не только наличия обычной техники генерирования, сбора и обработки данных, но и создания единой информационной образовательной инфраструктуры, т.е. создания системы сбора и обработки данных в заранее определенных точках образовательной цепи, обмена информацией между точками и передачи информации на различные уровни образования. Задача ИУО – управление информационным потоком по всей образовательной цепи на всех образовательных уровнях, обеспечение детальной информации для оптимизированного текущего и перспективного планирования, информационная поддержка стратегического управления образованием.

Большое разнообразие и объем собираемых данных требуют и системного подхода к их обработке. В мире объем самой разной информации, передаваемой через информационно-телекоммуникационную инфраструктуру, удваивается каждые 2-3 года. Проблема «информационных перегрузок» решается сегодня с помощью извлечения из всего массива данных необходимой для нужд пользователя (учреждение образования) информации путем применения совершенных средств отбора, дальнейшей обработки и своевременного обновления информации. Современные технологии позволяют решать вопросы сжатия внутриорганизационной и внешней информации, использования коммерчески выгодных интерфейсов, трансфера совместно используемых знаний между организационными подразделениями и федеральными структурами органов власти.

Чтобы полностью реализовать сулимые информационными технологиями преимущества необходимо рационализировать и модернизировать как управленческий процесс образования, так и организационную структуру учреждений образования.

Фактически деятельность любого учреждения образования представляет собой ничто иное, как совокупность выработанных в повседневной практике деловых процессов, в которые вовлечены финансовые, материальные, кадровые, информационные и прочие виды ресурсов. Именно деловые процессы определяют порядок взаимодействия отдельных сотрудников и целых подразделений, а также принципы построения информационных систем. Поэтому автоматизация сферы образования, исходя из делового процесса, наиболее логична, и самое главное, – вполне реальна.

Таким образом, непрерывный контроль над ИУО позволяет систематически совершенствовать процесс обучения. Поскольку вся ключевая информация об организации процесса обучения представлена в машинной форме, она может быть очень быстро оценена с применением компьютера. Преподаватели (учителя), руководители учреждений образования сами с помощью имеющихся средств могут легко вносить изменения в реализуемые процессы. При этом нужно обязательно учитывать человеческий фактор. Технико-организационная адаптация образовательных процессов должна осуществляться всегда в сочетании с кадровыми мероприятиями. Следовательно, систематическая переподготовка руководителей сферы образования и преподавателей (учителей) должна стать важной составной частью текущего совершенствования информационного управления образования.

Свечников С.В.

РАЗРАБОТКА МЕТОДОВ АВТОМАТИЧЕСКОГО ПОИСКА, АНАЛИЗА И КАТЕГОРИЗАЦИИ ИНТЕРНЕТ-РЕСУРСОВ ДЛЯ ОЦЕНКИ ЭФФЕКТИВНОСТИ ФУНКЦИОНИРОВАНИЯ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ

ssv@informika.ru

Федеральное государственное учреждение «Государственный научно-исследовательский институт информационных технологий и телекоммуникаций» (ФГУ ГНИИ ИТТ «Информика»)

г. Москва

За последние несколько лет активное развитие информационно-коммуникационных технологий привело к тому, что объем информационных ресурсов значительно вырос. Этот всевозрастающий объем информации, а также ее различные виды представлений (текстовая, графическая, аудио-, видеоинформация) приводят к проблемам, связанным с бесконтрольным доступом к сети интернет.

В сегодняшнем деловом мире использование интернета является неотъемлемым процессом, также как растет количество пользователей интернета, так растет и количество доступной информации в нем. Наряду с преимуществами интернет является и самым большим источником опасностей - почти каждый день появляются новые вредоносные материалы, такие как спам, агрессивный контент и шпионские программы.

Российский сегмент сети – один из самых быстроразвивающихся, количество пользователей интернета по различным данным около 26 миллионов человек, из них 2 миллиона детей [3].

Обеспечение учебных заведений и публичных библиотек доступом к сети интернет увеличивает количество учащихся, пользующихся различными сервисами и информационными источниками, предоставляемыми глобальной сетью. Такой бесконтрольный доступ к сети интернет может привести к серьезным угрозам для детей и учащихся.

Также интернет бесконтрольно используется в личных целях работниками умственного труда, имеющими доступ к глобальной сети, что снижает эффективность их работы и снижает производительность корпоративной сети [1].

При этом методы прямого регулирования (цензуры) неэффективны, встречают протест пользователей интернета и юридически несостоятельны, поскольку противоречат естественным правам граждан на свободу воли, высказываний и волеизъявления.

В связи с этим решение этой проблемы надо искать не в цензуре, а в предоставлении инструмента защиты от нежелательного контента, который пользователи могут использовать по своей воле и по своему усмотрению [2].

Для этого необходимо создание методов автоматического поиска, анализа и категоризации интернет-ресурсов, которые будут преодолевать указанные недостатки и упорядочивать информацию, представленную в сети интернет для управления доступом к ней.

Необходимость в системах для контроля доступа к интернет-ресурсам не вызывает сомнений. Организации несут значительные расходы не связанные с рабочим процессом, это происходит из-за того, что недобросовестные сотрудники используют интернет в личных целях. Основные расходы связаны с неэффективным использованием рабочего времени и затратами на оплату доступа в интернет.

Применение систем автоматического поиска, анализа и категоризации интернет-ресурсов позволяют значительно сократить расходы, связанные с неэффективным использованием рабочего времени за счет уменьшения нецелевого использования интернета и уменьшения веб-трафика.

В целях повышения гибкости и удобства процесса ограничения доступа к интернету, такие системы поддерживают тематическую категоризацию интернет-ресурсов.

Суть таких систем заключается в декомпозиции объектов информационного обмена, анализе содержимого этих компонентов, определении соответствия их параметров принятой политике использования интернет-ресурсов и осуществлении определенных действий по результатам анализа.

Основные представленные на российском рынке программные продукты в области автоматического поиска, анализа и категоризации интернет-ресурсов, принадлежат следующим компаниям.

Производитель	Страна	Программный продукт
Secure computing	США	Sentian
Surfcontrol	США	Surfcontrol web-filter
Websense	США	Websense Enterprise
Cobion	Германия	Proventia Web Filter

Другие системы или не поддерживают фильтрацию русскоязычных интернет-ресурсов или являются не пригодными для корпоративной эксплуатации.

Перечисленные решения представляют собой программы, которые устанавливаются в локальной сети организации и работают на принципе анализа и категоризации интернет-ресурсов.

Все перечисленные системы прекрасно фильтруют, в первую очередь, англоязычный контент. При работе с русскоязычным контентом указанные продукты демонстрируют:

- неполноту базы данных русскоязычных ресурсов;
- систематическую погрешность категорирования сайтов, связанную с неучетом российских социально-политических реалий;
- систематическую погрешность категорирования сайтов, связанную, как правило, с полностью автоматическим определением категорий русскоязычных сайтов;
- низкую оперативность обновления.

В связи с представленными недостатками существует необходимость создания такой системы, адаптируемой для русскоязычных интернет-ресурсов.

Для реализации системы автоматического поиска, анализа и тематической категоризации необходимо разработать модель автоматического поиска и метод тематического анализа текстовой информации, основанные на общеизвестных моделях информационного поиска (таких, как булевская модель, векторная модель, вероятностная модель), а также лингвистических и статистических методах анализа информации.

Задача автоматического поиска и тематической категоризации текстовой информации предполагает решение следующих подзадач:

- отнесение текстовой информации к той или иной категории;
- определение степени соответствия информации категории.

Представленные подзадачи связаны, в первую очередь, с анализом текстовой информации, т.е. ее содержанием.

Разрабатываемая модель автоматического поиска будет основана, в первую очередь, на векторной модели поиска, суть которой сводится к представлению документов и запросов в виде векторов.

Каждому терму (слову или словосочетанию) t_i в документе d_j и запросе q сопоставляется некоторый неотрицательный вес w_{ij}. Таким образом, каждый документ и запрос может быть представлен в виде k-мерного вектора:

$$\vec{d}_j \overset{def}{=} (w_{1j}, w_{2j}, ..., w_{kj})$$

где k - общее количество различных термов во всех документах.

Согласно векторной модели, близость документа d_j к запросу q оценивается как корреляция между векторами их описаний. Эта корреляция может быть вычислена, например, как скалярное произведение соответствующих векторов описаний [4].

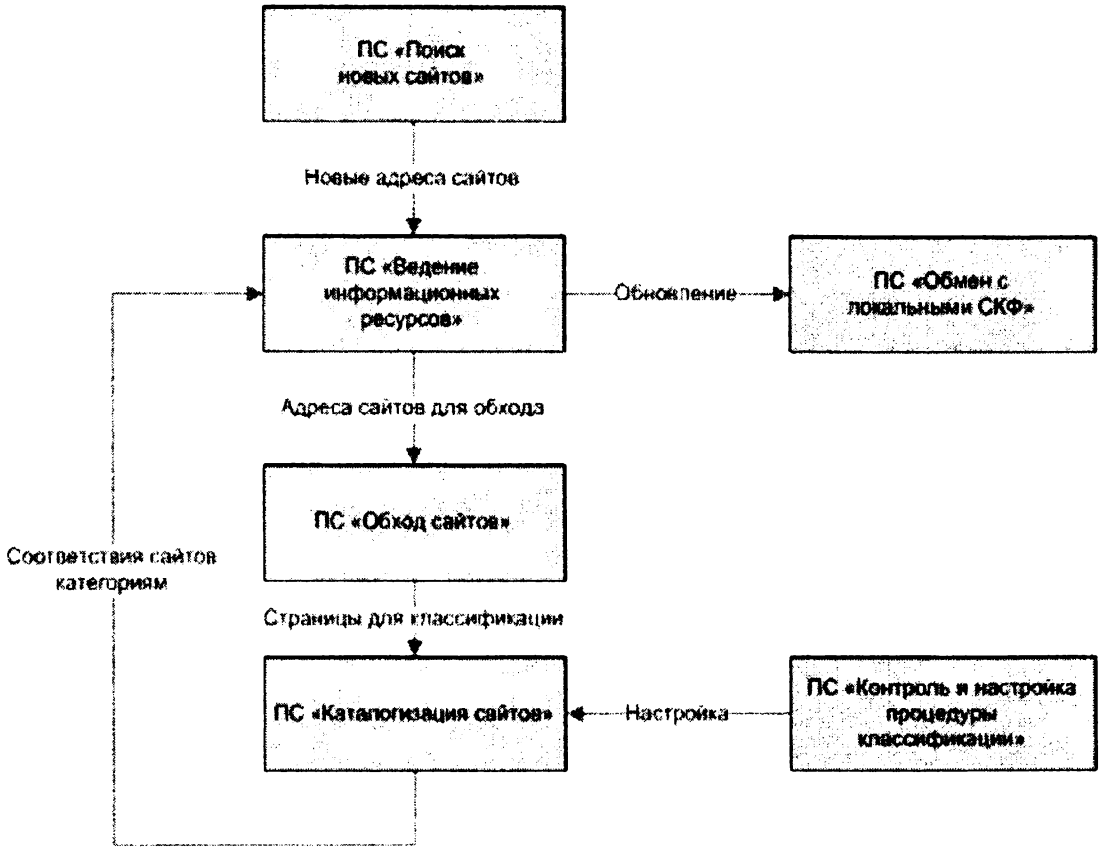
Предлагаемая модель также будет использовать принцип вероятностного ранжирования, который заключается в следующем - наивысшая общая эффективность поиска достигается в случае, когда результирующие документы ранжируются по убыванию вероятности их релевантности запросу. Сначала для каждого документа оценивается вероятность того, что он релевантен запросу, а затем по этим оценкам выполняется ранжирование документа.

Вся совокупность представленных на сегодняшний день методов тематического анализа текстовой информации делится на две группы:

- лингвистический анализ;
- статистический анализ.

Лингвистический анализ ориентирован на извлечении смысла текста по его семантической структуре. Статистический анализ – на частотном распределении слов в тексте. Разрабатываемый метод тематического анализа текстовой информации будет использовать сочетание обеих методик.

Основная структура разрабатываемой системы автоматического поиска, анализа и категоризации интернет-ресурсов и взаимодействие ее подсистем выглядит следующим образом:



Подсистема «Поиск новых сайтов» предназначена для поиска новых интернет-ресурсов. Результатом ее деятельности является набор новых адресов сайтов, пополняющих базу тематической категоризации.

Далее вся информация о новых ресурсах поступает в подсистему «Ведение информационных ресурсов». При этом есть только базовая информация о ресурсе, он не привязан к категориям.

После этого в работу включается подсистема «Обход сайтов». В рамках данной подсистемы осуществляется обход сайта и получение набора страниц, которые можно анализировать тематически.

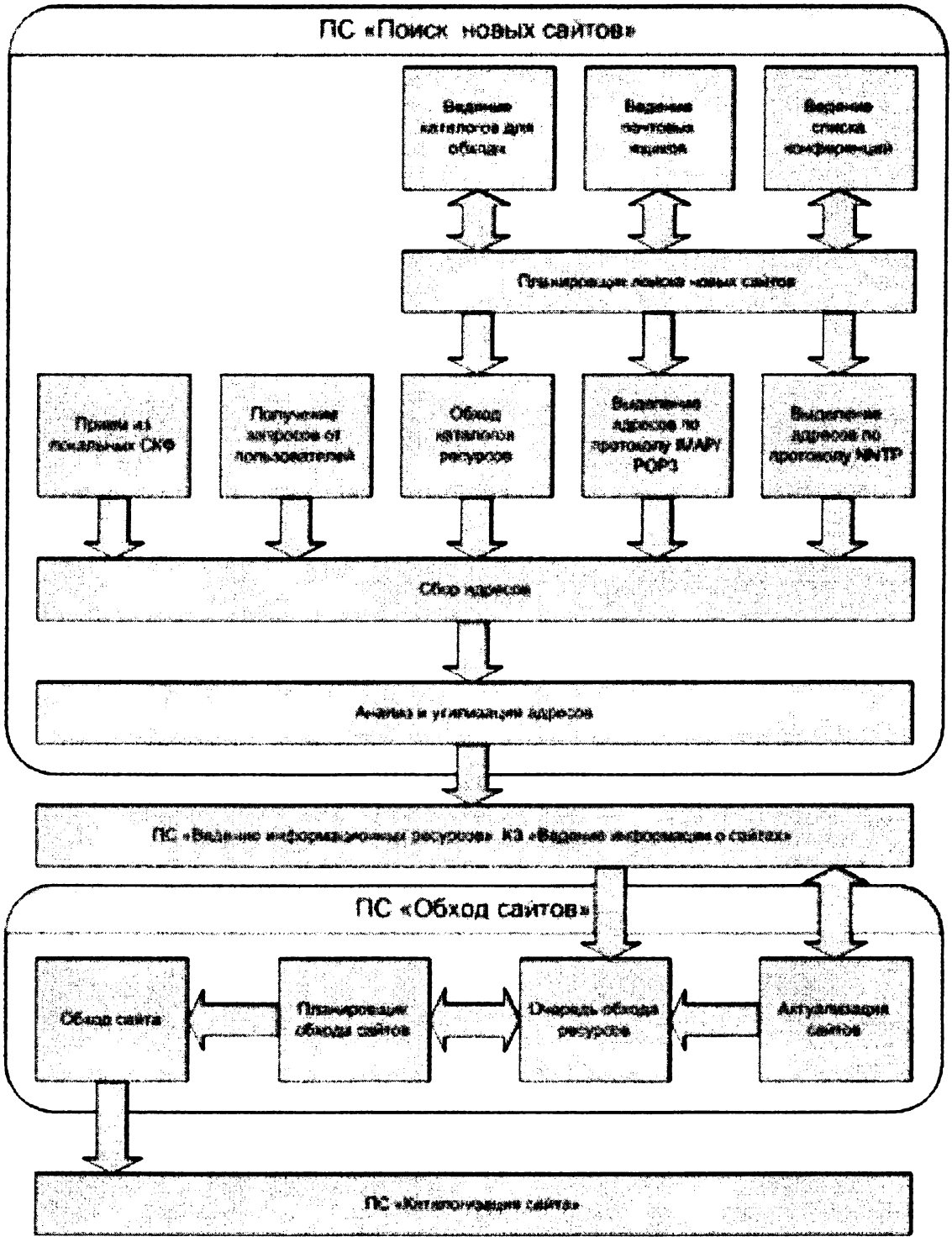
Следующим является подсистема каталогизации сайтов, которая анализирует тексты страниц, составляет их профиль и на основании этого решает об отнесении сайта к той или иной тематической категории.

Процессом, контролирующим качество классификации, управляет подсистема «Контроль и настройка процедуры классификации».

В результате классификации в рамках подсистемы «Ведение информационных ресурсов», сайты получают соответствие категориям.

Дополнительно ведется специализированный журнал отслеживания изменений о сайтах и категориях, который используется подсистемой «Обмен с системами контентной фильтрации (СКФ)» для обновления данных в базах СКФ и получения от них новых неизвестных адресов для анализа.

Более подробно процесс получения данных о сайтах и их категоризации представлен на следующей схеме:



В рамках данной схемы более детально расписаны источники получения новой информации и процесс передачи информации через очереди задач.

Литература

1. А.Абсалямов Борьба с киберслэкингом. Windows 2000 Magazine, №3 2000.
2. И.Е.Поляков Опыт создания системы фильтрации агрессивного web-контента Труды XII Всероссийской научно-методической конференции «Телематика 2005», 6-9 июня 2005г., Издательство во СПб.
3. Фонд «Общественное мнение», http://www.fom.ru/
4. G. Salton, M.J. McGill. Introduction to modern Information Retrieval. McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983