

МЕЖДУНАРОДНЫЙ СОЛОМОНОВ УНИВЕРСИТЕТ



Дмитрий Владимирович ЛАНДЭ

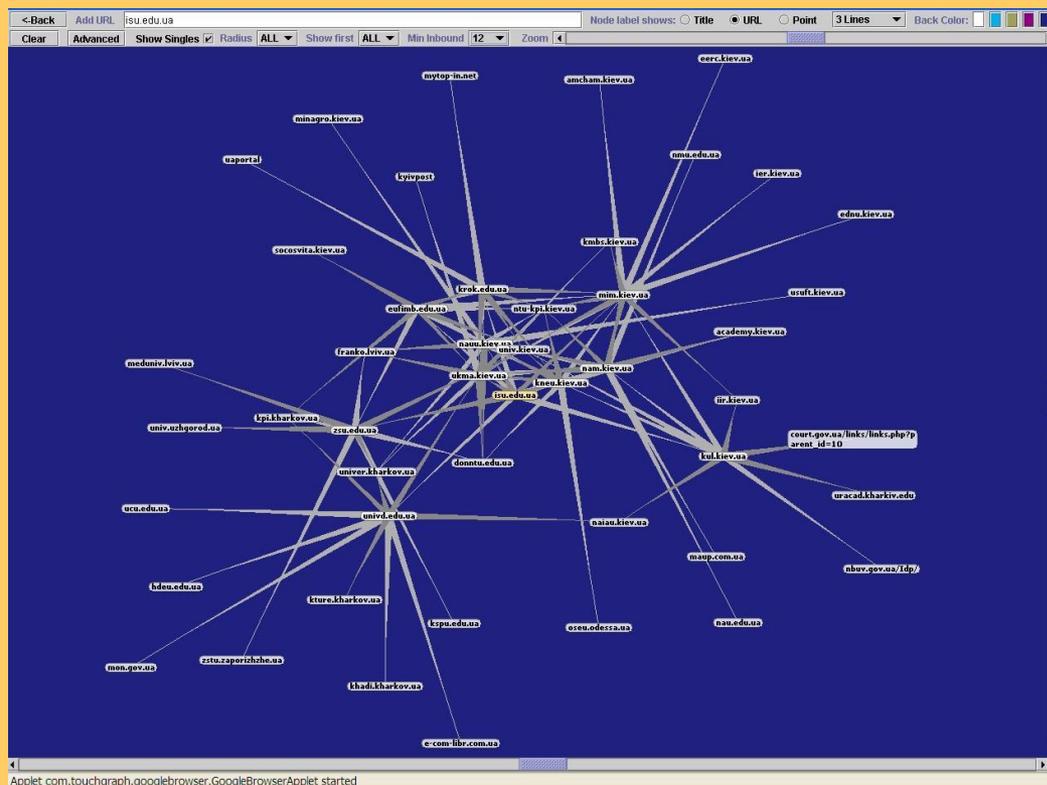
Лекция 7
“Кластерный анализ
и информационный поиск”



Понятие «кластерного анализа»

Кластерный анализ - метод группировки экспериментальных данных в классы. Наблюдения, попавшие в один класс, в некотором смысле ближе друг к другу, чем к наблюдениям из других классов.

(Глоссарий.ru)



Пример кластеров сайтов - «групп подобия по контенту»
(www.touchgraph.com)



Понятие информационного портрета

Портрет - модель реального объекта, выраженную его наиболее узнаваемыми чертами.

Информационный портрет документа - статистически значимая совокупность информационных характеристик.

В качестве информационного портрета темы можно рассматривать множество ключевых слов, наиболее точно (по статистическим и смысловым алгоритмам) отражающее информацию, соответствующую данной теме.

Тематической рубрике соответствует ее информационный портрет:

$$P_i = \{v_{ij}\}, (j=1, \dots, K),$$

где v_{ij} –весовой коэффициент, соответствующий j -му терму, K - количество термов в словаре системы.

Информационный портрет	
Уточнить запрос	
	Рубрики (32)
	Языки (2)
	Размер (3)
	Цифровая насыщенность (3)
	Страны источников (13)
	Источники (50)
	Слова (80)
AND	NOT
<input type="checkbox"/>	АВАЛ <input type="checkbox"/>
<input type="checkbox"/>	АГЕНТСТВ <input type="checkbox"/>
<input type="checkbox"/>	АКБ <input type="checkbox"/>
<input type="checkbox"/>	АКТИВ <input type="checkbox"/>
<input type="checkbox"/>	АКЦ <input type="checkbox"/>
<input type="checkbox"/>	АЛЬФ <input type="checkbox"/>
<input type="checkbox"/>	АМЕРИКАНСК <input type="checkbox"/>
<input type="checkbox"/>	БАН <input type="checkbox"/>
<input type="checkbox"/>	БАНК ** <input type="checkbox"/>
<input type="checkbox"/>	БАНКОВСК <input type="checkbox"/>
<input type="checkbox"/>	БИЗНЕС <input type="checkbox"/>
<input type="checkbox"/>	БУМАГ <input type="checkbox"/>
<input type="checkbox"/>	ВАНК <input type="checkbox"/>
<input type="checkbox"/>	ВАЛЮТ <input type="checkbox"/>
<input type="checkbox"/>	ВАЛЮТН <input type="checkbox"/>



Взвешивание потока документов в пространстве информационного портрета

$M = \{m_{ij}\}$ ($i = 1, \dots, N; j = 1, \dots, K$) - матрица соответствия потока документов D информационному портрету l .

$D = \{d_i\}$ ($i = 1, \dots, K$). d_i – определяется как $TF * IDF$.

Близость D и P_i – $sim(D, P_i)$ – скалярное произведение K -мерных векторов.

Алгоритм взвешивания:

```
while не исчерпан список актуальных на момент  $t$  документов do  
  for каждого документа do  
    while не исчерпан список информационных портретов  
      for каждого информационного портрета do  
        Определение веса документа  
        if вес больше порогового значения  
          then do приписывание документу рубрики, соответствующей  
            информационному портрету  
          end if  
        end for  
      end while  
    end for  
  end while  
Итоговый подсчет веса потока документов  
Визуализация гистограммы
```



Латентное семантическое индексирование

Метод кластерного анализа **LSI** (латентного семантического индексирования), базируется на сингулярном разложении матриц (**SVD**). Сингулярным разложением матрицы A называется ее разложение вида $A=USV^T$, где U и V – ортогональные матрицы, а S – диагональная матрица, элементы которой $s_{ij} = 0$, если i не равно j , а $s_{ii} \geq 0$. В рассматриваемом примере (таблиц взаимосвязей) матрица $A = M^T M$ – квадратная, однако метод LSI применяется и к прямоугольным матрицам, но в этих случаях размерность матрицы S соответствует рангу матрицы A .

В соответствии с методом LSI в рассмотрение берутся k наибольших сингулярных значений, а каждому такому сингулярному значению матрицы A соответствует кластер взаимосвязанных документов. A аппроксимируется матрицей $A_k = \sum u_i s_{ii} v_i^T$.

Метод LSI применим и к ранжированию выдачи информационно-поисковых систем, основанному на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых популярных на сегодня в области информационного поиска.

Ввиду своей вычислительной трудоемкости (равной $O(N^2)$, N – размерность A), этот метод LSI применяется только для относительно небольших матриц.



Взаимосвязь тем и метод k-means

Суть алгоритма k-means: случайным образом выбирается k векторов-строк, которые определяются как центроиды кластеров. Затем k кластеров наполняются – для каждого из оставшихся векторов-строк определяется близость к центроиду соответствующего кластера. После этого вектор-строка приписывается к тому кластеру, к которому он наиболее близок. После этого строки-векторы перегруппируются. Затем для каждого из новых кластеров заново определяется центроид. После этого заново выполняется процесс наполнения кластеров и т. д., пока процесс не стабилизируется или не зациклится.

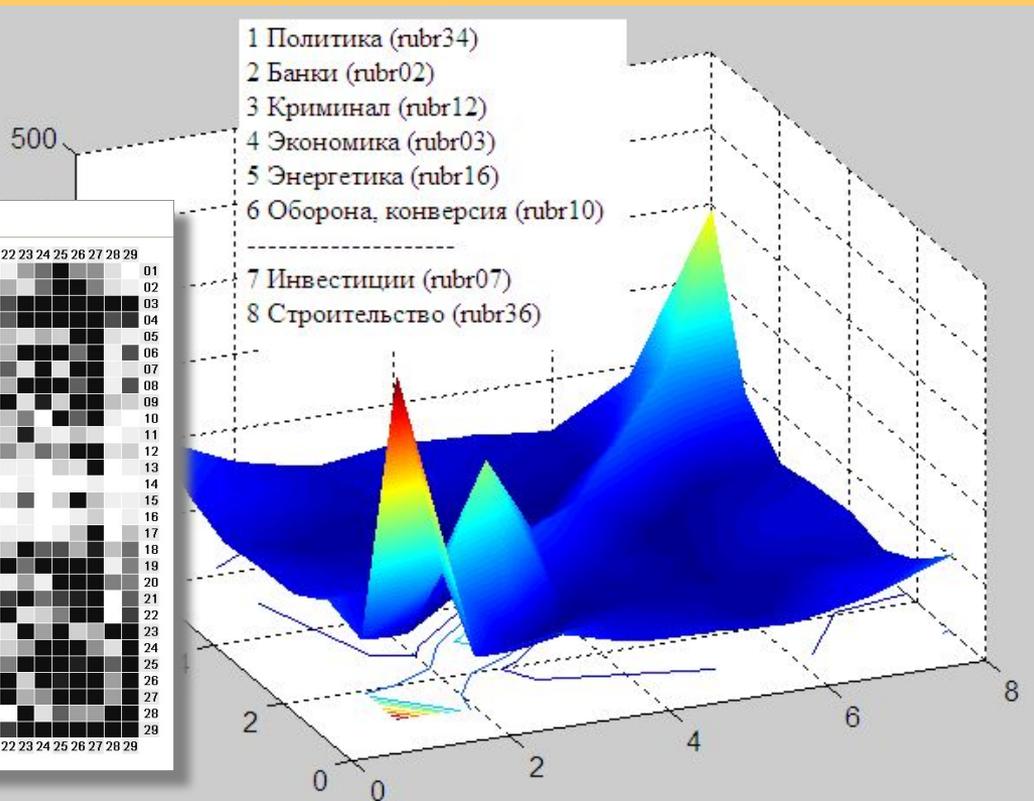
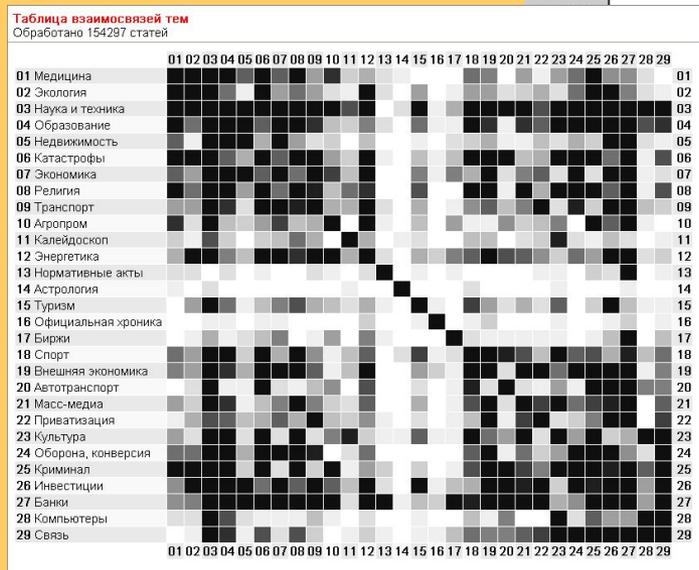
```
Произвольный выбор центроидов  $k$ -кластеров
while процесс формирования не стабилизировался do
  for каждого вектора-строки do
    найти кластер  $c$ , центроид которого наиболее близок к вектору-строке
    приписать вектор-строку кластеру  $c$ 
  end for
  for каждого кластера  $c$  do
    вычисление центроида кластера по входящим в него элементам
  end for
  for каждого вектора-строки do
    переставить элементы в векторе-строке,
    соответствующие выполненной перенумерования
  end for
end while
```



Группировка тем метод k-means

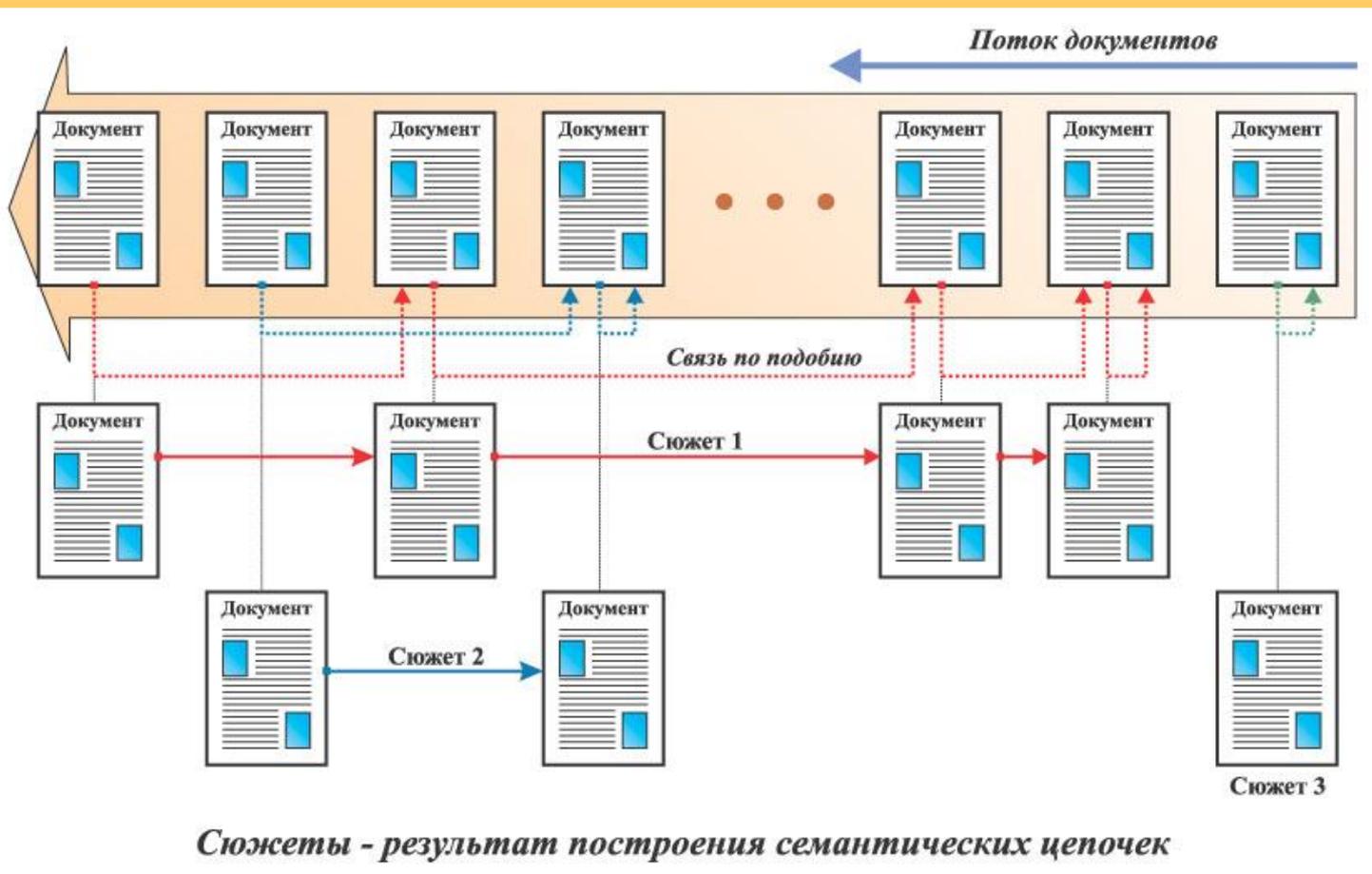
В отличие от метода LSI, *k-means* идеально подходит для кластеризации динамических информационных потоков.

Укрупнение рубрик – актуальная задача кластерного анализа и она может быть решена путем их группировки по признакам подобия. Выделение групп взаимосвязанных рубрик методом кластерного анализа *k-means*:





Метод, основанный на применении сетевого подхода - выявление сюжетов





Построение адаптивных интерфейсов уточнения запросов



Шаг 1



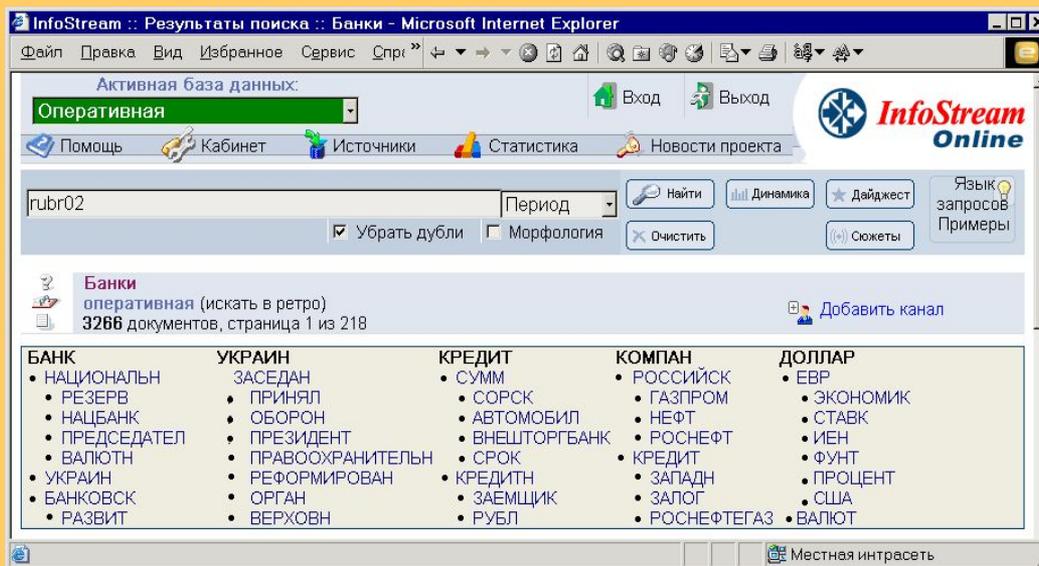
Шаг 2



Шаг 3



Шаг 4





Спасибо за внимание!

Ландэ Д.В

dwl@visti.net

<http://poiskbook.kiev.ua>

**МЕЖДУНАРОДНЫЙ СОЛОМОНОВ
УНИВЕРСИТЕТ
Киев, Украина**