

ФЕДЕРАЛЬНОЕ АГЕНСТВО ПО ОБРАЗОВАНИЮ
Байкальский государственный университет экономики и права

Факультет экономической кибернетики
Кафедра информатики и кибернетики

К ЗАЩИТЕ ДОПУСКАЕТСЯ:

Заведующий кафедрой

_____ д. э. н., профессор, Амбросов Н. В.
(подпись) (ученая степень, звание, Ф. И. О.)

ДИПЛОМНЫЙ ПРОЕКТ

на тему: Каталог ресурсов интернет с элементами
искусственного интеллекта

Студент:	_____	<u>Витязев Ярослав Михайлович</u>
	(подпись)	(Ф. И. О.)
Руководитель:	_____	<u>Ступин Виталий Валерьевич</u>
	(подпись)	(Ф. И. О.)
Нормоконтролер:	_____	<u>Иньшина Наталья Дмитриевна</u>
	(подпись)	(Ф. И. О.)

Иркутск, 2007 г.

ЗАДАНИЕ
на дипломный проект

Студенту: Витязеву Ярославу Михайловичу.

Тема дипломного проекта: Создание каталога ресурсов интернет с элементами искусственного интеллекта.

Утверждена приказом по Байкальскому государственному университету экономики и права _____.

Руководитель дипломной работы канд. физ.-мат. наук, доцент В. В. Ступин.
/ученая степень, звание, ИОФ/

Срок сдачи студентом работы _____.

КАЛЕНДАРНЫЙ ПЛАН
о выполнении дипломного проекта

Этапы выполнения	План выполнения	Фактическое выполнение	Подпись руководителя
Обзор предметной области	15.01.2007	15.01.2007	
Проектирование каталога	20.01.2007	20.01.2007	
Разработка каталога	15.02.2007	15.02.2007	
Размещение каталога	18.03.2007	18.03.2007	
Завершение работ	01.06.2007	01.06.2007	

Исполнитель _____
(дата, подпись)

Витязев Ярослав Михайлович
(Ф. И. О.)

Оглавление

Введение.....	4
1. Каталоги ресурсов интернет.....	6
1.1.Определение каталога ресурсов интернет.....	6
1.2.Типичные проблемы реализации каталогов.....	9
1.3.Сравнительный анализ наиболее популярных каталогов.....	20
1.4.Использование элементов искусственного интеллекта для улучшения качества работы каталога.....	27
2. Проектирование каталога.....	31
2.1.Цели и задачи проектирования.....	31
2.2.Ограничения на ресурсы.....	31
2.3.Необходимые функции.....	31
2.4.Реализация функций каталога с использованием элементов искусственного интеллекта.....	38
2.5.Информационное обеспечение.....	43
2.6.Программное обеспечение.....	53
2.7.Техническое обеспечение.....	58
2.8.Интерфейсы каталога.....	59
3. Размещение каталога в сети Интернет и его продвижение.....	76
3.1.Перечень этапов внедрения.....	76
3.2.Обзор и выбор хостинговой компании.....	77
3.3.Обзор и выбор методов продвижения.....	79
3.4.Результаты эксплуатации каталога ресурсов интернет.....	84
Заключение.....	93
Список источников.....	94
Приложение 1: Статистические отчеты Google Analytics.....	96
Приложение 2: Таблицы для расчета экономической эффективности.....	101

Введение

Экспоненциальный рост объема информации, содержащейся в Интернете, является причиной все более и более возрастающей трудности поиска необходимых документов и организации их в виде структурированных по смыслу хранилищ. Начиная еще с докомпьютерной эпохи, в качестве эффективного средства смысловой организации массивов документов, обеспечивающего возможность удобного доступа к ним, используются иерархические каталоги. В настоящее время этот подход применяется разнообразными компьютерными системами поддержки поиска и доступа к документам. Вероятно, из всех типов таких систем, лидерами по количеству использующих их людей, являются веб-каталоги, такие как DMOZ, Yandex, Yahoo и Rambler. В качестве других примеров можно назвать рубрицированные хранилища патентов (например, Всемирной Организации Интеллектуальной Собственности WIPO) или разнообразные компьютерные библиотечные каталоги. Такие системы незаменимы для эффективного поиска и навигации в огромных массивах документов, однако поддержка их полноты, производимая главным образом вручную, становится все более трудоемкой в условиях взрывного роста числа документов [1].

Интеллектуализация созданных в настоящее время каталогов ресурсов интернет находится на зачаточном уровне, многие функции существующих в настоящее время каталогов даже не автоматизированы. Наиболее развитыми и полными каталогами в Интернете являются: Open Directory Project, каталог Yahoo, каталог Alexa. В русскоязычном интернете следует выделить следующие каталоги: каталог Яндекс, каталог-рейтинг Рамблер ТОП-100, каталог и рейтинг Mail.ru. Следует заметить, что только в русскоязычном Интернете количество каталогов ресурсов интернет превышает 1700 [2]. В основном, это частично автоматизированные неинтеллектуализированные тематические каталоги. Зачастую такие каталоги создаются для «обмана» поисковых систем с целью повышения рейтинга сайта в выдаче поисковой машины при помощи обмена ссылками.

Созданные в настоящее время каталоги не используют потенциал современных технологий, а ориентируются, как правило, на одном-двух направлениях: в некоторых каталогах содержится большое количество тщательно отобранных модераторами ресурсов, однако сама процедура регистрации и модерации реализована неэффективно, в других же каталогах проблемой является рубрикация и классификация ресурсов, в третьих – непродуманная система ранжирования, невозможность интеграции с внешними сервисами, отсутствие возможности синдикации контента, осуществления поиска. Все эти проблемы актуальны для каталогов, однако, ни в одном современном каталоге эти проблемы не решены полностью.

Данный проект посвящен разработке современного каталога ресурсов интернет, решающих не только эти проблемы, но и ряд других, не менее важных, выявленных в процессе работ над созданием каталога.

В рамках проекта рассмотрены существующие проблемы каталогов, сформированы критерии для оценки каталогов; осуществлен обзор наиболее популярных каталогов и проведен их анализ, сделаны выводы и определены ключевые проблемы, которые необходимо решить; сформирован проект каталога, на основе которого создана его программная реализация, успешно работающая в интернете, начиная с середины марта 2007 года; проведены мероприятия по продвижению каталога в сети интернет, на страницах каталога размещены рекламные материалы и произведена оценка экономической эффективности проекта. Уже сейчас каталог широко используется целевой аудиторией и составляет конкуренцию для ряда других каталогов.

1. Каталоги ресурсов интернет

1.1. Определение каталога ресурсов интернет

Каталог ресурсов интернет (англ. web directory) — структурированный набор ссылок на сайты с кратким их описанием. Сайты внутри каталога разбиваются по темам, а внутри тем могут быть ранжированы или по определенному индексу или по дате добавления, или по алфавиту, или по другому параметру. Это один из старейших сервисов Интернета [3].

Подавляющее большинство рейтингов посещаемости ресурсов имеют классификатор сайтов, но ранжирование при этом основано на посещаемости сайтов.

В зависимости от широты тематики ссылок каталоги могут быть общими и специализированными (тематическими). В зависимости от типа используемых ссылок, каталоги могут быть с прямыми ссылками и ссылками, установленными через т. н. «редирект». Также каталоги подразделяют на т. н. «белые» и «серые», в зависимости от требований к регистрации. По типу регистрации интернет-ресурсов выделяют модерируемые и немодерируемые каталоги. Выделяют и, так называемые, каталоги-рейтинги, в зависимости от используемых алгоритмов ранжирования интернет-ресурсов.

Как правило, каталоги ресурсов интернет строятся на основе клиент-серверной архитектуры. Использовать их по прямому назначению можно при помощи интернет-браузеров — специального программного обеспечения, в большинстве случаев, поставляемого вместе с операционной системой.

Таким образом, основа практически всех каталогов интернет-ресурсов — HTML-документы [15]. Для стилового оформления этих документов используются либо внутренние средства [X]HTML, либо каскадные таблицы стилей.

Однако, весьма трудно вручную поддерживать актуальное состояние каталогов (см. рис. 1), поэтому ручное редактирование исходных [X]HTML-текстов мало распространено.

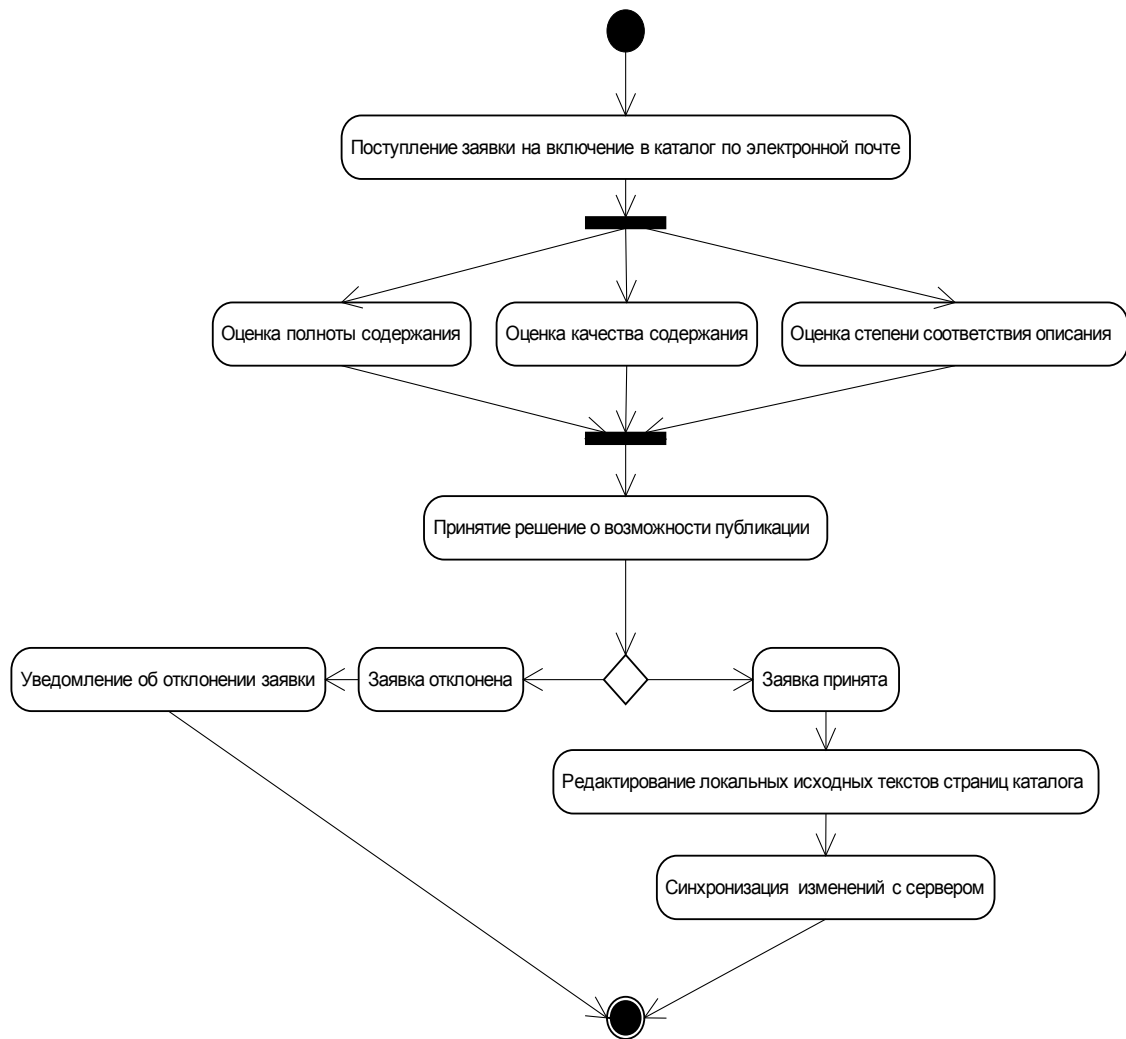


Рис. 1: Диаграмма деятельности (UML): добавление информации об интернет-ресурсе вручную

Для избежания данной проблемы, как правило, используют сценарные языки программирования для написания специальных динамических интерфейсов: интерфейса пользователя и интерфейса администратора каталога. При этом, любой пользователь может подать заявку на регистрацию интернет-ресурса в каталоге, после рассмотрения и одобрения которой интернет-ресурс будет опубликован в каталоге (см. рис. 2).

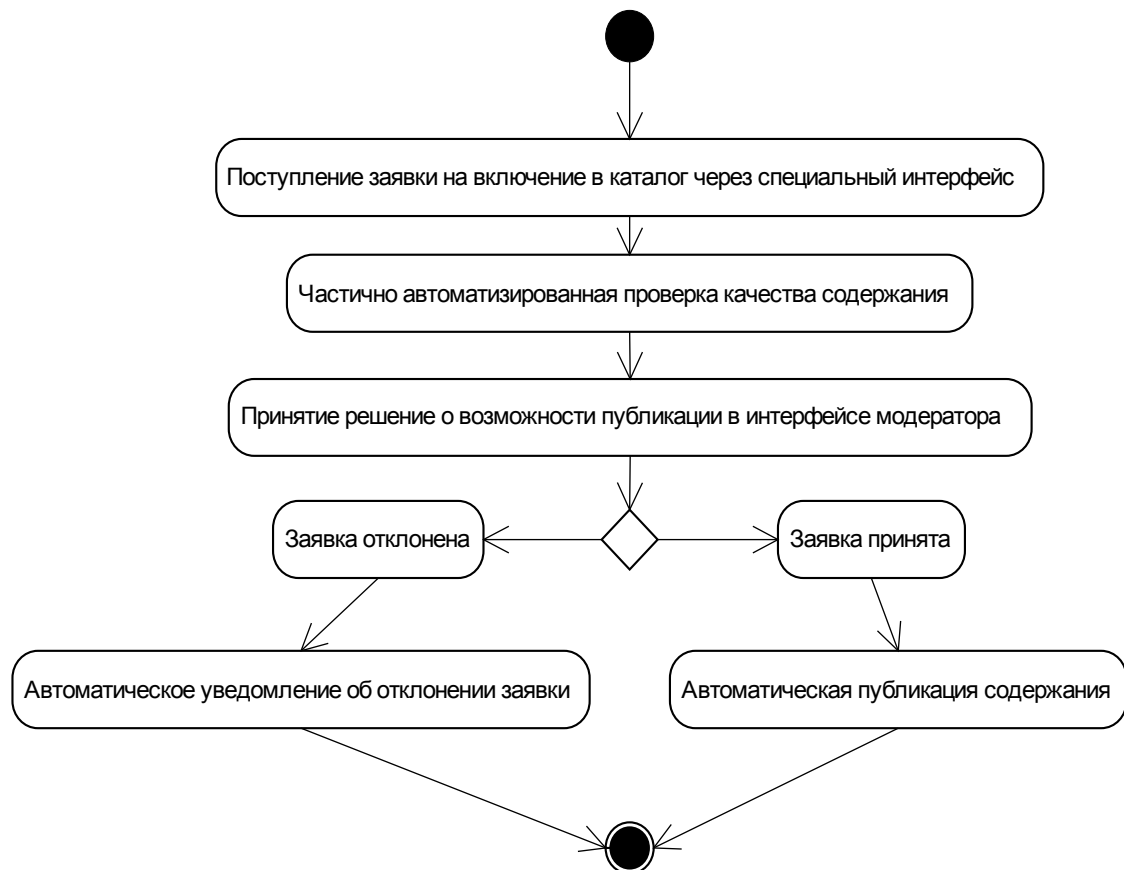


Рис. 2: Диаграмма деятельности (UML) в частично автоматизированной схеме

Также, для хранения данных о зарегистрированных сайтах и манипуляции ими в динамических каталогах ресурсов интернет используются системы управления базами данных. Как правило, используют СУБД MySQL [13] в виду ее широкого распространения и отсутствие ограничений на использование в коммерческих целях.

Для того, чтобы каталог ресурсов интернет был доступен для конечных пользователей, необходимо:

- 1) наличие зарегистрированного доменного имени;
- 2) наличие компьютера (сервера), подключенного к глобальной сети Интернет;
- 3) наличие двух IP-адресов в различных подсетях класса С.

В большинстве случаев, для организации каталогов ресурсов интернет используют услуги хостинг-провайдеров.

1.2. Типичные проблемы реализации каталогов

Существует ряд типичных для каталогов ресурсов интернет проблем, они будут описаны ниже. Некоторые из этих проблем подробно описаны в [4].

1.2.1. Нежелательное содержание и спам

Одной из актуальных и наиболее важных проблем является возможность публикации нежелательного содержания в каталоге. Данная проблема выявляется различно, в зависимости от типа каталога. В каталогах с предварительной модерацией редакторам приходится ежедневно обрабатывать такие заявки вручную. В каталогах с пост-модерацией или в немодерируемых каталогах появляется задача удаления такого содержания после публикации. Причем, нежелательный материал, в случае пост-модерации, может быть опубликован, соответственно, рядовой пользователь каталога может увидеть такой материал, что крайне нежелательно.

Обычно к нежелательному содержанию относят следующую информацию:

- большое количество ненормативной лексики;
- ресурсы, содержащие жестокость, расовую нетерпимость или пропаганду действий против личности, группы или организации.;
- о взломе компьютерных систем;
- о наркотиках и их атрибутике;
- порнографию и информацию только для взрослых;
- азартные игры или информация, относящаяся к казино;
- чрезмерное количество рекламы на сайте;
- любая информация, пропагандирующая нелегальную деятельность или нарушающая права других людей;
- всплывающие окна переднего плана, всплывающие окна заднего фона или окна, открывающиеся при уходе со страницы, которые мешают навигации по сайту, меняют настройки пользователя или предназначены для загрузки;
- наличие чрезмерного количества часто повторяющихся и несоответствующих

- щих содержанию страницы ключевых слов на самой странице или в коде страницы;
- содержание, вводящее пользователя в заблуждение или манипулирующее им, а также структура, позволяющая увеличить рейтинг страницы при поиске в каталоге;
- продажа или реклама определенных видов оружия, например, огнестрельного оружия, боеприпасов, складных ножей и кастетов;
- продажа или реклама пива или крепких алкогольных напитков;
- продажа или реклама табака или табачных изделий;
- продажа или реклама лекарств, отпускаемых по рецепту;
- продажа или продвижение товаров, являющихся копией или имитацией изделий от дизайнеров и т. п.

В общем случае, определение недоброкачественного содержания сводится к выявлению тематики, противоречащей политике каталогов. Как правило, в роли цензора выступает модератор.

В настоящее время разработан ряд методов для борьбы с публикацией недоброкачественного содержания. Эти методы подразделяют на ручные (обработка заявок модератором), частично-автоматизированные (когда используются модули, позволяющие автоматизировать часть операций и выявить недоброкачественное содержание) и полностью автоматизированные (когда ИС работает в экспертном режиме, и определяет недоброкачественное содержание: удаляет его или помечает специальным маркером самостоятельно, без привлечения модератора).

1.2.2. Некачественное и неполное описание

Существует и сходная с предыдущей проблема: проблема некачественного описания интернет-ресурса. В том случае, если описательную информацию об интернет-ресурсе предоставляет рядовой пользователь каталога, отправленное описание следует проверять. При этом, можно оценивать ряд факторов, например: длину описания, соотношение букв к знакам пунктуации (так можно отличить набор ключевых слов от осмысленного описания), соотношение

заглавных букв к строчным (большое количество рекламных описаний содержит текст с большим количеством заглавных букв) и т. п.

1.2.3. Проблема дублирования описания ресурса

Документ (один и тот же) в интернете, в идеале, должен иметь один уникальный адрес. Однако, на практике, это далеко не так. Существует множество документов на различных интернет-серверах с практически идентичным содержанием. При появлении скриптовых языков программирования и различных технологий для серверной стороны интернет-приложений, данная проблема стала наиболее острой. Достаточно просто породить неограниченное количество дублирующихся документов. С этим связана одна из основных проблем информационно-поисковых систем: проблема определения нечетких дублей документов.

1.2.4. Процедура модерации интернет-ресурса

Модерация (применительно к каталогам ресурсов интернет) — это процедура одобрения к публикации интернет-ресурса, включающая в себя процессы определения доступности документа, его соответствия описательной информации, проверки на недоброкачественное содержание, публикации или отклонения заявки на публикацию интернет-ресурса в каталоге.

Модерация требует значительных усилий со стороны персонала каталога [8]. Зачастую процесс модерации ресурсов в каталоге неавтоматизирован, несмотря на то, что практически все функции модерации могут подлежать автоматизации, а некоторые из них могут быть интеллектуализированы и выполняться без привлечения модератора (см. рис. 3).

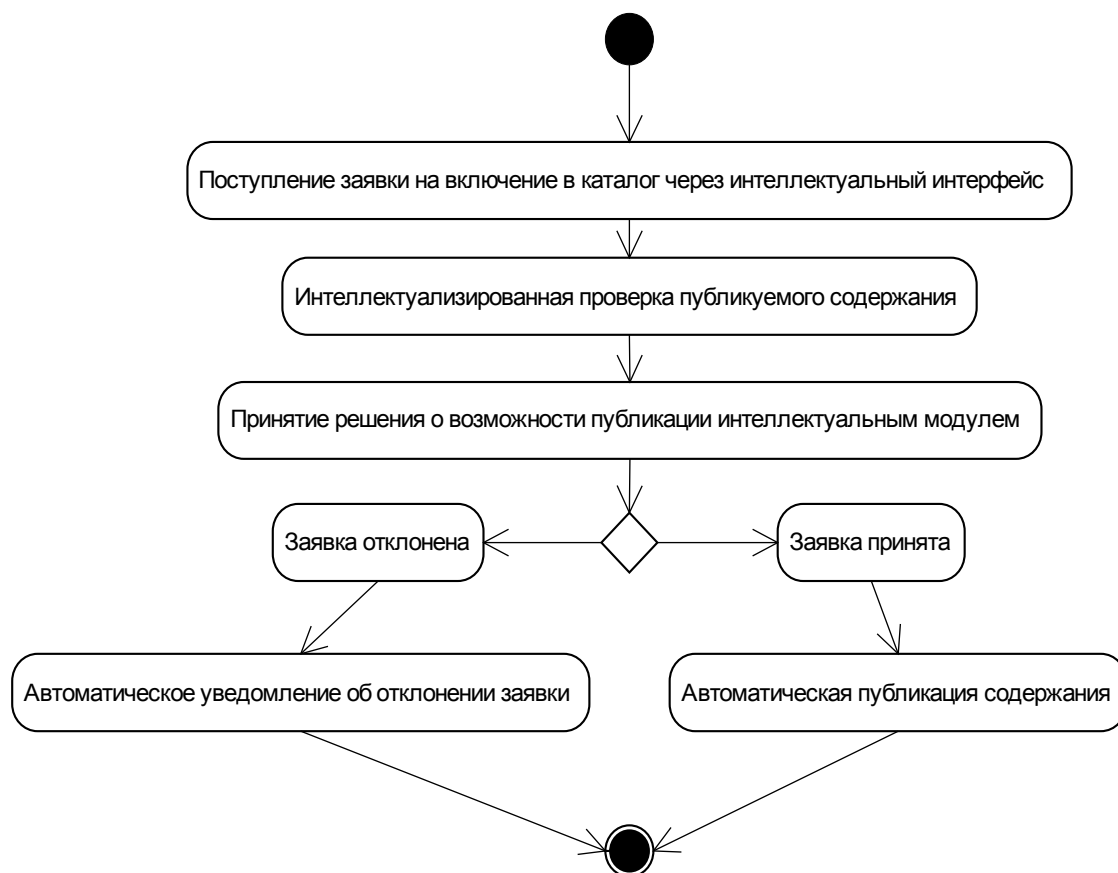


Рис. 3: Модерация интернет-ресурса

1.2.5. Проблемы рубрикации и классификации

Одной из основных проблем в каталогах ресурсов интернет является проблема рубрикации интернет-ресурсов. Существует ряд подходов, которые могут быть применены для решения данной проблемы.

Классическим подходом является таксономическая рубрикация. Такой вид рубрикации подразумевает наличие, заранее заполненного административным персоналом каталога, перечня разделов. Как правило, таксономический рубрикатор является иерархическим, и на практике не содержит более трех уровней вложенности. Таксономический подход ограничивает свободу пользователя, регистрирующего ресурс, т. к. список разделов задан заранее, и редактировать его не представляется возможным. Также, при таком подходе может возникнуть проблема отнесения интернет-ресурса к двум различным релевантным разделам (обычно в каталогах допускается для одного ресурса использовать только один, наиболее подходящий, раздел).

Одним из современных подходов, получившим широкое распростране-

ние является фолксономический подход, позволяющий осуществлять рубрикацию при помощи тегов. Тег — это специальный маркер, ассоциированный с ресурсом, указываемый пользователем в свободной форме. Фолксономическая рубрикация практически не накладывает никаких ограничений на пользователя, поэтому она получила высокую популярность и используется на многих интернет-ресурсах нового поколения. Однако, существуют проблемы, связанные с фолксономическим подходом, для которых на данный момент не существует формальных методов решения: одной из таких проблем является указание нерелевантных тегов — пользователь может указать тег, который не связан с действительным содержанием интернет-ресурса (однако, это можно отследить, например, при помощи эвристических алгоритмов). Другой проблемой является указание смежных тегов для одних и тех же смысловых значений, например, тегов в разных формах: «ИТ», «информационные технологии», «информационная технология», «IT» — все эти теги должны ссылаться на один и тот же элемент фолксономической структуры, а определить связь между ними типичными алгоритмами не представляется возможным, такую проблему нужно так же решать эвристическими способами.

Также, существуют и другие методы классификации интернет-документов, в том числе и малопригодные для интернет-ресурсов методы, такие как УДК, ББК и т. п.

Ряд интернет-каталогов использует и другие, не получившие широкое распространение методы классификации, например, фасетную классификацию (используется в каталоге Яндекса) — такой подход требует значительных ресурсов для указания фасетных признаков и его использование затруднено.

1.2.6. Проблема поддержки актуального состояния описания

Одной из важных проблем является проблема поддержки актуального состояния описания интернет-ресурса. Как правило, каталоги ресурсов интернет не предоставляют возможности пользователям актуализировать состояние описания (информацию о ресурсе можно добавить, а редактировать нельзя). Однако, многие интернет-ресурсы являются динамичными и их содержание со

временем изменяется — эти изменения должны быть отражены и в каталоге. Реализация такого интерфейса не является сложной (рис. 4).

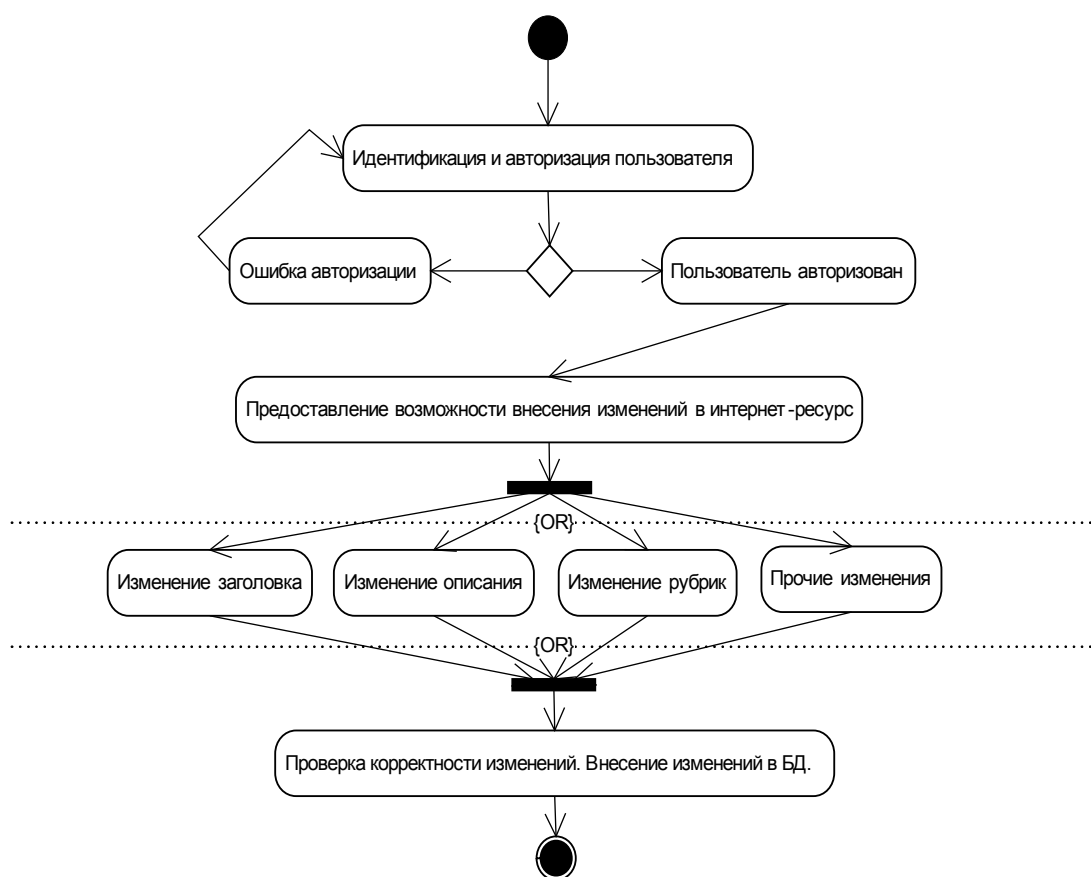


Рис. 4: Поддержка актуального состояния описания интернет-ресурса

1.2.7. Проблема автоматических и частично-автоматизированных регистраций интернет-ресурсов в каталоге

В последнее время получили развитие сервисы, предоставляющие возможность в полуавтоматическом режиме, за короткий промежуток времени, зарегистрировать сайт в нескольких тысячах каталогов ресурсов интернет. Такие сервисы являются популярными среди специалистов по поисковой оптимизации интернет-ресурсов (SEO). SEO используют их для повышения рейтингов ссылочного ранжирования поисковых систем (как известно, практически все современные алгоритмы поисковых систем используют ссылочное ранжирование как один из основных компонентов для упорядочивания интернет-ресурсов в поисковой выдаче). При этом, зачастую, при помощи таких сервисов происходит спам каталогов ключевыми словами, появление которого в каталоге крайне нежелательно.

Одним из эффективных методов для борьбы с подобными сервисами и прочими нежелательными публикациями является установка на формы регистрации теста Тьюринга (CAPTCHA, рис. 5): этот тест позволяет достаточно качественно (на данном уровне развития НТП) определять человек заполняет форму или «компьютерный робот» [5].

Но, все-таки, существует ряд алгоритмов, позволяющих «взломать» такую защиту: некоторые простые CAPTCHA поддаются распознаванию при помощи примитивных эвристических алгоритмов; также, возможен обход такой защиты способом, известным как «использование порно-сайта для взлома CAPTCHA» (взломщик создает страницу в интернете, в которую импортирует изображение, которое нужно распознать для осуществления регистрации, и на этой же странице размещает надпись вида «для доступа к порнографическому содержанию вам необходимо ввести текст с картинки» и затем уже использует, распознанный другим человеком текст, для прохождения теста — такой способ взлома показал достаточно высокую эффективность).



Рис. 5: Пример CAPTCHA

1.2.8. Проблема авторства

Существует и проблема определения авторства интернет-ресурса. Большинство каталогов не позволяют выявлять автора интернет-ресурса, в связи с этим может случиться следующее: пользователь (возможно конкурент) регистрирует интернет-ресурс с некоторым содержанием (зачастую недоброкачественным или неинформативным), а реальный владелец сайта не может изменить его, т. к. доступ к ресурсу в каталоге «захвачен» другим пользователем.

Существует метод, позволяющий разрешить эту проблему. Пользователю, претендующему на авторство, предлагается на сервере сайта разместить файл с уникальным именем (или же на странице сайта прописать уникальный

мета-тег): известно, что данную операцию может сделать только владелец сайта. Интерфейс каталога делает запрос к файлу или странице с мета тегом и проверяет достоверность. В случае подтверждения достоверности размещения действительному владельцу возвращаются права на изменение содержания. Такой подход применяется в ИПС «Google», toodoo.ru и ряде других авторитетных проектов.

1.2.9. Проблема доступности данных

Пользователям каталога следует отображать только доступные интернет-ресурсы (т. е. в каталоге должны отсутствовать т. н. «битые» гиперссылки на зарегистрированные сайты).

Известно, что в случае доступности интернет-ресурса, веб-сервер возвращает код ответа, начинающийся на цифру «2», как правило, это ответ «HTTP/1.0 200 Ok». Можно использовать этот факт для определения доступности интернет-ресурса в автоматическом режиме (рис. 6).

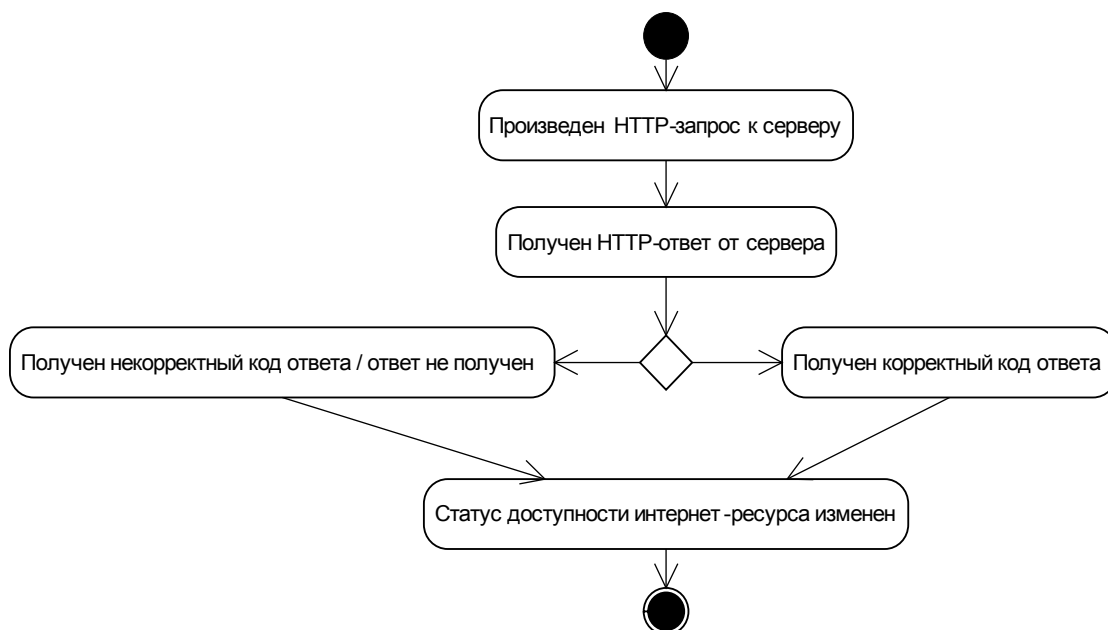


Рис. 6: Автоматизированный процесс проверки доступности интернет-документа

1.2.10. Проблема представления данных

Представление информации в браузере в понятном для человека виде — лишь один из немногих способов представления информации об интернет-ре-

сурсе. Классические интернет-ресурсы ограничивались лишь этим способом: информация была представлена лишь в виде HTML-файлов с заданным стилевым оформлением.

Современное понимание семантического веба не позволяет ограничиваться лишь таким представлением. Консорциумом W3 и DCMi был создан ряд стандартов (см. рис. 7) представления данных, которые должны использоваться в интернет-ресурсах для осуществления дополнительных возможностей [17]. К таким возможностям относятся:

1) агрегация и синдикация — возможность использовать информацию, предоставляемую интернет-ресурсом на других интернет-ресурсах и других приложениях (широкое распространение получили так называемые новостные RSS-ленты);

2) получение метаданных — возможность третьим программным средствам и ИПС получать информацию о документе, его семантических и прочих связях с другими документами и предметными областями;

3) получение результатов поиска — возможность для современных ИПС получать поисковую выдачу по определенным запросам в удобном для синдикации формате (такой подход позволяет строить ИПС с более простыми интерфейсами, и, наиболее просто, интегрировать поисковые функции в приложения).

На практике данные технологии применяются редко (ввиду сложности реализации и низкой на данном этапе развития НТП экономической отдачи от реализации такой функциональности).

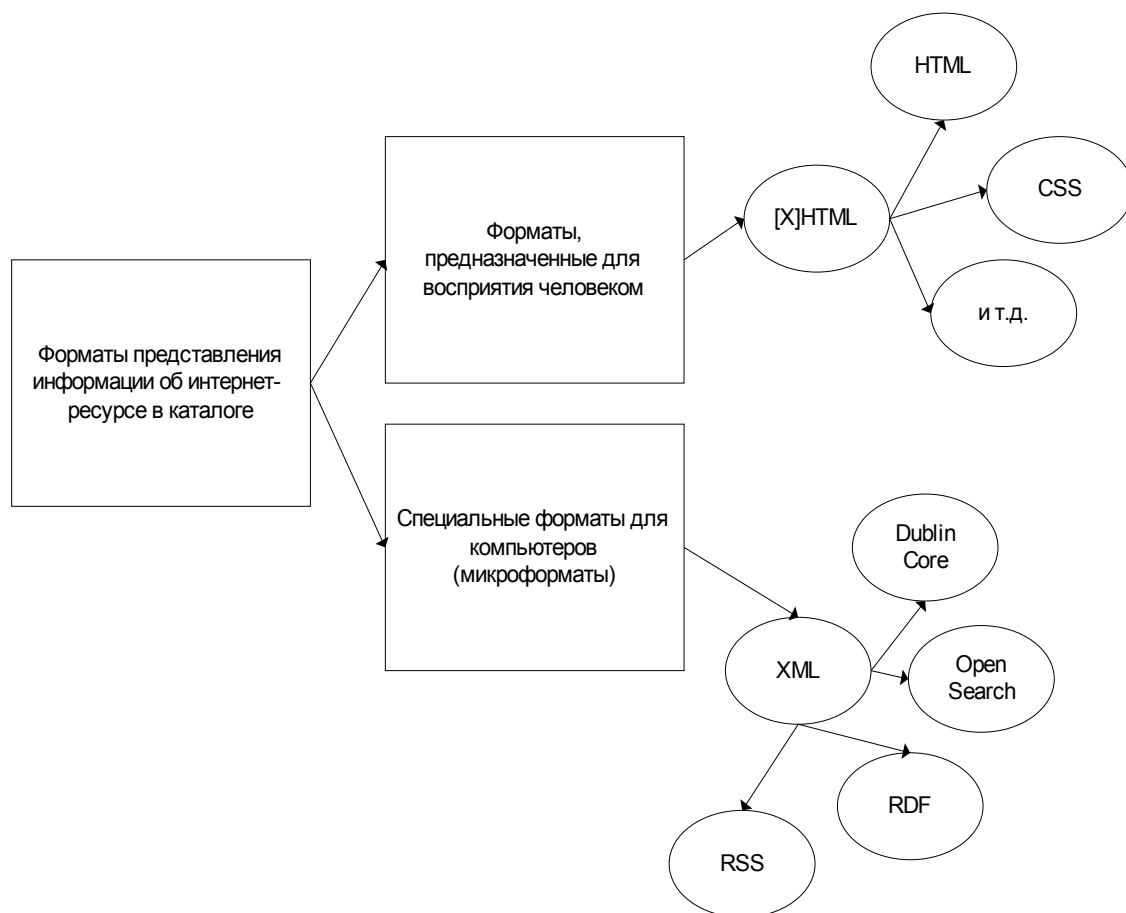


Рис. 7: Разновидности форматов представления информации интернет-ресурса

1.2.11. Востребованность каталогов

Каталоги ресурсов интернет незаменимы при поиске информации в интернет, когда у пользователя нечетко сформирована цель поиска. Каталоги позволяют легко найти смежные ресурсы и ресурсы с некоторыми связями (например, фасетными связями: регион, сектор экономики, тип ресурса и т. п.) В данный момент существует большое количество интернет-ресурсов, однако качественных каталогов среди них единицы.

В современном интернете существует тенденция к созданию семантического интернета. Каталоги ресурсов в нем будут одним из неотъемлемых связующих звеньев [6,7].

Семантическая паутина (англ. Semantic web) — новая концепция развития Всемирной паутины и сети Интернет, принятая и продвигаемая Консорциумом Всемирной паутины. Иногда также упоминается как семантический веб.

Семантическая паутина — это надстройка над существующей Всемир-

ной паутиной, которая призвана сделать размещённую в сети информацию более понятной для компьютеров. Известно, что почти вся информация в Интернете находится в текстовой форме. Не секрет, также, что прогресс в области обработки человеческих языков (англ. Natural Language Processing) идёт очень медленно. Компьютеры не могут воспринять и осмыслить словесную информацию, размещённую в Интернете, и в ближайшее время, видимо, не смогут. Тогда встаёт вопрос — как же заставить компьютеры понимать смысл размещённой в сети информации и научить компьютеры пользоваться ею? На этот вопрос и призвана ответить концепция семантической паутины. Слово «семантическая» в данном случае означает «осмысленная», «понятная».

В настоящее время компьютеры принимают довольно ограниченное участие в формировании и обработке информации в сети Интернет. Функции компьютеров, в основном, сводятся к хранению, отображению и поиску информации. В то же время создание информации, её оценку, классификацию и актуализацию — всё это по-прежнему выполняет человек. Как включить компьютер в эти процессы? Если компьютер пока нельзя научить понимать человеческий язык, то нужно использовать язык, который был бы понятен компьютеру. То есть, в идеальном варианте вся информация в Интернете должна размещаться на двух языках: на человеческом языке для человека и на компьютерном языке для понимания компьютера. Семантическая паутина — это концепция сети, в которой каждый ресурс на человеческом языке был бы снабжен описанием, понятным компьютеру. Каталог ресурсов интернет легко снабдить таким описанием, используя данные об интернет-ресурсах в специально формализованной модели метаданных.

1.3. Сравнительный анализ наиболее популярных каталогов

1.3.1. Формирование критериев оценки

Исходя из существующих видов каталогов и применяемых в них подходов и технологий, а также проблем каталогов был сформирован перечень критериев для оценки каталогов ресурсов интернет (см. табл. 1).

«Полнота» является одним из важных критериев. Определяется путем просмотра разделов таксономического справочника каталога, и, если хотя бы в одном разделе 1-2 уровня опубликовано менее 2 сайтов, каталог считается недостаточно полным, в обратном случае — полным.

Метод регистрации ресурсов определяется существованием модерации. Каталог может быть модерлируемым (когда ресурс проверяется перед публикацией) или немодерируемым (в обратном случае). Данный критерий также важен: в случае немодерируемого каталога, содержание обычно не является качественным, и содержит, в основном, спам.

Количество таксономических разделов и связанная с ним характеристика уровень таксономических разделов являются также важными критериями. Количество разделов должно быть небольшим, ровно как и уровень их вложенности. Для оценки можно использовать любой из этих критериев. При большом уровне вложенности трудно найти необходимый ресурс в каталоге.

Многоязычность определяет (ограничивает) целевую аудиторию каталога. Это существенное ограничение, поэтому необходимо его учесть при анализе.

Способность своевременно обслуживать входящий поток требований (заявок на публикацию интернет-ресурсов) определяет возможность скоровременного включения публикуемого интернет-ресурса в страницы каталога. Может существовать ряд причин, по которым публикация может задерживаться, например: неоптимальная реализация интерфейса администратора, дополнительные издержки на заполнение полей вручную, на верификацию сайтов и т. п. Наличие автоматизированных функций определяет насколько оперативно административный персонал каталога может управлять

интернет-ресурсами и модерацией. Под наличием автоматизированных функций следует понимать наличие АРМ модератора / администратора каталога. Данный критерий связан с предыдущим.

Наличие интеллектуальных функций определяет использование каталогом каких-либо подходов к интеллектуализации: использование интеллектуальных алгоритмов, эвристических подходов, интеллектуальных интерфейсов, баз знаний. Использование таких функций повышает эффективность каталога: как на стороне пользователя, так и на стороне администрации каталога (в зависимости от реализации).

Существует ряд методов рубрикации и категоризации интернет-ресурсов. Структура каталога определяется именно использованием этих методов, поэтому этот критерий является важным и должен быть учтен.

Под возможностью интеграции с внешними сервисами подразумевается предоставление метаданных, новостных лент, информации об интернет-ресурсах (в формате удобном для синдикации), специальных интерфейсов для поисковых систем, возможность интеграции с сервисами социальных закладок и т. п.

В каждом каталоге интернет-ресурсы внутри рубрик ранжируются по определенным правилам: именно способ упорядочивания сайтов внутри рубрик определяет какие сайты будут чаще всего посещать пользователи. Некоторые каталоги используют специальные подходы, некоторые — ранжируют интернет-ресурсы лишь по дате регистрации.

Некоторые интернет-ресурсы не позволяют после регистрации изменять информацию о сайте даже его владельцу, т. е. не предоставляют для этого специальный интерфейс. Это вызывает определенные проблемы у владельцев сайтов, поэтому этот критерий также важен.

Перечень критериев для оценки каталогов

1	Полнота
2	Метод регистрации ресурсов
3	Количество таксономических разделов
4	Уровень таксономических разделов
5	Многоязычность
6	Способность своевременно обслуживать ВПТ
7	Наличие автоматизированных функций (АРМ модератора)
8	Наличие интеллектуальных функций
9	Методы рубрикации и категоризации
10	Возможность интеграции с внешними сервисами
11	Методы ранжирования ресурсов
12	Возможность впоследствии актуализировать информацию об интернет-ресурсе в каталоге

Связь критериев и проблем вы можете видеть на рис. 8. Проблемы представлены в виде эллипсов, критерии — в виде прямоугольников.

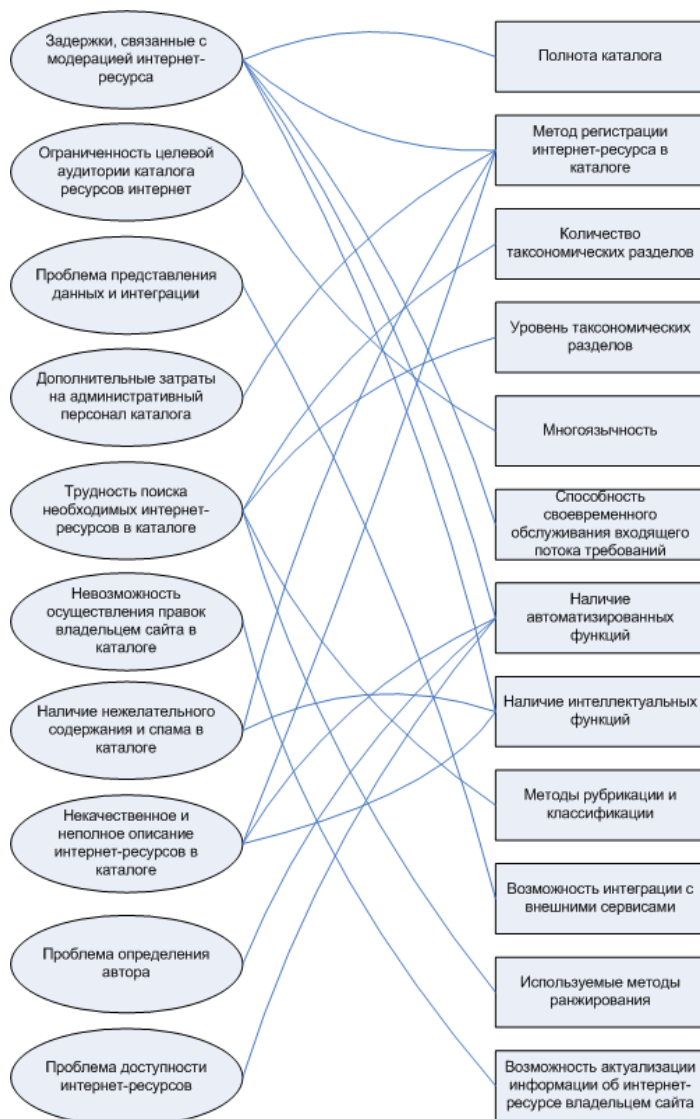


Рис. 8: Связь проблем и критериев

1.3.2. Выбор каталогов для проведения анализа

Для анализа выбрано 10 каталогов, наиболее часто упоминающихся в русскоязычной части интернета по данным поисковых систем Яндекс и Google на 12.02.2007 (табл. 2). Выбор осуществлялся следующим образом: по ключевым словам в логической связке «или»: «каталог сайтов», «каталог ресурсов интернет», «регистрация в каталоге сайтов», «добавить сайт в каталог» было выбрано 5 первых результатов поиска в поисковых машинах Яндекс и Google без повторов (всего 10 уникальных результатов).

Таблица 2

Перечень каталогов для анализа

№	Название	Адрес URL
1	Каталог Яндекса	http://yasa.yandex.ru
2	Каталог Google (ODP)	http://www.google.ru/dirhp?hl=ru
3	Рейтинг Rambler's TOP100	http://top100.rambler.ru/top100/
4	Каталог Апорт	http://www.aport.ru/
5	Каталог Yahoo	http://dir.yahoo.com/
6	The List Of Russian Webservers	http://weblist.ru/
7	Russia On The Net	http://www.ru/
8	Каталог «Всего.ру»	http://www.vsego.ru/
9	Каталог «Иван Сусанин»	http://www.susanin.net/
10	Каталог «@mail.ru»	http://list.mail.ru/

1.3.3. Сравнительный анализ

Проведен анализ каталогов по ранее определенным критериям, результаты приведены в виде таблицы (см. табл. 3).

Таблица анализа каталогов

Р/К	Полнота	Метод рег-ии	Уровень так. разд.	Много-язычность	Способ-ть обл. ВПТ	Авт. ф-ии	Инт. ф-ии	Методы рубр.	Интеграция с вн. с.	Мет. ранж.	Возм. акт. инф.
http://yaca.yandex.ru	+	мод	4	-	-	+	-	таксон., фасетный	-	собств., дата	-
http://www.google.ru/dirhp?hl=ru	+	мод	6	+	+	+	-	таксон., фасетный	+	собств., дата	-
http://top100.rambler.ru/top100/	+	н/м	5	-	+	-	-	таксоном.	-	дата	+
http://www.aport.ru/	+	мод	6	-	-	+	-	таксон., фасетный	-	собств., дата	+
http://dir.yahoo.com/	+	мод	8	-	-	+	-	таксоном.	-	собств., дата	+
http://weblist.ru/	-	мод	6	-	+	-	-	таксоном.	-	собств., дата	-
http://www.ru/	+	мод	5	-	+	+	-	таксоном.	-	дата	-
http://www.vsego.ru/	+	н/м	3	-	+	+	-	таксоном.	-	посещ., дата	+
http://www.susanin.net/	-	н/м	3	-	+	-	-	таксоном.	-	посещ., дата	-
http://list.mail.ru/	+	н/м	7	-	+	-	-	таксоном.	-	посещ., дата	+

Пояснения к таблице: критерии располагаются по столбцам, по строкам располагаются интернет-ресурсы.

Расшифровка критериев представлена в виде таблицы (см. табл. 4).

Расшифровка критериев

№	Название	Пояснения
1	Полнота	«+» - полный (содержит репрезентативное количество ресурсов), «-» - неполный
2	Метод регистрации ресурсов	«мод» - модерация (ресурс проходит модерацию перед публикацией или после нее), «н/м» - немодерируемый каталог
3	Уровень таксономических разделов	Число, отражающее уровень вложенности разделов
4	Многоязычность	«+» - многоязычный, «-» - одноязычный
5	Способность своевременно обслуживать ВПТ	«+» - входящий поток заявок может быть полностью обслужен в течение суток (по экспериментальным данным)
6	Наличие автоматизированных функций	«+» - функция имеется в наличии, «-» - функция отсутствует
7	Наличие интеллектуальных функций	«+» - функция имеется в наличии, «-» - функция отсутствует
8	Методы рубрикации и классификации	«таксон.» - таксономическая, «фасетн.» - фасетная, «фолк.» - фолксономическая
9	Возможность интеграции с внешними сервисами	«+» - функция имеется в наличии, «-» - функция отсутствует
10	Методы ранжирования	«собств.» - собственный, «дата» - ранжирование по дате регистрации, «посещ.» - основанное на показателях
11	Возможность актуализации информации после регистрации ресурса	«+» - функция имеется в наличии, «-» - функция отсутствует

1.3.4. Результаты

В результате проведения анализа было выявлено, что наиболее популярные, в российской части интернета, каталоги содержат достаточное количество ресурсов, некоторые из них являются модерируемыми, некоторые не модерируются: используется как система пост-модерации, так и система предварительной модерации. В основном, каталоги строятся на основе таксономиче-

ских справочников (при этом, используются глубокие уровни вложенности — более трех уровней — которые на практике применять затруднительно). Каталоги, в основном, являются русскоязычными (исключительный случай: каталог Google, см. [12]). Более половины каталогов могут обслужить заявку на регистрацию сайта в течение суток, в некоторых каталогах процесс обработки заявки занимает более месяца (например, в каталоге Яндекса при регистрации на бесплатной основе, см. [11]). Многие популярные каталоги содержат автоматизированные функции, т. е. часть работы выполняется в автоматизированном режиме. Ни один из рассмотренных каталогов не использует интеллектуальные функции для осуществления своей работы. Некоторые каталоги используют, наряду с таксономическим подходом к классификации, и фасетный подход. Фолксономический подход практически не используется. Только один каталог из десяти позволяет осуществить интеграцию с внешними сервисами. Менее половины каталогов используют собственные механизмы для внутреннего ранжирования ресурсов, остальные ранжируют интернет-ресурсы по дате добавления в каталог (как правило, в прямом и обратном порядке). Половина каталогов позволяет изменять информацию об интернет-ресурсе после его регистрации, остальные же каталоги не предоставляют такого интерфейса (однако, изменение возможно путем направления заявки на электронную почту администратору каталога).

Таким образом, отчетливо видны основные проблемы каталогов ресурсов интернет, которые необходимо решать.

В связи с большим количеством интернет-ресурсов, и, соответственно, большим количеством заявок на публикацию интернет-ресурсов в каталогах возникает проблема модерации интернет-ресурсов. Решать эту проблему можно различными способами. Однако, решение ее за счет привлечения людских ресурсов не является эффективным.

Существует и проблема рубрикации: зачастую таксономический справочник перегружают большим количеством разделов и уровней иерархии, что негативно сказывается на удобстве использования интерфейса. Фасетные же признаки способны лишь помогать пользователю при фильтрации интернет-

ресурсов внутри таксономических рубрик по определенным критериям. Очевидно, что совмещение таксономического справочника и фасетной категоризации позволяют сужать область поиска, и, соответственно, повышать эффективность использования каталога.

Деятельность администраторов каталогов ресурсов интернет зачастую даже не автоматизирована, соответственно, в рассмотренных случаях об интеллектуализации даже не идет и речи.

Создатели каталогов практически всегда замыкают его реализацию «на себе» и не позволяют производить интеграцию с внешними сервисами. Таким образом, каталог не получает развития, а его целевая аудитория сужается.

Одной из актуальных и наиболее важных проблем является проблема ранжирования интернет-ресурсов, которая во многих случаях является не решенной. Ранжирование по дате регистрации интернет-ресурса — всего лишь один из возможных подходов. Ранжирование в каталоге должно осуществляться по нескольким критериям, возможно, в виде агрегированного весового рейтинга (учитывающего различные факторы, влияющие на качество интернет-ресурса).

Одна из очевидных функций, необходимых для корректного функционирования каталога — изменение информации об интернет-ресурсах владельцем сайта, после регистрации — в половине реализаций отсутствует. Зачастую именно отсутствие этой функции производит высокую нагрузку на административный персонал каталога и приводит к дополнительным тратам.

1.4. Использование элементов искусственного интеллекта для улучшения качества работы каталога

Интеллектуальные информационные технологии – одна из наиболее перспективных и быстро развивающихся научных и прикладных областей информатики, уже сейчас эта область дает обществу практически значимые результаты, многие из которых способствуют кардинальным изменениям в сферах их применения. Целью интеллектуальных информационных технологий являются, во-первых, расширение круга задач, решаемых с помощью компьютеров, особенно в слабоструктурированных предметных областях, и во-вторых, повышение уровня интеллектуальной информационной поддержки пользователей.

Существует ряд подходов к интеллектуализации программного обеспечения. Интеллектуализация каталога ресурсов интернет практически не отличается от интеллектуализации других видов программного обеспечения, поэтому подходы будут едиными. Как правило, эти подходы зависят от способа представления знаний (т. е. модели знаний).

В настоящее время в интеллектуальных информационных системах применяются семь классов моделей знаний: логические, продукционные, фреймовые, сетевые, объектно-ориентированные, комплексные и специальные. Именно на основе одного из этих классов строятся интеллектуальные информационные системы. Основной задачей разработчика/проектировщика является выбор класса модели представления знаний [7].

В логических моделях знания представляются в виде совокупности правильно построенных формул какой-либо формальной системы. Такие модели являются наиболее простыми и накладывают ограничения на реализацию: все правила-знания должны быть истинными, в обратном случае полезность такой модели обесценивается.

Центральным звеном продукционной модели является множество продукций или правил вывода. Каждая такая продукция может быть представлена выражением (записью) со следующими компонентами: сфера применения продукции (условие), необходимое предусловие, ядро продукции (собственно необходимая операция заданная в виде «если..., то...»), постусловие продукции (изменения, которые необходимо внести после выполнения продукции). Системы, основанные на продукционной модели, состоят (как правило) их трех компонентов: базы правил (продукций), базы фактов (декларативные знания о предметной области) и интерпретатора продукций.

Фундаментом фреймовой модели служит понятие фрейма – структуры данных, представляющей некоторый концептуальный объект или типовую ситуацию. Фрейм идентифицируется уникальным именем и включает в себя несколько слотов. Каждому слоту соответствует определенная структура данных. Представление предметной области в виде иерархической структуры фреймов хорошо отражает внутреннюю и внешнюю структуры объектов этой

предметной области. Организация вывода во фреймовой системе базируется на обмене сообщениями между фреймами, активации и выполнении присоединенных процедур. Реализация фреймовой модели является сравнительно сложной.

Наиболее общий способ представления знаний, при котором предметная область рассматривается как совокупность объектов и связывающих их отношений, реализован в сетевой модели знаний. В качестве носителя знаний в этой модели выступает семантическая сеть, вершины которой соответствуют понятиям (объектам), а дуги – отношениям между понятиями. Поскольку фактически сетевая модель объединяет множество методов представления предметной области с помощью сетей, сопоставление данной модели с прочими способами представления знаний затруднительно. Очевидные достоинства сетевой модели заключаются в ее высокой общности, а также легкости понимания такого представления. В то же время в семантической сети имеет место смешение групп знаний, относящихся к совершенно различным ситуациям при назначении дуг между вершинами, что усложняет интерпретацию знаний. Другая проблема, присущая сетевой модели, состоит в трудности унификации процедур вывода и механизмов управления выводами на сети.

Объектно-ориентированная модель знаний получила широкое применение в современных технологиях проектирования разнообразных программных и информационных систем. Такой подход позволяет моделировать одну и ту же предметную область с различных точек зрения, однако при таком подходе возможны потери информации при переходе от одного представления к другому. Также, существует специальная метамодель (MDA, ODP), которая ограничивает способы использования такого подхода.

Класс специальных моделей знаний объединяет модели, отражающие особенности представления знаний и решения задач в отдельных, относительно узких предметных областях. В качестве примера подобного способа формализации знаний можно привести модель «объект-признак», используемую в автоматизированных системах поиска аналогов.

Применение на практике того или иного способа формализации обуславливается спецификой задачи, для решения которой планируется использовать

базу знаний. По мнению специалистов [7] наиболее перспективны смешанные или комплексные модели, интегрирующие преимущества различных базовых моделей представления предметной области.

Подходы к интеллектуализации каталогов ресурсов интернет не ограничиваются лишь представлением предметной области в виде базы знаний.

Существует и ряд других подходов, более «мелких», однако, не менее эффективных. В ряде случаев можно применять эвристические алгоритмы для совершения ряда действий. Например, широкое распространение в ИПС получили эвристические алгоритмы для нормализации слов, так называемые «стеммеры» (получающие без словаря, эвристическим способом слово в единственном числе, именительном падеже).

Знания для подсистемы искусственного интеллекта каталога можно получать при помощи экспертов. Особенно эффективным является подход получения знаний процедурного характера и записи таких знаний в алгоритмической форме. Таким образом легко интеллектуализировать множество функций каталога, выполняемых человеком; например, функцию регистрации интернет-ресурса и связанную с ней функцию модерации. Достаточно уточнить перечень критериев, которыми пользуется эксперт, проводя модерацию интернет-ресурса, определить степень их важности и записать порядок осуществления проверки соответствия критериям в виде алгоритма.

2. Проектирование каталога

2.1. Цели и задачи проектирования

Основной целью проекта является создание удобного для конечных пользователей средства поиска информации в сети интернет в виде структурированного хранилища, поддерживаемого, по возможности, с минимальным участием человека.

В рамках основной цели выделен ряд задач:

- 1) создание основной программной части каталога (обеспечивающей работу основных функций);
- 2) создание дополнительной программной части, использующей современные достижения в области информационных технологий и интернет;
- 3) продвижение и реклама каталога в сети интернет;
- 4) продажа рекламы на страницах каталога.

2.2. Ограничения на ресурсы

Временные рамки проекта (создание программной части каталога и документации) жестко регламентированы. Дата начала проекта: 1 марта 2006 года. Дата окончания проекта: 1 апреля 2007 года. Ориентировочная продолжительность проекта: 13 календарных месяцев.

Трудовые ресурсы проекта ограничены. Разработкой проекта и программной части занимается автор. По мере необходимости, привлекаются третьи лица для консультаций и осуществления аутсорсинга услуг.

Каталог разрабатывается в стиле близком к ХР (методология экстремального программирования). При этом большее внимание будет уделяться практическим аспектам создания проекта и меньшее внимание будет уделяться документированию проекта. Такой подход показал достаточно высокую эффективность в предыдущих проектах разработки каталогов автором.

2.3. Необходимые функции

Для реализации каталога в полном объеме необходима поддержка ряда

основных и дополнительных функций. Эти функции будут описаны ниже.

2.3.1. Функции и модули автора ресурса

Основная задача автора интернет-ресурса — регистрация интернет-ресурсов в базе данных каталога и последующее управление ими. В таблице 5 приведен полный перечень необходимых функций для автора ресурса.

Таблица 5

Функции автора ресурса

Наименование функции	Описание
Регистрация учетной записи в каталоге	Позволяет пользователю каталога зарегистрировать собственную учетную запись, указав свой адрес электронной почты и желаемый пароль.
Смена пароля	Позволяет в панели управления учетной записью изменить пароль на новый.
Восстановление пароля	Позволяет изменить забытый пароль на новый при помощи адреса электронной почты.
Удаление учетной записи пользователя	Позволяет удалить учетную запись пользователя и всю с ней связанную информацию (рекурсивно).
Регистрация (первичная) интернет-ресурса	Позволяет зарегистрировать интернет-ресурс в каталоге путем указания адреса URL и его заголовка. При этом происходит автоматическая проверка корректности введенных данных.
Отображение перечня интернет-ресурсов, связанных с учетной записью пользователя	Отображает интернет-ресурсы, зарегистрированные текущим пользователем и позволяет переходить к управлению этими ресурсами.
Изменение заголовка	Позволяет изменить заголовок интернет-ресурса на новый, при этом его корректность проверяется автоматически.
Изменение описаний	Позволяет добавлять новые описания интернет-ресурсов или редактировать ранее добавленные. Используется редактор <i>FCK Editor</i> на основе JavaScript с интерфейсом, похожим на MS Word. Корректность описания проверяется автоматически.
Изменение перечня ассоциаций с таксономическим справочником	Позволяет связать интернет-ресурс со справочником иерархических разделов
Изменение перечня ассоциаций с фолксономическим справочником	Позволяет указать теги (текстовые метки) для интернет-ресурса

Окончание табл. 5.

Наименование функции	Описание
Формирование заявки на разрешение публикации интернет-ресурса в каталоге	Позволяет сформировать запрос к интеллектуальному модулю на предмет возможности осуществления публикации интернет-ресурса в каталоге.
Запрет регистрации интернет-ресурса в каталоге	Возможность запретить публикацию в каталоге интернет-ресурса.
Просмотр и анализ рейтинга интернет-ресурса	Позволяет просмотреть взвешенный рейтинг интернет-ресурса и все его составляющие, при этом по каждой компоненте рейтинга выводится комментарий, объясняющий как его можно повысить.
Установка и проверка обратных ссылок	Позволяет установить обратную ссылку на каталог и учесть этот факт в системе рейтинга каталога.

2.3.2. Функции и модули рядового пользователя каталога

Рядовой пользователь каталога обзорекает каталог и пользуется общедоступными функциями, не требующими авторизации. Полный перечень необходимых функций приведен в таблице 6.

Таблица 6

Функции рядового пользователя

Наименование функции	Описание
Навигация по таксономическим разделам	Позволяет производить навигацию по иерархическим разделам каталога и осуществлять просмотр связанных с разделами интернет-ресурсов. Поддерживает постраничную навигацию и результаты в виде RDF/XML.
Навигация по фолксономическим тегам	Позволяет производить навигацию по облаку таксономических тегов каталога и осуществлять просмотр связанных с разделами интернет-ресурсов. Поддерживает постраничную навигацию и результаты в виде RDF/XML.
Поиск по каталогу	Позволяет осуществлять полнотекстовый поиск по заголовкам интернет-ресурсов каталога. Поддерживается булев поиск с возможностью задания критериев, приоритета слов, квантификаторов исключения и необходимости наличия. Поддерживается технология OpenSearch XML.

Окончание табл. 6.

Наименование функции	Описание
Отображение последних зарегистрированных ресурсов	Позволяет осуществлять обзор последних добавленных в каталог интернет-ресурсов, которые получили одобрение к публикации.
Отображение ресурсов с наибольшим рейтингом	Позволяет осуществлять обзор ресурсов с наибольшим взвешенным рейтингом.
Отображение информации об интернет-ресурсе	Формирует персональную страницу для каждого интернет-ресурса, содержащую ряд информационных блоков, отображающих всю доступную в каталоге информацию об интернет-ресурсе.
Панель настроек для пользователей	Позволяет производить некоторые настройки для удобства пользования каталогом: изменить размер шрифта, отключить отображение рекламных материалов.
Панель социальных закладок	Позволяет добавлять интернет-ресурс, описанный в каталоге в ряд наиболее популярных сервисов интернет-закладок.
Формирование скриншота сайта	Позволяет отображать на странице описания интернет-ресурса скриншот сайта (изображение сайта, сделанное в браузере).
Отображение рейтингов интернет-ресурса	Позволяет отображать взвешенный рейтинг интернет-ресурса и все его компоненты на странице описания интернет-ресурса.
Интеллектуальная социальная модерация	Позволяет отправить сайт на модерацию интеллектуальным модулем.
Комментарии к интернет-ресурсу	Возможность добавлять комментарии и просматривать комментарии, которые были добавлены ранее.

2.3.3. Общие функции и модули каталога

Существует ряд общих функций, которые необходимы для корректного функционирования каталога, перечень этих функций представлен в таблице 7.

Таблица 7

Общие функции и модули

Наименование функции	Описание
Отображение шаблона	Позволяет отображать общий шаблон для всех страниц каталога.
Отображение страницы	Позволяет отображать страницы и модули каталога и формировать их содержание.

Окончание табл. 7.

Наименование функции	Описание
Человекопонятные URL	Позволяет скрыть динамические адреса и привести их в более пригодный для понимания человеком вид.
Модуль для отображения информационных потоков RSS XML	Позволяет формировать информационные потоки на основе формата RDF RSS/XML.
Модуль открытого поиска OpenSearch XML	Позволяет осуществлять поиск по каталогу при помощи технологии открытого поиска. Результаты поисковых запросов транслируются в формате XML.
Модуль отображения метаданных в формате Dublin Core XML	Позволяет отображать информацию об интернет-ресурсе в понятном для компьютера виде. Эта информация может быть использована рядом интеллектуальных агентов поисковых машин и позволит строить более качественные результаты поиска.
Интеллектуальная модерация интернет-ресурса	Позволяет проверять интернет-ресурс по ряду критериев на наличие спама, бранных выражений и прочих критериев на основе эвристических правил записанных алгоритмически.
Поддержка технологии AJAX	Позволяет изменять информацию на страницах без перезагрузки страниц (используется асинхронный запрос к веб-серверу).
Модуль интеллектуальных документов	Предоставляет методы для создания, отображения, интерпретации и поиска документов, создаваемых интеллектуальными модулями.

2.3.4. Функции и модули административного персонала каталога

Административный интерфейс каталога является минимальным. Перечень функций, необходимых администратору представлен в таблице 8.

Таблица 8

Функции и модули администратора

Наименование функции	Описание
Панель управления администратора	Предоставляет доступ к функциям администрирования.
Модуль управления таксономическим справочником	Позволяет наполнять и изменять таксономический справочник.
Модуль ручной модерации интернет-ресурсов	Позволяет вручную производить модерацию интернет-ресурсов и накладывать санкции на интернет-ресурсы.

2.3.5. Система рейтинга ресурсов

Интернет-ресурсы в каталоге следует ранжировать по ряду критериев, таких как: популярность ресурса, степень соответствия описания в каталоге действительному содержанию сайта, степень заполненности информации о ресурсе в каталоге; также следует учитывать штрафные санкции. Для этого нужна собственная система рейтинга интернет-ресурсов.

Интернет-ресурсы в каталоге упорядочиваются по убыванию взвешенного рейтинга (Weighted Rating).

WR — взвешенный рейтинг-агрегат, рассчитывающийся на основе других рейтингов каталога (формула 1). Значения взвешенного рейтинга являются вещественными числами. Взвешенный рейтинг не может принимать значения меньшие нуля и большие 1000.

$$WR = ((PR \times 80 + VR \times 5 + CR \times 20 + ER \times 9 + SR \times 8) / 5) / FR \quad (1)$$

PR — рейтинг PageRank. Данный рейтинг является зарегистрированной торговой маркой компании Google. Автор каталога никак не связан с авторством данной технологии. Все права принадлежат компании Google, inc. Мы используем данный рейтинг по мере возможности как один из компонентов для расчета взвешенного рейтинга. Данные рейтинга берутся при помощи того же способа, который используется в тулбаре Google.

PageRank выполняет объективную оценку значимости веб-страниц путем расчета уравнения с более 500 переменными и 2 миллиардами терминов. Вместо подсчета прямых ссылок PageRank интерпретирует ссылку страницы А на страницу В как голос страницы А в пользу страницы В. Затем PageRank оценивает значимость страницы по числу полученных голосов [16].

PageRank используется в каталоге для осуществления оценки ссылочного ранжирования интернет-ресурса. Вес рейтинга в модели (WR): 80 — это обусловлено тем, что PR является целочисленной величиной в диапазоне от нуля до десяти. Вес в модели следует трактовать как указано в формуле 2.

$$PR_w = (PR \times 10) \times 8; \quad (2)$$

{ приведение к диапазону 0-100, вес 8 }

VR (Value Rating) — содержательный рейтинг используется в каталоге для оценки уровня заполненности возможных данных ресурса. Если данный рейтинг низкий, скорее всего, автор сайта указал слишком мало семантических данных или короткое описание интернет ресурса. Рейтинг является целочисленным, его значения лежат в диапазоне от нуля до ста.

Рейтинг рассчитывается исходя из половины возможных заполненных данных. То есть, если автор сайта заполнил описание сайта хотя бы наполовину возможной длины, указал 3 тега и, по крайней мере, 2 таксономических раздела — содержательный рейтинг его сайта будет равным 100. Если же информация заполнена на четверть — VR будет равным 50 и т. д.

CR (Correlation Rating) — рейтинг соответствия описания интернет-ресурса в каталоге его действительному содержанию. Рассчитывается в процентном эквиваленте. При расчете данного рейтинга используются эвристические механизмы (проводится нормализация слов с использованием стеммеров: в настоящее время поддерживаются только русскоязычные и англоязычные слова и фразы). Этот рейтинг целочисленный и принимает значения в диапазоне от нуля до ста.

Данный рейтинг имеет наибольший вес в модели, однако, при значениях рейтинга соответствия больше 50 начисляются штрафные баллы в виде FR-рейтинга. Объяснение: если описание сайта в каталоге соответствует сайту более чем наполовину, наиболее вероятно, что данное описание создавалось специально для «накрутки» этого рейтинга; подобные действия нужно пресекать.

ER (Expandable Rating) — расходуемый рейтинг используется в каталоге как бонус для новых сайтов и сайтов, содержание которых в каталоге регулярно обновляется. В момент регистрации или при внесении изменений в описание сайта, ресурсу присваивается расходуемый рейтинг, равный 70 пунктам.

Данный рейтинг расходуется ежедневно: один раз в сутки значение ER для всех сайтов уменьшается на 5 единиц, вплоть до нуля. Рейтинг ER не может

принимать значения меньше нуля и больше ста пунктов.

SR (Static Rating) — этот рейтинг в каталоге меньше всего подвержен влиянию внешних факторов. Статический рейтинг присуждается ресурсу за определенные действия, связанные с развитием каталога. Например, для любого интернет-ресурса, размещенного в каталоге, можно повысить рейтинг путем размещения обратных ссылок. Рейтинг SR целочисленный и принимает значения в диапазоне 0—100 пунктов.

FR (Fine Rating) — рейтинг штрафов, используется в каталоге как одна из санкций для нарушителей. Например, если владелец сайта разместил обратную ссылку в нарушение правил размещения, его следует наказать повышением штрафного рейтинга. FR является действенной мерой штрафования сайтов.

По умолчанию, для всех ресурсов в каталоге, штрафной рейтинг равен единице (т. е. нет штрафов). За каждое нарушение штрафной рейтинг повышается на единицу. В случае грубого нарушения, размер штрафа может быть более единицы. Если вы обратили внимание, в формуле расчета взвешенного рейтинга, FR используется как делитель. Т. е., если ресурс получил хотя бы один штрафной рейтинг ($FR = 2$), его взвешенный рейтинг делится пополам: если $FR = 3$, взвешенный рейтинг делится на три и т. д.

2.4. Реализация функций каталога с использованием элементов искусственного интеллекта

Был выбран ряд подходов (см. раздел 1.4) к интеллектуализации, применимых для каталога ресурсов интернет: учитывая опыт разработки систем подобного рода автором, от способа представления знаний в виде сетевой модели пришлось отказаться. Несмотря на то, что такое представление предметной области для каталога при первом приближении кажется наиболее подходящим, использование сетевого представления знаний (или же его упрощенного аналога — семантических отношений) на практике затруднено. Хостинг-провайдеры не разрешают устанавливать дополнительное необходимое интеллектуальное программное обеспечение для работы с базами знаний, а собственноручная реализация работает не достаточно эффективно: образуется

большая избыточность информации, выборки из реляционных баз данных по сложным запросам (коих в такой реализации большинство, например, одна из тривиальных выборок: «выбрать свойства №{5, 6, 7} сайта №3, для которого значение свойства №2 равно X») являются весьма ресурсоемкими, что непозволительно в условиях ограниченности ресурсов сервера и растущей посещаемости каталога.

Достаточно эффективным способом оказался смешанный: продукционно-алгоритмический способ представления знаний. В котором, в качестве условий продукций задается сфера ее применения (например, регистрация сайта, указание заголовка, указание тегов), в качестве пре-условия задаются ограничения, а ядро продукции представляет собой набор необходимых действий, записанных алгоритмически, интегрированных с объектной моделью каталога; в пост-условии задаются дополнительные действия, которые необходимо выполнить после выполнения ядра продукции.

Таким образом, при таком подходе в базе знаний хранится информация о том, какие события могут быть обработаны подсистемой искусственного интеллекта, какие действия необходимо предпринять в рамках события, какие ограничения (как правило, на входные данные) необходимо проверить и, наконец, что нужно сделать в случае успеха или отказа в выполнении ядра продукции. Такая модель является полной и способна использоваться в каталоге. Более того, такая модель легко расширяется: пополнить базу знаний новыми продуктами не сложно.

Описанный выше способ успешно работает в предыдущем проекте каталога автора (доступном по адресу <http://dir.ikernel.org>).

Если же абстрагироваться от деталей реализации, то общую схему работы такой модели можно представить следующим образом: инженер по знаниям (администратор, разработчик каталога) получают знания от эксперта в виде описаний его действий при выполнении определенной функции в каталоге. После чего, эти знания представляются в формальном виде (пригодном для интерпретации решателем) и помещаются в базу знаний. С другой стороны, пользователь (или же интеллектуальный модуль) обращается к интеллектуальному ин-

терфейсу каталога, определяющему какую из продукций вызвать в данном контексте обращения и передает управление блоку решателя. Решатель производит необходимые проверки начальных условий, выполняет заданную ранее алгоритмическую часть и пост-условие, после чего формирует «документ», который может быть интерпретирован инициировавшим продукцию интеллектуальным модулем, и, по мере необходимости, представлен в понятном для пользователя виде. Общую схему работы такой модели вы можете видеть на рис. 9.

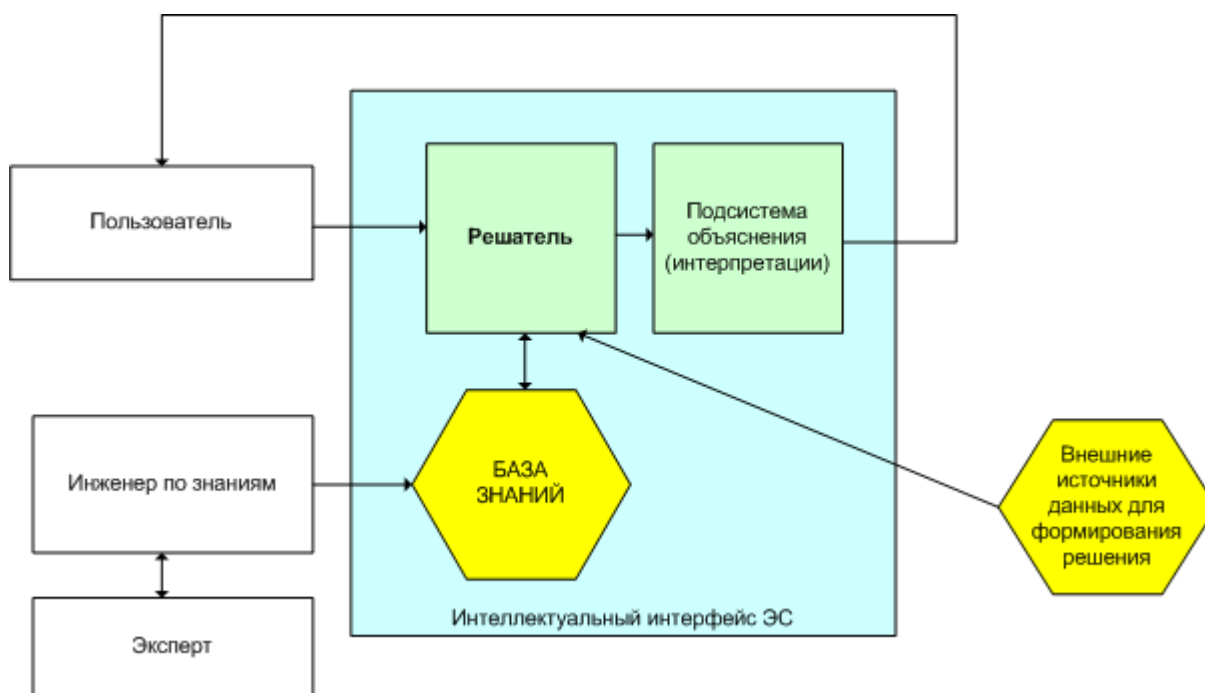


Рис. 9: Схема работы подсистемы ИИ

Помимо представления знаний экспертов в каталоге используются и другие подходы к интеллектуализации. Например, широкое применение получил «стемминг» (нормализация слов): использован алгоритм, известный как «Стеммер Портера», позволяющий эвристическим способом путем исключения широкоупотребимых окончаний, суффиксов и прочими способами получать нормальную форму слова. Стеммер используется как при поиске, так и для расчета рейтингов интернет-ресурса (при определении степени соответствия описания его действительному содержанию).

В каталоге используется небольшая база для проверки синтаксиса предложений на русском языке, работающая на основе эвристических правил. Она не претендует на звание полноценной, однако, позволяет исправлять

наиболее распространенные синтаксические ошибки непосредственно перед публикацией и оформлять текст в соответствии с некоторыми типографскими правилами.

Также, благодаря разработанной подсистеме интерпретации, документы, формируемые решателем могут быть легко интерпретированы интеллектуальным модулем и представлены в понятной для человека форме.

Как было отмечено выше, в каталоге используется продукционно-алгоритмическая база знаний. В этой базе знаний хранится структура, при помощи которой решатель может определить события (активаторы продукций), которые могут быть обработаны на интеллектуальном уровне, а также методы обработки этих событий, описанные алгоритмически (табл. 9). Вся база знаний описана на языке РНР, продукции хранятся непосредственно в исходных текстах и определены заранее. В данный момент не существует возможности их пополнения автоматическим способом, их можно задать/изменить только «вручную», отредактировав исходный текст.

Таблица 9

Таблица «Интеллектуальные функции каталога»

Событие	Описание метода обработки
Регистрация интернет-ресурса (первичная регистрация)	Верификация исходных данных (проверка на ограничения), проверка заголовка интернет-ресурса на недоброкачественность (используется расширяемый словарь «отпечатков» нежелательных слов (Obscenity filter heuristic).
Добавление/редактирование заголовка сайта	Верификация исходных данных, проверка заголовка на недоброкачественность, типографская проверка текста.
Добавление/редактирование описания	Верификация исходных данных, проверка описания на недоброкачественность, типографская проверка текста, проверка на «спам».
Добавление/изменение тега (ключевого слова)	Верификация исходных данных, проверка на недоброкачественность, анализ «сходных» тегов.
Расчет статического рейтинга интернет-ресурса	Получение содержания страницы с установленной обратной ссылкой, поиск ссылки на эту страницу с главной страницы сайта, разбор дерева документа (DOM), получение перечня возможных ссылок; оценка страницы (количество внешних ссылок, рейтинг PageRank), расчет статического рейтинга на основе этих показателей.

Окончание табл. 9.

Событие	Описание метода обработки
Расчет рейтинга соответствия интернет-ресурса	Получение содержания страницы, выбор информационного наполнения страницы и его представление в виде перечня слов, исключая повторы нормализация полученных слов (эвристический стеммер Портера); сравнение полученного перечня с перечнем (полученным аналогичным образом) в каталоге (заголовок, описание, ключевые слова). Расчет соотношения совпадений (слов внутри каталога к словам на сайте).
Запрос интерпретации документа пользователем	Получение документа из общей базы документов; «десераилизация» массива (восстановление объекта из строки в БД), разбор документа интеллектуальным модулем, его создавшим, представление документа на понятном человеку языке.

В таблице выше, в поле «Описание метода обработки», указывались некоторые подходы; ниже будут даны их более подробные описания.

Верификация исходных данных подразумевает наличие определенных ограничений (пре-условия продукции). Например, в случае редактирования заголовка сайта, такими ограничениями служат: ограничение длины заголовка равное 150 символам, ограничение алфавита заголовка (русские и английские буквы, цифры, знаки пунктуации и кавычки). Могут и существовать дополнительные ограничения, задаваемые алгоритмически (например, требование к отсутствию дублирующихся записей при указании разделов).

Проверка исходных данных на недоброкачественность подразумевает эвристическую проверку текста на вхождения в нее т. н. «отпечатков» сообщений (или же слов). В том случае, если обнаруживается хотя бы одно вхождение «отпечатка» в текст, его следует считать недоброкачественным и производить отказ в его приеме; уведомлять об этом пользователя.

Типографская проверка (автоматическое форматирование) текста позволяет оформлять неверный типографически текст в соответствии с правилами оформления. К таким правилам относятся правила расстановки кавычек, замена знаков дюйма (") на корректные кавычки: лапки и елочки. Замена повторяющихся неуместных знаков пунктуации, расстановка неразрывных

пробелов перед предлогами и т. п. Такие ошибки часто возникают при публикации интернет-ресурсов в каталоге и такой подход позволяет снизить их количество, что, в свою очередь, повышает восприятие текста пользователями.

Проверка на «спам» позволяет на основе базы отпечатков сообщений рекламного характера выявлять их вхождения в текст, и определять текст как содержащий спам, в некоторых случаях. Также, используется частотная проверка: если одно из слов текста (за исключением предлогов и слов с длиной менее 3 символов) встречается в тексте слишком часто (соотношение к другим словам более 30%), текст считается спамом, и, наиболее вероятно, носит рекламный характер, активно упоминающий бренд или название услуги. Аналогичным образом ведется учет регистра: рекламные объявления зачастую содержат большую часть текста в верхнем регистре, и, в том случае, если соотношение символов в верхнем регистре к символам в нижнем регистре составляет более 30%, текст считается спамом. Использование таких методов позволило практически избавиться содержание каталога от спама.

Анализ сходных тегов позволяют уменьшить дублирование тегов, сходных по своему смыслу, но разных по написанию. Используется эвристический алгоритм, известный как «стеммер Портера», при поиске сходных тегов в таблице базы данных. Найденные теги, отсортированные в порядке убывания частоты их упоминания, предлагаются к использованию.

Стеммер Портера позволяет без словаря находить начальную форму слов (именительный падеж, единственное число), все (возможные) формы слова. На практике, в каталоге используется при поиске (при осуществлении запроса сначала осуществляется поиск по заданному запросу, и, в том случае, если поиск возвращает ноль результатов, осуществляется запрос по всем возможным формам запросов на основе стеммера), при поиске сходных тегов, при расчете рейтингов (для учета соотношения слов).

2.5. Информационное обеспечение

2.5.1. Структура интернет-ресурса как объекта каталога

Если рассматривать интернет-ресурс, как объект, то в нем можно выде-

лить свойства, присущие каждому экземпляру (см. табл. 10).

Таблица 10

Структура интернет-ресурса как объекта каталога

Название	Описание
URL	Уникальный идентификатор ресурса – адрес интернет-ресурса в сети.
Заголовок сайта	Каждая страница, в соответствии с рекомендацией консорциума W3 должна иметь собственный заголовок.
Описание сайта	Каждый сайт в каталоге должен иметь описание, понятное для конечного пользователя.
Язык сайта	Для каждого сайта присущ определенный язык (natural language); однако, некоторые сайты могут быть и многоязычными.
Признак блокировки	По различным причинам, интернет-ресурс в каталоге может быть доступен или заблокирован для конечных пользователей.
Владелец / автор сайта	Пользователь, зарегистрировавший сайт в каталоги и имеющий право впоследствии управлять им.
Даты	Объект, содержащий информацию о дате регистрации сайта, дате последнего изменения и модерации для записи.

Также, для каталогов ресурсов интернет характерны определенные отношения для группировки ресурсов по определенным признакам. Это таксономические и фолксономические отношения.

Под реализацией таксономических отношений подразумевается создание иерархического справочника разделов, содержащего не более трех вложенных уровней. Любой интернет-ресурс в каталоге можно будет привязать не более чем к трем таким разделам.

Под фолксономическими отношениями следует понимать категоризацию при помощи тегов (определенных текстовых маркеров), задаваемых пользователями в свободной форме. Любой интернет-ресурс может иметь не более пяти уникальных тегов.

2.5.2. Модель данных для каталога

На основе перечня хранимых данных и необходимых функций можно

сформировать первичную модель данных каталога, которая, впоследствии, может быть доработана.

Основной таблицей является т. н. «База интернет-ресурсов», ее структура представлена в таблице 11.

Таблица 11

Таблица «База интернет-ресурсов»

Поле	Тип	Ноль	По умолчанию
id	int(11)	Нет	
dc_identifier	varchar(70)	Нет	
dc_title	varchar(120)	Нет	
dc_lang	enum('ru', 'en', 'un')	Нет	un
sys_enabled	boolean	Нет	
sys_owner_id	int(11)	Нет	0
sys_date_created	int(11)	Нет	0
sys_date_modified	int(11)	Нет	0
sys_date_moderated	int(11)	Да	0

Таблица «База интернет-ресурсов» предназначена для хранения записей о зарегистрированных интернет-ресурсах. В этой таблице хранится основная информация и некоторые значения, используемые системой каталога для внутренних нужд.

Комментарии к полям:

- 1) id — уникальный идентификатор ресурса;
- 2) dc_identifier — URL сайта;
- 3) dc_title — заголовок сайта;
- 4) dc_lang — язык сайта (двухсимвольный код языка);
- 5) sys_enabled — признак доступности (публикации) сайта;
- 6) sys_owner_id — идентификатор владельца сайта;
- 7) sys_date_created — дата регистрации сайта (метка времени).
- 8) sys_date_modified — дата изменения сайта (метка времени).
- 9) sys_date_moderated — дата последней модерации сайта (метка времени).

Структура таблицы отношений представлена в таблице 12.

Таблица «Отношения»

Поле	Тип
id	int(11)
base_id	int(11)
ent_id	int(11)
rel_name	enum('dc_creator', 'dc_description', 'dc_subject', 'dc_keyword', 'dc_comment')

Данная таблица предназначена для хранения отношений различных видов между записью интернет-ресурса и связанными элементами. Необходимость такой таблицы обусловлена возможностью существования множественных отношений (вида «один ко многим»). Сам тип отношения задается отдельным полем. Такая структура удобна для осуществления массовых операций с записью интернет-ресурса и связанными с ним полями. Например, при удалении учетной записи пользователя каталога или удалении интернет-ресурса достаточно удалить все элементы, имеющие значение «base_id» равное идентификатору интернет-ресурса, а также удалить рекурсивно все записи из смежных таблиц, указанных в поле «rel_name».

Комментарии к полям:

- 1) id – идентификатор отношения;
- 2) base_id – идентификатор ресурса;
- 3) ent_id – идентификатор сущности внешней таблицы;
- 4) rel_name – имя (тип) отношения.

Для полноты понимания необходимости такой таблицы, приведем пример хранения в ней отношения «описание» («dc_description») для интернет-ресурса №7. Интернет-ресурс с идентификатором №7 уже зарегистрирован в базе данных каталога. В таблицу описаний добавляется запись {«id», «value», «value_lang»} => {DESC_ID, «Описание сайта», «ru»}; после чего, в таблицу отношений добавляется запись, связывающая интернет-ресурс №7 с только что добавленной записью {«id», «base_id», «ent_id», «rel_name»} => {REL_ID, «7», DESC_ID, «dc_description»}. Теперь интернет-ресурс связан с вновь добавленным описанием. Такая организация взаимосвязей более удачна,

нежели, если бы связи осуществлялись внутри множества таблиц: во-первых, можно одним не сложным запросом получить все связанные элементы для конкретного интернет-ресурса, во-вторых, такой подход позволяет довольно-таки легко рекурсивно удалять все связанные с интернет-ресурсом записи, например при операции «Удаление интернет-ресурса» или операции «Удаление учетной записи пользователя».

Структура таблицы комментариев представлена в таблице 13.

Таблица 13

Таблица «комментарии»

Поле	Тип
id	int(11)
rid	int(11)
oid	int(11)
time	int(11)
name	varchar(100)
comment	text

Данная таблица предназначена для хранения комментариев, которые могут оставлять зарегистрированные пользователи.

Комментарии к полям:

- 1) id — идентификатор комментария;
- 2) rid — идентификатор интернет-ресурса;
- 3) oid — идентификатор автора комментария;
- 4) time — метка времени добавления комментария;
- 5) name — имя автора, отображаемое на странице вывода комментария;
- 6) comment — содержание комментария.

Структура таблицы описаний интернет-ресурсов представлена в таблице 14.

Таблица «Описания интернет-ресурсов»

Поле	Тип
id	int(11)
value	text
value_lang	varchar(2)

Данная таблица предназначена для хранения описаний интернет-ресурсов. Описания хранятся в виде XHTML-представление в поле текстового типа. С данным полем ассоциирован полнотекстовый индекс, позволяющий незатруднительно подключить полнотекстовый поиск по описаниям интернет-ресурсов.

Комментарии к полям:

- 1) id — идентификатор описания;
- 2) value — содержание описания;
- 3) value_lang — двухсимвольный код языка, на котором создается описание.

Структура таблицы фолксономического справочника-рубрикатора представлена в таблице 15.

Таблица «фолксономический рубрикатор»

Поле	Тип
id	int(11)
title	varchar(60)
usage	int(11)

В данной таблице хранятся так называемые «теги»: фолксономические метки, используемые пользователями для категоризации контента. Для каждой метки определено поле «usage», указывающее на количество использований данного тега. Как только значение этого поля становится равным нулю, тег автоматически удаляется.

Комментарии к полям:

- 1) id — идентификатор тега;

2) title — название тега;

3) usage — количество связей ресурсов с тегом.

Структура таблицы таксономического справочника представлена в таблице 16.

Таблица 16

Таблица «таксономический рубрикатор»

Поле	Тип
id	int(11)
pid	int(11)
title_en	varchar(150)
title_ru	varchar(150)

Данная таблица наполняется администрацией каталога и содержит перечень иерархических разделов каталога. Для организации вложенности разделов используется подход, известный как «id-parentid».

Комментарии к полям:

1) id — идентификатор раздела;

2) pid — идентификатор родительского раздела;

3) title_en, title_ru — наименования разделов соответственно на английском и на русском языках.

Структура таблицы «Обратные ссылки» представлена в таблице 17.

Таблица 17

Таблица «Обратные ссылки»

Поле	Тип
id	int(11)
rid	int(11)
url	varchar(150)
atf	smallint(6)
r_pagerank	smallint(6)
r_static	smallint(6)
ext_links	smallint(6)
sys_date_lastchecked	int(11)
sys_enabled	(1)
linkid	smallint(6)

Данная таблица необходима для учета установки обратных ссылок поль-

зователями каталога. В ней содержатся специфичные поля, такие как «atf» — «attempts to find» — счетчик количества неудачных попыток обнаружения ссылки (если значение превышает 3, запись автоматически удаляется). Также, в данной таблице хранятся данные необходимые для расчета рейтинга, выреченного при помощи размещения обратной ссылки.

Комментарии к полям:

- 1) id — идентификатор обратной ссылки;
- 2) rid — идентификатор интернет-ресурса;
- 3) url — адрес страницы с обратной ссылкой;
- 4) atf — количество попыток обнаружения ссылки на странице;
- 5) r_pagerank — рейтинг PageRank страницы с обратной ссылкой;
- 6) r_static — статический рейтинг для обратной ссылки;
- 7) ext_links — количество внешних ссылок на странице;
- 8) sys_date_lastchecked — время последней проверки обратной ссылки;
- 9) sys_enabled — доступность обратной ссылки (признак);
- 10) linkid — идентификатор рекламного материала, размещенного на странице.

Структура таблицы «Документы интеллектуальных модулей» представлена в таблице 18.

Таблица 18

Таблица «Документы интеллектуальных модулей»

Поле	Тип
id	int(11)
rid	int(11)
from_sid	int(11)
to_sid	int(11)
action_id	int(11)
content	text
date	int(11)
lifetime	int(11)
need_to_be_processed	(1)

Данная таблица предназначена для хранения документов, которые ис-

пользуют интеллектуальные модули в своей работе. У каждого документа есть свой владелец и адресат. Отдельной таблицы для хранения идентификаторов интеллектуальных модулей не существует (эти значения определены заранее и известны в программном коде).

Комментарии к полям:

- 1) id — идентификатор документа;
- 2) rid — идентификатор ресурса, связанного с документом;
- 3) from_sid — идентификатор интеллектуального модуля, создавшего документ;
- 4) to_sid — идентификатор интеллектуального модуля, который должен получить документ;
- 5) action_id — идентификатор действия (может быть разрешен через систему разрешения задач);
- 6) content — содержание документа (сериализованный массив);
- 7) date — дата создания документа;
- 8) lifetime — время «жизни» документа в секундах;
- 9) need_to_be_processed — признак обработки документа.

Структура таблицы рейтингов интернет-ресурсов каталога представлена в таблице 19.

Таблица 19

Таблица «Рейтинги интернет-ресурсов»

Поле	Тип
rid	int(11)
r_weighted	float
r_pagerank	tinyint(4)
r_value	tinyint(4)
r_correlation	tinyint(4)
r_expandable	tinyint(4)
r_static	tinyint(4)
r_fine	tinyint(4)

Таблица предназначена для хранения рейтингов интернет-ресурсов. Основной рейтинг, использующийся для ранжирования в каталоге —

взвешенный (название поля «r_weighted»), он рассчитывается на основе других компонент.

Комментарии к полям:

- 1) rid — идентификатор ресурса;
- 2) r_weighted — взвешенный рейтинг;
- 3) r_pagerank — Google PageRank рейтинг;
- 4) r_value — содержательный рейтинг;
- 5) r_correlation — соответствия рейтинг;
- 6) r_expandable — расходуемый рейтинг;
- 7) r_static — статический рейтинг;
- 8) r_fine — штрафной рейтинг.

Структура таблицы статистики посещений интернет-ресурсов представлена в таблице 20.

Таблица 20

Таблица «Статистика посещений интернет-ресурсов»

Поле	Тип
rid	int(11)
date	int(11)
hits	smallint(5)

В данной таблице хранятся данные о «хитах» (неуникальных посещениях интернет-ресурсов в каталоге). Данные хранятся в рамках суток. То есть минимальный интервал, который можно получить из этой таблицы должен быть обязательно кратен одним суткам.

Комментарии к полям:

- 1) rid — идентификатор ресурса;
- 2) date — метка времени (суточная) доступа к ресурсу;
- 3) hits — количество просмотров страниц.

Структура таблицы с информацией о пользователях каталога представлена в таблице 21.

Таблица «Пользователи каталога»

Поле	Тип
id	int(11)
email	varchar(150)
password	varchar(40)

В данной таблице хранится информация, необходимая для авторизации пользователей. В качестве имени пользователя (логина) используется адрес электронной почты. Пароль хранится в виде хеш-отображения по собственному алгоритму, основанному на преобразованиях MD5 и SHA-1.

Комментарии к полям:

- 1) id — идентификатор пользователя;
- 2) email — адрес электронной почты пользователя;
- 3) password — хеш пароля пользователя (отображение MD5).

Общая схема базы данных с указанием связей между таблицами не приведена специально по следующей причине: связи между таблицами и ссылочная целостность не может гарантироваться на уровне СУБД [13] ввиду ограничений хостинга (используется MySQL 4 версии, тип таблиц MyISAM). Ссылочная целостность отслеживается на уровне бизнес-логики приложения.

Для работы с СУБД используется слой абстракции DbSimple2 (разработка Д. Котерова), выравнивающий диалекты вызовов системных функций и добавляющий ряд функциональных возможностей, которые могут понадобиться при развитии проекта, такие как ситуационное и временное кэширование запросов, использование технологии memcache и т. п.

2.6. Программное обеспечение

2.6.1. Программная платформа каталога

В качестве программной платформы для построения каталога используется «Независимое ядро» («Independent kernel», <http://ikernel.org>). Данное ядро предназначено для построения современных информационных систем для сети Интернет и включает в себя множество готовых решений, не

требующих дополнительных разработок. Эта платформа идеально подходит для построения интернет-каталога на ее основе за счет своей простоты, использования объектно-ориентированного подхода вкупе с компонентным подходом. Ядро включает в себя ряд классов для работы с СУБД, кэшированием на любом уровне, сайтами (страницы, шаблоны вывода, интеграция с модулем Apache rewrite и т. п.) Платформа работает на основе технологии клиент-сервер и написана на языке PHP пятой версии. Само ядро является многокомпонентным, причем управлением необходимыми компонентами занимается разработчик (что позволяет оптимизировать скорость работы платформы для собственных нужд за счет отключения ненужных компонент), разрешением зависимостей занимается сама платформа.

Для лучшего представления работы платформы следует привести пример отображения страницы сайта (с подключением и вызовом необходимых компонент) и описать его. Следует сразу заметить, что пример упрощен для улучшения восприятия (рис. 10).

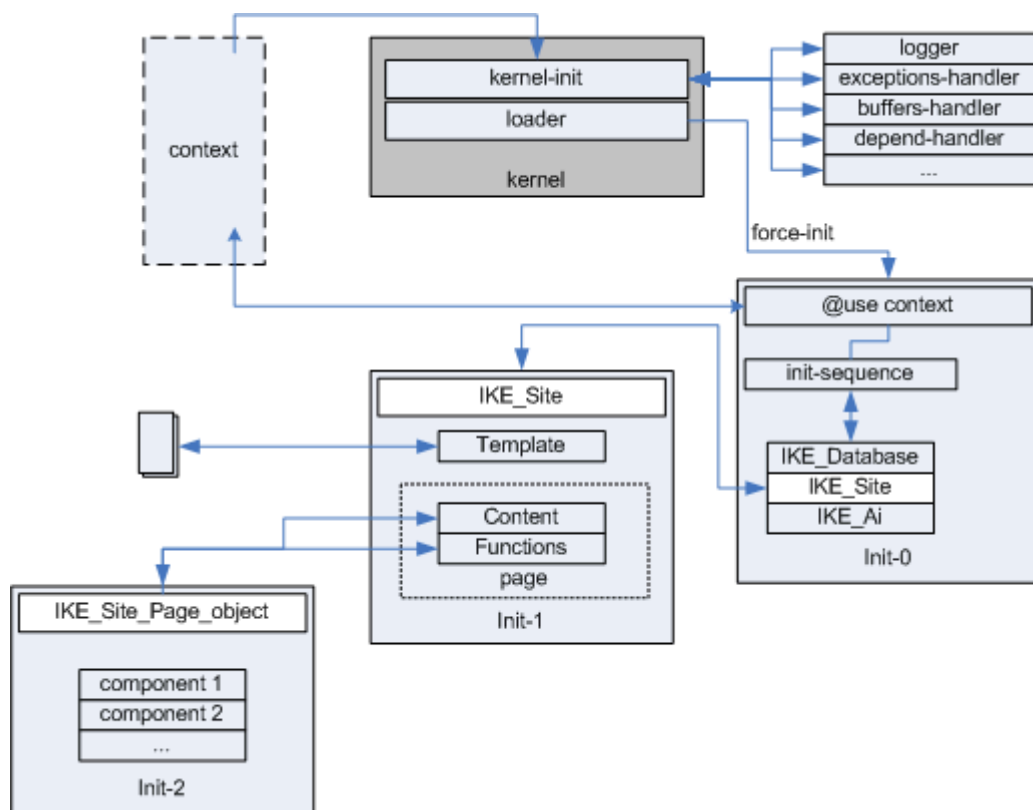


Рис. 10: Архитектура платформы

Изначально пользователем формируется HTTP-запрос (вида GET или

POST). Веб-сервер обрабатывает запрос и перенаправляет его в контекст вызова (context). После чего происходит инициализация ядра (kernel-init); в данное время подгружаются необходимые компоненты уровня ядра, которые являются неотъемлемыми и обеспечивают его корректное функционирование, к таким компонентам относятся: логирование действий, управление перехватом исключений, управление буферами, разрешение зависимостей и ряд других. После инициализации компонент ядра управление передается основному загрузчику (loader). Основной загрузчик запускает процедуру инициализации; в соответствии с контекстом определяется тип запроса и перечень необходимых к загрузке компонент. На данном этапе в память уже загружен модуль, отслеживающий зависимости между компонентами, поэтому разработчику не нужно их отслеживать самостоятельно. Управление передается последовательно каждому необходимому компоненту. Рассмотрим, в данном случае, сложный компонент «Интернет-сайт». При вызове данного компонента в соответствии с контекстом определяется необходимый шаблон и страница сайта. Под страницей понимается отдельный компонент, агрегирующий функции и информационное наполнение. После передачи управления компоненту страницы сайта, управление возвращается в вызванный модуль «Интернет-сайт», полученные данные объединяются с шаблоном вывода, а полученный результат записывается в буфер вывода ядра. После чего управление передается следующим модулям. По окончании выполнения всех модулей буфер вывода передается как ответ веб-сервера, затем его содержимое очищается и платформа завершает свою работу.

2.6.2. Классы каталога ресурсов интернет

Как было отмечено ранее, платформа использует объектно-ориентированный подход. Поэтому для создания каталога необходимо разработать ряд классов, позволяющих осуществлять необходимые для корректной работы каталога функции. Детальная архитектура классов не описывается по причине большого количества методов и свойств каждого класса. Основная задача — дать общее укрупненное представление об

архитектуре каталога. Применительно к каталогу ресурсов интернет, для платформы разработан следующий ряд классов.

`IKE_Ai_Specialist_Abstract` — абстрактный класс, предоставляющий определение общего интерфейса для всех интеллектуальных модулей каталога.

`IKE_Ai_Document` — класс, предоставляющий набор свойств и методов для создания документов, используемых интеллектуальными модулями.

`IKE_Ai_Captcha` — класс, предоставляющий возможность генерирования уникального изображения для прохождения теста Тьюринга (для защиты от автоматических регистраций).

`IKE_Ai_Specialist_RegistrationManager` — класс, предоставляющий методы для работы с регистрацией интернет-ресурсов в каталоге: регистрация сайтов, обновление описаний, управление публикацией и т. д.

`IKE_Ai_Specialist_AdvertsManager` — класс, предоставляющий методы для отображения и управления рекламными объявлениями на страницах каталога.

`IKE_Ai_CensorshipManager` — класс, предоставляющий методы для отслеживания недоброкачественного содержания на страницах каталога с возможностью установки штрафов, блокировки ресурса или его удаления.

`IKE_Ai_FeedbackManager` — класс, предоставляющий методы для осуществления обратной связи с пользователями и администратором каталога.

`IKE_Ai_RatingManager` — класс, предоставляющий методы для работы с рейтингом интернет-ресурса: расчет компонент рейтинга при помощи некоторых эвристических алгоритмов, расчет взвешенного рейтинга, установка рейтингов и ранжирование.

`IKE_Ai_RelationshipManager` — класс, предоставляющий методы для работы с таксономическим и фолксономическим рубрикатором, включающим в себя: методы для добавления тегов и разделов, выборки интернет-ресурсов по заданному тегу или разделу.

`IKE_Ai_SearchManager` — класс, предоставляющий методы для поиска интернет-ресурса в базе данных каталога, используется полнотекстовый метод поиска с возможностью задания критериев поиска непосредственно в строке

запроса (т. н. MySQL fulltext boolean search).

IKE_Ai_StatisticsManager — класс, предоставляющий методы для учета посещений (просмотров) страниц каталога: в данное время работает лишь учет просмотров страниц (т. н. хитов), также, позволяет выводить информацию о посещениях и использовать ее при расчете рейтинга.

IKE_Ai_WebDocumentsManager — класс, предоставляющий методы для работы с интернет-документами: получение документов, формирование запросов, получение значения PageRank™, оценка размера документов и интеграция с внешними сервисами.

IKE_Ai_LinguaStemmer — класс, предоставляющий метод для получения нормальной формы слова (стеммер), позволяющий работать на 2 языках: русском и английском.

Перечень классов, не разработанных специально для каталога, но использующихся в каталоге, представлен ниже.

IKE_Data_Directory_AiOperations — класс предоставляющий методы для «решателя» базы знаний на основе продукционно-алгоритмической модели: позволяет определять перечень возможных условий и выполнять продукции внутри контекста.

IKE_Data_Directory_Database (наследованный от DbSimple2) — класс, предоставляющий уровень абстракции от диалекта вызовов функций СУБД в языке PHP, предоставляющий единый интерфейс для работы с различными СУБД, а также возможности для логирования запросов и их кэширования.

IKE_Data_OwnerAlert — класс, необходимый для уведомления администратора каталога о возникновении критических ошибок.

IKE_Data_PageRankGrabber - низкоуровневый класс, необходимый для получения значения рейтинга PageRank™.

IKE_Types_String_Utf8 — позволяет использовать кодировку UTF-8 на страницах сайта прозрачным способом.

Основной класс IKE_Site — предоставляет возможность использовать шаблоны, страницы, URL, управлять навигационным меню. Является основным

классом для построения интернет-ресурсов. В рамках класса `IKE_Site` существует ряд подклассов, отвечающих за различные функции: отображение страниц — `IKE_Site_Module_Page`, обработка запросов — `IKE_Site_Module_Parse`, подготовка к отображению — `IKE_Site_Module_Prepare`.

2.6.3. Требования к программному обеспечению

При проектировании и начальной разработке каталога (локальная разработка) используется рабочая станция со следующим программным обеспечением: среда разработки Eclipse, веб-сервер Apache 2.x, СУБД MySQL, интерпретатор PHP 5 версии [14] – под управлением ОС Linux (Linux Ubuntu).

На этапах тестирования, не включающих начальное тестирование и этапе эксплуатации каталога используется веб-сервер со сходными характеристиками ПО:

- ОС Debian GNU/Linux (или любой UNIX-подобный аналог);
- веб-сервер Apache 2.x вкуче с проксирующим веб-сервером nginx;
- СУБД MySQL 4.1;
- интерпретатор PHP 5.

2.7. Техническое обеспечение

При проектировании и начальной разработке каталога используется рабочая станция со следующими характеристиками:

- процессор Intel Pentium IV HT 3.0 Ghz;
- оперативная память 1024 Mb RAM;
- жесткий диск 120 Gb HDD;
- монитор LG L1952HQ;
- клавиатура, манипулятор «мышь».

На этапах внешнего тестирования и эксплуатации необходим веб-сервер, предоставленный хостинг-компанией со следующими характеристиками:

- частота процессора не менее 1GHz (с возможностью потребления не менее 5% времени);
- оперативная память не менее 1024 Mb RAM (с расчетом на shared-

вариант хостинга);

- RAID-система;
- Дополнительные характеристики:
- выделенное дисковое пространство не менее 250 Мб;
- входящий трафик не менее 1Gb/мес.;
- исходящий трафик: не менее 10Gb/мес.;
- не менее 1 БД MySQL.

Далее, в разделе «Обзор и выбор хостинговой компании» будет произведен выбор провайдера услуг хостинга, удовлетворяющего данным характеристикам.

2.8. Интерфейсы каталога

Для отображения информации для пользователей, администратора каталога, специализированных «сетевых роботов» и ИПС используется ряд интерфейсов, которые будут описаны ниже. Далее по тексту будет приведен ряд изображений экранных форм. Все экранные формы отобразить в тексте дипломного проекта не предоставляется возможным ввиду их большого количества.

2.8.1. Интерфейс рядового пользователя каталога

Интерфейс проектируется исходя из функциональных возможностей и потребностей пользователей. В данном случае в интерфейсе должны быть отражены следующие возможности:

- главная навигационная страница (см. рис. 11);
- быстрый доступ к таксономическому справочнику;
- быстрый доступ к наиболее популярным тегам;
- форма поиска по каталогу;
- информация о последних зарегистрированных ресурсах;
- возможность доступа к справочному разделу;
- возможность доступа к другим страницам каталога.

Быстрый доступ к таксономическому справочнику подразумевает наличие на главной странице перечня иерархических разделов первого уровня, после перехода в один из разделов отображается страница с перечнем подразде-

лов и интернет-ресурсов, опубликованных в данном разделе.

Быстрый доступ к наиболее популярным тегам подразумевает наличие на главной странице так называемого «облака тегов»: особого вида представления структуры с учетом весов его элементов. Теги различаются размером (размером кегля шрифта) и цветом (используются градации серого): наиболее популярные теги отображаются большим кеглем и более темным цветом. Таким образом, пользователи могут легко отделить популярные теги от менее популярных.

Форма поиска по каталогу размещается в первой (видимой на разрешении монитора 640x480) части страницы и содержит поле для ввода поискового запроса, кнопку осуществления запроса и краткий комментарий (пример запроса).

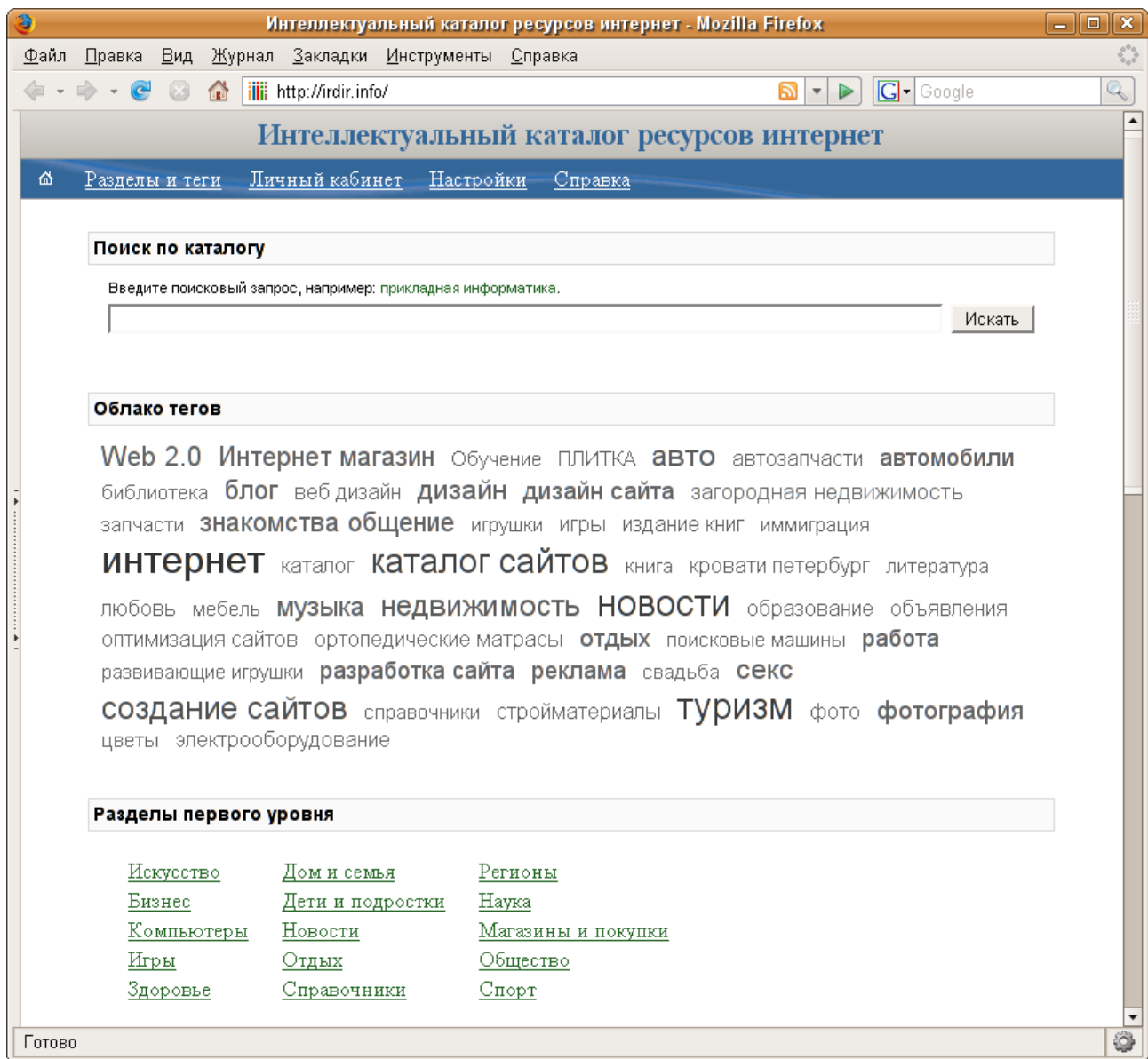


Рис. 11: Первая часть главной страницы каталога

Блок «информация о последних зарегистрированных ресурсах» содержит три компонента: последние зарегистрированные ресурсы в обратном хронологическом порядке, последние зарегистрированные ресурсы с наибольшим взвешенным рейтингом и интернет-ресурсы с наибольшим взвешенным рейтингом. Для каждой компоненты отображаются следующие данные: заголовок интернет-ресурса, его взвешенный рейтинг и дата регистрации (в наиболее понятном для пользователя в формате: в виде количества месяцев/недель/дней/часов/минут, прошедших с момента регистрации).

На главной странице должна быть размещена ссылка (в меню навигации) на справочный раздел каталога, в котором детально описаны технологиче-

ские особенности каталога и комментарии по использованию элементов интерфейса каталога.

Возможность доступа к другим страницам каталога подразумевает наличие навигационного меню, позволяющего переходить к различным страницам. Для отображения некоторых элементов этого меню следует использовать технологию асинхронного доступа к веб-серверу (т. н. AJAX). В частности, такой подход будет полезен для быстрого отображения перечня разделов и облака тегов на любой из страниц каталога (рис. 12).

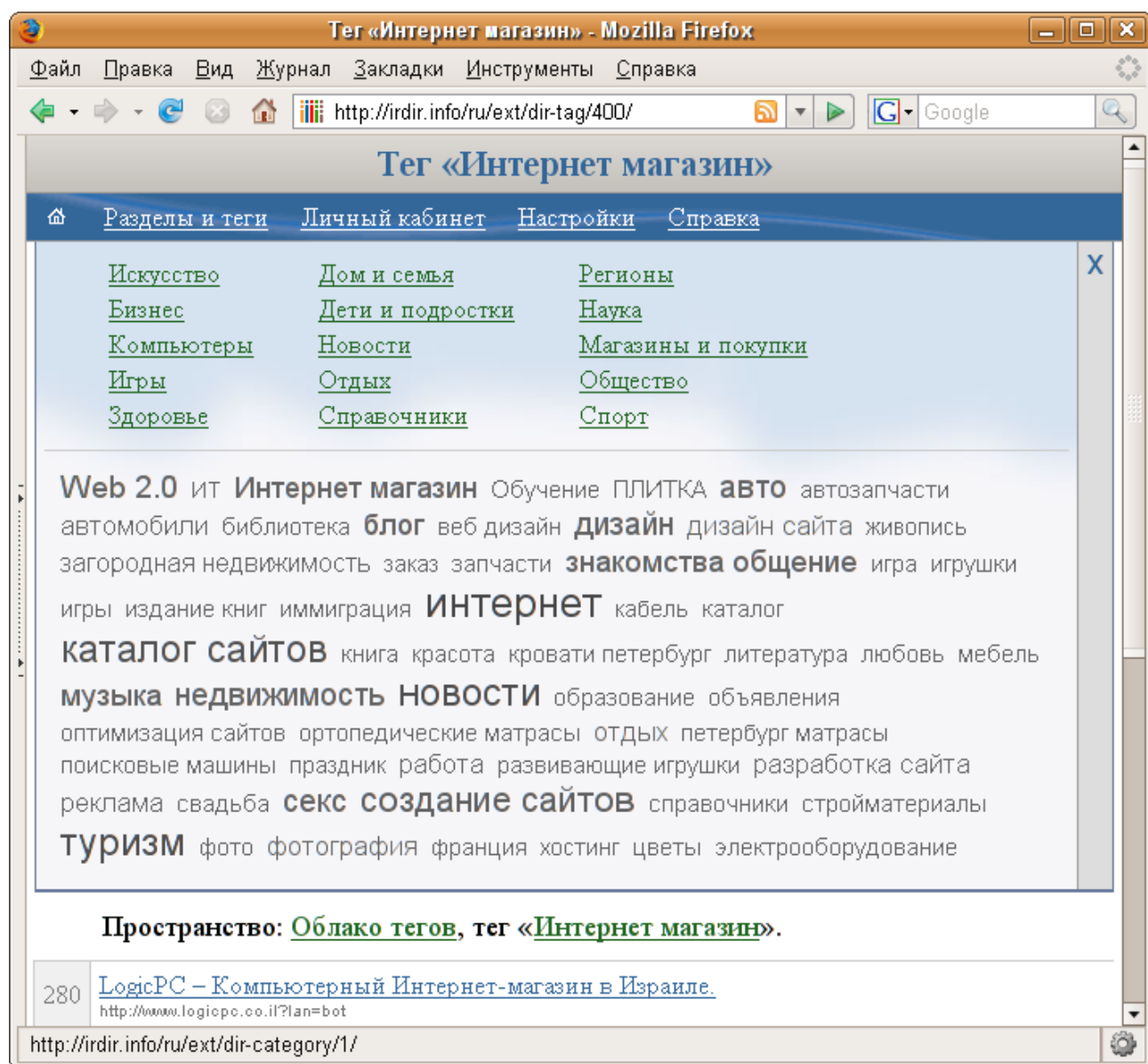


Рис. 12: Использование AJAX на страницах каталога

Также, необходимы следующие элементы интерфейса и страницы:

— страница поиска;

- базовая форма поиска в каталоге;
- страница результатов поиска (SERP) с возможностью постраничной навигации;

Страница поиска открывается при осуществлении поискового запроса либо с главной страницы каталога, либо при использовании интерфейса OpenSearch. При осуществлении запроса повторно отображается форма поиска, с уже введенным (текущим) поисковым запросом. А также результаты поиска (search engine results page). Внизу страницы отображается перечень страниц с результатами. На каждой странице выводится не более 15 результатов (найденных интернет-ресурсов). Для перехода к следующим страницам результатов можно использовать, как манипулятор «мышь», осуществив переход на интересующую страницу, так и клавиатуру, нажав одновременно кнопку «Ctrl» и стрелку вправо или влево, для осуществления последовательной постраничной навигации (см. рис. 13).

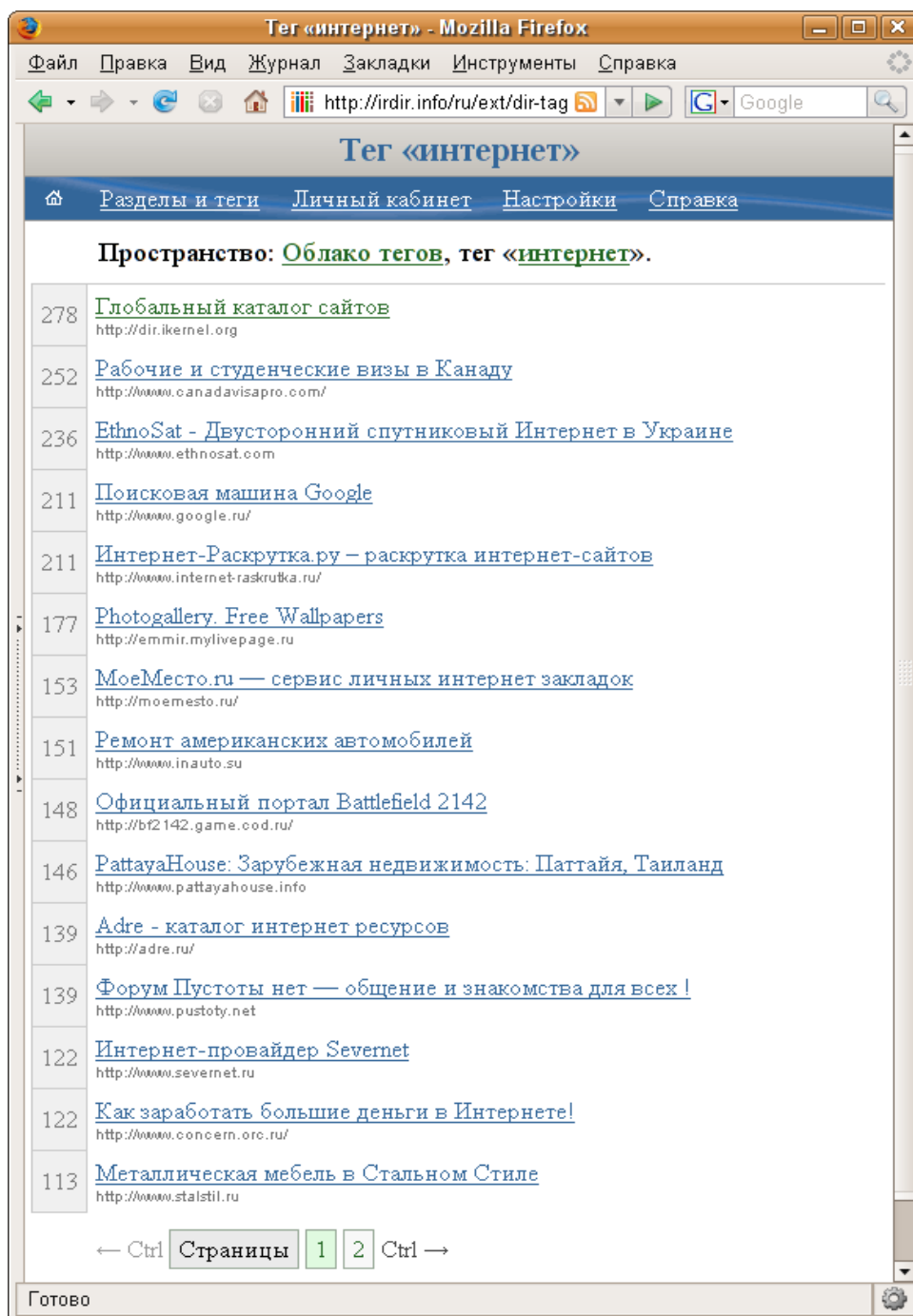


Рис. 13: Перечень интернет-ресурсов (SERP)

Страницы для фолксономической категоризации:

- страница фолксономического рубрикатора;
- страница «облако тегов», агрегирующая все теги каталога;
- страница для определенного тега с возможностью постраничной навигации.

Страница фолксономического рубрикатора по своей структуре аналогична странице поиска за следующим исключением: при начальном доступе к стра-

нице отображается т. н. «облако тегов».

Страницы для таксономической категоризации:

- страница таксономического рубрикатора;
- страница, отображающая таксономические разделы всех уровней;
- страница для отображения определенного раздела с возможностью постраничной навигации и возможностью перемещения по уровням иерархической структуры в обратном порядке (т. н. «хлебные крошки»).

Данная страница аналогична странице поиска за следующими исключениями: в том случае, если текущий раздел содержит подразделы, они должны быть выведены до результатов (SERP). Также, до результатов должны быть выведены «хлебные крошки» — навигационное меню, предназначенное для навигации по иерархическим разделам вертикально.

— Страница с описанием интернет-ресурса содержит следующие компоненты (элементы интерфейса):

- заголовок интернет-ресурса;
- описание интернет-ресурса;
- ссылка на интернет-ресурс;
- информация о рейтингах ресурса;
- скриншот интернет-ресурса;
- таксономические разделы ресурса;
- фолксономические теги ресурса;
- модуль интеграции с сервисами социальных закладок;
- отображение комментариев пользователей.

Данная страница является одной из наиболее важных компонент интерфейса каталога (см. рис. 14). Она обязательно должна содержать заголовок описываемого интернет-ресурса и он должен быть четко выделен относительно остального содержания. Описание располагается ниже заголовка, при этом, если владелец сайта опубликовал несколько описаний, отображается одно из них, выбранное равномерно случайно.

Ссылка на интернет-ресурс должна быть легко доступной для пользователей. Желательно представить ее в трех видах: простой гиперссылкой с URL

сайта, навигационной кнопкой «Посетить сайт» и ссылкой со скриншота сайта.

Информация о рейтингах интернет-ресурса должна быть представлена таблично, она носит дополнительный информационный характер. Взвешенный рейтинг должен быть выделен относительно остальных компонент. В том случае, если интернет-ресурс содержит дополнительные штрафы, также следует выделить и штрафной рейтинг (красным цветом).

На странице должна быть навигационная панель, позволяющая производить семантическую навигацию: должны быть размещены ссылки на разделы, в которых описан интернет-ресурс и ссылки на теги, с которыми связан интернет-ресурс.

Некоторые пользователи используют сервисы социальных закладок. Эти сервисы позволяют хранить закладки на интересующие их пользователей сайты без привязки к определенному компьютеру. Для таких пользователей необходимо разместить панель, при помощи которых они могли бы легко добавить интернет-ресурс в свои сети социальных закладок.

Недвижимость. Из рук в руки. - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

http://irdir.info/ru/ext/dir-resource/88/

Google

Недвижимость. Из рук в руки.

Разделы и теги Личный кабинет Настройки Справка

Описание интернет-ресурса

Издательская группа ИЗ РУК В РУКИ предлагает вашему вниманию internet-версии московского, Санкт-Петербургского и белорусского выпусков еженедельного журнала Фотонедвижимость. Из рук в руки.

Вы сможете просмотреть объявления о продаже и аренде квартир, продаже новостроек, коммерческой, зарубежной недвижимости, опубликованные в последнем, свежем, выпуске еженедельника выбранного вами региона (все объявления проиллюстрированы цветными фотографиями), а также получить много другой полезной информации.

Семантическая навигация

Данный интернет-ресурс описан в следующих разделах:

- [Дом и семья](#), подраздел [Недвижимость](#).

Комментарии пользователей

В настоящее время этот интернет-ресурс не содержит ни одного комментария. Вы можете [добавить комментарий](#).

Комментарии в виде [RSS-ленты](#). Элементы Дублинского ядра — [Dublin core RDF](#).


Готово

Адрес сайта:
<http://www.realty-photo.ru/>

Семантические теги:
[недвижимость](#), [квартиры](#).

Социальные закладки

Скриншот сайта



Посетить

Рейтинги ресурса [?]

Взвешенный	WR	345
Google PageRank™	PR	5
Содержательный	VR	33
Соответствия	CR	58
Расходуемый	ER	0
Статический	SR	0
Штрафной	FR	1

Интеллектуальная модерация

[Пожаловаться на спам](#)

Также, на странице описания ресурса должен отображаться перечень комментариев, оставленных пользователями каталога. В том случае, если к описанному интернет-ресурсу нет комментариев, должна отображаться ссылка, позволяющая добавить комментарий при переходе по ней.

2.8.2. Интерфейс владельца сайта

Интерфейс владельца сайта должен позволять регистрировать учетные записи пользователей в каталоге, при помощи которых впоследствии могут осуществляться регистрации интернет-ресурсов и последующее управление ими.

Страница авторизации должна содержать поля для ввода идентификационных данных: e-mail пользователя и его пароль. Также, должна существовать возможность восстановления пароля через электронную почту (путем верификации заявки на смену пароля).

Страница регистрации пользователя должна содержать три поля: e-mail пользователя и два поля для ввода пароля (второе поле для исключения ошибок).

После регистрации или авторизации владельца сайта, пользователю должен предоставляться доступ к основной странице управления его учетной записью, при помощи которой он может производить следующие действия:

- управление сайтами;
- регистрация нового интернет-ресурса;
- перечень всех зарегистрированных ресурсов этим пользователем;
- управление учетной записью;
- изменение пароля пользователя;
- удаление учетной записи пользователя;
- обратная связь;
- возможность отправки сообщения об ошибке в каталоге;
- обратная связь с администрацией каталога.

Для каждого описанного выше действия должна существовать персональная страница или же группа страниц.

При регистрации нового интернет-ресурса пользователь должен заполнить первичную форму, содержащую ряд полей:

- URL-адрес сайта,
- заголовок сайта (Title),
- основной язык сайта (выбор из перечня);

— также, необходимо указать проверочный код для прохождения теста Тьюринга, позволяющего избежать автоматических регистраций ресурсов в каталоге.

При просмотре перечня интернет-ресурсов, зарегистрированных данным пользователем, отображается следующая информация:

- идентификатор сайта;
- статус публикации в каталоге;
- URL и язык сайта;
- заголовок сайта;
- перечень возможных действий, содержащий:
 - управление интернет-ресурсом в каталоге;
 - удаление интернет-ресурса из каталога.

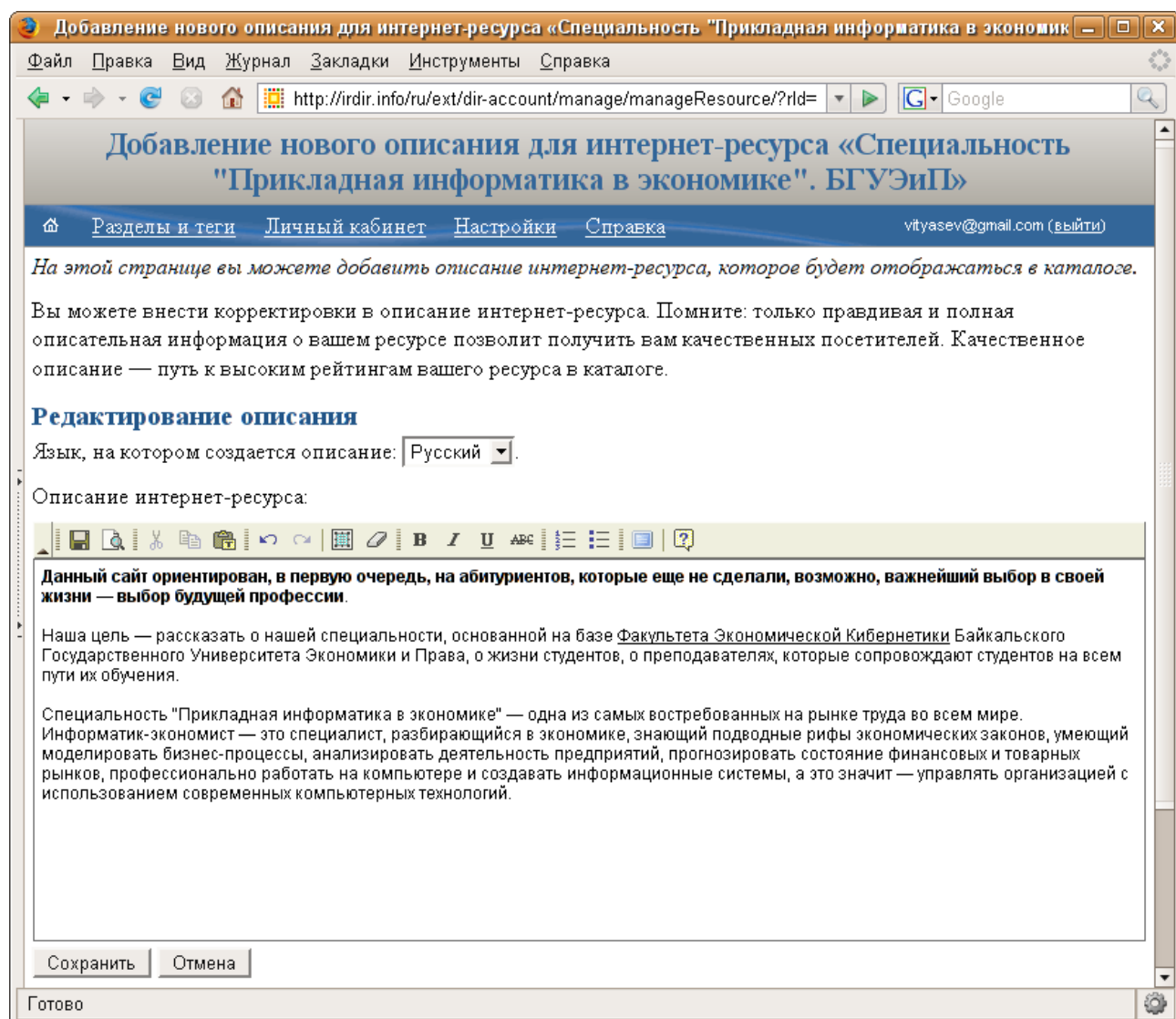
На странице управления интернет-ресурсом доступны следующие действия (на каждое из действий должна быть собственная страница или группа страниц):

- общие функции управления интернет-ресурсом;
- управление заголовком интернет-ресурса;
- управление описаниями;
- управление разделами (таксономия);
- управление тегами (фолксономия);
- управление публикацией ресурса;
- заявка на разрешение публикации;
- запрет публикации в каталоге;
- управление рейтингом ресурса;
- анализ рейтинга интернет-ресурса;
- обмен ссылками для повышения рейтинга.

На странице управления заголовком интернет-ресурса можно изменять заголовок в текстовом поле и сохранять изменения.

Управление описаниями позволяет задать для интернет-ресурса до четырех различных описаний, на основной странице выводится перечень отформатированных описаний с возможностью управления ими или их удаления.

Под управлением подразумевается возможность редактирования этих описаний в визуальном RichText-редакторе, работающем в режиме онлайн (рис. 15).



Заявка на разрешение публикации интернет-ресурса в каталоге делает запрос к интеллектуальному модулю на предмет возможности осуществления такого действия. В том случае, если ресурс может быть опубликован (согласно базе эвристических правил), пользователю выводится сообщение об этом факте и интернет-ресурс становится доступным в каталоге. В обратном случае пользователю выводится причина, по которой заявка была отклонена.

В том случае, если пользователю необходимо временно заблокировать к показу интернет-ресурс в каталоге, а удалять его нецелесообразно, можно воспользоваться функцией запрета публикации. Для этого на специальной странице необходимо подтвердить свое намерение запретить публикацию. Для восстановления предыдущего состояния необходимо воспользоваться функцией формирования заявки на разрешение публикации.

Пользователь может анализировать рейтинг своего интернет-ресурса и частично (ограниченно, строго в определенных пределах) влиять на его значение. Для этого существует две специальных страницы.

Анализ рейтинга интернет-ресурса позволяет в табличном виде просматривать значения взвешенного рейтинга и всех его компонент, также, справа от значения приводятся комментарии, позволяющие понять причину формирования именно такого значения и определить воздействия, необходимые для повышения определенного компонента рейтинга.

Обмен ссылками для повышения рейтинга позволяет увеличить значение Статического рейтинга (Static rating) на определенную величину, вплоть до ста пунктов. Для этого необходимо на любой из страниц регистрируемого ресурса разместить рекламный материал (рекламирующий каталог). Все материалы имеют различные веса. Размер бонусного Статического рейтинга определяется исходя из ряда параметров: количества внешних ссылок на странице, уровня рейтинга PageRank страницы, вида рекламного материала. Разрешается размещать не более трех рекламных материалов для каждого зарегистрированного в каталоге интернет-ресурса. На основной странице отображается перечень размещенных рекламных материалов и размер полученного статического рейтинга с каждого размещенного материала.

2.8.3. Интерфейс администратора каталога

Интерфейс администратора каталога — минимальный. В нем нет дополнительных возможностей, связанных с модерацией интернет-ресурсов. Модерацией занимаются интеллектуальные модули каталога, но не администратор. Поэтому интерфейс администратора каталога содержит единственный (минимально необходимый) модуль: «Управление таксономическим рубрикатором».

Этот модуль позволяет создать иерархическую структуру каталога и изменять (дополнять) ее, по мере необходимости (рис. 16).

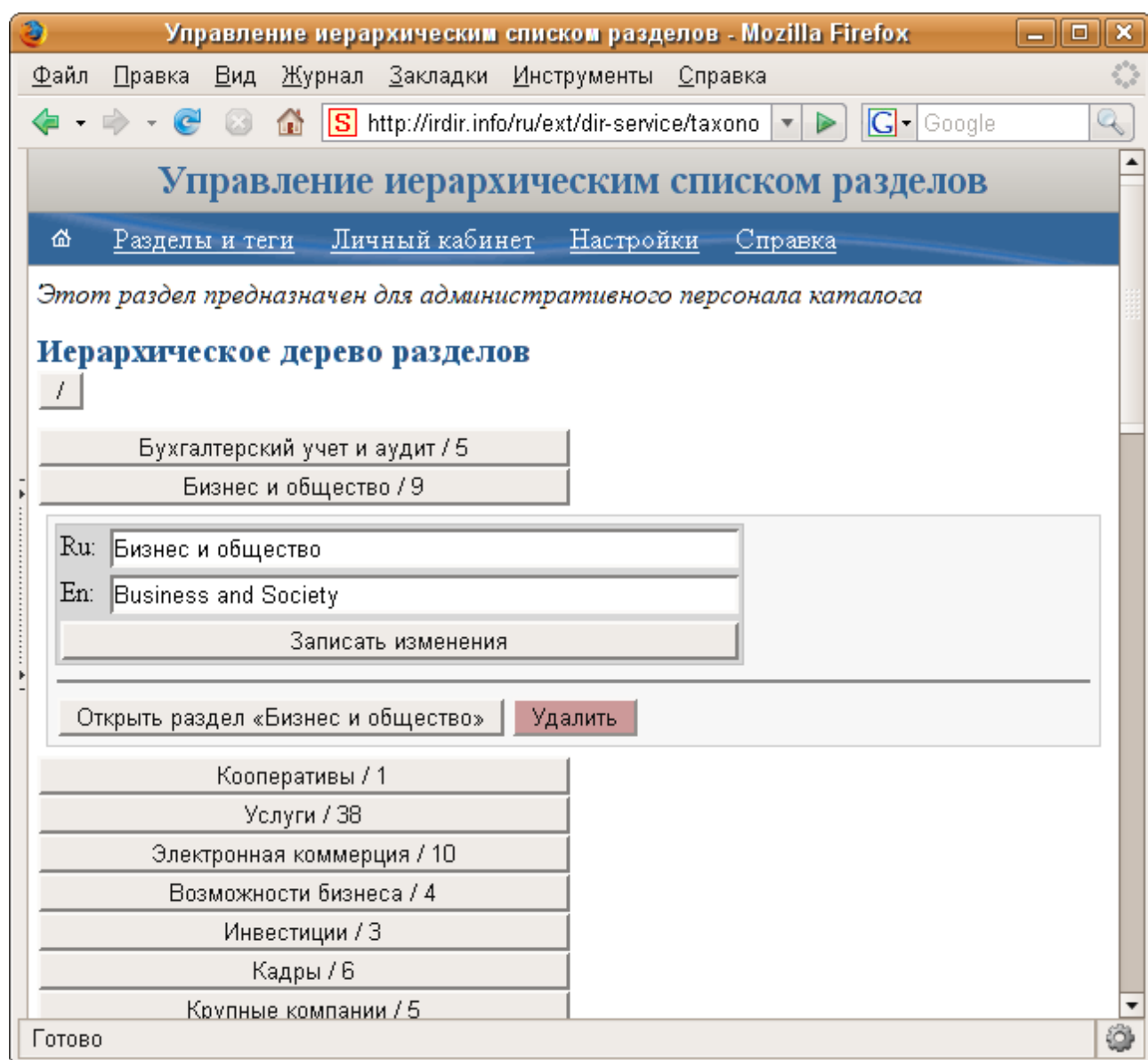
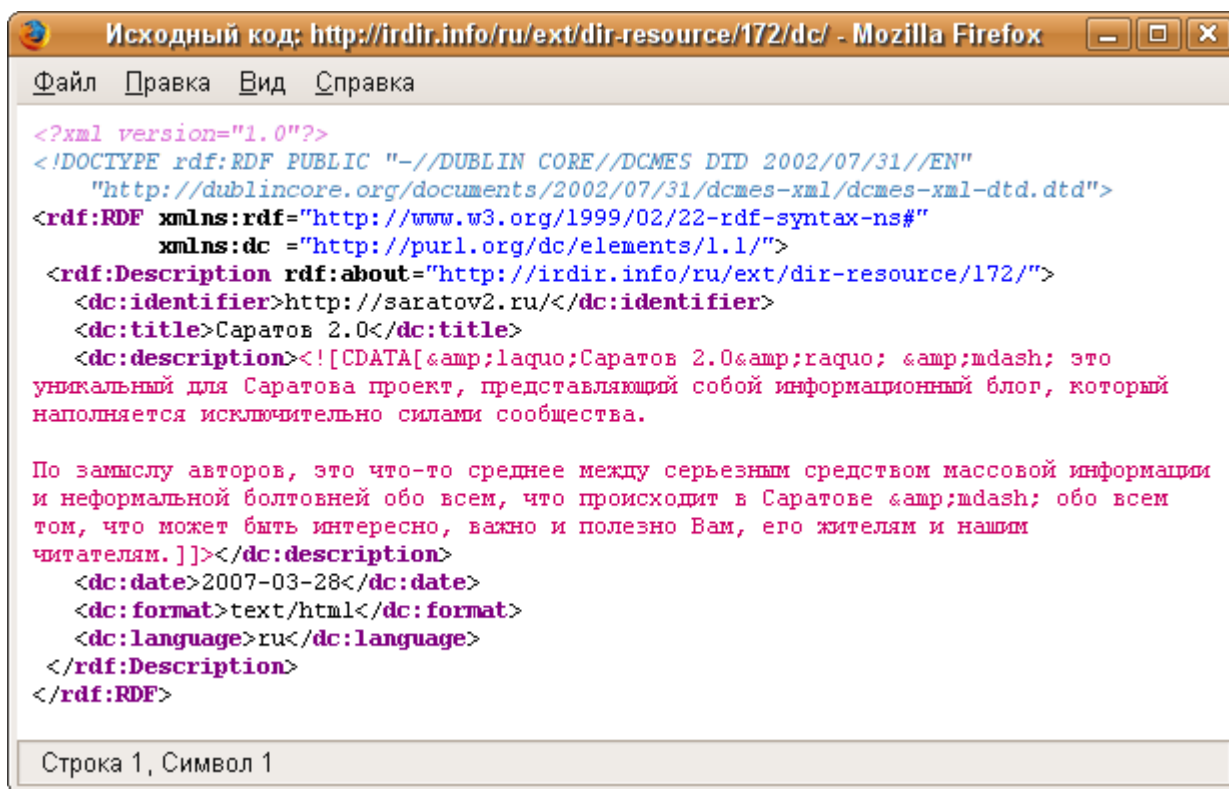


Рис. 16: Управление таксономическим рубрикатором

2.8.4. Интерфейс предоставления метаданных

Каталог использует модель метаданных «Дублинское ядро». Эти метаданные необходимо предоставлять для интеллектуальных агентов информационно-поисковых систем и для прочих ИС. Формат «Дублинское ядро» жестко регламентирован DCMI. В каталоге используется его основная (т. н. «неквалифицированная») модель. Данные предоставляются в RDF/XML-совместимом формате. На страницах каталога, где есть метаданные вставлен элемент «auto-discovery», позволяющий ИПС самостоятельно находить метаданные и обрабатывать их без применения дополнительных эвристических методов. Пример предоставляемых метаданных представлен на рис. 17.



```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF PUBLIC "-//DUBLIN CORE//DCMES DTD 2002/07/31//EN"
"http://dublincore.org/documents/2002/07/31/dcmes-xml/dcmes-xml-dtd.dtd">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://irdir.info/ru/ext/dir-resource/172/">
<dc:identifier>http://saratov2.ru/</dc:identifier>
<dc:title>Саратов 2.0</dc:title>
<dc:description><![CDATA[&laquo;Саратов 2.0&raquo; &mdash; это
уникальный для Саратова проект, представляющий собой информационный блог, который
наполняется исключительно силами сообщества.

По замыслу авторов, это что-то среднее между серьезным средством массовой информации
и неформальной болтовней обо всем, что происходит в Саратове &mdash; обо всем
том, что может быть интересно, важно и полезно Вам, его жителям и нашим
читателям. ]]></dc:description>
<dc:date>2007-03-28</dc:date>
<dc:format>text/html</dc:format>
<dc:language>ru</dc:language>
</rdf:Description>
</rdf:RDF>
```

Рис. 17: Пример предоставляемых метаданных в формате DC

2.8.5. Интерфейс открытого поиска

Каталог поддерживает технологию открытого поиска OpenSearch. Для этого существует специальный интерфейс на основе RDF/RSS 1.0 - «OpenSearch XML». Поиск по каталогу может быть интегрирован в любой браузер (см. рис. 18), поддерживающий технологию OpenSearch. Результаты поиска по каталогу могут быть использованы ИПС, поддерживающими этот формат.

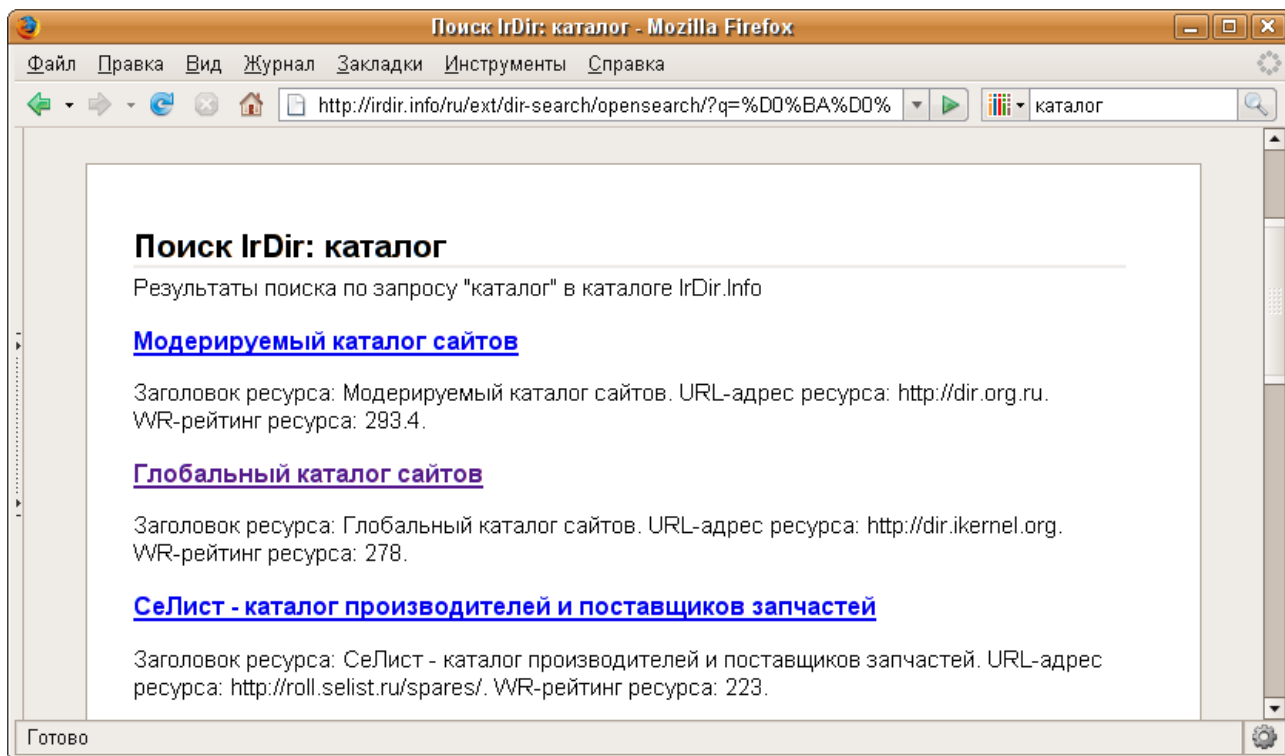


Рисунок 18: Интерфейс OpenSearch в Mozilla Firefox

2.8.6. Интерфейс новостных потоков

В каталоге существует множество интерфейсов новостных потоков. Практически на каждой странице каталога есть новостная лента. Использовать ее можно при помощи различных агрегаторов (самостоятельного ПО; ПО, интегрированного в браузер или же при помощи интернет-агрегаторов). Предоставляется следующая информация в формате RDF/RSS 1.0:

- 1) RSS-лента последних поступлений в каталог;
- 2) RSS-лента новых поступлений в определенный раздел;
- 3) RSS-лента новых поступлений по определенному тегу;
- 4) RSS-лента новых поступлений по определенному поисковому запросу.

Таким образом, пользователи могут отслеживать новые поступления интернет-ресурсов в каталог в соответствии с личными потребностями. Например, на рис. 19 показано использование одной из лент в RSS-агрегаторе Reader от компании Google, inc.

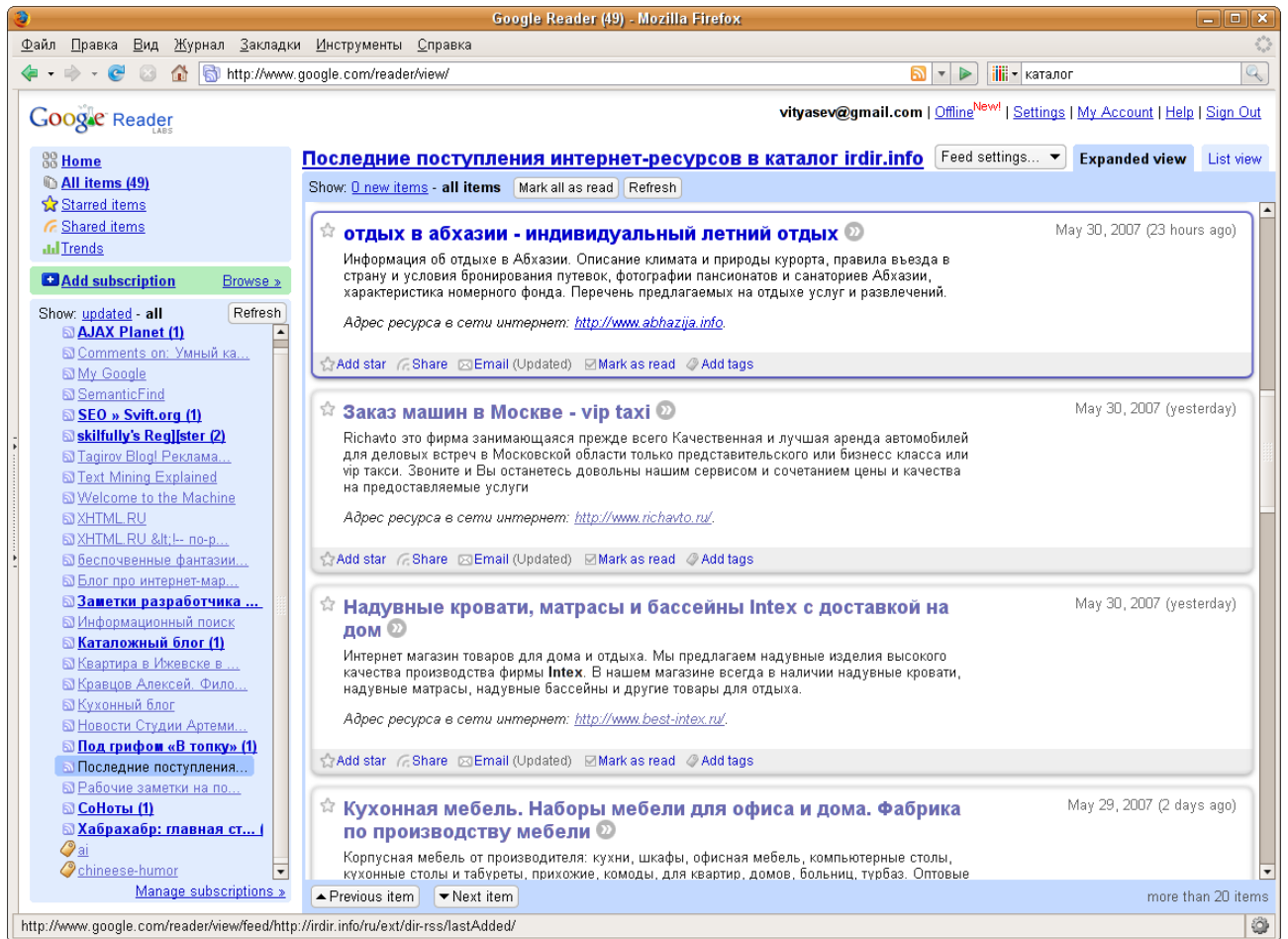


Рис. 19: Отображение ленты последних поступлений каталога в Google Reader

3. Размещение каталога в сети Интернет и его продвижение

3.1. Перечень этапов внедрения

Этапы внедрения обусловлены процессом разработки каталога и зависят от модели жизненного цикла каталога. Процесс состоит из нескольких этапов:

- 1) размещение каталога на внешнем интернет-сервере;
- 2) тестирование в закрытом режиме (альфа-версия);
- 3) тестирование в режиме эксплуатации (бета-версия);
- 4) эксплуатация.

После разработки всех необходимых функциональных модулей и подключения требуемых интерфейсов происходит локальное тестирование каталога и его отладка. По мере необходимости вносятся коррективы в разрабатываемые модули.

После этапа локального тестирования программная часть каталога размещается на интернет-сервере (см. подраздел 3.2).

Доступ к размещенному в интернет каталогу изначально ограничивается: на данном этапе важно обеспечить доступ к реализации ограниченному числу лиц (альфа-тестерам). Желательно, чтобы эти лица разбирались в предметной области и умели пользоваться сходными продуктами. Для подобного проекта достаточно пяти альфа-тестеров. Этим специалистам предлагается воспользоваться всеми доступными функциями каталога и написать отчет в свободной форме о проблемах, с которыми они столкнулись. На основе этой информации каталог дорабатывается, исправляются найденные ошибки.

После внесения корректив доступ к каталогу открывается всем желающим, рассылаются пресс-релизы, осуществляются публикации в интернет-СМИ. Основная цель данного этапа — собрать отзывы с большого количества пользователей (как профессиональных, так и не разбирающихся в предметной области) при помощи средств обратной связи, расположенных на страницах каталога. Эта информация позволит увидеть «сильные» и «слабые» стороны реализации и понять что стоит улучшить, в каком направлении следует развивать проект, что нужно исправить. Этап длится не более двух недель. На данном

этапе по прежнему не гарантируется целостность пользовательских данных.

По мере завершения этапа бета-тестирования каталог переходит на этап эксплуатации. Отключаются дополнительные возможности обратной связи, которые были необходимы на этапе бета-тестирования. Публикуется пресс-релиз о готовности каталога к эксплуатации.

3.2. Обзор и выбор хостинговой компании

Хостинг — услуга по предоставлению дискового пространства для физического размещения файлов сайта на сервере, постоянно находящегося в сети (обычно Интернет).

В услугу также может входить предоставление места для почтовой корреспонденции, баз данных и т. п., а также поддержка функционирования соответствующих сервисов.

Обычно предоставляется хостинговыми компаниями. Предоставлению хостинга, как правило, сопутствует услуга по регистрации домена.

Хостинг часто разделяется на платный и бесплатный. Обычно компания, предоставляющая бесплатный хостинг, зарабатывает путем показа рекламы на страницах, размещенных на нем. Частные лица для своих домашних страничек обычно используют бесплатный хостинг, а коммерческие организации — почти всегда платный хостинг. Общественные организации могут использовать как платный, так и бесплатный хостинг. Бесплатный хостинг, как правило, медленнее платного, предоставляет только базовые услуги и иногда ненадежен.

Также можно разделить услуги хостинга по типу предоставляемого ресурса:

— виртуальный хостинг — предоставляется место на диске и/или внешний трафик, среда исполнения веб-сервисов единая для многих пользователей.

— виртуальный выделенный сервер (VPS или VDS) — предоставляется место на диске, внешний трафик, часть общей памяти, процессорное время сервера, выглядит для пользователя как аренда целого сервера, но физически на одном реальном сервере располагается несколько вирту-

альных серверов. Обычно предназначен для проектов средней тяжести и начинающих реселлеров.

— выделенный сервер — предоставляется целый сервер с определенным дисковым пространством, памятью, процессорами и внешним трафиком. Используется для «тяжёлых» веб-проектов, которые не могут сосуществовать на одном сервере с другими проектами и требуют под себя все ресурсы сервера.

С юридической точки зрения, услуга хостинга относится к телематическим услугам связи или, в случае, когда хостинговая компания размещает у себя оборудование пользователя и обеспечивает его подключение к Интернету (колонкация) — к услугам передачи данных. Эти виды деятельности подлежат лицензированию. Лицензии выдаются Федеральной службой по надзору в сфере связи.

В виду высокой стоимости выделенных серверов будут рассматриваться только варианты виртуального хостинга, предоставляемые на платной основе.

Выделены следующие критерии, помимо требований к аппаратному и программному обеспечению, удовлетворяющие минимальным требованиям к виртуальному серверу и способу оплаты:

- Форма оплаты : система WebMoney;
- Платформа : UNIX;
- Сервисы : PHP; Cron; FTP-доступ; Доступ к .htaccess;
- Ежемесячная плата : не более 1000 рублей.

Поиск возможных площадок производился при помощи онлайн-сервиса HostObzor [10], были получены претенденты (см. табл. 22).

Обзор хостинг-площадок

Провайдер	План	Объем	Трафик	Цена (руб.)
UKRHOSTING	e-Commerce Gol	2000	н/о	985.25
UKRHOSTING	e-Commerce Gol	2000	н/о	985.25
robo-host.ru	RUS VIP2 (Dire	5000	н/о	980.00
CryoHosting	H5000	5000	н/о	980.00
RHP Company	Adult-5	400	30	952.00
eServer.ru	Максимум	3000	н/о	952.00
Adamant	Тариф А30	3000	н/о	947.24
WebXL	R3-Norm	5000	70	927.50
Mail.Ru	Максимальный	1500	н/о	924.00
mchost.ru	VIP-1	5000	н/о	900.00
MacHoster.Ru	Оптимум	8000	н/о	900.00
Hostraid	Raid	10000	н/о	900.00
DualHost	VIP	2000	н/о	900.00
TutHost	Гигабайт 3	3000	н/о	900.00
Rus-Host	EURO-5	500	н/о	900.00
bhost.ru	Максимальный	2500	н/о	900.00
Hostpro	Безлимитный	10000	н/о	898.50
host-web-site	ПЛАН 2002	2002	н/о	898.50
DinoHost	Gold	2500	н/о	892.50
HostZona.Ru	Vip	500	н/о	892.50
ГОСТ хостинг	тов. Берия	10000	н/о	870.00
Via	SEVEN	5000	н/о	870.00
ProstoHost	Medium-3	1500	н/о	870.00
ИНО.RU	elite + IP	800	н/о	870.00
Majordomo.ru	Профи	1500	н/о	870.00
PeterHost.Ru	VEGA	2000	н/о	870.00
DreamSee	(2)	2048	н/о	868.00
ZAHOSTI.RU	План VIP	3072	н/о	840.00
ArmHosting.Net	Special-Россия	4000	н/о	840.00
robo-host.ru	USA VIP2 (Dire	5000	н/о	840.00
FlyOne Telecom	FLY1500	1500	н/о	840.00
Fedora-Hosting.c	Core 6	3000	н/о	840.00
robo-host.ru	USA VIP2 (CPa	5000	н/о	840.00
AdvantA.org	R-2 (Россия)	2048	н/о	839.70
OH Web Hosting	E-commerce Pro	2048	51	838.60

На основе данного обзора выбор был сделан в пользу компании «Majordomo»: данная компания предоставляет качественный сервис и гарантирует безотказную работу своего оборудования, а также обеспечивает круглосуточную поддержку пользователей по электронной почте в любой день недели. При этом стоимость услуг, предоставляемых данной компанией является ниже средней рыночной стоимости по выдвинутым критериям.

3.3. Обзор и выбор методов продвижения

Одним из эффективных методов продвижения интернет-ресурсов в сети

интернет является реклама.

Под термином «интернет-реклама» обычно понимают: любую рекламу, размещаемую в Интернете; специфические формы рекламы, использующие технические особенности сети Интернет; рекламу интернет-ресурсов (прежде всего, веб-сайтов) независимо от того, где и как размещается сама реклама.

3.3.1. Баннерная реклама

Статические баннеры мало отличаются от традиционных рекламных форм, они полностью аналогичны рекламе в печатных СМИ. Наличие под баннером гиперссылки и возможность анимированного изображения не приводят к принципиальным отличиям такой рекламы. Качественные отличия начинаются при переходе от статических баннеров к системам баннерного обмена (СБО). Размещая на странице баннерообменный код, вебмастер не знает точно, каково будет содержимое баннера и его ссылки. У вебмастера бывает возможность лишь опосредованно управлять тематикой появляющихся баннеров через настройки СБО. Рекламодатель, напротив, не знает точно, на каких именно сайтах будет размещён его баннер. Подобное размещение рекламы «вслепую» было невозможно для традиционных, офлайн-методов рекламы. Таким образом, СБО, являясь удобным технологическим посредником между рекламодателем и рекламодателем, вносит дополнительные сложности в контроль рекламной деятельности.

3.3.2. Спам

Специфичность этого вида интернет-рекламы состоит не в анонимности рекламодателя и не в том, что спам — незапрошенная рассылка (многие виды рекламы являются незапрошенными). Особенность спама в том, что спамеры возлагают значительную часть затрат по доставке рекламы на потребителей и интернет-провайдеров, ничем это не компенсируя. Именно данная особенность сделала спам самой выгодной рекламой по соотношению затраты/отклик. Этот показатель для спама составляет порядка 0,01-0,05 \$/отклик, в то время как для других видов рекламы — порядка 1-10 \$/отклик. Спам носит массовый характер и встречается в виде почтовых рассылок,

сообщениях в форумах, деятельностью рекламных ботов в чатах и т. п.

3.3.3. Оптимизация для поисковых машин

Для некоторых типов сайтов поисковики приносят до половины и больше всех посетителей (то есть, потенциальных клиентов). Необходимым условием этого является присутствие ссылки в первых строках результатов поиска по наиболее популярным запросам. Поскольку результаты поиска обычно отсортированы по релевантности, перед оптимизатором стоит задача повысить релевантность кода веб-страниц к наиболее распространённым поисковым запросам.

3.3.4. Всплывающие (pop-up) окна и sruware

Аналогично спаму, для распространения используются ресурсы потребителя. Но метод не столь дешёв, как спам. К тому же, sruware во многих случаях признаётся вредоносной программой.

Просмотр рекламы за плату или подписка на рекламу. Этот метод не показал особой эффективности в сравнении с другими и в настоящее время используется мало.

3.3.5. Регистрация в каталогах

Этот вид рекламы не очень специфичен для Интернета — в офлайне тоже есть каталоги и справочники, внесение в которые даёт свою долю клиентов. Интернет-каталоги отличаются от офлайновых своим количеством, которое уже перешло в качество.

Очень давно, несколько лет назад, о каталогах сайтов люди даже не думали, а просто старались набивать свои ресурсы различной и интересной информацией, которая им и помогала привлекать новых пользователей на свои сайты.

Сейчас каталоги сайтов — это просто неотъемлемая часть интернета, как международного так и в каком-либо отдельном регионе. Мы знаем большое количество сборников ссылок, из них даже составляются целые базы, в которых находится по несколько тысяч каталогов, но только маленькая часть, а это при-

мерно процентов 2-5 из них созданы для людей, а не для поисковых машин.

3.3.6. Участие в рейтингах

Для некоторых сайтов этот метод приносит существенную часть клиентов. Для большинства — незначительную часть. К тому же, не существует добросовестных способов подняться в рейтинге за деньги. Поэтому данный способ не укладывается в типовую экономическую схему «деньги-реклама-клиенты-деньги». Тем не менее, его с натяжкой можно отнести к методам рекламы.

3.3.7. Реферальные и «партнёрские» программы

Хотя подобные методы рекламы и маркетинга (см., например, MLM) давно известны в офлайне, в Интернете несравненно удобнее учитывать рефералов и привлечённых клиентов. Поэтому можно сказать, что данный метод получил в Интернете новую жизнь.

3.3.8. Контекстная реклама

Контекстная реклама — вид динамического размещения интернет-рекламы, при котором рекламное объявление близко к контексту веб-сайта, где оно размещается. При этом может размещаться как баннер, так и текстовое сообщение.

Особую популярность приобрел частный вид контекстной рекламы — реклама на странице результатов поиска самой поисковой системы, называемая поисковой рекламой.

Так же популярно размещение контекстной рекламы на информационных сайтах и каталогах, когда рекламное объявление является частью содержания страницы.

Контекстная реклама вызывает больший интерес посетителей, чем иные виды рекламы, что выражается в более высоком индексе CTR. Поэтому данный вид рекламы — это возможность для рекламодателя показывать свое рекламное сообщение только нужным потенциальным клиентам (целевой аудитории).

Обычно для определения контекста и отбора объявлений используется движок той или иной поисковой машины.

3.3.9. Поисковая реклама

Поисковая реклама — частный случай контекстной рекламы, применяемый в поисковых системах. Отличительной особенностью является то, что выбор демонстрируемых рекламных сообщений определяется с учетом поискового запроса пользователя.

Оплата поисковой рекламы может основываться на разных принципах: по числу показов рекламного сообщения, по числу кликов пользователей поисковой системы, по принципу аукциона ключевых слов.

Данный вид интернет-рекламы относится к числу наиболее эффективных, поскольку тематика демонстрируемых рекламных сообщений максимально соответствует текущим интересам пользователя.

3.3.10. Выбор методов

Цель продвижения каталога в сети интернет — получение большего количества пользователей каталога и авторов сайтов, которые могут регистрировать свои интернет-ресурсы в каталоге. Ввиду низкой экономической эффективности методы продвижения «Поисковая реклама», «Контекстная реклама», «Баннерная реклама» использоваться не будут — стоимость таких методов сравнительно велика. Метод продвижения «Спам» является весьма спорным и рискованным, поэтому придется отказаться от его использования (несмотря на его высокую эффективность). Реклама при помощи pop-up (pop-under) окон раздражает пользователей, а партнерские программы не достаточно эффективны. Поэтому будут использованы следующие методы продвижения:

— Регистрация в каталогах и рейтингах — данный метод позволяет зарегистрировать интернет-ресурс вместе с его описанием в тысячах каталогов ресурсов интернет. При этом посетители могут быть привлечены как из самих каталогов, так и из поисковых систем — многие каталоги ставят прямые ссылки на регистрируемый интернет-ресурс, что позволяет получить более высокие позиции в выдаче поисковых систем за счет ссылочного ранжирования.

— Поисковая оптимизация интернет-ресурса — при грамотном

проектировании шаблонов вывода и информационной структуры каталога можно достичь высоких позиций в выдаче поисковых систем. В основном, речь идет о низкочастотных запросах (однако, при этом, каталог должен содержать большое количество релевантных страниц). По высокочастотным запросам продвигать каталог практически не имеет смысла – эта сфера занята так называемыми «черными» оптимизаторами, которые «контролируют» выдачу поисковых систем по определенным высокочастотным запросам – уровень конкуренции слишком высок.

— Публикация пресс-релизов в социальных сетях и специализированных интернет-СМИ — такой подход может обеспечить краткосрочный лавинообразный рост количества посетителей. Однако, некоторые из них могут добавить каталог в «избранное» своего браузера или же запомнить его адрес, если он покажется им полезным. Также, такой подход обеспечивает качественную обратную связь, на основе обсуждений и комментариев пользователей этих ресурсов.

3.4. Результаты эксплуатации каталога ресурсов интернет

18 марта 2007 года каталог ресурсов интернет был запущен в режиме бета-версии. Было произведено начальное продвижение каталога в интернет, включающее ряд этапов:

- произведена поисковая оптимизация страниц каталога;
- произведена платная регистрация в каталогах ресурсов интернет и рейтингах на полу-автоматической основе;
- опубликовано интервью для сетевых изданий toodoo.ru, internet.ru с анонсами на сайтах news2.ru, habr.ru, а также персональном блоге автора.

После публикаций интервью и пресс-релизов наблюдался лавинообразный рост количества посетителей. В этот момент на сайте были размещены многочисленные формы обратной связи, позволившие пользователям указать на положительные стороны и недостатки реализации. Был сформирован перечень доработок, которые были незамедлительно произведены в некотором объеме.

Бета-тестирование продолжалось в течение трех недель. 11 апреля 2007 года состоялся «релиз» каталога: сняты дополнительные формы обратной связи, размещены рекламные материалы рекламного брокера «Бегун». Начиная с данного момента каталог работает в режиме эксплуатации.

3.4.1. Статистические данные

Начиная с 18 марта 2007 года по 20 мая 2007 года:

- каталогом воспользовались более 7 тысяч пользователей;
- пользователи просмотрели более 35 тысяч страниц каталога;
- каждый пользователь (в среднем) просмотрел 5 страниц;
- в среднем, пользователь просматривает каталог более 2,5 минут;
- основные источники трафика: поисковые машины и ссылающиеся сайты;
- в каталоге зарегистрировано более 850 интернет-ресурсов;
- в каталоге зарегистрировано более 800 владельцев сайтов;
- владельцы сайта указали более 1 тысячи тегов для фолксономической категоризации их интернет-ресурсов;
- пользователи оставили 15 комментариев к интернет-ресурсам;
- интеллектуальные модули создали более 2 тысяч документов для внутренних нужд.

Более детальную статистику в разрезе дней можно изучить в виде отчетов системы Google Analytics™, представленных в Приложении 1.

3.4.2. Оценка экономической эффективности проекта

Рекламные материалы размещены на страницах описаний интернет-ресурсов каталога. Эти материалы представлены в виде контекстно-зависимых объявлений. В качестве рекламного брокера используется компания «Бегун». Рекламные материалы представляют собой текстовые ссылки с кратким описанием, их количество на каждой странице: не более трех. Область рекламных материалов выделена цветом, отличным от цвета фона и других элементов страницы, т. о. рекламные материалы выделяются среди прочих элементов страницы (см. рис. 20).

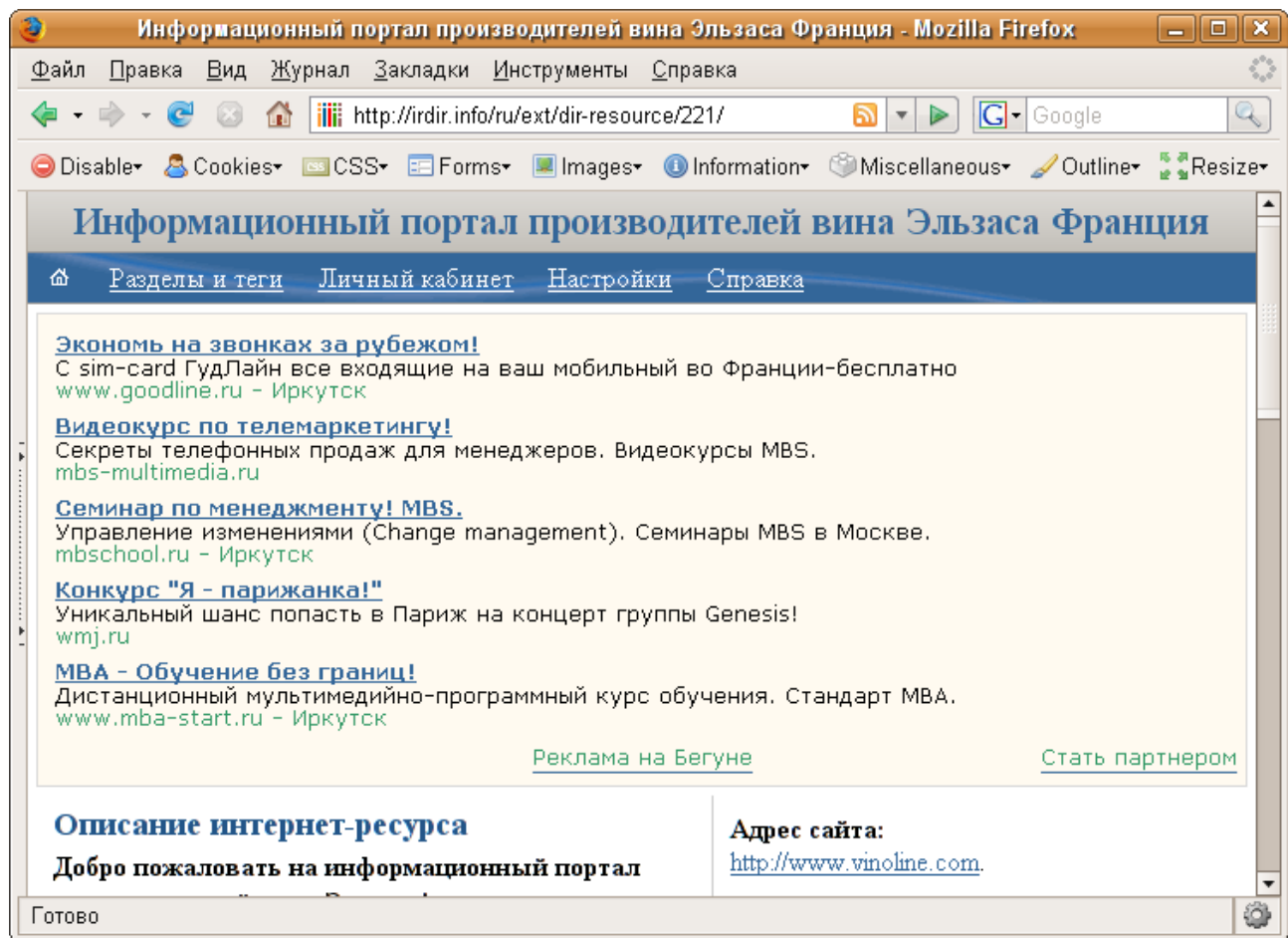


Рис. 20: Рекламные объявления в каталоге

Доход с каталога образуется за счет рекламодателей, сотрудничающих с компанией «Бегун» [9]. При этом стоимость перехода по ссылке определяется на аукционной основе и зависит, как правило, от количества рекламодателей, желающих рекламироваться по определенному поисковому запросу. В среднем, для рекламодателя, цена перехода составляет около 3-5 рублей. При этом, за каждый осуществленный переход на сайт рекламодателя с сайта партнера, компания «Бегун» удерживает комиссию в размере 50% от начальной стоимости клика. Остальные 50% получает владелец рекламной площадки (партнер). Также, стоит отметить, что при обналичивании полученного вознаграждения, партнер должен оплатить комиссию по переводу денежных средств в размере 0.08% (на основании правил использования системы Webmoney Transfer, определившей именно такой порог комиссии для транзакций титульных знаков WMZ — электронный эквивалент USD в терминах системы).

Исходя из статистических данных о просмотрах страниц можно оценить

тенденции роста посещаемости, а также, количество переходов по рекламным объявлениям.

Прогнозирование будет осуществляться при помощи интернет-приложения для прогнозирования по временному ряду, ранее разработанному автором данного проекта. Будет использоваться адаптивный метод прогнозирования: метод экспоненциального сглаживания Брауна.

Несмотря на то, что данный ряд (взяты наблюдения за последние 30 дней, рис. 21) содержит циклическую составляющую, ее характер не является ярко выраженным, поэтому применение метода экспоненциального сглаживания является вполне оправданным .

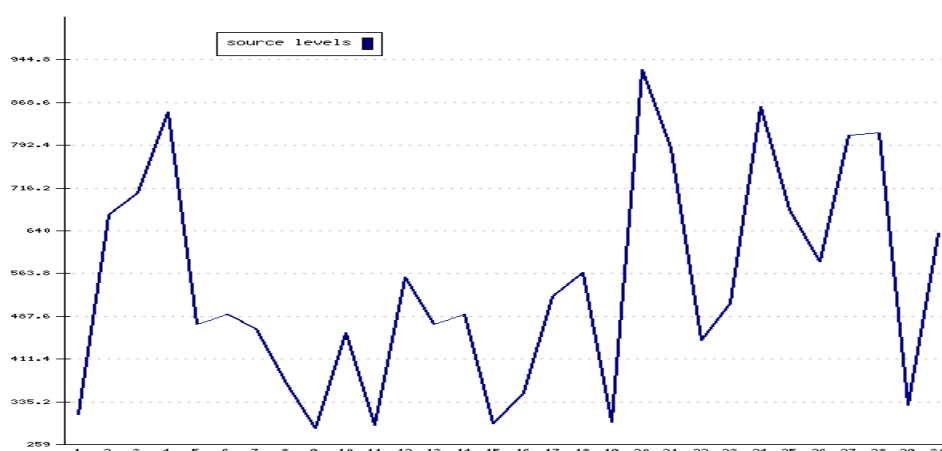


Рисунок 21: Исходные уровни ряда

Для модели задан ряд начальных параметров:

- тип функции: полином второй степени;
- период упреждения прогноза: 20 периодов;
- период сглаживания: 15 периодов (половина исходного ряда);
- уровень значимости (доверительная вероятность): 95% (0.05).

На основе исходных данных и начальных параметров модели был построен прогноз и получены прогнозные уровни (см. рис. 22).

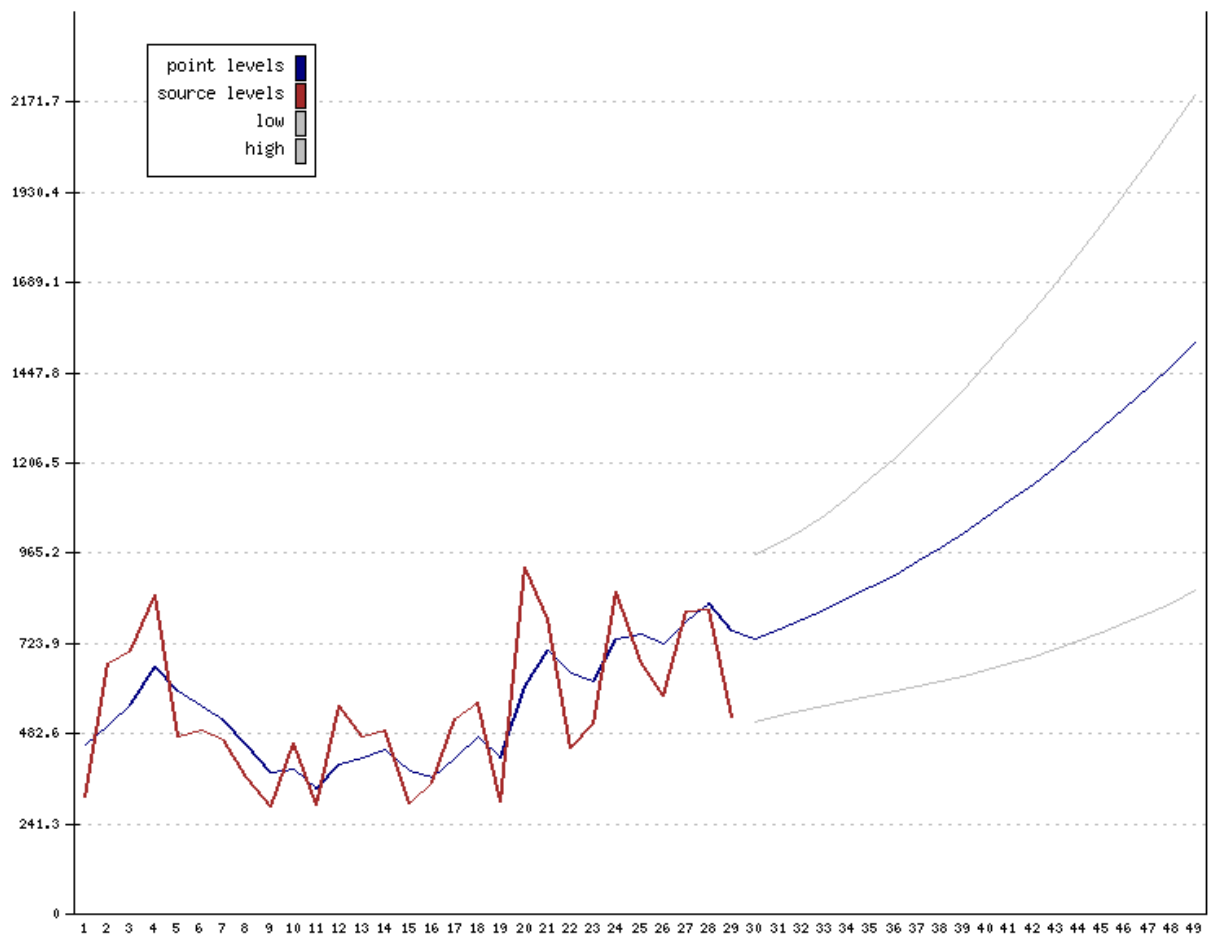


Рисунок 22: Результаты прогнозирования

Для последних трех прогнозных уровней ряда значения точечного и интервального прогноза составляют:

период 48:

нижняя граница: 805.268090768;
 точечный прогноз: 1354.87784933;
 верхняя граница: 2016.38014535;

период 49:

нижняя граница: 834.062279616;
 точечный прогноз: 1410.82411806;
 верхняя граница: 2103.56796451;

период 50:

нижняя граница: 864.711039485;
 точечный прогноз: 1468.81512207;

верхняя граница: 2192.99068321;

По данному прогнозу отчетливо видно, что ряд обладает тенденцией роста (даже если исходить из пессимистических соображений, принимая во внимание лишь нижнюю границу интервального прогноза).

Учитывая тот факт, что уровень цен на рекламные объявления составляет (в среднем) 4 рубля за переход, можно определить уровень дохода, производя соотношения со статистикой посещаемости и прогнозными уровнями. Также, следует заметить, что уровень CTR (click-through-ratio) для таких объявлений составляет не менее 5% (т. е. один переход по ссылкам осуществляется при просмотре 200 страниц сайта, в среднем).

Основываясь на таблице расчета дохода (см. ПРИЛОЖЕНИЕ 2) можно определить доход за последний и предпоследний месяцы, он, соответственно будет равен:

- доход за последний месяц: 543,58 р.
- доход за предпоследний месяц: 317,8 р.

Основываясь на этих данных, можно предположить, что ежемесячный рост дохода составляет 171% («грубая» оценка по двум периодам).

Таким образом, экстраполируя на 10 периодов упреждения, получаем следующие данные (один период эквивалентен одному месяцу, табл. 23):

Таблица 23

Доход по периодам

Период	Доход	Нарастающим итогом
1	317,8	317,8
2	543,44	861,24
3	929,28	1790,52
4	1589,07	3379,58
5	2717,3	6096,89
6	4646,59	10743,48
7	7945,67	18689,15
8	13587,1	32276,25
9	23233,94	55510,18
10	39730,03	95240,21

Следует заметить, что данная оценка дохода является «грубой», на практике, рост дохода не будет иметь такой формы, а будет носить «затухающий» характер, спустя несколько периодов роста. Рост в такой форме может обеспечиваться лишь постоянным продвижением проекта, например, используя соци-

альные сети и интернет-СМИ.

Перейдем к расчету затрат на создание и поддержку каталога. Приведем все затраты к табличному виду и поделим их на единоразовые и ежемесячные.

К единоразовым затратам относятся затраты на разработку проекта, разработку программной части, дизайна, верстки макетов, покупку доменного имени, начального размещения в сети интернет, все этапы тестирования каталога (табл. 24). Следует учесть, что разработка ведется на протяжении трех месяцев. Вознаграждение (ежемесячное) разработчика составляет 8 т. р., при этом на коммуникационные услуги (включая доступ к сети интернет) выделяется 1,5 т. р. в месяц.

Таблица 24

Единоразовые затраты

Вид затрат	Стоимость
Коммуникационные услуги	4500
Зарплата разработчику	24000
Разработка проекта	5000
Разработка дизайна и верстка	10000
Доменное имя	300
Этапы тестирования	5000
Накладные расходы	1500
Итого	50300

К ежемесячным затратам относятся затраты на поддержку пользователей, оплату услуг хостинговой компании, затраты на модернизацию каталога и поддержки его программной части в актуальном состоянии. Следует учесть, что затраты на поддержку рассчитываются исходя из 7-месячного периода. При этом ежемесячно на поддержку пользователей выделяется 500 рублей: этого вполне достаточно для осуществления обратной связи, поскольку большинство ответов на часто задаваемые вопросы пользователи могут получить при интерпретации документов от интеллектуальных модулей каталога (табл. 25).

Таблица 25

Ежемесячные затраты

Вид затрат	Стоимость
Поддержка пользователей	5000
Услуги хостинга	3000
Модернизация каталога	5000
Итого	13000

Таким образом, спустя 13 месяцев, совокупные затраты на проект будут составлять 63300 р. (единоразовые затраты в размере 50300 р. и ежемесячные затраты в размере 13000 р.). Чистая прибыль проекта определяется как доход проекта, умноженный на 99,92% (учет комиссии системы WMT) минус единоразовые и ежемесячные затраты за период проектирования, разработки и эксплуатации каталога. Чистая прибыль по истечению 13 месяцев (без учета дисконтирования) с момента разработки проекта составляет 31914 р. (разница между доходом размером в 95240 р. с учетом комиссии WMT и затратами в размере 63300 р.). Период окупаемости проекта составляет 13 месяцев (табл. 26, рис. 23).

Таблица 26

Доходы и затраты по периодам

Период	Доход накопленным итогом	Затраты накопленным итогом
1	0	21300
2	0	35300
3	0	50300
4	317,8	51600
5	861,24	52900
6	1790,52	54200
7	3379,58	55500
8	6096,89	56800
9	10743,48	58100
10	18689,15	59400
11	32276,25	60700
12	55510,18	62000
13	95240,21	63300

Расшифровку затрат по периодам можно посмотреть в Приложении 2.

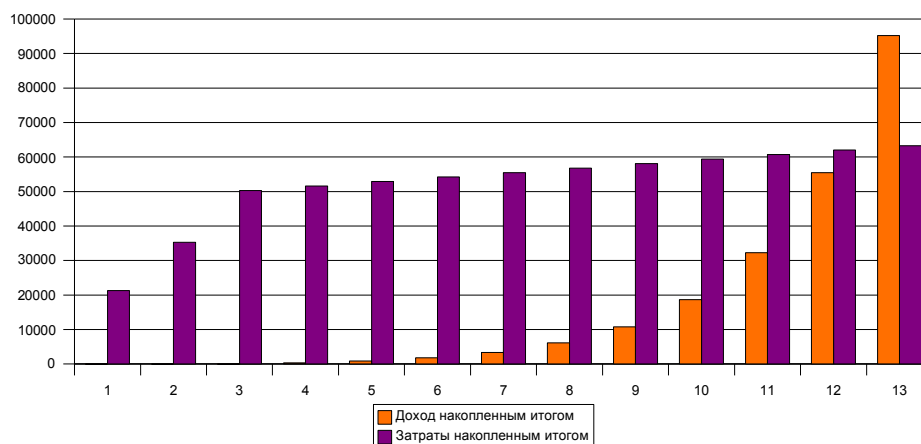


Рис. 23: Динамика доходов и затрат по периодам

Создание каталога с автоматизированными и интеллектуализированными функциями позволило снизить постоянные (ежемесячные) издержки и поддерживать их на достаточно низком уровне.

Учитывая доходы и расходы проекта нетрудно посчитать чистый доход по периодам и построить диаграмму для визуализации этого ряда (рис. 24).

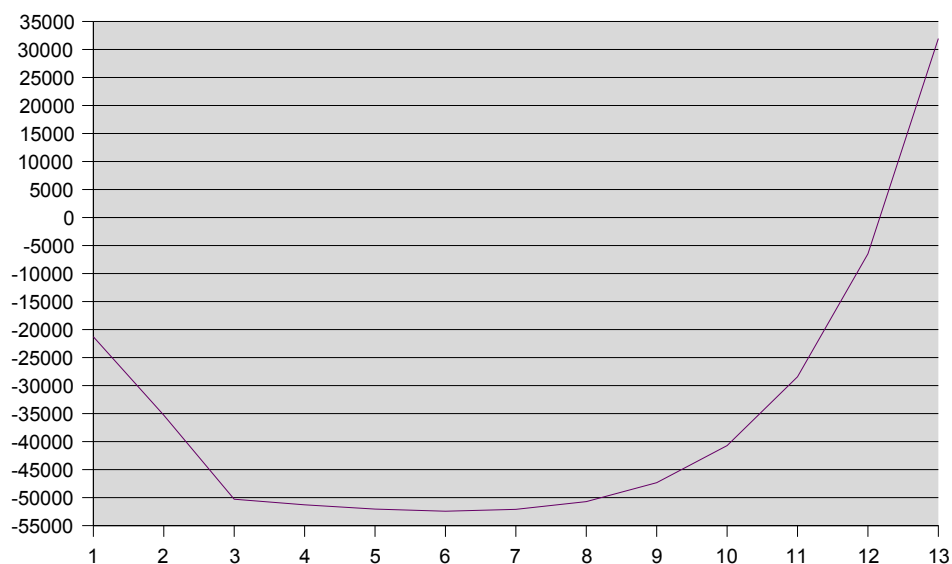


Рисунок 24: Чистый доход

Заключение

В процессе выполнения дипломного проекта была изучена специфика каталогов ресурсов интернет, применяемые подходы при построении информационных систем такого уровня, проведен анализ наиболее популярных каталогов и выявлены их сильные и слабые стороны.

На стадии проектирования каталога определена его структура, необходимые функции, элементы интерфейса. При разработке каталога использован ряд уникальных подходов, которые ранее не применялись при построении каталогов ресурсов интернет: объединенный таксономический и фолксономический подход к категоризации, многофакторная система ранжирования интернет-ресурсов, система метаданных «Дублинское ядро», технология открытого поиска OpenSearch, «интеллектуальная» модерация. Была разработана программная реализация каталога, проведены этапы тестирования, произведено продвижение каталога в сети Интернет.

В настоящее время каталог работает в режиме эксплуатации и обслуживает многочисленные запросы пользователей. На страницах каталога размещены рекламные материалы, при помощи которых обеспечивается окупаемость проекта.

Каталог доступен в сети интернет по адресу <http://irdir.info>.

Список источников

1. Сайт компании Яндекс: «Оптимизация процедуры автоматического пополнения веб-каталога» [Электронный ресурс] — Электрон. дан. — Режим доступа: http://company.yandex.ru/grant/2005/08_Kiselev_102710.pdf, свободный.
2. Сайт компании 1PS: «Продвижение сайтов, регистрация в каталогах» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://1ps.ru/>, свободный.
3. Интернет-энциклопедия «Википедия»: «Каталоги ресурсов интернет» [Электронный ресурс] — Электрон. дан. — Режим доступа: http://ru.wikipedia.org/wiki/Каталог_ресурсов_интернет, свободный.
4. Введение в искусственный интеллект: Учеб. пособие для студ. высш. учеб. заведений / Л.Н. Ясницкий . - М.: Издательский центр «Академия», 2005. - 176 с.
5. Программирование искусственного интеллекта в приложениях / М. Тим Джонс; Пер. с англ. Осипов А.И. - М.: ДМК Пресс, 2006 – 312 с.: ил.
6. Управление информационными потоками / Сборник трудов Института системного анализа Российской академии наук. Под ред. д. т. н. проф. В.Л. Арлазарова, д. т. н., проф. Н.Е. Емельянова. - М.: Едиториал УРСС, 2002. - 368 с.
7. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учебное пособие. - М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. - 304 с.: ил.
8. Портал интернет-разработчиков «Xpoint»: «Зачем нужна модерация» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://xpoint.ru/know-how/Articles/ZachemNuzhnaModeratsiya>, свободный.
9. Компания «Begun»: «Правила участия в партнерских программах» [Электронный ресурс] — Электрон. дан. — Режим доступа: http://www.begun.ru/legal/partner_rules.php, свободный.

10. Хостинг-обзор: «Эпицентр русскоязычного хостинга» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://hostobzor.ru>, свободный.
11. Компания «Яндекс»: «Как устроен каталог Яндекса» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://help.yandex.ru/catalogue/>, свободный.
12. Open directory project: «About the ODP» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://dmoz.org/about.html>, свободный.
13. MySQL. MySQL Reference Manual. MySQL AB, 2006.
14. PHP Group: «Руководство по PHP» [Электронный ресурс] — Электрон. дан. — Режим доступа: <http://www.php.net/manual/ru/>, свободный.
15. W3 Consortium, XHTML 1.0 Transitional documentation. W3C, 2007.
16. Google corporation, Google Technology. Google corp., 2007.
17. John Allsopp. Microformats: Empowering Your Markup for Web 2.0. - New York, Springer-Verlag, 2007. - 330 с.

Статистические отчеты Google Analytics

В этом приложении содержится ряд статистических отчетов, предоставленных системой статистики сайтов Google Analytics™ от компании Google®, inc.

Посещения для всех посетителей

Период: 17.03.2007–20.05.2007.

Домен: irdir.info.

Всего 7 358 посещений за указанный период.

Графическое представление находится на рис. 25.



Рис. 25: Посещения для всех посетителей

Детализация по дням:

17 марта 2007 г.	0,00%	(0)
18 марта 2007 г.	1,58%	(116)
19 марта 2007 г.	1,85%	(136)
20 марта 2007 г.	1,18%	(87)
21 марта 2007 г.	0,80%	(59)
22 марта 2007 г.	0,90%	(66)
23 марта 2007 г.	0,49%	(36)
24 марта 2007 г.	0,33%	(24)
25 марта 2007 г.	0,41%	(30)
26 марта 2007 г.	0,77%	(57)
27 марта 2007 г.	1,03%	(76)
28 марта 2007 г.	0,86%	(63)
29 марта 2007 г.	0,91%	(67)
30 марта 2007 г.	0,77%	(57)
31 марта 2007 г.	0,31%	(23)
1 апреля 2007 г.	0,29%	(21)
2 апреля 2007 г.	0,91%	(67)
3 апреля 2007 г.	1,09%	(80)
4 апреля 2007 г.	2,09%	(154)
5 апреля 2007 г.	2,62%	(193)
6 апреля 2007 г.	1,36%	(100)

7 апреля 2007 г. 0,53% (39)
8 апреля 2007 г. 0,64% (47)
9 апреля 2007 г. 1,16% (85)
10 апреля 2007 г. 1,45% (107)
11 апреля 2007 г. 1,44% (106)
12 апреля 2007 г. 1,44% (106)
13 апреля 2007 г. 1,85% (136)
14 апреля 2007 г. 0,94% (69)
15 апреля 2007 г. 0,94% (69)
16 апреля 2007 г. 1,45% (107)
17 апреля 2007 г. 1,66% (122)
18 апреля 2007 г. 1,51% (111)
19 апреля 2007 г. 1,71% (126)
20 апреля 2007 г. 1,98% (146)
21 апреля 2007 г. 1,29% (95)
22 апреля 2007 г. 1,82% (134)
23 апреля 2007 г. 2,12% (156)
24 апреля 2007 г. 2,54% (187)
25 апреля 2007 г. 2,01% (148)
26 апреля 2007 г. 1,81% (133)
27 апреля 2007 г. 1,71% (126)
28 апреля 2007 г. 1,54% (113)
29 апреля 2007 г. 1,03% (76)
30 апреля 2007 г. 1,35% (99)
1 мая 2007 г. 1,07% (79)
2 мая 2007 г. 1,54% (113)
3 мая 2007 г. 1,71% (126)
4 мая 2007 г. 1,98% (146)
5 мая 2007 г. 1,66% (122)
6 мая 2007 г. 1,74% (128)
7 мая 2007 г. 2,83% (208)
8 мая 2007 г. 2,56% (188)
9 мая 2007 г. 1,47% (108)
10 мая 2007 г. 2,47% (182)
11 мая 2007 г. 2,61% (192)
12 мая 2007 г. 1,86% (137)
13 мая 2007 г. 2,17% (160)
14 мая 2007 г. 2,62% (193)
15 мая 2007 г. 2,85% (210)
16 мая 2007 г. 2,85% (210)
17 мая 2007 г. 3,29% (242)
18 мая 2007 г. 2,79% (205)
19 мая 2007 г. 1,79% (132)
20 мая 2007 г. 1,66% (122)

Источники трафика

На рисунке 26 представлены основные источники трафика в виде диаграммы с краткими пояснениями.



Рис. 26: Источники трафика

Основные источники трафика

Основные источники трафика: поисковые системы и прямые запросы (рис. 27).

Источники	Посещения	% посещений
yandex (organic)	2 892	39,30%
(direct) ((none))	1 455	19,77%
google (organic)	489	6,65%
dir.org.ru (referral)	486	6,61%
go.mail.ru (referral)	359	4,88%

Рис. 27: Основные источники трафика

Просмотры страниц

Период: 17.03.2007–20.05.2007.

Домен: irdir.info.

Всего 35 704 просмотра страниц за указанный период.

Графическое представление приведено на рис. 28.



Рис. 28: Просмотры страниц

Детализация по дням:

17 марта 2007 г.	0,00%	(0)
18 марта 2007 г.	3,09%	(1 103)
19 марта 2007 г.	4,15%	(1 480)
20 марта 2007 г.	2,43%	(866)
21 марта 2007 г.	2,09%	(745)
22 марта 2007 г.	1,04%	(373)
23 марта 2007 г.	1,19%	(426)
24 марта 2007 г.	0,73%	(261)
25 марта 2007 г.	0,64%	(230)
26 марта 2007 г.	1,08%	(387)
27 марта 2007 г.	1,45%	(518)
28 марта 2007 г.	1,98%	(706)
29 марта 2007 г.	1,15%	(411)
30 марта 2007 г.	0,55%	(198)
31 марта 2007 г.	0,31%	(110)
1 апреля 2007 г.	0,89%	(319)
2 апреля 2007 г.	1,12%	(400)
3 апреля 2007 г.	1,52%	(543)
4 апреля 2007 г.	2,98%	(1 063)
5 апреля 2007 г.	3,44%	(1 229)
6 апреля 2007 г.	0,98%	(349)
7 апреля 2007 г.	0,63%	(226)
8 апреля 2007 г.	0,99%	(353)
9 апреля 2007 г.	1,51%	(540)
10 апреля 2007 г.	1,46%	(522)
11 апреля 2007 г.	1,48%	(527)
12 апреля 2007 г.	1,55%	(553)
13 апреля 2007 г.	2,29%	(817)
14 апреля 2007 г.	1,33%	(475)
15 апреля 2007 г.	2,04%	(730)
16 апреля 2007 г.	1,34%	(478)
17 апреля 2007 г.	2,74%	(979)
18 апреля 2007 г.	1,25%	(448)
19 апреля 2007 г.	1,28%	(457)
20 апреля 2007 г.	1,57%	(562)

21 апреля 2007 г. 0,87% (312)
22 апреля 2007 г. 1,87% (669)
23 апреля 2007 г. 1,98% (707)
24 апреля 2007 г. 2,39% (852)
25 апреля 2007 г. 1,33% (474)
26 апреля 2007 г. 1,38% (492)
27 апреля 2007 г. 1,30% (465)
28 апреля 2007 г. 1,03% (367)
29 апреля 2007 г. 0,81% (288)
30 апреля 2007 г. 1,28% (458)
1 мая 2007 г. 0,82% (294)
2 мая 2007 г. 1,56% (558)
3 мая 2007 г. 1,32% (473)
4 мая 2007 г. 1,38% (492)
5 мая 2007 г. 0,83% (296)
6 мая 2007 г. 0,98% (351)
7 мая 2007 г. 1,46% (523)
8 мая 2007 г. 1,59% (566)
9 мая 2007 г. 0,84% (299)
10 мая 2007 г. 2,60% (928)
11 мая 2007 г. 2,20% (787)
12 мая 2007 г. 1,25% (445)
13 мая 2007 г. 1,43% (511)
14 мая 2007 г. 2,41% (861)
15 мая 2007 г. 1,90% (677)
16 мая 2007 г. 1,64% (585)
17 мая 2007 г. 2,27% (811)
18 мая 2007 г. 2,28% (814)
19 мая 2007 г. 0,92% (329)
20 мая 2007 г. 1,78% (636)

ПРИЛОЖЕНИЕ 2

Таблицы для расчета экономической эффективности

В приведенной ниже таблице содержится расчет дохода по периодам (табл. 27), один период равен одним суткам.

Таблица 27

Расчет дохода по периодам

Период	Просмотры	Клики	Доход
1	1103	5,52	22,06
2	1480	7,4	29,6
3	866	4,33	17,32
4	745	3,73	14,9
5	373	1,87	7,46
6	426	2,13	8,52
7	261	1,31	5,22
8	230	1,15	4,6
9	387	1,94	7,74
10	518	2,59	10,36
11	706	3,53	14,12
12	411	2,06	8,22
13	198	0,99	3,96
14	110	0,55	2,2
15	319	1,6	6,38
16	400	2	8
17	543	2,72	10,86
18	1063	5,32	21,26
19	1229	6,15	24,58
20	349	1,75	6,98
21	226	1,13	4,52
22	353	1,77	7,06
23	540	2,7	10,8
24	522	2,61	10,44
25	527	2,64	10,54
26	553	2,77	11,06
27	817	4,09	16,34
28	475	2,38	9,5
29	730	3,65	14,6
30	478	2,39	9,56
31	979	4,9	19,58
32	448	2,24	8,96
33	457	2,29	9,14
34	562	2,81	11,24
35	312	1,56	6,24
36	669	3,35	13,38
37	707	3,54	14,14
38	852	4,26	17,04
39	474	2,37	9,48
40	492	2,46	9,84
41	465	2,33	9,3
42	367	1,84	7,34
43	288	1,44	5,76
44	458	2,29	9,16

Окончание табл. 27.

Период	Просмотры	Клики	Доход
45	294	1,47	5,88
46	558	2,79	11,16
47	473	2,37	9,46
48	492	2,46	9,84
49	296	1,48	5,92
50	351	1,76	7,02
51	523	2,62	10,46
52	566	2,83	11,32
53	299	1,5	5,98
54	928	4,64	18,56
55	787	3,94	15,74
56	445	2,23	8,9
57	511	2,56	10,22
58	861	4,31	17,22
59	677	3,39	13,54
60	585	2,93	11,7
61	811	4,06	16,22
62	814	4,07	16,28
63	329	1,65	6,58
64	636	3,18	12,72
65	756	3,78	15,12
66	738	3,69	14,76
67	761	3,81	15,22
68	786	3,93	15,72
69	813	4,07	16,26
70	843	4,22	16,86
71	874	4,37	17,48
72	907	4,54	18,14
73	943	4,72	18,86
74	980	4,9	19,6
75	1020	5,1	20,4
76	1062	5,31	21,24
77	1105	5,53	22,1
78	1151	5,76	23,02
79	1199	6	23,98
80	1249	6,25	24,98
81	1301	6,51	26,02
82	1355	6,78	27,1
83	1411	7,06	28,22
84	1469	7,35	29,38

В таблице 28 приведен расчет затрат по периодам с учетом типов расходов. В первые 3 периода происходит разработка, остальные периоды относятся к этапам после периода разработки.

Таблица 28

Затраты по периодам

Период	з/п разработчика	Коммуникационные услуги	Разработка проекта	Дизайн	Покупка домена	Тестирование	Накладные расходы	Постоянные издержки
1	8000	1500	5000	6000	300	0	500	0
2	8000	1500	0	4000	0	0	500	0
3	8000	1500	0	0	0	5000	500	0
4	0	0	0	0	0	0	0	1300
5	0	0	0	0	0	0	0	1300
6	0	0	0	0	0	0	0	1300
7	0	0	0	0	0	0	0	1300
8	0	0	0	0	0	0	0	1300
9	0	0	0	0	0	0	0	1300
10	0	0	0	0	0	0	0	1300
11	0	0	0	0	0	0	0	1300
12	0	0	0	0	0	0	0	1300
13	0	0	0	0	0	0	0	1300