ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
РАБОТЫ СО ЗНАНИЯМИ:
ОБНАРУЖЕНИЕ, ПОИСК, УПРАВЛЕНИЕ
(Вычислительные системы)
2008 год Выпуск 175

УДК 681.3: 004.8

РАЗРАБОТКА ЭКСПЕРИМЕНТАЛЬНОЙ СИСТЕМЫ ПОДБОРА ЭВРИСТИК ДЛЯ ВИРТУАЛЬНОГО КАТАЛОГА

А.М. Бездольный

Введение

В наши дни практически вся информация, накопленная человечеством, содержится в электронном виде в сети. В Интернете можно найти ресурсы, посвященные самым разнообразным темам и вопросам. Однако по мере накопления информации становится все сложнее и сложнее найти именно то, что необходимо. Ситуация осложняется тем, что все быстрее растет число людей, использующих сеть и доля неподготовленных пользователей среди них. Это приводит к тому, что проблема эффективного и простого для пользователя поиска информации становится особенно острой.

Основными на сегодняшний день подходами к поиску информации являются информационно–поисковые системы и каталоги. У каждого из этих подходов есть свои преимущества и недостатки. Виртуальный каталог — это новый подход к организации поиска в Интернете, который объединяет в себе простоту обычных каталогов и преимущества информационно–поисковых систем. Внешне виртуальный каталог выглядит так же как обычный: используется иерархическая структура онтологии предметной области (рубрикатор). Но отличие состоит в том, что вертуальный

каталог не хранит в себе ссылок на ресурсы. Рубрика вместе с дополнительной информацией определяет уточняющий запрос к обычной информационно–поисковой системе. Пользователь для каждой рубрики получает список ресурсов, являющийся обработанным результатом исполнения запроса к информационно–поисковой системе, что позволяет каждый раз обеспечивать актуальность и полноту.

Одной из основных задач при реализации виртуального каталога является задание онтологии предметной области и указание эвристик поиска для каждого элемента этой онтологии. Еще более сложной, и не менее важной, задачей является дальнейшая поддержка и обновление онтологии и эвристик.

Целью этой работы было облегчить и систематизировать процесс создания поддержки и обновления онтолгии и эвристик для виртуального каталога. Сам процесс составления иерархии онтологии является сложной много этапной эадачей, и поэтому в данной работе рассматривается только внесение данных о уже составленной рубрикации и процесс подбора эвристик поиска. Для достижения поставленной цели было решено разработать «Экспериментальную систему подбора эвристик» — специальный инструментарий для работы экспертов–разработчиков виртуального каталога.

Виртуальный каталог

Для оценки качества поиска можно использовать две характеристики — релевантность и пертинентность. Релевантность — это степень соответствия результатов поиска поисковому запросу. Таким образом, релевантность оценивает только формальное соответствие запроса и документов, полученных в результате поиска. Пертинентность — это степень соответствия результата поиска и информационной потребности пользователя. И эта величина определяется как отношение объема полезной для пользователя информации к общему объему выданной поисковой системой информации.

Основной задачей любой поисковой системы является обеспечение пертинентности выделяемых результатов поиска. Как было подробно описано в работе [6], основным на сегодняшний день

подходом к поиску — информационно–поисковым системам и каталогам — это удается обеспечить далеко не всегда. Каталоги понятны и легки в использовании, но не могут обеспечить необходимый уровень полноты и актуальности. Использование информационно–поисковых систем во многих случаях требует от пользователя специальной подготовки. Также часто очень сложно точно определить в поисковом запросе контекст поиска, а многие термины в зависимости от предметной области имеют очень разное значение.

Одним из возможных новых подходов к поиску в Интернете является виртуальный каталог. Так же, как и в обычных каталогах, основой виртуального каталога является иерархическая структура онтологии предметной области. Перемещаясь в вертикальном или горизонтальном направлениях по этой иерархии, пользователь точно указывает контекст для дальнейшего поиска. Такой подход позволяет предоставить простой и понятный интерфейс к поисковой системе и обеспечивает адекватность понимания информационной потребности пользователя.

Как уже было отмечено, виртуальный каталог не содержит в себе списка ссылок на ресурсы сети. Каждому элементу иерархии онтологии предметной области посталены в соответствие эвристики поиска. Эвристики используются для составления запроса к информационно–поисковой системе. Это позволяет использовать основные преимущества информационно–поисковых систем: полноту и актуальность полученных результатов.

Виртуальный каталог предоставляет пользователям более простой интерфейс для точной формулировки своей информационной потребности, что позволяет существенно увеличить пертинентность получаемых результатаов поиска. Но предоставление пользователю более богатых возможностей усложняет задачу обеспечения релевантности — для этого в вертуальном каталоге необходимо составить онтолгию предметной области и определить эвристики поиска.

Процесс подбора эвристик

Как уже было отмечено, составление рубрикации для виртуального каталога яваляется комплексной задачей; для этого необходимо составить онтологию предметной области и выделить

ее иерархию. В работе рассматривается работа с уже составленой рубрикацией и процесс подбора эвристик.

Для систематизации процесса в начале был проведен анализ порядка работы экспертов над подбором эвристик. Были выделены следующие этапы:

1) создание рубрикации;

2) для каждой рубрики изначально выбирается ее название в качестве эвристики;

3) для каждой рубрики проверяется корректность получаемых результатов поиска; если находятся несоответствия, то эксперт пытается подобрать новую эвристику, которая исключит попадание в результаты поиска неподходящих по тематике ресурсов.

Описанная последовательность действий не учитывает многие моменты. Например, при подборе эвристик, в процессе отсева неподходящих по тематике ресурсов из результатов поисковой выдачи, могут быть исключены и нужные ресурсы. Разработка более корректного процесса подбора эвристик продолжается. Изучение опыта создания виртуального каталога дает возможность проработать необходимые процедуры и устранить эти моменты. На данный момент для обеспечения полноты в виртуальном каталоге для каждой рубрики используется несколько эвристик и пользователю предоставляется объединенный список ресурсов.

Для составления нескольких эвристик и для того, чтобы избежать отсеивания подходящих по тематике ресурсов, экспертами производится подробное сравнеие результатов поиска с использованием каждой эвристики. Даже описанный выше упрощенный процесс сложно исполнять без поддержки специальным инструментарием, а сравнение результатов поиска с использованием разных эвристик само по себе является нетривиальной задачей и требует отдельной проработки.

Отдельной большой задачей на текущий момент является разработка процесса поддержки и обновления онтологии и эвристик. Точно описать этот процесс пока не представляется возможным, но можно сказать, что потребуется набор отчетов и других механизмов для отслеживания изменений в поведении системы виртуального каталога. Также прорабатываются способы итоговой

оценки и анализа списка выдаваемых ресурсов для каждой рубрики. Возможно это позволит автоматически или полуавтоматически численно оценивать полноту и релевантность выдаваемых ресурсов.

Задачи, решаемые экспериментальной системой подбора эвристик

Ранее мы рассмотрели порядок подбора эвристик поиска для вертуального каталога и обнаружили трудоемкость этого процесса. Для обеспечения задачи подбора эвристик было принято решение разработать специальный инструментарий для работы экспертов.

Перед началом разработки этого инструментария было необходимо определить возможные роли пользователей и какие действия могут быть исполнены с его помощью. На данный момент были выделены две роли пользователей:

- администратор — управляет работой виртуального каталога и работой инструментария,
- эксперт — задает рубрикацию и осуществляет подбор эвристик.

Основным пользователем экспериментальной системы подбора эвристик является эксперт. Анализируя процесс подбора эвристик, можно выделить основные действия, которые могут совершаться экспертом:

- создание рубрики,
- удаление рубрики,
- изменение названия или описания рубрики,
- просмотр всей рубрикации,
- просмотр спика эвристик, определенных для рубрики,
- создание эвристики,
- удаление эвристики,
- проверка эвристики,
- модификация эвритики,
- сравнение эвристик.

Поскольку процесс подбора эвристик является трудоемкой задачей, при его исполнении могут быть допущены ошибки. Поэтому изменения, сделанные экспертом, не должны сразу отражаться на работе системы виртуальный каталог. Исходя из этого

соображения можно составить список возможных действия для администратора экспертной системы подбора эвристик:

- сохранение текущего состояния рубрикации и эвристик,
- возврат к ранее сохраненному состоянию,
- экспорт рубрикации и эвристик в виртуальный каталог.

Требования к экспериментальной системе подбора эвристик

Ранее мы описали минимальный набор вариантов использования экспериментальной системы подбора эвристик. Этот набор является отражением существующего на данный момент процесса подбора эвристик экспертами. Следующим шагом перед разработкой инструментария являлось составление списка функциональных и нефункциональных требований.

Для составления списка требований необходимо было учесть то, что в разработке системы виртуального каталога может участвовать несколько экспертов, а также то, что одновременно в разработке может находиться несколько виртуальных каталогов для разных предметных областей.

После обсуждения и анализа процесса подбора эвристик и задач, которые должна решать экспериментальная система, были сформулированы требования к ней.

1. Удобный интерфейс создания и редактирования рубрикации.

2. Возможность привязывать к каждой рубрике неограниченное количество эвристик (уточняющих запросов).

3. Возможность увидеть какие ссылки появляются по каждой привязанной к рубрике эвристике.

4. Для каждой эвристики должна запоминаться сатистика по ссылкам, выдаваемым поисковой системой, и сохраняться степень точности эвристики.

5. Должна присутствовать возможность для каждой эвристики определить будет ли она использоваться для поиска в виртуальном каталоге.

6. Должен присутствовать быстрый экспорт в систему виртуального каталога.

7. Должна присутствовать возможность одновременно работать с несколькими предметными областями.

Кроме требований к функциональным возможностям системы был сформулирован ряд нефункциональных требований.

1. Экспериментальная система должна поддерживать совместную работу нескольких экспертов.

2. Система должна предоставлять простой и понятный интерфейс. Эксперты не должны разбираться в технических тонкостях работы всей системы.

3. Работа с экспериментальной системой должна быть возможна с любого компьютера, подключенного к Интернет.

4. Для работы с экспериментальной системой должно быть достаточно только установленного Интернет–обозревателя (Internet browser).

Также отдельно нужно отметить, что разрабатываемый инструментарий является экспериментальной системой. Сейчас продождается проработка процесса подбора эвристик и постоянно вносятся изменения в архитектуру системы виртуального каталога. Поэтому появляется дополнительное требование к архитектуре экспериментальной системы — она должна быть модульной и достаточно гибкой для легкой модификации, отражающей изменения в процессе подбора эвристик и в работе системы виртуального каталога.

Архитектура экспериментальной системы подбора эвристик

Для построения архитектуры сначала необходимо построить модель данных, с которыми работает разрабатываемая экспериментальная система подбора эвристик и выделить основные сущности.

На рис.1 изображена модель данных, с которыми работает инструментарий эксперта. Рассмотрим представленные сущности подробнее.

• Проект — в рамках этой сущности содержатся все остальные. В разных проектах могут вестись работы для различных предметных областей или различных исталляций виртуального каталога.

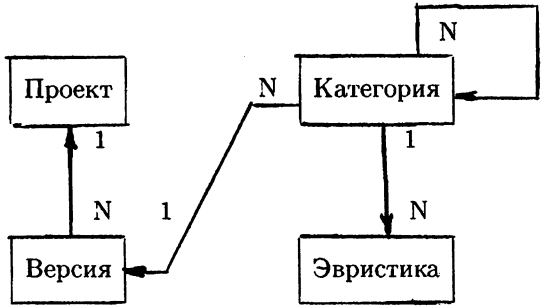

Рис.1. Модель данных

• Версия — это ветка проекта. Каждая ветка содержит рубрикацию и набор эвристик. В каждом проекте есть одна выделенная версия, над которой в данный момент происходит работа и в которую вносятся изменения. Все остальные версии являются ее снимками, сделанными в определенный момент времени для сохранения состояния и возможности отката. Каждая версия содержит метку времени о создании и текстовый комментарий. Также одна из сохраненных версий со специальной поменткой используется для эксорта в систему виртуального каталога.

• Категория — это элемент рубрикации. Категория имеет имя и описание. Для каждой категории может быть создан набор подкатегорий и набор эвристик.

• Эвристика — это уточняющий запрос, заданный для категории. Каждая эвристика содержит уточняющий запрос, наименование поисковой системы, для которой он созжан и оценку пригодности. Также для каждого запроса может быть задано использовать или нет этот запрос при экспорте.

При проектировании инструментария необходимо также учесть архитектурные особенности системы виртуального каталога. Виртуальный каталог разрабатывается как WEB–приложение и использует специальное хранилище для эвристик поиска. После подбора эвристик по команде пользователя экспериментальная система подбора экристик должна модифицировать данные в хранилище эвристик виртуального каталога. Также при проектировании и выборе технологий для разработки учитывались дополнительные положения о гибкости архитектуры разра-

батываемой системы. Необходимо было обеспечить возможность запуска на различных платформах, возможность легко и быстро расширять функциональности системы и, в том числе, расширять список доступных к использованию информационно–поисковых систем. И наконец, приоритетным является использование свободно распространяемых и открытых систем, технологий и библиотек.

Для обеспечения легкости доступа экспертов к экспериментальной системе подбора эвристик было принято решение разрабатывать систему в виде WEB–приложения. В качестве основного языка разработки было решено использовать язык Java (http://java.sun.com). Этот выбор обусловлен тем, что Java обладает множеством преимуществ. способных значительно упростить разработку системы:

- программы на языке Java могут быть запущены практически на всех платформах;
- на языке Java возможно использование современных шаблонов проектирования сложных систем;
- для языка Java существует огромное число библиотек и стандартных средств для создания мощных WEB–приложений;
- для языка Java существуют очень мощные и удобные инструменты разработки.

При проектировании архитектуры за основу были взяты шаблоны проектирования MVC (http://ru.wikipedia.org/wiki/Model–view–controller) и сервисно–ориентированной архитектуры. В связи с этим, в качестве каркаса системы было решено использовать технологию Equinox (http://eclipse. org/equinox), которая является открытой реализацией стандарта OSGi R4 и содержит в себе все необходимое для реализации модульных сервисно–ориентированных систем.

Для обеспечения разработки и последующего использования было решено использовать MySQL в качестве СУБД и Tomcat в качестве контейнера.

После анализа требований и модели данных была разработана общая модульная архитектура, в которой можно выделить следующие группы модулей.

• **Работа с базой данных.** Данный модуль содержит все функции для работы с данными системы и в рамках трехуровнего шаблона проектирования MVC реализует уровень модели. В разрабатываемой системе не предполагалось большой нагрузки на базу данных и поэтому было решено использовать технологию объективно–реляционной проекции для облегчения и ускорения разработки. В качестве реализации такой проекции была выбрана открыто распространяемая библиотека Hibernate (http://www.hibernate,org).

• **Модули основных функций.** Данные модули обеспечивают реализацию основных функций системы и в рамках трехуровнего шаблона проектирования MVC реализует уровень бизнес–логики. В них содержатся все методы для обработки сущностей системы. Другие модули не работают с моделью напрямую, а используют определенные здесь методы для вызова операций над сущностями. Это вызвано тем, что необходимо обеспечить одинаковое выполнение всех действий над сущностями и обеспечить целостность, транзакционность и выполнение всех проверок корректности данных.

• **Работа с поисковыми системами.** Модуль содержит необходимые интерфейсы и сервисные классы для работы с поисковыми службами. На данный момент в системе реализована поддержка работы с информационно–поисковыми системами Google и Яндекс. Для возможности расширения списка доступных для взаимодействия поисковых систем в этом модуле присутствует точка расширения. Используя ее можно реализовать дополнительные модули для работы с другими поисковыми системами.

• **Модули функций администрирования.** Данные модули реализуют все функции, необходимые для работы администраторов системы (на данный момент это создание проектов и экспорт в систему виртуальный каталог). На сегодняшний день эти модули реализованы в виде отдельных приложений.

• **Интерфейс пользователя.** Интерфейс экспериментальной системы был выполнен с использование технологии Rich Ajax Platform (RAP http://www.eclipse.jrg/rap). Эта технология позволяет создавать сложные и функциональные графические интерфейсы для WEB–приложений и при этом транслирует элементы

интерфейса в обычные страницы HTML и Java-Script. Для работы с приложениями, созданными с использованием RAP, не требуется загрузки дополнительного программного обеспечения. Выбор RAP в качестве технологии для разработки интерфейса позволил создать удобный интерфейс, удовлетворяющий всем функциональным и нефункциональным требованиям, и обеспечил необходимый уровень гибкости и расширяемости.

На рис.2 изображен интерфейс экспериментальной системы подбора эвристик после загрузки. Цифрами помечены основные управляющие элементы:

1. Строка меню. В этом меню содержатся команды для вызова всех функций системы.

2. Строка Toolbar. Здесь расположены пиктограммы для быстрого вызова основных функций системы.

3. Кнопки переключения между перспективами.

Остановимся подробнее на понятии «перспектива». Для многих систем, содержащих сложную функциональность или работающих с большим количеством типов объектов данных, приходится создавать очень сложный интерфейс с большим числом управляющих и отображающих информацию элементов. Но в подавляющем большинстве случаев функции любой системы можно разделить на несколько слабо пересекающихся групп по принадлежности к разным подзадачам. Обычно пользователь продолжительное время использует функции из одной группы. RAP и RCP предлагают механизм для разделения интерфейса на части по группам задач. Для каждой такой группы создается своя перспектива — предопределенный набор доступных элементов интерфейса.

В экспериментальной системе содержится три перспективы:

- онтология — перспектива для создания и редактирования рубрикации;
- эвристика — перспектива для создания, редактирования и сравнения эвристик для рубрик;
- проекты — перспектива для управления проектами. Здесь содержатся функции сохнанения и восстановления состояния проекта и экспорта в виртуальный каталог.

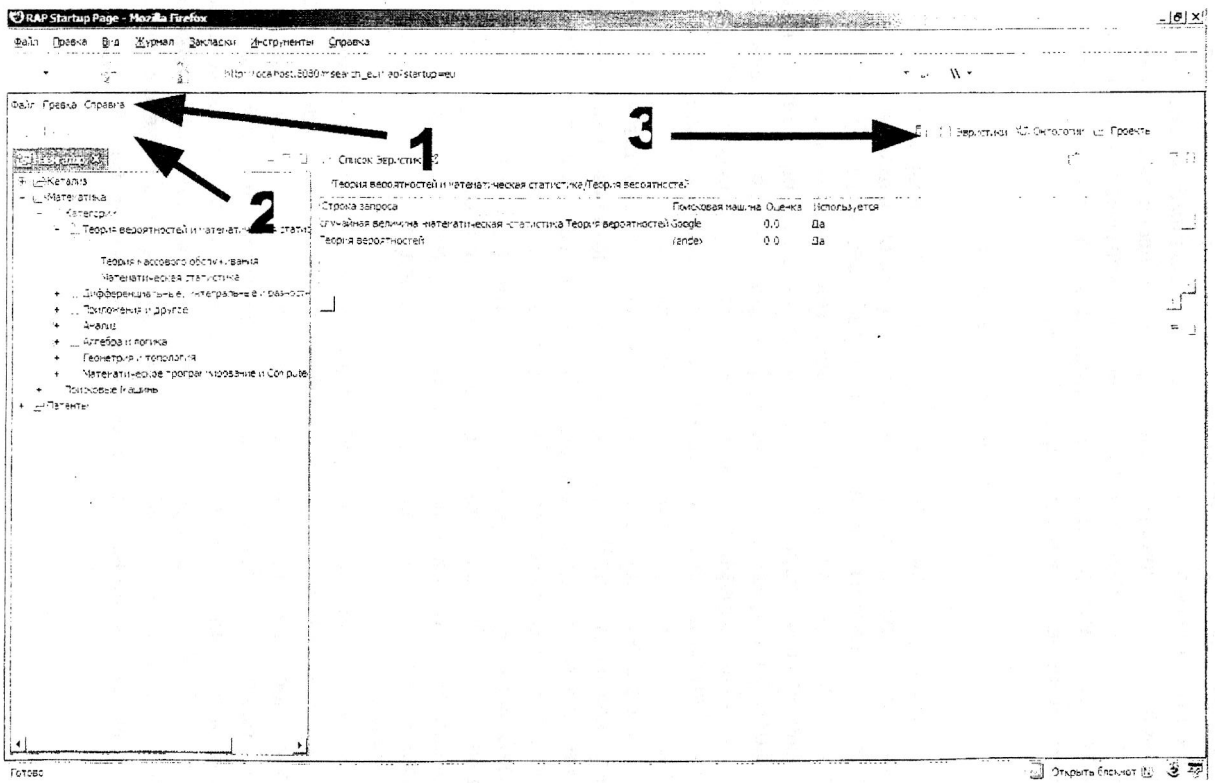

Рис.2. Интерфейс экспериментальной системы поиска эвристик

Еще одной особенностью интерфейса системы является возможность одновременно работать с несколькими объектами. Напрмер, можно одновременно открыть в нескольких закладках формы редактирования разделов или эвристик и быстро переключаться между ними. Эти закладки остаются открытыми, несмотря на переключение перспективы.

Заключение

Целью работы было облегчение процесса разработки системы виртуальный каталог в части определения онтологии и эвристик. Основным итогом проделанной работы стало создание экспериментальной системы подбора эвристик — инструментария для экспертов при разработке и поддержке системы виртуального каталога. Она представляет интерфейс для задания и изменения онтологии подбора и сравнения эвристик и реализует необходимые функции для совместной работы нескольких экспертов над созданием системы виртуального каталога.

Разработанная система в данный момент находится в тестовой эксплуатации. Эксплуатация системы показала, что разработанная архитектура и используемые технологии позволили успешно реализователь систему, Соответствующую сформулированным требованиям, и эти требования корректны в условиях поставленной задачи. В рамках тестовой эксплуатации системы проводится окончательная обкатка системы с целью уточнения списка необходимых доработок и исправлений для внедрения в последующие версии.

Сейчас продолжается работа по улучшению экспериментальной системы подбора эвристик. Планируется расширить функциональность системы и добавить возможность создания эвристик не только для рубрикации по иерархии онтологии предметной области, но и для разделения ресурсов по типам (личные страницы, статьи, библиотеки, организации и т.д.). Также, не менее важной задачей является доработка интерфейса системы с учетом пожеланий, высказанных экспертами при тестовой эксплуатации.

Литература

1. PAL'CHUNOV D.E. Logical Methods of Ontology Generation with the Help of GABEK//IV International GABEK Symposium. – Innsbruck. – 2002. – P.17.

2. PAL'CHUNOV D.E. GABEK for Ontology Hierarchy Generation// V International GABEK Symposium. – Innsbruck. – 2004. – P.5–6.

3. PAL'CHUNOV D.E. GABEK and FCA for object domain ontology creation// Abstracts of the 6th International GABEK Symposium. – Sterzing. – 2006. – P.28.

4. ПАЛЬЧУНОВ Д.Е. Моделирование мышления и формализация рефлексии 1: Теоретико–модельная формализация онтологии и рефлексии// Философия науки. – 2006. – Т.31, № 4. – С.86–140.

5. PAL'CHUNOV D.E. GABEK for Ontology Generation // Herdina Ph., Oberprantacher A., Zelger J. (eds.) Learning and Development in Organizations. Vol.2. – Berlin: Wien (LIT), 2007. – P.90–109.

6. ПАЛЬЧУНОВ Д.Е., СИДОРОВА Е.С. Виртуальный каталог// Тр. Всероссийской конф. "Знания–Онтология–Теория". Новосибирск, август 2007 г. – Новосибирск. – 2007. – С.166–175.

7. ПАЛЬЧУНОВ Д.Е. Решение задачи поиска информации на основе онтологии// Бизнес–информатика. – 2008. – № 1. – С.3–13.

Поступила в редакцию
16 июня 2008 года