Семинар Московской секции ACM SIGMOD, 24 мая 2007 года.

Анализ поведения пользователей Интернет: возможность автоматизации

Докладчик: к.ф.-м.н. Щербина А.А.

Структура доклада

- Методы исследования пользователей Интернет
- Системы data mining
- Возможности автоматизации
- Предлагаемая методика

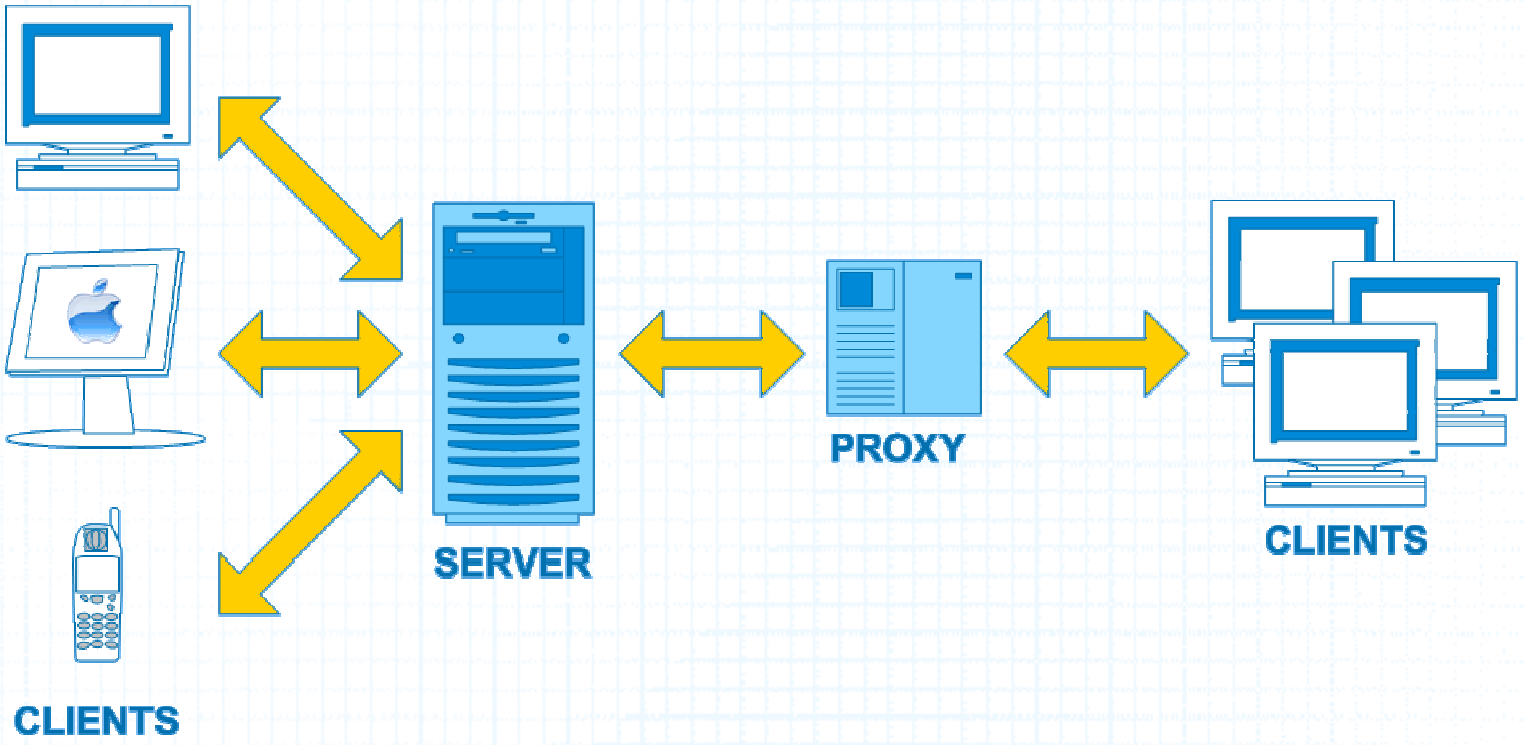
Статистические исследования

- Марковская цепь Vi = Vi-1 + ε
- P(L) = L^-3/2 (Распределение Гауса)
- Среднее количество страниц, посещаемых за одну сессию, составляет 3,86

Предпосылки для улучшения

- Рост значения Интернет
- Значительный рост объёма хранимых исторических данных по поведению пользователей
- Рост потребности в анализе пользователей и их предпочтений.
- Отсутствие возможностей для увеличения числа экспертов

Сбор информации



Существующие методы

- Статистический анализ
- Визуализация данных
- Поиск ассоциативных правил
- Существующие методы требуют участия эксперта на различных этапах обработки данных или анализа.
- Кластеризация
  - Существующие методы кластеризации пользовательских сессий не учитывают содержание посещенных страниц;
  - Для проверки результатов кластер-анализа используются эталонные разбиения или статистические параметры.

Ассоциативные правила

- Правила вида:
- $A => p_k$. Где $A = \{p_1,..,p_j\}$
- Поддержка – отношение числа сессий где выполняется А к общему числу сессий.
- Достоверность – отношение числа сессий, где выполняется правило к сессиям где выполняется А.

Применение кластеризации

- Уменьшение размерности (выбор представителей)
- Социологическое исследование
- Модификация сайтов
- Персонификация наполнения

Цель работы

Создание автономного программного средства, выполняющего классификацию поведения пользователей произвольного Интернет-сайта.

Поставленные задачи

- Разработка метода кластеризации пользовательских сессий.
- Автономное определение оптимального разбиения по кластерам. Все полученные кластеры должны соответствовать типам поведения пользователей.
- Создание системы классификации, включающей все этапы анализа.

Предлагаемый метод

Сбор данных на стороне сервера

- Существуют заранее накопленные данные
- Структура журналов отвечает общему стандарту CLF.
- Основной недостаток: Не всех пользователей можно идентифицировать.

Нечёткий к-центроидов метод

- $Jm(V;X) = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m} r(x_j, v_i)$

  Минимизируется функция Jm, где u_{ij} – характеристическая функция, v_i – центр i-го кластера.
- Только 30 элементов с наибольшей вероятностью используются для пересчёта центров.

Метод Dbscan

- К каждому элементу присоединяются соседние. Кластеры – набор пересекающихся ε окрестностей.
- Основной параметр – ε.

Расстояние редактирования

Представление сессии: $v_i = \{x_{i1}, x_{i2}, x_{i3}..x_{in}\}$, где x_{ij} – страница сайта.

Операции замены символа, вставки и удаления. Веса 2, 1 и 1 соотвественно.

Примеры строк: ‘cat’, ‘cash’

CAT -> CAS -> CASH

Общее расстояние 3.

Предлагаемое расстояние

- dir11/dir12/pagename1
- dir21/dir22/pagename2
  Если совпадают dir 11 и dir 21 то уменьшается стоимость замены

  Если совпадают dir 21 и dir 22 то стоимость снижается еще больше

Теорема 1. Предложенное расстояние является метрикой

Модификации расстояний

- Уменьшение расстояния редактирования на -10% за каждый совпавший символ.
- Все строки дополняются до одинаковой длины.
- Страницы заменяются на каталоги второго уровня.

Расстояние Манхэттена

- Каждая сессия это вектор
  $v_i = \{x_1, .. x_N\}$
- $x_j = 1$ если страница j входит в сессию.
- $x_j = 0$ иначе.

$$r(v_1, v_2) = \sum_{i=1}^{N} |x_{1i} - x_{2i}|$$

Методы верификации

- Индекс Данна MD(c)=min(r(j,i))/max(Dk) $\forall$i,j,k: i≠j, где D_k – диаметр кластера, r(j,i) – расстояние между кластерами i и j, c – количество кластеров
- Индекс разделения IR(c)= max((Dj + Di)/r(j,i)) $\forall$i,j: i≠j., где D_j – диаметр кластера, r(j,i) – расстояние между кластерами i и j, c – количество кластеров
- Энтропия разбиения. $PE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}\log(u_{ij})$
- Коэффициент разбиения (индекс Беждека).

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^2$$

где u_{ij} – характеристическая функция

Предлагаемая верификация

- Подсчёт уникальных ассоциативных правил
- M (P)= $\sum_{i=1}^{C} M_i / N_c$, где N_i количество элементов в i-ом кластере, M_i – количество уникальных правил в i-ом кластере, C – количество кластеров.

Реализация системы

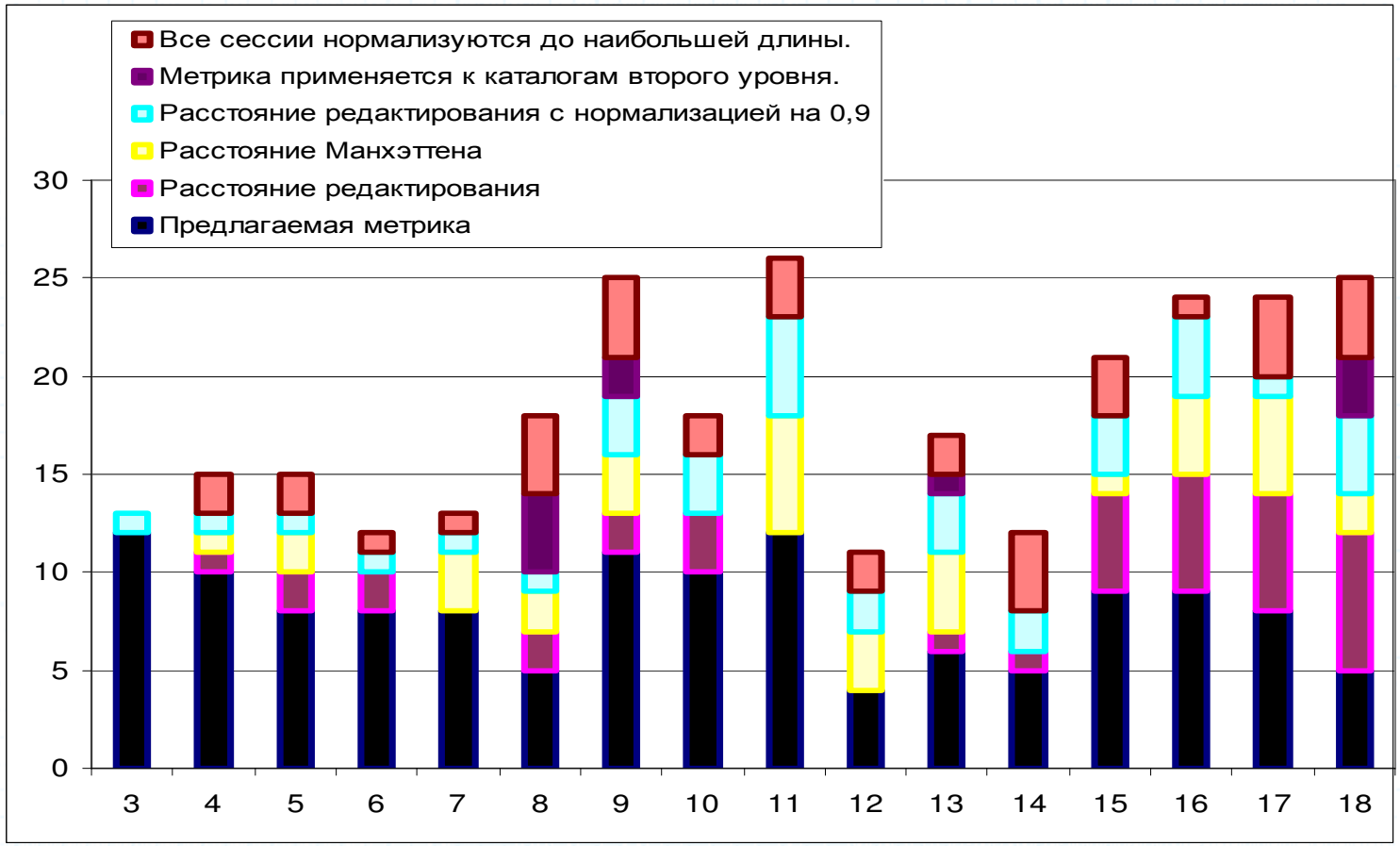
- Сессии представлены как численные векторы
- Расстояние редактирования рассчитывается методом динамического программирования
- Нечёткий C-Medoids метод
- Подсчет уникальных ассоциативных правил
- Хранение представителей в оперативной памяти
- Достигнутая эффективность: 15 минут на обработку данных за 1 день.

Экспериментальные результаты

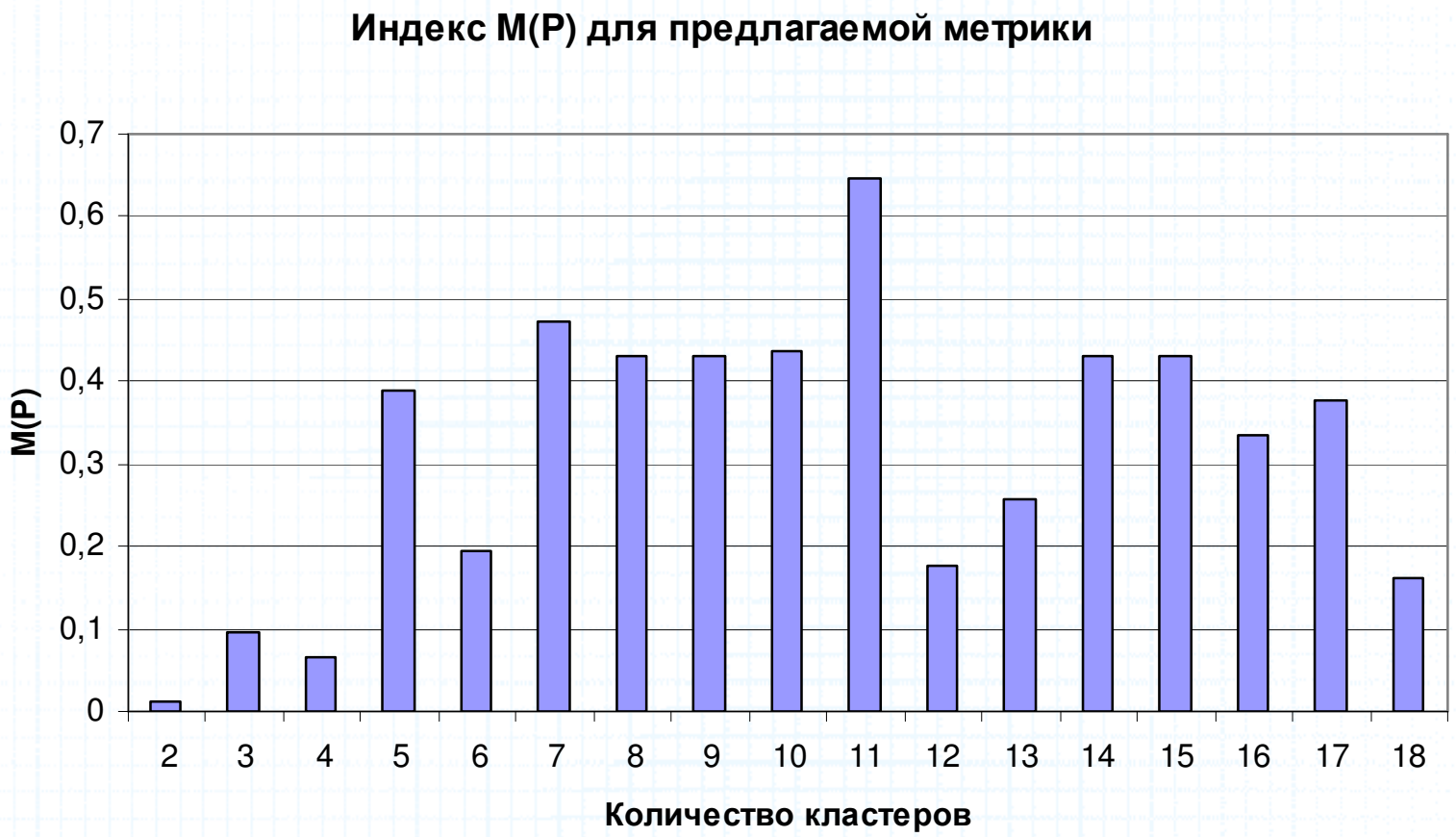
Данные Citforum.ru

- 70000 посещений в день
- 1300 сессий в день
- 10000 страниц
- Данные за три дня
- Сессии длительностью от 3 до 40 визитов

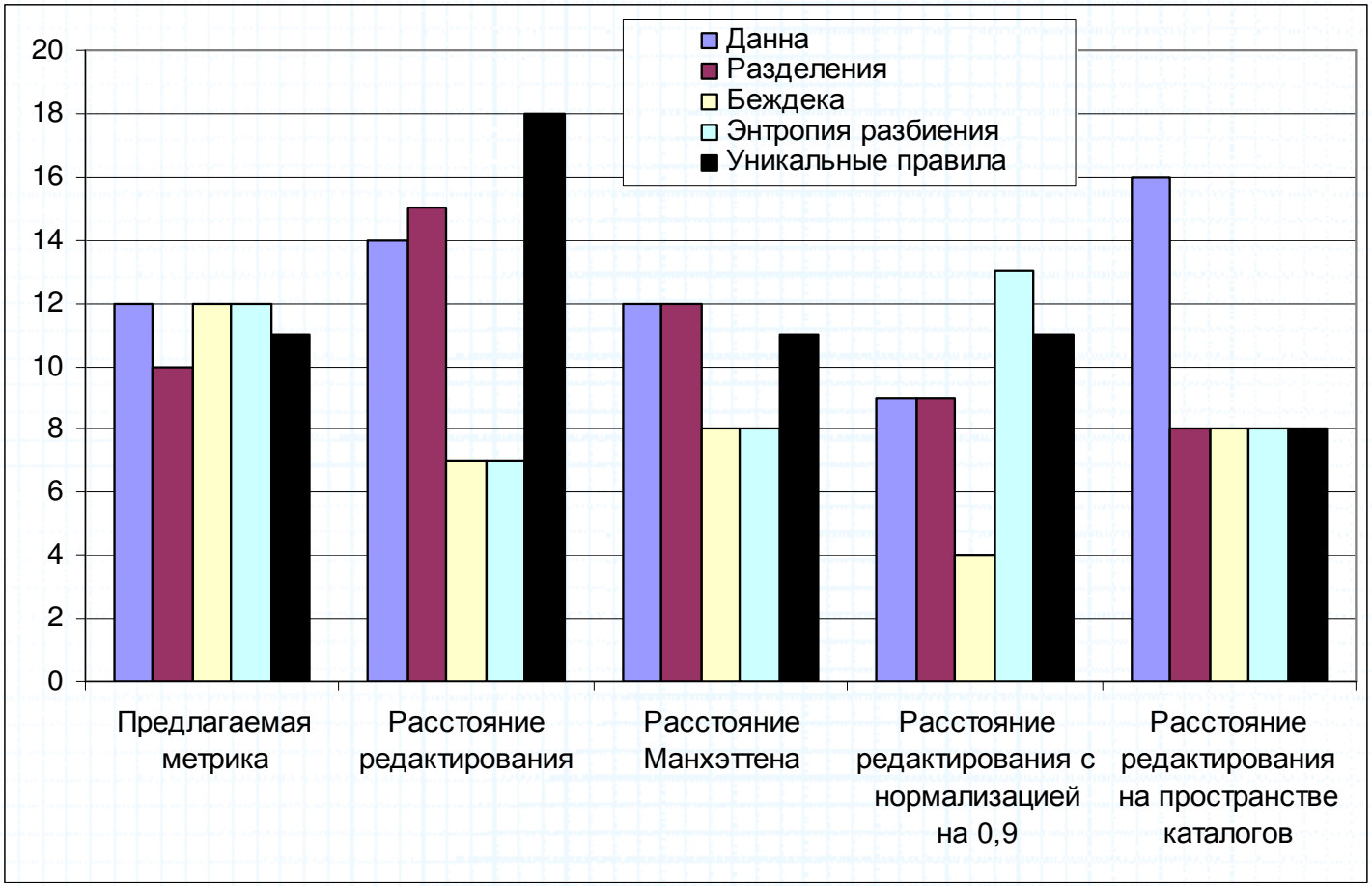
Уникальные правила



Индекс различимости



Сравнение индексов



Результаты работы

1. Разработан и апробирован метод сравнения сессий пользовательского доступа к веб-сайту. Основные достоинства предложенного метода состоят в следующем:
   - предложенное расстояние обеспечивает корректное сравнение сессий пользовательского доступа за счёт использования данных о последовательности посещения страниц и их размещения на сайте;
   - применение метода не требует предварительной индексации или обработки самих страниц сайта.
2. Разработан и апробирован метод автоматической верификации результатов кластеризации пользовательских сессий. Основные особенности предложенного метода верификации результатов кластеризации:
   - метод определяет качество разбиения сессий по кластерам в зависимости от наличия уникальных ассоциативных правил;
   - метод выявляет семантические отношения, характерные для каждого из кластеров.
3. Создана программная система, обеспечивающая автономную классификацию сессий пользовательского доступа на основании журнала веб-сайта.