В.Т. Вдовицын, Г.М. Керт, Н.Б. Луговая, Ю.В. Чуйко (Петрозаводск, ИПМИ КарНЦ РАН). **Применение алгоритмов поиска логических зависимостей для решения задач в области топонимики.**

Топонимической науке в силу ее специфики приходится оперировать огромным количеством исходного материала. Между тем естественные возможности человека в осмыслении этого материала ограничены. В связи с этим назрела настоятельная необходимость применения методов прикладной математики и информатики в топонимических исследованиях.

С целью содействия развитию российской топонимической науки, координации теоретических и прикладных исследований и разработок российских топонимистов в различных регионах страны нами был разработан тематический Web-сайт TORIS с интегрированной базой данных, содержащей информацию о топонимах северо-западного региона России (http://toris.krc.karelia.ru) [1]. Структура описания русской и прибалтийско-финской топонимии, включающая 18 характеристик, учитывает структурные, семантические и иные свойства топонима, а также экстралингвистические признаки объекта, именуемого данным топонимом. При этом часть характеристик (например, характеристика 'семантика') являются составными (т.е. представляют собой однородные группы параметров) и называются компонентами, которых может быть не более 4-х в описании топонима.

Нами сформулирован ряд задач в области топонимики, в основе решения которых лежит определение частотных характеристик топонимов и их компонентов, а также выявление характерных повторяющихся значений характеристик в описаниях топонимов [2]. Подобные задачи могут быть решены с использованием алгоритмов поиска значимых множеств и ассоциативных правил.

В формальной постановке задача нахождения логических зависимостей между характеристиками топонимов в виде ассоциативных правил может быть сформулирована следующим образом. Обозначим через:

- $T = \{t\}$ – исходный (анализируемый) набор топонимов, в котором для каждого топонима определено множество исследуемых характеристик;
- $\Theta = \{\omega : \exists t \in T : \exists \tau_i \in t : \tau_i = \omega\}$ – множество всех возможных значений исследуемых характеристик, описывающих топонимы из рассматриваемого набора;
- $Rule = \{Antecedent => Consequent \quad | \quad c, s\}$ – множество ассоциативных правил, 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.

Задача заключается в том, чтобы найти все правила $Rule$, для которых $s > minsupport$ и $c > minconf$ (где: $minsupport$ и $minconf$ – задаваемые нижние пороги поддержки и степени уверенности правила).

Для решения задач в области топонимики нами разработана программная система DMiner, в основе которой лежит алгоритм поиска значимых множеств PrefixSpan [3]. Система также включает ряд сервисных функций, связанных с подготовкой исходных данных для их обработки и интерпретации полученных результатов. С помощью разработанной системы был проведен ряд вычислительных экспериментов с наборами топонимов из базы данных сайта TORIS, в результате которых были получены ассоциативные правила, отражающие приоритеты в выборе компонентов в названиях географических объектов.

Работа выполняется при поддержке РГНФ (грант № 03-04-12033в).

Список литературы

[1] В.Т. Вдовицын, Г.М. Керт и др. Электронная коллекция информационных ресурсов по топонимии Европейского Севера России. //Сб. трудов Третьей Всеросс. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", RCDL'2001, Петрозаводск, 11-13 сентября 2001 г., Карельский научный центр РАН, 2001, с.109-201.

[2] Керт Г.М. Применение компьютерных технологий в исследовании топонимии (прибалтийско-финская, русская).– Петрозаводск: Карельский научный центр РАН, 2002.– 187 с.

[3] Jian Pei, Juawei Han and others. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In Proc. 2001 Int. Conf. Data Engineering (ICDF'01), Heidelberg, Germany, April 2001, pp.215-224.