Анализа данных трафика научного учреждения с использованием вычислительных систем

Рыговский И.А.
СибГУТИ, Новосибирск

Подготовка данных

Интенсивное развитие компьютерных сетей в России в условиях объективно ограниченного финансирования делает особенно актуальной проблему анализа работы сетей и их перспектив для выработки оптимальной стратегии их развития.

Трафик (объем передаваемой и принимаемой информации в единицу времени) сети является одним из важнейших фактических показателей работы сети.

С использованием системы пост-анализа был произведен анализ сетевого трафика ИВМиМГ СО РАН. В распоряжении были данные трафика научного института в течении 4 лет.

Для работы с данными были сформированы 2 набора данных с различной детализацией для каждого пользователя: день , час, 5 минут.

1) Индикаторный ряд, где для каждого отрезка времени установлена 0 или 1, где 0 – пользователь не активен; 1 – пользователь активен.

2) Числовая последовательность, где трафик делится на количество уровней, заданных пользователей. Процесс напоминает аналого-цифровое преобразование. При этом при выборе максимального значения необходимо исключать «выбросы», удовлетворяющие неравенству Маркова. Таким образом, преобразуем значения с реальным количеством потребляемого трафика к числовым значениям или «уровням», зависимым исключительно от поведения конкретного пользователя.

Анализ данных

Имея в распоряжении два основных набора данных, мы можем начать извлечение знаний из этих массивов, меняя их формат и структуру для уточнения модели.

Были сформулированы ряд гипотез, которые в дальнейшем проверялись и уточнялись:

- Выявление и анализ паттернов поведения пользователей
  - Поиск нормы поведения пользователей, отклонения от нормы
  - Анализ на различных интервалах времени (часы, дни, недели, месяцы)
- Поиск кластеров пользователей
  - относительно активности пользователей (используя индикаторный ряд)
  - относительно количества потребляемого трафика (используя числовую последовательность)
  - для различных типов трафика
- Поиск сходств и различий между известными группами пользователей

- Физических (время, рабочий день и т.п.)
- Социальных (вид деятельности, лаборатория)

Проверка гипотез возможна при использовании, как различных стандартных статистических методов, путём анализа полученных наборов данных, так и методов интеллектуального анализа данных, к примеру, для решения задачи кластеризации.

Набор объектов с фиксированным количеством параметром для каждого объекта являются удобным представлением данных для использования методов кластеризации. В качестве объектов могут использоваться не только пользователи, но и сами промежутки времени (время суток, день недели, сезон и т.п.), которые естественно отличаются между собой. Но изучая эти отличия можно выявить не только уже известные знания, но и найти новые, которые зависят от поведения пользователей, именно они могут помочь лучше понять пользователей.

Кластерный анализ данных, как правило, включается в себя несколько этапов, это: предварительный анализ, позволяющий выбрать нужные методы анализа; применение различных методов анализа для проверки, сформулированных ранее гипотез; анализ полученных результатов.

Наиболее эффективный метод предварительного анализа данных является визуальное представление данных, с помощью которого проводится первичная оценка структуры данных, из которой можно сделать первые выводы и определить дальнейшую стратегию анализа.

Использование метода многомерного шкалирования позволяет привести многомерные данные к двумерному виду, который легко можно представить графически.

На рисунке 1 можно увидеть, как в пространстве распределилось множество дней в течение 4 лет, параметрами у которых являлись сами пользователи, их активность в течение дня. Видно, что дни распределись на несколько ярко выделенных групп, распределённые по плоскости.

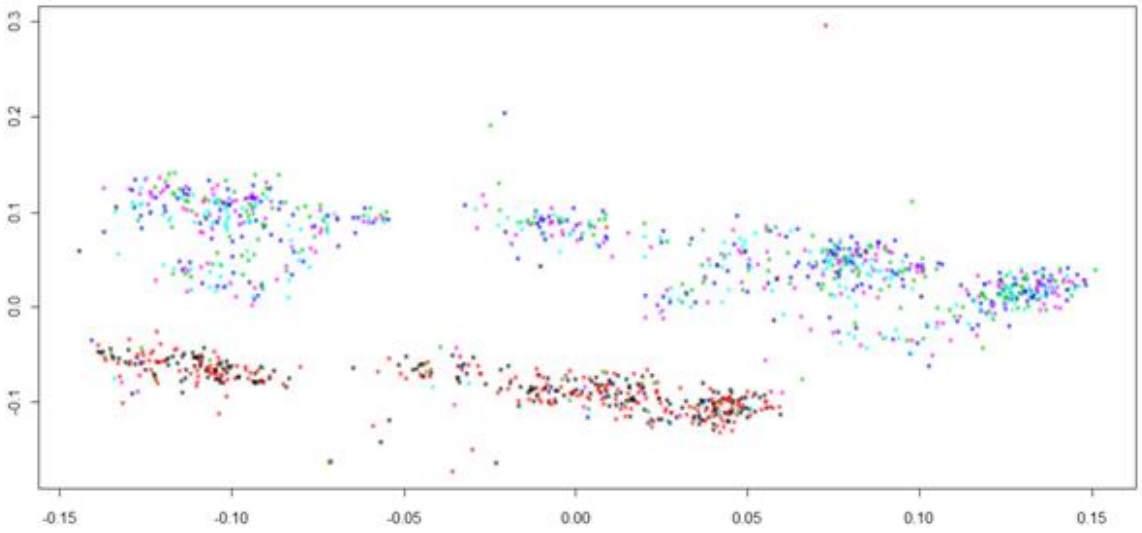

Рисунок 1. – Визуальное представление активности пользователей в течение 4-х лет по дням. Весь трафик, по дням недели.

Цветами на рисунке 1 отмечены дни недели, чёрный и красный цвет – выходные дни. Это позволяет проверить гипотезу о зависимости активности пользователей о днях недели, а также визуально оценить потребление трафика в различные дни недели.

Метод многомерного шкалирования выполняет ещё одну важную функцию – это выбор нужной метрики для дальнейшего применения методов кластеризации. К примеру, при разбивании дней на классы легко применить метод k-средних (k-means), где для расчёта расстояния между объектами используется Евклидово расстояние. А для разделения на классы часов, достаточно провести прямую линию и разделить отклонение от нормы от основной группы линейным уравнением. Второй вариант – использовать для вычисления принадлежности к кластеру объекта при помощи метрики расстояния ближайшего соседа или любой другой, не зависимой от геометрических размеров кластера. На рисунке 2 изображено распределение пользователей, параметрами которых являются уже сами промежутки времени.

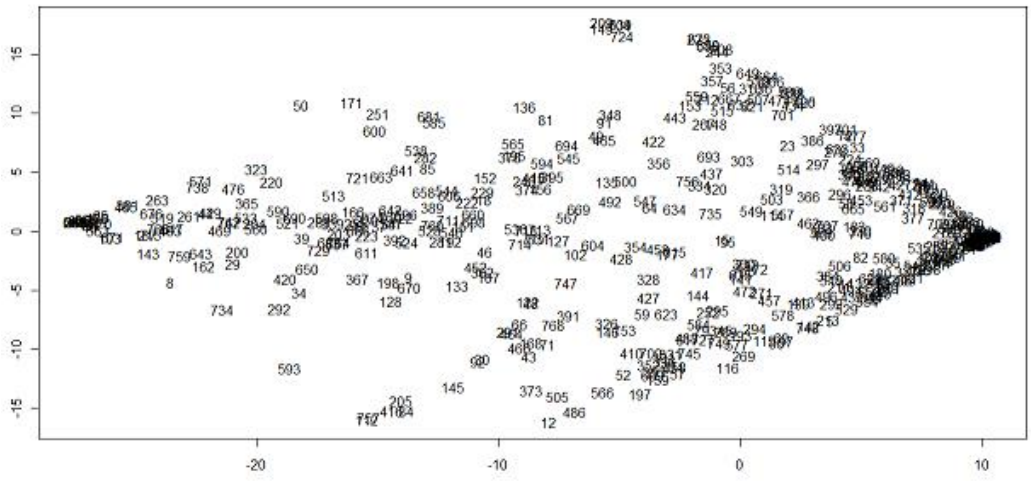

Рисунок 2. – Визуальное представление активности пользователей в течение 4 лет по дням. TCP трафик

Структура данных, в данном случае, не так очевидна, и есть смысл применять методы иерархической кластеризации для полного отображения всех возможных вариантов кластеризации, используя различные метрики для выявления кластеров и проверки поставленных гипотез.

Иерархическая кластеризация позволяет увидеть принадлежность пользователя не к одному, заранее заданному кластеру (в случаи k-means), а к нескольким одновременно. Это позволяет сделать гораздо более точные выводы о сходстве найденных кластеров, известных групп пользователей, найти новые факторы, которые могут влиять на принадлежность пользователя к определённому классу.

В данном случае, для нахождения расстояния между кластерами и объектом, использовался метод Уорда, как учитывающий роль каждого элемента кластера. А использование в качестве параметров часы или дни позволяет классифицировать пользователей по разным качественным признаком, в зависимости от поставленной задачи. Это может быть поиск классов по поведению в течении дня или в течении более продолжительного времени, например, в течении квартала или учебного года.

В дальнейшем полученную информацию можно сравнивать с уже известными группами, такими как: профессия, возрастная группа, различные социальные группы и т.п.

Анализ числовой последовательности, полученной при разделении трафика на уровни, по количеству скачанной информации, позволяет оценить схожесть в поведении пользователей, учитывая не просто активность пользователей, но и количество трафика. При этом необходимо подготовить данные, чтобы очистить их от физических, поведенческих факторов, оставив только необходимые для выявления схожести в поведении. К таким фактором можно отнести скорость канала, используемое программное обеспечение или предпочтения пользователя, которые не играют роли при выявлении его поведения, например, промежутки между активностью или её продолжительность. Для различных задачи и эффективности работы алгоритма можно выбирать разное количество уровней, на которые разбивается трафик.

Подсчитав количество промежутков с разным уровнем трафика, можно получить набор данных, где у каждого пользователями параметрами становятся отношение трафика одного уровня к общему количеству. Эти данные подойдут для кластеризации и классификации пользователей.

Кластеризацию можно проводить различными методами анализа данных, в том числе метод многомерного шкалирования, предварительно исключив из выборки трафик с минимальным уровнем, который незначительно повлияет на различие между пользователями.

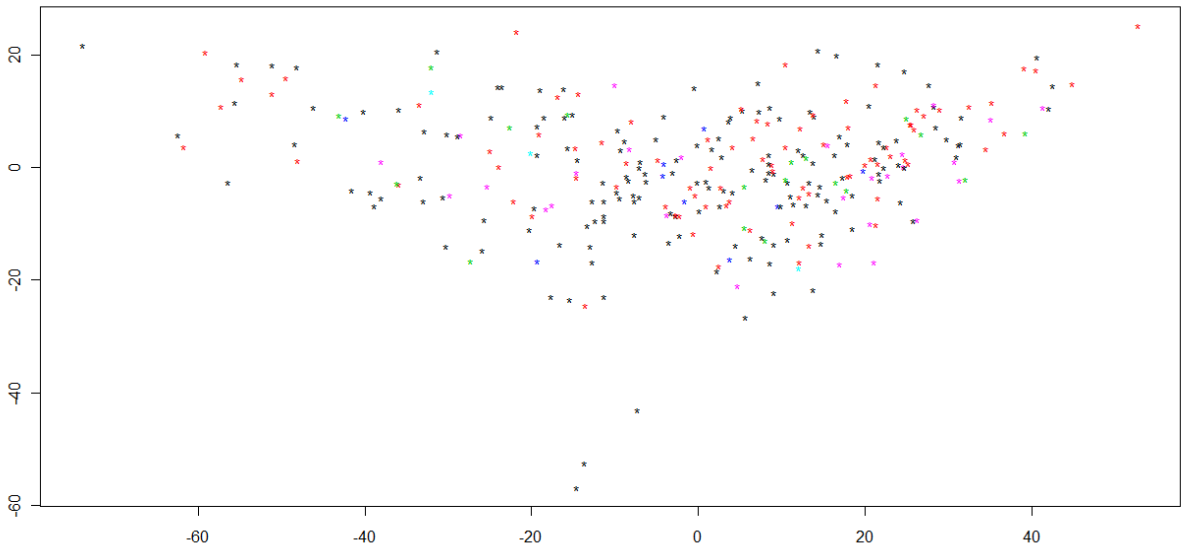

Рисунок 4. - Распределение пользователей на плоскости, относительно количества потребляемого трафика.

На рисунке 4 можно увидеть результат работы. Точками различного цвета окрашено различие между реальным потреблением трафика, а расстояние между точками – сходство в поведении пользователей. Таким образом видно, что пользователи с различным реальным потреблением трафика в действительности схожи по поведению. Эти данные могут дать значительную информацию для анализа поведения групп людей в сети, оценить поведение пользователей с различных углов зрения, учитывая особенности анализируемой сети и конкретных её пользователей, применяя индивидуальный подход.

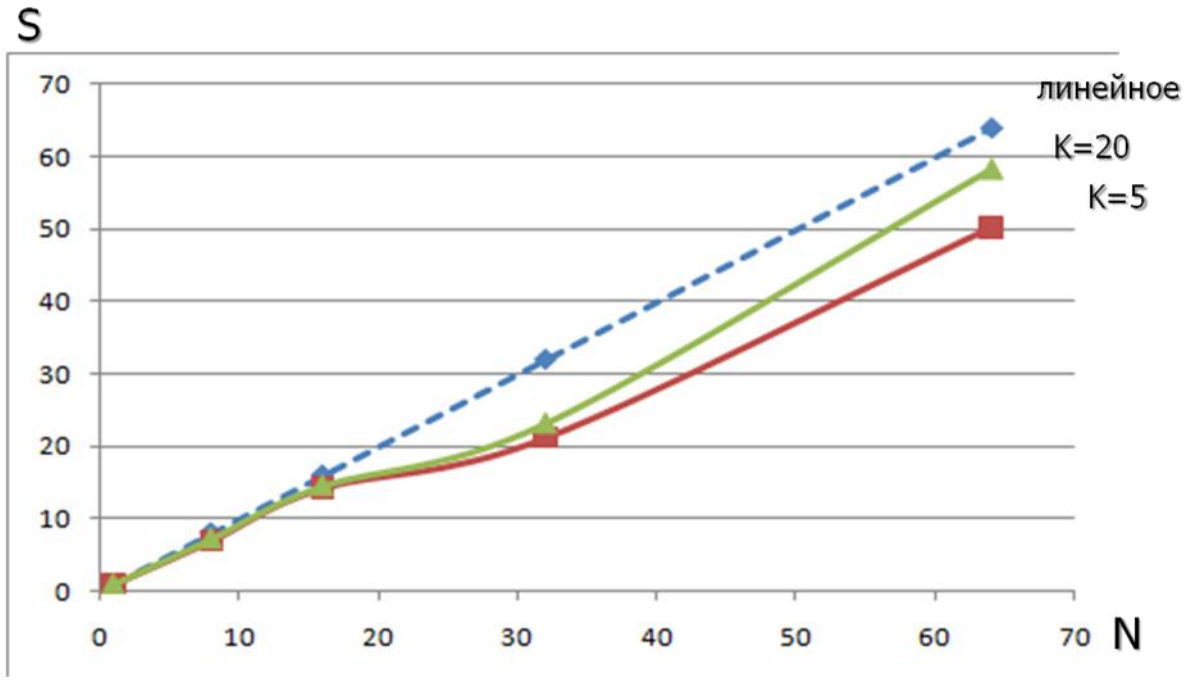

Рисунок 5.- Ускорение работы метода k-mean.

Для более детального анализа данных есть смысл использовать вычислительные системы и параллельные алгоритмы Data Mining. Были исследованы и реализованы параллельные методы, используемые выше. К примеру на Рисунке 5 видно, что для различного количества кластеров отличается ускорение. Метод k-means является достаточно простым для распараллеливания, так как практически отсутствует передача данных между узлами ВС. Реализация иерархических алгоритмов кластеризации значительно сложнее, из за необходимости на каждой итерации каждого узла иметь доступ к данным других кластеров. Для различных метрик следует использовать специальные типы данных для хранения рассчитанных расстояний между кластерами, к примеру наименьший объём передаваемой информации потребуется при использовании метрики «ближайшего соседа», а максимальный (на каждой итерации), при использовании методов Уорда.

При исследовании использовались ресурсы Центра параллельных вычислительных технологий ГОУ ВПО “СибГУТИ”, а также ресурсы Информационно-вычислительного центра НГУ.