Web технология динамической классификации квази-однородной электронной коллекции

© *Маркова Н.А., Обухова О.Л., Соловьев И.В., Чочиа А.П.*

Институт проблем информатики РАН
support@intergallery.ru

Аннотация

В докладе рассматриваются виртуальные коллекции, состоящие из независимых объектов, определяемых своими атрибутами. Предлагается формальная модель, определяющая динамическую классификацию объектов и задаваемый ею процесс навигации в коллекции как интерактивную последовательность шагов, уточняющих запрос в терминах атрибутного поиска. Наглядное визуальное представление содержимого коллекции, отвечающего текущему шагу, а также ряд специальных решений позволяет существенно повысить эффективность нахождения элементов коллекции.

1 Введение

В рамках многих востребованных приложений, реализуемых средствами web-технологии, – виртуальных музеев, виртуальных досок объявлений, Интернет – магазинов, – хранимая информация представляет собой виртуальные коллекции независимых друг от друга информационных объектов (ИО), наборы атрибутов которых частично пересечены. Такого рода коллекции являются квази - однородными, ИО в них «почти» однородны. Весьма полезным, с практической точки зрения, отступлением от однородности является разрешение ИО иметь не все атрибуты. Для квази - однородных коллекций не актуальны семантические и структурные проблемы, свойственные научным информационным системам, виртуальным архивам, библиотекам, энциклопедиям, поисковым машинам сети Интернет.

В то же время, для коллекций, объем которых составляет тысячи объектов, а пользователи существенно различаются по интересам, большинство из применяемых в настоящее время средств доступа неэффективно.

Пользователю предлагаются два базовых способа найти информацию – поиск и навигацию. Поиск осуществляется по запросу, указывающему параметры искомых объектов. Навигация – это последовательность шагов, где каждый шаг соответствует выбору некоторой группы объектов или индивидуального объекта.

Построение навигационной схемы, как правило, связано с классификацией ИО. Дерево навигации представляет собой рубрикатор соответствующей классификации. При удачном разбиении на рубрики минимизируется число шагов для доступа к ИО. К сожалению, разбиение не может быть «хорошим» для всех. Многие сайты коллекций предоставляют несколько вариантов иерархий. В реальной жизни количество вариантов, интересных для разных пользователей, весьма велико.

Поиск, в том виде, в котором он представлен в большинстве функционирующих коллекций, де-факто реализует принцип фасетной классификации ИО. Фасетная система классификации [1], в отличие от иерархической, позволяет выбирать признаки классификации независимо друг от друга. Признаки классификации называются фасетами (они соответствую «атрибутам» в терминах информационно - поисковых языков). Каждому фасету отвечает некоторое множество возможных значений. Группа пар фасет - значение называется фасетной формулой. Каждая фасетная формула определяет некоторое подмножество ИО, интересующее пользователя. Как правило, фасетную формулу визуальный интерфейс web-сайта задает с помощью набора фильтров. Каждый фасетный признак определяется строкой или выбором из списка значений («комбо-боксом»). Учитывая, что наличные ИО отнюдь не обязаны покрывать всевозможные сочетания классификационных признаков, типичным ответом на запрос пользователя будет: «*К сожалению, объекта, удовлетворяющего вашему запросу, не найдено. Попробуйте, пожалуйста, поискать еще раз, изменив параметры*»

Мы проанализировали несколько десятков типовых сайтов, представляющих доступ к квази - однородным коллекциям для различных предметных областей и использующих атрибутный поиск по совокупности фильтров - фасетов. Количество фасетов от 4 до 20, типично 8 - 10. Количество значений фасетов от 3 до 300. Как показывает практика, после десятка «промахов», пользователь переходит на более длительный, но надежный навигационный путь.

Нам хотелось бы объединить достоинства обоих подходов, избежав присущих им недостатков. Мы хотим предоставить пользователю на каждом шаге

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

навигации возможность выбора из динамически формируемой классификационной схемы, представляющей только те атрибуты (фасетные признаки), которые имеются у существующих объектов. При этом мы должны следить за длиной пути и трудоемкостью каждого шага, дабы процесс навигации был бы эффективным.

Предложим формализм хранения данных и процесса навигации коллекции в терминах фасетов, рассмотрим проектные решения визуального интерфейса, реализующего фасетную навигацию.

2 2 Модель динамической классификации

Представим формальную модель, определяющую содержание квази-однородной коллекции, а также процесс ее динамической классификации, представляющий основу для навигации.

Коллекция состоит из ИО, каждому из которых сопоставлены атрибуты. Атрибуты, значения которых могут быть перечислены или/и упорядочены таким образом, чтобы служить базой для классификации, – будем называть **фасеты**. К таким атрибутам относятся: *Вид*, *Категория*, *Автор*, *Цена*. Несколько объектов могут обладать одним и тем же значением (или входить в определенный диапазон значений) фасета, образуя тем самым классификационную группу. Атрибуты с уникальными значениями, интерпретация которых прерогатива пользователя – *Название*, *Описание*, *Изображение*, – будем называть **образами**. Образы – это данные для просмотра и детальной визуализации. Их рассмотрение выходит за рамки настоящего доклада

В коллекции $G = \{g\}$ фасет φ сопоставляет объекту некоторое значение из множества допустимых значений данного фасета A^{φ}

$\varphi : G \to A^{\varphi}$

$\Phi = \{\varphi\}$ - множество фасетов на G.

Также как и конструкции информационно-поисковых языков, фасеты предназначаются как для описания фактов с целью последующего хранения и поиска, так и для формирования информационных запросов [2].

2.1 Классификация объектов с помощью фасетных формул

Каждому объекту сопоставим фасетную формулу, представленную в виде последовательность пар: фасет - его значение на этом элементе, для всех фасетов из Φ. На практике для конкретного объекта имеет смысл рассматривать только те фасеты, которые на нем определены. Под фасетной формулой объекта (*FF)* будем понимать

$$FF(g) = <(\varphi, \varphi(g))> \;\middle|\; \varphi \in \Phi \wedge \varphi(g) \neq \varepsilon$$

Фасетная таблица, содержащая фасетные формулы для всех объектов коллекции, представляет данные для всех возможных вариантов их классификации:

$$FT(G) = \{FF(g) \,|\, g \in G\}$$

2.2 Динамическая классификационная группа - фасетная формула для формирования запроса

Использование коллекции сопряжено с выборкой объектов. Пользователь определяет выборку некоторого подмножества $RG \subset G$, т.е. свою классификационную группу, в терминах фасетов:

$$RG = \{g \in G \;|\; \varphi_i(g) \in A_i^{\varphi_i} \subset A^{\varphi}, i = 1,\ldots\}$$

Для простоты рассмотрения будем считать, что должны выполняться все фасетные отношения одновременно (связка «И»).

Определим фасетную формулу запроса как последовательность пар: фасет - подмножество значений, к которому должно принадлежать его значение для искомых объектов:

$$RFF = <(\varphi_i, A_i^{\varphi_i})> \;\middle|\; i = 1,\ldots$$

Операция выборки (*retrieve*) формирует подмножество в соответствии с фасетной формулой запроса:

$$RG = \text{retrieve}(G, RFF)$$

2.3 Навигация в терминах фасетных формул

Определим **навигацию** как итеративный процесс построения фасетной формулы запроса. На каждом шагу пользователь уточняет формулу, выбирая фасет и задавая подмножество его искомых значений. Естественно, что один и тот же пользователь в разных случаях, а тем более различные пользователи предпочтут различные траектории выбора фасетов, что соответствует всевозможным вариантам классификационных схем. Задача навигационного инструментария - обеспечить, чтобы формируемому запросу всегда соответствовала непустая выборка, тогда движение по любой траектории приведет хотя бы к одному ИО.

В начале, на 0-шагу, фасетная формула запроса пуста $RFF_0 = <>$, выборка соответствует всей коллекции $RG_0 = G$, а допустимые фасеты – всем фасетам коллекции $\Phi_0 = \Phi$

На очередном шагу рассматриваются только те фасеты, которые имеют непустые значения на объектах, принадлежащих к выборке предыдущего шага (фасет **доступен**), и только эти значения будут использоваться для уточнения подмножества искомых значений (значение фасета **доступно**).

$$\Phi_n = \{\varphi \in \Phi_{n-1} \,|\, \exists g \in RG_{n-1}\, \varphi(g) \neq \varepsilon \ \}$$

$$RFF_n = RFF_{n-1}; (\varphi_n \ , A_n^{\varphi n})$$

$$\left|\varphi_n \in \Phi_n \wedge A_n^{\varphi n} \subset \{\varphi_n(g) \,|\, g \in RG_{n-1}\}\right.$$

$$RG_n = \text{retrieve}(RG_{n-1}, RFF_n)$$

3. Эффективность навигации

Эффективность навигации целесообразно определять трудоемкостью нахождения ИО: в числе шагов и числе операций на каждом шагу.

Операции текущего шага включают выбор фасета и определение подмножества искомых значений. В простейшем случае пользователь выбирает доступный фасет, а затем выбирает его доступное значение. Но, как отмечалось во Введении, и значений и фасетов может быть довольно много.

Два вида решений могут способствовать повышению эффективности навигации. Структурные решения упорядочивают множества значений фасетов. Презентационные решения предполагают специальные приемы визуального интерфейса.

В соответствии с принципом, сформулированным американским психологом Миллером [3], для того, чтобы выбор был эффективным, количество элементов в нем не должно быть больше семи-девяти.

Два возможных пути сокращения числа вариантов выбора – уменьшение числа возможных значений фасета и их группировка, – иллюстрируют следующие структурные решения.

1. Многоаспектность – многозначность фасетов. Открытость в систематизации объектов (то, что при занесении нового объекта ему может быть присвоено новое значение фасета), как правило, приводит к тому, что число возможных значений фасета неограниченно возрастает. Число новых значений может быть сокращено, если разрешить сопоставлять объекту более одного значения. Например, «пейзаж» + «библейский сюжет».

2. Естественная структуризация значений фасета. Во многих практически значимых приложениях множество значений для фасета имеет иерархическую структуру. Выбор значения из иерархии существенно быстрее, чем из списка.

3. Искусственная структуризация значений фасета. Там, где значения фасета семантически не разбиваются на классы, а значений много, возможна искусственная иерархия, основанная на формальных признаках. Например, фасет «*Автор*» может быть надстроен уровнями по первой букве, первым трем буквам.

В визуальном интерфейсе каждому фасету, а также его значению или группе значений сопоставляется визуальный элемент.

Эффективности навигации будет способствовать следующие решения:

1. Недоступные фасеты и недоступные их значения невидимы (в некоторых случаях их можно показывать закрытыми на ввод).

2. Возможность разворачивания-сворачивания визуальных элементов. Если на экране несколько визуальных элементов – фасетов с входящим в них визуальными элементами – значениями, то выбор фасета и его значения производятся за одно указание подчиненного элемента. Для активного (вошедшего в фасетную форму) фасета визуальный элемент можно свернуть, дабы не загромождать экран.

3. Численная ориентировка. Текущий шаг навигации, а также варианты следующего шага сопровождаются числом объектов в соответствующей выборке (цифрами или другой визуальной формой). Если число объектов невелико, возможно, пользователь остановит процесс навигации и перейдет к просмотру.

4. Визуальный интерфейс фасетной навигации

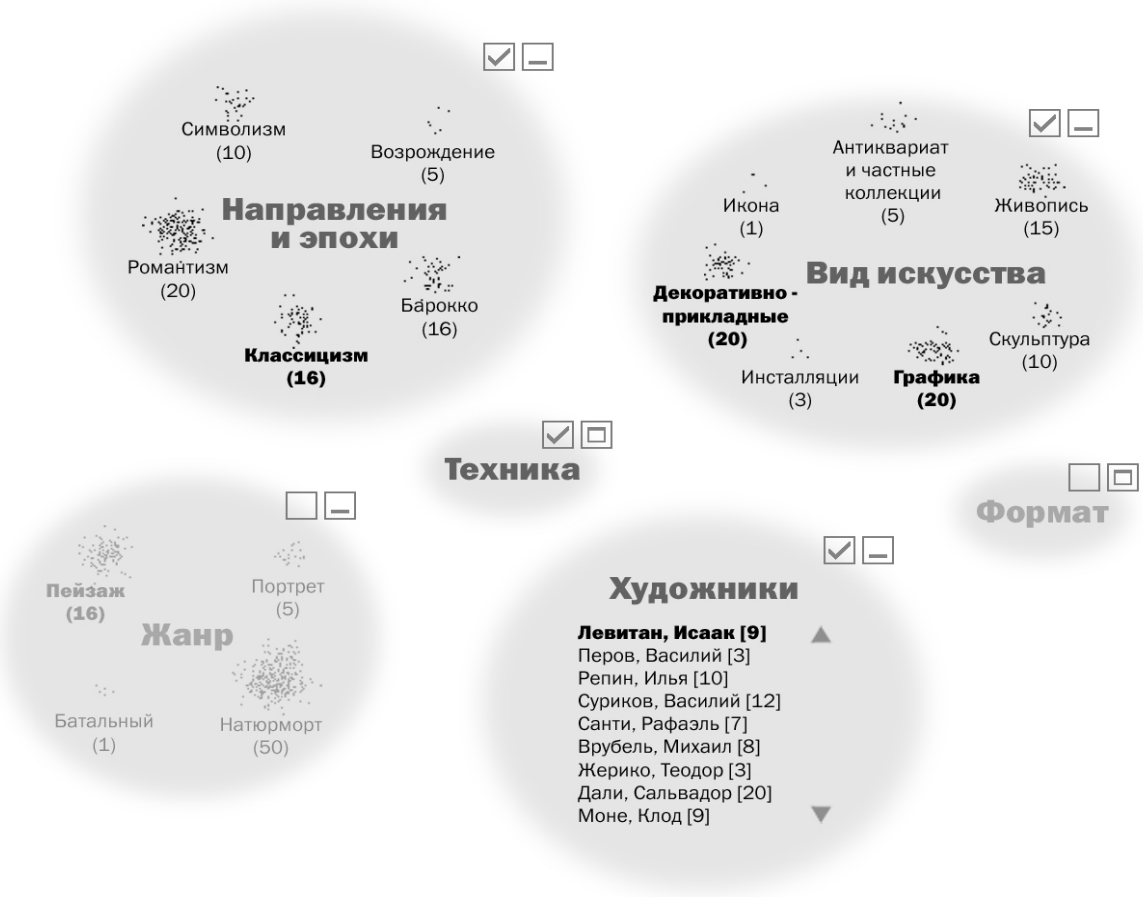
Представим одну из возможных форм визуального интерфейса фасетной навигации, реализующую перечисленные презентационные решения.

Фасеты представлены в виде «облаков», которые изначально не активны.

В представленный момент не активны фасеты Жанр, Формат. Они, соответственно, не участвуют в формировании фасетной формулы. Пользователь выбирает фасет и его значение за одно указание – «клик» на значение. Набор доступных значений в каждом фасете зависит от текущей фасетной формулы запроса.

После выбора фасета - его значения, соответствующая пара добавляется к фасетной формуле запроса. Тем самым за одно указание выполняется шаг навигации. Картина во всех открытых «облаках» меняется, дабы соответствовать новой фасетной формуле.

Рядом со значением отображается количество объектов. В цифровом виде и в виде точек разной густоты, что дает легко интерпретируемую картинку.



Заключение

Навигация представляет лишь часть функций, необходимых для работы электронной коллекции. Вопросы занесения данных, хранения информации, администрирования и технической поддержки, о которых в настоящей статье не говорилось, существенным образом переплетены с проблемами реализации навигации.

В дальнейшем планируется реализация нескольких вариантов интерфейса для различных по числовым характеристикам (числе и объемам фасетов, числе объектов и их разнородности) коллекций, которые помогут оценить эффективность предложенных решений.

Планируется продемонстрировать предлагаемые идеи фасетной навигации на примере конкретной реализации - виртуальной коллекции произведений искусства. Ориентировочные параметры реализации: 10 фасетов, количество значение фасета от 7 - 300, число объектов 10000. Реализация будет осуществлена в среде ОС: Microsoft Windows 2000/2003; Веб-сервер: Microsoft Internet Information Server 5.0/6.0, контейнер серверного приложения: Microsoft ASP.NET 2.0; СУБД: Microsoft SQL Server 2005. В качестве средств разработки используется Microsoft Visual Studio 2005, ASP.NET 2.0.

Литература

[1] *Ранганатан Ш.Р.* Классификация двоеточием //Основная классификация. М.: ГПНТБ СССР, 1970.

[2] ГОСТ 7.74-96 СИБИД. Информационно-поисковые языки. Термины и определения ГОСТ от 27.11.1996 N 7.74-96

[3] *Миллер Дж.* Магическое число семь, плюс или минус два. - В кн.: Инженерная психология. М. 1964.

Web-technology of dynamic classification in quasi-homogeneous digital collections

Natalia Markova, Olga Obuhova, Ivan Soloviev, Anton Chochia

Merging the benefits of attributive search and navigation would allow the user to navigate a digital collection by progressively selecting desired facet values of information objects.

The paper presents facet navigation based on dynamic classification - fast and easy drill down in collections by selecting attributes. The formal representations of “facet formulas”, “facet table” and “facet request” are proposed. Several design decisions that can improve efficiency of facet navigation are discussed. A sample of visual interface snapshot illustrates the main ideas of the paper.