Визуализация в информационном поиске

Константин Воронцов
k.v.vorontsov@phystech.edu

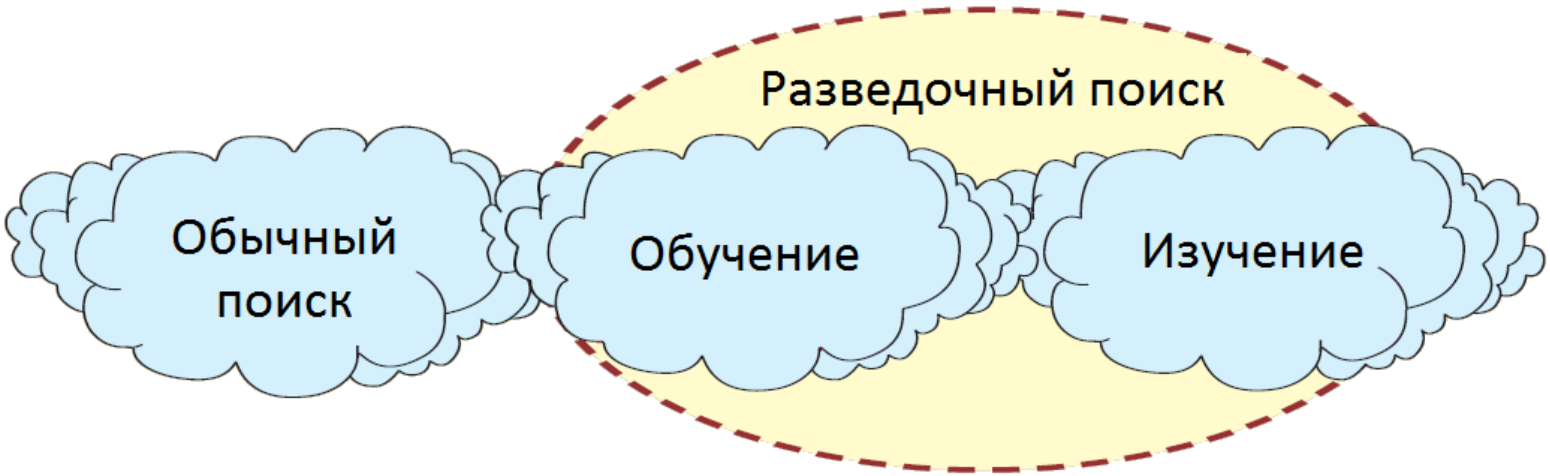
МФТИ • 15 сентября 2018

Содержание

1. **Разведочный информационный поиск**
   - Концепция разведочного поиска
   - Сценарии разведочного поиска
   - Технологии визуализации
2. **Метафоры визуализации текстов**
   - Карты и связи
   - Иерархии и структуры
   - Потоки и эволюции
3. **Замыслы**
   - Пространство «время–темы»
   - Иерархическая суммаризация
   - Фрактальное пространство знаний

Концепция разведочного поиска (exploratory search)

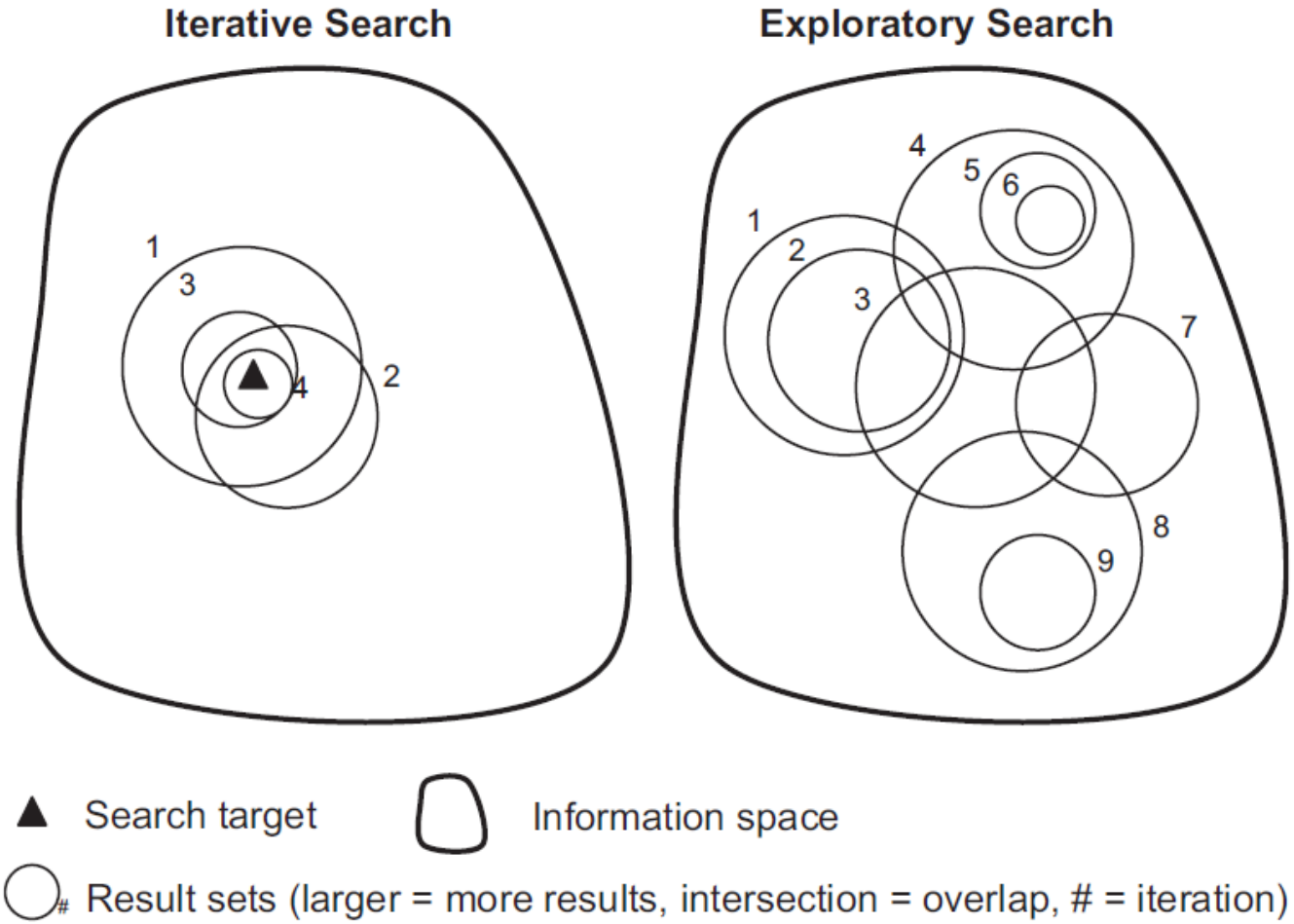
- пользователь может не знать ключевых терминов
- запросом может быть текст произвольной длины
- информационная потребность — систематизация знаний



навигация в сети,	самообразование,	исследование,
поиск фактов,	тематический поиск	экспертиза,
упоминаний,	систематизация	реферирование,
конкретных ответов	знаний	мониторинг тем

Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

От итераций «query-browse-refine» к разведочному поиску



R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

От ближнего чтения (close reading) к дальнему (distant reading)

Концепция дальнего чтения Франко Моретти

«*Дальнее чтение* — не ограничение, а способ представления знаний: меньше элементов, чётче понимание их взаимосвязей, акцент на формах, отношениях, структурах, моделях»

Мантра Шнейдермана

«Сначала крупный план, затем масшабирование и фильтрация, детали по требованию»

B.Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Visual Languages, 1996.

F.Moretti. Graphs, Maps, Trees: Abstract Models for a Literary History. 2005.

S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ / фрагмент документа / коллекция документов

Цели поиска:

- к каким темам относится мой запрос?
- что есть по этим темам нового/важного/популярного?
- в каком порядке лучше знакомиться с темой?
- какова тематическая структура этой предметной области?
- какие области являются смежными?

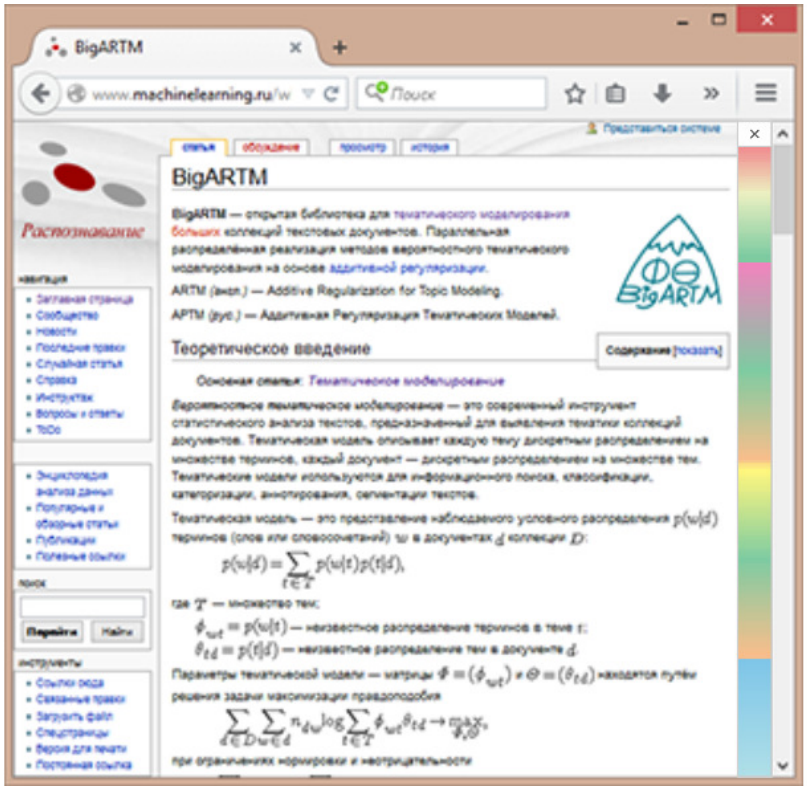
Сценарий поиска:

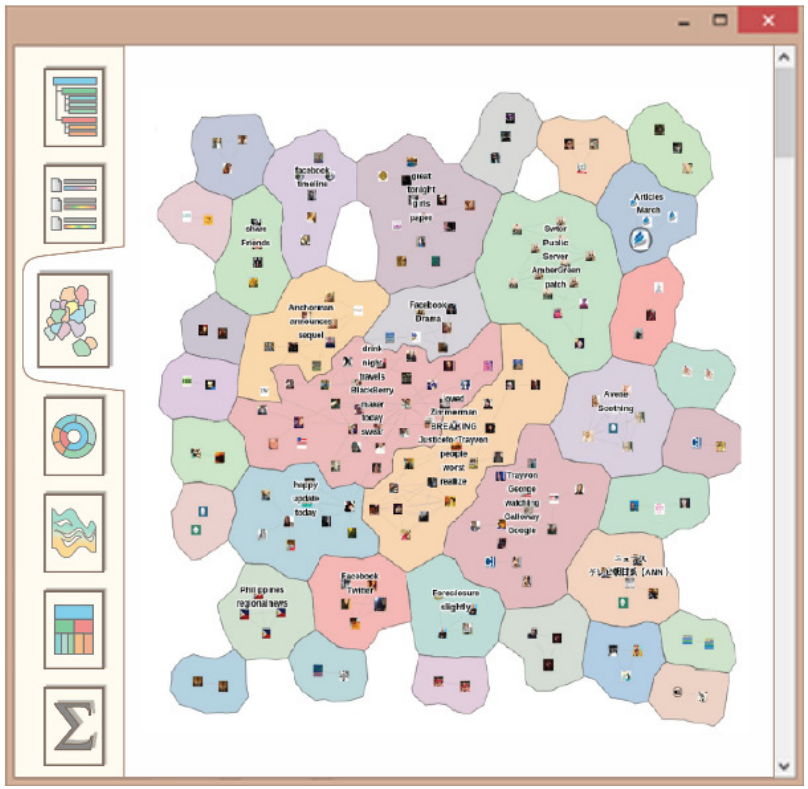
1. имея любой текст под рукой, в любом приложении,
2. получаем карту содержащихся в нём тем-подтем
3. и карту предметной области в целом

Переход от текстового запроса к визуализации (концепт)

Тематическая сегментация: структура документа-запроса

Визуализация: тематическая карта области



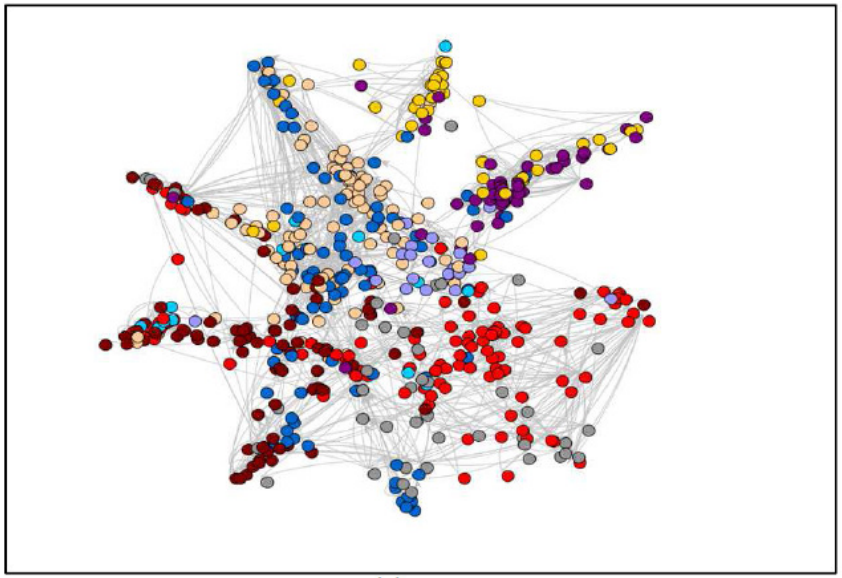


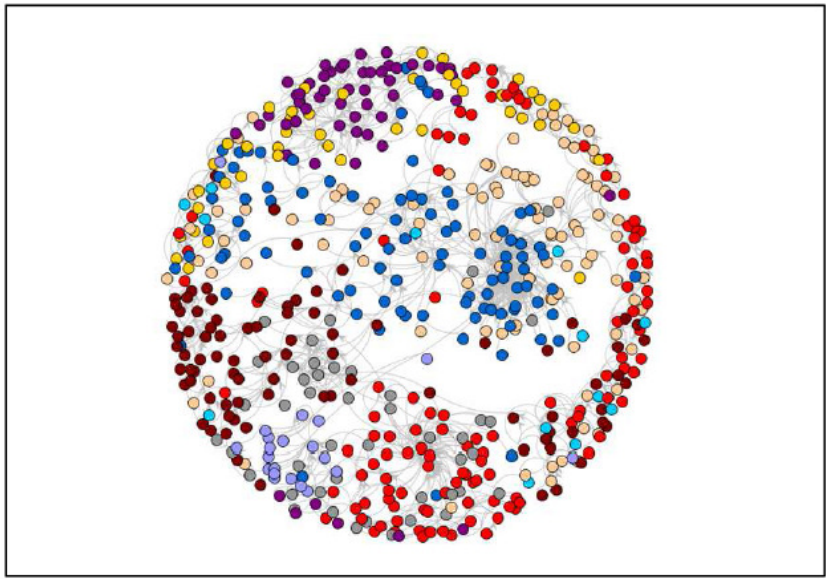
http://textvis.lnu.se

Интерактивный обзор 400 средств визуализации текстов

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Карта сходства: кластерная структура текстовой коллекции

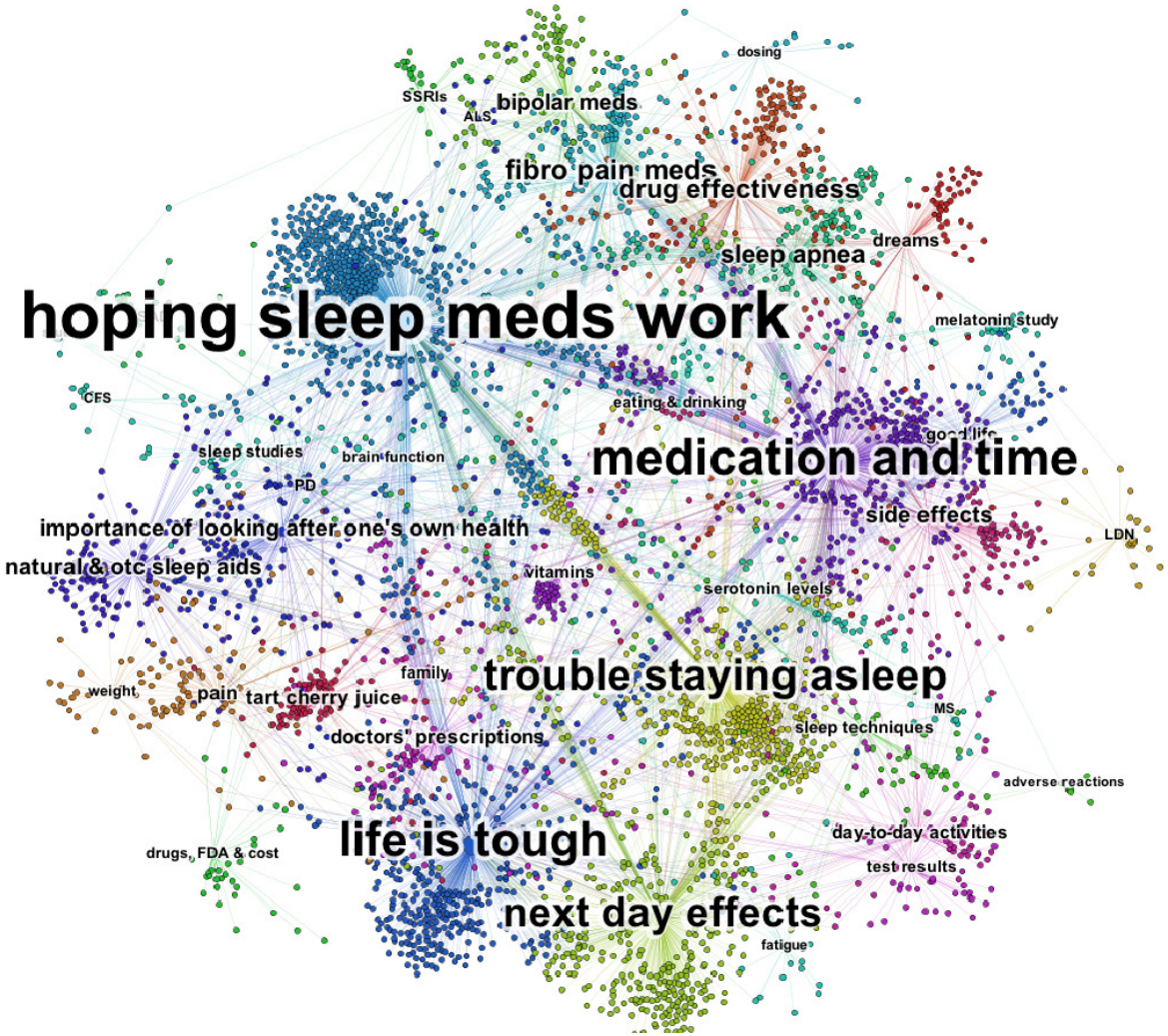




- Точки — это документы (или их фрагменты)
- Кластеры — это группы тематически схожих документов
- Форму облака точек можно настраивать

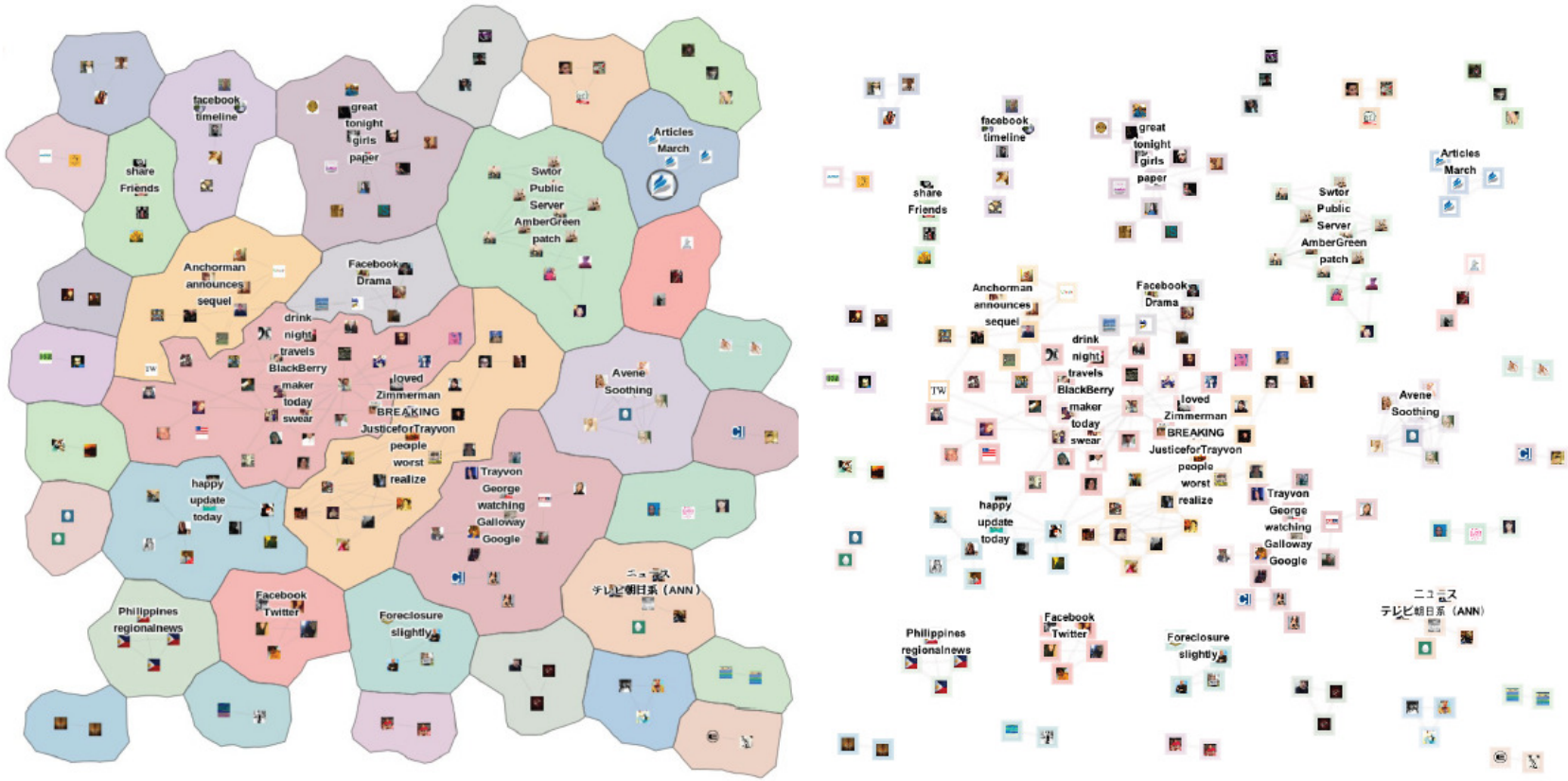
Tuan M. V. Le, Hady W. Lauw. Probabilistic Latent Document Network Embedding. IEEE International Conference ICDM. 2014.

Пример: тематика обсуждений на www.PatientsLikeMe.com



Chen A., Eichler G. Topic Modeling and Network Visualization to Explore Patient Experiences. 2013.

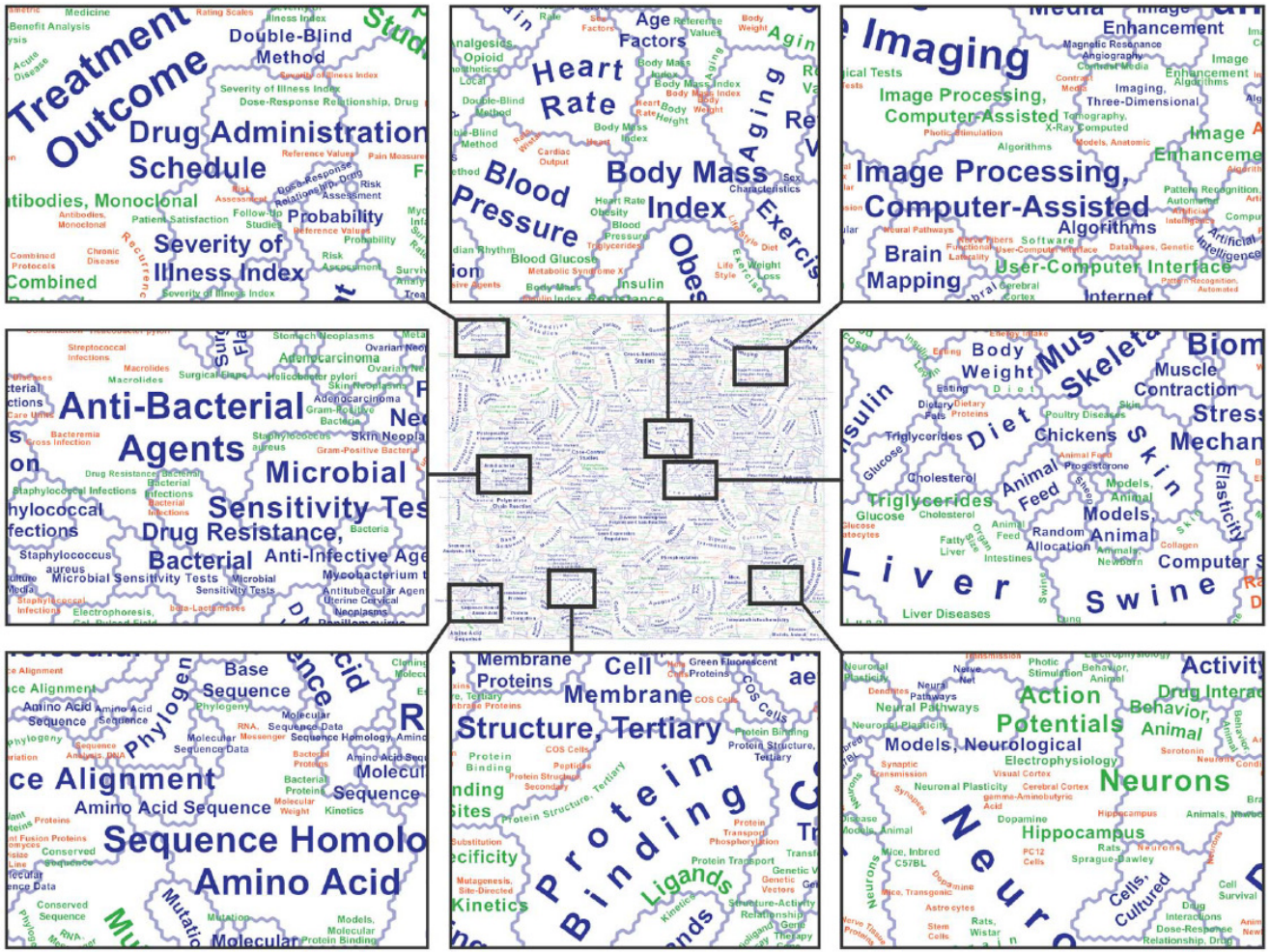
Географическая метафора: кластерная структура коллекции



«A map metaphor visualization (left) seems more appealing than a plain graph layout (right), and clusters seem easier to identify.»

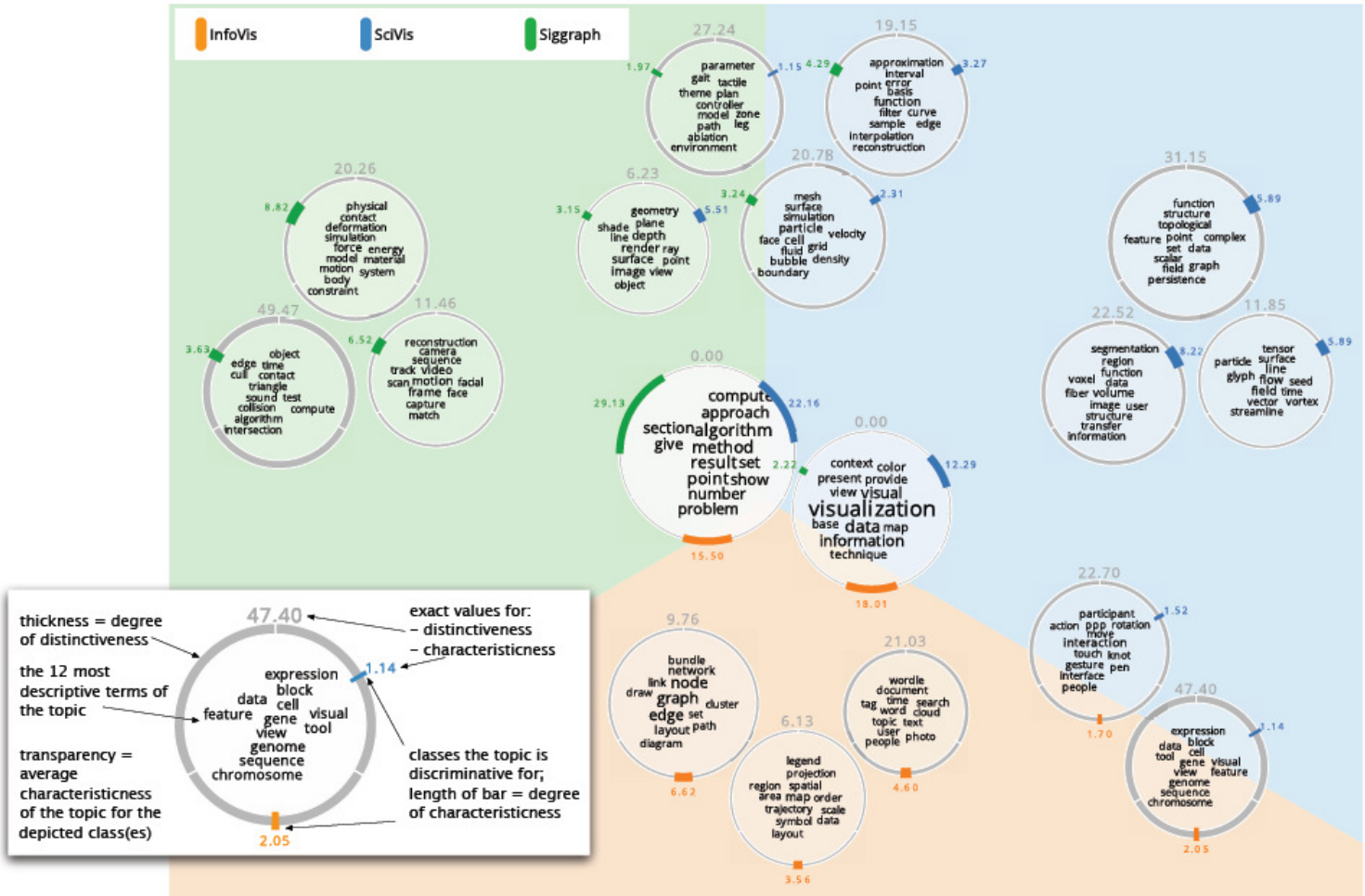
E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

Пример карты медицинских знаний



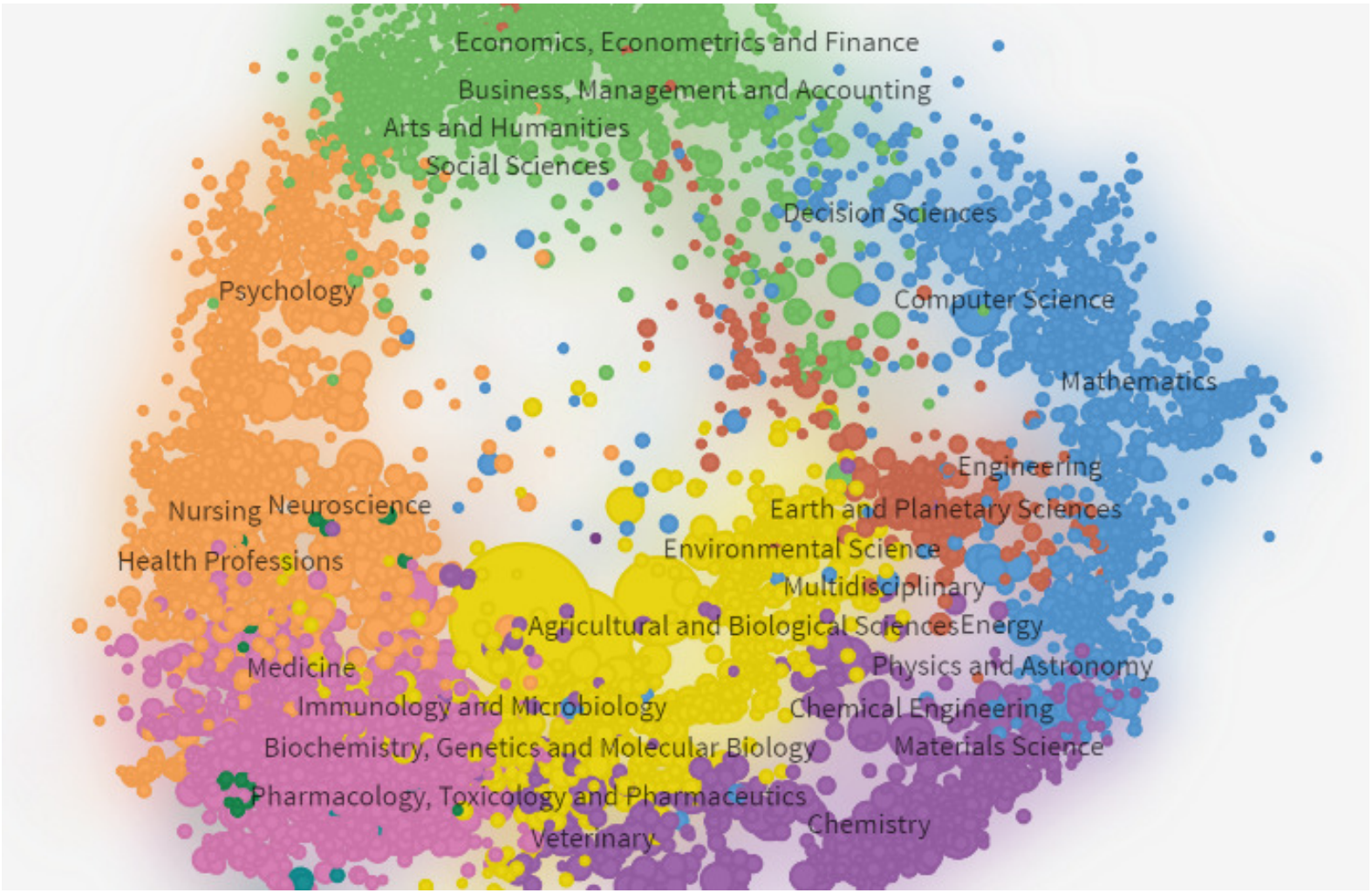
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Тематический анализ источников



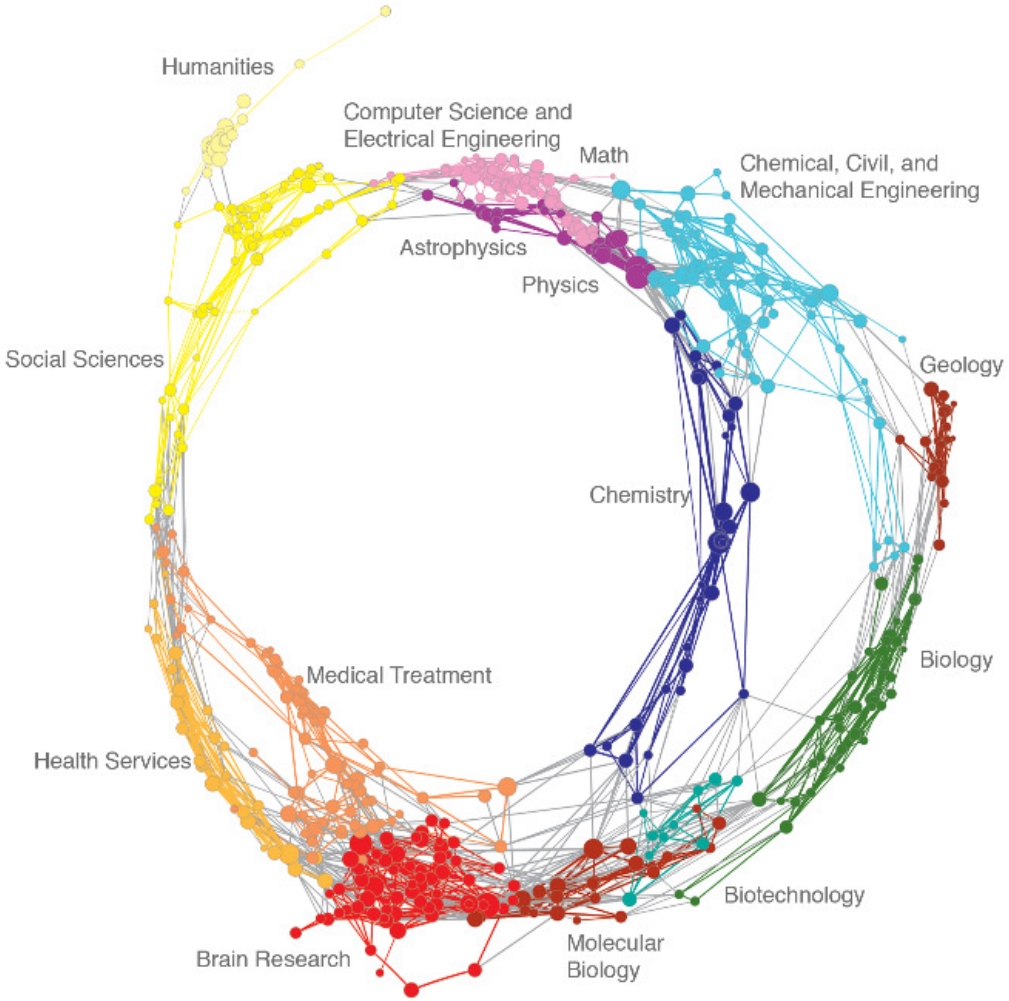
Oelke D., Strobelt H., Rohrdantz C., Gurevych I., Deussen O. Comparative exploration of document collections: a visual analytics approach. EuroVis. 2014.

Пример карты науки



http://onlinelibrary.wiley.com/browse/subjects

Ещё один пример карты науки



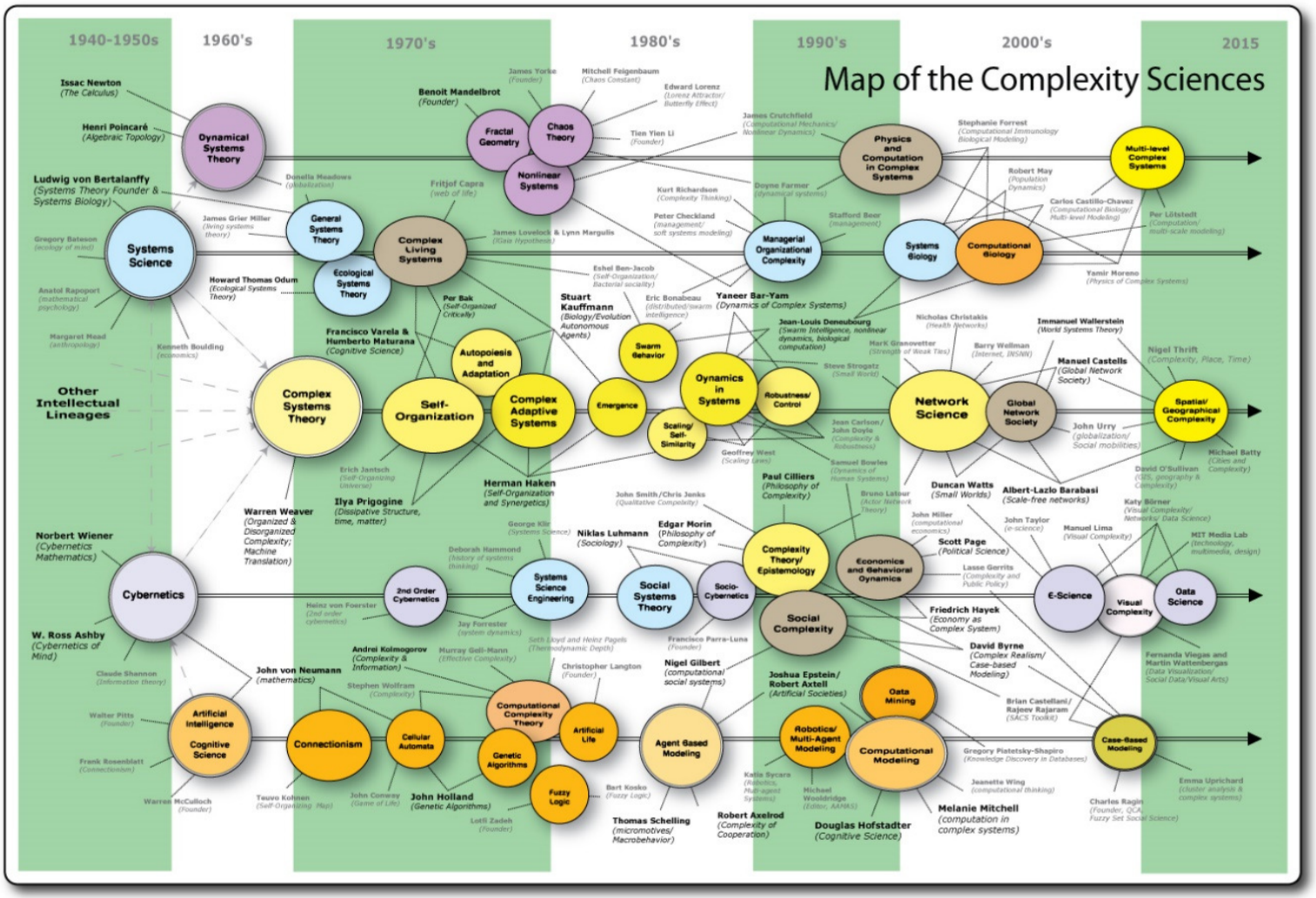
http://scimaps.org

Важное наблюдение:
области знания самопроизвольно располагаются по кругу, значит,
их можно располагать и вдоль прямой линии.

Недостатки:

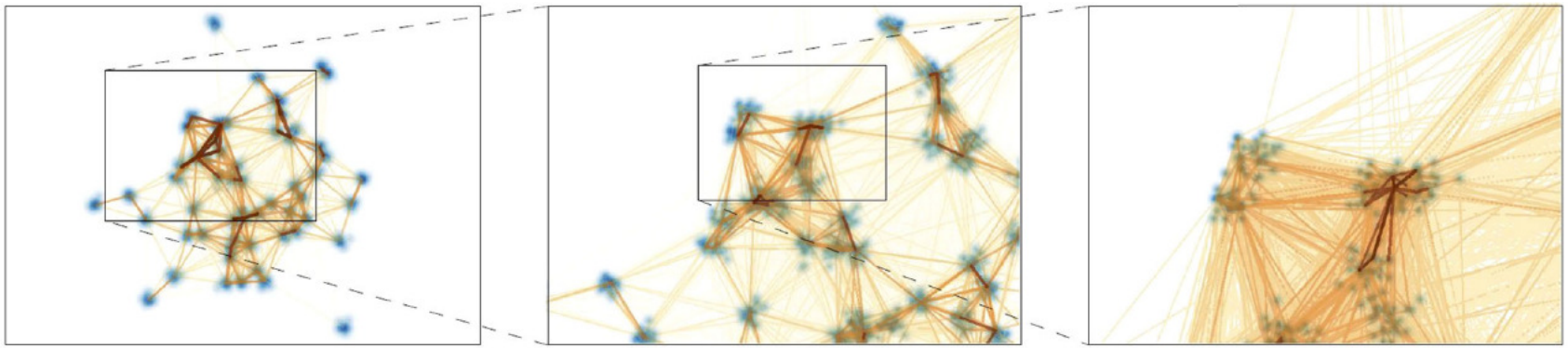
- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

Пример карты предметной области, построенной вручную



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

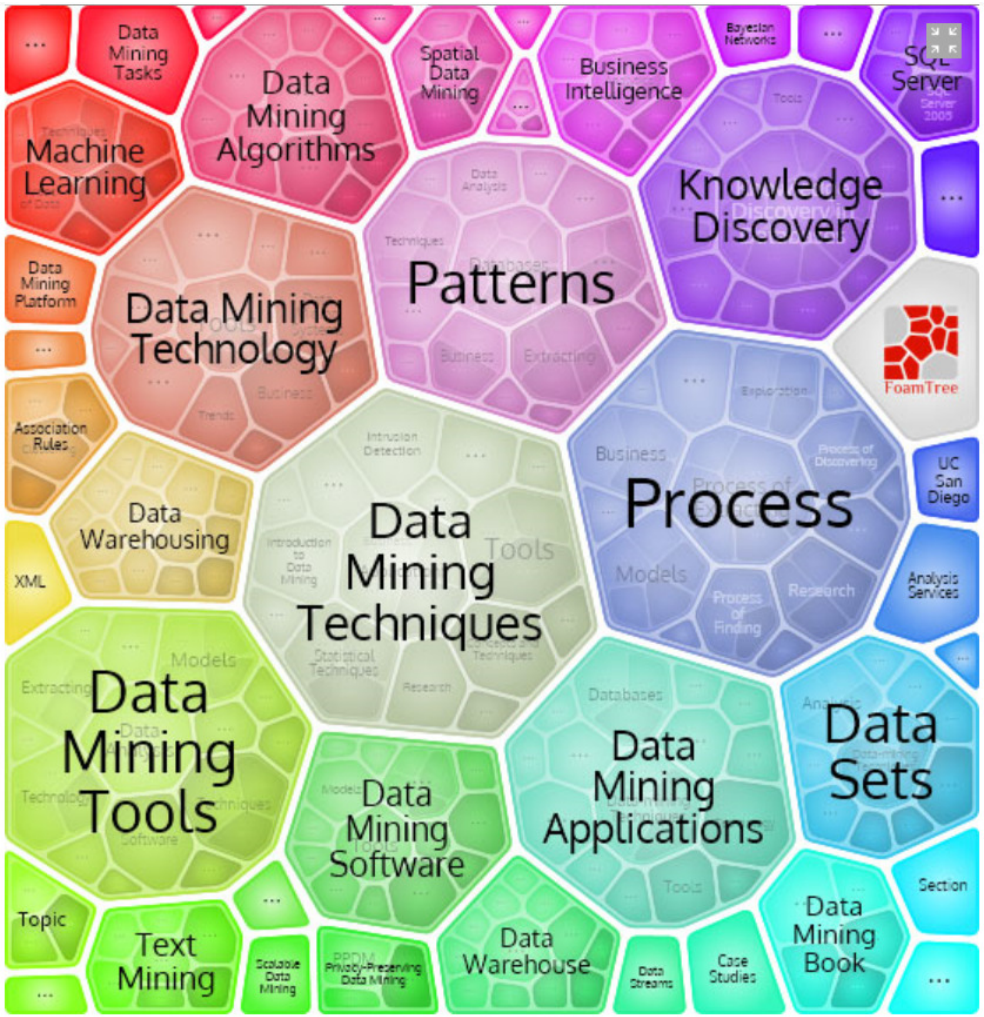
Фрактальная природа тематических кластерных структур



- Кластеры
  - кластеров
    - кластеров
      - кластеров...

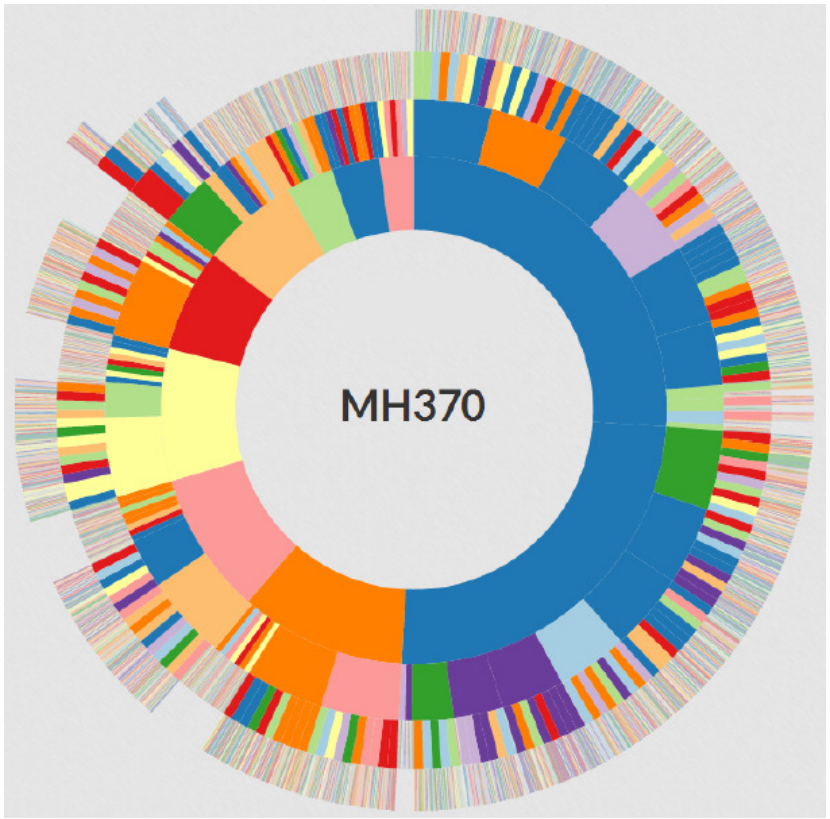
M.Zinsmaier, U.Brandes, O.Deussen, H.Strobelt. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

Пример иерархической карты области *Data Mining*



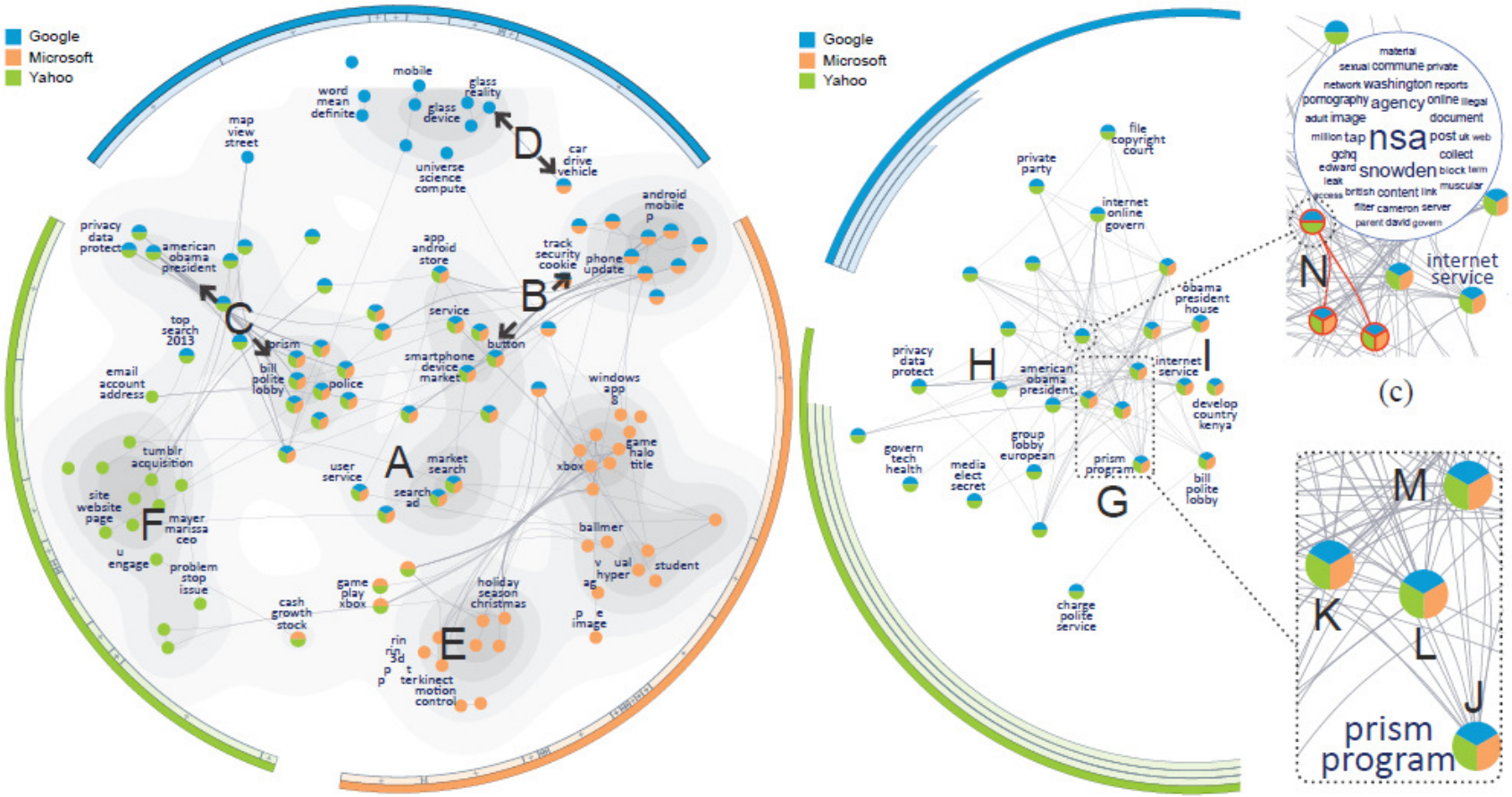
FoamTree: https://carrotsearch.com/foamtree

Тематическая иерархия: альтернативное представление



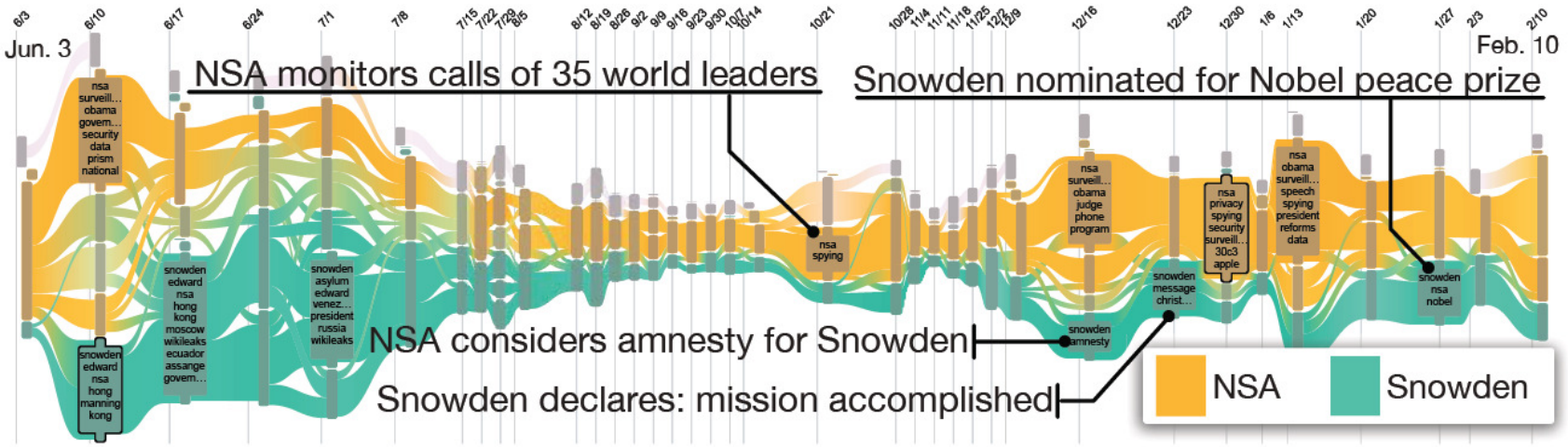
Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Тематическая иерархия: дерево, области и круговая диаграмма



Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, Baining Guo. TopicPanorama: a full picture of relevant topics. IEEE VAST, 2014.

Динамика тем во времени: эволюция предметной области



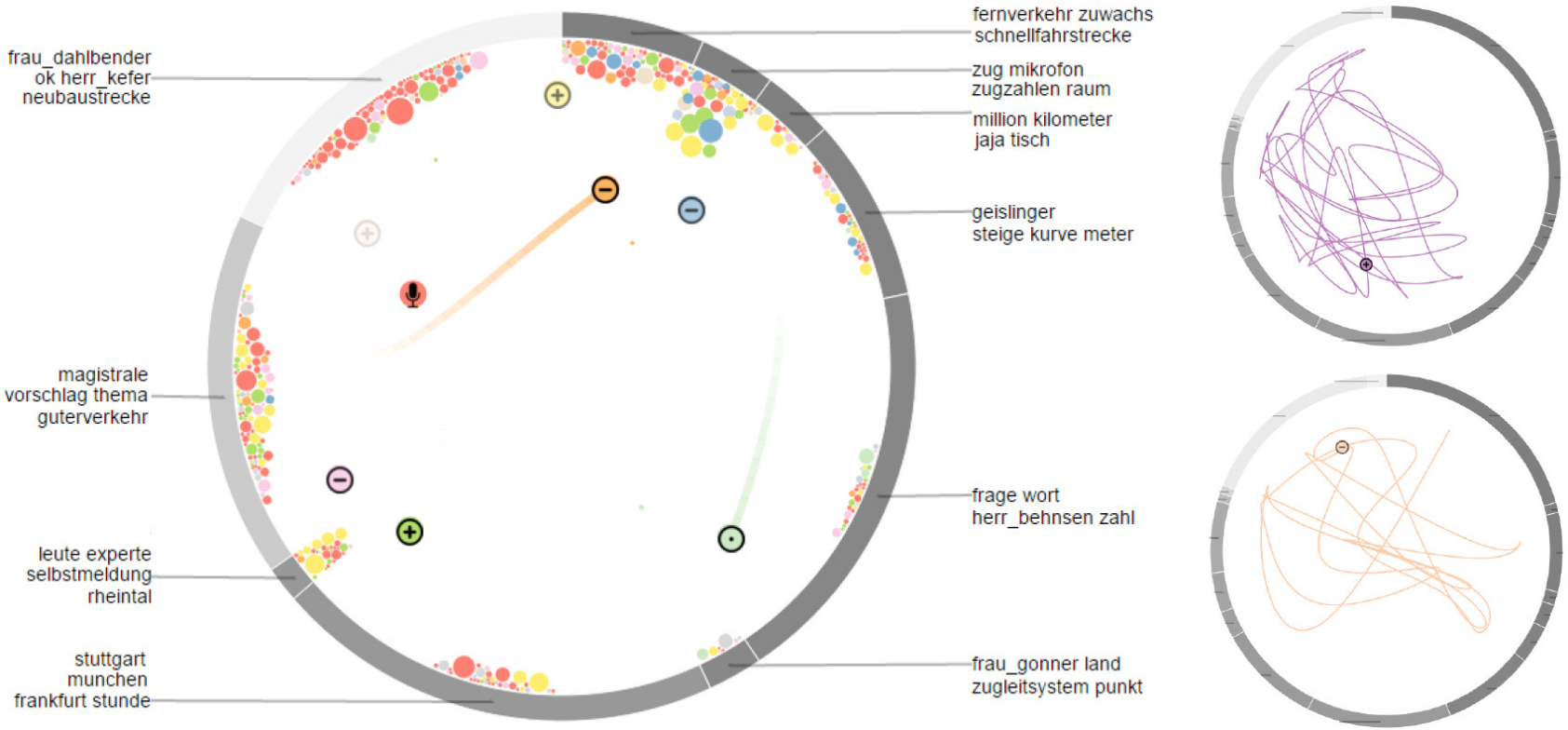
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

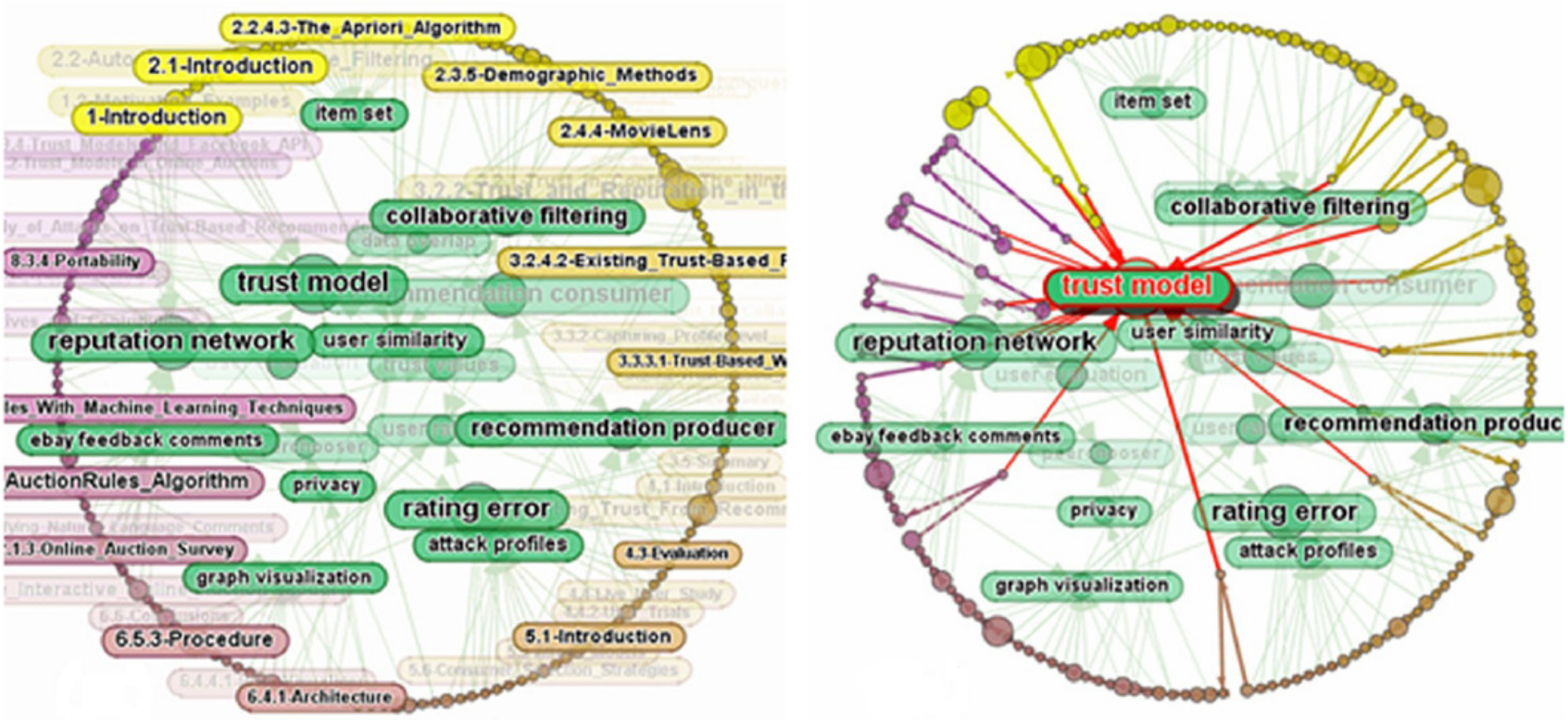
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

Динамика тем внутри полемического диалога



M.El-Assady1, V.Gold, C.Acevedo, C.Collins, D.Keim. ConToVi: Multi-Party Conversation Exploration Using Topic-Space Views. 2016.

Динамика тем внутри документа: тематическая сегментация

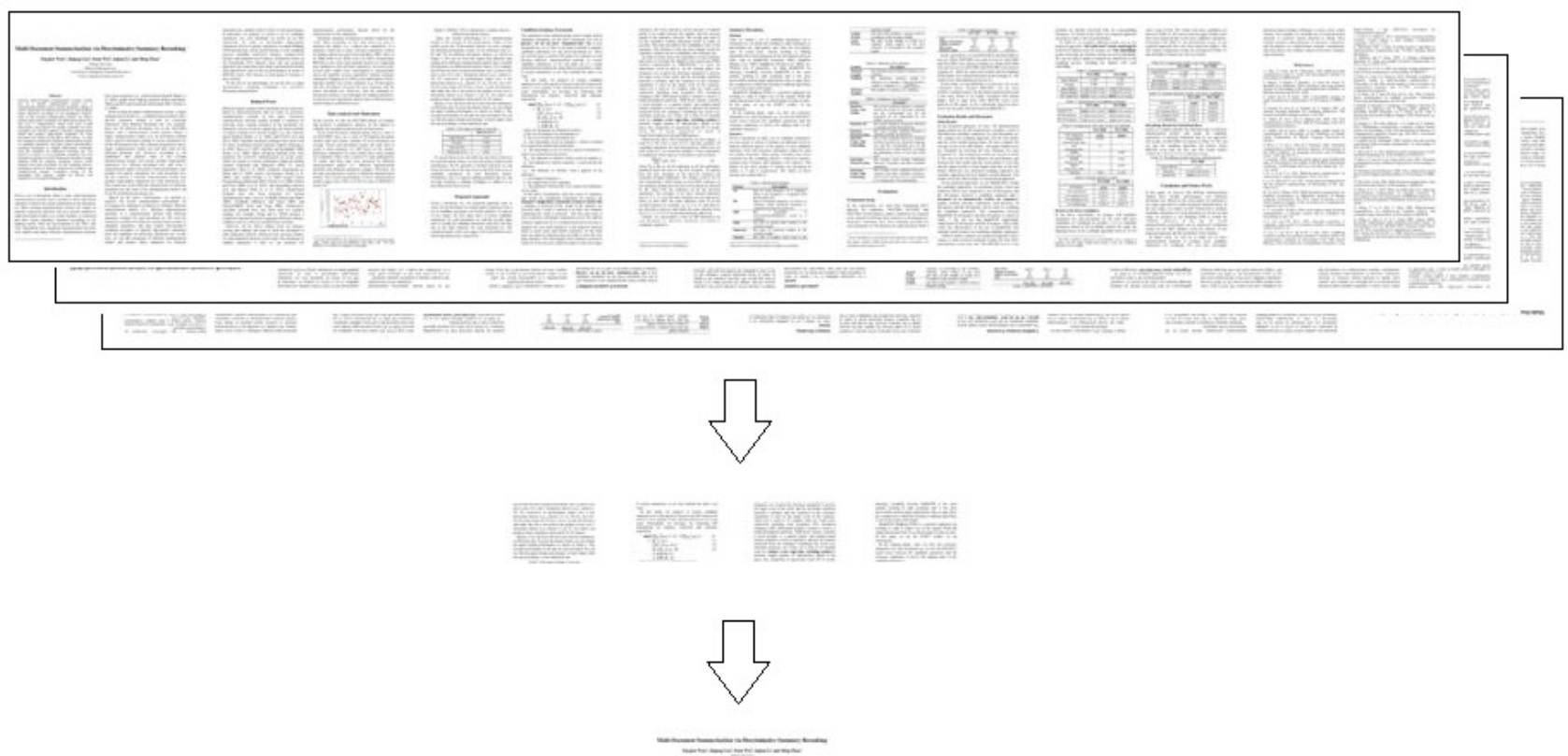


Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Визуализация тематического разведочного поиска (концепт)

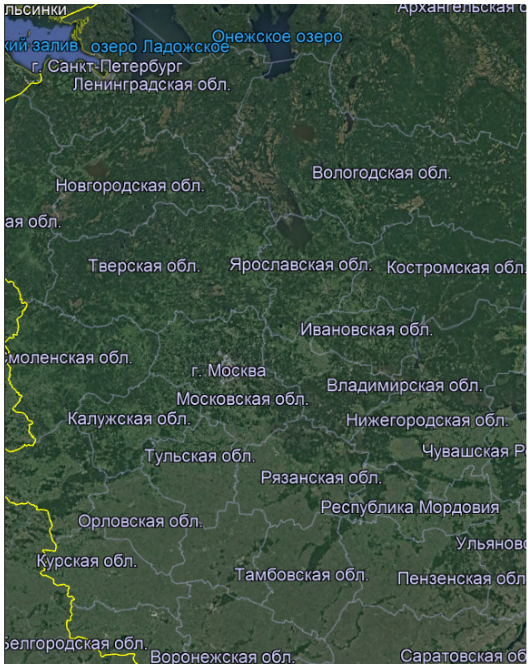
- Интерпретируемые оси: время–темы или сложность–темы
- Спектр тем: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- Темы делятся на подтемы иерархически
- Интерактивность: реализация мантры Шнейдермана
- При любом масштабе на карте достаточно много текста

От простой суммаризации к иерархической



Аналогия с геоинформационными системами

Если представить, что вся Земля плотно покрыта текстами, написанными человечеством, то с любой высоты мы должны прочитать сжатое, и в то же время понятное, краткое содержание текста, находящегося в поле зрения.



Эволюция информационного поиска

- Книги, библиотеки — долго искать, долго понимать
- Поисковые машины — быстро искать, долго понимать
- Визуальный поиск — быстро искать, быстро понимать