УДК 004.457

А.А. Масюк, С.Э. Сараджишвили
Санкт-Петербург, Россия

ТЕМАТИЧЕСКАЯ КАТЕГОРИЗАЦИЯ РЕСУРСОВ В СИСТЕМАХ КОНТЕНТНОЙ ФИЛЬТРАЦИИ

A.A. Masyuk, S.E. Saradgishvili
St.-Petersburg, Russia

THEME CATEGORIZATION RESOURCES IN THE CONTENT FILTERING

Изучены методы фильтрации и управления доступом к ресурсам сети Интернет и создание системы контентной фильтрации, которая должна предоставлять механизм распределения прав доступа к сетевым ресурсам, основанный на централизованной базе данных категоризированных Интернет-ресурсов и списках ключевых слов.

КОНТЕНТНАЯ ФИЛЬТРАЦИЯ. ФИЛЬТРАЦИЯ ПО КЛЮЧЕВЫМ СЛОВАМ. ДИНАМИЧЕСКАЯ ФИЛЬТРАЦИЯ. URL-ФИЛЬТРАЦИЯ. УПРАВЛЯЕМЫЙ ДОСТУП. ПОЛИТИКА ФИЛЬТРАЦИИ. ТЕМАТИЧЕСКАЯ КАТЕГОРИЗАЦИЯ.

The article covers investigation of filtering mechanisms and management of access to Internet-resources. The main purpose of this investigation is to develop the content-filtering system, which should provide the mechanism of access rights distribution to net-resources based on categorized and centralized data base of Internet-resources and key-words.

CONTENT FILTERING. FILTERING BY KEYWORDS. DYNAMIC FILTERING. URL-FILTERING. MANAGEMENT OF ACCESS. POLICY OF FILTRATION. THEME CATEGORIZATION.

Для решения задач обеспечения сетевой безопасности существует большой выбор межсетевых экранов, антивирусов и других программных и программно-аппаратных комплексов.

Однако недостаточно внимания и ресурсов уделялось проблеме управления доступом к формально безопасному, с точки зрения других компонент системы обеспечения безопасности, содержимому сайтов. Данную задачу решают при помощи контентной фильтрации, которая технологически реализуется различными способами, но конечная ее цель – изолировать пользователя от нежелательной информации. Контентная фильтрация может выполняться аппаратными и программными средствами.

Контентная фильтрация – это функция управления доступом пользователей к специфическим типам информации на основе анализа содержимого информационных объектов (веб-запросы пользователей, содержимое веб-страниц) и определения соответствия их параметров принятой политике безопасности [1, 2]. Под политикой безопасности понимается набор правил по доступу к ресурсам сети, которые назначаются для пользователей и групп пользователей.

Система контентной фильтрации (СКФ) – это технологическое решение, основной задачей которого является предоставление сервиса управления доступом пользователей к ресурсам Интернет. СКФ отвечает за управление доступом пользователей к ресурсам Интернет в зависимости от категории запрашиваемых ресурсов и принятой организационной политики [1, 2].

Существующие СКФ хорошо работают с англоязычными ресурсами, но имеют ограниченные возможности работы с данными на других языках. В ситуации с русскоязычными ресурсами контентные фильтры имеют ряд существенных недостатков:

некорректная работа с русскоязычными ресурсами по причине отсутствия специальных инструментов работы с информацией на русском языке;

бедность базы данных категоризированных русскоязычных ресурсов.

Методы категоризации ресурсов в большинстве существующих СКФ имеют нелинейную сложность и чрезмерно требовательны к вычис-

лительным ресурсам. По этой причине большинство СКФ не могут категоризировать ресурсы при первом обращении: пользователь получает доступ, а ресурс категоризируется позже.

Во многих случаях методы категоризации требуют участия человека в проверке результатов для обеспечения приемлемого качества. Обычно участие человека в таких системах сводится к исключению из списков ресурсов, положительно категоризированных из-за несовершенства метода категоризации.

Распространенной ошибкой является определение ресурса как относящегося к категории, на основе единственного термина, встретившегося в тексте, несмотря на то что ресурс имеет тематику, существенно отличающуюся от тематики категории. Таким же образом ресурс может быть отнесен к нескольким категориям одновременно. В подобных методах проблемой является отсутствие интегрального показателя значимости, который позволил бы правильно соотносить ресурсы и категории.

Изложенные основания предопределяют актуальность исследования и разработки модели тематической категоризации, лишенной перечисленных недостатков.

Статья посвящена исследованию методов управления доступом к ресурсам сети Интернет и созданию модели тематической категоризации для системы контентной фильтрации. Система должна предоставлять механизм распределения прав доступа к сетевым ресурсам, основанный на централизованной базе данных категоризированных Интернет-ресурсов и списках ключевых слов (терминов).

Для разрабатываемой СКФ необходимо создать модель категоризации, инвариантную языку анализируемых ресурсов. Метод категоризации, лежащий в основе модели, должен иметь линейную сложность и минимизировать участие человека в оценке качества категоризации. Оценка должна быть формализована.

Методы контентной фильтрации. Существует большое количество методов контентной фильтрации. В СКФ обычно используется более одного метода, что необходимо для обеспечения лучших результатов работы системы, поскольку различные методы позволяют достигать требуемых результатов только на определенном типе фильтруемых данных. Однако можно выделить три наиболее общих класса методов фильтрации [2, 3]:

- фильтрацию с использованием справочников ключевых слов;
- фильтрацию на основе списков IP/URL;
- динамическую фильтрацию.

Фильтрация по ключевым словам – наиболее простой способ фильтрации, поэтому она часто применяется как отдельно (в простых системах для домашнего использования), так и в комбинации с другими методами фильтрации. Этот метод позволяет включать блокировку страницы либо сайта целиком при наличии в них слов или словосочетаний из справочника. Метод прост в реализации и использовании, но имеет существенный недостаток: он может блокировать ресурсы, в которых фильтруемые слова используются в другом контексте (т. н. избыточная фильтрация).

IP/URL-фильтрация позволяет блокировать ресурсы по справочнику IP-адресов или URL (возможен смешанный режим). Справочник может наполняться как вручную, так и автоматически, на основе алгоритма предварительного анализа ресурсов.

Динамическая фильтрация – широкий класс методов, в которых содержимое ресурса анализируется в момент поступления запроса к ресурсу. Доступ к ресурсу блокируется, если его содержимое определяется как несоответствующее политике безопасности.

Основное отличие IP/URL-фильтрации и динамической фильтрации заключается в моменте анализа содержимого ресурса и поведении системы в случае, когда при первом обращении ресурс не классифицирован системой.

Сравнение существующих систем контентной фильтрации. В таблице приведено сравнение наиболее распространенных СКФ по ряду наиболее важных критериев (по принципу наличия или отсутствия функциональности).

Из обзора существующих решений можно сделать вывод, что имеются системы различного класса и различной функциональности, но нет единого подхода к реализации одних и тех же функциональных возможностей даже среди систем одного класса.

Лидирующие системы контентной фильтрации основываются на принципе анализа и категоризации Интернет-ресурсов, что признано наиболее эффективным методом фильтрации нежелательных данных. Эти системы используют регулярно обновляемые базы URL, гибкие

Сравнение систем контентной фильтрации

Критерий	Дозор-Джет	Dans-Guardian	Smooth-Guard-ian	Cyber Patrol	Cyber Snoop	Net Nanny	Cyber Sitter	Wiz-guard	Cyber Sentinel
Фильтрация по IP/URL	+	+	+	+	+	+	+	+	+
Фильтрация по терминам входящих данных	+	+	+	+	+	-	-	-	+
Фильтрация по терминам исходящих данных	+	-	-	-	+	-	-	+	+
Фильтрация по портам	+	-	-	-	-	-	-	-	-
Наличие локальной БД	+	+	+	+	+	+	+	-	+
Наличие центральной БД	+	-	-	+	+	+	+	+	+
Установка на рабочую станцию	-	+	+	+	+	+	+	+	+
Установка на шлюз	+	+	+	-	-	-	-	-	-
Управление временем работы	+	-	+	+	-	-	-	-	-
Сбор статистики	+	-	+	+	+	+	+	+	+
Встроенная поддержка русского языка	+	-	+	-	-	-	-	-	-
Наличие графического интерфейса	-	-	-	+	+	+	+	+	+

настройки фильтра и развитые системы отчетности.

Все рассмотренные системы фильтрации либо не используют центральные БД, либо используют закрытые БД ресурсов, свободный доступ к которым невозможен. Содержание таких БД обычно является наибольшей ценностью для компаний-разработчиков подобных систем.

На данный момент не обнаружено ни одной сколько-нибудь полной БД русскоязычных ресурсов со свободным доступом. По этой причине было принято решение о создании собственной БД и разработке системы тематической категоризации, которая будет наполнять эту БД и поддерживать ее в актуальном состоянии.

Определение тематической категоризации ресурсов. Категория – группа, к которой может быть отнесен сайт на основе некоторых признаков. Категории представлены иерархическим деревом, а в простейшем случае – списком. Категоризация Интернет-ресурсов осуществляется специальной системой тематической категоризации, предоставляющей данную информацию клиентам СКФ [2].

На данный момент можно выделить несколько основных видов категоризации ресурсов:

- использование регулярно обновляемых баз данных категоризированных ресурсов (система категоризации работает со списком категорий, категоризирует новые ресурсы и обновляет связи между категориями и существующими ресурсами);
- категоризация данных «на лету» путем анализа содержимого страниц;

• использование данных о категории, информацию о принадлежности к которой предоставляет сам сайт.

Категоризация данных и формирование баз категорий обычно производится в полуавтоматическом режиме – сначала выполняются анализ содержимого и определение категории с помощью специально разработанных средств. На втором этапе полученная информация часто проверяется людьми, принимающими решение о том, к какой категории можно отнести тот или иной сайт.

Многие компании автоматически пополняют базу категорий по результатам работы клиентов, если обнаруживается сайт, не отнесенный ни к одной из категорий.

В настоящее время используются два основных способа подключения предопределенных баз данных категоризированных ресурсов:

• использование локальной базы категорий с регулярным ее обновлением (данный метод очень удобен для больших организаций, имеющих выделенные серверы фильтрации и обслуживающие большое количество запросов);

• использование базы категорий, размещенной на удаленном сервере (данный метод часто применяется в различных устройствах – межсетевых экранах, ADSL-модемах и т. п.

Использование удаленной базы категорий немного увеличивает нагрузку на каналы связи, но обеспечивает использование актуальной базы категорий.

Стоит заметить, что эти два метода можно удачно скомбинировать, используя центральную базу и небольшой временный список ссылок на клиенте.

К преимуществам применения предопределенных баз категорий можно отнести то, что предоставление или запрет доступа производится еще на этапе выдачи запроса клиентом, что может существенно снизить нагрузку на каналы передачи данных. А главный недостаток использования данного подхода – задержки в обновлении баз категорий сайтов, поскольку для анализа требуется некоторое время. Кроме того, некоторые сайты часто меняют свое наполнение, из-за чего информация о категории, хранящаяся в базе адресов, становится неактуальной. Некоторые сайты также могут предоставлять доступ к разной информации, в зависимости от имени пользователя, географического региона, времени суток и т. п.

Далее будет выполнен обзор нескольких существующих моделей тематической категоризации и приведена новая модель, базирующаяся на методе вычисления относительной значимости терминов.

Простейшая модель категоризации. Тематический профиль – это совокупность данных (перечень терминов), необходимая для принятия решения о принадлежности документа на основе анализа его текстовых данных к заданной категории.

Под термином понимается слово, словосочетание, логическая формула из слов и словосочетаний, содержащая логические операторы.

При автоматическом построении профиля в качестве терминов синтезируются только слова и словосочетания. Кроме того, при автоматическом построении профиля не синтезируются перечни терминов-исключений, когда документ автоматически относится или исключается из категории. Эти операции должен осуществлять эксперт на основе анализа результатов категоризации.

Перед началом категоризации проводится очистка ресурса: удаляется навигационная часть, теги html и скрипты. В случае более глубокого анализа возможно удаление слов, не несущих смысловой нагрузки [4].

Пусть дан ресурс D, представимый как множество элементов текста $D = \{d_k\}$ и категория C, состоящая из двух подмножеств $D = \{Ca, Cb\}$, где $C = \{Ca_i\}$ – множество терминов, которые должны присутствовать в ресурсе D для его отнесения к категории C; $Cb = \{Cb_j\}$ – множество терминов, которые должны отсутствовать в ресурсе D для его отнесения к категории C. Вычислив количество элементов в пересечениях множеств $|D \cap Ca| = Ra$ и $|D \cap Cb| = Rb$, можно сделать вывод о принадлежности ресурса к категории: если $Ra > \tau$ и $Rb < \vartheta$, то ресурс D относится к категории C. Пороговые значения τ и ϑ задаются экспертом, либо вычисляются в процессе обучения.

Представленная модель применялась ранее в системах, предназначенных для домашнего использования. Системы, реализующие данную модель, нетребовательны к вычислительным ресурсам и легко администрируются. Основной проблемой является невозможность учета количества вхождений термина в текст ресурса и веса термина, – все термины имеют одинаковый приоритет, что часто не соответствует требованиям, предъявляемым к системам фильтрации.

Семантическая категоризация. Семантический анализ – процесс выявления смыслового содержания слов и словосочетаний в предложении. Семантический анализ обеспечивает нормализацию синтаксической структуры предложений, распознавание терминов, классификацию терминов по семантическим признакам, с учетом синонимических и гипонемических (отношение «общее – частное») классов, выявление определений терминов.

Степень соответствия найденных документов запросу пользователя характеризуется понятием *релевантность*. Оно не является специфичным для систем информационного поиска. Это понятие появилось из философских теорий, объясняющих относительную связь между источниками информации, и изучается многими направлениями науки. Для организации наиболее релевантного поиска предлагается использовать онтологии.

Онтологии являются новыми интеллектуальными средствами для поиска ресурсов в сети Интернет, новыми методами представления и обработки знаний и запросов. Они способны точно и эффективно описывать семантику данных для некоторой предметной области и решать проблему несовместимости и противоречивости понятий [5, 6].

Формально определим онтологию как множество $O = (L, C, F_l, F_c, R_h)$, где $L = \{(w_i, x_i)\}_{i=1,n}$ – словарь терминов предметной области; w_i – термин; x_i – рейтинг термина w_i относительно других терминов в категории $C = \{c_i\}_{i=1,m}$; $F_l(L) \to C$ – функция интерпретации терминов; $F_c(c_i) \to L$ – функция интерпретации категорий; R_h – отношения иерархии между категориями (концепциями) в онтологии.

Запрос на определение соответствия ресурса категории представляется в виде множества терминов из L: $u = \bigcup_m w_m$.

Итоговая формула для $P(c_i \mid u)$ выглядит следующим образом [7]:

$$P(c_i \mid u) = \sum_{w \in u} \left(\frac{P(w \mid c_i)}{\sum_{c' \in C} P(w \mid c')} \cdot \frac{count(w, L)}{\sum_{w' \in u} count(w', L)} \right),$$ где

$P(w \mid c_i) = x_w^i$ – вероятность вхождения термина w в категорию c_i, вес данного термина в данной категории; $count(w, L)$ – отношение количества вхождений термина w к общей сумме вхождений всех терминов.

Основными недостатками семантических методов категоризации является их привязка либо к определенному языку, либо к группе языков. В приведенной выше модели для каждого языка требуется составление онтологии по каждой группе связанных категорий. Для других подходов семантической категоризации нередко требуется разработка отдельных моделей для групп языков, пример такой системы – POESIA [8].

Следует отметить, что системы фильтрации на основе семантической категоризации дают очень хорошие результаты для тех групп языков, для которых они были разработаны (около 98 % ресурсов категоризируются правильно).

Тематическая категоризация на основе вычисления весовых коэффициентов терминов, принадлежащих категории. Пусть дано множество ресурсов D, разделенное на два непересекающихся подмножества T_r и T_s, называемые *обучающей* и *тестовой выборкой*. На основании обучающей выборки строится классификатор категорий, а на тестовой выборке проверяется качество категоризации. Пусть также дано соответствие между ресурсами и категорией C в виде $\Phi : D \to \{0,1\}$, устанавливающее значение 1, в случае, если ресурс принадлежит категории и нуль – в противоположном случае [9, 10].

Используя подмножество T_r необходимо построить функцию $\Phi' : D \to \{0,1\}$, аппроксимирующую Ф так, чтобы число ошибок на T_s было наименьшим: $E = \sum_{T_s} |\Phi - \Phi'| \to \min$.

Пусть T – множество терминов, выделенное из ресурса категории C. Тогда ресурс можно представить в виде вектора $d_j = (w_{1j}, ..., w_{|T|j})$, где $w_{ij} \in [0,1]$ – нормированный вес термина t_i в ресурсе d_j. В таком случае, категорию можно представить в виде вектора той же размерности, что и вектор ресурса: $C = (c_1, ..., c_{|T|})$, где c_i – вес термина t_i в категории C [11].

Для получения веса термина используется частотный метод вычисления степени соответствия $w_{ij} = \frac{T_{i,j}}{|T_i|} \cdot \frac{|D|}{|D_j|}$, где $T_{i,j}$ – число терминов t_i в ресурсе d_j; T_i – общее число терминов в ресурсе d_j; D_j – число ресурсов, в которых встречается термин t_i; D – общее количество ресурсов категории, $i \in [1, D]$, $j \in [1, T]$. Таким образом, чем чаще термин встречается на странице некоторого ресурса, но реже встречается во всех ресурсах, тем выше будет его вес в данном ресурсе.

Решение о принадлежности ресурса к категории будем принимать, если степень соответствия $CSV(c,d_j)=cd_j=\sum_i c_i d_{ij}$ достигнет некоторого порога τ. Таким образом, получаем: $\Phi'(c,d_j)=\begin{Bmatrix}1,\ CSV(c,d_j)\ge\tau\\ 0,\ CSV(c,d_j)<\tau\end{Bmatrix}$.

Основной проблемой для данной модели является процесс обучения, заключающийся в подборе весовых коэффициентов и порога, начальное значение которого должно задаваться экспертом. Оценка качества категоризации производится с использованием метрик информационного поиска, таких, как точность, полнота и *F*-мера [12]. Процесс обучения и уточнения коэффициентов должен производиться регулярно в связи с увеличением количества ресурсов, относимых к категориям. Методы подбора весовых коэффициентов обладают малой вычислительной масштабируемостью, что не позволяет использовать данную модель в больших системах.

Разработанная модель тематической категоризации на основе метода вычисления относительной значимости терминов. В основе разработанной модели категоризации [2] лежат законы Ципфа–Мандельброта [13]. Пусть дано множество ресурсов $D=\{d_i \mid i\in[1,M]\}$, каждый ресурс d_i с точки зрения модели представляет собой множество терминов $d_i=\{t_j \mid j\in[1,N_i]\}$.

Далее рассмотрим статистические величины, отражающие информационную значимость терминов в множестве ресурсов. Частота встречаемости термина t_j в ресурсе d_i: $DF(t_j,d_i)=\mathrm{Log}_2\frac{count(t_j,d_i)}{\sum_{n\in[1,N_i]}count(t_n,d_i)}$, где $count(t_j,d_i)$ – количество вхождений термина t_j в ресурс d_i. Частота $DF(t_j,d_i)$ является вероятностью выбрать термин t_j в ресурсе d_i при случайном выборе всех вхождений терминов, имеющихся в тексте ресурса.

Инверсная частота встречаемости термина во множестве ресурсов определяет количество информации, получаемое при снятии неопределенности наступления события встречи термина в одном из ресурсов множества: $IDF(t_j,D)=\mathrm{Log}_2\frac{|D|}{|d\in D\mid t_j\in d|}$. Поскольку $\frac{|D|}{|d\in D\mid t_j\in d|}\ge 1$, то $IDF(t_j,D)\ge 0$. Для часто встречающихся терминов $IDF(t_j,D)$ близка к нулю, а для редких терминов она стремится слева к $\mathrm{Log}_2|D|$: $0\le IDF(t_j,D)\le\mathrm{Log}_2|D|$.

Рассмотрим подмножество текстов $D'\subset D$, представляющее собой тематическую группу ресурсов. Пусть $d'\in D'$ и $t_j\in d'$, тогда возможно вычислить величину тематической $IDF(t_j,D')$, а разность $\Delta I(t_j,D,D')=IDF(t_j,D)-IDF(t_j,D')$ определяет изменение информативности термина при отнесении d' к множеству D'. Значения разности могут быть как положительными, так и отрицательными, а поскольку отнесение ресурса к тематическому множеству означает снятие информационной неопределенности относительно тематики документа, но не внесение большей неопределенности, то $\Delta IDF(t_j,D,D')$ можно записать следующим образом: $\Delta IDF(t_j,D,D')=\begin{cases}\Delta I(t_j,D,D'),\ \Delta I(t_j,D,D')>0\\ 0,\ \Delta I(t_j,D,D')\le 0\end{cases}$, тогда можно определить границы значений: $0\le\Delta IDF(t_j,D,D')\le IDF(t_j,D)$ [14].

Согласно [13, 15] под значимостью термина t_j в ресурсе d, входящего в множество D, понимается величина $DFIDF(t_j,d,D)=DF(t_j,d)\cdot IDF(t_j,D)$. По аналогии рассмотрим $DFTIDF(t_j,d,D,D')=DF(t_j,d)\cdot\Delta IDF(t_j,D,D')$, которая характеризует значимость t_j в ресурсе d с учетом того, что $d\subset D'$. Величину $DFTIDF(t_j,d,D,D')$ будем называть характеристикой относительной значимости. Поскольку $0\le\Delta IDF(t_j,D,D')\le IDF(t_j,D)$, то $DFTIDF(t_j,d,D,D')\le DFIDF(t_j,d,D)$.

Пусть задано множество категорий $C=\{c_k \mid k\in[1,N]\}$, Каждая категория представляет собой множество терминов $c_k=\{t_j\}$. Введем метакатегорию терминов, содержащихся во всех категориях $C'=\bigcup_{k\in[1,N]}c_k$. На этапе обучения для всех терминов должны быть рассчитаны $IDF(t_j,C')$ и $\Delta IDF(t_j,c_k,C')$.

Тогда мы можем вычислить суммарную относительную значимость терминов, содержащихся в ресурсе d по отношению к категории c_k: $S(d,c_k,C')=\sum_{t\in c_k\cap d}DFTIDF(t,d,c_k,C')$. Учитывая, что документ может не содержать все термины категории, можем записать нормированный вариант $SN(d,c_k,C')=\frac{S(d,c_k,C')}{\sum_{t\in d}DFIDF(t,d,C')}$, $0\le SN(d,c_k,C')\le 1$. Последнее соотношение позволяет использовать нормированную относи-

тельную значимость для принятия решения о категоризации ресурса: ресурс d относится к категории c_k тогда и только тогда, когда $SN(d, c_k, C') > \tau$, где τ – пороговое значение, задаваемое экспертом, либо определяемое автоматически в процессе обучения.

В статье предложена модель тематической категоризации, в основе которой лежит метод вычисления относительной значимости терминов.

Модель базируется на законах Ципфа–Мандельброта, инвариантных языку категоризируемых ресурсов, и предполагает практически полную автоматизацию процесса категоризации. Для обучения и работы системы требуется только множество категорий и обучающая выборка ресурсов для установки значения порога.

Оценку качества категоризации предложено осуществлять при помощи метрик информационного поиска, таких, как точность и полнота [12]. Вычисляя метрики, можно устанавливать оптимальное значение порога, которое может быть различным: в некоторых случаях потребность в доступе является наиболее важной, поэтому допустимы ложные срабатывания, т. е. полнота важнее точности. В случаях жесткого ограничения доступа точность важнее полноты.

Показано, что процедура категоризации имеет линейную сложность, трудоемкость линейно зависит от количества категорий и терминов. Это позволяет говорить о высокой скорости категоризации, что в свою очередь дает возможность рассматривать предложенную модель как универсальную: категоризация может производиться как на выделенном сервере, так и на локальных машинах пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. **Масюк, А.А.** Контентная фильтрация и управление доступом к ресурсам сети Интернет [Текст] / А.А. Масюк, С.Э. Сараджишвили // Фундаментальные исследования и инновации в технических университетах: Матер. XIV Всерос. конф. по проблемам науки и высшей школы. –СПб.: Изд-во Политехнического ун-та, 2010. –Т. 1. –С. 127–128.

2. **Масюк, А.А.** Контентная фильтрация и управление доступом к ресурсам сети Интернет в образовательных учреждениях [Текст] / А.А. Масюк, С.Э. Сараджишвили // Научно-технические ведомости СПбГПУ. Сер. Информатика. Телекоммуникации. Управление. –2010. –№ 4. –С. 153–162.

3. **Масюк, А.А.** Методы контентной фильтрации [Текст] / А.А. Масюк, С.Э. Сараджишвили // Фундаментальные исследования и инновации в технических университетах: Матер. XIV Всерос. конф. по проблемам науки и высшей школы. –СПб.: Изд-во Политехнического ун-та, 2010. –Т. 1. –С. 128–130.

4. **Некрестьянов, И.С.** Обнаружение структурного подобия HTML-документов [Текст] / И.С. Некрестьянов, Е.Ю. Павлова // Тр. IV Всерос. конф. RCDL'2002. –Дубна, 2002. –С. 38–54.

5. **Добров, Б.В.** Формирование базы терминологических словосочетаний по текстам предметной области [Текст] / Б.В. Добров, Н.В. Лукашевич, С.В. Сыромятников // Тр. V Всерос. науч. конф. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – RCDL'2003. –СПб., 2003.

6. **Загоруйко, Н.Г.** На пути к автоматическому построению онтологии [Электронный ресурс] / Н.Г. Загоруйко, А.М. Налетов, И.М. Гребенкин. –Режим доступа: http://www.dialog-21.ru/Archive/2003/Zagorujko.htm

7. **Захарова, И.В.** Об одном подходе к реализации семантического поиска документов в электронных библиотеках [Текст] / И.В. Захарова // Вестник УГАТУ. Сер. Управление, вычислительная техника и информатика. –Уфа: Изд-во Уфимского государственного авиационно-техн. ун-та. –2009. –Т. 12. –№ 1 (30). –С. 133–138.

8. Website of POESIA Project [Электронный ресурс] / Режим доступа: http://www.poesia-filter.org/

9. **Поляков, И.Е.** Опыт создания системы фильтрации агрессивного web-контента [Текст] / И.Е. Поляков // Тр. XII Всерос. науч.-метод. конф. Телематика 2005. –СПб: Изд-во СПбИТМО.

10. **Sebastiani, F.** Machine Learning in Automated Text Categorization [Электронный ресурс] / F. Sebastiani. –Режим доступа: http://nmis.isti.cnr.it/sebastiani/

11. **Свечников, С.В.** Высокорелевантный поиск и автоматическая категоризация ресурсов Интернета [Текст] / С.В. Свечников // Сб. Интернет-порталы: содержание и технологии. Вып. 4 / ФГУ ГНИИ ИТТ «Информика». –М.: Просвещение, 2007. –С. 538–548.

12. **Поляков, П.Ю.** RCO на РОМИП 2006 [Текст] / П.Ю. Поляков, В.В. Плешко // Тр. IV семинара по оценке методов информационного поиска. –СПб.: НИИ Химии СПбГУ, 2003. –С. 72–79.

13. **Попов, А.** Поиск в Интернете – внутри и снаружи [Электронный ресурс] / А. Попов. –Режим доступа: http://www.citforum.ru/pp/search_03.shtml

14. **Тихомиров, И.А.** Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке [Текст] / И.А. Тихомиров, И.В. Соченков // Вестник Новосибирского гос. ун-та. Сер. Информационные технологии. –Новосибирск: НГУ, 2008. –Вып. 2. –Т. 6. –С. 94–100.

15. **Han, E.** Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification [Text] / E. Han, G. Karypis, V. Kumar // Proc. of the 16th International Conf. on Machine Learning. –Denver, 1999. –P. 41–56.