководитель: д.т.н., профессор, Ф.А. Попов

ИНФОРМАЦИОННАЯ СИСТЕМА УЧЕТА, АНАЛИЗА И ТЕМАТИЧЕСКОЙ РУБРИКАЦИИ ИНТЕРНЕТ-РЕСУРСОВ В ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЯХ

Цель проекта

Цель проекта – проектирование и реализация комплексной информационной системы учета, анализа и тематической рубрикации ресурсов Интернет в телекоммуникационных сетях. В основе проекта лежит задача – повышение эффективности и безопасности использования ресурсов Интернет путем выявления пользовательских предпочтений и тематической рубрикации англо- и русскоязычных сайтов.

Сегодня количество пользователей Интернета в России по различным оценкам достигает 68,7 миллионов человек в возрасте от 18 лет и старше [1]. Ежедневно интернетом пользуется 89% подростков 12-17 лет. С возрастом частота использования Интернета подростками растет, достигая максимума у 17 - летних юношей, среди которых каждый день в Интернет выходят 96% [2]. Среди опрошенных 89% считают, что взрослые должны следить за тем, какие сайты посещают их несовершеннолетние дети; 74% уже следят за тем, чтобы дети не посещали «опасные» сайты, а 14% установили программы для обеспечения доступа детей только к безопасным сайтам [3]. Увеличивается количество детей в Сети, за счет повсеместного проникновения мобильный устройств и сервисов, оснащения учебных заведений компьютерными классами и библиотеками, предоставляющими доступ к сети Интернет. Такой бесконтрольный доступ к источникам информации несет в себе серьезную опасность, в первую очередь, психическому и психологическому здоровью несовершеннолетних пользователей.

С другой стороны, сотрудники компаний зачастую используют неконтролируемый доступ к ресурсам Интернет в рабочее время для решения личных вопросов, общения и развлечения, что отрицательно влияет на эффективность их работы и снижает производительность корпоративной сети.

В связи с этим актуальной становится задача разработки инструмента для защиты пользователей корпоративных сетей от нежелательного и (или) опасного контента, и его несанкционированного использования. Реализация такого инструмента тесно связана с тематической рубрикацией ресурсов Интернет, имеющей много различный применений, среди которых наибольший интерес представляют определение пользовательских предпочтений и фильтрация нежелательной информации.

Общее описание проекта

Доступ в Интернет в большинстве современных организаций, как правило, осуществляется через прокси-сервер. Такой подход получил широкое распространение благодаря ряду преимуществ: снижение нагрузки на внешние каналы связи за счет кеширования популярных ресурсов, защита компьютеров внутренней сети от некоторых видов сетевых атак, обеспечение анонимности клиентов, возможность вести учет трафика и контролировать доступ пользователей к ресурсам, фильтровать рекламу при помощи внешних компонент и т.д.

Разработанная программная система позволяет решать следующие задачи:

1. проводить синтаксический разбор журнальных файлов прокси-сервера и их представление в удобной для понимания человеком форме;
2. анализировать информацию об обращениях из внутренней сети к источникам данных в Интернет;
3. осуществлять тематическую рубрикацию ресурсов Интернет по каталогу с иерархической структурой;

4. составлять тематические портреты пользователей, на основе анализа просматриваемых ими ресурсов;
5. выявлять группы пользователей со схожими предпочтениями относительно выбора источников информации и их тематик;
6. выявлять и анализировать информационные ресурсы, оказавшие наибольшее влияние на формирование тематического портрета пользователя или группы пользователей;
7. формировать списки доступа для фильтрации ресурсов «нежелательной» тематики: пропаганда насилия, экстремизм, нелегальный и запрещенный контент и пр.;
8. представлять отчеты с указанием количественных характеристик входящего трафика в разрезе отдельных пользователей и групп; агрегированные отчеты по источникам данных (информационным ресурсам) и периодам времени;
9. представлять отчеты о тематическом составе входящего трафика, агрегированные по пользователям, группам пользователей и периодам времени.

Использование методов классификации позволяет назначить каждый информационный ресурс одной из тематических категорий (рубрик). Тем самым решается задача определения качественного состава входящего трафика и ограничения доступа к ресурсам «нежелательной» тематики.

На рисунке 1 представлена диаграмма, отражающая качественный (тематический) состав входящего трафика организации в целом за 6 месяцев. Из диаграммы видно, что бóльшая часть входящего трафика организации приходится на категории «Высокие технологии» и «Развлечения». Такое представление существенно удобнее для восприятия, по сравнению с отчетами в разрезе отдельных информационных ресурсов. Кроме того, информационная система позволяет провести анализ списка ресурсов, попавших в ту или иную категорию (подкатегорию), и настроить параметры фильтрации входящего трафика. Запросы к информационным ресурсам, отнесенным к категории запрещенных для просмотра, могут быть автоматически заблокированы и (или) перенаправлены на указанную web-страницу.

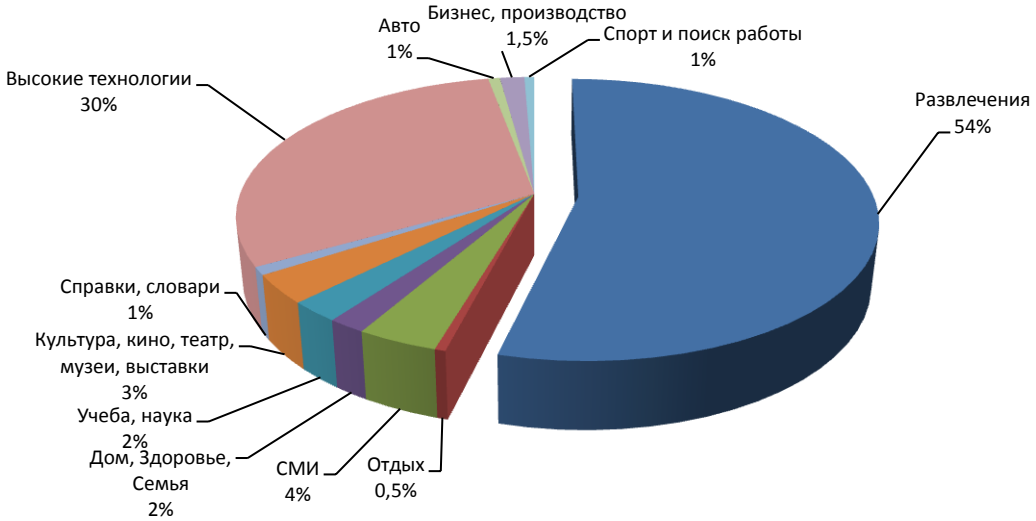

Рисунок 1. Распределение трафика организации по тематическим категориям

Также в рамках программной системы, решается задача определения пользовательских предпочтений, путем составления «тематических портретов» пользователей. Тематический портрет пользователя строится на основе категорий просматриваемых им ресурсов. Анализ тематических портретов позволяет выявлять группы пользователей, интересующихся определенной тематикой (или несколькими тематиками).

В результате, руководитель организации получает возможность видеть целостную картину, позволяющую оценить эффективность работы отдельных сотрудников и групп. Для

специалистов в области системного и сетевого администрирования рассматриваемая ИС может являться инструментом мониторинга и управления входящим трафиком.

Отличия от аналогов

Системы анализа и фильтрации трафика можно классифицировать по двум основным признакам: способ и время анализа трафика. По способу анализа трафика все системы можно разделить на два больших класса: 1) анализирующие лишь метаинформацию о ресурсе; 2) анализирующие содержимое ресурса (с выделением содержательной части ресурса). По времени анализа все системы также можно разделить на два класса: 1) системы онлайн-анализа, т.е. анализирующие информацию во время запроса пользователем Интернет-ресурса; 2) системы офлайн-анализа, обрабатывающие информацию после того, как пользователь получил доступ к ресурсу.

В рамках проекта разработаны алгоритмические, программные и архитектурные решения для учета, анализа и тематической рубрикации ресурсов Интернет в локальных сетях, анализирующие содержательную часть web-страницы в отложенном режиме. Отличительными особенностями являются:

- возможность проводить рубрикацию информационных ресурсов по каталогу с иерархической структурой;
- возможность редактирования тематических категорий и настройки классификатора;
- возможность обнаружения и анализа групп пользователей со схожими предпочтениями относительно выбора информационных ресурсов, за счет использования разработанной методики составления тематических профилей пользователей.

Научная составляющая

Для решения задач учета, анализа и тематической рубрикации ресурсов Интернет с целью повышение эффективности и безопасности их использования путем выявления пользовательских предпочтений и тематической рубрикации англо- и русскоязычных сайтов применялись теория информационного поиска и методы машинного обучения.

Общественная и экономическая полезность

По предварительным оценкам внедрение ИС учета, анализа и тематической рубрикации ресурсов Интернет позволит повысить эффективность использования внешних каналов связи, за счет контроля и управления входящим сетевым трафиком. Ограничение доступа к развлекательным ресурсам может послужить стимулом для повышения эффективность труда сотрудников организации. Контентная фильтрация опасных ресурсов позволит существенно обезопасить работу в Интернете детей и подростков.

Следует отметить, что ИС учета, анализа и тематической рубрикации ресурсов Интернет работает в автоматическом режиме, что позволит снизить возможные расходы на поддержание списка разрешенных (или запрещенных) ресурсов в актуальном состоянии, за счет привлечения труда экспертов.

Техническое обоснование проекта

Ядро ИС учета, анализа и тематической рубрикации ресурсов Интернет состоит из пяти функционально законченных подсистем и модулей (рис. 2). Рассмотрим их более подробно.

Модуль загрузки и синтаксического разбора log-фалов, предназначен для загрузки и обработки журнальных файлов прокси-сервера. В log-файлах содержится информация об обращениях клиентов внутренней сети к ресурсам Интернет, такая как время обращения, идентификатор клиента, адрес узла, к которому осуществлялся запрос, объем полученной информации, и ряд служебных данных. Модуль осуществляет синтаксический разбор (парсинг) содержимого файлов, преобразует данные в удобный для последующей обработки формат, и заносит их в таблицу базы данных.

Модуль обработки URL-адресов предназначен для получения списка URL-адресов источников данных, которые в дальнейшем будут использоваться для анализа и агрегации. Отбор URL-ресурсов производится на основе данных о количестве обращений к ресурсу, количестве пользователей, объеме полученной информации. Таким образом, находится «базовый URL», остальные URL усекаются по маске до домена заданного уровня. Список обработанных URL-адресов заносится в таблицу БД для последующего рубрицирования.

Подсистема тематической классификации производит рубрицирование ресурсов по каталогу с иерархической структурой. В основе механизма тематической классификации лежат алгоритмы определения содержательной части web-страницы и классификации текстов.

Подсистема построения и кластеризации тематических профилей пользователей. Как уже было сказано выше, тематический профиль пользователя формируется исходя из категорий посещаемых им ресурсов Интернет. К полученным тематическим профилям пользователей применяются методы машинного обучения (кластеризации). В результате формируются группы (кластеры) пользователей, имеющих общие интересы. Например, в один кластер могут быть объединены пользователи, предпочитающие определенный новостной сайт, играющие в онлайн-игры, общающиеся в одной из социальных сетей, работающие в ОС Linux, совершающие покупки онлайн, собирающиеся в отпуск за границу, просматривающие фильмы в онлайн-кинотеатрах и т.д.

Подсистема контентной фильтрации формирует списки доступа к ресурсам Интернет и поддерживает их в актуальном состоянии.

Административный интерфейс служит для настройки параметров системы, обучения классификаторов, форматирования отчетов и выполнения иных административных функций.

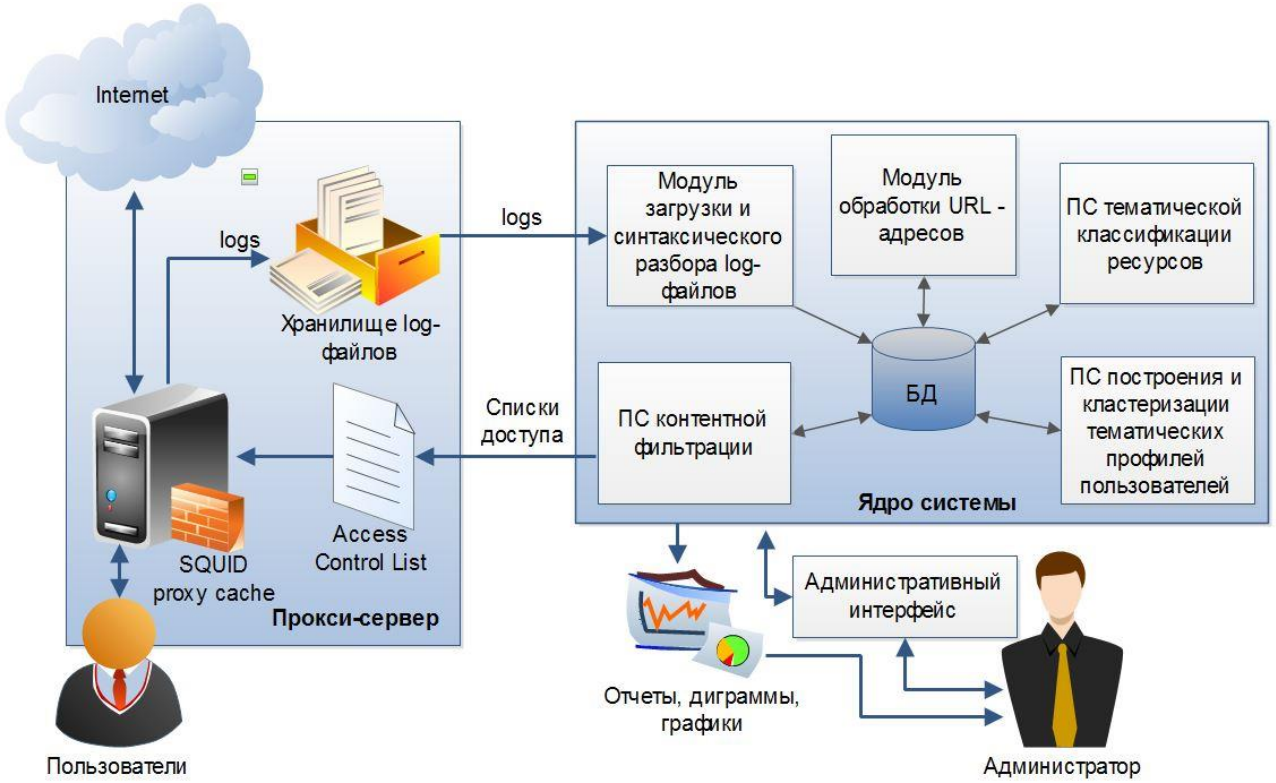

Рисунок 2. Структурная схема программной системы

Программная система разработана на объектно-ориентированном языке программирования Python 2.7 с использованием свободно-распространяемых библиотек для научных расчетов SciPy и NumPy. Для хранения и управления данными используется СУБД PostgreSQL 9.3.0.

Заключение

Предварительный анализ результатов эксперимента показал, что внедрение разработанных решений для учета, анализа и тематической рубрикации веб-страниц и их использования в системе контентной фильтрации позволит сократить объемы входящего сетевого трафика на 30-40%, за счет ограничения доступа пользователей к непрофильным и потенциально опасным ресурсам Интернет. Составление тематических профилей пользователей и их кластеризация позволит строить кластеры пользователей со схожими предпочтениями относительно выбора информационных ресурсов. Внедрение разработанных решений в области системного и сетевого администрирования позволит снизить нагрузку на сетевое оборудование и внешние каналы связи, прогнозировать и обосновывать расходы на развитие телекоммуникационной инфраструктуры.

Список использованных источников

1. Интернет в России: динамика проникновения. Зима 2013 - 2014 гг. Аналитический бюллетень / А. Рыжова, С. Борисова, Ю. Чеканова. – М.: ООО «инФОМ», 2014. – [Электронный ресурс]. – Режим доступа: http://fom.ru/SMI-i-internet/11417
2. Цифровая компетентность подростков и родителей. Результаты всероссийского исследования / Г.У. Солдатова, Т.А. Нестик, Е.И. Рассказова, Е.Ю. Зотова. — М.: Фонд Развития Интернет, 2013. — 144 с.
3. Безопасность детей в Сети.– Фонд Общественное мнение. – [Электронный ресурс]. – Режим доступа: http://fom.ru/SMI-i-internet/11115