Визуализация поиска информации в репозитории ОИЯИ

© Т. Н. Заикина © И. А. Филозова © Ж. Мусульманбеков
Объединенный институт ядерных исследований
Дубна
ztanya@jinr.ru Irina.Filozova@jinr.ru genis@jinr.ru

Аннотация

Обсуждаются методы создания пользовательского интерфейса с визуализацией поиска информации на базе данных репозитория научных публикаций и документов Объединенного института ядерных исследований (ОИЯИ) JINR Document Server (JDS). Представлена разработка двух вариантов визуализации поиска и представления документов в JDS.

1 Введение

В настоящий момент наметилась тенденция кроме традиционного поиска информации использовать методы визуализации навигации и представления данных. Применение методов и средств визуализации в электронных библиотеках открывает пользователям новые возможности “увидеть” их содержимое с разных точек зрения и существенно повышает информативность и эффективность электронных библиотек [12].

В 2010 г. был создан репозиторий научных документов ОИЯИ – JINR Document Server (JDS). Существующий пользовательский интерфейс JDS предоставляет пользователям стандартный способ вывода результатов поиска. Было принято решение создать графический вариант пользовательского интерфейса JDS с визуализацией поиска информации для предоставления дополнительных возможностей пользователям.

В разделе 2 рассматриваются проблемы и особенности предметной области, описан процесс создания и основные стратегии визуализации информации, выделены особенности применения визуализации информации в электронных библиотеках. В разделе 3 описаны методы визуализации документов JDS, реализованные в виде двух прототипов. Так же в приложении приводится краткое описание программного продукта Prefuse, выбранного для создания визуального интерфейса.

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

2 Визуализация информации

Основатели направления визуализации информации — Стюарт Кард, Джок Маккинли и Бен Шнейдерман (Stuart K. Card, Jock D. Mackinlay, и B. Shneiderman) — дают следующие определения [2]:

Визуализация — использование созданных с помощью компьютера интерактивных, визуальных представлений данных.

Визуализация информации — использование созданных с помощью компьютера интерактивных, визуальных представлений абстрактных данных для лучшего восприятия.

Ключевая задача визуализации информации заключается в создании понятного пользователю графического отображения набора данных, а также в использовании интерактивных технологий, которые бы упростили работу с данными и позволили пользователю изучать их интуитивно [6].

Визуализация информации основывается на знаниях таких областей как компьютерная графика, человеко-машинное взаимодействие, графический дизайн, когнитивная психология, статистика, картография и изобразительное искусство. Синтез соответствующих идей из этих областей с новыми методологиями и технологиями позволяет человеку с помощью компьютерного взаимодействия справляться с огромными потоками данных в современном мире.

Продвижение визуализации информации в значительной степени было обусловлено исследованием поиска информации. Основная проблема поиска информации — это повысить его эффективность и результативность.

Визуализация информации — процесс анализа и преобразования абстрактных данных в визуальную форму для того, чтобы улучшить наше восприятие больших объемов данных.

Визуализация тысячи объектов с различных точек зрения позволяет увидеть более полную картину и получить больше информации об отображаемых объектах, что позволяет более эффективно исследовать полученную информацию, а также повышая осведомленность пользователя об объектах.

Хорошо продуманная визуализация позволяет:

- Предоставить возможность охватить огромные объемы данных.
- Снизить затраты времени на поиск.
- Обеспечить более глубокое понимание сложного набора данных.
- Отображать отношения между данными.
- Рассматривать набор данных с различных точек зрения одновременно.

2.1 Процесс создания визуализации

Процесс создания визуализации — это процесс преобразования информации в визуальное представление, с которым могут взаимодействовать пользователи.

На первом шаге необработанную информацию следует реорганизовать в набор данных, выделить объекты, которые будут визуализироваться, и отношения между этими объектами. Объекты — это предметы интереса, которые важно визуализировать, а отношения — это связи между объектами, которые и формируют структуру визуализации. И объекты и отношения имеют атрибуты. Атрибутом является свойство объекта, которое не может существовать независимо от него. Так, например, цвет яблока это атрибут яблока, а температура воды это атрибут воды. Но иногда решить, что должно быть атрибутом, а что объектом не так просто. Например, у нас есть объект «рабочие», у которых атрибутом может быть заработная плата, но так же объектом у нас может быть и «количество зарабатываемых денег», в таком случае необходимо определить отношение между объектами «рабочие» и «количество зарабатываемых денег» [4].

Второй шаг — «сердце» процесса визуализации. Преобразование набора объектов в графическую форму. Графическая форма состоит из символов, которые соответствуют набору объектов.

Третий шаг — внедрение графического отображения в представление, которое будет выведено на экране и обеспечивать множество видов преобразований, таких как навигация, масштабируемость изображения и т.д.

Четвертый шаг – взаимодействие пользователя с визуальным представлением, с последующей интерпретацией информации.

Данный процесс преобразует информацию в визуальные представления.

Процесс визуального отображения — начальная точка для создания визуализации. Чем разнообразнее информация, которую нам нужно визуализировать, тем в большем количестве методов создания визуальных представлений мы нуждаемся. Важным шагом при разработке методов визуализации информационных данных является их структуризация. Структура обеспечивает высокий уровень организации набора данных и часто является решающей в выборе метода для создания подходящей визуализации.

2.2 Структуры информации

Структура отображает пространственное расположение атрибутов информационных данных и исходный вид для последующей визуализации. Выделяют следующие структуры информации [4]:

Табличная структура

Таблицы состоят из строк (объектов) и столбцов (атрибутов). Данный тип структуры информации часто используется для многомерных данных, так как каждый атрибут определяет размерность пространства данных, в рамках которой каждый объект является одной точкой.

Структура деревьев

Деревья, отображающие иерархии являются самым распространенным способом для отображения информационных данных. Например, организационная структура файловой системы может быть представлена иерархически. Визуализация иерархий является одним из самых зрелых и активных ответвлений в визуализации информации. Древовидная структура данных, играя существенную, самостоятельную роль в визуализации информации, как правило, является частью представления более сложных структур [3].

Структура сетей

Существует много эффективных алгоритмов визуализации для иерархических структур сетей. Обычно используемая стратегия состоит в том, чтобы упростить сеть, извлекая древовидную структуру и затем применить эффективный древовидный алгоритм визуализации.

Структура коллекций документов и текста

Структуры текста или коллекций документов могут быть распределены по другим информационным структурам. Например: заголовки — структура деревьев, метаданные — табличная структура, ссылки — структура сети.

2.3 Стратегия обзора (overview strategy) в визуализации информации

Методы разработки визуальных представлений больших объемов данных — основная проблема визуализации информации. Чем больше информации, тем сложнее ее всю отобразить на возможном пространстве экрана. Нам бы просто не хватало пикселей. Но даже если бы пикселей было достаточно, визуализация выглядела бы слишком загроможденной. Обычно, разработанная визуализация отображает все детали только небольшой группы объектов из всего набора. Аналогом может служить «заглядывать в большую комнату сквозь замочную скважину», — проблема «замочной скважины». Для визуализации большого объема данных, Бен Шнейдерман (B. Shneiderman) предложил правило, которое гласит: сначала обзор, масштабирование и фильтрация, затем разделение по назначению. Идея заключается в том, чтобы сначала ознакомить пользователя с полным набором данных, а затем обеспечить механизмы взаимодействия, позволяю-

щие пользователю приблизить интересующую его информацию и отфильтровать ненужную, и в заключение быстро извлекать и отображать подробную информацию о группе выбранных пользователем объектов или объекте [4].

Для обзора большого количества информации используются два подхода:

1. Сокращение количества данных до их отображения.
2. Сокращение физических размеров отображаемых данных.

Сокращение количества отображаемых данных

Один из методов сокращения отображаемых данных — объединение. Объединение объектов в группы создает новый набор объектов с меньшим количеством входных данных.

Первой задачей объединения информации в группы является выявление объектов, которые следует объединить в группы. Объекты могут быть объединены по общим значениям атрибутов или с помощью алгоритмов кластеризации, или по ближайшему соседу. Следующая задача — определение новых атрибутов уже для группы объектов. В идеале, атрибуты группы должны отображать атрибуты объектов, из которых она состоит. Агрегирование может многократно применяться для создания групп и подгрупп древесных структур. Также графическое представление групп должно отображать их содержимое.

Сокращение размеров графически отображаемых данных

Чем больше объектов нам необходимо отобразить, тем меньше по площади должно быть графическое отображение объекта (правило Эдварда Тафти (Edvard R. Tafte: “data: ink ratio”)). Использование данного правила помогает убрать излишки графического представления объектов в визуализации [5].

2.4 Стратегии навигации в визуализации информации

Для поддержки навигации по пространству отображаемых данных используются интерактивные методы. Существует три навигационные стратегии:

- Масштабирование и панорамирование (Zoom + Pan);
- Обобщение и детализация (Overview + Detail);
- Фокус + Контекст (Focus + Context).

Далее данные стратегии описаны более детально.

Масштабирование и панорамирование

Визуализация начинается с общего вида информационных объектов. Пользователь, интерактивно выделяя интересующий его объект, может приблизить отображение, а потом опять вернуться к просмотру общего вида, после чего, проделать те же действия, но уже с другими интересующими его объектами визуализации.

Обобщение и детализация

Эта стратегия использует множественные представления для одновременного вывода на экран общего вида и детализации отдельных объектов.

Фокус + Контекст

Основная идея «фокус + контекст» состоит в том, чтобы позволить видеть интересующий объект с возможностью представления его в деталях и одновременно общий вид всей окружающей его информации [11].

«Фокус + Контекст» состоит из трех основных предпосылок: Во-первых, пользователь нуждается и в обзоре (контекст) и в подробной информации (фокус) одновременно. Во-вторых, информация, необходимая в обзоре, может отличаться от детализируемой. В-третьих, эти два типа информации могут быть объединены в пределах одного представления.

Положительные и отрицательные стороны каждого из подходов

Преимуществами стратегии «Масштабирование и панорамирование» (Zoom+Pan) являются:

+Эффективное использование пространства экрана.

+Многоуровневое масштабирование.

Недостатками являются:

-Потеря общего вида при приближении.

-Медленная навигация.

Преимуществами стратегии «Обобщение и детализация» (Overview+Detail) являются:

+Стабильный краткий обзор.

+Масштабируемость, фокусирование взгляда.

Недостатками являются:

-Визуальный разрыв между общим видом и детализацией.

-Отображения краткого обзора и детализации конкурируют за место на экране, небольшой обзор.

Недостатками стратегии «Фокус + Контекст» (Focus+Context) являются:

+Детали визуализации соединены с окружающим контекстом.

Недостатками являются:

-Ограниченная масштабируемость.

-Искажение.

2.5 Интерактивное взаимодействие

Стратегии взаимодействия используются для поддержки масштабируемости при визуализации информации со сложной структурой. Графически предпочтительно выводить все данные на дисплей в форме, которая эффективно отображает суть визуализации. Сделать это без интерактивного взаимодействия обычно невозможно, даже для данных небольшой сложности. Стратегии взаимодействия устраняют ограничения экрана, позволяя пользователям в интерактивном режиме исследовать большее количество информации. Далее описаны основные стратегии взаимодействия, которые желательно учитывать в каждом проекте визуализации [4].

Выбор (Selecting)

Одна из основных потребностей визуализации — интерактивный выбор отдельных объектов или их подмножества. Пользователи выбирают объекты для идентификации данных, которые представляют для них интерес. Например: просмотр подробной информации об объекте, выделение объектов, которые затенены или закрыты, группировка ряда связанных объектов, или извлечение объектов для будущего использования. Пользователи могут выбирать объекты, непосредственно указывая на их графическое изображение или группируя объекты с помощью мыши и т.д.

Фильтрация (Filtering)

Интерактивная фильтрация позволяет пользователям уменьшить количество информации на дисплее и сфокусироваться на интересующей информации.

2.6 Визуализация и электронные библиотеки

В настоящий момент электронные библиотеки в основном используют стандартный способ вывода результатов поиска, однако, такие поисковые интерфейсы не предоставляют возможность проводить исследование информации.

Цель применения визуализации информации к интерфейсам электронных библиотек состоит в том, чтобы переключить восприятие информации с более медленного чтения (текста) на более быстрые перцептивные способы, такие как распознавание образов.

Исследования, проведенные в данной области, определили несколько правил, как сделать использование электронных библиотек более эффективным [8, 9]:

• Провести исследование, основанное на подробном анализе информационных потребностей и задач пользователей;

• Обеспечить взаимодействие с данными, позволить пользователям фильтровать документы и исследовать связи (отношения) между этими документами;

• Используя ссылки — выделять только значимые слова и, для лучшего восприятия, отображать их не перегружая текст;

• Представлять одну и ту же информацию с разных точек зрения;

• Поддерживать поиск текста, изображения, видео;

• Предоставлять персональные корзины для пользователей, для хранения выбранных наборов документов.

Выделяют три возможности предоставляемые пользовательскими интерфейсами с визуализацией информации в электронных библиотеках [8]:

• Идентификация полученных результатов, обнаружение связей между полученными документами и усовершенствование поиска;

• Обзор документов электронной библиотеки и простота просмотра;

• Отображение результатов взаимодействия пользователя с доступными документами для оценки и улучшения качества электронных библиотек.

3 Разработка визуализации информации в JDS

Визуализация информации в JDS реализована в виде графического пользовательского интерфейса с визуализацией поиска информации, предоставляющего пользователю ряд дополнительных возможностей.

Прежде всего, формируется соответствующий XML файл, использующий данные, хранящиеся в базе данных JINR Document Server. Далее XML файл используется программной библиотекой Prefuse (см. приложение), с помощью которой создается визуализация.

Разработаны два визуальных представления, каждое из которых отображает определенную информацию. Были выбраны следующие методы: радиальный граф (radial graph) и древовидная карта (treemap).

3.1 Визуальное представление информационных ресурсов JDS с помощью метода древовидной карты

Данный метод визуализации был введен Б. Шнейдерманом и Б. Джонсоном (B. Shneiderman, B. Johnson) в 1991 году. Первоначально метод был применен к визуализации распределения дискового пространства между файлами в иерархической структуре директорий, но постепенно получил широкое применение во многих областях визуализации от финансовой информации до результатов спортивных состязаний.

Базовая идея метода состоит в том, чтобы изобразить дерево, каждая вершина которого имеет имя и численный атрибут в виде прямоугольника. Для изображения поддеревьев используется рекурсивная процедура разбиения прямоугольника, соответствующего всему дереву, на прямоугольники меньшей площади без незаполненных пространств и наложений. Для этого площадь прямоугольника разбивается горизонтальными и вертикальными линиями на прямоугольники, площадь которых пропорциональна значению данного атрибута. Такая разновидность метода называется методом продольно-поперечных разрезов (Slice-and-Dice).

Сильные и слабые стороны метода

К сильным сторонам метода древовидной карты относятся такие свойства, как:

• Использование 100% площади экрана;

• Наследование иерархических уровней;

• Отображение атрибутов узла (размерам площади узла и цвет).

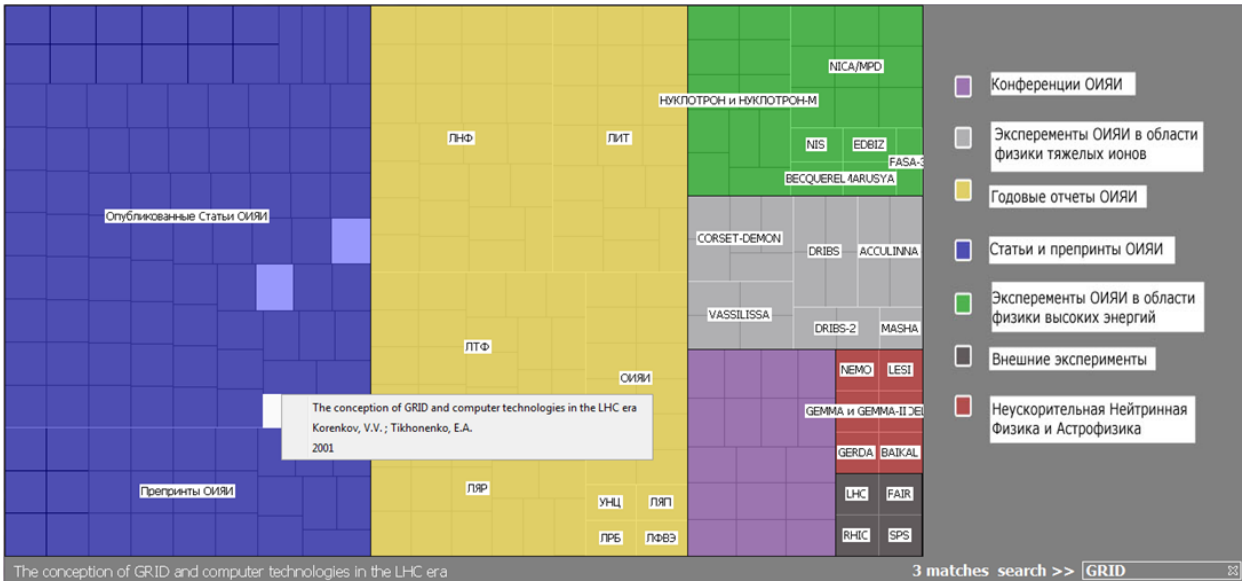

Рис. 1 *Визуальное представление информационных ресурсов JDS с помощью метода древовидной карты.*

К слабым сторонам метода древовидной карты относятся такие свойства, как:

• Размеры узлов по площади не всегда удобны для сравнения;

• Проблема подписи каждого из узлов;

• Трудно различимые границы между узлами.

Метод древовидной карты крайне эффективен при изображении численных атрибутов элементов (размер, стоимость, значение), организованных в большие иерархии. Так как нам необходимо отображать иерархии большого набора данных, а одним из основных атрибутов является количество публикаций, сравнив сильные и слабые стороны, мы сочли данный метод наиболее подходящим для нашей задачи.

Данный метод использовался для создания визуальных отображений по тематикам:

(1) *предметных коллекций;*

(2) *хранящихся в теге 650.*

(1). В визуализации публикаций по тематикам предметных коллекций, каждая предметная коллекция отображается соответствующим цветом (Рис. 1).

Чем больше пространства окрашено определенным цветом, тем больше публикаций в данной тематике. Данная визуализация наглядно отображает распределение публикаций по тематикам предметных коллекций. Следующий уровень иерархии отображается в виде записи, и конечные публикации отображаются в виде интерактивных прямоугольников. Представление также содержит строку поиска, найденная публикация выделяется соответствующим цветом. Публикация является ссылкой на саму себя. При наведении курсора на публикацию появляется информация об ее авторе, соавторе, дате опубликования. При нажатии на правую клавишу мыши можно масштабировать изображение. Для начала, как тестовый вариант, был реализован прототип визуализации на выборочных данных. Были использованы все предметные коллекции, но с ограниченным числом публикаций (Рис.1).

(2). Идея в визуализации по тематикам, хранящихся в теге 650 заключается в отображении публикаций по тематикам arXiv.org, позволяющая пользователю наглядно оценить количество публикаций в той или иной области.

3.2 Визуальное представление информационных ресурсов JDS с помощью метода радиального графа

Радиальное рисование графа отличается от обычного тем, что его уровни имеют вид концентрических окружностей. В радиальном графе только один узел располагается по центру, а остальные узлы располагаются вокруг него. Центральный узел — фокус, а остальные узлы располагаются на концентрических кольцах вокруг него. Узел располагается на кольце, которому соответствует наименьшее расстояние до фокуса. Любые два узла соединенные ребрами считаются соседями. Соседние узлы располагаются от центрального на самом ближайшем внутреннем кольце, их соседи на втором кольце и т.д. [29].

С помощью метода радиального графа отображаются авторы и соавторы публикаций. Автор отображается в центре, соавторы на втором концентрическом круге, а соавторы соавторов на третьем. В визуализации используется только три уровня иерархии, чтобы избежать перенасыщения.

Выбранный объект (автор) размещается на середине экрана, и его соавторы выделяются соответствующим цветом. При наведении курсора на автора/соавтора появляется информация о количестве его публикаций (Рис. 2).

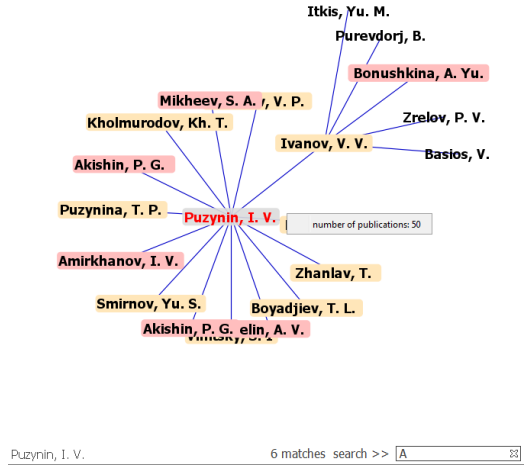

Рис.2. Визуальное представление информационных ресурсов JDS с помощью метода радиального графа.

Выводы

В ходе работы было проведено исследование предметной области. Описан процесс создания визуализации информации, выделены особенности применения визуализации информации в электронных библиотеках. Особое внимание было уделено методам и стратегиям создания визуализаций. Рассмотрены такие стратегии как масштабирование и панорамирование, фокус+контекст, обзор и детализация. Приводятся примеры различных структур информации. Спроектированы прототипы визуальных представлений для предоставления дополнительных возможностей поиска и исследования информации пользователям JDS. В дальнейшем планируется развитие и внедрение визуальных представлений в репозиторий.

Приложение: Пакет Prefuse

Prefuse — расширяемая библиотека, позволяющая разработчикам создавать интерактивные информационные приложения визуализации, используя язык программирования Java. Prefuse может использоваться для создания, как автономных приложений, так и для визуальных компонент, внедренных в большие приложения. Особенными компонентами Prefuse являются [7]:

- Таблицы, графы и древесные структуры данных, поддерживающие произвольные атрибуты данных, индексацию данных;
- компоненты для схемы размещения: цвета, размеры, методы искажения, анимации и т.д.;
- преобразование представлений, включающих панорамный вид, изменение масштаба изображения;
- динамические запросы для интерактивной фильтрации данных;
- интегрированный текстовый поиск.

Архитектура Prefuse основана на абстракциях, и предоставляет библиотеки для создания визуализаций информации.

Prefuse основан на информационной эталонной модели визуализации: от сбора данных к представлению интерактивных дисплеев [1, 7, 10]. Процесс визуализации или конвейер, соответствующий этой модели проиллюстрирован на Рис. 3.

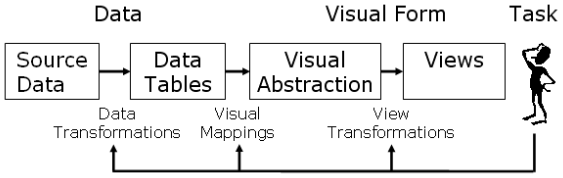

Рис. 3. Диаграмма, изображающая эталонную модель визуализации информации в Prefuse

Как следует из этой схемы, конвейер визуализации в Prefuse включает несколько этапов :

- Вначале необходимо иметь исходные данные, для визуализации. Это может быть колонка чисел, структура каталога файлов, или любой другой набор данных.
- Исходные данные используются для того, чтобы создать таблицы данных (внутреннее представление данных), поскольку именно они должны быть визуализированы. Процесс движения от исходных данных к таблицам может включать не только считывание из файла или базы данных, но и любое число преобразований этих данных.
- Таблицы данных являются объектами визуального отображения при создании визуальной абстракции. Визуальная абстракция — модель данных, которая включает визуальные особенности, такие как пространственная схема размещения, цвет, размер и форма. Визуальная абстракция содержит всю информацию, которая может быть визуализирована.
- Пользовательское взаимодействие в процессе визуализации (обычно через мышь и ввод с клавиатуры) — это обратная связь, с помощью которой можно изменять или обновлять процесс в любой стадии «конвейера» визуализации. Prefuse позволяет перемещать объекты и изменять масштаб изображения в представлении.

Литература

[1] Adam Smith. Prefuse. University of Pittsburgh: Interdisciplinary Visualization Research Lab, Teaching: CS2620 Interdisciplinary Modeling and Visualization, 2010. http://vis.cs.pitt.edu/teaching/cs2620/lectures/L06_PrefuseTutorial.pdf

[2] Card, S., Mackinlay, J., Shneiderman, B. Readings in Information Visualization: Using Vision to

Think. San Francisco, CA: Morgan Kaufmann, 1999.

[3] Chaomei Chen. Information Visualization Beyond the Horizon Second Edition. London: Springer-Verlag London Limited, 2006.

[4] Chris North. Information visualization. Blacksburg, VA: Virginia Polytechnic Institute and State University, 2005.

[5] Edvard R. Tufte. Envisioning information. Graphics Press, 2005.

[6] IEEE Information Visualization Conference 2011. http://www.visweek.org/visweek/2011/info/infovis-welcome/infovis-welcome

[7] Jeffrey Heer, Stuart K. Card, James A. Landay. Prefuse: a toolkit for interactive information visualization. Portland Oregon: SIGCHI Conference on Human Factors in Computing Systems, 2005.

[8] Judith Gelernter.Visual Classification with Information Visualization (lnfoviz) for Digital Library Collections. New Jersey: Knowledge Organization 34(3), 2007.

[9] Panayiotis Zaphiris, Kulvinder Gill, Terry Hoi-Yan Ma, Stephanie Wilson, Helen Petrie.Exploring the use of Information Visualization for Digital Libraries. London, United Kingdom: Center for Human Computer Interaction Design, 2005.

[10] Prefuse visualization toolkit. http://www.prefuse.org/

[11] З. В. Апанович. ОТ РИСОВАНИЯ ГРАФОВ К ВИЗУАЛИЗАЦИИ ИНФОРМАЦИИ. Новосибирск, 2007.

[12] И.А. Филозова, Т.Н. Заикина, «ВИЗУАЛИЗАЦИЯ В НАУЧНОЙ ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ» Электронный журнал «Системный анализ в науке и образовании», №4 2011.

Using Information Visualization at JINR Document Server

Tatyana Zaikina, Irina Filozova, Genis Musulmanbekov

Development of visual interfaces applied for visualization of search and representation of documents in the JDS repository of scientific publications of Joint Institute for Nuclear Research, is discussed. Two variants of visualization prototypes are described.