Выявление семантических характеристик в слабоструктурированных текстовых данных

Нурк Сергей Юрьевич, 445 гр.

Научный руководитель: к.ф.-м.н, доц. К.В. Вяткина

Введение

Информация в интернете

- Понятна для человека
- Загадка для компьютера

Структурирование информации необходимо

- Тематические каталоги
- Вертикальный поиск

Экстракция

- Качественная, но дорогая
- Автоматическая, но ...

Что интересует в данных

Признаки

- Словарные
- Количественные
- Другие

Задача

- **Цель:** узнать до куда удастся дойти, не взяв ничего с собой
- **Постановка:** по набору слабоструктурированных сущностей как можно более качественно выявить словарные признаки, характерные для исходных объектов

Общая схема

- Предобработка, сбор статистики употребления слов
- Анализ статистики
- Восстановление ответа

Анализ статистики

- Синтез методов text-minig и кластеризации графов
- “Пляшем от слов”
- Единица анализа – вектор использования
- Этапы
  - Нормировка
  - Подсчет similarity measures (мер сходства)
  - HCS-кластеризация (Highly Connected Subgraphs)

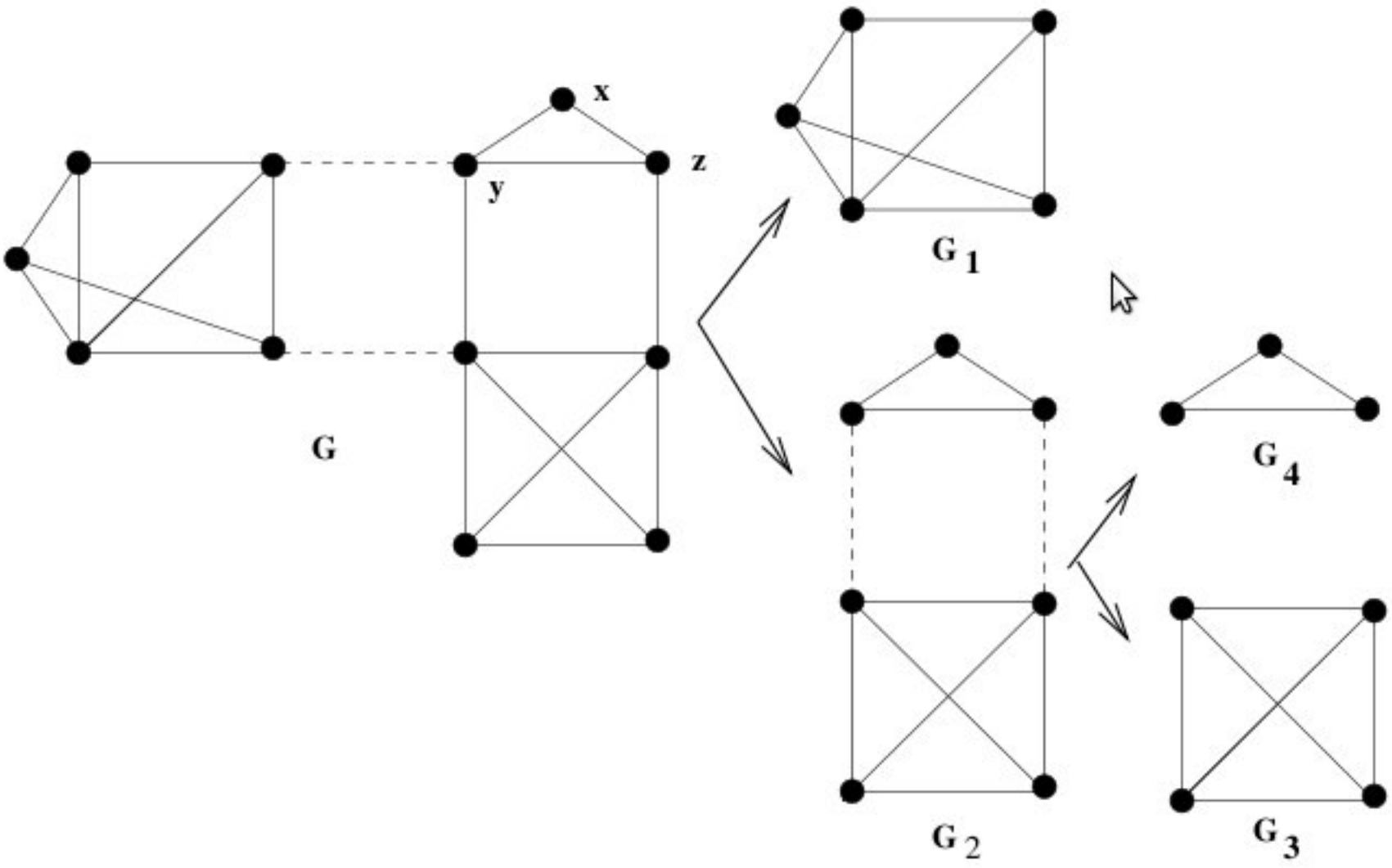
HCS-clustering

- Работает с не взвешенными ненаправленными графами
- Почему именно HCS
  - Устойчивость
  - Теоретически обоснованные причины надеяться на качество кластеров
  - Отсутствие настройки
  - Эвристики, ускоряющие алгоритм и улучшающие качество кластеров

HCS Simple

```
HCS(G) {
  (H; H'; C ) ← MINCUT(G)
  if (|C| > n/2)
     return G
  else {
     HCS(H)
     HCS(H')
  }
}
```

Пример



Результаты на реальных данных

Название 1: кухн

Словарь 1:немецк, французская, грузинск, русск, американск, европейск, китайск, мексиканск, восточн, итальянск, кавказск, японск, смешанн, авторск

Название 2: ориентир

Словарь 2: академическ, проспект просвещен, маяковск, ладожск, стар деревн, восстан, владимирск, чернышевск

Название 3: unknown

Словарь 3: пицц, клуб, центр, каф, кофейн, трактир, бар, ресторан

Название 4: кредитн карт

Словарь 4: mastercard, maestro, visa, visa electron, american express

Итоги

- Положительный ответ на вопрос о возможности успеха подобного анализа
- Разработан подход для выявления словарных признаков в наборе слабоструктурированных сущностей
- Эффективная программная реализация на языке Java
- Получены достаточно качественные результаты на тестовой выборке