Технология извлечения знаний из использования Интернет

Технология извлечения знаний из использования Интернет

Определение

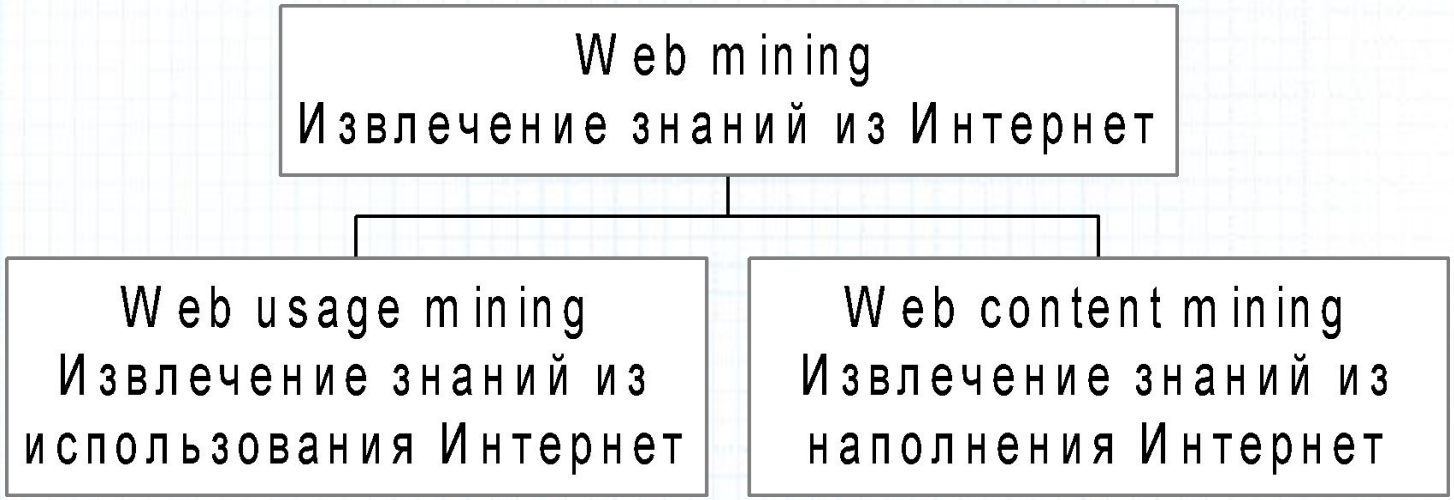
Извлечение знаний – поиск нетривиальных потенциально полезных знаний в больших объёмах данных.

Основные области применения

- Финансы
- Страхование
- Медицина
- Биология
- Интернет

OLAP/Data mining

Структура web mining



Структура web content mining

Web usage mining

Извлечение знаний из использования Интернет – поиск нетривиальных потенциально полезных знаний в деятельности пользователей Интернет.

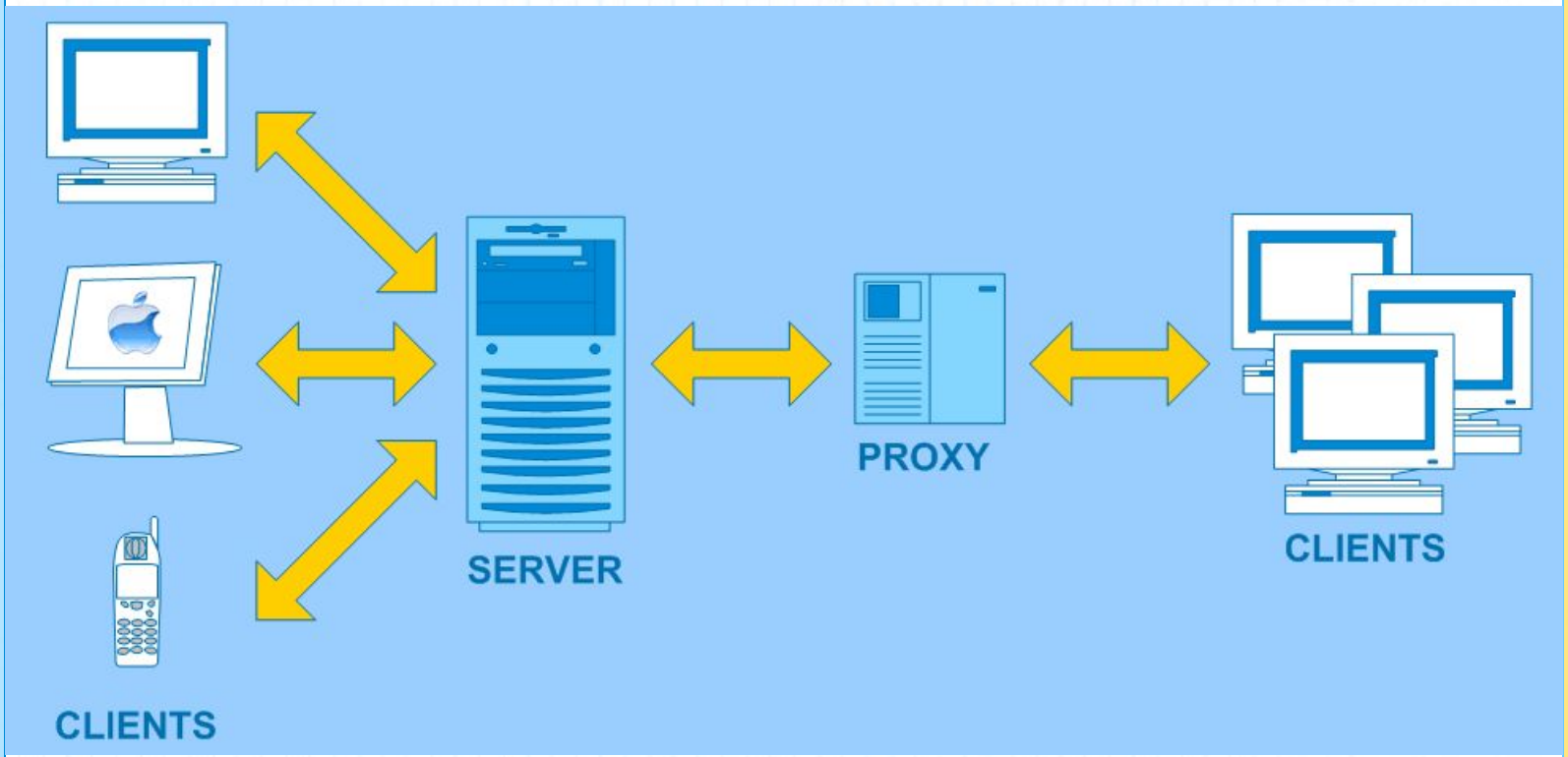
Применения Web usage mining

- Персонификация контента
- Улучшение работы сети
- Модификация сайтов
- Исследования сети

Этапы Web usage mining

- Сбор данных
- Обработка данных
- Применение методов Data mining
  - Кластеризация
  - Поиск ассоциативных правил
  - Поиск наиболее частых подпоследовательностей

Сбор информации



Обработка данных

- Очистка данных
- Заполнение пути
- Выделение пользовательских сессий

Ассоциативные правила

- Правила вида:
- A=>b. Где A - ДНФ
- Поддержка – отношение тех элементов где A к общему числу
- Уверенность – отношение элементов, где выполняется правило к элементам с A

Цель кластеризации

- Уменьшение размерности (выбор представителей)
- Генерация гипотез
- Проверка гипотез
- Прогнозные модели

Методы кластеризации

- Иерархические
- Алгоритмы оптимизации
- Основанные на плотности
- Нечёткие методы

Иерархические методы

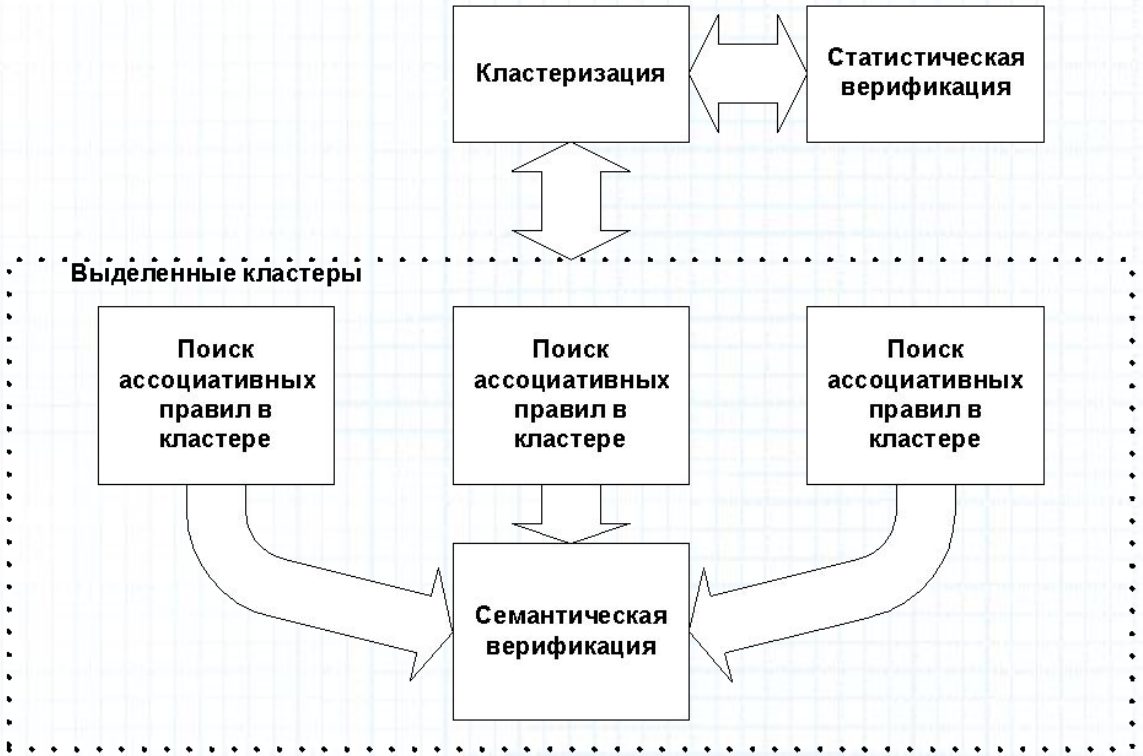
- N кластеров
- На каждом шаге объединение двух самых «близких» кластеров
- Расстояние: по наиболее близкими или наиболее удалённым точкам, по центрам.

Нечёткий c-medoids метод

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

  Минимизируется это значение
- Только 30 элементов с наибольшей вероятностью используются для пересчёта центров.

Верификация кластеризации



Методы верификации

- Сопоставление эталонного разбиения и кластеров
- Статистические
- Связанные с нечётким разбиением
- Комбинированные методы

Предлагаемый метод

- Сессии представлены как численные векторы
- Используется расстояние редактирования
- Расстояние модифицируется с учётом положения страниц
- Нечёткий C-Medoids метод

Данные Sigla.ru

- 70000 посещений в день
- 1300 сессий в день
- 50 страниц
- Данные за три дня
- Сессии с длинной от 3 до 40 визитов

Расстояние Евклида

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

Расстояние редактирования

Примеры строк: ‘cat’, ‘cash’

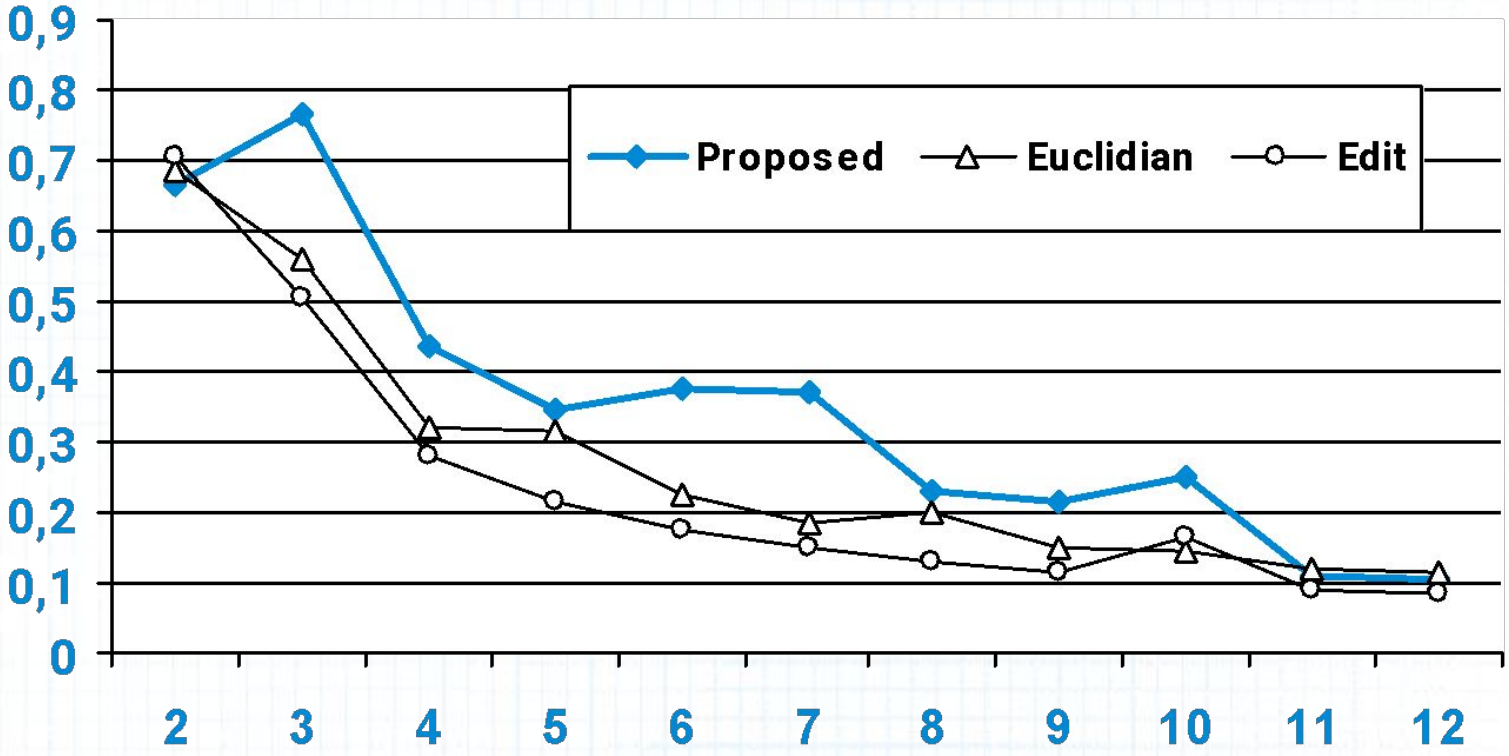
CAT -> CAS -> CASH

Общее расстояние 3.

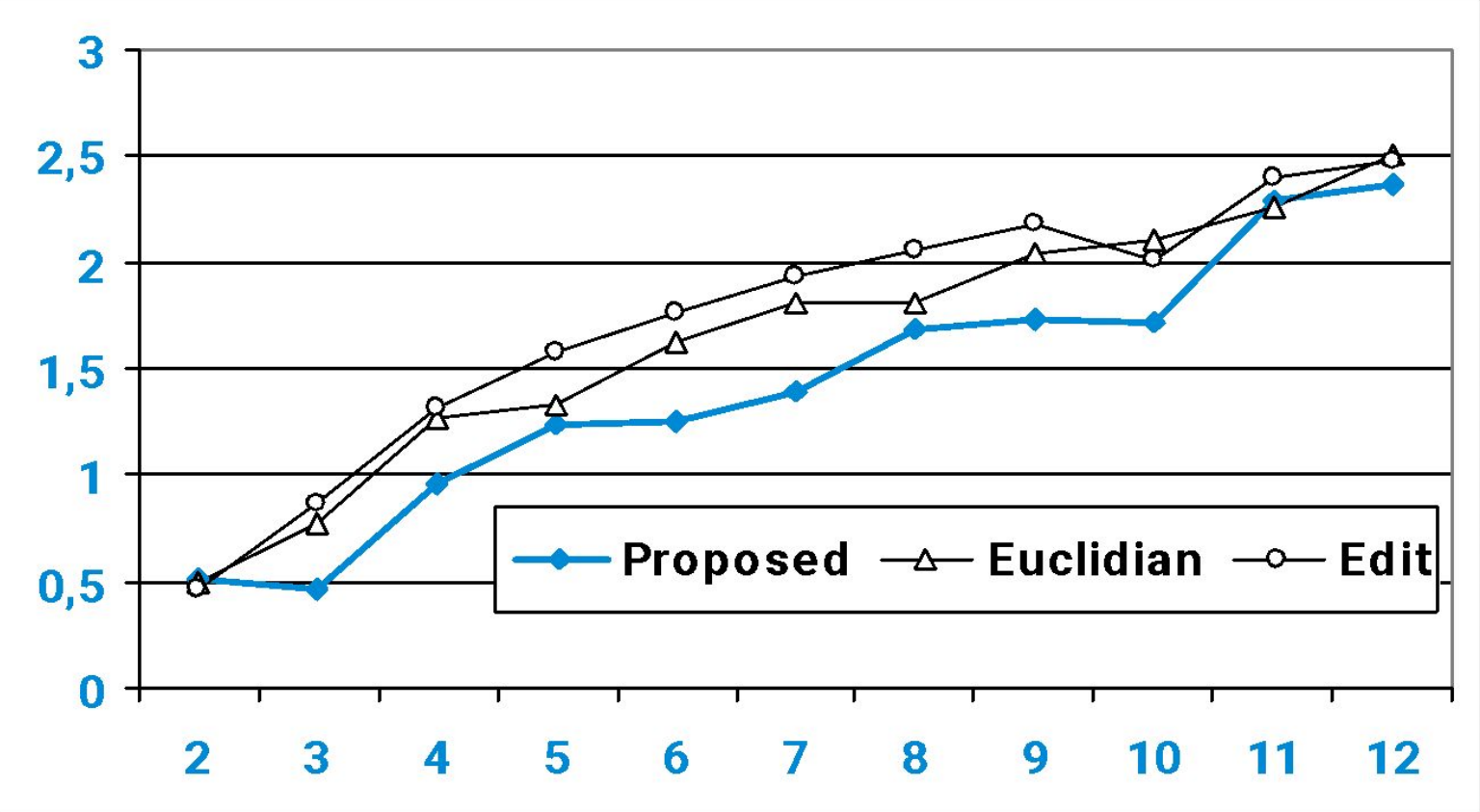
Модификация расстояния

- 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

Индекс Беждека



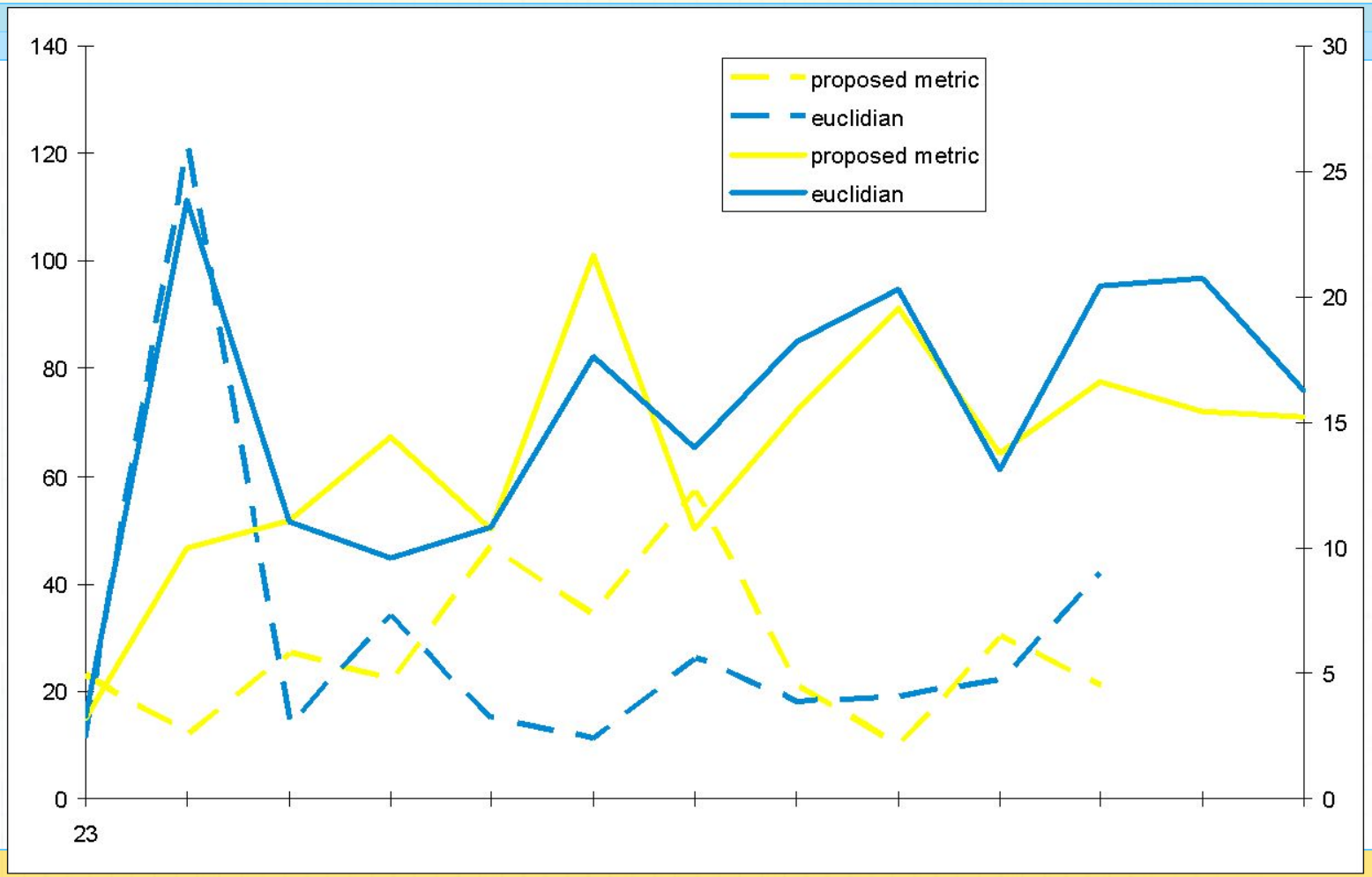
Энтропия разбиения



Предлагаемая верификация

- Подсчёт уникальных ассоциативных правил
- Индекс = количество уникальных правил/количество кластеров

Предлагаемый метод



Спасибо!

Ваши вопросы?..