Применение системы DMiner для решения задач в области топонимики

©ИПМИ КарНЦ РАН

Топонимической науке в силу ее специфики приходится оперировать огромным количеством исходного материала. На Европейской части Севера России топонимистами КарНЦ РАН накоплен значительный объем топонимических данных, представляющий собой значительную информационную, культурологическую, историческую и лингвистическую ценность [1].

Между тем, естественные возможности человека в осмыслении этого материала ограничены. Назрела настоятельная необходимость использования вычислительной техники в топонимических исследованиях. Создание топонимической базы данных TORIS стало важным шагом в применении вычислительной техники для обработки результатов топонимических исследований [1].

TORIS (http://toris.krc.karelia.ru) – это тематический Web-сервер по топонимии Европейского Севера России, который создавался с целью содействия развитию российской топонимической науки, координации теоретических и прикладных исследований и разработок российских топонимистов в различных регионах страны, оперативного обмена информацией, организации телеконференций и электронных сборников научных публикаций. Для его создания использованы Russian Apache для Unix, некоммерческие версии СУБД miniSQL и PostgreSQL, а также технология CGI-сценариев [2]. Работа по созданию и развитию Web-сайта TORIS поддержана грантами РГНФ (N 00-04-12020В, N 03-04-12033В).

В настоящий момент TORIS-сервер КарНЦ РАН включает в себя следующие компоненты [3]:

- реляционную базу данных TORIS по топонимии Европейского Севера России;
- Web-интерфейс, как компонент, обеспечивающий механизм доступа к содержимому топонимической базы данных;
- библиографическую базу данных, которая содержит информацию о существующих публикациях по топонимии;
- набор электронных публикаций по теме исследования, новости, контактную информацию и другие информационные материалы;

- программное обеспечение для поддержки постоянно действующей тематической телеконференции.

Топонимическая база данных TORIS с Web-интерфейсом является главным компонентом сервера. Она является реляционной базой данных, созданной на базе СУБД PostgreSQL версии 6.5 для Solaris. Она включает 13 связанных таблиц, в которых размещаются записи, в совокупности описывающие каждый топоним по 25 характеристикам. На данный момент в базе данных содержится информация по 1194 топонимам. Это составляет малую часть огромного набора материалов по топонимии Европейского Севера России, накопленного в результате многолетних исследований топонимистов КарНЦ РАН. Постоянно происходит пополнение базы данных TORIS по мере занесения филологами-топонимистами новых наборов топонимов из собранных архивов.

Средства СУБД в сочетании с Web-интерфейсом облегчают исследователям работу с собранным материалом, предоставляя возможность осуществлять стандартные операции поиска, ввода и корректировки данных. Кроме того, наличие базы данных с фиксированной структурой обеспечивает однозначность схемы формализации исходного материала, что является необходимым условием для возможности проведения дальнейшей более сложной автоматизированной обработки данных путем применения методов Data Mining.

Впервые идеи применения методов Data Mining для анализа топонимической базы данных TORIS были представлены в работах [3, 4]. Г.М. Кертом в [5] поставлен ряд задач в топонимических исследованиях, которые могут быть решены с помощью ЭВМ. В основе многих из них лежит определение частотных характеристик топонимов и их компонент, выявление характерных повторяемых элементов. Подобные задачи могут быть решены с помощью методов Data Mining поиска значимых множеств и ассоциативных правил.

1 Постановка задачи

Для получения информативных результатов анализ должен проводиться в наборах записей топонимической базы данных, относящихся к территориям, для которых достаточно полно и равномерно собрана информация о топонимах. В данный момент в базе данных TORIS этому требованию отвечает собранный С.А. Агарковой материал на территории Кемского района Карелии, составляющий 397 записей по русским топонимам. Поэтому задача анализа формулируется для данного набора топонимов.

В первую очередь необходимо выбрать набор анализируемых характеристик топонимов, таким образом, чтобы данные характеристики удовлетворяли семантическому ограничению, то есть не были явным образом зависимы друг от друга. Вместе с тем, для них должно быть возможно наличие некоторой скрытой зависимости. Такими характеристиками являются поля, значениями которых являются компоненты топонима, и поле, обозначающее объ-

ект топонима. Тогда для набора из данных значений для каждого топонима задача нахождения зависимостей между ними в виде ассоциативных правил формулируется следующим образом. $T = \{t\}$ – исходный набор, в котором для каждого топонима определено множество свойств $t = \{\tau_i\}$, включающее в качестве элементов τ_i значения компонентов и объект, которому принадлежит данный топоним. $\Theta \equiv \{\omega : \exists t \in T : \exists \tau_i \in t : \tau_i = \omega\}$ – множество всех имеющихся компонентов и объектов, элементами которого описываются выбранные характеристики любого топонима в рассматриваемом наборе. $Rule = \{Antecedent => Consequent \quad | \quad c, s\}$ – правило, где $Antecedent \subset \Theta$ и $Consequent \subset \Theta$,

$$s = s(Rule, T) = \frac{|\{t \in T : Antecedent \subset t \quad \&\& \quad Consequent \subset t\}|}{|\{t \in T\}|}$$

– поддержка правила $Rule$ в T,

$$c = c(Rule, T) = \frac{|\{t \in T : Antecedent \subset t \quad \&\& \quad Consequent \subset t\}|}{|\{t \in T : Antecedent \subset t\}|}$$

– степень уверенности правила $Rule$ в T.

Необходимо найти набор всех правил $Rule$, таких что $s > minsupport$ и $c > minconf$, где $minsupport$ и $minconf$ – задаваемые нижние пороги поддержки и степени уверенности правила.

2 Решение задачи с помощью системы DMiner

Для решения данной задачи необходимо выполнить следующие действия:

1. загрузка в рабочую базу данных информации о компонентах и объектах топонимов Кемского района;
2. поиск значимых множеств;
3. генерация из них набора ассоциативных правил.

Все эти шаги выполняются с использованием базовых модулей системы DMiner.

Для задания параметров загрузки анализируемой информации в рабочую базу данных в данном случае используется текстовый файл. Причиной выбора данного способа загрузки является тот факт, что структура базы данных TORIS не позволяет использовать интерактивную настройку параметров для загрузки данного набора полей (все компоненты хранятся в отдельной таблице, на которую ссылаются значения соответствующих различным компонентам полей основной таблицы).

В приложении А приведен файл, в котором определяются параметры загрузки данных. В качестве исходной базы данных используется локальная копия базы данных TORIS на базе СУБД MySQL. Рабочая база данных создается на той же платформе в той же СУБД. В качестве ключевого поля

выбран идентификатор топонима в основной таблице ling. Значения полей берутся в нижнем регистре, так как в данной задаче при сравнении элементов не должен учитываться регистр их написания. Для каждого топонима определяются способ извлечения из базы данных значений полей, определяющих его объект и компоненты. На данном шаге в рабочую базу данных было загружено 397 записей, состоящих из 364 различных элементов.

В качестве параметра при поиске значимых множеств задается значение нижнего порога поддержки, равное 0.5%. При общем объеме исходного набора в 397 записей это означает, что значимое множество должно содержаться хотя бы в двух записях исходного набора. В качестве типа множества выбирается "сочетание", так как в данной задаче не важен порядок следования компонентов топонима и количество вхождений компонентов в один топоним. Элементы исходного набора, загруженного в рабочую базу данных не соответствуют требованиям, определяющим данный тип множества (например, в топониме Луда Большая Варблуда, компонент "луда" встречается два раза). Поэтому при анализе выбирается опция, определяющая необходимость приведения элементов исходного набора к требуемому виду. В результате выполнения процедуры поиска для данного набора характеристик топонимов получено 99 значимых множеств, состоящих из более чем одного элемента.

На основе полученных значимых множеств, определяющих повторяемые сочетания компонентов топонима и значения объекта топонима, выполняется генерация ассоциативных правил. В качестве нижнего порога степени уверенности правила принимается значение 2%. В результате выполнения процедуры генерации получено 52 ассоциативных правила, которые приведены в приложении В. Наиболее интересными из них являются последние, отражающие приоритеты в выборе компонентов в названиях, даваемым таким объектам местности, как озеро, тоня, остров, губа, болото, мыс. Важно отметить, что данные результаты справедливы только по отношению к использованной в исследовании исходной выборке данных. Однако они иллюстрируют работоспособность методов Data Mining в применении к топонимическим исследованиям.

A Файл параметров анализа

```
#Source database host and JDBC driver
jdbc:mysql://localhost/toris org.gjt.mm.mysql.Driver
#Source database login and password
dminer ******
#Target database host and JDBC driver
jdbc:mysql://localhost/workbase org.gjt.mm.mysql.Driver
#Target database login and password
dminer ******
#Source and Target database encoding
ISO8859-5 Cp1251
#Table for entries, being analized, and code table
```

ALL_EN Decodes
#Key table.field in source database
ling.id
#Lowercase characters (1 - yes, 0 - no)
1
#Amount of fields, being analized
5
#Fields, being analized, and how to recieve them, having certain
#entry from key table
Объект = obj.rus FROM ling, geogr, obj WHERE (geogr.obj = obj.id) AND (geogr.id = ling.object) AND (geogr.dis = 61)
Компонент = comp.name FROM ling, comp, geogr WHERE (comp.id = ling.comp1) AND (geogr.id = ling.object) AND (geogr.dis = 61)
Компонент = comp.name FROM ling, comp, geogr WHERE (comp.id = ling.comp2) AND (geogr.id = ling.object) AND (geogr.dis = 61)
Компонент = comp.name FROM ling, comp, geogr WHERE (comp.id = ling.comp3) AND (geogr.id = ling.object) AND (geogr.dis = 61)
Компонент = comp.name FROM ling, comp, geogr WHERE (comp.id = ling.comp4) AND (geogr.id = ling.object) AND (geogr.dis = 61)

B Полученный набор ассоциативных правил

(Компонент = анг) ==> (Объект = озеро, Компонент = озеро) c = 100.0%, s = 0.76%
(Объект = озеро, Компонент = анг) ==> (Компонент = озеро) c = 100.0%, s = 0.76%
(Компонент = озеро, Компонент = анг) ==> (Объект = озеро) c = 100.0%, s = 0.76%
(Компонент = южное) ==> (Объект = озеро) c = 100.0%, s = 0.76%
(Компонент = ламбина) ==> (Объект = озеро) c = 100.0%, s = 1.51%
(Компонент = большое) ==> (Объект = озеро) c = 100.0%, s = 1.26%
(Компонент = малое) ==> (Объект = озеро) c = 100.0%, s = 0.76%
(Компонент = анг) ==> (Объект = озеро) c = 100.0%, s = 0.76%
(Компонент = амбарное) ==> (Объект = озеро) c = 100.0%, s = 1.01%
(Компонент = анг) ==> (Компонент = озеро) c = 100.0%, s = 0.76%
(Компонент = малый) ==> (Объект = остров) c = 100.0%, s = 1.26%
(Компонент = горелый) ==> (Объект = остров) c = 100.0%, s = 1.01%
(Компонент = домашняя) ==> (Объект = губа) c = 100.0%, s = 0.76%
(Компонент = озеро) ==> (Объект = озеро) c = 93.47%, s = 7.3%
(Компонент = луда) ==> (Объект = остров) c = 91.24%, s = 7.81%
(Компонент = березовец) ==> (Объект = остров) c = 80.16%, s = 1.01%
(Компонент = наволок) ==> (Объект = мыс) c = 77.06%, s = 2.52%
(Компонент = голомянные) ==> (Компонент = луды) c = 75.25%, s = 0.76%
(Компонент = голомянные) ==> (Объект = острова) c = 75.25%, s = 0.76%
(Объект = болото) ==> (Компонент = мох) c = 66.89%, s = 1.01%
(Компонент = мох) ==> (Объект = болото) c = 66.89%, s = 1.01%
(Компонент = большой) ==> (Объект = остров) c = 66.52%, s = 1.51%
(Компонент = корга) ==> (Объект = тоня) c = 60.32%, s = 0.76%

(Компонент = луды) ==> (Объект = острова) c = 59.92%, s = 1.51%
(Компонент = остров) ==> (Объект = остров) c = 57.39%, s = 1.01%
(Компонент = мох) ==> (Компонент = большой) c = 50.33%, s = 0.76%
(Объект = озеро) ==> (Компонент = озеро) c = 36.23%, s = 7.3%
(Объект = мыс) ==> (Компонент = наволок) c = 35.74%, s = 2.52%
(Компонент = большой) ==> (Компонент = мох) c = 33.48%, s = 0.76%
(Компонент = луды) ==> (Объект = остров) c = 30.16%, s = 0.76%
(Компонент = луды) ==> (Компонент = голомянные) c = 30.16%, s = 0.76%
(Объект = острова) ==> (Компонент = луды) c = 29.96%, s = 1.51%
(Объект = остров) ==> (Компонент = луда) c = 26.28%, s = 7.81%
(Объект = острова) ==> (Компонент = голомянные) c = 15.08%, s = 0.76%
(Объект = озеро, Компонент = озеро) ==> (Компонент = анг) c = 10.41%, s = 0.76%
(Компонент = озеро) ==> (Объект = озеро, Компонент = анг) c = 9.73%, s = 0.76%
(Компонент = озеро) ==> (Компонент = анг) c = 9.73%, s = 0.76%
(Объект = губа) ==> (Компонент = домашняя) c = 8.38%, s = 0.76%
(Объект = озеро) ==> (Компонент = ламбина) c = 7.49%, s = 1.51%
(Объект = озеро) ==> (Компонент = большое) c = 6.25%, s = 1.26%
(Объект = тоня) ==> (Компонент = корга) c = 5.39%, s = 0.76%
(Объект = остров) ==> (Компонент = большой) c = 5.08%, s = 1.51%
(Объект = озеро) ==> (Компонент = амбарное) c = 5.01%, s = 1.01%
(Объект = остров) ==> (Компонент = малый) c = 4.24%, s = 1.26%
(Объект = озеро) ==> (Компонент = озеро, Компонент = анг) c = 3.77%, s = 0.76%
(Объект = озеро) ==> (Компонент = южное) c = 3.77%, s = 0.76%
(Объект = озеро) ==> (Компонент = малое) c = 3.77%, s = 0.76%
(Объект = озеро) ==> (Компонент = анг) c = 3.77%, s = 0.76%
(Объект = остров) ==> (Компонент = горелый) c = 3.4%, s = 1.01%
(Объект = остров) ==> (Компонент = остров) c = 3.4%, s = 1.01%
(Объект = остров) ==> (Компонент = березовец) c = 3.4%, s = 1.01%
(Объект = остров) ==> (Компонент = луды) c = 2.56%, s = 0.76%

Список литературы

[1] Г.Керт, В.Вдовицын, А.Веретин, Компьютерный банк топонимии Европейского севера России: TORIS Препр. докл. На заседании Президиума КарНЦ РАН 24 апреля 1998 г.

[2] Вдовицын В.Т., Керт Г.М., Сорокин А.Д., Луговая Н.Б., Беляева Н.А., Чуйко Ю.В., Тематический Web-сайт по топонимии Европейского Севера России // Тезисы докладов Всероссийской научной конференции "Научный сервис в сети Интернет" (18-23 сентября 2000 г., г.Новороссийск). – М.: Изд-во МГУ, 2000, с.167-168.

[3] V. T. Vdovitsyn, G. M. Kert, N. B. Lugovaya, J. V. Chuiko, The Creation and the Development of the Thematic Web-site on the Toponymy of the European North of Russia // Proc. of the FDPW'2000, vol. 3, University of Petrozavodsk, 2001, pp.132-136.

[4] В.Т. Вдовицын, Г.М. Керт, Н.А. Беляева, Н.Б. Луговая, А.Д. Сорокин, Ю.В. Чуйко, Электронная коллекция информационных ресурсов по топонимии Европейского Севера России // Сборник докладов Третьей Всероссийской конференции RCDL'2001 "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Петрозаводск, 11-13 сентября 2001 г. – Карельский научный центр РАН, 2001, с.199-201.

[5] Керт Г.М., Применение компьютерных технологий в исследовании топонимии (прибалтийско-финская, русская). – Петрозаводск: Карельский научный центр РАН, 2002.