УДК 502.5:004.82

А. А. Платонов, Б. Х. Санжапов

ОРГАНИЗАЦИЯ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В СИСТЕМЕ АВТОМАТИЗИРОВАННОГО ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА ГОРОДСКИХ ТЕРРИТОРИЙ

Рассматривается общий подход к анализу текста на естественном языке с целью извлечения данных о фактах, их числовых значениях и географических координатах. С целью понимания принципов работы такого анализа производится разработка модели организации и хранения данных.

К л ю ч е в ы е с л о в а: обработка естественного языка, геоинформационные системы, извлечение знаний.

The authors develop the general approach to text analysis in natural language in order to receive data about facts, their numerical values and geographical coordinates. The development of data structure and storage model is done to understand the principals of work of this analysis.

K e y w o r d s: natural language processing, geo-information systems, knowledge extraction.

В рамках решения задачи анализа экологической обстановки в Волгоградской области возникает потребность в сборе и систематизации данных о состоянии экологии в регионе в целом и в отдельных его частях. При наличии возможности проведения измерений и сбора данных непосредственно на месте, эта задача легко решаема. Однако вследствие ограниченности ресурсов такого рода сбор данных невозможен [1, 2].

В качестве решения этой проблемы предлагается использование открытых источников данных, таких как сводки погоды, отчеты об экологическом мониторинге и т. п. При этом эти источники должны быть открытыми, т. е. публично доступными. Но здесь возникает другая сложная проблема, от решения которой зависит успешность подхода — анализ текста на естественном языке, поскольку данные в указанных источниках представлены на естественном языке[1] [3].

В целом задача обработки естественного языка достаточно хорошо проработана, хотя результаты работы большинства методов еще далеки от идеала. В данной работе производится попытка разработки обобщенной модели данных, которая могла бы использоваться парсером текста на естественном языке для хранения извлеченных данных.

Задача разработки модели данных является подзадачей разработки парсера — анализатора текста на естественном языке. Целью работы парсера является выявление экологических фактов в предоставленном фрагменте текста, числовых показателей для этого факта (или относительных оценок [4]) и географического положения [5, 6] места проявления выявленного факта. Так

[1] Implementing a Natural Language Processing Framework to Query OpenStreetMap Features in ArcGIS, April 24, 2013. Gary R. Huffman. MGIS Capstone Project The Pennsylvania State University. (Получено из: https://gis.e-education.psu.edu/sites/default/files/capstone/Huffman_596Bpaper_20130420.pdf)

же для анализа собранных данных должна быть собрана информация о моменте времени, в который данный факт был измерен (оценен). Обобщенная схема работы такого парсера представлена на рис. 1.

Рис. 1. Общая схема работы парсера

Для конечной системы анализа важны такие сведения как:
идентификатор факта (в простом случае его название);
числовое значение факта;
географическое положение места измерения факта;
временная метка измерения.

Это минимальный набор сведений, необходимый для проведения пространственного и/или временного анализа.

Уже на данном этапе мы разбиваем все факты на две категории: *Явления* и *Параметры* (рис. 2). Такое разделение позволяет анализировать экологическую обстановку по разного рода явлениям, которые могут быть охарактеризованы различными параметрами. При этом возможен анализ получаемой картины изолированно по параметрам, так как их отношение к какому-либо явлению не обязательно. Также такое разделение позволяет анализировать явления, имеющие одинаковые параметры. Т. е. система в конечном итоге становится многоаспектной. С технической точки зрения хранение значений одного параметра, относящегося к разным явлениям, позволяет избежать чрезмерной избыточности данных.

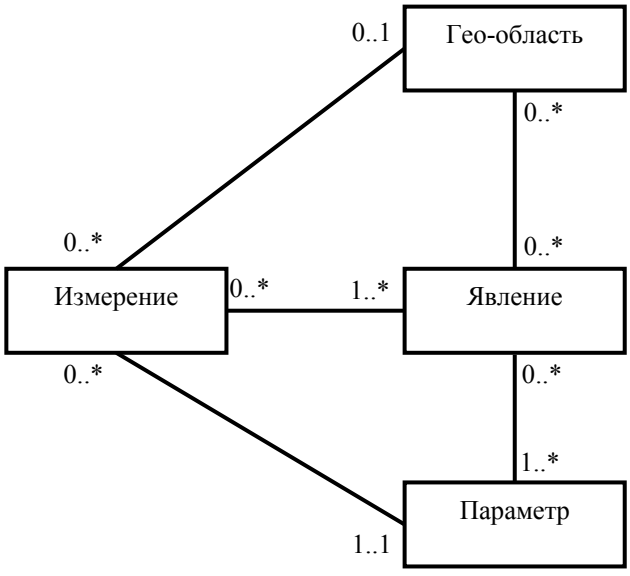

Рис. 2. Модель хранения данных об измерении

Как отмечалось ранее, конечная система анализа должна оперировать данными о географическом положении места измерения и данными о моменте времени измерения. Для учета географического положения в модель хранения данных вводится сущность *Гео-область* (см. рис. 2). Хранение информации о времени измерения, как и данных самого измерения будет осуществляться в атрибутах сущности *Измерение*.

Так как планируется работа с текстом на естественном языке, в модели предусматривается сущность, позволяющая хранить лингвистическую информацию, — *Название* (рис. 3). Эта сущность на данном этапе служит для простой интеграции с системой лингвистической обработки парсером текста, так как на данном этапе не выбран конечный алгоритм обработки естественного языка и используемые для этого средства [5—7].

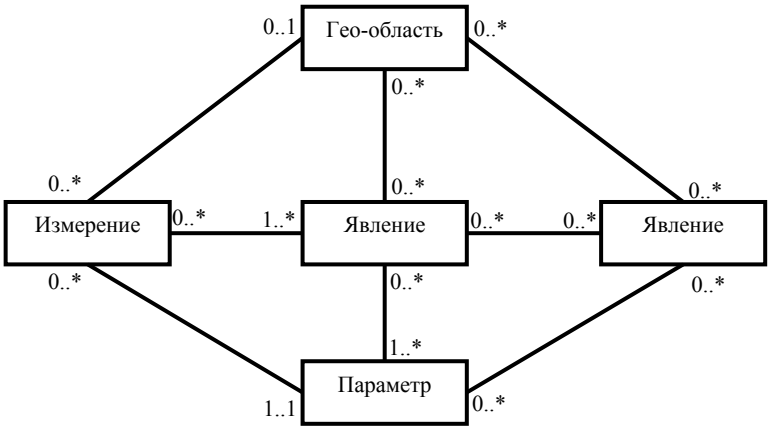

Рис. 3. Модель системы хранения данных парсера

В данной модели вертикальные связи на рисунках используются для отображения логической связи между сущностями. Важны для дальнейшего анализа лишь связи сущности *Измерение* [8]. Все сущности: *Явление*, *Параметр* и *Гео-область* имеют свои имена, которые хранятся в сущности *Название*. Это позволяет при обнаружении парсером в тексте любого анализа провести оценку взаимосвязанности [9] элементов модели по логическим связям и идентифицировать их для фиксации выявленных данных в *Измерении*.

Для сущностей *Явление*, *Параметр* и *Гео-область* в модели подразумеваются иерархические отношения. Например для *Гео-области* «Волгоградская область» вложенной гео-областью является «город Волгоград», который, в свою очередь, также делится на городские районы и т. д. Именно по этой иерархии и будут распространятся собранные данные по измерениям различных фактов. Если нужно оценить уровень загазованности воздуха в городе Волгоград, то достаточно проинтегрировать данные показания по его улицам.

Таким образом, была разработана первоначальная модель для хранения данных об экологической обстановке, которые будет собирать парсер текста на естественном языке. В дальнейшем для системы необходимо провести разработку конечной системы анализа, проанализировать существующие методы анализа текстов на естественном языке и разработать алгоритм работы парсера. Также, с учетом алгоритма работы парсера и системы анализа, должна быть расширена и модель хранения данных [10, 11].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Анализ данных и процессов : учеб. пособие / А. А. Берсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. 3-е изд., перераб. и доп. СПб. : БХВ-Петербург, 2009. 512 с.

2. *Виноград Т.* Программа, понимающая естественный язык. М. : Мир, 1976.

3. *Санжапов Б. Х., Садовникова Н. П.* Поддержка принятия решений при планировании развития городских территорий на основе экосистемного подхода // Вестник ВолгГАСУ. Сер.: Стр-во и архит. 2013. № 31(50). Ч. 2. С. 577—584.

4. *Солнцев Л. А.* Геоинформационные системы как эффективный инструмент поддержки экологических исследований: электронное учебно-методическое пособие. Нижний Новгород : Нижегородский госуниверситет, 2012. 54 с.

5. *Барахнин В. Б., Куперштох А. А.* Алгоритм координатного индексирования электронных научных документов // Вычислительные и информационные технологии в науке, технике и образовании. Казахстан, Павлодар, 20—22 сентября 2006 г. Т. I. С. 228—232.

6. *Санжапов Б. Х., Садовникова Н. П.* Согласование целей при эколого-экономическом обосновании градостроительного проекта с учетом ограничений на значения характеристик, входящих в систему средств, в условиях нечеткой информации // Вестник ВолгГАСУ. Сер.: Стр-во и архит. 2011. Вып. 21(40). С. 151—159.

7. *Lampoltshammer T. J.* Natural Language Processing in Geographic Information Systems — Some Trends and Open Issues // International Journal of Computer Science & Emerging Technologies. 2012. Vol. 3. Issue 3. P. 81—88.

8. *Lutz C., Seylan I., Wolter F.* Mixing Open and Closed World Assumption in Ontology-Based Data Access: Non-uniform Data Complexity // Proc. of the 2012 International Workshop on Description Logics (DL 2012).

9. *Steinberg M., Brehm J.* Utilizing Open Content for Higher-Layered Rich Client Applications // International Journal On Advances in Intelligent Systems. 2009. Vol. 2. № 2—3. P. 303—316.

10. *Younis E. M. G., Jones Ch. B., Tanasescu V., Abdelmoty A. I.* Hybrid Geo-spatial Query Methods on the Semantic Web with a Spatially-Enhanced Index of DBpedia // Geographic Information Science. 7th International Conference, GIScience 2012, Columbus, OH, USA, September 18—21, 2012. Proceedings. P. 340—353.

11. *Bird S., Klein E., Loper E.* Natural language processing with Python. O'Reilly Media, Inc., 2009. 204 p.

1. Analiz dannykh i protsessov : ucheb. posobie / A. A. Bersegyan, M. S. Kupriyanov, I. I. Kholod, M. D. Tess, S. I. Elizarov. 3-e izd., pererab. i dop. SPb. : BKhV-Peterburg, 2009. 512 s.

2. *Vinograd T.* Programma, ponimayushchaya estestvennyi yazyk. M. : Mir, 1976.

3. *Sanzhapov B. Kh., Sadovnikova N. P.* Podderzhka prinyatiya reshenii pri planirovanii razvitiya gorodskikh territorii na osnove ekosistemnogo podkhoda // Vestnik Volgogradskogo gosudarstvennogo arhitekturno-stroitel'nogo universiteta. Ser.: Stroitel'stvo i arhitektura. 2013. № 31(50). Ch. 2. S. 577—584.

4. *Solntsev L. A.* Geoinformatsionnye sistemy kak effektivnyi instrument podderzhki ekologicheskikh issledovanii: elektronnoe uchebno-metodicheskoe posobie. Nizhniy Novgorod : Nizhegorodskii gosuniversitet, 2012. 54 s.

5. *Barakhnin V. B., Kupershtokh A. A.* Algoritm koordinatnogo indeksirovaniya elektronnykh nauchnykh dokumentov // Vychislitel'nye i informatsionnye tekhnologii v nauke, tekhnike i obrazovanii. Kazakhstan, Pavlodar, 20—22 sentyabrya 2006 g. T. I. C. 228—232.

6. *Sanzhapov B. Kh., Sadovnikova N. P.* Soglasovanie tseley pri ekologo-ekonomicheskom obosnovanii gradostroitel'nogo proekta s uchetom ogranichenii na znacheniya kharakteristik, vkhodyashchikh v sistemu sredstv, v usloviyakh nechetkoi informatsii // Vestnik Volgogradskogo gosudarstvennogo arhitekturno-stroitel'nogo universiteta. Ser.: Stroitel'stvo i arhitektura. 2011. Vyp. 21(40). S. 151—159.

7. *Lampoltshammer T. J.* Natural Language Processing in Geographic Information Systems — Some Trends and Open Issues // International Journal of Computer Science & Emerging Technologies. 2012. Vol. 3. Issue 3. P. 81—88.

8. *Lutz C., Seylan I., Wolter F.* Mixing Open and Closed World Assumption in Ontology-Based Data Access: Non-uniform Data Complexity // Proc. of the 2012 International Workshop on Description Logics (DL 2012).

9. *Steinberg M., Brehm J.* Utilizing Open Content for Higher-Layered Rich Client Applications // International Journal On Advances in Intelligent Systems. 2009. Vol. 2. № 2—3. P. 303—316.

10. *Younis E. M. G., Jones Ch. B., Tanasescu V., Abdelmoty A. I.* Hybrid Geo-spatial Query Methods on the Semantic Web with a Spatially-Enhanced Index of DBpedia // Geographic Information Science. 7th International Conference, GIScience 2012, Columbus, OH, USA, September 18—21, 2012. Proceedings. P. 340—353.

11. *Bird S., Klein E., Loper E.* Natural language processing with Python. O'Reilly Media, Inc., 2009. 204 p.

Поступила в редакцию
в мае 2014 г.

Ссылка для цитирования:

Платонов А. А., Санжапов Б. Х. Организация представления знаний в системе автоматизированного экологического мониторинга городских территорий // Интернет-вестник ВолгГАСУ. Сер.: Строительная информатика. 2014. Вып. 11(32). Ст. 6. Режим доступа: http://www.vestnik.vgasu.ru/

For citation:

Platonov A. A., Sanzhapov B. Kh. [Knowledge representation organization in automated environmental monitoring system for urbanized territories]. *Internet-Vestnik VolgGASU*, 2014, no. 11(32), paper 6. (In Russ.). Available at: http://www.vestnik.vgasu.ru/