САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-Механический факультет

Кафедра Системного Программирования

Анускина Ирина Михайловна

Анализ потока запросов к системе хранения

Бакалаврская работа

Заведующий кафедрой:

д.ф. – м.н., профессор Терехов А. Н.

Научный руководитель:

д.ф. – м.н., профессор Новиков Б. А.

Рецензент:

к.ф. – м.н., доцент Михайлова Е. Г.

Санкт-Петербург

2014 г.

SAINT-PETERSBURG STATE UNIVERSITY

Mathematics and Mechanics Faculty

Software Engineering Department

Irina Anuskina

Query stream analysis in the storage system

Bachelors' Thesis

Head of Department:

Professor Andrey Terekhov

Scientific advisor:

Professor Boris Novikov

Reviewer:

Associate Professor Elena Mikhailova

Saint-Petersburg

2014

Оглавление

Введение

В современном обществе, в условиях возрастающей ценности информации, крайне важно организовать систему хранения, которая обеспечит бесперебойный доступ к актуальным данным. Возможность доступа к данным и управления ими является необходимым условием для выполнения бизнес-процессов.

Доступ к данным невозможен как в случае выхода из строя каких-либо вычислительных средств, входящих в инфраструктуру системы хранения, так и в случае отсутствия необходимой производительности для выполнения прикладных задач.

Зачастую во время пиковых нагрузок быстродействие систем хранения снижается. Вычислительных ресурсов, которых хватает в период средней загруженности системы, может оказаться недостаточно во время активного обращения к данным на устройствах хранения.

Сократить задержки во время пиковых нагрузок возможно путем вовлечения дополнительных вычислительных ресурсов и расширения инфраструктуры системы хранения, однако это приведет к дополнительным затратам на закупку нового оборудования, а также повлечет за собой расходы на его последующее обслуживание – администрирование, мониторинг, устранение неисправностей. Поэтому нужно иметь возможность выравнивать пики нагрузки системы без привлечения дополнительных ресурсов.

Для достижения этих целей необходимо отслеживать пиковые нагрузки, анализировать их поведение и иметь возможность прогнозировать их дальнейшее появление. Для этого требуется понимать, к каким данным происходит обращение в период таких максимальных активностей пользователей системы.

На основании информации о востребованности тех или иных данных в период пиковых нагрузок, можно оптимизировать структуру системы хранения и тем самым увеличить ее быстродействие и сократить время доступа к данным. Например, можно осуществлять загрузку актуальных данных на более быстродействующие устройства хранения, такие как SSD-диски. И наоборот, данные,

которые были востребованы пользователями единожды, но при этом требуют долгосрочного хранения, можно помещать на устройства с низкой скоростью доступа, освободив тем самым дорогостоящие быстродействующие носители для более востребованных данных.

Получить информацию о том, к каким данным происходит обращение и в какой период времени, можно на основании журнала запросов к базе данных. Журнал запросов хранит информацию о том, в какой интервал времени выполнялся тот или иной запрос, сколько времени на его выполнение было затрачено и сколько вычислительных ресурсов было задействовано во время его исполнения.

Однако информации о каждом запросе в отдельности недостаточно, для того чтобы отследить реальные активности пользователей системы. Как правило бизнес-процессы, оказывающие нагрузку на систему хранения, сопровождаются некоторым набором запросов к данным.

Поэтому важно рассматривать запросы не каждый в отдельности, а пытаться находить группы взаимосвязанных запросов, которые выполняются приблизительно в одни и те же временные интервалы. Предполагается, что такие группы запросов представляют из себя некоторые бизнес-активности или группы взаимодействующих активностей. Впоследствии обнаруженные бизнес-активности можно анализировать, изучать периодичность и прогнозировать их дальнейшее поведение.

1 Постановка задачи

На основании вышеперечисленных аспектов можно сформулировать цель данной работы и выделить её задачи.

Целью данной работы является выявление бизнес-активностей, которые оказывают пиковые нагрузки на систему хранения, на основании журнала запросов, а также изучение их поведения и периодичности.

Для достижения цели данной работы были поставлены следующие задачи.

1. Выделить фоновые активности, которые вносят малый вклад в общую нагрузку системы:
    - отделить запросы без ярко-выраженных пиков активности;
    - отделить запросы с малой суммарной активностью.
2. Осуществить поиск групп взаимосвязанных запросов – бизнес-активностей:
    - сформировать характеристические вектора запросов;
    - понизить размерность характеристических векторов запросов;
    - произвести анализ на множестве характеристических векторов запросов.
3. Проанализировать влияние дополнительной информации, получаемой из атрибутов источников запросов на выявление периодов бизнес-активностей и их прогнозирование.

2 Обзор существующих методов

В данном разделе представлены существующие решения для поиска бизнес-активностей на основании анализа журнала запросов, рассмотрены методы, которые могут быть использованы для поиска групп взаимосвязанных запросов, а также описаны подходы для изучения периодичности событий и их предсказания, которые представляют интерес для анализа поведения обнаруженных бизнес-активностей.

2.1 Поиск бизнес-активностей

Задача поиска бизнес-активностей на основании журнала запросов была предложена в работе [1]. В ней было выдвинуто предположение о том, что запросы, которые выполняются приблизительно в одно и то же время представляют из себя некоторые бизнес-активности.

Для поиска групп взаимосвязанных запросов в данной работе было предложено ввести величину взаимосвязанности, которая вычисляется как отношение числа временных интервалов, в которые запросы появлялись вместе, возведенное в квадрат, к произведению числа интервалов, в которых появлялся каждый из запросов в отдельности:

$$M(q_1, q_2) = \frac{|q_1 \cap q_2|^2}{|q_1| * |q_2|}$$

Если эта величина превышает некоторого заданного порога, то два запроса считаются связанными, и могут быть объединены. Чем выше установленный порог, тем более сильно-связанное множество запросов получается на выходе. Поиск осуществляется итеративно, изначально находятся группы, состоящие из двух запросов, затем из трех и т.д. Так на шаге, когда найдена группа, состоящая из n запросов, величина взаимосвязанности вычисляется следующим образом:

$$M(G, q) = \frac{(G \cap q)^2}{avgsize(G) * |q|}$$

$$avgsize(G) = \left\{ \frac{|q_1| * \dots * |q_n|}{n}, q_i \in G, 1 \leq i \leq n \right\}$$

В результате апробации данного подхода выяснилось, что он позволяет обнаруживать порядка 60% известных бизнес-активностей.

2.2 Метод кластерного анализа

Кластеризация позволяет организовать данные в однородные группы. Каждая группа состоит из объектов, которые похожи между собой и отличаются от объектов других групп. При таком подходе каждый объект рассматривается как вектор, компоненты которого представляют какие-либо численные характеристики данного объекта:

$$X_i = (x_1, x_2, \dots, x_n)$$

Задача состоит в том, чтобы разбить множество характеристических векторов:

$$\{X_1, X_2, \dots, X_k\}$$

на непересекающиеся подмножества – кластеры, так, чтобы каждый кластер состоял из векторов близких по некоторой метрике ρ, а объекты разных кластеров существенно отличались. Метрика позволяет определить степень схожести между объектами. Наиболее распространенной метрикой является метрика Евклида (евклидово расстояние) [2] :

$$\rho(X_i, X_j) = \sqrt{\sum_{k}^{n} (x_{ik} - x_{jk})^2}$$

Спектр применения кластерного анализа очень широк, и на сегодняшний день существует множество различных подходов к реализации данного метода [3].

Существуют подходы кластеризации [4], [5], в основе которых лежит понятие информационной энтропии [6]. Поскольку данные в рамках одного кластера достаточно однородны, то и значение энтропии каждого отдельного кластера должно быть мало.

При таком подходе группировка объектов по кластерам осуществляется таким образом, чтобы минимизировать общую энтропию по всем найденным кластерам.

В рамках данной работы метод кластерного анализа позволил бы организовать запросы к системе хранения в однородные группы, в пределах которых запросы имеют схожие вектора, которые коррелируют между собой, а, следовательно, могут принадлежать одному бизнес-процессу. В качестве компонент характеристических векторов в данном случае могли бы выступать активности запроса в каждый из рассматриваемых временных интервалов.

Недостаток данного подхода заключается в том, что качество кластеризации значительно зависит от размерности пространства характеристических векторов, поэтому для получения приемлемых результатов требуется предварительно осуществить процедуры понижения размерности характеристических векторов.

2.3 Анализ периодичности и предсказание бизнес-активностей

В ходе совместного исследовательского проекта кафедры информационно-аналитических систем СПбГУ и компании EMC, целью которого было обнаружение шаблонов обращений к системе хранения на основании журнала запросов и изучение их периодичности, был разработан алгоритм предсказания бизнес-активностей.

Идея алгоритма заключается в следующем. Активности запросов подразделяются на четыре типа: высокие, с ярко выраженным пиком, средние, которые встречаются как правило в рабочие часы, низкие, встречающиеся в выходные или праздничные дни, и нулевые – крайне редкие активности. Временной ряд [7] активностей запроса кодируется с помощью четырех символов, обозначающих соответствующий тип активности в тот или иной промежуток времени. Далее на

основании такого символьного представления выполняется анализ периодичности [8] символов путем полного перебора, отсеиваются символы со слишком длинными периодами, а также символы с малым числом повторений. Затем полученные данные используются для генерации предсказаний.

На основании временного ряда активностей запроса за последние три недели, извлекаются периодические символы, затем каждый символ повторяется с найденным периодом на следующую неделю. После чего каждый предсказанный символ заменяется на среднее значение того уровня активности, который он обозначает.

Аналогичным образом на основании временного ряда активностей за месяц, могут быть сгенерированы предсказания на следующие пять месяцев

Для оценки качества реализованного алгоритма предсказания используется две метрики – точности и полноты.

$$precision = \frac{relevantPredicted}{relevantPredicted + errorPredicted}$$

$$recall = \frac{relevantPredicted}{relevantPredicted + nonPredicted}$$

$relevantPredicted -$ количество релевантно предсказанных активностей,

$errorPredicted -$ количество ложно предсказанных активностей

$nonPredicted -$ активности, которые не были предсказаны

Как видно из формул, приведенных выше, точность – это доля прогнозируемых активностей, которые действительно имели место быть, полнота – доля реальных активностей, которые были предсказаны.

В ходе тестирования алгоритма было обнаружено, что средняя точность предсказания составляет 63%, средняя полнота предсказания 47%.

3 Выделение фоновых активностей

Первостепенной задачей в рамках данной работы является выделение фоновых активностей, которые вносят малый суммарный вклад в общую нагрузку системы.

Анализ активностей значительно усложняется ввиду большого количество запросов. При этом значительная часть из этих запросов представляет из себя регулярные ежедневные активности, которые не представляют интереса для нашего изучения. Такие активности не оказывают большой нагрузки на систему и не требуют дополнительных вычислительных ресурсов. Кроме того, такие активности создают общий шум, на фоне которого сложно обнаруживать интересующие нас бизнес-процессы.

Для выделения регулярных ежедневных запросов необходимо проанализировать распределение активностей на протяжении всего рассматриваемого промежутка времени. Имея общую картину того, как распределены активности по каждому из дней в рамках журнала запросов, можно выделить дни с ярко-выраженными пиками активности и в дальнейшем исследовать запросы, которые выполнялись именно в эти дни.

Для выделения запросов без ярко-выраженных пиков в данной работе было принято решение интерпретировать характеристические вектора запросов как выборки значений и вычислять размах выборки, поскольку эта величина позволяет определить степень разброса данных. В данном случае выборка интерпретируется как вариационный ряд:

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

Размах выборки, вычисляется как разность между крайними элементами вариационного ряда:

$$R = x_{(n)} - x_{(1)}$$

На конце вариационного ряда может находиться потенциальный пик активности запроса, и если размах близок к нулю, то активность запроса в целом равномерна, чем эта величина больше отличается от 0, тем более ярко выраженные пики имеет запрос.

Однако может оказаться так, что на конце вариационного ряда окажется некоторый случайный выброс, при этом размах выборки будет отличен от нуля, однако запрос окажется в целом регулярным. Альтернативой этому решению является вычисление интерквантильного размаха выборки:

$$Q = x_{(3n/4)} - x_{(n/4)}$$

Такая величина является робастной – нечувствительной к различным отклонениям и случайным выбросам.

4 Поиск бизнес-активностей

После выделения фоновых активностей необходимо осуществить поиск бизнес-активностей, которые оказывают пиковые нагрузки на систему хранения. Для этого, как и предполагалось, требуется выделить группы взаимосвязанных запросов.

Каждый запрос может быть представлен как вектор, компоненты которого представляют из себя численные характеристики запроса за выбранный интервал времени, который, как правило, составляет один час. Этими характеристиками могут быть суммарное время выполнения запроса или суммарное процессорное время, затраченное на исполнение запроса.

Таким образом для поиска бизнес-активностей необходимо произвести анализ на множестве характеристических векторов запросов.

4.1 Понижение размерности характеристических векторов запросов

Поиск бизнес-активностей на множестве характеристических векторов запросов значительно усложняется виду большой размерности данных векторов, которая зависит от размера выбранных временных интервалов. Так, имея данные журнала запросов за последние три месяца и временные интервалы продолжительностью в один час, мы будем получать вектора, состоящие более чем из двух тысяч компонент.

Необходимо понизить размерность векторов, используя более продолжительные интервалы времени, нежели просто промежуток в один час. В данной работе было принято решение использовать в качестве таких интервалов дни недели, поскольку они имеют более естественную природу, так как многие бизнес-процессы организованы в зависимости от дней недели, и в этом случае удастся отследить пиковые нагрузки с периодом в одну неделю.

Таким образом, необходимо для каждого запроса суммировать активности по каждому из семи дней недели и получать характеристические вектора, состоящие из семи компонент.

4.2 Кластеризация запросов методом K-means

На множестве характеристических векторов с пониженной размерностью было решено произвести кластеризацию с помощью распространенного алгоритма K-means.

Перед запуском алгоритма вектора необходимо нормировать, для этого была выбрана самая распространённая евклидова норма:

$$\|x\| = \sqrt{\sum_i |x_i|^2}$$

После нормализации все значения компонент вектора приводятся к диапазону $[0,1]$.

Алгоритм K-means представляет собой итеративную процедуру, в процессе которой выполняются следующие шаги.

1. Выбирается число кластеров k.
2. Из исходного множества данных случайным образом выбирается k векторов, которые будут служить начальными центрами кластеров.
3. Для каждого оставшегося вектора исходной выборки определяется ближайший к ней центр кластера. Таким образом образуются начальные кластеры.
4. Пересчитываются центры кластеров. Каждый центр – это вектор, элементы которого представляют собой средние значения компонент, вычисленные по всем векторам кластера.

Последние два шага итеративно повторяются. На каждой итерации происходит изменение границ кластеров и смещение их центров. В результате минимизируется расстояние между элементами внутри кластеров. Остановка алгоритма производится тогда, когда границы кластеров и расположения центров не перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор векторов.

4.3 Кластеризация запросов с использованием критерия минимальной энтропии

Для кластера запросов можно определить энтропию, как меру его однородности.

Классическое определение энтропии случайной величины X имеет вид:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Где $(x_1, x_2, \dots, x_n)$ - множество возможных значений случайной величины X, $p(x_i) -$ вероятность принятия значения x_i.

Для кластера C, состоящего из векторов $X_1, \dots, X_m$, где $X_i = (x_{i1}, \dots, x_{in})$, энтропия вычисляется следующим обрзом:

$$H(\mathrm{C}) = -\sum_{i=1}^{n} \sum_{x} p_i(x) \log_2 p_i(x), \text{где}$$

$p_i(x) -$ частота появления значения x в i компоненте каждого вектора

Из определения видно, что энтропия позволяет определить насколько однородны вектора, входящие в кластер по каждой из компонент.

Чтобы оценить общее качество кластеризации можно просуммировать значения энтропии, умноженные на веса кластеров, по каждому из имеющихся кластеров. Вес кластера определяется как отношение числа векторов, содержащихся в нем, к общему числу векторов.

$$H = \sum_{i=1}^{k} H(C_k) * \frac{size(C_k)}{N}, \text{где } N - \text{общее число векторов.}$$

Чем меньше полученное значение энтропии, тем более однородные кластеры. Поэтому критерий минимизации энтропии можно использовать при кластеризации запросов.

Первоначально необходимо выделить начальные кластеры запросов. Это можно осуществить, сформировав группы векторов, близких друг другу по какой-либо из элементарных метрик, например, евклидовой.

Для каждого из оставшихся запросов можно определить ближайший к нему кластер, такой, что при добавлении к нему данного запроса общее значении энтропии по всем кластерам окажется минимальным.

5 Анализ качества прогнозирования бизнес-активностей

В рамках данной работы была поставлена задача оценить влияние дополнительных атрибутов, получаемых из журнала активностей объектов базы данных на качество предсказания алгоритма, описанного в разделе 2.3.

Одним из таких атрибутов является идентификатор сессии соединения с базой данных. Таким образом, каждый объект имеет разбиение своей активности по сессиям.

Было выдвинуто предположение о том, что поведение объектов в рамках одной сессии более регулярное, и качество алгоритма предсказания, запущенного отдельно на таких объектах, может улучшиться. Поэтому было принято решение сгруппировать объекты, активные в рамках одной и той же сессии, и запустить алгоритм предсказания на таких группах, а в дальнейшем сравнить полученные результаты точности и полноты предсказания с теми, которые были достигнуты в рамках проекта и представлены в разделе 2.3.

Однако атрибут идентификатора сессии для осуществления такого анализа не подходит, поскольку при каждом новом обращении к объекту его значение может изменяться. В то же время сессия характеризуется еще и такими атрибутами, как идентификатор машины, идентификатор сервиса, с которых осуществлялось обращение к объекту. Такие атрибуты уже могут быть использованы для группировки объектов и анализа качества предсказания.

6 Эксперименты

6.1 Поиск бизнес-активностей

В ходе экспериментов использовались данные о запросах к системе хранения компании, занимающейся отправкой портовых грузов. Эти данные были получены из журнала запросов и объединены по временным интервалам продолжительностью в 1 час. Они содержат такую информацию как идентификатор запроса, суммарное время выполнения, суммарное процессорное время, число повторений запроса, начало и конец временного интервала (табл. 1).

Идентификатор запроса	**Суммарное время выполнения**	**Суммарное процессорное время**	**Число повторений запроса**	**Начало временного интервала**	**Конец временного интервала**
b1gasr76z864c	2.184406	2.182057	375	2011-07-16 22:00:04	2011-07-16 23:00:06
70697c540n0f2	2.850525	0.932858	375	2011-07-16 22:00:04	2011-07-16 23:00:06
6r558vfksuvf5	4.152529	3.765698	375	2011-07-16 22:00:04	2011-07-16 23:00:06
…	…	…	…	…	…
4qx1uxkn6jm1n	56.092396	37.351832	74	2011-10-24 08:00:04	2011-10-24 09:00:11

Таблица 1: Описание данных

6.1.1 Выделение фоновых активностей

Для проведения экспериментов были выделены дни с пиковыми нагрузками, в которые суммарное время выполнения запросов превышало 200000 секунд (рис. 1).

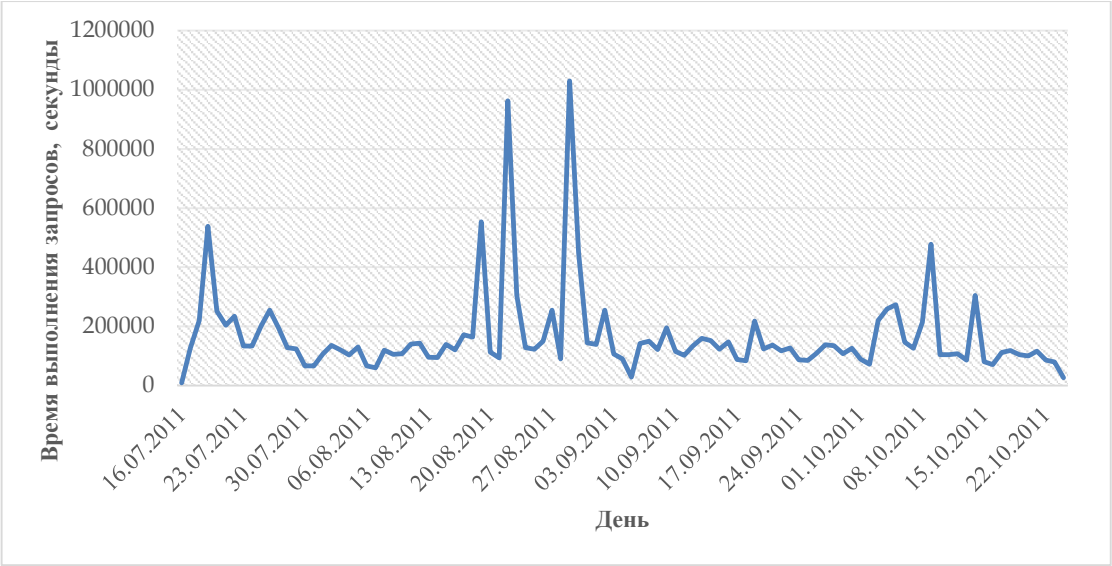

Рисунок 1: Распределение активностей запросов

Из 1325 запросов было выделено 845 запросов, выполняющихся в дни с пиковой нагрузкой.

На множестве характеристических векторов запросов, которые выполнялись в дни с пиковой нагрузкой был произведен кластерный анализ. С целью понижения размерности векторов активности запросов были просуммированы по каждому из дней недели (рис. 2).

Идентификатор запроса	16.07.11 22:00:04 - 16.07.11 23:00:06	16.07.11 23:00:06 - 16.07.11 00:00:08	16.07.11 00:00:08 - 16.07.11 01:00:11	16.07.11 01:00:11 - 16.07.11 02:00:13	...	24.10.11 05:00:56 - 24.10.11 06:00:58	24.10.11 06:00:58 - 24.10.11 07:00:00	24.10.11 07:00:00 - 24.10.11 08:00:04	24.10.11 08:00:04 - 24.10.11 09:00:11
asmwy9n0tynp2	0,022731	0,032134	0,011957	0,043338	...	0,008401	0,008415	0,025352	0,064584

> 2000 компонент

Идентификатор запроса	Понедельник	Вторник	Среда	Четверг	Пятница	Суббота	Воскресенье
asmwy9n0tynp2	79,943503	93,599126	15,454859	12,022606	11,2883	23,080351	4,644218

7 компонент

Рисунок 2. Понижение размерности характеристических векторов запросов

6.1.2 Алгоритм кластеризации K-means

Для кластеризации использовался инструмент Weka[1]. На вход алгоритма подавались вектора, состоящие из семи компонент и нормированные с помощью евклидовой нормы. Каждая компонента вектора соответствует суммарному времени выполнения запроса в тот или иной день недели. Число кластеров было выбрано равным 25.

На выходе алгоритма были получены кластеры, включающие запросы, выполняющиеся преимущественно в тот или иной день недели (табл. 2).

Кластер	Число запросов	Процент от общего числа запросов
Понедельник	59	7%
Вторник	39	5 %
Среда	14	2%
Четверг	11	1%
Пятница	40	5%
Суббота	15	2%
Воскресенье	30	4%

Таблица 2. Распределение запросов по кластерам. Алгоритм K-means

В кластеры попали запросы, чья средняя активность в соответствующий день недели составляет более 65% от общей нагрузки по всем дням. Так в кластер, характеризующий понедельник, попали запросы, чья активность в понедельник составляет порядка 80% от общей нагрузки (рис. 3, рис. 4).

[1] http://www.cs.waikato.ac.nz/ml/weka/

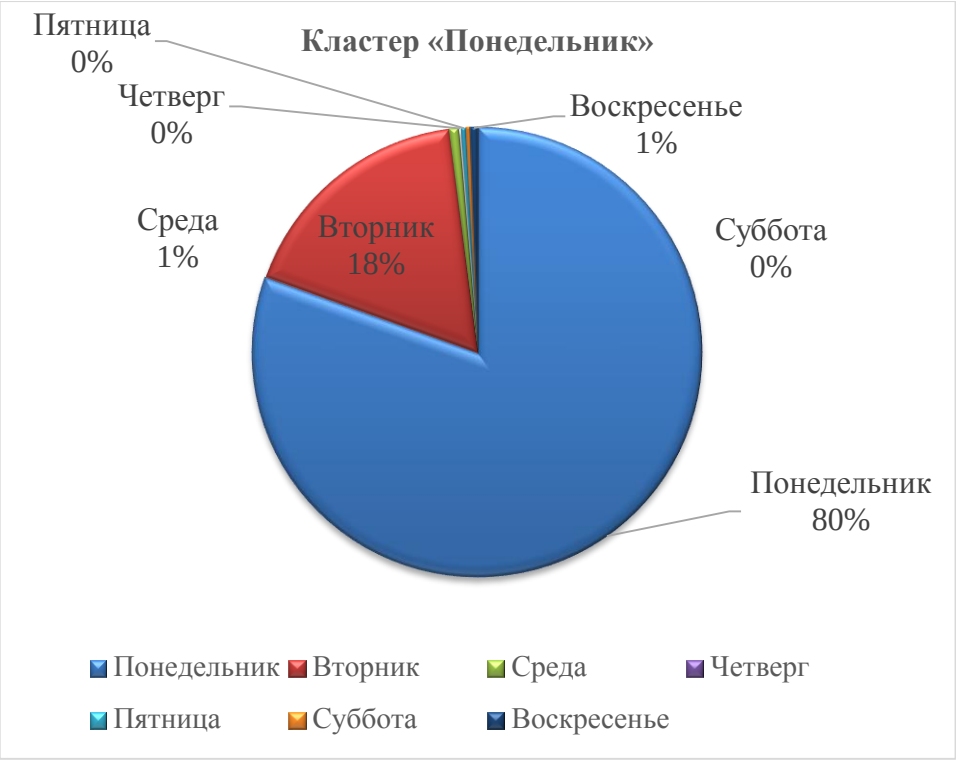

Рисунок 3: Средняя активность запросов в кластере “Понедельник” – 1

Рисунок 4: Средняя активность запросов в кластере “Понедельник” - 2

В ходе кластеризации удалось отсеять запросы, активность которых распределена по всем дням без ярко выраженных пиков. Такие запросы попали в отдельные кластеры, и они не представляют интереса для дальнейшего исследования бизнес-активностей.

6.1.3 Кластеризация с использованием критерия минимальной энтропии

Для экспериментов были использованы те же данные, которые были описаны в начале данного раздела. Вектора запросов были сформированы путем суммирования активностей по дням недели. Вектора запросов без ярко выраженных пиков и с малой суммарной активностью были исключены из рассмотрения.

В ходе экспериментов были выделены начальные кластеры запросов. Из 845 векторов было сформировано 29 групп, каждая из которых включает в среднем по 3-5 запросов. Запросы объединялись по ближайшему расстоянию Евклида.

Оставшиеся 724 запроса были распределены по образовавшимся кластерам на основании критерия минимальной общей энтропии по всем кластерам.

Результирующие кластеры, полученные в рамках экспериментов представлены ниже (табл.3).

Кластер	**Число запросов**	**Процент от общего числа запросов**
Понедельник	73	9%
Вторник	35	4 %
Среда	30	4%
Четверг	43	5%
Пятница	82	10%
Суббота	17	2%
Воскресенье	29	3%
Понедельник-среда	24	3%
Понедельник-вторник	37	4%
Понедельник-пятница	9	1%
Среда-Четверг	9	1%
Четверг-Пятница	68	8%

Пятница-Суббота	9	1%
Среда-Пятница	11	1%
Вторник-Пятница	14	2%

Таблица 3. Распределение запросов по кластерам. Критерий минимальной энтропии.

Как видно из таблицы результатов, в ходе экспериментов удалось обнаружить не только кластеры запросов, которые выполняются преимущественно в тот или иной день недели, но запросы, активность которых распределена по каким-либо двум дням недели, в остальные же дни такие запросы практически неактивны.

На следующей диаграмме представлена активность запросов кластера «Понедельник» (рис. 5). На диаграмме отображены столбцы активности для каждого запроса кластера. Каждому запросу соответствует 7 столбцов разных цветов, представляющих активность данного запроса в каждый из дней недели.

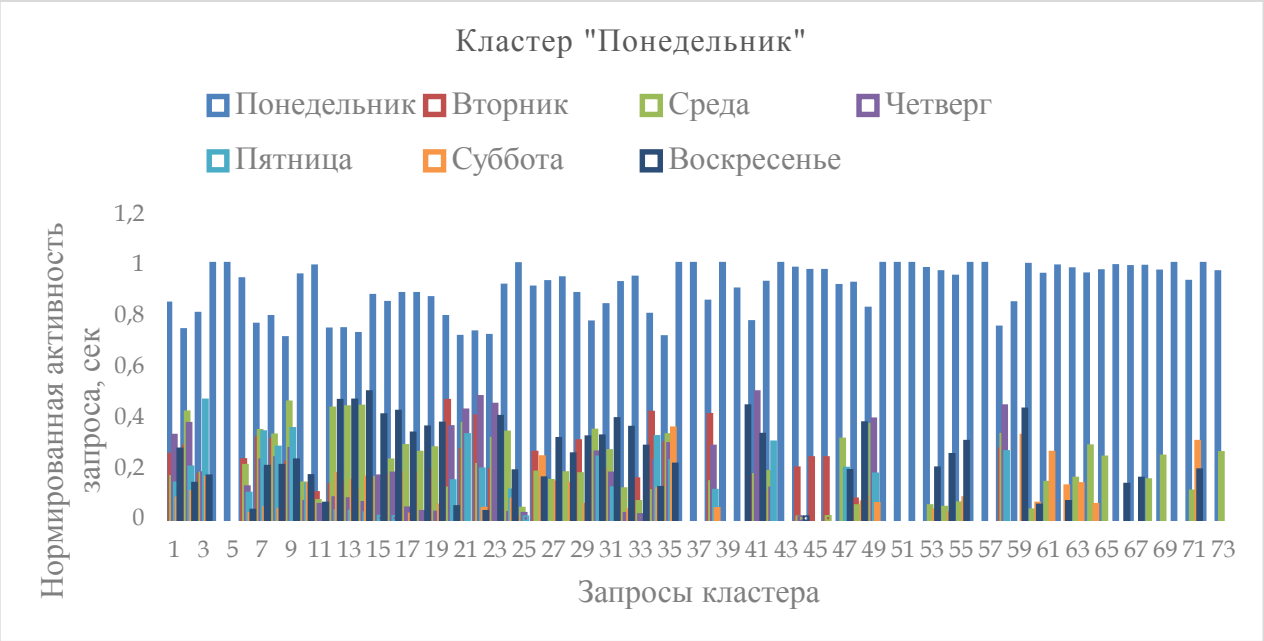

Рисунок 5. Активность запросов кластера "Понедельник"

Аналогичные данные представлены ниже для кластера запросов, чья активность распределена преимущественно в понедельник и четверг, в остальные же дни такие запросы имеют малую активность. (рис. 6)

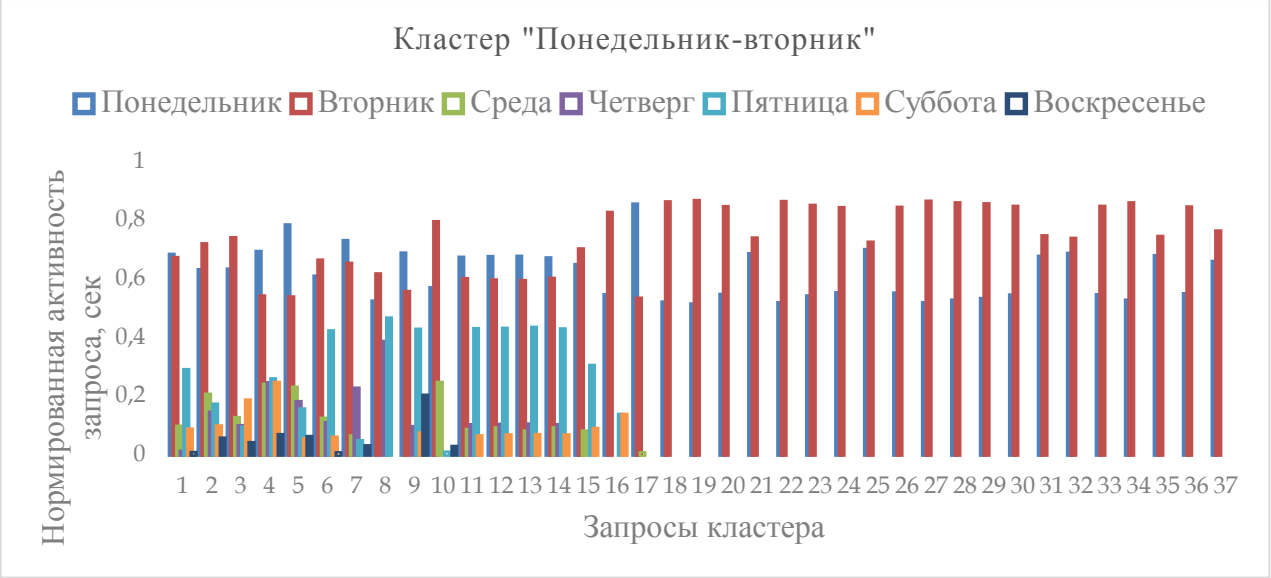

Рисунок 6. Активность запросов кластера "Понедельник-вторник"

6.2 Анализ качества прогнозирования бизнес-активностей

В ходе данных экспериментов использовались данные из совместного исследовательского проекта кафедры информационно-аналитических систем СПбГУ и компании EMC.

Для экспериментов были выбраны данные об активности объектов за промежуток 07.10.12-07.11.12, содержащие 668 записей и данные за промежуток 08.05.2012 – 03.02.2013, содержащие 1446 записей.

Данные об объектах содержат следующую информацию: идентификатор объекта (таблицы или другого объекта базы данных), начало временного интервала активности объекта, идентификатор сессии, время выполнения всего запроса, время выполнения операторов запроса, работающих с текущим объектом (табл. 4).

OBJECT_ID	RUN_TIME	SESSION_CONTEXT_ID	TOTAL_QUERY_TIME	TOTAL_IO_WAIT
372767	07-10-2012 00:00	26813212	7606000	200000

Таблица 4. Описание данных об объектах

Данные о сессиях содержат следующую информацию: идентификатор сессии, имя пользователя, идентификатор компьютера, идентификатор терминала, аббревиатура сервиса (табл. 5).

SESSION_CONTEXT_ID	SYSTEM_USER_NAME	MACHINE	TERMINAL	SERVICE_NAME
6455828	"SYSTEM_USER_NAME_1"	"MACHINE_2"	"TERMINAL_1"	"SERVICE_NAME_0"

Таблица 5. Описание данных о сессиях

На основании данных о трехнедельной активности объектов создаются характеристические вектора с суммированием активностей по 1-часовым интервалам. В качестве характеристики активности был выбран атрибут объекта TOTAL_QUERY_TIME.

По данным о пятимесячной активности объектов аналогичным образом создаются характеристические вектора, но с суммированием активностей по 6-часовым интервалам, ввиду большего объема данных и высокой размерности векторов.

Реализованный в рамках проекта алгоритм на основании данных о трехнедельной активности объектов выдает предсказание на следующую неделю, По данным о пятимесячной активности объектов выдает предсказание на следующий месяц.

Для группировки объектов был выбран атрибут “SERVICE_NAME”, который имеет 28 различных значений. Цель экспериментов состояла в том, чтобы отобрать объекты с одинаковым атрибутом “SERVICE_NAME”, и запустить алгоритм предсказания только на этих объектах (рис. 7). После этого сравнить полученные результаты, с результатами, имеющимися для всех объектов.

OBJECT_ID	RUN_TIME	SESSION_CONTEXT_ID	TOTAL_QUERY_TIME	TOTAL _IO_WAIT
…	…	…	…	…
372767	07-10-2012 00:00	26813212	7606000	200000
372767	07-10-2012 00:00	329	94215541000	18401000
…	…	…	…	…

SESSION_CONTEXT_ID	SYSTEM_USER_NAME	MACHINE	TERMINAL	SERVICE_NAME
…	…	…	…	…
26813212	"SYSTEM_USER_NAME_1849"	"MACHINE_2078"	"TERMINAL_1"	"SERVICE_NAME_23"
329	"SYSTEM_USER_NAME_1579"	"MACHINE_2013"	"TERMINAL_1880"	"SERVICE_NAME_23"
…	…	…	…	…

Рисунок 7. Группировка объектов с одинаковым атрибутом «SERVICE_NAME»

Эксперименты были проведены как на данных о 3-недельной активности объектов с 1-часовыми временными интервалами, так и на данных о 6-месячной активностью объектов с 6 часовыми интервалами. Показатели точности и полноты предсказания, полученные на выходе экспериментов представлены ниже (табл. 6)

Данные	Точность (1-часовые интервалы)	Полнота (1-часовые интервалы)	Точность (6-часовые интервалы)	Полнота (6-часовые интервалы)
Все объекты	68,45	45,15	64,23	49,85
Объекты, сгруппированные по атрибуту “SERVICE_NAME0”	73,24	56,59	62,66	48,28

Таблица 6. Точность и полнота предсказания

Ниже представлены показатели точности и полноты предсказания, полученные при запуске алгоритма предсказания на объектах сгруппированных по различным атрибутам «номер сервиса» (рис. 8).

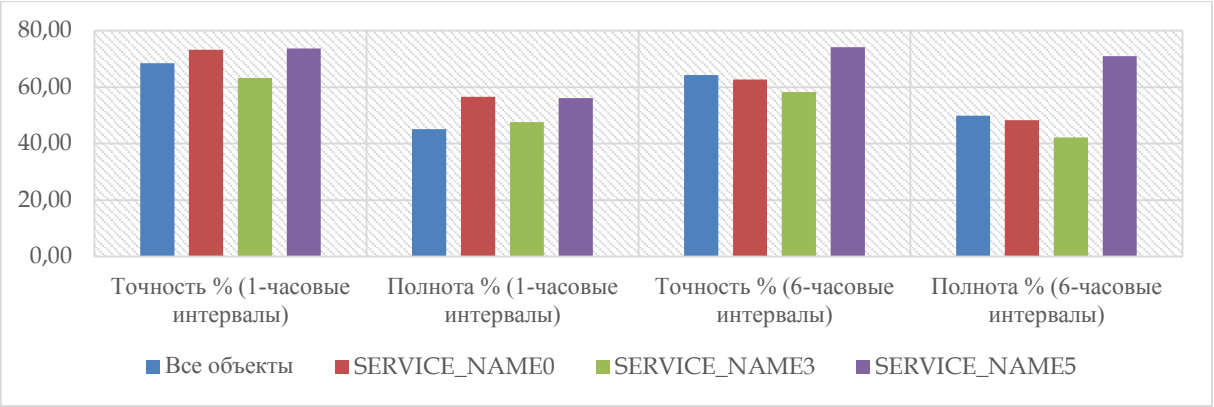

Рисунок 8. Точность и полнота предсказания для объектов с различными атрибутами SERVICE_NAME”

Также в ходе экспериментов были вычислены показатели точности и полноты для каждого дня, для которого делается предсказание, в отдельности. Для предсказаний на основании трехнедельных данных было получено 7 значений каждого показателя, для предсказаний по пятимесячным данным было получено 30 значений (рис. 9, рис. 10).

Рисунок 9. Точность и полнота предсказания для каждого дня (6-часовые интервалы)

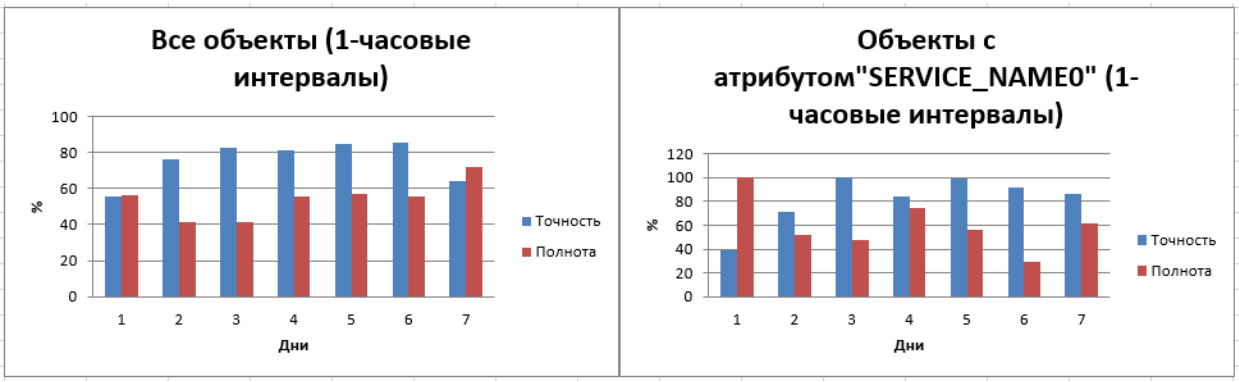

Рисунок 10. Точность и полнота предсказания для каждого дня (1-часовые интервалы)

По результатам экспериментов удалось обнаружить, что за счет группировки объектов с одинаковыми атрибутами “SERVICE_NAME”, в случае 1-часовых интервалов точность предсказания улучшилась в среднем на 2%, полнота на 8%. В случае 6-часовых интервалов точность улучшилась чуть более чем на 1%, полнота улучшилась на 4%.

Заключение

В ходе данной работы были достигнуты следующие результаты.

1. Реализован подход выделения фоновых активностей системы хранения, который отсеивает порядка 36% от общего числа запросов.
2. Реализованы процедуры предварительной нормализации и понижения размерности характеристических векторов запросов к системе хранения.
3. Проведены эксперименты с альтернативными подходами поиска групп взаимосвязанных запросов – бизнес активностей:
    - кластеризация запросов с помощью алгоритма K-means;
    - кластеризация запросов с использованием критерия минимальной энтропии.
4. Проведены эксперименты по изучению влияния такого атрибута, как идентификатор сервиса, с которого происходит обращение к объектам системы хранения, на качество прогнозирования бизнес-активностей, по результатам которых удалось обнаружить улучшение точности предсказания в среднем на 1%, полноты предсказания в среднем на 4%.

Список литературы

[1] B. Novikov, E. Michailova, D. Vasilik, E. Ivannikova and A. Pigul, "Mining logs for long-term patterns," in *Databases and Information Systems VII, Selected Papers from the Tenth International Baltic Conference, DB&IS 2012*, January 2013.

[2] P. Grabusts, "The choice of metrics for clustering algorithms," in *Proceedings of the 8th International Scientific and Practical Conference*, Rezekne, Latvia, 2011.

[4] T. Li, S. Ma and M. Ogihara, "Entropy-Based Criterion in Categorical Clustering," Department of Computer Science, University of Rochester, Rochester, NY.

[6] T. M. Cover and J. A. Thomas, "Entropy, Relative Entropy and Mutual Information," in *Elements of Information Theory*, John Wiley & Sons, Inc, 1991.

[7] P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, New York: Springer, 2002.

[8] C. Berberidis, W. G. Aref, M. Atallah, I. Vlahavas and A. K. Elmagarmid, "Multiple and Partial Periodicity Mining in Time Series Databases".