УДК 004.82

ПОДХОД К АНАЛИЗУ БОЛЬШИХ ДАННЫХ НА ОСНОВЕ ПРИМЕНЕНИЯ ОНТОЛОГИИ

В.С. Кавицкая, д. т. н. В. В. Любченко

Одесский национальный политехнический университет
Украина, г. Одесса
kavickaya@mail.ru

Рассматривается проблема анализа больших данных. Определяются основные задачи, для решения которых могут быть применены методы анализа больших данных. Предлагается использование онтологии для эффективного решения задачи структурирования данных при анализе больших данных, а также задачи формирования новых знаний при анализе больших данных на основе применения ассоциативных правил.

Ключевые слова: онтология, обработка знаний, большие данные, ассоциативные правила.

В настоящее время наблюдается большой интерес к технологиям анализа больших данных, связанный с постоянным ростом данных, которыми приходится оперировать в любой организации. Как правило, когда говорят о термине «большие данные», то используют определение трех V, что означает:

— Volume — большой объем данных;
— Velocity — необходимость анализировать данные с приемлемой скоростью;
— Variety — многообразие и часто недостаточную структурированность данных.

Основная проблема при анализе больших данных заключается в том, что большая часть данных представлена в формате, плохо соответствующем традиционному структурированному формату. При этом данные хранятся во множестве разнообразных хранилищ.

Анализ ситуации в решении проблемы анализа больших данных на основе литературных источников [1, 2] позволяет сделать вывод о том, что, несмотря на возросший интерес и большое количество работ в данном направлении, в области анализа больших данных остается ряд нерешенных проблем. Существующие методики и подходы позволяют работать с данными большого объема, разнообразного состава, часто обновляемых и находящихся в разных источниках, однако остро стоит проблема эффективности анализа больших данных, обусловленная ограничениями на время обработки и используемые ресурсы. В связи с этим, целью работы является повышение эффективности анализа больших данных на основе применения онтологии.

Методы анализа больших данных должны позволять решать следующие задачи:

— структурировать данные;
— извлекать знания из структурированных данных;
— извлекать знания из неструктурированных данных;
— анализировать и получать новые знания из множества различных источников;
— запоминать, оценивать, собирать и исследовать данные.

Основополагающими задачами при анализе больших данных является возможность структурирования данных и формирования знаний самостоятельно, на основе имеющихся данных.

Выделяют следующие стадии структурирования данных [3]:

— определение входных и выходных данных;
— составление словаря терминов;
— выявление объектов, понятий и их атрибутов;
— выявление связей между понятиями;
— выделение метапонятий и детализация понятий;
— построение пирамиды знаний;

— определение отношений;

— определение стратегии принятия решений.

Онтология, в свою очередь, определяется как [4]

$$O = \{X, R, F\},$$

где X — конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология O;

R — конечное множество отношений между концептами (понятиями, терминами), заданной предметной области;

F — конечное множество функций интерпретации (аксиоматизации).

Так как построение онтологии охватывает все стадии структурирования данных, то предлагается ее применение в качестве адекватного инструмента для решения задачи структурирования.

Получение новых знаний, которые в явном виде не формируют эксперты, на основе имеющихся данных является важным аспектом. Для решения задачи получения новых знаний при анализе больших данных предлагается использовать онтологию с применением ассоциативных правил.

Ассоциативные правила позволяют находить закономерности между связанными концептами. В основе алгоритмов поиска ассоциативных правил лежит понятие частого набора, который также можно назвать частым набором концептов, часто встречающимся множеством. Под частотой понимается простое количество транзакций, в которых содержится данный набор концептов. Тогда частыми наборами концептов будут те из них, которые встречаются чаще, чем в заданном числе транзакций [5].

Ассоциативным правилом называется импликация $X \Rightarrow Y$, где $X \subset I$, $Y \subset I$, $X \cap Y \neq 0$; $I = \{i_1, i_2, i_3 \ldots i_n\}$ — множество (набор) концептов; транзакция T — это набор концептов из X, $T \subseteq X$.

Каждая транзакция T представляет собой бинарный вектор, где t_k=1, если концепт i_k присутствует в транзакции, иначе t_k=0. Транзакция T содержит X, некоторый набор концептов из I, если $X \subset T$.

Другими словами, целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор концептов X, то на основании этого можно сделать вывод о том, что другой набор концептов Y также должен появиться в этой транзакции. Установление таких зависимостей дает возможность находить очень простые и интуитивно понятные правила, а также говорить о получении новых знаний, которые не формируются в явном виде.

Таким образом, применение онтологии для анализа больших данных позволяет эффективно извлекать знания из неструктурированных данных, анализировать, оценивать, исследовать знания, а также получать новые знания за счет использования ассоциативных правил.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Géczy P. Big data characteristics // A MultidiscIplinary Journal of Global Macro Trends.— 2014.— Vol. 3.— P. 94—104.

2. Bhat U. Moving towards non-relational databases / U. Bhat, S. Jadhav // International Journal of Computer Applications.— 2010 — Vol. 1 (13). — P. 40—46.

3. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

4. Gavrilova T., Laird D. Practical design of business enterprise ontologies // In Industrial Applications of Semantic Web. — Springer, 2005. — P. 61—81.

5. Hipp J., Guntzer U., Nakaeizadeh G. Algorithms for association rule mining — a general survey and comparison // In Proc. ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining.— 2000.— Vol. 2.— P. 58—64.

V. S. Kavitska, V. V. Liubchenko

Approach to the big data analysis based on the use of ontologies.

The problem of big data analyzing is considered. The main problems that should be solved by the method of analysis of big data are determined. The use of ontologies for effective solution of the problem of structuring data in the analysis of big data is proposed. The use of ontologies for effective solution of the problem of formation of new knowledge in the analysis of big data through the application of association rules is proposed.

Keywords: *ontology, knowledge processing, big data, association rules.*