Анализ потока запросов к системе хранения

Бакалаврская работа

Анускина Ирина Михайловна

461 группа

Научный руководитель: Новиков Б.А., д.ф.-м.н., профессор

Рецензент: Михайлова Е.Г., к.ф.-м.н., доцент

Анализ потока запросов

- **Актуальность**
  - Возрастающая ценность информации
  - Потребность в бесперебойном доступе к актуальным данным

- **Проблема**

  Сокращение быстродействия системы хранения во время пиковых нагрузок

- **Предположение**

  Необходимо отслеживать пиковые нагрузки, анализировать их поведение и прогнозировать дальнейшее появление

Постановка задачи

- **Цель работы** - выявление бизнес-активностей, которые оказывают пиковые нагрузки на систему хранения, на основании журнала запросов, а также изучение их поведения и периодичности
- **Задачи:**
  - Выделить фоновые активности, которые вносят малый вклад в общую нагрузку системы
  - Осуществить поиск групп взаимосвязанных запросов – бизнес-активностей
  - Проанализировать влияние дополнительной информации, получаемой из атрибутов источников запросов на выявление периодов активностей и их прогнозирование

Существующие решения

- Исследования кафедры информационно-аналитических систем СПБГУ
  - Mining Logs for Long-Term Patterns
- Исследовательский проект компании EMC и кафедры информационно-аналитических систем
  - Long-Term Predictions for Periodic Access Patterns

Данные журнала запросов

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

Выделение фоновых активностей

- Выделение запросов с малой суммарной активностью
- Выделение запросов без ярко-выраженных пиков
  - Выборка активностей запроса:

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

$$Q = x_{(3n/4)} - x_{(n/4)} \to 0$$

Понижение размерности характеристических векторов запросов

- Суммирование активностей по дням недели

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

Поиск бизнес-активностей

- Кластерный анализ на множестве характеристических векторов запросов
  - K-means, с использованием метрики Евклида:

$$(X_i, X_j) = \sqrt{\sum_{k}^{n} (x_{ik} - x_{jk})^2}, j = 1, \dots, k$$

  - Критерий минимальной энтропии:

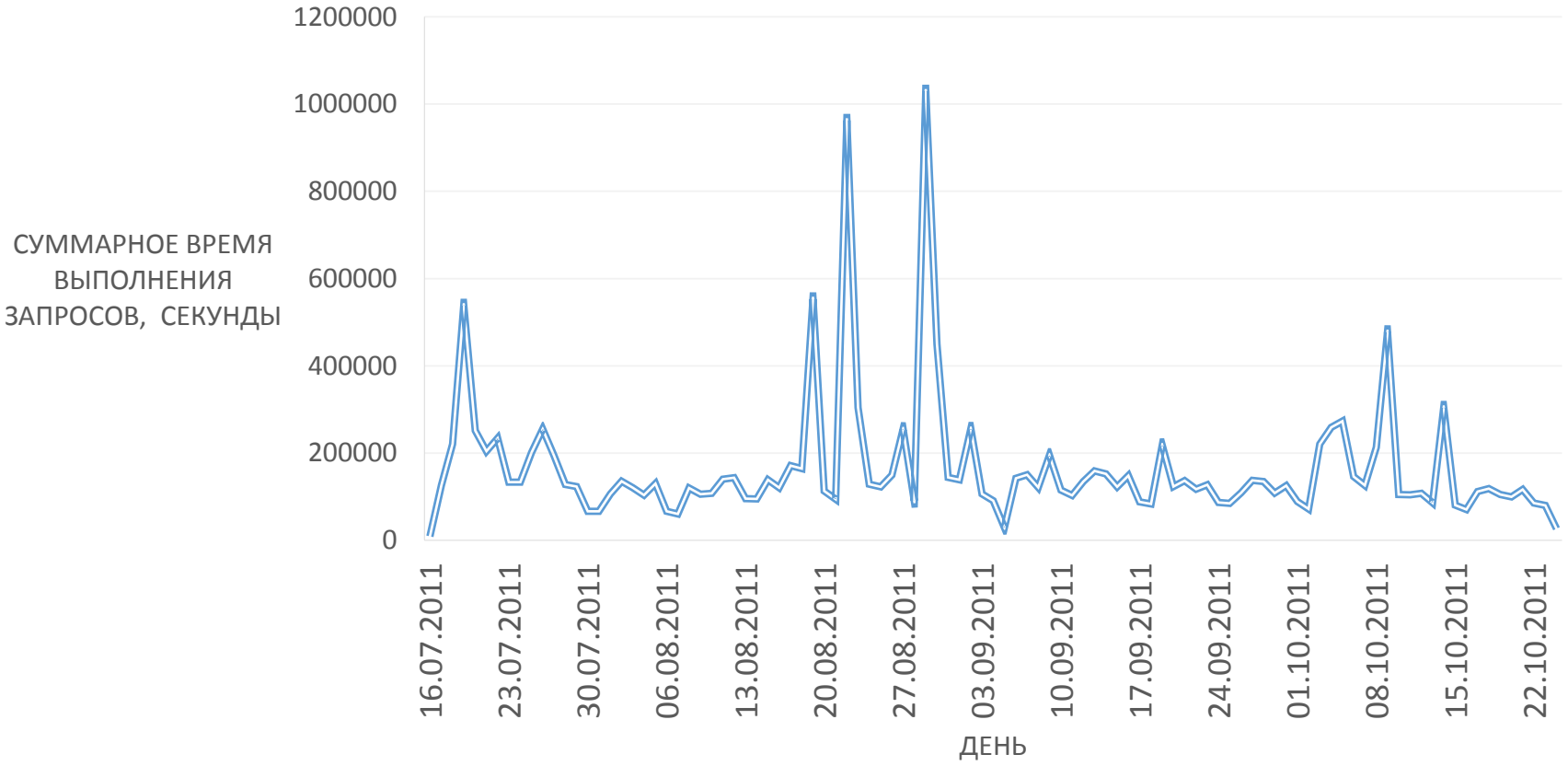
$$H = \sum_{i=1}^{k} H(C_k) * \frac{size(C_k)}{N}$$

$$N - \text{общее число векторов}$$

$$H(C_k) = -\sum_{i=1}^{n} \sum_{x} p_i(x) \log_2 p_i(x)$$

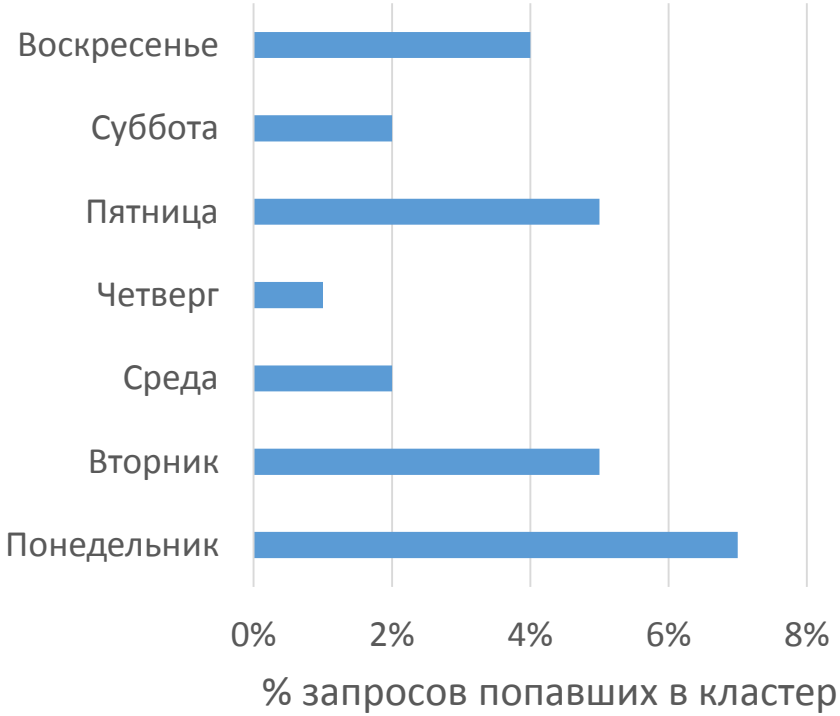
Эксперименты

- Распределение активностей запросов



Кластеризация, K-means

- Распределение запросов по кластерам

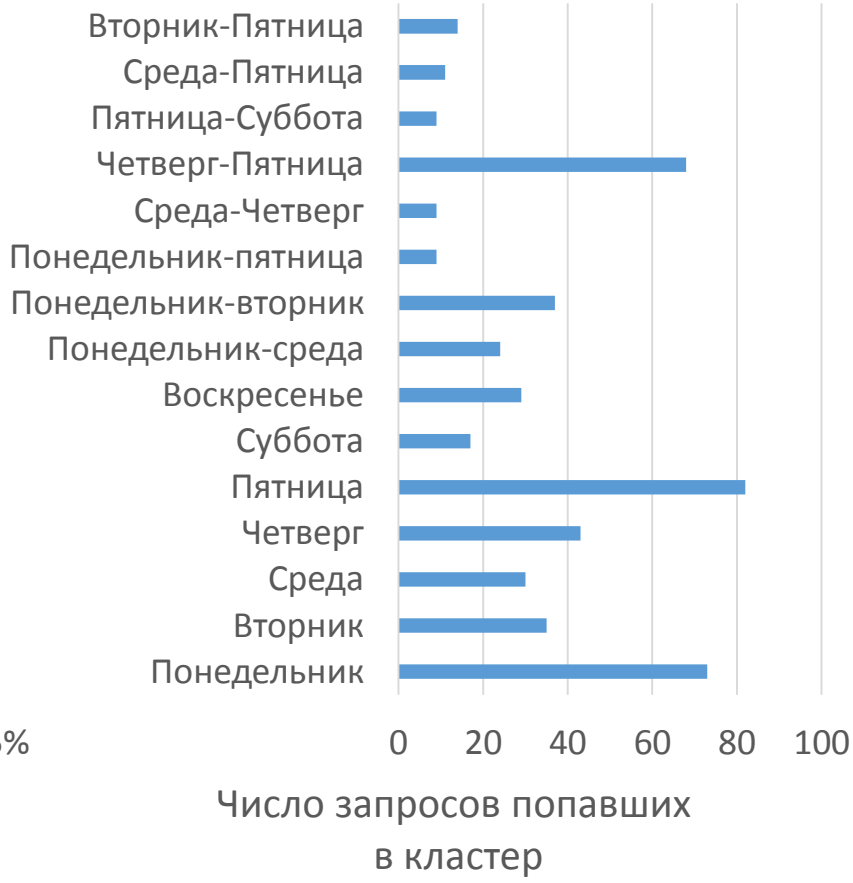


Кластеризация, K-means

- Распределение активностей запросов попавших в кластер «Вторник»

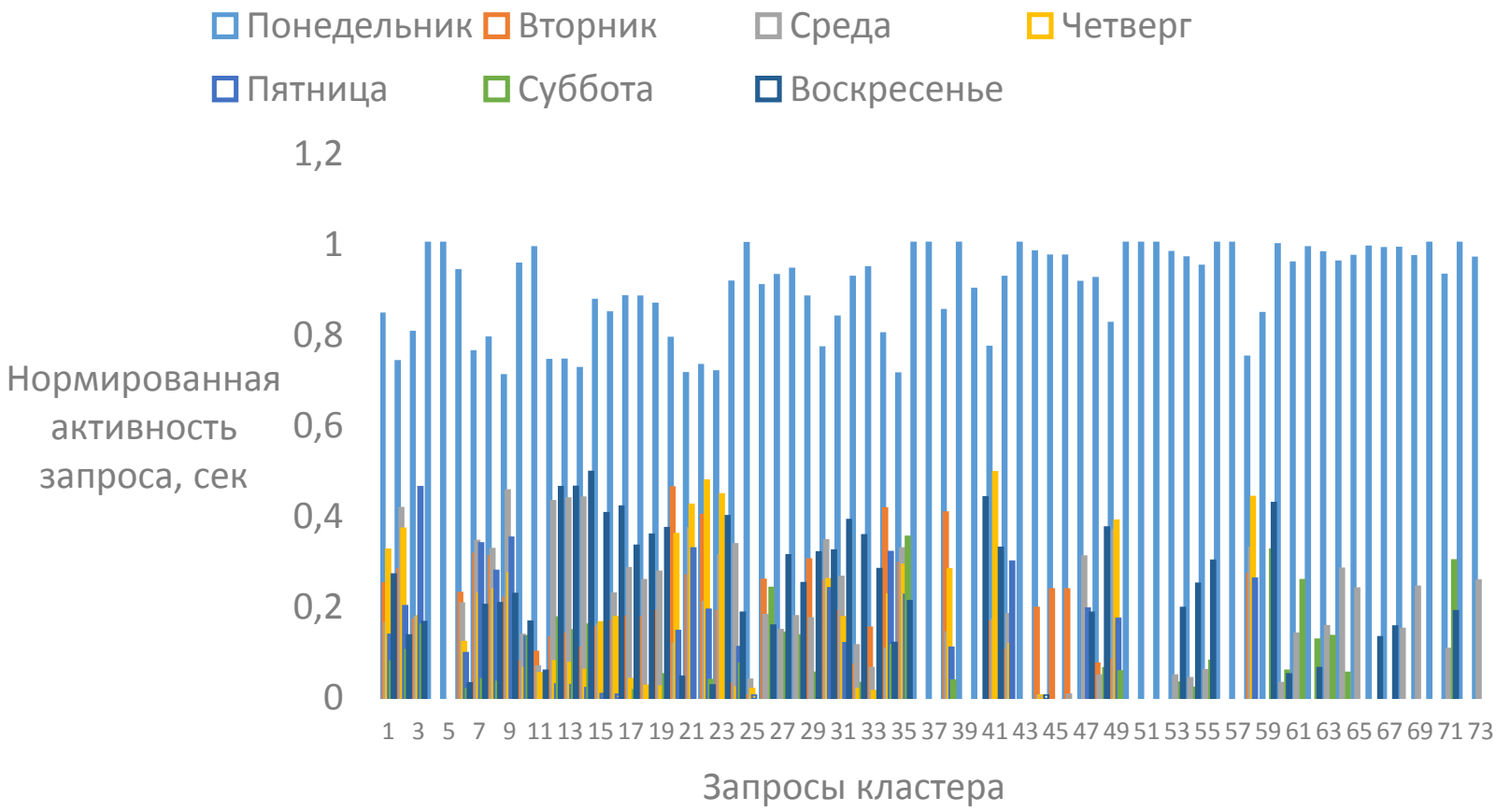
Кластеризация с использованием критерия минимальной энтропии

- Распределение запросов по кластерам



Кластеризация с использованием критерия минимальной энтропии

- Активность запросов попавших в кластер «Понедельник»



Анализ качества прогнозирования активностей

- Метрики качества предсказания

$$precision = \frac{a}{a + b}$$

$$recall = \frac{a}{a + c}$$

$a -$ релевантно предсказанные активности,
$b -$ ложно предсказанные активности
$c -$ активности, которые не были предсказаны

Улучшение качества прогнозирования

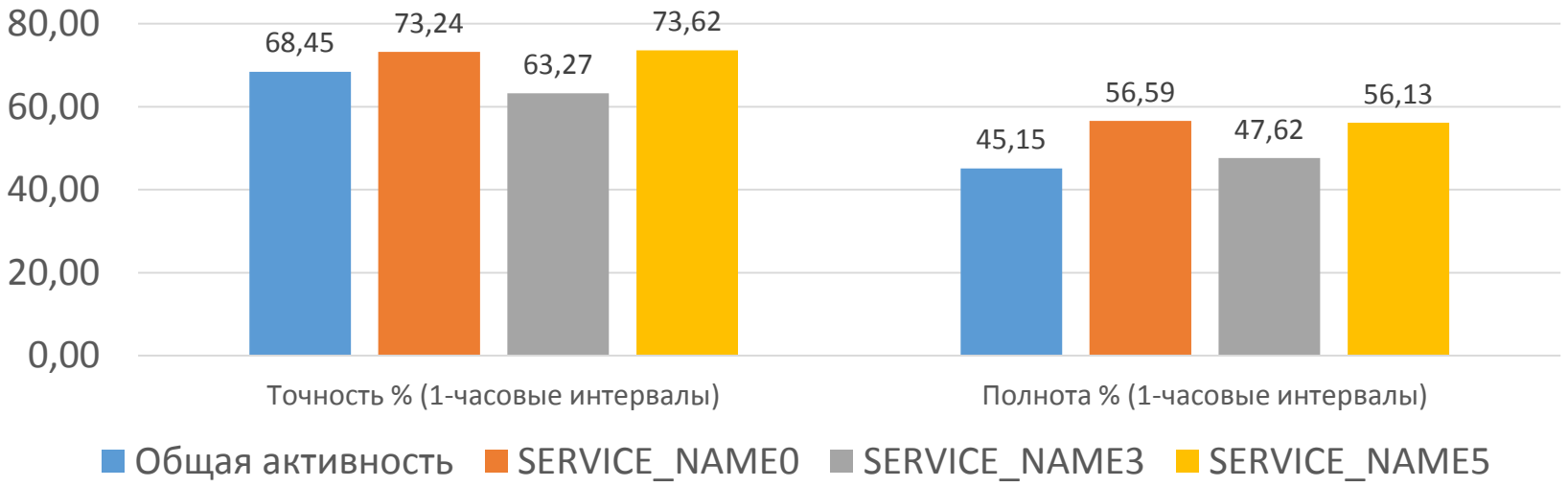
- Группировка активности объектов в рамках одной сессии соединения с базой данных

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"
…	…	…	…	…

Эксперименты

- Запуск алгоритма прогнозирования на различных данных



Результаты

- Реализован подход выделения фоновых активностей системы хранения, который отсеивает порядка 36% от общего числа запросов к системе
- Реализована процедура предварительного понижения размерности характеристических векторов запросов к системе хранения
- Проведены эксперименты по поиску групп взаимосвязанных запросов – бизнес активностей с альтернативными подходами
- Проведены эксперименты по изучению влияния атрибута идентификатора сервиса, с которого происходит обращение к объектам системы хранения, на качество прогнозирования активностей, по результатам которых удалось обнаружить улучшение точности предсказания в среднем на 1%, полноты предсказания в среднем на 4%