Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен

© Варфоломеев А.Г., Кравцов И.В., Москин Н.Д.

Петрозаводский государственный университет
avarf@mainpgu.karelia.ru

Аннотация

Данная статья описывает проект специализированного Интернет-ресурса, представляющего уникальную коллекцию текстов фольклорных песен Заонежья конца XIX века, а также средства формализации этих текстов с помощью представления их в виде графов. Особенностью ресурса является использование технологии OLAP для быстрого выполнения аналитических запросов к базе данных.

1 Введение

В рамках предыдущих конференций RCDL нами были предложены модель для отображения семантики фольклорных песен и прототип информационной системы, реализованный в Delphi [1, 2]. Информационная система создана на основе уникальной, малоизученной коллекции текстов бесёдных песен Заонежья конца XIX века, собранной Р.Б. Калашниковой [3]. Главная идея, лежащая в основе нашей системы – это представление песни в виде формальной модели (графа), открывающее дорогу для применения математических методов классификации и анализа структуры песен. Спустя год, прошедший после конференции RCDL в Дубне, система претерпела большие изменения. Ведется работа над созданием более гибкой реализации системы, которая позволит применять различные способы формализации одной и той же песни. Кроме того, в систему был добавлен морфологический словарь, который значительно упростил процесс представления песни в виде графа, а также реализован ряд методов кластерного анализа и агрегации графов. Все это, безусловно, заслуживает подробного описания, однако в данной статье хотелось бы сделать акцент на другой стороне нашей работы: представлении коллекции и средств ее исследования научному сообществу в сети Интернет.

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

2 Формализация песен с помощью графов

В фольклористике существуют разнообразные подходы в исследовании песен. Наиболее распространенный из них нашел свое отражение в работах таких видных ученых, как Г.И. Мальцев и А.Т. Хроленко [5, 9], где семантика песни раскрывается при помощи понятия “традиционная формула”. Формульная семантика является ключом к решению таких важных задач фольклористики, как классификация песен и поиск инвариантов. Тематические формулы могут выступать в роли объектов графа, но они же, с другой стороны, определяют отношения между подчиненными объектами – меньшими формулами или словами песни. В итоге песня может быть представлена в виде иерархии графов.

Проиллюстрируем представление песни в виде графа на примере хороводной песни начала XIX века:

Широкая борода!
Не ходи мимо сада.
Не ходи, не гуляй,
Мил дорожки не тори,
Худой славы не спусти.
Худа славушка пройдет,
Никто замуж не возьмет:
Ни приказный, ни купец,
Ни удалый молодец.
Отцу, матери бесчестье,
Роду-племени укор,
С плеч головушка долой!

Мне нельзя идти домой;
Скажу так, скажу сяк,
Скажу изнова опять.
Я во садику была, во зеленом гуляла;
Сладки яблочки щипала, наливчатыя.
Я наливчатыя, самы рассыпчатыя,
Я на блюдечко клала, на серебряный поднос;
В высок терем подошла
И милому поднесла.

Милый яблоков не принял;
Ничего не говорит:
Не отказывает, не приказывает.
Только и знает мой миленькой,
Что сердит на меня.
Рассержусь же я, младёшенька,
Я сама ль на него,
Ещё ль покрепче ль того.

В этой песне можно выделить три центральных объекта: девушка, парень и «широкая борода», остальные объекты вспомогательные. Сюжет

делится на 3 части - в тексте они отделены пунктиром. Таким образом, песню можно представить в виде графа с тремя вершинами и тремя связями между ними (рис. 1). Остальные рисунки расшифровывают эти связи, представляя их также в виде графов:

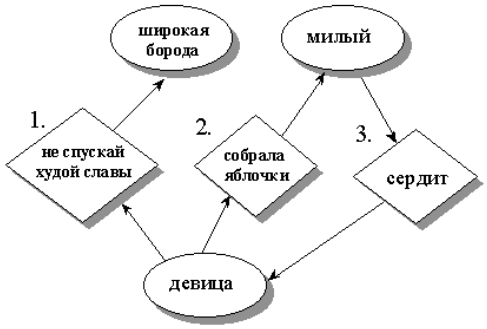

рис.1

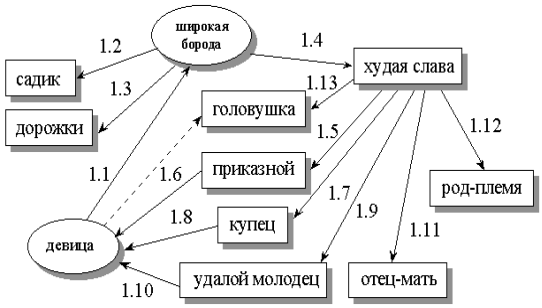

рис.2

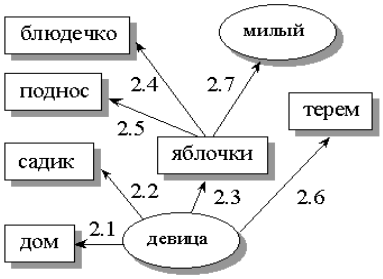

рис.3

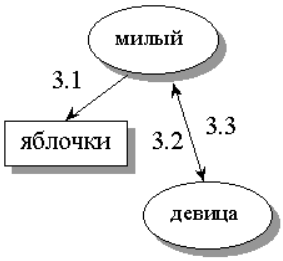

рис.4

Такая методика далеко не исчерпывает потенциальные возможности графов при анализе фольклорных песен. Существует и другие методы формализации текстового материала. Один из них был предложен И. П. Севбо [7]. В его основе лежат "постулаты русской традиционной грамматики и грамматики зависимостей". Каждое слово, независимо от его веса в предложении, является объектом графа, а отношения – это синтаксические связи между словами. Такой подход является более формализованным. Он используется в определении стиля того или иного автора.

Возможны и другие интерпретации объектов и отношений в графе. Важно, что мы понимаем под элементарной единицей текста. Либо это самостоятельная синтаксическая или семантическая единица, либо это часть более крупной единицы, представленная на некотором смысловом уровне. Так, в «Лекциях по структурной поэтике» Ю. М. Лотман [4] выделяет такие элементарные единицы поэтического текста, как фонема, слово, морфологический элемент (грамматическая категория с лексической семантикой), строфа, композиционная часть стиха, стихотворение в целом. Эти единицы могут находиться в различных отношениях друг с другом.

3 Разработка Интернет-ресурса

3.1 Цели проекта

Различные способы представления песен в виде графа будут, очевидно, порождать и различные классификации песен. Тем интересней было бы их сравнение, выделение наиболее адекватных способов для тех или иных ситуаций. И лучший способ для этого - представить нашу информационную систему в виде Интернет-ресурса, доступного ученым всего мира.

Основными целями создания Интернет-ресурса на основе информационной системы по фольклорным песням являются:

- публикация информации о проекте и предоставление коллекции песен;
- демонстрация применения математических методов для классификации и анализа песен;
- обеспечение удаленного доступа к информационной системе для потенциальных пользователей, предоставление им возможности работы в системе со своими материалами.
- разработка и апробация методики создания подобного наукоемкого ресурса, изучение его функциональности и полезности для научного сообщества.

С одной стороны, для Интернет-ресурса требуется выбрать наиболее понятные пользователю методы классификации, с другой – показать наиболее эффективные, но, возможно, объемные и сложные методы. Поэтому в первую очередь требуется создать наглядные представления исходной, промежуточной и результирующей информации, предоставить методику работы с приложением, предложить удобную навигацию, а также интуитивно понятные интерфейсы ввода и просмотра данных.

Для построения рассматриваемого ресурса будет использована клиент-серверная архитектура, таким образом, основные вычисления и обработка данных будут происходить на сервере. Данная концепция позволит довольно эффективно работать в сети и не потребует от пользователя дополнительной установки программного обеспечения, кроме обычного Web-браузера.

3.2 Использование технологии OLAP

Для организации быстрой и функционально удобной работы мы предлагаем использовать технологию OLAP (On-Line Analytical Processing) [6, 8]. В основном OLAP-средства используются в коммерческих приложениях, где необходим анализ больших объемов данных, собираемых операционными базами данных. Накопленные данные организуются в специальные хранилища (Data Warehouse). OLAP-средства рассчитаны на быструю обработку нерегламентированных запросов аналитиков к хранилищам данных. Данные в хранилищах заранее агрегированы и выстроены в многочисленные иерархии в зависимости от решаемых задач. Их можно представить, в отличие от реляционных баз данных, в виде многомерных кубов (гиперкубов), что позволяет говорить о многомерных базах данных.

Согласно Д. Кодду многомерное представление данных состоит из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению [10].

Используя методологию ROLAP, согласно которой многомерные данные хранятся в реляционных таблицах, можно моделировать структуру данных гиперкубов с помощью двух типов таблиц: таблиц фактов и таблиц измерений. Измерения – это параметры, шкалы, по которым будут распределяться данные, а факты – это записи, фиксирующие полученные распределения.

рис.5

Пусть, например, нас интересует, каков типологический состав объектов, выделенных нами в песнях. Для этого мы создаем все необходимые измерения и объединяем их в отдельное пространство, которое будем называть пространством классификации (рис. 5).

Точка в этом пространстве показывает наличие объекта-слова определенного класса в той или иной песне. Такая точка соответствует записи в таблице фактов, причем вместо символьных значений, которые показаны на рисунке, в ней хранятся числовые поля. Числовые поля (своеобразные координаты в пространстве классификации) соответствуют ключевым полям таблиц измерений, формирующих гиперкуб. Структура реляционных таблиц, описывающих смоделированный куб, представлена на рисунке 6.

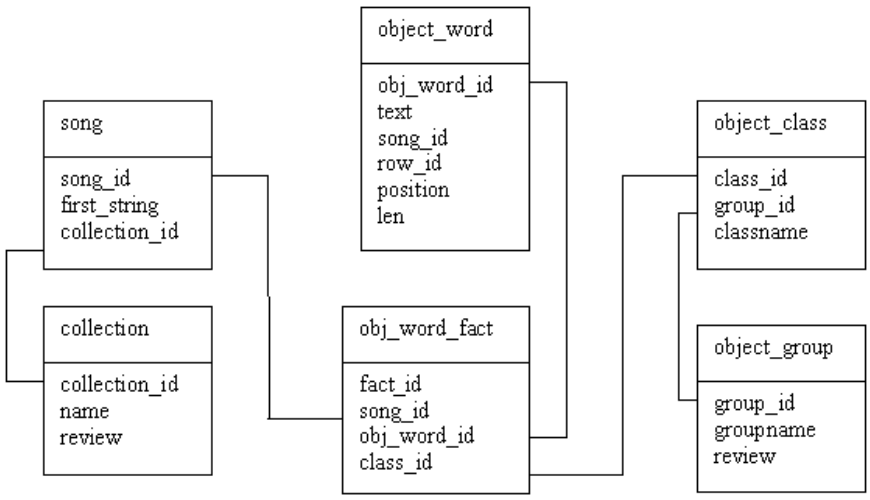

рис. 6

Такую схему таблиц с заданной иерархией измерений еще называют «звездой» или «снежинкой». В данном случае:

obj_word_fact – таблица фактов;
song, *object_word*, *object_class* – таблицы измерений;
collection, *object_group* – таблицы иерархии для измерений;

В результате структура данных становится денормализованной, потому что отдельные поля (например, *song_id*) дублируется в нескольких таблицах. В то же время сокращается время получения запросов к часто используемой информации, такой как, например, разбиение объектов по классам для нескольких песен. Для получения необходимой информации требуется анализ только таблицы фактов

Таблица иерархии для классов объектов (*object_group*) описывает их объединение в группы, соответствующие различным разбиениям (см. рис. 7). Например, классификация по типам (люди, животные, еда, постройки и т.д.) или вхождение более мелких объектов в более крупные. Во втором случае измерениями являются шкалы самих объектов: слова, словоформы, предложения, целые песни, а точка в пространстве описывает факт вхождения. Если необходимо выяснить только вхождение слов в предложения, то мы делаем срез гиперкуба (его подпространство) по интересующим измерениям.

Удобство использования подобных структур состоит в том, что при добавлении новых сущностей или параметров классификации достаточно лишь определить новые или использовать некоторые старые измерения и создать пространство для новой классификации, не изменяя всей остальной структуры данных. Также, если необходимо добавить какие-то оценочные, весовые и любые другие данные к построенным разбиениям, мы добавляем соответствующие поля в таблицы фактов (приписывание к точкам в пространстве дополнительных параметров). То есть, если нам нужно количество объектов «девица» в песне, то появляется поле *number* в таблице фактов *obj_word_fact.*

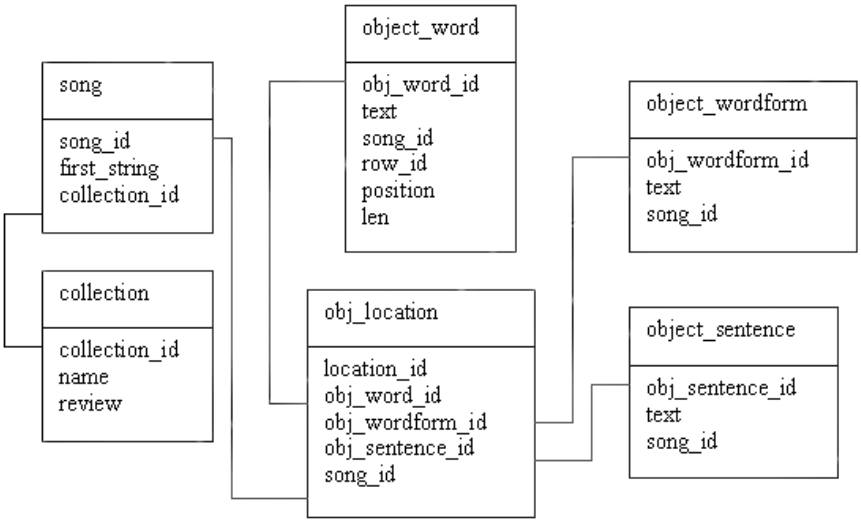

рис. 7

3.3 Организация Интернет-ресурса

В процессе разработки ресурса используются свободно распространяемые программные и языковые средства: язык разметки гипертекста HTML, язык сценариев PHP, СУБД MySQL и web-сервер Apache. Такой набор средств является широко распространенным для Web-приложений различного рода: от форумов до сайтов электронной коммерции.

Логически Интернет-ресурс разделен на три части. Первая будет содержать информацию о проекте и методологию работы с приложением, также будет добавлен форум для обсуждения исследований. Вторая часть будет представлять собой хранилище коллекций песен и инструменты работы с ними. Третья охватит все созданные, а также проектируемые методы и инструменты исследования песенного массива.

Программный код системы в свою очередь разделен на две части. Используя методологию объектно-ориентированного программирования, одна часть представляет собой набор классов, описывающих объекты анализа, хранящиеся в базе

данных, и функции работы с ними. В основном, это классы контейнеры, содержащие список песен, список объектов-слов, набор типологических классов, словарь и т.д. Все функции работы с базой данных инкапсулируются внутри этих классов в качестве их методов.

Вторая часть кода генерирует экранные страницы, используя экземпляры необходимых классов. Например, при поиске объектов-слов в песне с использованием словаря в программном коде будут созданы: экземпляр класса «песня», который загрузит ее из базы данных, экземпляр класса «список объектов-слов», в которые будут заноситься найденные объекты, и экземпляр класса «словарь», с помощью которого будет осуществляться поиск.

Элементы будущего Интернет-ресурса можно посмотреть по адресу http://cs.karelia.ru/~kravcov.

Литература

[1] А. Г. Варфоломеев, Н. Д. Москин. Об электронной коллекции фольклорных песен с теоретико-графовой формализацией текстов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сборник аннотаций стендовых докладов Третьей Всероссийской конференции RCDL'2001. Петрозаводск, 2001. С.20

[2] А. Г. Варфоломеев, Н. Д. Москин, И. В. Кравцов. Информационная система по фольклорным песням Заонежья как инструмент формализации и классификации песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Четвертой Всероссийской научной конференции RCDL'2002. Т.2. Дубна, 2002. С.143-147.

[3] Р. Б. Калашникова. Бесёды и бесёдные песни Заонежья второй половины XIX века. Петрозаводск, 1999

[4] Ю. М. Лотман и тартуско-московская семиотическая школа. Москва, 1994

[5] Г. И. Мальцев. Традиционные формулы русской необрядовой лирики. Ленинград, 1989

[6] А. А. Сахаров. Концепция построения и реализации информационных систем, ориентированных на анализ данных // СУБД, №4, с. 55-70, 1996

[7] И. П. Севбо. Графическое представление синтаксических структур и стилистическая диагностика. Киев, 1981

[8] А. Федоров, Н. Елманова. Введение в OLAP. Ч.1. Основы OLAP // КомпьютерПресс, №4, 2001

[9] А. Т. Хроленко. Поэтическая фразеология русской народной лирической песни. Воронеж, 1981

[10] Л. В. Щавелёв. Оперативная аналитическая обработка данных: концепции и технологии // http://www.olap.ru/basic/olap_and_ida.asp

The project of the specialized Internet-resource for presentation and analysis of folklore songs

A.Varfolomeyev, I.Kravtsov, N.Moskin

This article describes the project of the specialized Internet-resource representing a unique collection of the folklore songs of North Russia and also the flexible means of the text formalization with the help of presentation them as graphs (sets of some objects and relations between them). Such structures can automate the process of solving different significant problems, which have appeared at the time of the investigation of this collection (song classification, the defining of invariants and standard themes etc.).

The potential users of the information system would have the opportunity to work with the own materials, compare the results and choose the most adequate one. Feature of a resource is use of OLAP technology for fast performance of analytical queries.