Средства навигации в полнотекстовых базах данных и порталах♣

© Прохоров М.Е. Бартунов О.С.

Государственный Астрономический Институт им. П.К.Штернберга
mike@sai.msu.ru oleg@sai.msu.su

Аннотация

Объемы информации в современных информационных порталах и родственных им системах становятся все больше и для успешной работы с ними необходимы эффективные средства навигации. Даже в таких «классических» источниках информации, как книги и газеты эти средства присутствуют (тоже в самой «классической» форме): оглавления, различные индексы, нумерация страниц. Однако, средства навигации в электронных системах информации гораздо более разнообразны.

1 Введение

Основные объекты, с которыми работает пользователь портала (синонимом «работы» в данном контексте является «чтение»), это документы (статьи, книги, публикации и пр.) и списки (списки публикаций, каталоги, рубрикаторы, результаты работы поисковых систем и т.д.). Оба типа документов могут быть как статическими, так и динамическими. Если объем отдельного документа или списка не позволяет посматривать его целиком, а число документов превышает десяток, то для работы с такой системой требуются средства навигации. В большинстве современных информационных систем указанные ограничения превышаются на порядки и средства навигации являются неотъемлемыми их составляющими. Среди них встречаются как широко известные и даже «классические», так и новые малоизвестные механизмы, разработка оптимальных свойств которые в настоящее время еще не закончена.

Данная статья построена следующим образом: в первом разделе приведена классификация методов навигации в порталах и базах данных и дано их краткое описание. Остальные части посвящены более подробному описанию различных групп навигационных механизмов с реальными примерами их использования. В конце работы даны приложения и список литературы.

Большинство примеров реализации навигационных механизмов взято с сайтов поисковой системы Google [11], свободной энциклопедии WikiPedia [5,35] и сайта проекта Astronet - «рабочей площадки» авторов данной публикации [1].

2 Механизмы навигации классификация и краткое описание

2.1 Классификация

Электронные документы, как статические, так и динамические, реализуют метафору классического печатного документа, а порталы являются реализациями метафор классической библиотеки (с каталогами и индексами). В связи с этим в электронных системах реализуются все механизмы навигации, существующие в «бумажных» изданиях Эффективность электронных реализаций некоторых механизмов оказывается очень высокой, другая часть средств навигации из «бумажного» наследия в электронных системах практически не используется. Электронные реализации предлагают ряд средств навигации, которые в «бумажных» книгах и библиотеках были невозможны. Ниже мы последовательно рассмотрим эти механизмы.

В качестве основного признака для классификации механизмов навигации мы будем использовать тип используемой в них информации. Дополнительными признаками будет зона действия механизма (внутри документа, портала или архива, в рамках всей сети) и его связь с типом документа (только статическая, динамическая или любая).

Среди механизмов навигации можно выделить следующие основные группы:

- разметка встроенная в текст;
- классическая навигация внутри документа: (постраничная, оглавление, индексы и т.д.);
- навигация по рубрикаторам и ключевым словам;
- с использованием поисковых механизмов;
- кластеризация документов.

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

2.2 «Классические» механизмы

Разметка встроенная в текст.

Этот способ навигации объединяет ряд очень сильно различающихся по сложности механизмов, которые однако, выглядят как «обычные» HTML-ссылки [21] – выделенный фрагмент текста, активизация которого приходит к переходу на другую страницу, к другому документу (порталу, сайту, результату работа поисковой системы и т.п.).

Навигация внутри документа.

Постраничная навигация – перемещение по разбитому на страницы тексту. Такое разбиение чаще всего используется при просмотре списков и таблиц. Стандартными действиями здесь являются переходы на одну страницу вперед/назад, в начало/конец документа, на страницу с заданным номером и т.д.

Алфавитная навигация – используется в словарях, справочниках, энциклопедиях списках и алфавитно-упорядоченных списках.

Хронологическая навигация – используется списках выполнения работ, периодических изданий, исторических и биографических справочниках и т.д. Основными механизмами здесь являются хронологически упорядоченные списки и «метафоры» календарей. Данный тип навигации может использоваться и в рамках портала в целом.

Оглавление – показывает структуру документа и позволяет совершать переходы на его различные внутренние части.

Индексы (авторские, терминологические, объектные и пр.) – указывают в каких местах документа встречаются упоминания того или иного автора, термина и пр. Индексы могут создаваться автоматически по заданному списку слов, такая реализация близка к поисковым системам и выдает все места появления слов, в результате чего индекс будет избыточно полным. Ручное создание индексов позволяет включать в них только реально важные ссылки и делает их гораздо более информационно-ценными, но ручная индексация очень трудоемкий процесс.

Описываемые ниже методы относятся к более высокому уровню, т.к. действуют не в пределах одного документа, а в рамках портала/библиотеки или всей сети.

Классификационные методы навигации.

Данная группа методов основана на различных способах приписывания метаинформации к документам. К основным методам этой группы относятся:

Рубрикация – присвоение документу одного или нескольких значений из произвольного числа рубрикаторов системы. Рубрикаторы можно использовать для решения целого ряда задач, например, для различных видов таксономии – тематической классификации публикаций, разделению их по типу аудитории, уровню сложности и пр. Обычно рубрикаторы имеют иерархическую (древовидную) структуру. Очень важным свойством системы рубрикаторов является их «ортогональность» - разные рубрикаторы описывают независимые свойства объекта. Этим свойством должен обладать базовый набор системных рубрикаторов. Однако наряду с ними могут существовать «фасетные» или «личные» рубрикаторы, в которых это свойство отсутствует. (Например, личный рубрикатор является копией системного из которого исключены не востребованные данным пользователем разделы, а наиболее важные - уточнены или дополнены.)

Ключевые слова – привязка к публикации набора терминов (слов или словосочетаний) наиболее точно характеризующих ее содержание. В отличие от рубрикации ключевые слова берутся либо из фиксированного списка с линейной структурой, либо являются произвольными терминами, приведенными к стандартной форме.

Фолксономия – альтернатива таксономии, отличие данного метода заключающаяся в том, что рубрикаторы или списки ключевых слов с нем создаются пользователями (читателями) – так называемая «*социальная классификация*».

Букмаркинг – одна из разновидностей фолксономии. Метод основывается на составлении и публикации пользователями списков наиболее интересных или важных публикаций на различные темы. Кроме навигационных возможностей данных механизм является мощным средством привлечения интереса к порталу. Наличие в портале встроенного механизма создания подобных списков во многих случаях очень полезно.

Тэггинг – еще один вариант фолксономии или – присвоение публикациям ключевых слов (тэгов) читателями. В этом методе обычно ограничения на множество ключевых слов не накладываются.

Языковая (или *терминологическая*) *подпись* – «сжатие» документа до набора слов заданного размера, который наиболее полно описывает уникальные особенности конкретного документа. Это достаточно новое направление в развитии информационных средств анализа текстов. Данный метод по внешнему виду близок к системам ключевых слов, но построение подписи производится 1) автоматически, 2) на основе анализа текста, а не смысла документа.

Использование сведений о популярности материалов.

Мы выделяет в отдельную группу методы навигации, которые используют информацию о *популярности документов*, т.е. о количестве и типе обращений к ним за различные интервалы времени. Эти методы позволяют найти самые читаемые и самые «популярные» документы, авторов, разделы и пр. Другой тип подобных сведений – *интернет-цитируемость* документа – появление на него ссылок в интернет. Если информацию о числе обращений к документам портала можно получать внутренними методами (которые обычно оказываются более эффективными), то для

определения интернет-цитируемости требуется использование глобальной (или региональной) поисковой машины, внешней по отношению к порталу или базе данных.

Оценки популярности на основе числа обращений и цитируемости обладают неустранимой временной задержкой – от момента публикации документа до пика его популярности обращений к нему, а тем более до появления ссылок проходит некоторое время. Поэтому для априорной оценки ожидаемой популярности документа в момент его публикации можно использовать аналоги *импакт-фактора* – среднюю популярность уже опубликованных документов для данного портала или раздела портала.

Методы основанные на поиске.

Другая группа методов основывается на полнотекстовом поиске. К ним относится сам поиск в простейшем виде – обнаружение в текстах портала или базы дынных заданного слова в точной форме, так и его различные более сложные варианты (поиск слов в различных языковых формах, поиск фраз и т.д.).

Поисковые системы могут использоваться при создании других средств навигации: например, при автоматической генерации индексов, при создании языковых подписей документов и т.д.

2.3 Новые механизмы навигации

Перечисленные выше средства навигации с той или иной полнотой реализованы во всех современных порталах и полнотекстовых системах. Поэтому особый интерес представляют следующие уровни навигации, основанные на анализе метаданных всего множества публикаций.

Тезаурусы.

Систему рубрикаторов можно использовать не только как источники метаинформации для классификации публикаций, но и в качестве еще одного средства навигации в портале. Для этого необходимо реализовать для них интерактивное графическое представление. Очень полезным в этой системе оказывается применение текстового поиска по названиям рубрик.

Дополнение рубрикаторов с набором ключевых слов, используемых в портале, и с объяснением встречающихся в них терминов из словарей и энциклопедий, превращает их в *тезаурус* – онтологическую иерархию – которую можно использоваться как для непосредственного изучения структуры данной тематической области, так и в качестве независимого навигационного механизма.

Кластеризация и метрические методы.

Эти методы основываются на группировке (кластеризации) публикаций на основе использования различных типов метаинформации всего набора публикаций или некоторого их подмножества. В этих методах обычно используется попарное сравнение публикаций портала использующее некий критерий их «похожести», а на его основе определить эффективное расстояние между публикациями и ввести для них метрику.

В качестве критериев сходства могут использоваться следующие параметры:

наборы рубрик,
наборы ключевых слов,
языковые подписи документов,
количество взаимных ссылок.

Возможны и более сложные варианты критериев. Кластерный анализ данных о сходстве публикаций позволяет выделить реальные содержательные группы публикаций, т.е. ввести независимую («естественную») систему их классификации. Графическое представление подобных кластеров и связей между ними (*информационные карты*) является информационным и навигационным средством нового типа.

Контекстно-зависимые ссылки.

Последняя группа средств, которую мы хотим рассмотреть – создание ссылок на основе связанной с контекстом информации. Под ними подразумевается размещение на страницах портала ссылок на дополнительные источники, которые могут оказаться интересными для читателей данной страницы.

Простейшим вариантом контекстно-зависимых ссылок является автоматическое создание списков «смотри также», содержащих ссылки на близкие по тематике статьи, статьи с теми же ключевыми словами, статьи тех же авторов и т.п.

Второй вариант – дополнение контекстно-зависимыми ссылками результатов работы поисковой системы.

Последний, очень интересный вид ссылок данной группы, – контекстно-зависимая реклама.

Размещаемая информация может основываться только на содержании данной страницы, но может использовать индивидуальную информацию о читателе (область интересов, что читал сегодня, что – вчера, с какой страницы пришел и т.д.). Второй вариант существенно более трудоемок, но гораздо богаче по возможностям.

3 Механизмы навигации: подробности

3.1 Разметка встроенная в текст

В текст может включаться рад ссылок различных типов. Внешне все они будут выглядеть как классические ссылка в HTML [21] – активный фрагмент текста, обычно отличный по оформлению от обычного текста, взаимодействие с которым приводит к открытию переходу новую страницы (из того же или другого документа, на данном или внешнем сайте, в текущем или новом окне браузера).

Кроме стандартных ссылок на источники в документах порталов могут встречаться дополнительные их типы:

- ссылки на внутренние публикации портала по их уникальному идентификатору;

- ссылки на статьи словарей, энциклопедий, глоссариев по имени используемого термина;
- предопределенные поисковые ссылки и т.п.

Цвет и гарнитура шрифта ссылок могут содержать дополнительную информацию: например указывать на внутренние и внешние ссылки, на присутствующие и отсутствующие в портале статьи и т.п.

Поведение ссылок может также быть статическим и динамическим. Статическими являются ссылки, которые включаются в документ «как есть» или разрешаются один раз, в момент публикации документа. Динамические ссылки разрешаются каждый раз при показе документа. Этот тип ссылок соответствует так называемым «висячим ссылкам» введенным в своё время в языке SGML [20,31]. (Для повышения эффективности работы портала часто используется промежуточная стратегия работы с динамическими документами – созданные динамические страницы некоторое время хранятся в кэше и при последовательных обращениях берутся оттуда.)

Ниже приведен пример ссылок, встроенных в текст, с сайта [5]. Разным цветом показаны ссылки на имеющиеся статьи (синим) и на еще не написанные (красным).

> Расположена на правом берегу реки Лаба, напротив впадения в неё притока Фарс, в степной зоне, в 33 км северо-восточнее города Курганинск.

3.2 Навигация внутри документа

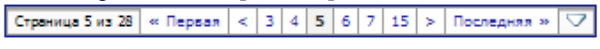
Описанные в этом разделе средства навигации представляют собой реализацию в электронном виде стандартных частей печатных документов, описанных, например, в [8].

Постраничная навигация.

Предназначена для перемещения по разбитому на равные части тексту. Такое разбиение чаще всего используется и лучше всего реализуется для просмотра списков и таблиц, в которых легко достигается разделение на равные части. Разбиение обычного текста несколько более сложная задача.

Стандартными действиями в постраничных навигаторах являются перемещения вперед/назад на одну или несколько страниц, переход на первую или последнюю страницу списка, на страницу с заданным номером.

На рисунках ниже приведены постраничные навигаторы с сайтов [14,11,1].

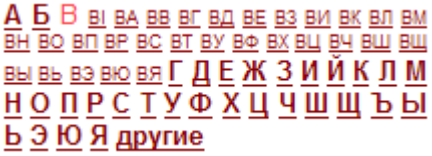


Механизму постраничной навигации присуща проблема *неустойчивости разбиения* на страницы. При фиксированном размере страницы (он задается разработчиком портала или может переопределяться пользователем) идентификация страниц производится по их номерам. В большинстве случаев, просматриваемый список пополняется с начала, а страницы нумеруются в прямом порядке, т.е. начало списка попадает на первую страницу. В этом случае добавление новой записи приводит перемещению последней записи каждой страницы в начало следующей. Таким образом, по мере пополнения списка содержание всех страниц изменяется – теряется смысл в сохранении ссылок на страницы.

Данную проблему можно решить, если изменить порядок нумерации страниц на обратный (на страницу с номером один попадает конец списка). В этом случае содержание каждой страницы остается неизменным, но возникает другая проблема – последняя по номеру страница, на которую в этом варианте попадает начало списка, будет неполной, меньшей по размеру, чем остальные.

Алфавитная навигация – используется в словарях, справочниках, энциклопедиях и других источниках, в которых содержание упорядочено в алфавитном порядке. На рисунке приведен пример алфавитного навигатора из [1].

Если списки на отдельную букву оказываются слишком блинными для удобного использования, то внутри них может использоваться постраничная навигация или основной список строится по двух- (или много-) буквенной системе. Пример последнего варианта алфавитного навигатора их [4] показан на рисунке.



Хронологическая навигация.

Используется для ориентации среди множества (множеств) событий, для которых существует явная упорядоченность по времени.

Простейшими примерами подобных, хронологически упорядоченных событий, являются списки периодических публикаций, расписания

семинаров, исторические и биографические справочники и т.п.

Другим распространенным интерфейсом хронологической навигации являются календари. На рисунке выше дан пример навигационного календаря с [1].

Для ориентации в более сложных наборах данных (например, для подробного представления последовательности событий на достаточно долгих интервалах времени) используются специальные демонстрационные или интерактивные графические приложения, см. [33].

Хронологическая навигация может использоваться как внутри документов определенных типов, так и в рамках портала в целом.

Оглавление.

Применяется с достаточно больших структурированных документах типа книги или отчета и отображает их внутреннюю структуру. По внешнему виду аналогично печатным оглавлениям, но обычно не содержит нумерации страниц (только ссылки на них).

При наличии в документе оглавления, остальные страницы обычно содержат ссылки на него.

Индексы.

Индексы – списки ссылок на страницы документа, где встречаются упоминания объектов разных типов: терминов по теме документа, имен и ли фамилий, физических объектов, аббревиатур и т.д. Сами списки обычно таются в алфавитном порядке.

Имея в распоряжении списки искомых терминов можно строить индексы автоматически – используя полнотекстовый поиск. В этом случае в индекс будут включены все места упоминания объектов в документе. В некоторых случаях это бывает полезно (например, индексы фамилий, наименований объектов или аббревиатур), в других – индекс получается слишком насыщенным (типичная ситуация для индекса терминов). В последнем случае индекс приходится создавать вручную или полуавтоматически – редактируя или добавляя флаги к автоматически создаваемым индексам – в обоих случаях это гораздо более трудоемкая работа.

3.3 Классификационные методы навигации

Рубрикация.

Рубрикатор – разновидность словаря, содержанием которого является перечень и предметных рубрик и их классификационных индексов (номеров). Одним из наиболее распространенных видов рубрикаторов являются иерархические тематические рубрикаторы (УДК, ГРНТИ, ББК, МКИ и др.). Рубрикация – присвоение документу одного или нескольких значений из произвольного числа рубрикаторов (внутренних или внешних). Рубрикатор по отдельной теме может представлять из себя плоский список разделов, но в более общем случае рубрикаторы обладают древовидной иерархической структурой (рубрики, подрубрики и т.д.). Подобная структура позволяет объединять произвольное число независимых рубрикаторов в один без потери их содержания. Классификация документов портала с помощью рубрикатора позволяет посматривать не весь портал, а только определенную его часть. На рисунке показаны два рубрикатора с [5]

[–] Статьи
- [+] Время
- [+] Информация
- [+] Культура
- [+] Наука
- [+] Образование
- [+] Общество
- [+] Природа
- [+] Религия
- [+] Страны и народы
- [+] Техника
- [+] Философия
- [+] Человек

[–] Метакатегории
- [+] Категории по алфавиту
- [+] Категории по времени
- [+] Категории по городам
- [+] Категории по религиям
- [+] Категории по тематике
- [+] Категории по морям
- [+] Категории по народам
- [+] Категории по океанам
- [+] Категории по регионам
- [+] Рейтинги
- [+] Списки
- [+] Категории по странам

Интересным и удобным расширением данного метода является создание «личных» рубрикаторов, которые содержать только часть наиболее важных для пользователя портала рубрик в порядке, также определяемом пользователем. Это позволяет быстрее перемещаться по порталу.

Ключевые слова.

Механизм основан на привязке к публикации набора терминов (слов или словосочетаний) наиболее точно характеризующих ее содержание. В отличие от рубрикации ключевые слова не объединяются в какую-либо иерархическую структуру, а берутся либо из фиксированного списка с линейной структурой, либо являются произвольными терминами, приведенными к стандартной форме. Количество ключевых слов, присваиваемых одному документу, обычно ограничивается.

Примером списка ключевых слов научного издания может служить список журнала «Astronomy and Astrophysics» [18]. На рисунке показан фрагмент списка ключевых слов [2].

- археоастрономия
- Архимед
- ассоциации, звездные
- астероидная опасность
- астероиды
- астигматизм
- астробиология
- астрогнозия
- астрограф
- астроклимат

Методы «социальной классификации».

Следующая группа классификационных методов является в определенном смысле альтернативой рубрикации и заданию ключевых слов. В них выполняются подобные действия, но при этом для классификации используются не предопределенные

разработчиками портала рубрикаторы и списки, а создаваемые пользователями самостоятельно.

Фолксономия.

Фолксономия – метод, основанный на практику совместной категоризации ресурсов посредством произвольно выбираемых ключевых слов. Это понятие относится к добровольному (спонтанному) сотрудничеству группы людей с целью организации информации в категории. Получающиеся при этом результаты заметно отличается от традиционных формальных методов фасетной классификации. Это, а также активный характер подобной классификации, привлекает к ней внимание. Как правило, такая практика возникает только в неиерархических сообществах, подобных формирующимся вокруг общедоступных web-сайтов.

Так как организаторы фолксономической информации обычно являются её же основными пользователями, фолксономия производит результаты, более точно отражающие совокупную концептуальную модель информации.

Подробнее данный подход описан в работе [25].

Тэггинг – базовый метод фолксономии основанный на присвоение объектам классификации ключевых слов (тэгов) читателями. В большинстве случаев этот термин можно считать синонимом фолксономии.

Букмаркинг – одна из разновидностей фолксономии. Метод основывается на составлении и публикации пользователями списков наиболее интересных или важных публикаций на различные темы.

Кроме навигационных возможностей данные механизмы является мощным средством привлечения интереса к порталу. Наличие в портале встроенных механизмов для ведения подобной работы может быть в ряде случаев очень полезно.

Примеры сайтов, использующих средства «социальной классификации»: служба закладок del.icio.us [13], фотоблог Flikr [15], российские сайты Мой Круг [10], Habrahabr [17].

Языковая подпись.

Языковая (или терминологическая) подпись – «сжатие» документа до набора слов заданного размера, который наиболее полно описывает уникальные особенности его содержания. Это достаточно новое направление в развитии средств анализа текстов. В этом методе активно используются поисковые технологии и предопределенные словари терминов данной области, словари стоп-слов и общеязыковой лексики.

При большой степени сжатия данный метод порождает наборы ключевых слов. При увеличении длины подписи и снижении требований к исключению стоп-слов и неспециальной лексики результатом метода становится резюме документа.

Отличие данного метода от присвоения ключевых слов состоит 1) в автоматическом проведении процедуры, 2) в том, что подпись строится на основе анализа текста, а не его смысла (как при работе эксперта). Механизмы построения подписей и автоматического резюмирования описаны в работах [22,23,34].

Дальнейшее использование языковой подписи не отличается от использования ключевых слов (например, при размещении контекстно-зависимой информации или при кластеризации публикаций).

3.4 Методы, использующие сведения о популярности материалов.

Исследования показывают, что заметная часть посетителей порталов интересуется какие публикации являются наиболее популярными, т.е. наиболее интересными. Составление списков наиболее затребованных публикаций за последнюю неделю, месяц и т.п. повышает общее число обращений к материалам портала. (Данный эффект является кумулятивным – составление списка самых популярным публикаций еще сильнее повышает их популярность.)

При сборе статистики обращений к документам чаще всего вычисляются следующие величины:

- хиты – общее число страниц загруженных пользователями;
- хосты – число уникальных IP-адресов с которых были обращения к порталу;
- «посетители», «пользователи» или «визиты» подсчитываются по уникальным идентификаторам посетителей портала (cookies). Эта величина обычно слабо отличается от числа хостов, т.к. несколько пользователей может придти с одного IP адреса или, наоборот, один и тот же пользователь может придти с нескольких IP адресов, при их динамическом выделении.

Альтернативным критерием популярности портала является интернет-цитируемость – число ссылок на него с других сайтов. Этот критерий можно определить только с помощью глобальной или, как минимум, региональной поисковой системы.

Оценки популярности и цитируемость ресурса можно определить только спустя некоторое время, после его публикации в сети. Для получения мгновенной оценки рейтинга публикации используются так называемые импакт-факторы [9].

Для сбора сведений о популярности публикаций портала могут использоваться внешние (рейтинговые системы типа rambler.ru, mail.ru и т.п.) или внутренние механизмы (основанные на анализе протоколов web-серверов портала). Оба подхода имеют широкое распространение, у каждого из них есть собственные преимущества и недостатки. С одной стороны, использование внешней рейтинговой системы позволяет сравнивать показатели различных сайтов. С другой – исследования показывают, что между результатами внешних и внутренних счетчиков посещений существует заметное систематическое различие.

Наиболее популярной рейтинговой системой в России и в русскоязычном интернете является [29], а системой интернет-цитирования – [12].

3.5 Методы основанные на поиске

Полнотекстовый поиск в порталах и архивах документов является общепринятым и широко используемым методом работы с массивами текстов. При реализации подобных методов основными являются два подхода: 1). использование внешней поисковой машины или системы, предоставляющей подобный сервис (например, www.google.gov или www.yandex.ru), 2) использование внутренней поисковой машины, работающей внутри хранилища текстов. Первый подход позволяет использовать готовые решения, но при этом приходится обходиться имеющимися (хотя и широкими) возможностями существующих поисковых машин или индексирующих программ. Второй – существенно более трудоемок, но обеспечивает максимальную гибкость и удовлетворение требований разработчиков. Следует отметить, что максимальная эффективность поиска обеспечивается, когда полнотекстовые поисковые механизмы встраиваются непосредственно в базу данных на низком уровне. Примерами таких баз данных являются Oracle [28] и PostgreSQL [27]. Описание общих принципов работы современных поисковых машин можно найти в [19, и ссылки там], функционирование механизмов полнотекстового поиска, встроенных в PostgeSQL описаны в [3].

Поиск в текстовых документах может действовать в нескольких режимах:

- поиск в тексте фрагментов точно совпадающих с заданны образцом – это базовая операция, все более сложные варианты поиска основаны на ней;
- поиск с учетом форм слов в данном языке;
- поиск с учетом синонимов;
- поиск с использованием специальных словарей терминов;
- логические (булевы) выражения для поиска (простейшие варианты: «все слова» или «любое из слов»);
- поиск с использованием регулярных выражений (шаблонов) [16];
- поиск фраз – слов в указанном порядке или с заданными интервалами между словами.

Второй, не менее важно частью поисковых систем, являются средства релевации результатов поиска, позволяющие наиболее интересные для пользователя результаты помещать в начале списка найденных документов. Типичный запрос, состоящий из 2-3 слов к глобальной поисковой системе порождает результат содержащий десятки тысяч ссылок (часто еще больше). Без разумного упорядочения подобная информация практически бесполезна для пользователя. Четкого критерия «важности» и «интересности» документов не существует, поэтому в этой области используется ряд эмпирических и полуэмпирических подходов (см., например, [30]).

3.6 Тезаурусы как средства навигации

Согласно одному из определений, тезаурус – «иерархический список терминов с определениями» [32]. С одной стороны, наличие определений сближает тезаурус со словарями или энциклопедиями (в зависимости от объема пояснительной информации). С другой – иерархическая структура делает их похожими на многоуровневые рубрикаторы.

В тезаурусах используется ряд различных видов связей межу входящими в них терминами:

- USE/USE FOR – эквивалентность терминов (синонимия). Данная связь является асимметричной, для чего среди синонимов обязательно выделяется главный термин. Иногда этот вид связей используют также для указания антонимов.
- BT (Broader Term)/NT (Narrower Term) – связь между терминами с более узким и более широким значениями. Иногда этот тип связей подразделяют на несколько категория: частное – общее, индивидуальное – видовое (коллективное). Иерархическая (строго древовидная) структура тезауруса определяется именно этим видом связей.
- RT (Related Term) – симметричная ассоциативная связь между терминами, указывающая на их сходство или взаимную зависимость. Ассоциативные связи делают структуру тезауруса более сложной (граф более общего вида, чем дерево).
- Связи межу эквивалентными терминами на разных языках, если тезаурус является многоязычным.

Тематический тезаурус, в котором задаются и поясняются термины, а также определяются связи между ними, является отображением системы знаний в данной области (domain of knowledge).

Требования, которым должны удовлетворять информационно-поисковые тезаурусы определены в стандартах [6,7]. Более подробное описание структуры тезауруса и применяющихся в нем интерфейсов приведено в тезаурусе NASA [26].

В системах навигации тезаурус может использоваться, по крайней мере, двумя способами: 1). как специальный словарь синонимов (и антонимов) для расширения полнотекстовых поисковых запросов, 2). в графическом представлении, как средство визуализации структуры области знаний и для перемещения между тематическими рубриками (в этом качестве тезаурус близок по возможностям к рубрикаторам).

Ниже показаны фрагмент страницы навигации по тезаурусу и фрагмент вертикального списка терминов (еще ниже) с сайта [1].

Разделы (id=300000001) [5397]

```
Английский для астрономов (id=300000103) [1]
Астрометрия(id=300000007) [114]
  Астрономические системы координат (id=30..
  Время, служба времени(id=300000034) [61]
  Эфемериды(id=300000036) [13]
Астрономические инструменты (id=3000000...
  Космические аппараты (id=300000031) [246]
    Межпланетные станции (id=1174148) [130]
  Радиотелескопы (id=300000032) [28]
  Телескопы (id=300000033) [79]
```

147 фраз со словами начинающимися с буквы "Q"

Q-branch
mural quadrant
quadratic effect
quadratic mean error
quadratic Stark effect
quadratic Zeeman effect
western quadrature
quadruple moment
quadruple star
magnetic quadrupole

3.7 Кластеризация

Основой этих методов является разделение публикаций портала на группы исходя из различных типов имеющейся о них информации.

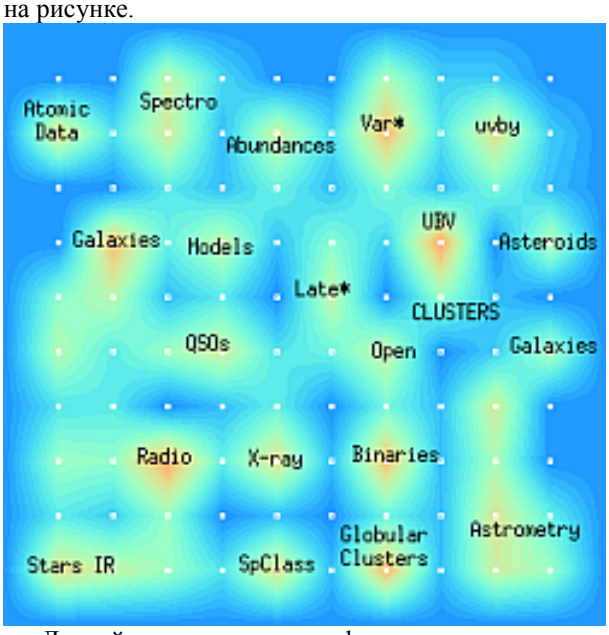
Деление на группы может проводиться непосредственным образом, например, с помощью методов распознания образов. В этом случае на выходе мы получаем заключение к какой из групп относится каждый объект (или вероятность его вхождения в соответствующую группу).

Другие методы основываются на введении расстояний между информационными объектами (задание метрики) или критерия их близости. Имея матрицу попарных расстояний можно решать либо провести глобальное разделение на группы всей совокупности объектов, либо строить «карту окрестности» отдельного объекта, отображающую взаимное расположение его ближайших соседей. В последнем случае вычисление полной матрицы взаимных расстояний становится необязательным.

Вычисление расстояний между объектами (публикациями) может производиться различными способами и на основе различной информации. Например, можно напрямую сравнивать наборы рубрик, ключевых слов или языковые подписи публикаций. Можно подсчитывать количество взаимных ссылок, в этом случае расстояния будут некоммутативными (расстояние от А до Б не будет равно расстоянию от Б до А).

Возможны и более сложные методы. Например, публикации могут считаться тем ближе, чем больше у них общих близких соседей (при этом расстояния до соседей вычисляются одним из описанных выше способов).

Глобальная кластеризация публикаций портала, основанная на различных видах метаданных, вводит альтернативные «естественные» системы рубрикации. Графические представления результатов подобной кластеризации носят названия «информационных карт» или «карт Кохонена» (Kohonen). Примером подобного сервиса, является генератор информационных карт астрономических публикаций (из журналов Astronomy and Astrophysics) [24]. Подобная карта может использоваться как для классификации предметной области, так и для навигации в портале – перехода к группам близких по смыслу публикаций. Пример созданной им карты показан на рисунке.

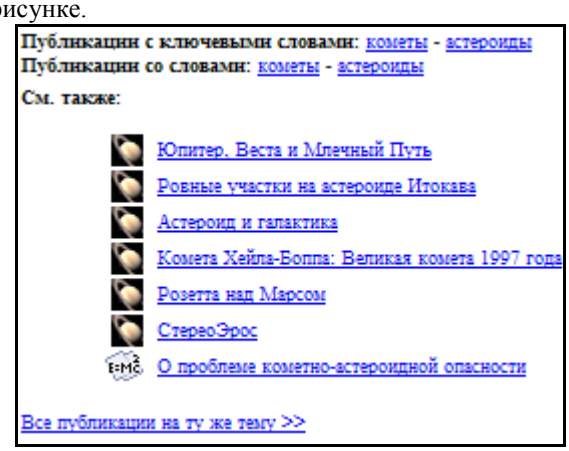


Другой пример графического средства навигации – карта «окрестностей», на которой изображены «соседи» данной публикации. Подобная система реализована в глоссарии сайта [1], пример карты показан на следующем рисунке.

На данной карте показаны «соседи» термина аберрация света 1-го и 2-го уровней. Клик по рамке с названием «соседа» приводит к построению его карты, нажатие на основной термин – к открытию соответствующей статьи глоссария.

3.8 Контекстно-зависимые ссылки

Простейший пример контекстно-зависимых ссылок – генерация рекомендация типа «смотри также». Пример подобных ссылок с [1] показан ан рисунке.



Размещаемая информация может основываться только на содержании данной страницы, но может использовать индивидуальную информацию о читателе: область интересов, что читал сегодня, что вчера, с какой страницы пришел и т.д. Второй вариант существенно более трудоемок, но гораздо богаче по возможностям. Особенно интересным может быть использование подобных средств на страницах показа результатов поиска, где для ее построения можно дополнительно использовать сведения о сделанном запросе, т.е. отреагировать на наиболее актуальные потребности пользователя.

Последний пример контекстно-зависимой информации – интернет реклама близкая по теме к основным материалам страницы. Хороший анализ контекста в данном случае охватывает нужную целевую аудиторию (т.е. эффективность подобной рекламы оказывается высокой) и не отталкивает основную массу читателей от сайта. Ниже показан умеренно удачный по нашему мнению пример контекстной рекламы от Google [11] взятый с астрономического сайта [1].

4 Заключение

Кроме технических, информационных и эргономических аспектов успешного функционирования следует учитывать также различие интересов посетителя портала и его создателя или владельца. Для первого основной задачей является как можно быстрее найти интересующую его в данный момент информацию и покинуть портал. Владелец портала заинтересован в том, чтобы посетитель прочел как можно больше материалов портала, не обязательно связанных с первоначальной целью его посещения. Очевидно, что эти цели несколько противоречат друг другу: если посетитель сразу получает ответ на свой запрос и только не наго, то он проводит на портале мало времени, хотя при этом он может часто на него возвращаться. Если для достижения своей цели посетителю потребуется выполнить слишком много шагов, то он больше не воспользуется порталом, хотя первый раз проведет на портале достаточно много времени. Одним из наиболее перспективных средств удержания пользователя на портале является публикация дополнительной информации (типа, смотри также, публикации того же автора, похожие публикации, самые популярные публикации, новости и пр.), а также внутренняя реклама публикаций портала.

Последнее тесно связано с «реанимацией» старых публикаций портала. Данная проблема состоит в следующем: на любом портале существуют самые популярные и посещаемые места для размещения: головная страница, новости, первые страницы основных тематических разделов. Для больших порталов на этих страницах помещается только небольшая доля всего контента, при этом они обычно заняты наиболее свежими публикациями. В результате не менее хорошие старые публикации оказываются фактически недоступны читателям. Для того, чтобы они время от времени появлялись перед посетителями нужна специальная политика их показа и особые механизмы ее осуществления, основанные на перечисленных выше возможностях.

Еще одна причина, по которой на портале должны присутствовать разнообразные средства навигации, причем как с простейшими, так и с развитыми интерфейсами, является реальное разделение посетителей на тех, кто знает, что он собирается искать на портале и пользуется привычными для себя возможностями портала, и на тех для у кого цель поисков плохо определена и им нужны простейшие системы поиска с достаточно разнообразные справочные материалы.

Литература

[1] Астронет, web сайт, 2000-2007. http://www.astronet.ru/

[2] Астронет: Ключевые слова, 2000-2007. http://www.astronet.ru/db/keywords.html

[3] Бартунов О.С., Сигаев Ф.Г. Введение в полнотекстовый поиск в PostgreSQL. 2007. http://citforum.ru/database/postgres/fts/

[4] Большая Советская Энциклопедия, Рубрикон, 2001-2006. http://www.rubricon.com/bse_1.asp

[5] Википедия, свободная энциклопедия (русская версия), 2007. http://ru.wikipedia.org/

[6] ГОСТ 7.24-90. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению. 1990.

[7] ГОСТ 7.25-2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. 2001.
[8] ГОСТ 7.32-2001, "Отчет о научно-исследовательской работе. Структура и правила оформления", 2001.
[9] Импакт-фактор, статья из Википедии, свободной энциклопедии, 2007. http://ru.wikipedia.org/wiki/импакт-фактор, http://en.wikipedia.org/wiki/Impact_factor
[10] "Мой Круг", сервис Яндекса, 2005-2007. http://moikrug.ru/
[11] Поисковая система Google, web сайт, 2007. http://www.google.com/
[12] Сервис интернет-цитирования Яндекс, 2001-2007. http://yaca.yandex.ru/
[13] Служба закладок (Social bookmarking), web-сайт, 2007. http://del.icio.us/
[14] Форум клуба пользователей фото продукции Minolta, web сайт, 2004-2007. http://www.forum.minolta-club.ru/
[15] Фотоблог Flickr, web-сайт, 2007. http://flickr.com/
[16] Фридл Дж. "Регулярные выражения". (Перевод "Mastering Regular Expression"). Питер-пресс. 2003.
[17] Хабрахабр, web-сайт, 2006-2007. http://habr.ru/
[18] Astronomy and Astrophysics Keywords List, Springer, 2005-2006. http://www.aanda.org/content/view/142/191/lang,en/#keywords
[19] Boswell W. Google Search Engine Basics. 2007. http://websearch.about.com/od/enginesanddirectories/a/toptenengines.htm
[20] Berglund A. CERN SGML User's Guide. 1986.
[21] HTML 4.01 Specification, 1999. http://www.w3.org/TR/html401/
[22] Johnson F.C., Paice C.D., Black W.J., Neal A.P. Journal of Document and Text Management, 1, 215-241. (1993)
[23] Johnson F.C., Paice C.D., Black W.J., Neal A.P. "The application of linguistic processing to automatic abstract generation" in "Readings in information retrieval", eds. K.S.Jones, P.Willett, Morgan Kaufmann Publishers Inc., San Francisco, USA, 1997, pp.538-552.
[24] Kohonen interface to astronomical literature, 2006-2007. http://simbad.u-strasbg.fr/A+A/map.pl
[25] Mathes A., "Folksonomies - Cooperative Classification and Communication Through Shared Metadata", 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
[26] NASA Thesaurus, NASA/SP-2007-7501/VOL1,2. 2007 http://www.sti.nasa.gov/98Thesaurus/vol1.pdf http://www.sti.nasa.gov/98Thesaurus/vol2.pdf
[27] Official Site of PostgreSQL, Web site, 1996-2007. http://www.postgresql.org/
[28] Oracle Datebese, From Wikipedia, the free encyclopedia, 2007. http://en.wikipedia.org/wiki/Oracle_database
[29] Rambler's Top 100, 1996-2007. http://top100.rambler.ru/
[30] SEOmoz: Google Search Engine Ranking Factors, Web site, 1996-2007. http://www.seomoz.org/article/search-ranking-factors
[31] SGML декларация, 2006. http://lib.custis.ru/index.php/SGML_Declaration
[32] Thesaurus. From Wikipedia. 2007. http://en.wikipedia.org/wiki/Thesaurus
[33] TimeLine AJAX widget, Web site, 2003-2007. http://simile.mit.edu/timeline/
[34] Turney P. "Learning to Extract Keyphrases from Text". Technical Report ERB-1057, Institute for Information Technology, National Research Council of Canada. 1999. http://cogprints.org/1802/00/ERB-1057.pdf
[35] Wikipedia, The Free Encyclopedia (English version), Web site, 2007. http://wikipedia.org/

Navigation into Full-text Data Bases and Portals

Prokhorov M.E., Bartunov O.S.

The English abstract for your paper for the 7th Russian Conference on Digital Libraries, taking place in Pereslavl, on the October 15th - 18th 2007. Recommended size is about half of column.

Large amount of information in present day portals and information systems requires effective methods for navigation. Navigation tools presence even into "classical" paper publications (naturally, in "classical" form). However, modern electronic navigation mechanisms sufficiently more various and effective.

♣ Данная работа поддержана грантами РФФИ 05-07-90225 и 06-07-89182.