МЕТОДЫ DATA MINING ДЛЯ АНАЛИЗА БОЛЬШИХ МАССИВОВ ДАННЫХ В ГЕТЕРОГЕННОЙ ГРИД НА БАЗЕ BOINC

Е.Е. Ивашко, А.С. Головин

Введение. Человек выполняет анализ данных для получения новой информации, необходимой при принятии решений в различных ситуациях. С развитием вычислительной техники все большую роль приобретает автоматизированный анализ больших объемов данных. Для этих целей были разработаны специальные алгоритмы и подходы, объединенные термином Data Mining.

В сферу применения Data Mining входят все области, в которых собираются большие объемы данных для получения из них полезных знаний. Одним из направлений Data Mining является поиск ассоциативных правил отражающих нетривиальные взаимосвязи между наборами данных.

Согласно определению [1], Data mining — это процесс обнаружения в сырых данных (1) ранее неизвестных, (2) нетривиальных, (3) практически полезных и (4) доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Рост объемов собираемой информации и развитие средств и методов ее надежного хранения привели к повышению актуальности разработки новых методов и алгоритмов анализа больших и сверхбольших наборов данных. Так, например, в работе [2] автором высказано утверждение, что выявление закономерностей в больших массивах данных становится основным инструментом для исследования и получения новых знаний в передовых областях науки в наше время. Действительно, стремительный рост объемов данных, предназначенных для обработки, характеризует не только IT-компании (например, Google [3]) и научную сферу (см., например, [4]), но и широкий спектр организаций в самых различных областях [5]. В современной науке и технике возникло отдельное направление, связанное с анализом больших и сверхбольших наборов данных, Big Data [6]. Анализ таких объемов данных требует привлечения технологий и средств реализации высокопроизводительных вычислений.

BOINC-грид. Как правило, для обработки больших массивов данных используются суперкомпьютеры или вычислительные кластеры. Для достижения большей производительности вычислительные кластеры объединяются высокоскоростными каналами связи в специализированные грид-системы. Однако с развитием сети Интернет появился и другой подход в построении грид-систем, позволяющий объединить значительное число источников сравнительно небольших вычислительных ресурсов для решения задач обработки больших и сверхбольших объемов данных. В большинстве случаев такие системы построены на использовании свободных вычислительных ресурсов частных лиц и организаций, добровольно присоединяющихся к этим системам (volunteer computing). Однако существуют и примеры построения подобных частных (в масштабах организации или группы организаций) распределенных систем (см., например, [7]). Наиболее эффективно использование таких распределенных систем для проведения серий независимых вычислительных экспериментов (см., например, [8]).

BOINC (Berkeley Open Infrastructure for Network Computing) — это открытая программная платформа для организации грид-систем и систем распределенных вычислений, разработанная в университете Беркли [9]. Это ПО стало основой для большого числа мировых научных проектов [10,11]. Платформа BOINC отличается простотой в установке, настройке и администрировании, а также обладает хорошими возможностями по масштабируемости, простоте подключения вычислительных узлов, использованию дополнительного ПО, интеграции с другими грид-системами и др.

Платформа BOINC имеет архитектуру «клиент-сервер», при этом клиентская часть может работать на компьютерах с различными аппаратными и программными характеристиками. Ключевым объектом системы является проект — автономная сущность, в рамках которой производятся распределенные вычисления. BOINC-сервер поддерживает одновременную работу большого числа независимых проектов; каждый вычислительный узел может одновременно производить вычисления для нескольких BOINC-проектов. Проект однозначно идентифицируется своим URL-адресом. BOINC предоставляет возможность гибкой настройки клиентской части, регулируя максимальный размер загружаемых файлов, время выполнения рабочих заданий, загрузку CPU или GPU, используемый объем оперативной памяти и дискового пространства.

Серверная часть BOINC основана на последовательном выполнении ряда служб, наиболее важные из которых — это служба планирования, выполняющая распределение заданий между вычислительными узлами, и служба освоения, обрабатывающая промежуточные результаты, полученные от вычислителей.

Ассоциативные правила. Одним из наиболее популярных методов Data Mining обнаружения знаний являются различные методы поиска ассоциативных правил. Ассоциативные правила позволяют описывать закономерности между связанными событиями.

Пусть I = {i1 , i2 , . . . , in} — это набор из n различных предметов. D — набор транзакций различной длины над I. Каждая транзакция T из D содержит набор предметов i1, i2, ..., ik из I. Ассоциативным правилом называется импликация X => Y, где $X \subset T$, $Y \subset T$ и $X \cap Y = \varnothing$. X называется условием правила, а Y —

следствием правила. Каждый предметный набор имеет меру статистической значимости, называемую поддержкой. Поддержкой определенного набора элементов называется количество транзакций, содержащих этот набор. Набор элементов является часто встречающимся, если его поддержка (support) больше или равна заданному порогу, который называется минимальной поддержкой (minsupp). Правило X => Y имеет поддержку s, если s% транзакций из D, содержат это правило. Достоверность (confidence) правила показывает какова (статистическая) вероятность того, что из X следует Y. Т.е. правило X => Y справедливо с достоверностью c, если c% транзакций из D, содержащих X, также содержат и Y. Достоверность определяется как отношение support(X ∪ Y)/support(X).

Задача поиска ассоциативных правил заключается в нахождении всех правил, чьи поддержка и достоверность, больше чем некоторые заданные пользователем порог минимальной поддержки и достоверности соответственно.

Впервые задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины.

На Рис. 1 представлена схема реализации алгоритма Partition, предназначенного для поиска ассоциативных правил. Этот алгоритм был адаптирован для выполнения в гетерогенной грид на базе BOINC. Выполнение алгоритма состоит из трех этапов, два из которых выполняются параллельно на вычислительных узлах грид-сети. На завершающем этапе происходит объединение промежуточных результатов.

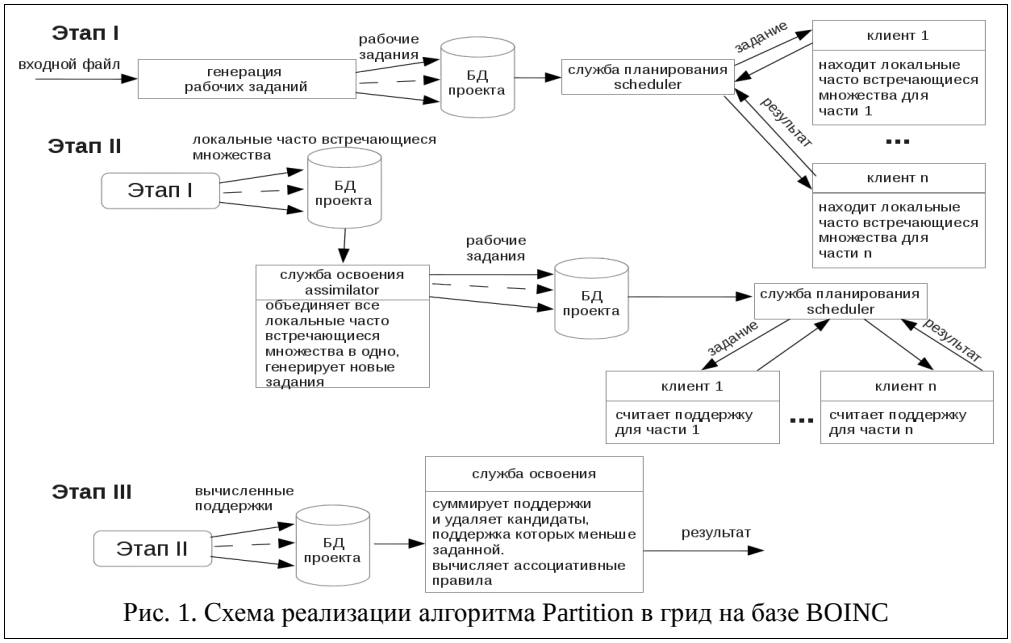

Рис. 1. Схема реализации алгоритма Partition в грид на базе BOINC

Рассмотрим работу алгоритма Partition в BOINC подробнее.

- Программа генерации заданий (разработанная специально для проекта) создает подзадачи и необходимые для их расчета входные файлы. Указанная программа получает на вход исходный файл с транзакционной базой данных и ряд параметров: значение минимальных поддержки и достоверности, а также параметр, отвечающий за разбиение базы данных на части. Последний параметр может ограничить размер каждой части в байтах или задать количество этих частей. По окончании работы программа сохраняет рабочие задания в базе данных проекта.
- BOINC создает для каждого из подзаданий один или несколько одинаковых экземпляров (в зависимости от настроек проекта).
- Планировщик BOINC распределяет подзадания различным клиентским программам.
- Каждый BOINC-клиент загружает с сервера входные файлы, являющиеся частями исходной транзакционной базы данных. Далее клиент запускает приложение, которое на первом этапе вычисляет локальные часто встречающиеся наборы для загруженной части, а на втором — поддержку глобальных кандидатов для своей части транзакционной базы данных.
- После расчета подзадания BOINC-клиент загружает выходные файлы на сервер.
- Клиентская программа отчитывается о выполнении подзадания (возможно, после небольшой задержки, необходимой для снижения нагрузки на программу-планировщик сервера).
- Служба проверки результатов проверяет выходные файлы и определяет наличие канонического результата.
- Когда найдено каноническое решение, служба освоения (разработанная специально для проекта) обрабатывает результаты, например, помещая их в отдельную базу данных или отсылая на электронную почту. В ходе выполнения общей программы поиска ассоциативных

правил служба освоения запускается 2 раза. После выполнения первого этапа служба освоения формирует из полученных локальных частовстречающихся наборов множество всех глобальных кандидатов. Кроме того служба освоения, формирует новую порцию рабочих заданий и сохраняет их в базе данных проекта. После выполнения второго этапа служба освоения запускается повторно. На этот раз данная служба суммирует полученные поддержки для каждого кандидата, удаляет те, чьи поддержки меньше заданного минимального порога и вычисляет ассоциативные правила.

- Когда все экземпляры подзадания завершены, служба удаления файлов удаляет ненужные больше входные и выходные файлы, а также очищает базу данных от информации о каждом подзадании и его экземплярах.

Еще раз обратим внимание на то, что некоторые службы являются стандартными и не зависят от конкретного проекта и его реализации. Однако другие службы необходимо разрабатывать отдельно для каждого проекта. В рамках данной работы кроме приложения для клиента BOINC был разработаны генератор рабочих заданий и служба освоения.

Результаты экспериментов. Для оценки производительности разработанного ПО был проведен ряд экспериментов. Вычисления проводились с использованием грид-сегмента ЦКП КарНЦ РАН «Центр высокопроизводительной обработки данных» [12]. На момент проведения экспериментов в состав грид-сегмента входили 84 вычислительных узла с различными аппаратными и программными характеристиками, а также разными настройками, связанными с организацией вычислений. В частности, два узла обслуживали проекты BOINC в монопольном режиме, а на десяти вычислительных узлах кластера, также входивших в грид, расчеты регулярно приостанавливались для выполнения расчетов пользовательских задач, запускаемых на кластере с помощью системы управления заданиями. Суммарная пиковая производительность грид составила 1,04 TFLOPS.

В качестве исходных данных использовались тестовые наборы Frequent Itemset Mining Dataset Repository [13]. Характеристики наборов данных представлены в табл. 1.

Табл. 1. Характеристики используемых наборов данных

	Файл	Количество транзакций	Средняя длина транзакции	Минимальная поддержка
I	T10I4D100K.dat	100000	10	1%
II	T25I20D100K.dat	100000	25	1,5%
III	T40I10D100K.dat	100000	40	5%

Результаты проведенных экспериментов показали, что время поиска ассоциативных правил на используемых наборах данных достигает минимального значения при использовании 28-30 вычислительных узлов с ускорением в 6-9 раз (см. Рис. 5).

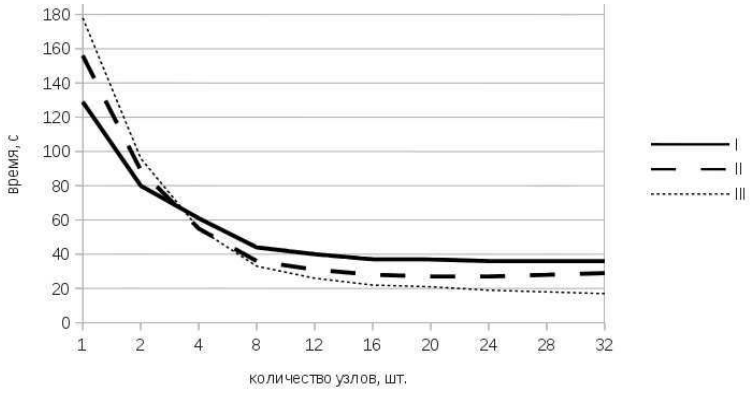

Рис. 5. Время выполнения для различного числа используемых вычислительных узлов

На данный момент на основе разработанной программы проводятся эксперименты по оценке производительности поиска ассоциативных правил в больших базах данных (50 — 100 Гб). Кроме того, реализуется тестовый проект по анализу данных об используемых тарифах и услугах пользователей сотовой связи компании Beeline в республике Карелия. Исходные данные состоят из записей о дате заключения контракта, названии тарифа, дата подключения и названиях услуг и др. ***(эти результаты будут представлены в финальной версии статьи).***

Заключение. В статье представлены результаты исследований, связанных с реализацией на BOINC-грид алгоритмов Data Mining по поиску ассоциативных правил в больших наборах исходных данных. Описана реализация алгоритма, предназначенного для работы в распределенной среде, представлены результаты экспериментов по оценке производительности разработанного ПО на тестовых базах данных.

ЛИТЕРАТУРА:

1. Дюк В., Самойленко А. Data Mining: учебный курс (+CD) СПб: Изд. Питер, 2001. 368 с.
2. The Fourth Paradigm: Data-Intensive Scientific Discovery, 2009, URL: http://research.microsoft.com/en-us/collaboration/fourthparadigm
3. Обзор технологий, google.ru, URL: http://www.google.ru/intl/ru/about/corporate/company/tech.html
4. Loek Essers: CERN pushes storage limits as it probes secrets of universe, URL: http://news.idg.no/cw/art.cfm?id=FF726AD5-1A64-6A71-CE987454D9028BDF
5. Yevgeniy Sverdlik. Making Way for Big Data // DatacenterDynamics Focus, April/May 2012, Volume 3, Issue 21.
6. *Л. Черняк.* Большие Данные — новая теория и практика// *Открытые системы. СУБД*. — М.: Открытые системы, 2011. — № 10. — ISSN 1028-7493.
7. Прорывная технология машинного перевода и вокруг нее. PC WEEK, №9, 12 апреля 2011 г.
8. Е. Е. Ивашко, Н. Н. Никитина. Организация квантовохимических расчетов с использованием пакета Firefly в гетерогенной грид-системе на базе BOINC // Вычислительные методы и программирование, Том 13, 2012 г., с. 8 — 12.
9. BOINC: Программное обеспечение с открытым исходным кодом для организации добровольных распределённых вычислений и распределённых вычислений в сети. URL: *http://boinc.berkeley.edu/index.php*
10. Проект добровольных вычислений Climateprediction.net. URL: http://climateprediction.net
11. Проект добровольных вычислений SETI@home. URL: http://setiathome.berkeley.edu
12. Центр высокопроизводительной обработки данных ЦКП КарНЦ РАН / Институт прикладных математических исследований Карельского научного центра РАН. URL: *http://cluster.krc.karelia.ru*
13. Frequent Itemset Mining Dataset Repository, URL: *http://fimi.ua.ac.be/*
14. Foster I. The Grid: Blueprint for a New Computing Infrastructure. — Morgan Kaufmann Publishers. — ISBN ISBN 1-55860-475-8.