RCDL'13, 14-17 октября, 2013

Методика учета интересов пользователя при работе в сети Internet на основе его профиля и ассоциативных связей

М.М.Шарнин,[1] А.В.Петров,[2] И.П.Кузнецов[1]

[1]Институт Проблем Информатики РАН, Москва

[2]Tinkoff Digital, Москва

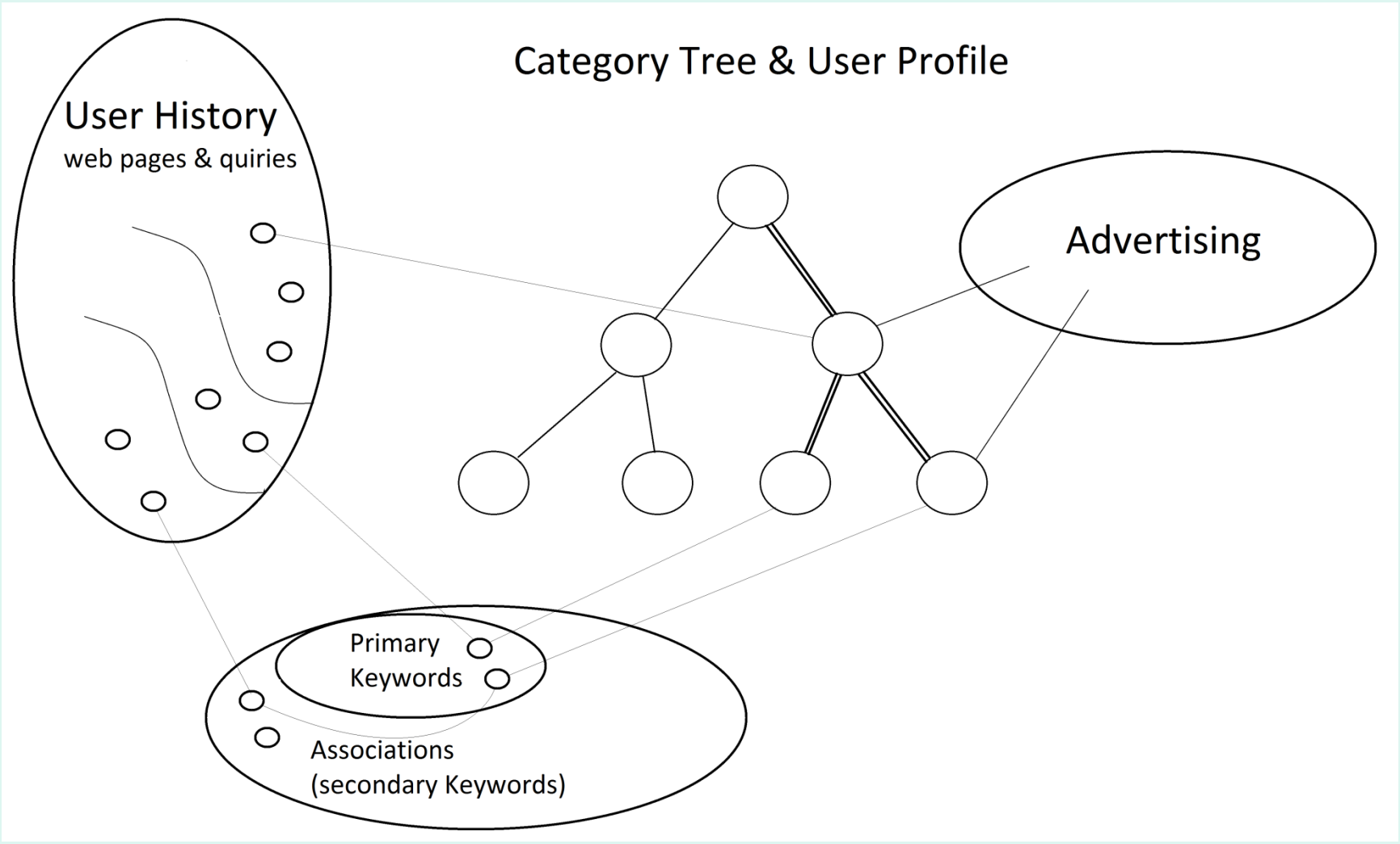
План Доклада

- **Представление интересов пользователей**
- **Первичные и вторичные ключевые термины (ассоциации) дерева категорий**
- **Оптимальные веса ключевых терминов для автоматической классификации**
- **Расчет весов первичных и вторичных ключевых терминов (ассоциаций)**
- **Определение интересов пользователя по истории истории его посещений и запросов**
- **Заключение**

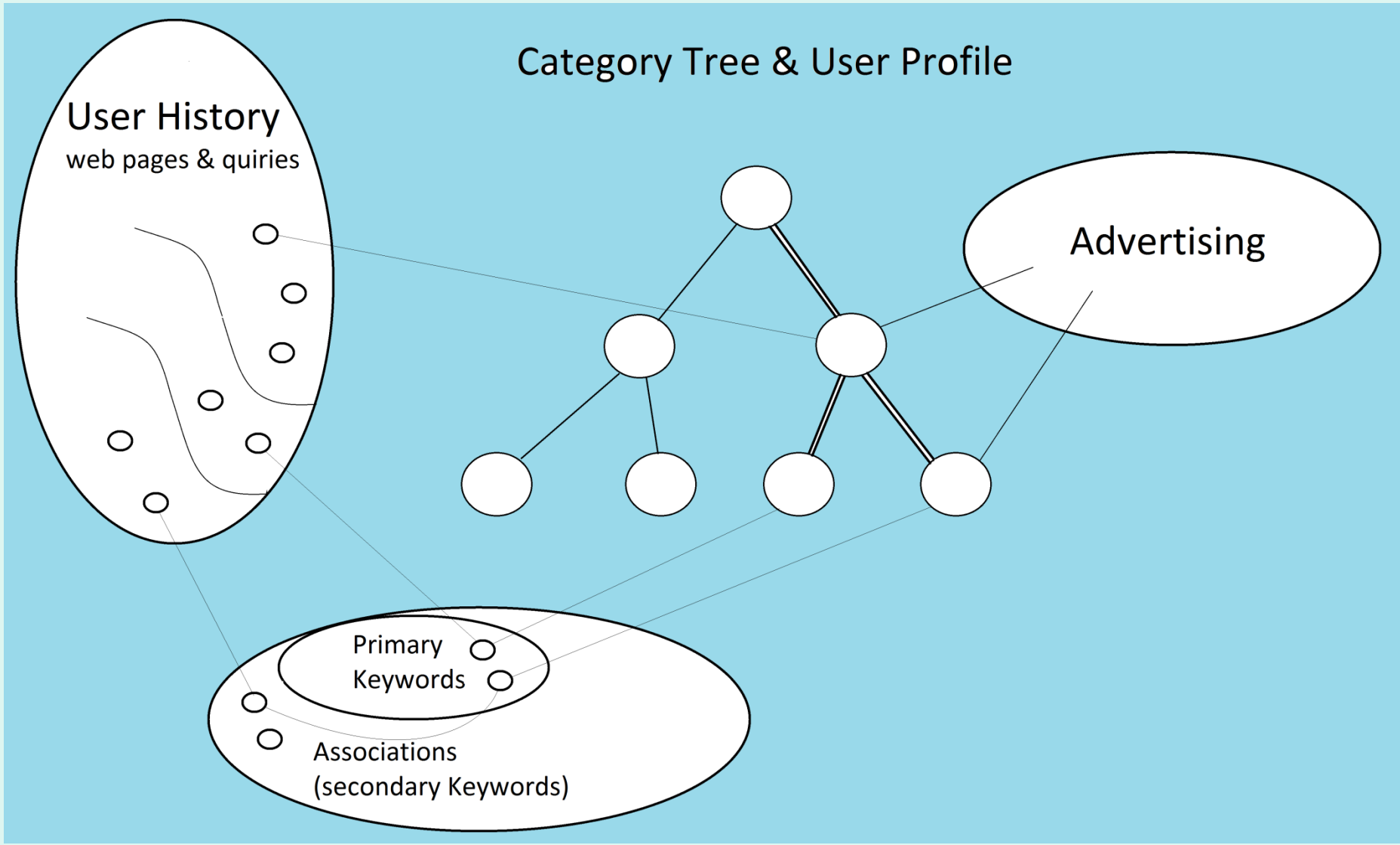
Представление интересов пользователя

- Интересы пользователей могут быть представлены как набор пар <ключевой термин, вес>
- Покупательница может иметь следующие интересы: <платье, 70> <косметика, 40> <сумочка, 20>
- Интересы спортивного болельщика: <спорт, 30> <футбол, 60> <воллейбол, 20>

Структура Данных



Структура Данных

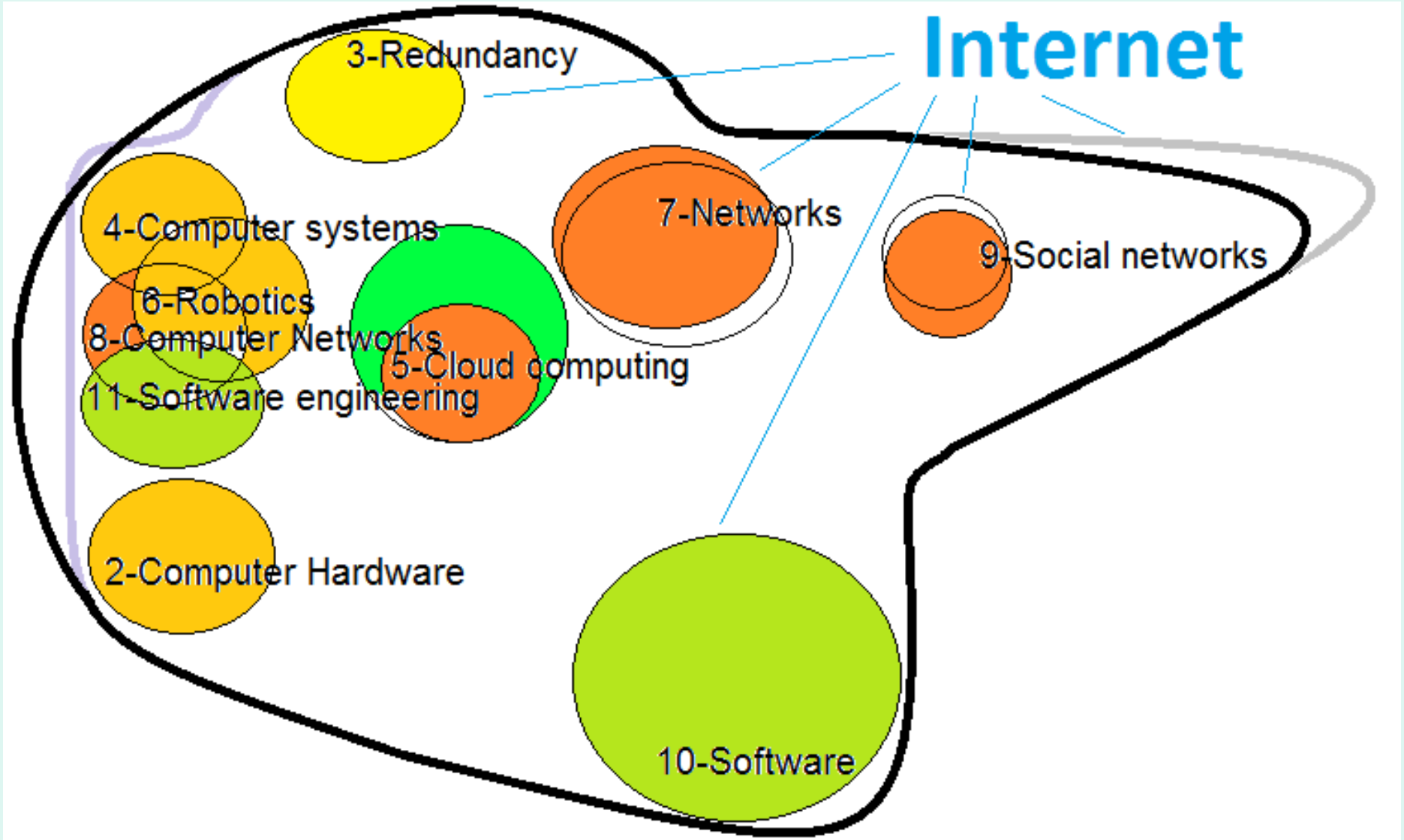


Мера семантической близости

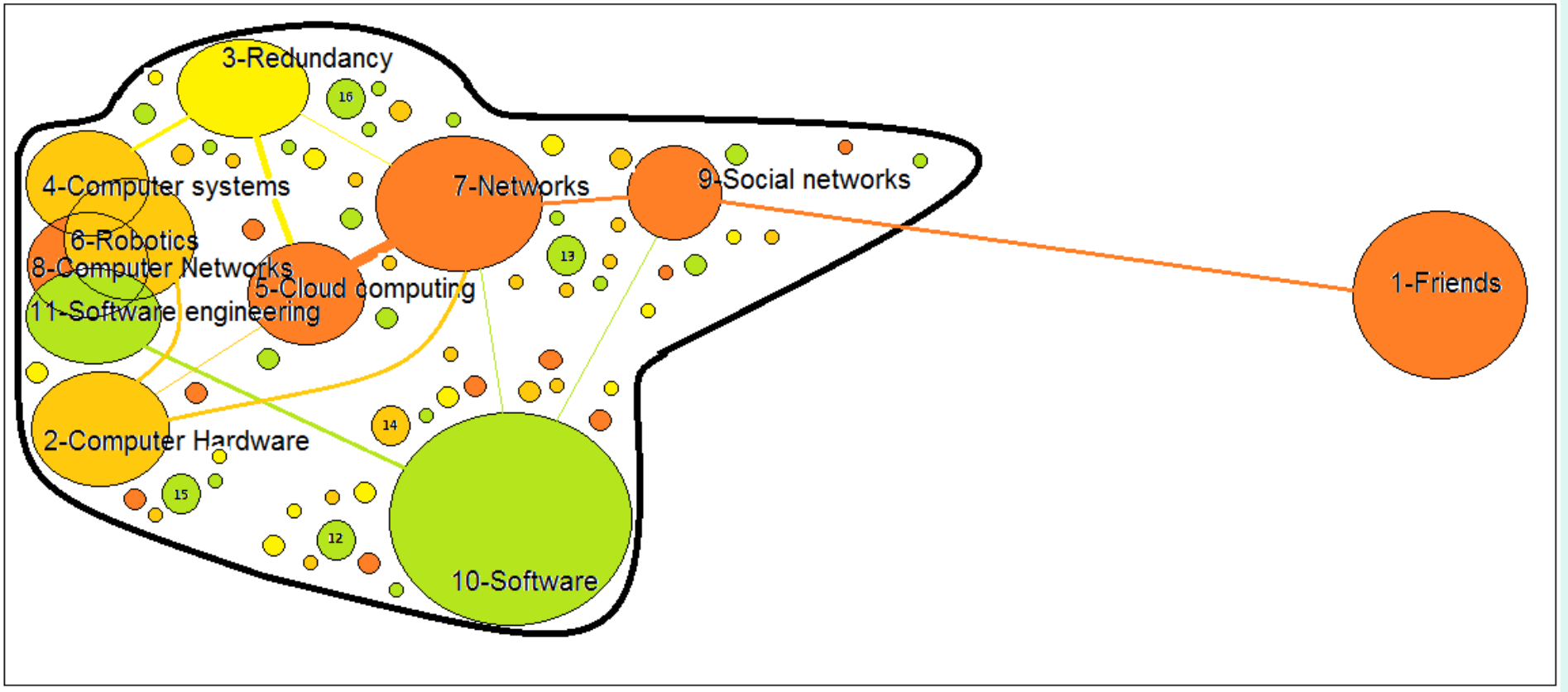
- Модели дистрибутивной семантики
- Статистический анализ Интернет текстов
- Специальные лексико-синтаксические шаблоны (x “- это” y | y “включая” x | x “такой как” y)
- Контекстные вектора терминов
- Косинусная мера семантической близости

$$\frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Визуализация Семантической Близости



Визуализация Ассоциаций



Первичные и вторичные термины дерева категорий

- Интересы пользователей классифицируется по дереву категорий Интернет-директории
- Каждая категория связана с набором первичных и вторичных ключевых терминов через тройки: <термин, категория, вес>
- **Первичные ключевые термины** определяются по обучающей выборке категоризованных документов
- **Вторичные ключевые термины** are определяются методами Дистрибутивной Семантики по Интернет текстам

Оптимальные веса ключевых терминов для классификации

- Each category is linked to a set of keywords by triples: <keyword, category, **weight**>
- The probability of correct classification:
  Pi = P(Category | Keyword)
- Optimal **weight = Log(Pi / 1 - Pi),** for two alternative categories (Nitzan and Paroush, 1982)
- For 3 or more categories:

$$\textbf{weight} = Log\left(\frac{P(keyword\ \&\ category)}{P(keyword)\,P(category)}\right)$$

Расчет весов первичных ключевых терминов

- Использовались Интернет директории Yandex, Google, Yahoo и Keywen, содержащие ссылки на категоризованные веб-сайты (обучающая выборка)
- Расчет вероятностей вхождения ключевых терминов в контекст категории
- Расчет весов первичных ключевых терминов категории (по вышеописанной формуле)

Расчет весов вторичных ключевых терминов

- Поиск Интернет текстов из заданной предметной области (категории), содержащих имя категории
- Определение и ранжирование ключевых терминов, специфичных для предметной области
- Использование методов дистрибутивной семантики для определения и ранжирования ассоциативных связей между ключевыми терминами

Выравнивание весов первичных и вторичных ключевых терминов

- Веса первичных и вторичных ключевых терминов рассчитываются двумя разными методами:

1). по вероятности вхождения термина в документы категории в обучающей выборке

2). по вероятности вхождения термина в Интернет контекст категории

- Первичные ключевые термины входят в состав вторичных ключевых терминов
- Рассчет **весовых коэффициетов** для выравнивания весов, рассчитанных разными методами

Определение интересов пользователя по истории его посещений и запросов

- Определение категорий веб-страниц, посещенных пользователем
- Определение наиболее посещаемых категорий веб-страниц
- Построение профиля интересов пользователя по категориям и терминам, имеющим доверительный интервал больше некоторого порога

Заключение

- В работе рассмотрен метод определения интересов пользователя по истории его запросов и посещений Интернет сайтов
- Описанная программа успешно определяет интересы пользователя, в том числе, по коротким запросам, не содержащим терминов из обучающей выборки

Спасибо за внимание!

М.М.Шарнин (mc@keywen.com),[1] А.В.Петров,[2] И.П.Кузнецов[1]

[1]Институт Проблем Информатики РАН, Москва

[2]Tinkoff Digital, Москва