ИНФОРМАЦИОННАЯ СИСТЕМА ПО ФОЛЬКЛОРНЫМ ПЕСНЯМ ЗАОНЕЖЬЯ КАК ИНСТРУМЕНТ ФОРМАЛИЗАЦИИ И КЛАССИФИКАЦИИ ПЕСЕН

А.Г.Варфоломеев, Н.Д.Москин, И.В.Кравцов
Петрозаводский государственный университет
avarf@mainpgu.karelia.ru

INFORMATION SYSTEM ON RUSSIAN FOLKLORE SONGS OF KARELIA AS A TOOL OF FORMALIZATION AND CLASSIFICATION OF SONGS

Aleksey G. Varfolomeyev, Nikolay D. Moskin, Ignat V. Kravtsov.
Petrozavodsk State University, Petrozavodsk, Russia.

The authors elaborate the project of information system dedicated to folklore songs of North Russia. For increase of research potential of information system it is necessary to decide a task of formalization of the song content. In our opinion the most adequate mathematical structure describing a folklore song is the set of oriented graphs connected among themselves. In each graph the vertexes are the objects of the text, the edges are relations between objects. Our information system will contain both complete texts of songs and their formal representations by graphs that will allow to carry out comparisons of songs, to find invariants, to decide tasks of classification.

Стремительное развитие Интернет-технологий дает ученым принципиально новые возможности в представлении своих результатов научному сообществу. Особенно это касается ученых социально-гуманитарных направлений, работающих с большими комплексами источников. Действительно, если математику зачастую достаточно выложить в Сеть свои публикации, физику – подробно описать эксперимент и сослаться на стандартные методы обработки полученных данных, то историк обычно опирается в своих исследованиях на неопубликованные архивные документы и авторские методики работы с ними. В прежние времена, когда печатная публикация результатов исследования была единственной формой представления научной работы, лишь немногим удавалось издать обширные приложения с первичными материалами [1], в которых приводились либо исходные статистические данные, либо “регесты” - таблицы, содержащие фрагменты архивных источников. В таких случаях подробное описание методики работы с источниками давало возможность проверить выводы автора, воспроизведя его расчеты, а то и провести собственное исследование по иной методике. Привычное для естественных наук требование вос-

производимости эксперимента оказывалось применимым и в гуманитарных областях знаний.

Web-технологии сети Интернет позволяют пойти дальше, объединив вместе текст с описанием методики исследования и выводами автора, базу данных с первичными материалами и инструменты для обработки данных по той или иной методике. Органичное сочетание в Web-страницах текста, графики и программ (скриптов), которые могут обращаться к базам данных и динамически порождать новые Web-страницы, делают Web-публикации потенциально более содержательными и полезными для научного сообщества, чем их традиционные бумажные аналоги. Центральным элементом такой публикации становится вовсе не текст, а информационная система, вводящая в научный оборот новые данные.

На кафедре информатики и математического обеспечения Петрозаводского государственного университета ведется работа над проектом по созданию информационной системы “Бесёдные песни Заонежья XIX – начала XX века” [2]. Информационная система создается на основе личного архива Р.Б.Калашниковой, насчитывающего около 500 текстов песен. Целью проекта является не только полнотекстовая база данных песен, но и создание средств сравнения и классификации песен по формальным признакам.

Текст песни, как и любой объект исследования, может быть охарактеризован некоторым числом количественных и качественных признаков – жанром, годом записи, наличием обрядовых символов, числом мотивов и т.д. Сам текст, кроме того, можно охарактеризовать его словарем, частотой употребления слов и синтаксических конструкций [3], однако эти традиционные для стилеметрии признаки не будут значимыми ввиду маленького объема текста. Намного более важную роль в формализации песни должна играть её структура как связного текста, которая может быть передана таким математическим объектом, как граф. Узлы графа соответствуют объектам и могут быть разных типов, а дуги соответствуют связям между объектами и могут также быть разных типов (тогда на одном множестве узлов возникает несколько графов). И дуги, и узлы могут также снабжаться числами (весами), задающими, например, силу связи.

Если рассматривать текст с точки зрения синтаксиса, то в роли объектов будут выступать все слова (члены предложения), соединенные между собой сочинительными и подчинительными связями. Такие графы известны давно [4], они используются для автоматического перевода или сравнения стилей текстов. Однако текст бесёдной песни, на наш взгляд, прежде всего описывает некую сцену и события, происходящие на ней, то есть, говоря на языке информатики, предметную область. В роли объектов предметной области выступают “сущности” (существительные и местоимения текста), а связи задаются, в частности, глаголами.

Связи между объектами имеют разный характер. На наш взгляд, их можно разбить на две группы: локальные и глобальные. Локальная связь встречается в тексте песни один раз, выражая какое-то конкретное действие, и, как правило, подкреплена глаголом или отглагольной формой. Это может быть *простая связь* (например, девушка ждет парня, парень целует девушку) или *сравнение* (например, парень – сокол)

Связи из второй группы, названные нами глобальными, находятся над повествованием, они не отражены в тексте песни, однако незримо присутствуют в нем. Так, можно выделить связь типа *равенство*, когда два объекта относятся к одному и тому же действующему лицу песни, формально могут быть объединены в один объект, но их разделение выявляет более тонкую семантику художественного произведения. Например, парень – добрый молодец. Другим видом глобальной связи является *принадлежность*, возникающая в том случае, если один из объектов являются частью другого. Например, девушка – коса, дерево – ветка.

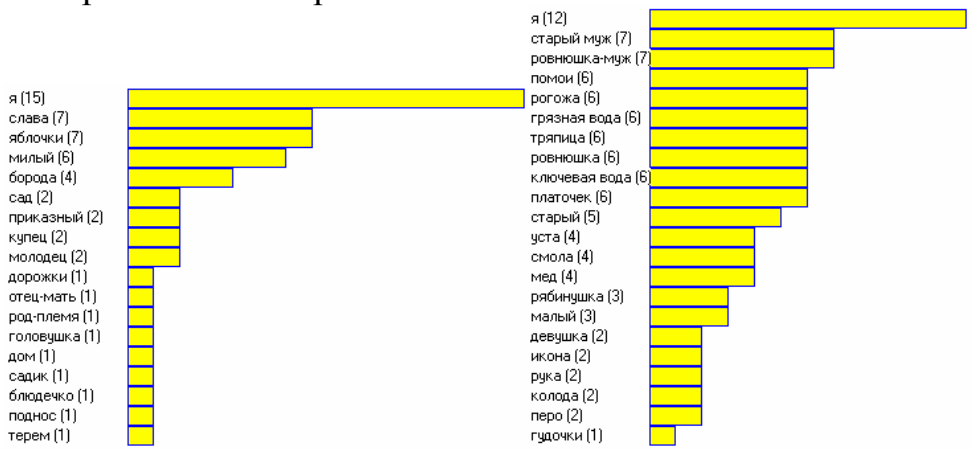
Таким образом, тексту песни можно сопоставить два графа с одинаковым множеством узлов, один из которых задает “сцену” песни (глобальные связи между объектами), а другой – действие, происходящее на сцене (локальные связи).

Анализ полученных графов предполагает как исследование отдельно взятой структуры, так и сравнение графов различных песен между собой. Существует два пути для сравнения графов – определение расстояния между графами и параметризация, то есть представление графа точкой в n-мерном пространстве параметров, а затем уже выяснение, насколько эти точки близки между собой.

Расстояние на множестве графов можно задать разными способами, лишь бы только оно обладало общими свойствами расстояний – было неотрицательным, равнялось бы нулю только в том случае, когда оба графа полностью совпадают и т.д.. Например, можно найти для двух графов максимальный по количеству узлов общий подграф (то есть часть графа), а затем сравнить число узлов в этом подграфе с числом узлов в обоих графах [5]. Чем разница между числами будет больше, тем больше графы отличаются друг от друга. Другой принцип определения расстояния основан на понятии “операции редактирования графа”. Такой операцией может быть удаление, вставка или изменение типа узла или дуги. Подсчет минимального числа таких операций, необходимого для превращения одного графа в другой, также может дать нам представление, насколько два графа похожи друг на друга [6].

Сравнение графов с помощью параметризации можно провести, например, на основе рангового распределения объектов по числу их связей. Гистограмму такого распределения разумно аппроксимировать гиперболой с двумя или тремя параметрами. Набор числовых значений этих параметров будет представлять граф, а вместе с ним и песню, точкой на плоскости

или в пространстве. Ниже на рисунке приведены гистограммы двух песен, резко отличающиеся друг от друга и по внешнему виду, и по набору параметров гиперболических кривых.



Другим способом параметризации графа может служить сопоставление его вершин заранее зафиксированным типам объектов (например, люди, животные, растения, части тела, и т.д.) и измерение частот распределения объектов песен по типам. В таком случае каждой песне ставится в соответствие вектор частот, и затем к этим векторам может быть применен один из методов кластерного анализа.

Прототип информационной системы уже реализован с помощью среды визуального проектирования Delphi. Система позволяет в диалоговом режиме определять объекты и связи песен, просматривать графы, а также сравнивать их между собой с помощью гиперболической аппроксимации ранговых распределений и с помощью кластерного анализа частотных распределений объектов по типам. В процессе формализации, то есть представления песни в виде графа, пользователи могут обращаться к небольшой экспертной системе, которая помогает им, задавая “наводящие” вопросы, относить объекты и связи к тем или иным типам.

В настоящий момент ведется работа по реализации системы в виде Web-приложения. Информационная система, размещенная в Интернете, будет служить инструментом формального сравнения фольклорных песен между собой. Пользователи, зарегистрировавшись в системе, получат возможность добавлять к базе данных новые песни и исследовать их с помощью большого числа стандартных методов классификации.

Работа такой системы должна быть основана, на наш взгляд, на концепции OLAP (On-Line Analytical Processing) – оперативной аналитической обработке данных по нерегламентированным запросам пользователей. [7] Основой OLAP является представление данных в виде многомерных кубов, обобщающих кросс-таблицы СУБД Paradox или перекрестные запросы MS Access. Над этими кубами выполняются некоторые операции (выделение среза, переход к более или менее детальным измерениям, и т.д.), в результате которых аналитик может быстро получать ответы на во-

просы, возникающие в ходе исследования. Концепция OLAP получила широкое распространение в проектировании систем поддержки принятия решений в сфере бизнеса, однако может быть очень полезной и при создании информационно-аналитических систем для научных исследований.

Литература

1. Литвак Б.Г. Опыт статистического изучения крестьянского движения в России в XIX веке. М., 1967; Миронов Б.Н. Хлебные цены в России за два столетия. Л., 1985; Витов М.В., Власова И.В. География сельского расселения Западного Поморья в XVI – XVIII вв. М., 1974.
2. “Бесёдами” назывались молодежные вечеринки у русского крестьянского населения Олонецкой губернии. Песни на бесёдах носили ярко выраженный игровой характер и сопровождались танцами. См.: Калашникова Р.Б. Бесёды и бесёдные песни Заонежья второй половины XIX века. Петрозаводск, 1999.
3. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. Таллин, 1987; Мартыненко Г.Я. Основы стилеметрии. Л., 1988.
4. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. Киев, 1981.
5. Horst Bunke, Kim Shearer. A graph distance metric based on the maximal common subgraph // Pattern Recognition Letters. Vol. 19. 1998.
6. Kaizhong Zhang, Jason Wang, Dennis Shasha. On the editing distance between undirected acyclic graphs and related problems // Proc. Combinatorial Pattern Matching. 1995.
7. Сахаров А.А. Концепция построения и реализации информационных систем, ориентированных на анализ данных // СУБД. 1996. № 4. С. 55-70; Федоров А., Елманова Н. Введение в OLAP. Ч.1. Основы OLAP // КомпьютерПресс. №4. 2001. См. также материалы сайта www.olap.ru.