УДК 621.39

С.В. Свечников

ТЕМАТИЧЕСКАЯ КАТЕГОРИЗАЦИЯ ИНТЕРНЕТ-РЕСУРСОВ В СЕТИ ИНТЕРНЕТ

Предложен подход для создания алгоритмов поиска, анализа и категоризации интернет-ресурсов. Представлено решение задач индексации и автоматической категоризации веб-сайтов интернета за счет выделения терминов и присвоения им весовых коэффициентов, что позволяет достаточно быстро и эффективно оценить контент интернет-ресурса.

Введение. В настоящее время в России существует необходимость в применении систем тематической категоризации, а вместе с ними и систем для управления доступом к интернет-ресурсам. Открытое информационное пространство содержит большое количество сайтов различного содержания, и наряду с полезной информацией интернет содержит ресурсы, объективно опасные для нравственного здоровья общества, оказывающие негативное воздействие, в первую очередь на подрастающее поколение. Российский сегмент сети – один из самых быстроразвивающихся, количество пользователей интернета в России, по различным данным, достигает 25 миллионов человек, из них 2 миллиона детей [1]. 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 производительность корпоративной сети [2]. При этом методы прямого регулирования (цензуры) неэффективны, встречают протест пользователей и юридически несостоятельны, поскольку противоречат естественным правам граждан на свободу высказываний и волеизъявления.

В связи с этим решение этой проблемы надо искать не в цензуре, а в предоставлении инструмента для защиты от нежелательного контента, который пользователи могут использовать по своей воле и по своему усмотрению [3]. Таким инструментом является разрабатываемая система тематической категоризации интернет-ресурсов.

Реализация системы тематической категоризации интернет-ресурсов предполагает решение следующих задач:

– индексация интернет-ресурсов (преобразование интернет-ресурсов к единому формату);

– автоматическая категоризация интернет-ресурсов, обучение системы и отнесение текстовой информации к заранее определенной категории;

– оценка качества категоризации с использованием метрик из информационного поиска.

Представленные подзадачи связаны, в первую очередь, с анализом текстовой информации веб-страницы, т.е. с ее содержанием (контентом).

Пусть дано множество интернет-ресурсов D, разделенное на два непересекающихся подмножества T_r и T_s, называемых обучающей и тестовой выборкой. На основании обучающей выборки строится классификатор категорий, а на тестовой выборке проверяется качество категоризации. Пусть также дано соответствие между интернет-ресурсами и некоторой категорией c в виде $\Phi : D \to \{0,1\}$, устанавливающее для каждого интернет-ресурса значение 1 в случае принадлежности интернет-ресурса категории и 0 – в противном случае [3, 4, 5, 6].

Необходимо построить, используя только информацию из обучающей выборки T_r, функцию $\Phi' : D \to \{0,1\}$, аппроксимирующую Φ, чтобы число ошибок E на тестовой выборке T_s было наименьшим:

$$E = \sum_{T_s} \left| \Phi - \Phi' \right| \to min . \qquad (1)$$

Пусть T – множество терминов, каким-либо образом выделенное из интернет-ресурсов категории c. Тогда интернет-ресурс можно представить в виде терминологического вектора:

$$d_j = (w_{1j}, ... w_{|T|j})^T , \qquad (2)$$

где w_{ij} - вес термина t_i в интернет-ресурсе d_j.

Описания каждой из категорий представим в виде векторов той же размерности, что и векторы интернет-ресурсов:

$$c = (c_1,...,c_{|T|})^T, \quad (3)$$

где c_i - вес термина t_i в описании категории c.

При таком подходе существуют два ограничения:

– нет дополнительной информации о категориях, к которым прикрепляются интернет-ресурсы;

– нет никакой внешней информации о интернет-ресурсе, кроме той информации, которая содержится в нем.

Каждый интернет-ресурс – это вектор, где номера терминов (слов) – его координаты, а веса терминов – значения координат, размерность вектора – это количество терминов, встречающихся в интернет-ресурсе. Так как учитываются все термины, векторы получаются большого размера, что затрудняет процесс индексации, поэтому необходимо уменьшить размерность вектора. Для уменьшения размерности вектора в качестве терминов используем не слова, а устойчивые словосочетания, не учитываем редкие слова, которые не несут полезной информации, не рассматриваем часто встречающиеся слова.

Процесс индексации представим следующим образом:

– очистка страницы;

– выделение терминов;

– исключение терминов, не несущих смысловой нагрузки;

– замена общих и специфичных терминов;

– присваивание терминам весовых коэффициентов.

Сначала проводится очистка страницы интернет-ресурса, т.е. удаляется навигационная часть, теги html, скрипты, стоп-слова – частотные слова языка, не несущие смысловой нагрузки (предлоги, союзы, частицы, местоимения, некоторые глаголы), за счет этого уменьшается объем поисковой базы и повышается производительность поиска [7, 8, 9]. После этого в тексте с помощью функции анализа контента интернет-ресурса выделяются термины – логические выражения, состоящие из слов и словосочетаний, связанные операторами AND, OR, NOT. Для исключения терминов, не несущих смысловой нагрузки, используется пометка «исключение», которая показывает, что термин не относится к теме. Оставшиеся термины могут также обладать недостатками: существуют термины, которые слишком специфичны или, наоборот, - значение которых слишком обще, поэтому их необходимо заменить более подходящими. Это увеличивает полноту индексирования. Для замены специфичных терминов используется тезаурус RCO, который представляет собой словарь общей лексики с семантическими отношениями между словами [10]. Использование тезауруса повышает качество анализа текста и полноту поиска информации, позволяя расширять запрос синонимичными, более общими и более частными понятиями. Общие термины заменяются сочетаниями терминов или несколькими связанными терминами, имеющими более определенное значение. После того как были определены термины, необходимо провести лемматизацию – приведение терминов к нормальной форме (мужской род, единственное число). Тем самым уменьшается словарь терминов и повышается скорость работы индексации.

Заключительным этапом является присвоение терминам весовых коэффициентов. Исходное представление интернет-ресурса выглядит следующим образом: интернет-ресурс = коллекция слов (терминов) T. Каждый термин $t_i \in T$ имеет определенный вес w_{ij} по отношению к интернет-ресурсу $d_j \in D$, т.е. встречаемость этого слова на странице интернет-ресурса. Порядок слов учитывать не будем. На основании этих признаков каждому слову сопоставляется его вес.

Таким образом, каждый ресурс можно представить в виде вектора весов его терминов $d_j = \{w_{1j},...,w_{|T|j}\}$. Веса документов нормируем так, чтобы $w_{ij} \geq 0$ и $w_{ij} \leq 1$, где $i \in (0,|T|)$ и $j \in (0,|D|)$.

Для вычисления веса термина на странице интернет-ресурса используем классический частотный метод вычисления степени соответствия интернет-ресурса, так как этот метод относительно прост и имеет несложный алгоритм, что принципиально при обработке больших объемов документов.

Вычисляем вес термина следующим образом:

$$w_{ij} = tf_{ij} \cdot \frac{1}{df_j}, \quad (4)$$

где tf_{ij} (частота термина) - это отношение числа терминов t_i в интернет-ресурсе d_j к общему количеству терминов в этом интернет-ресурсе, таким образом, оценивается важность термина t_i в пределах одного интернет-ресурса:

$$tf_{ij} = \frac{T_{ij}}{|T_i|}, \quad (5)$$

где $j = 1, ..., T$, $i = 1, ..., D$,

T_{ij} - число терминов t_i в интернет-ресурсе d_j,

T_i - общее число терминов в интернет-ресурсе d_j.

df_j (частота интернет-ресурса) – это отношение количества интернет-ресурсов категории, в которых встретился термин t_i, к общему количеству интернет-ресурсов категории:

$$df_j = \frac{|D_j|}{|D|}, \quad (6)$$

где $j = 1, ..., T$,

D_j - число интернет-ресурсов, в которых встретился термин t_i,

D - общее количество интернет-ресурсов категории.

Таким образом, чем чаще термин встречается на странице интернет-ресурса, но реже встречается во всех интернет-ресурсах, тем выше будет его вес в данном интернет-ресурсе.

Наиболее трудоемкой частью реализации системы является разработка процесса, отвечающего за автоматическую категоризацию интернет-ресурсов, обучение системы на уже категоризированных интернет-ресурсах и определение соответствия категории.

Алгоритм автоматической категоризации интернет-ресурсов заключается в следующем:

– вычисляется мера близости страницы интернет-ресурса и категории – степень соответствия ресурса категории;

– для каждой страницы выбирается категория, наиболее близкая к ресурсу;

– в случае если значение степени соответствия ресурса превышает некоторое пороговое значение категории, ресурс добавляется в категорию;

– в случае если значение степени соответствия ресурса не превысило порогового значения категории, ресурс не добавляется в категорию и решение о принадлежности его к категории определяет эксперт.

Степень соответствия (*CSV*) между категорией *c* и интернет-ресурсом d_j определяем как скалярное произведение между их векторными представлениями:

$$CSV(c, d_j) = c \cdot d_j = \sum_i c_i d_{ij}. \quad (7)$$

Будем принимать решение о принадлежности интернет-ресурса к категории, если степень соответствия достигнет заданного порога τ. Таким образом, получаем:

$$\Phi'(c, d_j) = \begin{cases} 1, & CSV(c, d_j) \geq \tau \\ 0, & CSV(c, d_j) < \tau \end{cases}. \quad (8)$$

После того как вычислены степень соответствия между категорией и интернет-ресурсом, а также пороговое значение категории, необходимо провести обучение. Цель обучения - настройка весовых коэффициентов и порогового значения таким образом, чтобы процедура категоризации относила положительные примеры к категории, а отрицательные примеры - не относила, т.е. чтобы суммы весовых коэффициентов всех положительных примеров были равны либо превышали пороговое значение, а суммы для отрицательных примеров были ниже порога.

Для оценки качества категоризации интернет-ресурсов применяем метрики из информационного поиска, такие как полнота, точность, F-мера [11].

Пусть D_r – множество интернет-ресурсов, категоризированных экспертами, а D_a – множество интернет-ресурсов, категоризированных автоматически.

Полнота категоризации интернет-ресурсов по категории вычисляется как отношение количества правильно категоризированных интернет-ресурсов системой к общему числу интернет-ресурсов, относящихся к этой категории:

$$r = \frac{|D_a \cap D_r|}{|D_r|}. \quad (9)$$

Точность категоризации интернет-ресурсов по категории вычисляется как отношение количества правильно категоризированных интернет-ресурсов системой к общему числу интернет-ресурсов, автоматически категоризированных системой:

$$p = \frac{|D_a \cap D_r|}{|D_a|}. \quad (10)$$

Для идеального алгоритма полнота и точность должны быть равны 100%.

F-мера, т.е. сводная оценка качества категоризации, определяется как гармонически среднее полноты и точности:

$$F = \frac{2 \cdot r \cdot p}{r + p}. \quad (11)$$

Алгоритм осуществления сбора и обработки данных интернет-ресурсов выглядит следующим образом (рисунок 1).

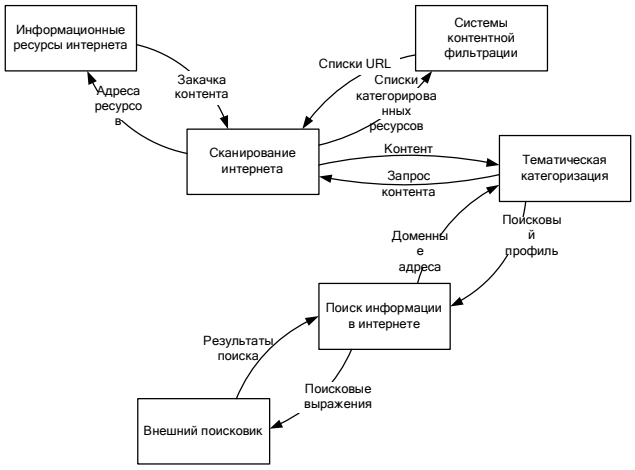

Рисунок 1 – Сбор и обработка данных интернет-ресурсов

Блок поиска информации в интернете получает от блока тематической категоризации поисковые профили и преобразует их в поисковые выражения для внешних поисковиков. Внешний поисковик передает результаты поиска обратно, далее выделяются доменные адреса информационных ресурсов и передаются блоку тематической категоризации.

Блок тематической категоризации запрашивает контент информационных ресурсов, адреса которых были добавлены в базу, но еще не были категорированы, или те, для которых пришло время повторной категоризации.

Блок сканирования интернета получает от систем контентной фильтрации списки доменных адресов для категоризации и передает обратно списки категорированных ресурсов.

Блок сканирования интернета по полученным адресам скачивает контент информационных ресурсов интернета и возвращает этот контент для проведения категоризации.

Для решения задач обработки информации интернет-ресурсов используются следующие два алгоритма:

1. Поступление и обработка нового ресурса.

2. Актуализация существующих интернет-ресурсов.

Первый алгоритм работает по следующему сценарию: находится новый ресурс, проверяется, есть ли он в каталоге, после чего проводятся обход сайта (загрузка некоторого количества страниц), анализ этих страниц и присвоение категории сайту; если этого количества страниц недостаточно для присвоения категории, проводится расширенный обход сайта (загружается большее количество страниц).

Актуализация существующих интернет-ресурсов проходит следующим образом: проводится новый обход сайта, проверяется, изменилась ли страница с момента последнего обхода, если изменилась, то проводится классификация загруженных страниц, если их достаточно, проводится классификация сайта, одновременно с этим уточняется принадлежность категориям.

Основная структура разрабатываемой системы тематической категоризации интернет-ресурсов и взаимодействие ее подсистем выглядят следующим образом (рисунок 2):

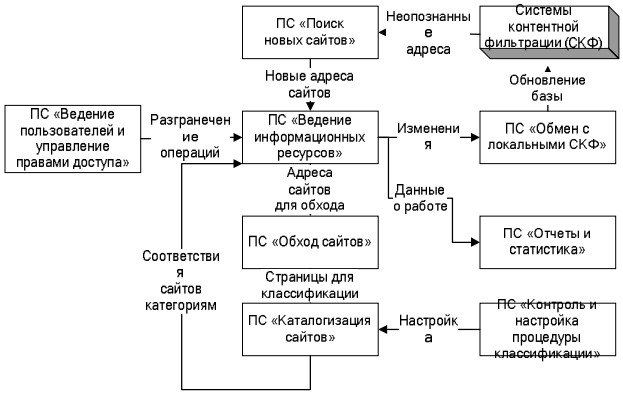

Рисунок 2 – Структура системы и взаимодействие ее подсистем

Подсистема «Поиск новых сайтов» предназначена для поиска новых интернет-ресурсов. Результатом ее деятельности является набор новых адресов сайтов, пополняющих базу тематической категоризации. На этом этапе информация о ресурсе является базовой, т.е. не содержит данных о категориях, к которым отнесен ресурс.

Далее вся информация о новых ресурсах поступает в подсистему «Ведение информационных ресурсов», где в результате классификации сайты получают соответствие категориям.

После этого в работу включается подсистема «Обход сайтов», в рамках которой осуществляются обход сайта и получение набора страниц для анализа тематики.

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

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

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

Подсистема «Ведение пользователей и управление правами доступа» позволяет использовать систему в многопользовательском режиме, с разграничением прав доступа между поль-

зователями и контролем действий, выполняемых пользователем.

Подсистема «Отчеты и статистика» собирает информацию от всех модулей и предоставляет ее для анализа.

Заключение. В статье предложены средства тематической категоризации интернет-ресурсов, которые позволяют:

– составлять тематический каталог интернет-ресурсов за счет поиска веб-сайтов и поступления их от локальных систем контентной фильтрации при посещении пользователями вебстраниц;

– обеспечивать высокую точность категоризации интернет-ресурсов за счет составления тематических профилей при описании категорий;

– проводить актуализацию существующих интернет-ресурсов, т.е. заново осуществлять обход сайта, проверять изменения страниц, уточнять принадлежность сайта категориям;

– осуществлять обмен данными с локальными системами контентной фильтрации для защиты от нежелательного контента.

Библиографический список

1. Фонд «Общественное мнение», http://www.fom.ru/.

2. *Абсалямов А.* Борьба с киберслэкингом. Windows 2000 Magazine, №3. 2000.

3. *Плешко В.В., Ермаков А.Е., Голенков В.П.* RCO на РОМИП 2004 // Российский семинар по оценке методов информационного поиска (РОМИП 2004) – Пущино. 2004. — С. 43-61

4. *Плешко В.В., Ермаков А.Е., Митюхин В.А.* RCO на РОМИП 2003: отчет об участии в семинаре по оценке методов информационного поиска // Труды первого российского семинара по оценке методов информационного поиска; под ред. И.С. Некрестьянова. – Санкт-Петербург: НИИ химии; СпбГУ. – 2003. – С. 42-51.

5. *Поляков И.Е.* Опыт создания системы фильтрации агрессивного web-контента // Труды XII всероссийской научно-методической конференции «Телематика 2005», 6-9 июня 2005 г. Издательство СПб.

6. *Sebastiani F.* Machine Learning in Automated Text Categorization, http://nmis.isti.cnr.it/sebastiani/.

7. *Некрестьянов И.С., Павлова Е.Ю.* Обнаружение структурного подобия HTML-документов // Труды четвертой всероссийской конференции RCDL'2002, 38-54, Дубна, Россия, 2002.

8. *Ziv Bar-Yossef, Sridhar Rajagopalan.* Template Detection via Data Mining and its Applications // In Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA.

9. *Gupta S., Kaiser G., Grimm P., Chiang M.*, Starren J. Automating Content Extraction of HTML Documents // World Wide Web Journal, January 2005.

10. Russian Context Optimizer. Технологии анализа и поиска текстовой информации, http://www.rco.ru/.

11. *Поляков П.Ю., Плешко В.В.* RCO на РОМИП 2006 // Труды четвертого российского семинара по оценке методов информационного поиска. Санкт-Петербург: НИИ химии. СпбГУ. – 2003. – С. 72-79.