Вероятностные тематические модели
Лекция 12. Визуализация

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 2 декабря 2020

Содержание

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.

Многомерное шкалирование (MDS, multidimensional scaling)

Дано:

конечное множество T-мерных векторов $\theta_d \in \mathbb{R}^T$, $d \in D$,
$R_{dd'}$ — попарные расстояния между ними.

Найти:

двумерные представления $(x_d, y_d) \in \mathbb{R}^2$ для каждого θ_d.

Критерий: аппроксимация исходных расстояний двумерными

$$\sum_{dd' \in D} w_{dd'} \big(r_{dd'}(X, Y) - R_{dd'}\big)^2 \to \min_{(X,Y)},$$

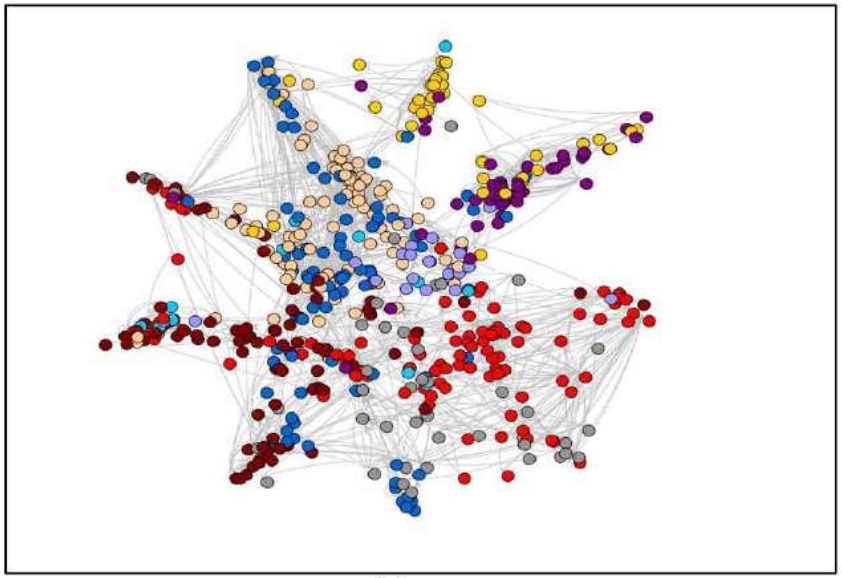
где $r_{dd'}(X, Y)$ — евклидово расстояние между точками в $\mathbb{R}^2$:

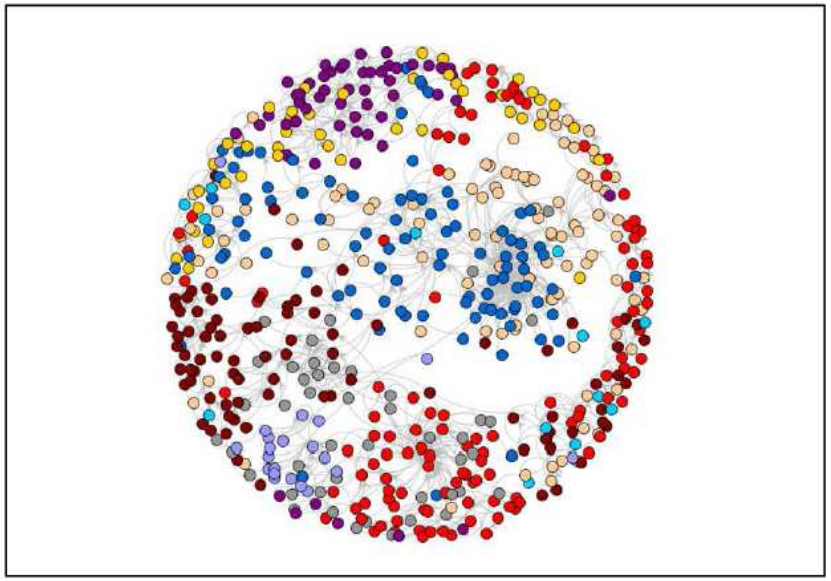
$$r^2_{dd'}(X, Y) = (x_d - x_{d'})^2 + (y_d - y_{d'})^2.$$

Наиболее популярный современный метод решения — tSNE.

Laurens van der Maaten, Geoffrey Hinton. Visualizing Data using t-SNE. 2008

Карта сходства: кластерная структура текстовой коллекции

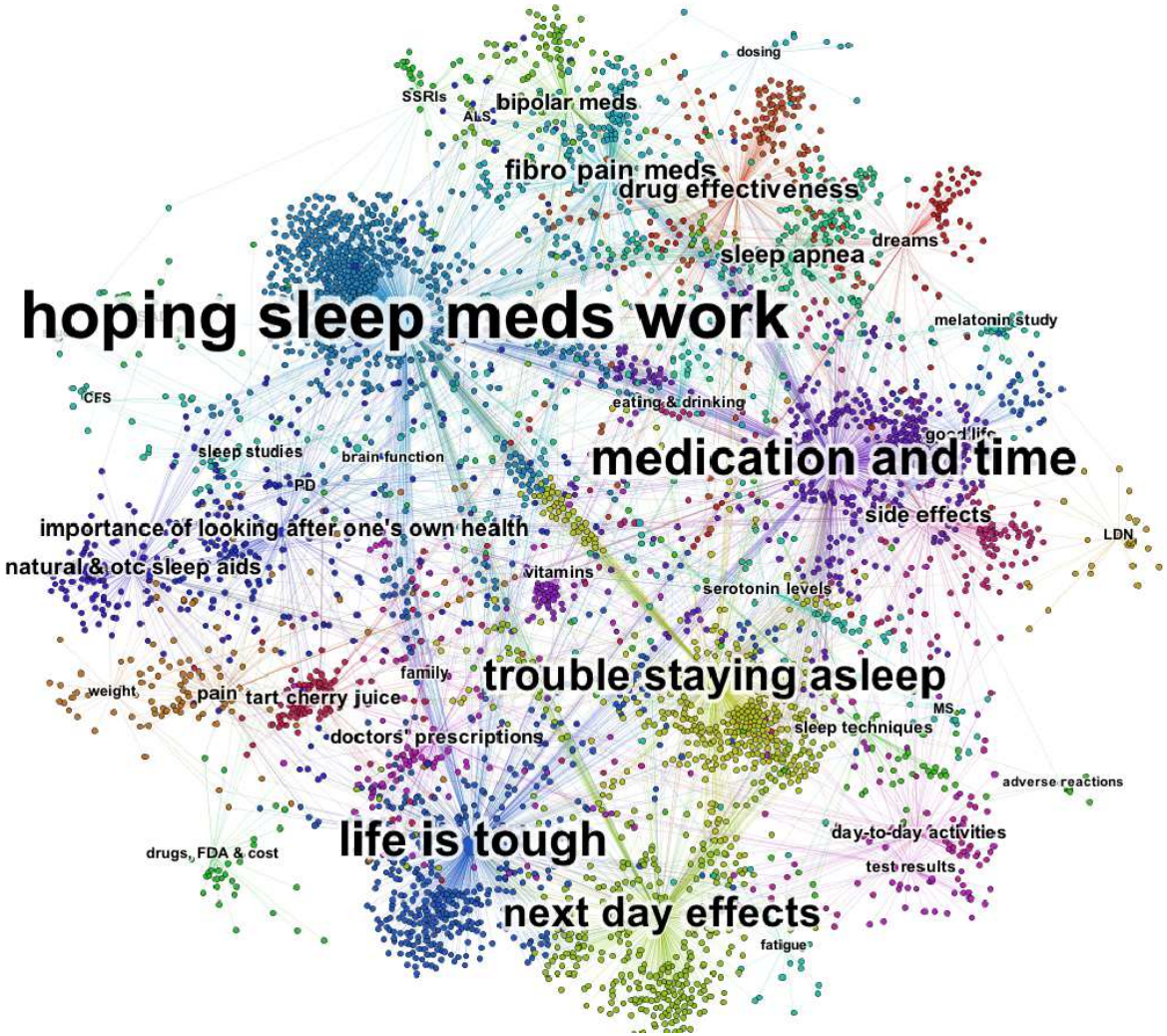




- Точки — это документы (или их фрагменты)
- Кластеры — это группы тематически схожих документов
- Форму облака точек можно настраивать

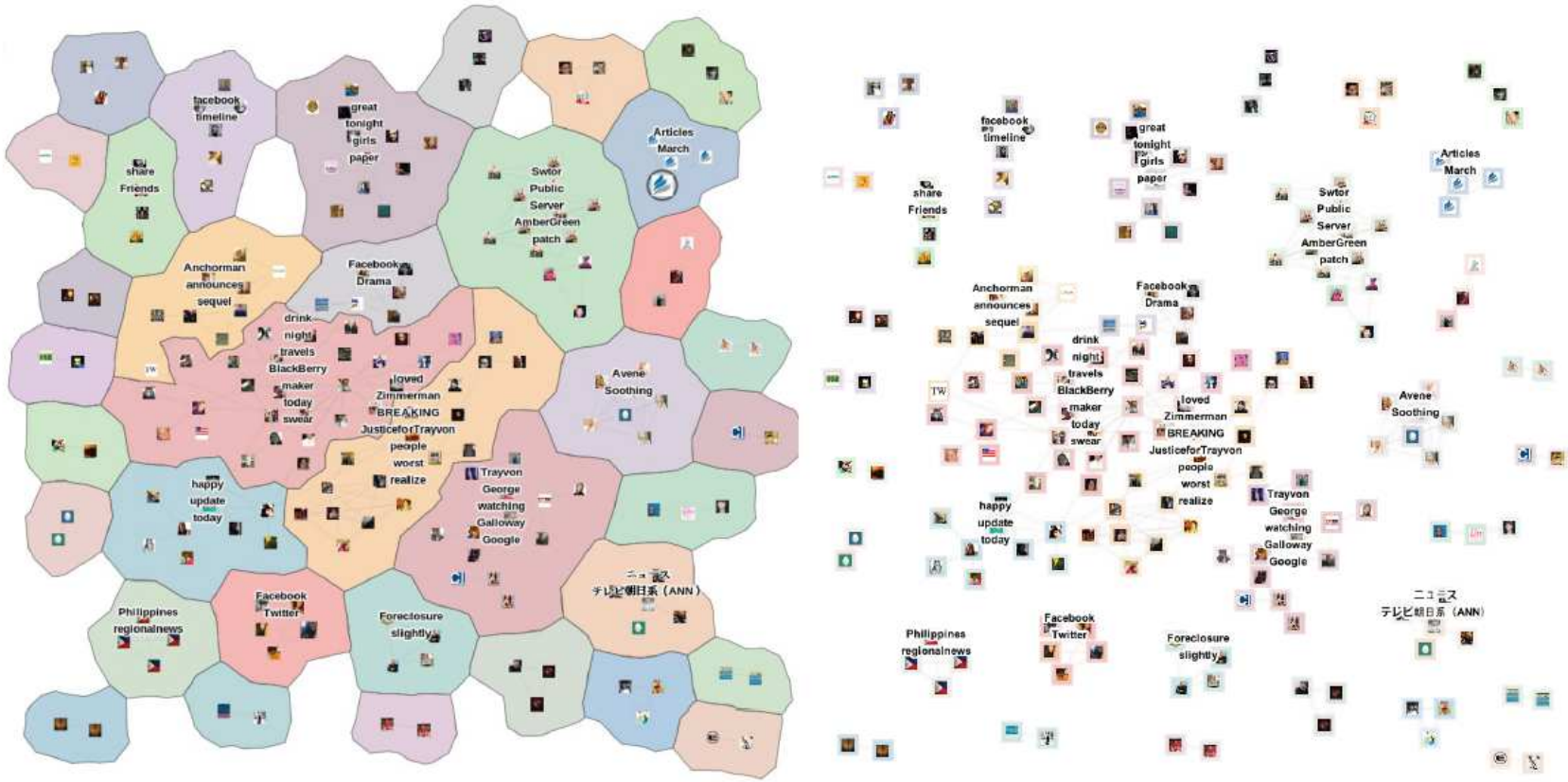
Tuan M. V. Le, Hady W. Lauw. Probabilistic Latent Document Network Embedding. IEEE International Conference ICDM. 2014.

Пример: тематика обсуждений на www.PatientsLikeMe.com



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

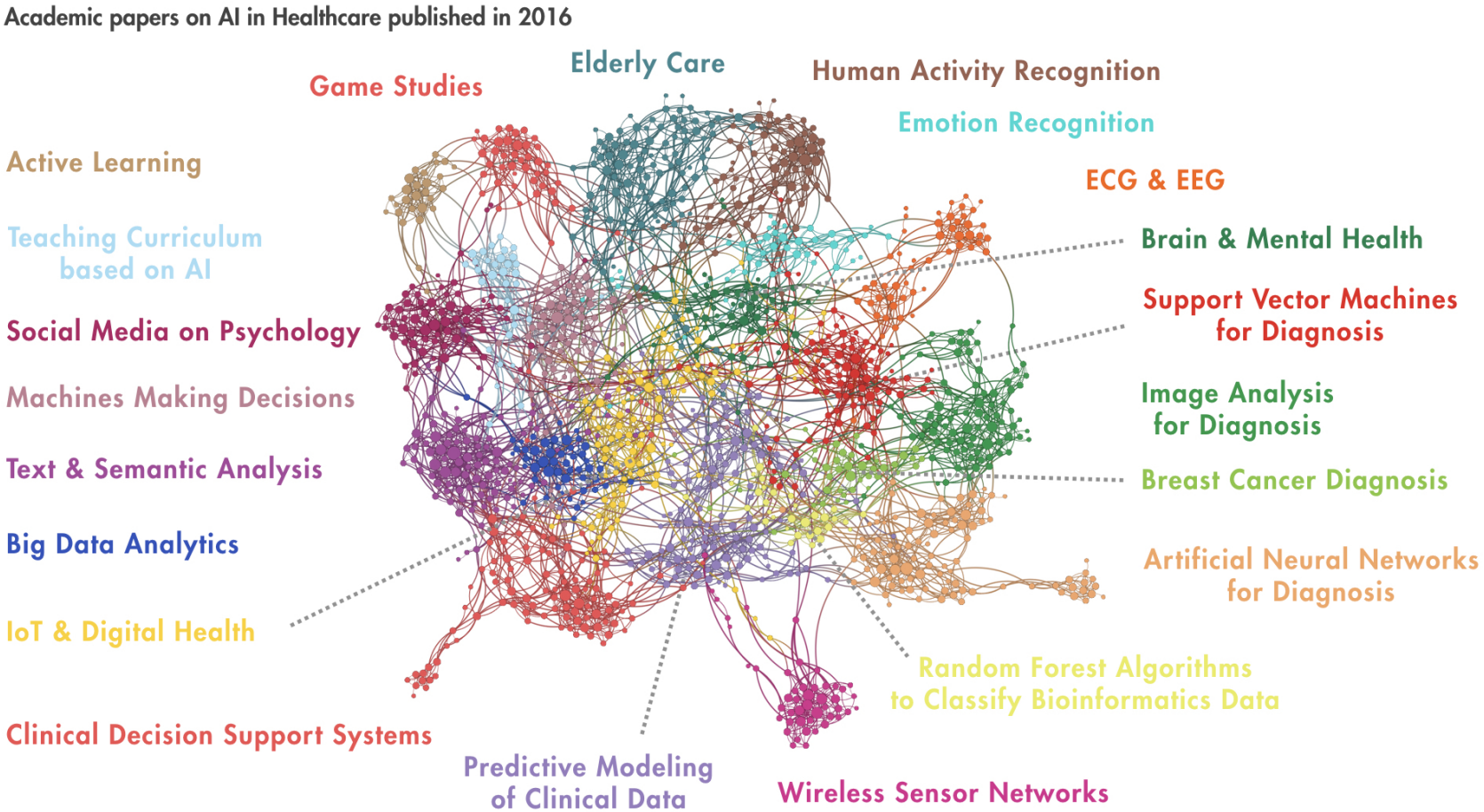
Географическая метафора: карта кластерной структуры



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

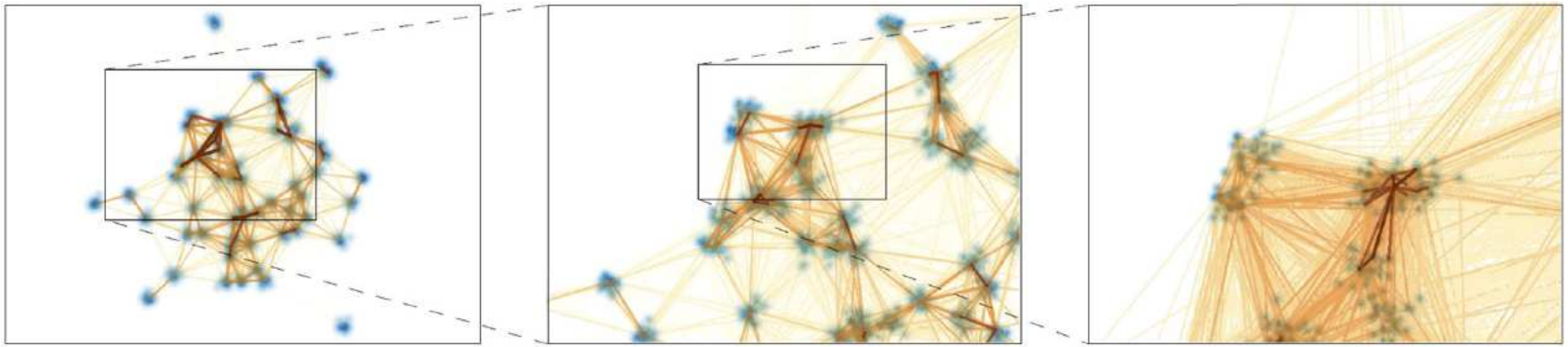
E.R.Gansner, Y.Hu, S.North. Visualizing streaming text data with dynamic maps. 2012.

Ещё одна карта тематической кластерной структуры



C.Folgar, J.McCuan. The 3 most-cited studies in healthcare and AI. Quid, 2017.

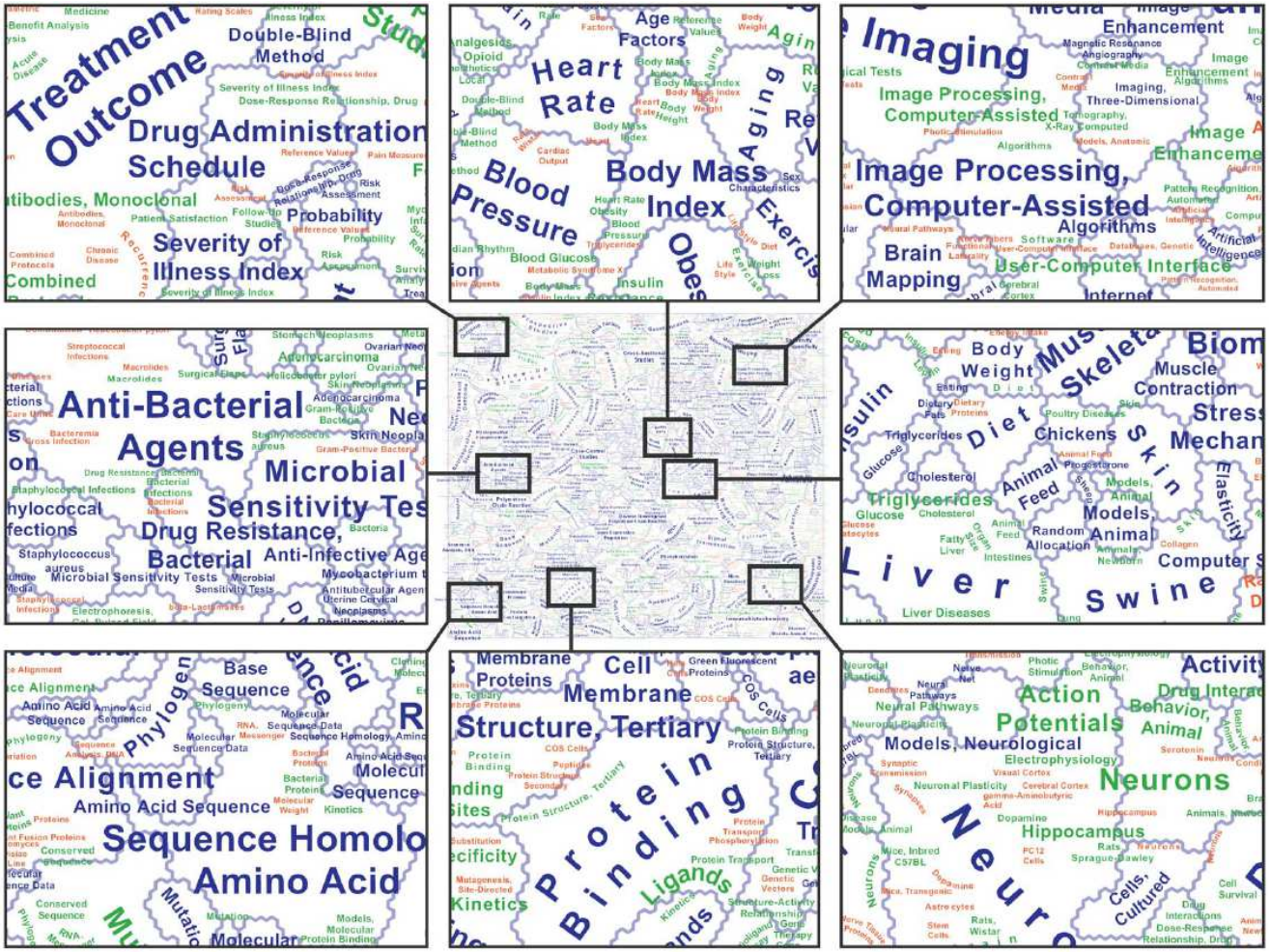
Фрактальная природа тематических кластерных структур



- Кластеры
  - кластеров
    - кластеров
      - кластеров...

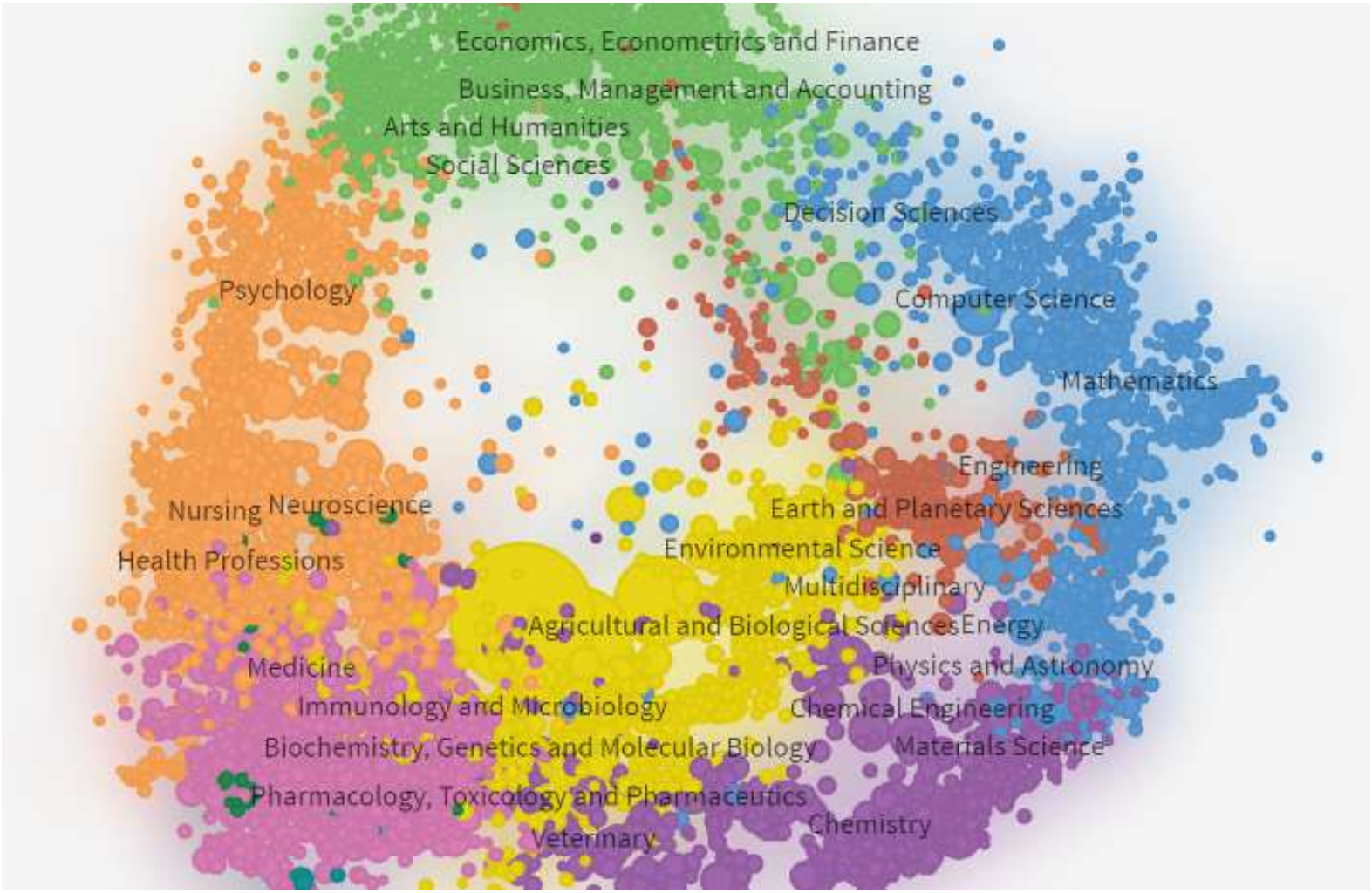
M.Zinsmaier, U.Brandes, O.Deussen, H.Strobelt. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

Пример карты медицинских знаний



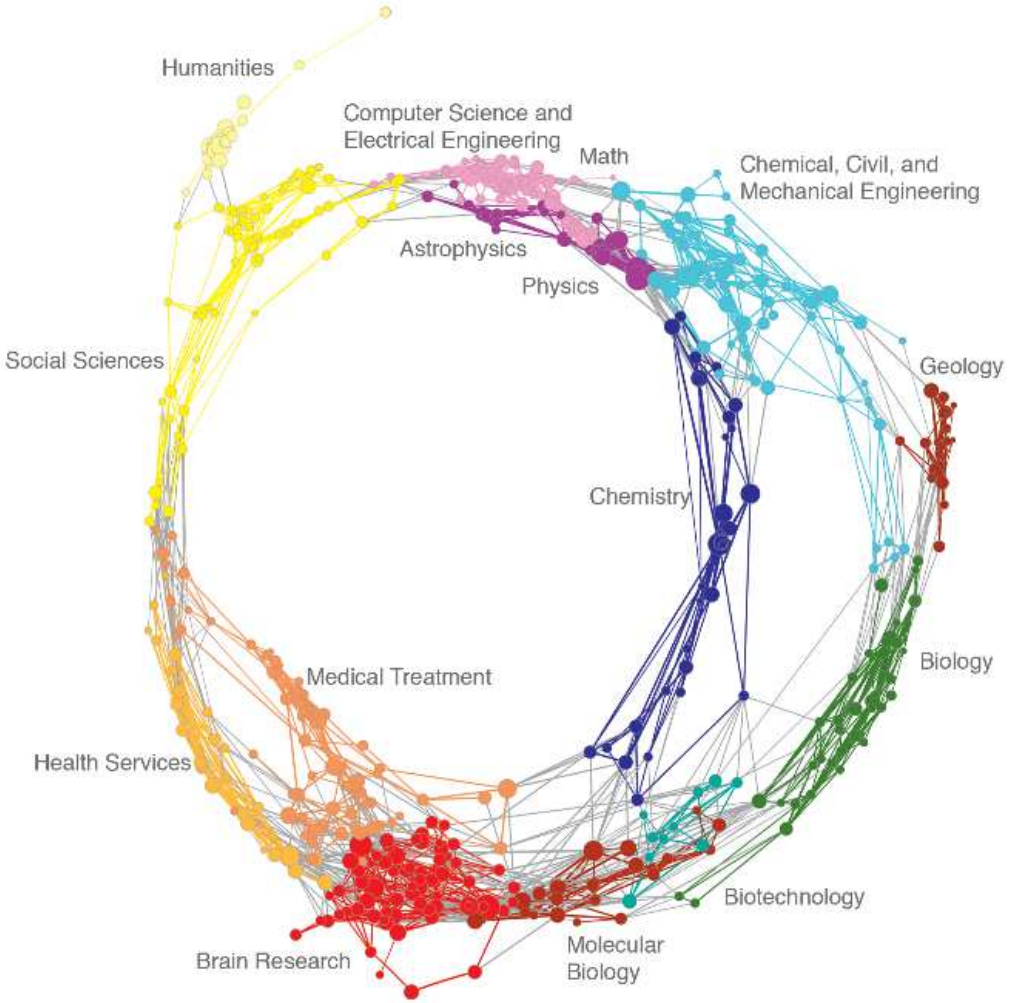
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Пример карты науки



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

Ещё один пример карты науки



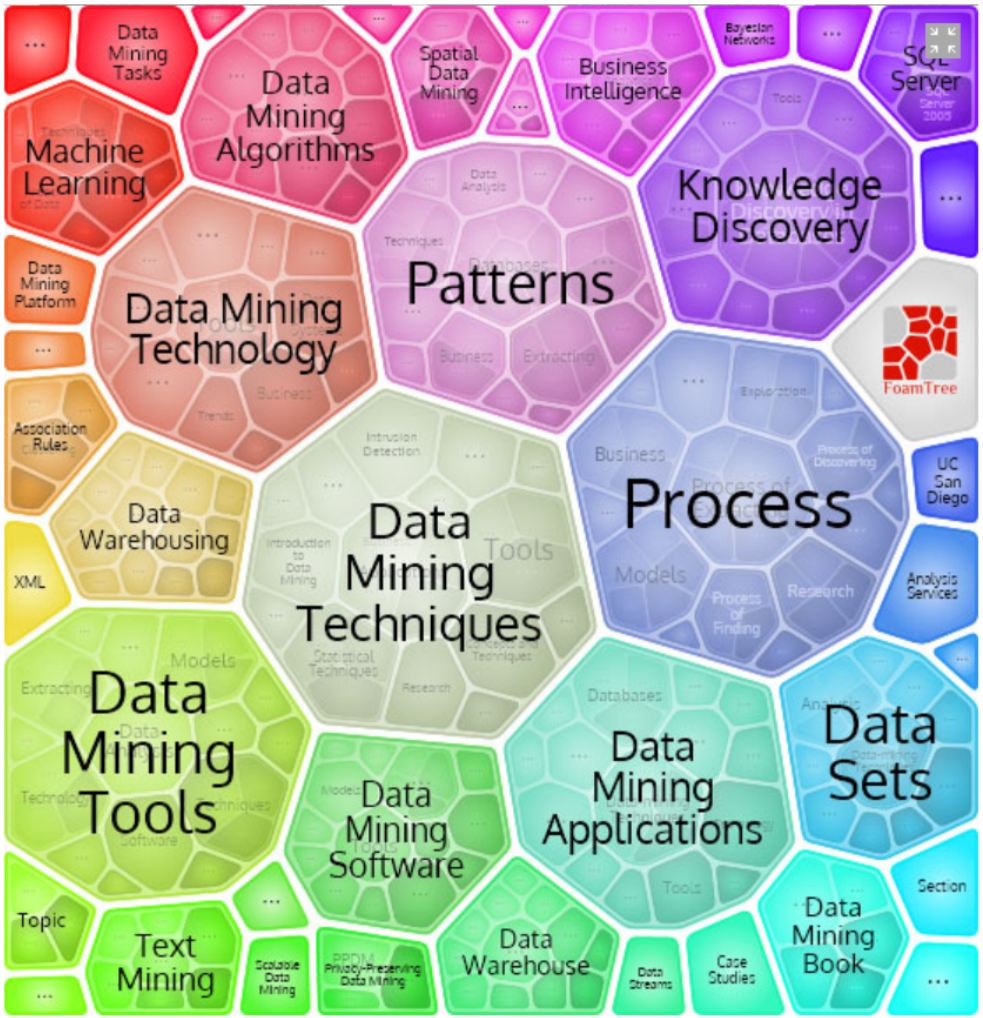
Важное наблюдение:
области знания самопроизвольно располагаются по кругу, значит,
их можно располагать и вдоль прямой линии.

Недостатки:

- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

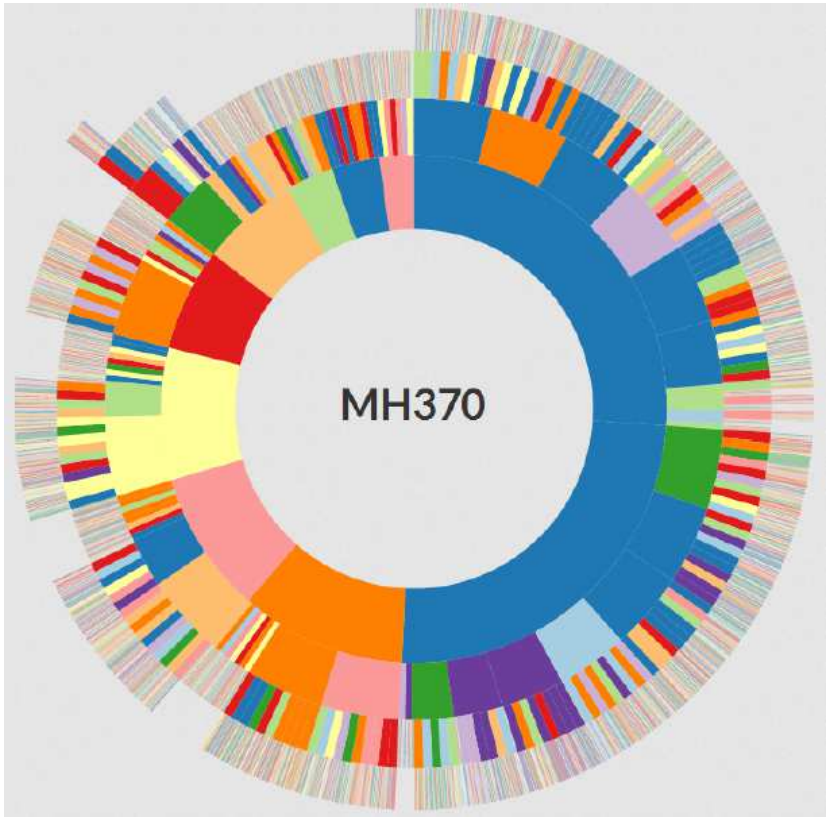
http://scimaps.org

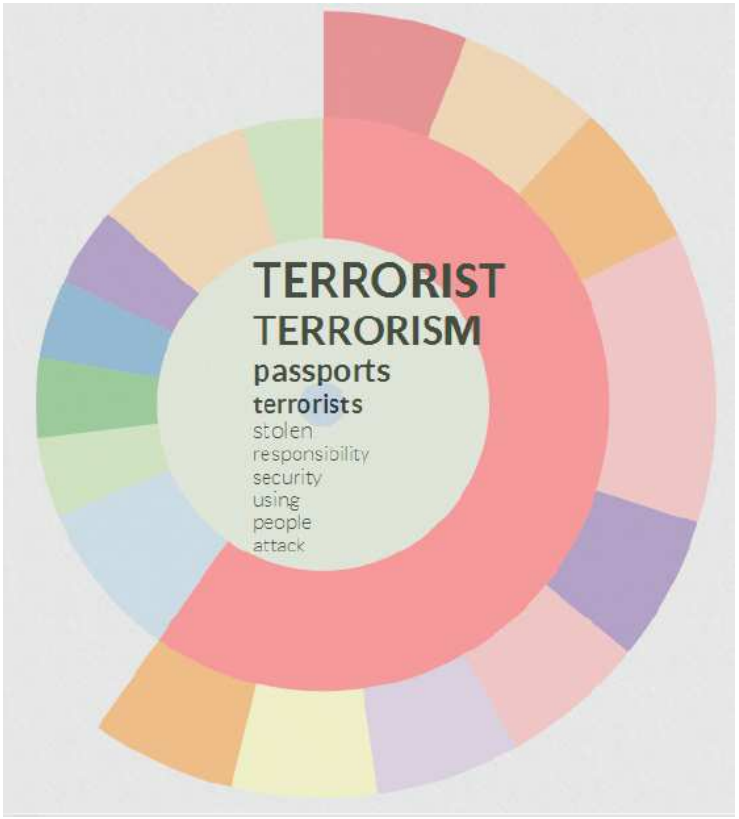
Пример иерархической карты области *Data Mining*



FoamTree: https://carrotsearch.com/foamtree

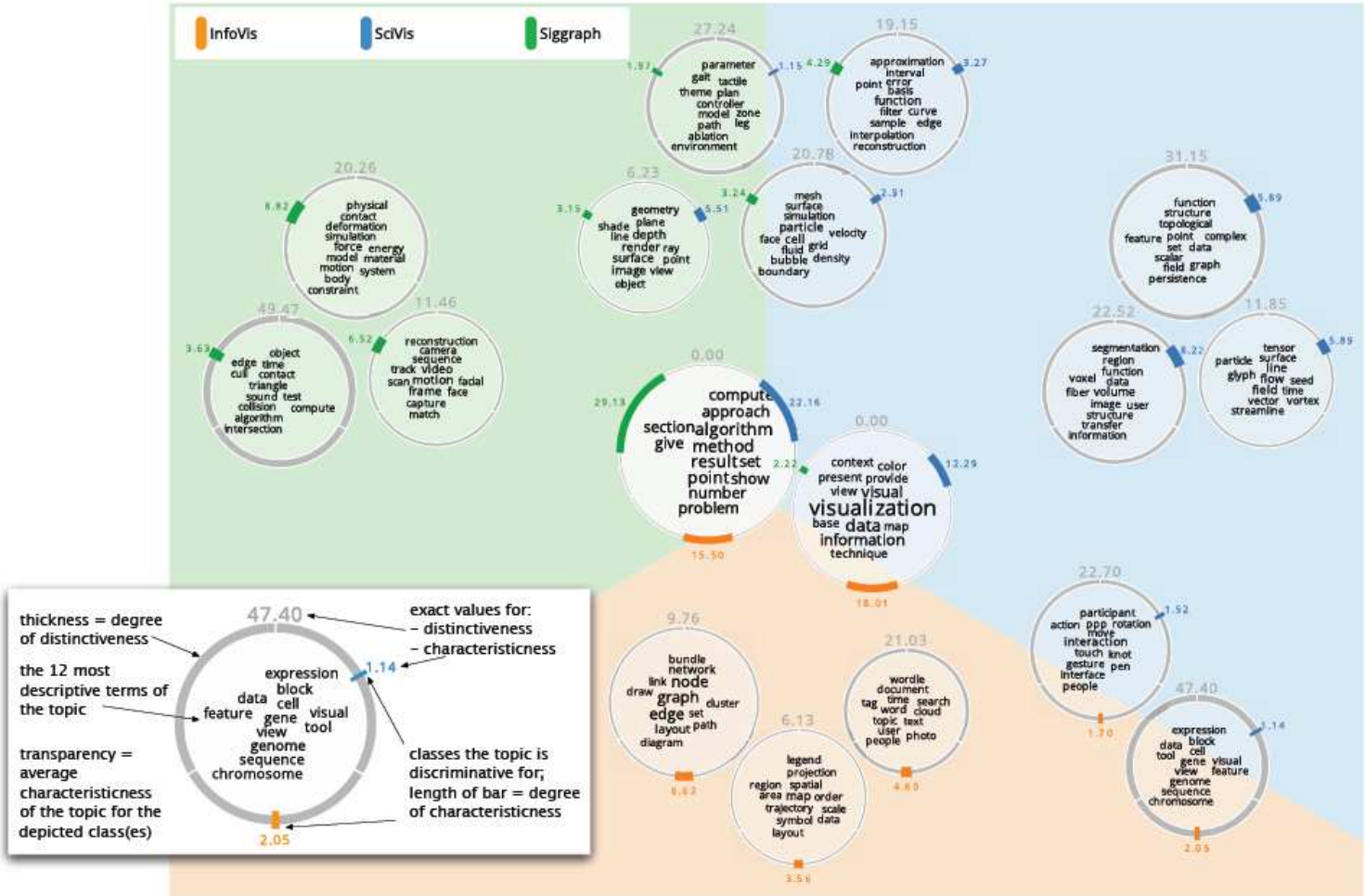
Тематическая иерархия: альтернативное представление





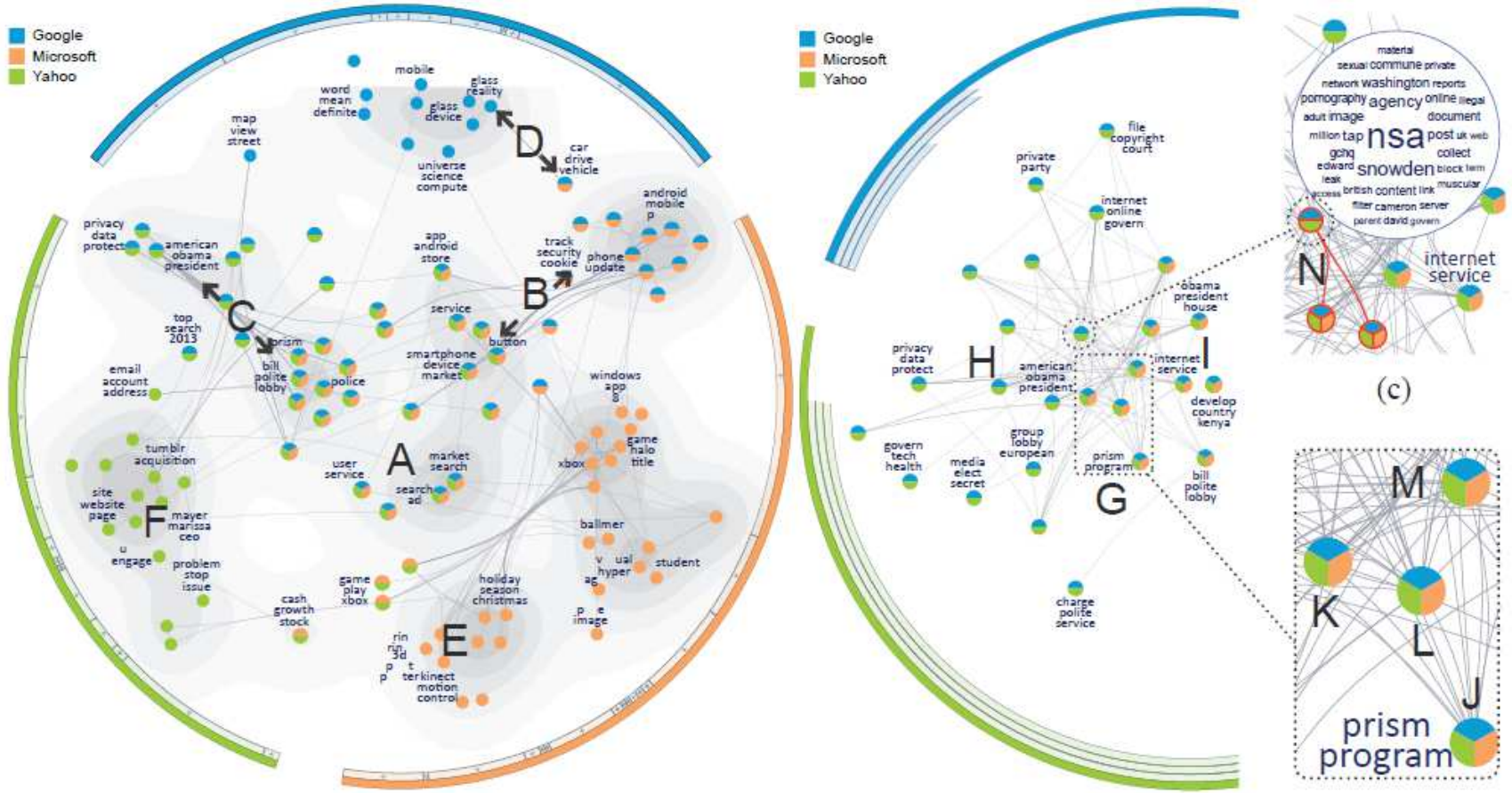
Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Тематический анализ источников



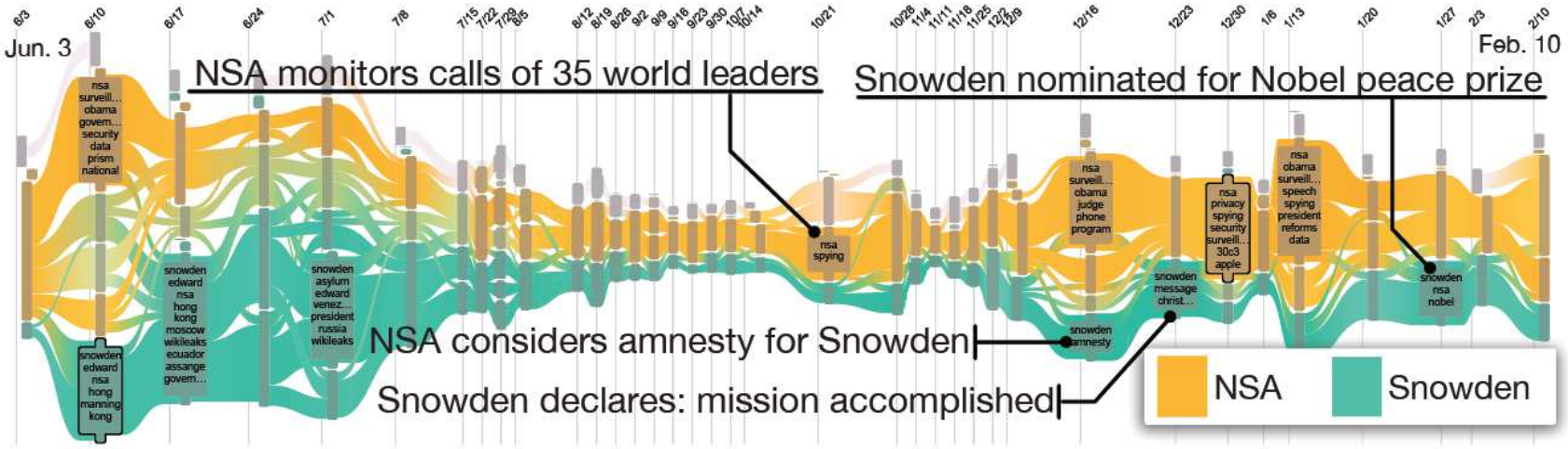
Oelke D., Strobelt H., Rohrdantz C., Gurevych I., Deussen O. Comparative exploration of document collections: a visual analytics approach. EuroVis. 2014.

Тематический анализ источников



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

Динамика тем: эволюция предметной области



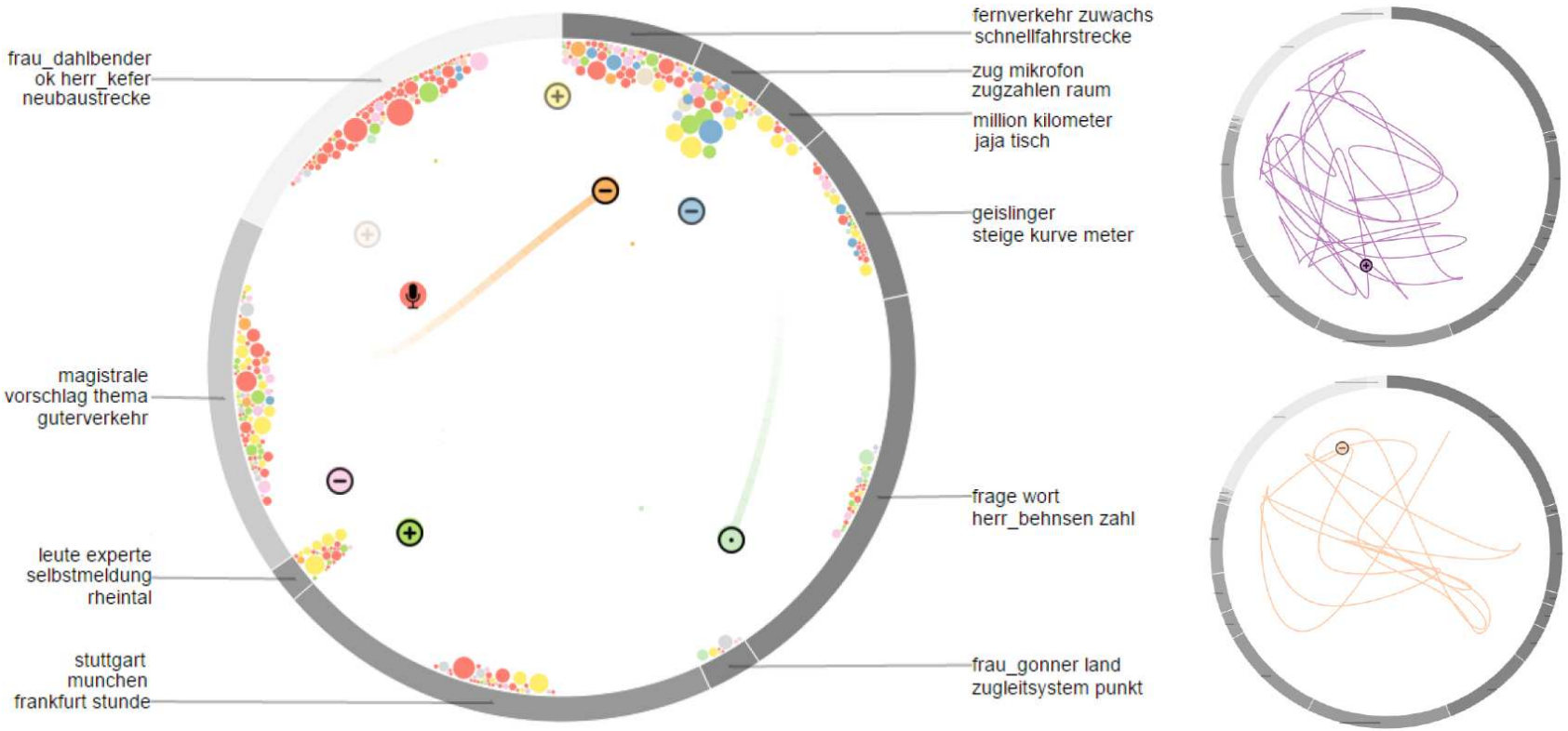
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

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

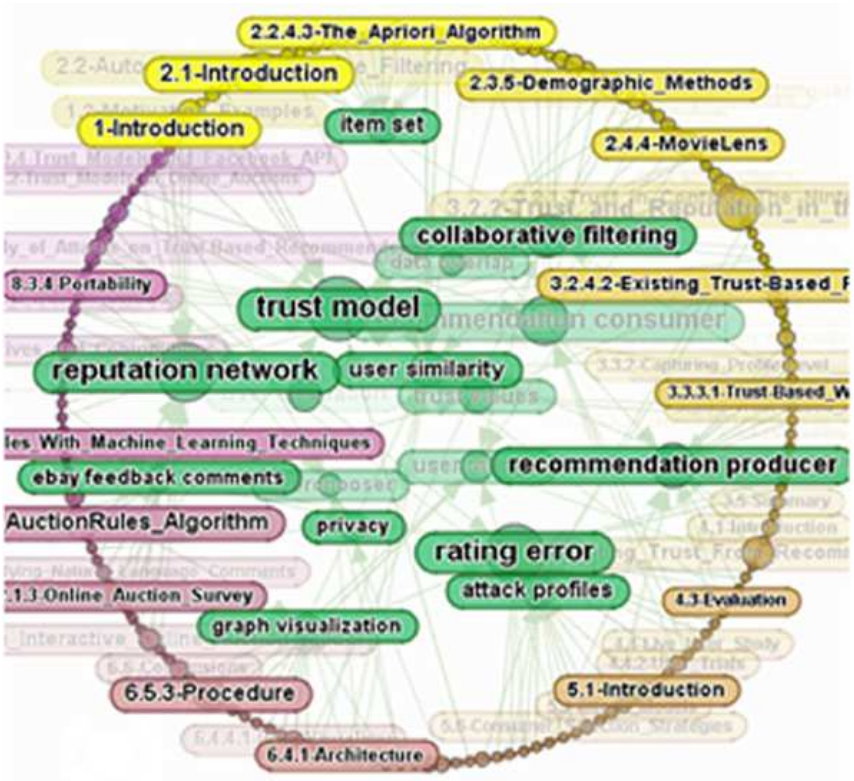
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

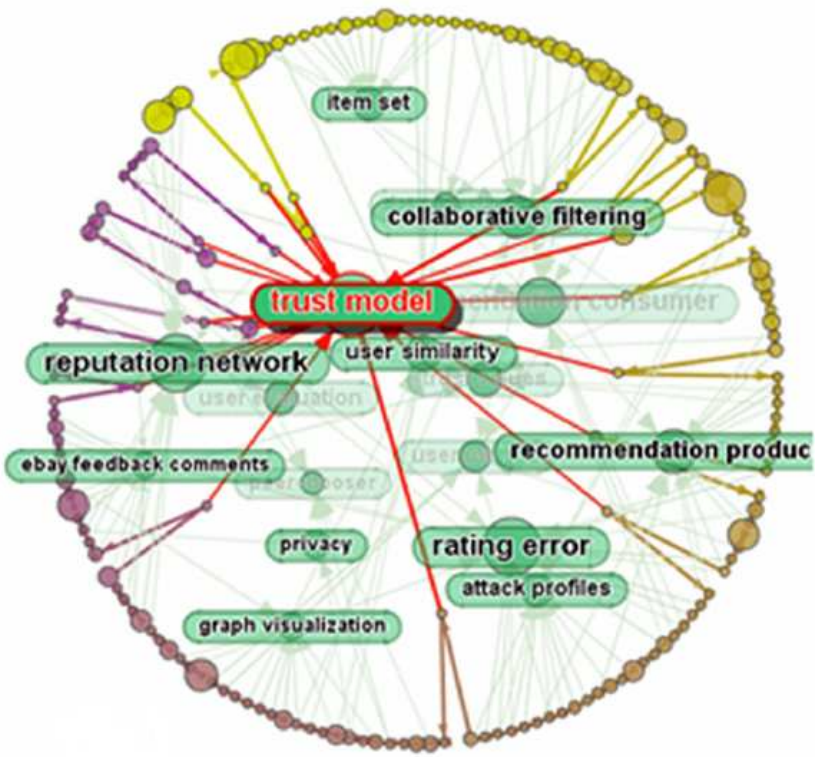
Динамика тем внутри полемического диалога



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.

http://textvis.lnu.se

Интерактивный обзор 440 средств визуализации текстов

Shixia Liu, Weiwei Cui, Yingcai Wu, Mengchen Liu. A survey on information visualization: recent advances and challenges. 2014.

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

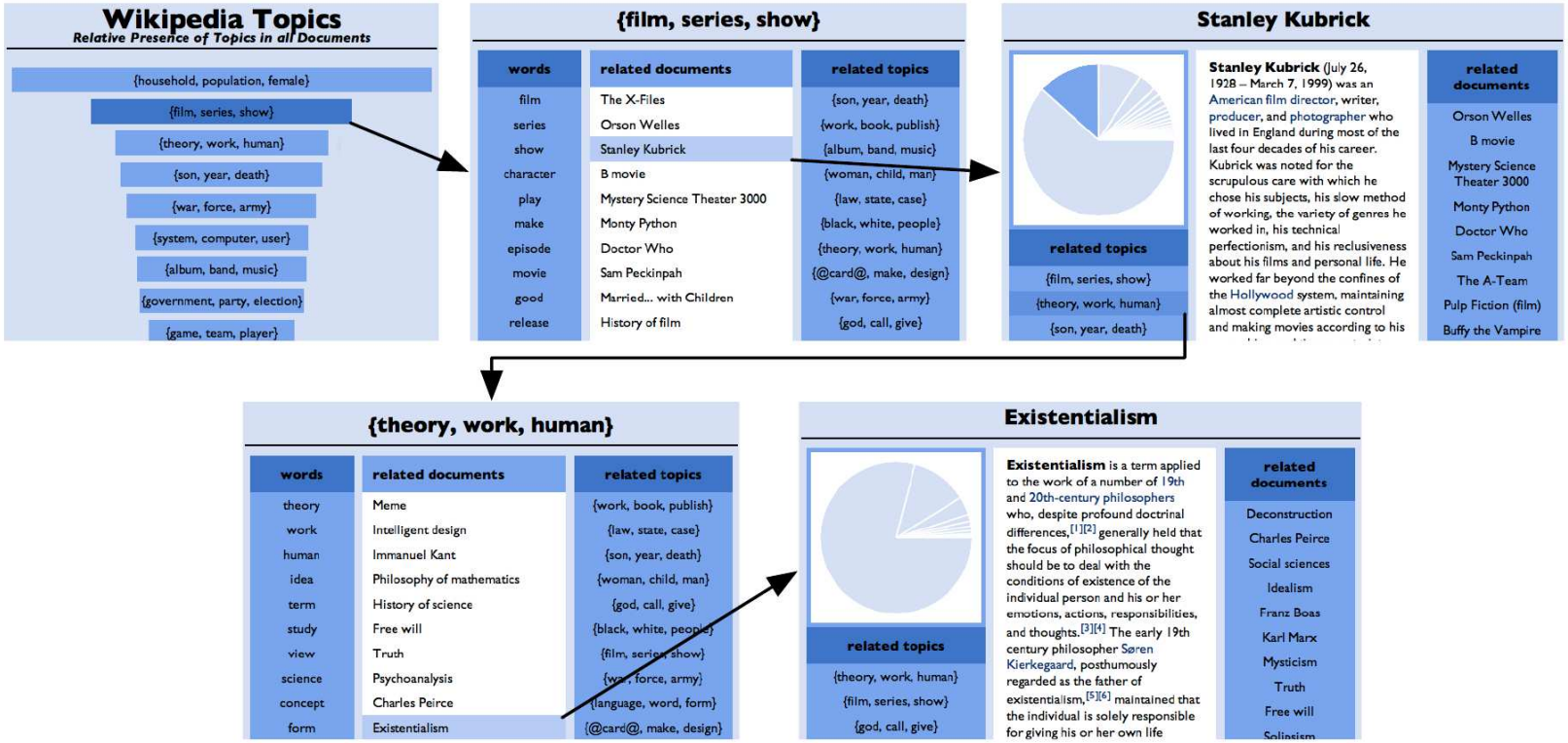
Что можно визуализировать

Одна из целей тематического моделирования — систематизация результатов информационного поиска.

- текстовое представление темы: название, топ-слова, топ-термы, топ-документы, аннотация, близкие темы
- масштабируемая тематическая карта коллекции
- иерархия тем
- граф связей между темами
- текст документа: темы слов или термов, сегментация
- графическая тематическая сегментация документа
- динамика тем во времени: временные ряды, реки тем
- иерархия + динамика

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

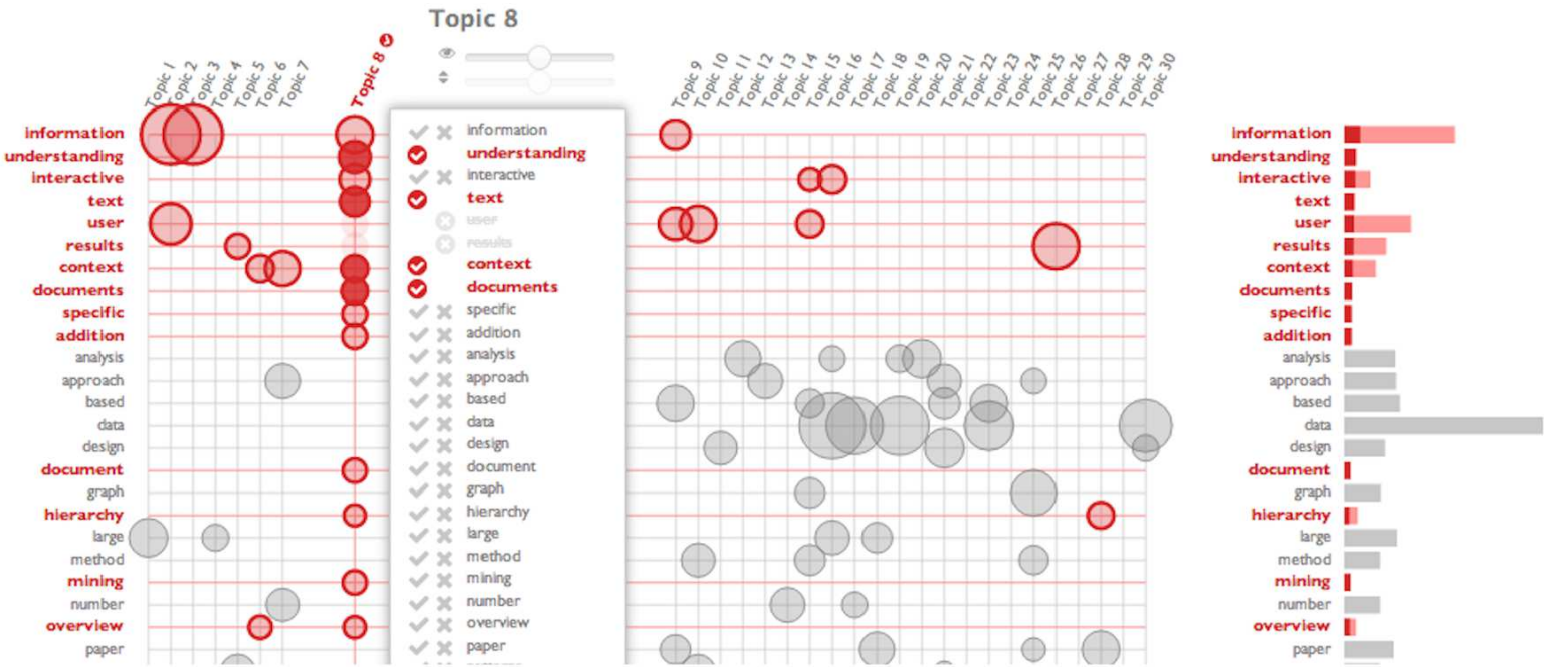


https://github.com/ajbc/tmv

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

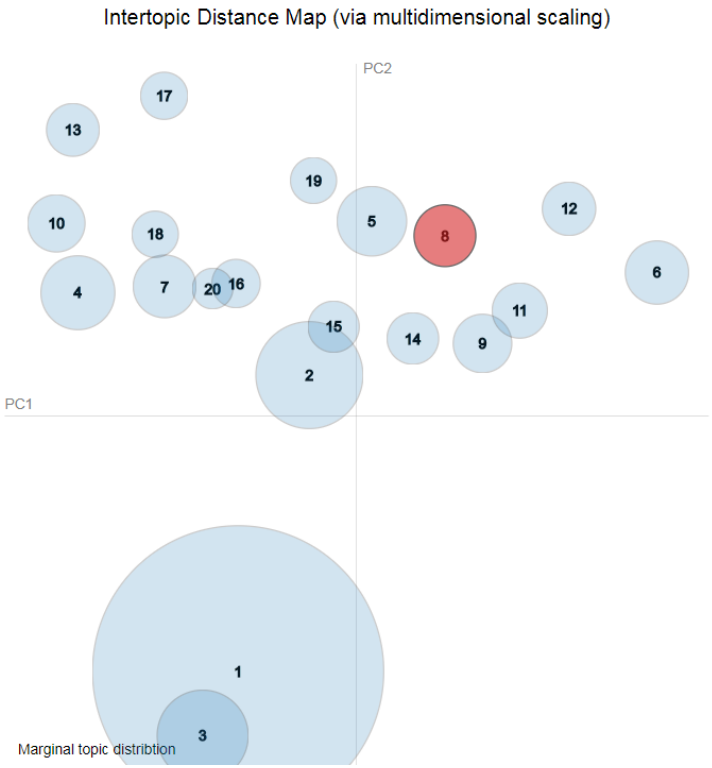


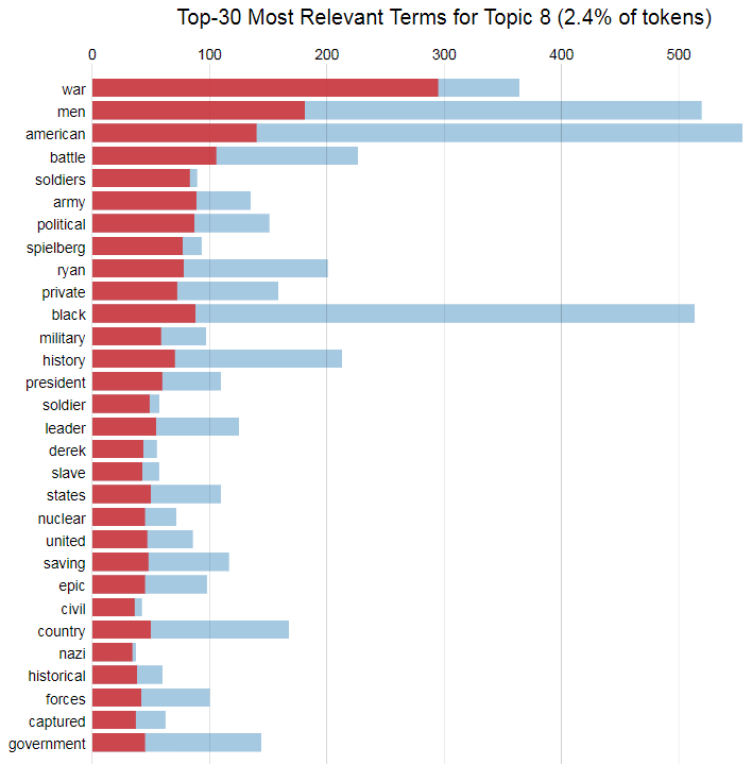
https://github.com/uwdata/termite-visualizations

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. IWCAVI 2012.

Система LDAvis

Карта сходства тем и сравнение $p(w|t)$ с $p(w)$:



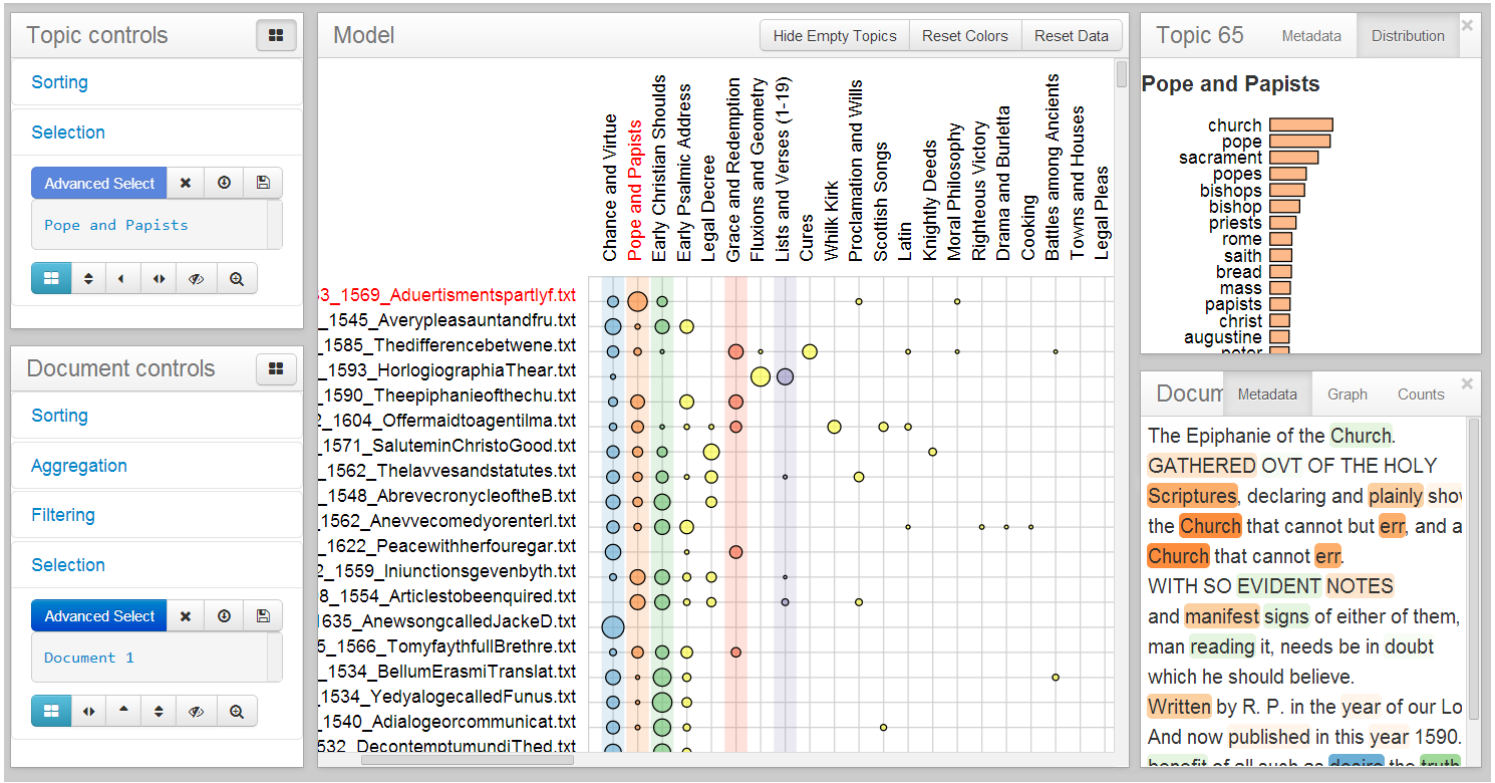


https://github.com/cpsievert/LDAvis

C.Sievert, K.Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

Система Serendip

Визуализация матриц Φ, Θ и тематики слов в текстах:



http://vep.cs.wisc.edu/serendip

E.Alexander, J.Kohlmann, R.Valenza, M.Witmore, M.Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. IEEE VAST 2014.

VisARTM: визуализация для BigARTM

- Web-приложение для визуализации ARTM моделей
- Открытый код: https://github.com/bigartm/visartm
- Автоматическое перестроение моделей через BigARTM
- Текстовые интерактивные визуализации документов, тем, термов, модальностей
- Графическая визуализация иерархических моделей
- Графическая визуализация темпоральных моделей
- Тематические спектры
- Сбор асессорских оценок

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

VisARTM: Визуализация документа

Химические коммуникации планктона

Эколог Егор Задереев о типах химических сигналов, миграциях зоопланктона и образовании покоящихся яиц

Text | Bag of words

Что исследователи знают о химической коммуникации планктона в воде? Какими сигналами обменивается зоопланктон? Как размножается зоопланктон? Об этом рассказывает кандидат биологических наук Егор Задереев.

Планктон — это организмы, местоположение которых в водной толще в основном определяется течениями. То есть это что-то маленькое, то, что переносится течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рачки. То, как водные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В наземных экосистемах, мы знаем, есть феромоны, различные сигнальные системы, которые хорошо исследованы. Мы используем их для создания ловушек, например, для вредителей — феромонные ловушки. Вода — это среда, которая благоприятна для химической коммуникации.

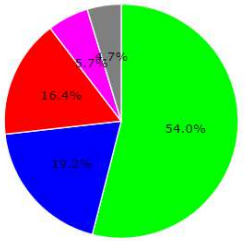
[post id="33793"]

Химические сигналы от хищников заставляют зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые ежесуточно происходят в океанах, морях и озерах. Зоопланктон ночью поднимается к поверхности, а днем уходит на глубину. Днем свет сверху помогает хищникам ловить животных, и животные уходят на глубину, а ночью поднимаются к поверхности, чтобы есть. Было показано, что эти вертикальные миграции регулируются двумя факторами. Первый — это освещенность. Очевидно, что, если не будет света, не будет сигнала. А второй — это химия, которую выделяют хищники.

В 2006 и 2009 годах выходили хорошие обзоры по химическим коммуникациям. То есть а) это очень маленькие молекулы, и б) они работают в очень низких концентрациях. Это до сих пор удивляет и поражает, потому что сообщества зоопланктона и вообще планктона в водных экосистемах — это сотни видов водорослей, рачков, которые живут в озерах, в морях, взаимодействуют между собой. А между ними есть очень сложная, судя по тому, что мы получаем в лаборатории, и разветвленная сеть химических сигналов и коммуникаций, которые влияют на разные поведенческие, физиологические и продуктивные функции. И эта сложная цепь, сеть взаимодействий до сих пор слабо исследована.

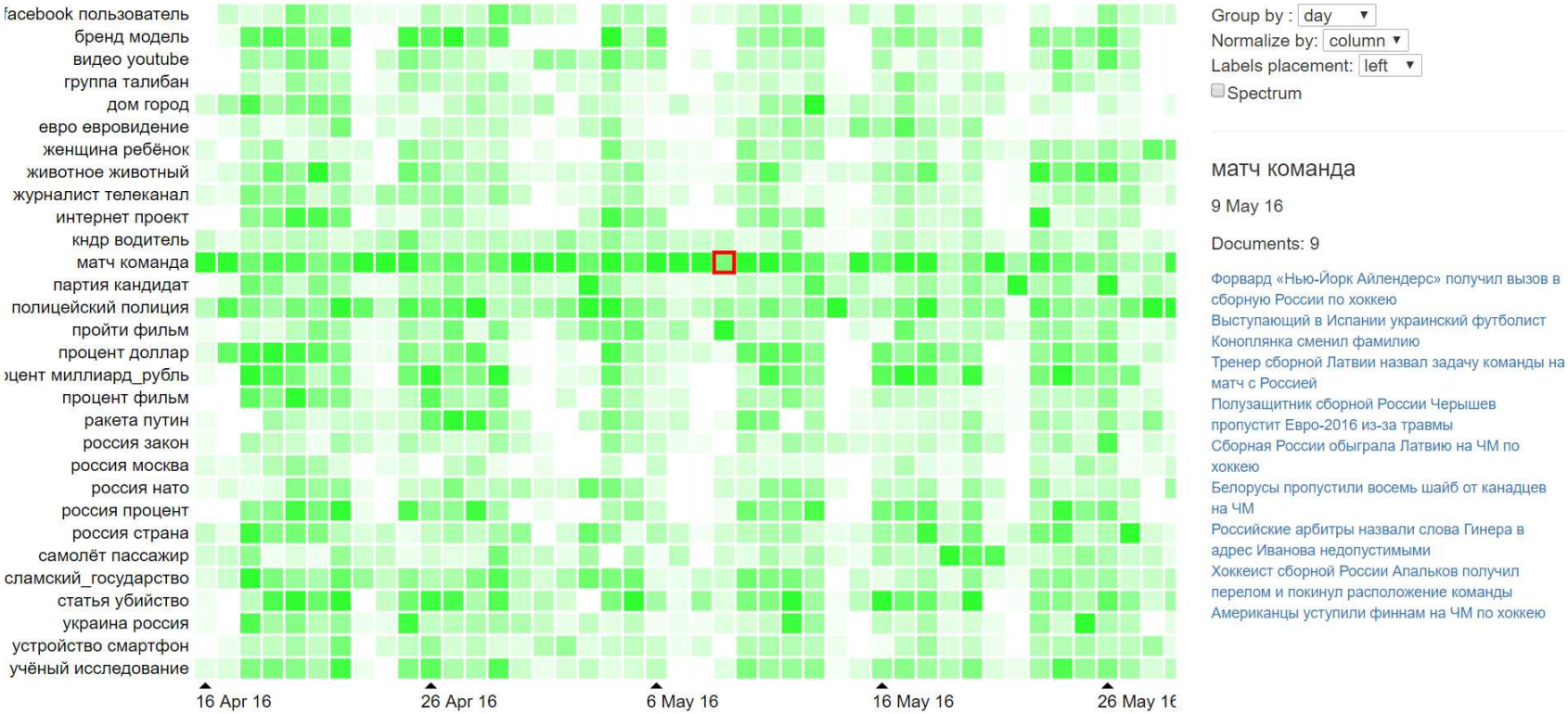
Dataset: postnauka
Time: Dec. 14, 2014, 3 p.m.
View original
index_id: 1866
text_id: 36719.txt
Terms count: 0
Unique terms count: 0
Model: flat-20
Highlighting: Words

Topic distribution

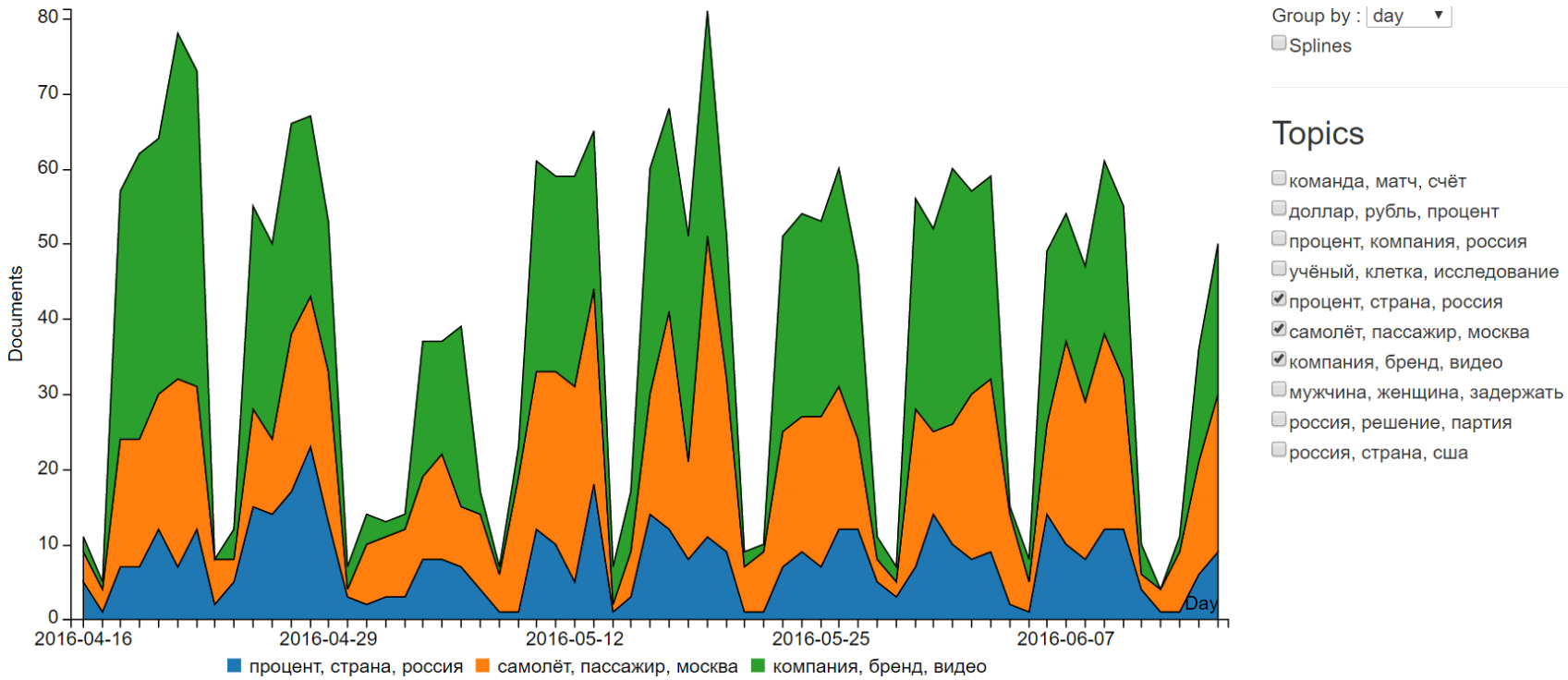


- земля, микроорганизм, вид
- вид, эволюция, ген
- материал, квантовый, структура
- город, социальный, пространство
- Other

VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация иерархической модели

Тексты научно-просветительского ресурса Postnauka.ru:
2976 документов, 43196 слов, 1799 тэгов

Belyy A.V., *Seleznova M.S.*, *Sholokhov A.K.*, *Vorontsov K.V.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Что такое «спектр тем» и зачем он нужен

Визуализация иерархии тем во времени (концепт):

- Интерпретируемые оси «время–темы»
- Близкие темы должны находиться рядом
- *Тематический спектр* — одномерная линейная проекция (например, науки: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные)

Построение спектра тем. Постановка задачи

Тематический спектр — такая перестановка тем $t_1, \ldots, t_{|T|}$, что сумма расстояний между соседними темами минимальна:

$$\sum_{i=2}^{|T|} \rho(t_i, t_{i-1}) \to \min$$

Функция расстояния $\rho(t, t')$ между темами, примеры:

- Манхэттенское: $\rho(t, t') = \sum_{w \in W} \left| \phi_{wt} - \phi_{wt'} \right|$
- Хеллингера: $\rho^2(t, t') = \frac{1}{2} \sum_{w \in W} \left(\sqrt{\phi_{wt}} - \sqrt{\phi_{wt'}} \right)^2$
- Жаккара: $\rho(t, t') = 1 - \frac{|W_t \cap W_{t'}|}{|W_t \cup W_{t'}|}$, $W_t = \left\{ w : \phi_{wt} > \frac{1}{|W|} \right\}$

Построение спектра тем — это задача коммивояжёра

Задача TSP (traveling salesman problem)

Найти путь минимальной суммарной стоимости, соединяющий T городов так, чтобы в каждом городе побывать один раз.

Алгоритм Лина–Кернигана в реализации Хельсгауна — лучший для решения задачи TSP, по данным *Encyclopedia of operations research* на 2013 год.
Вычислительная сложность $T^{2.2}$.

Другие алгоритмы оказались не только медленнее, но и хуже по качеству тематических спектров.

Keld Helsgaun. An effective implementation of the Lin–Kernighan traveling salesman heuristic. EJOR, 2000.

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

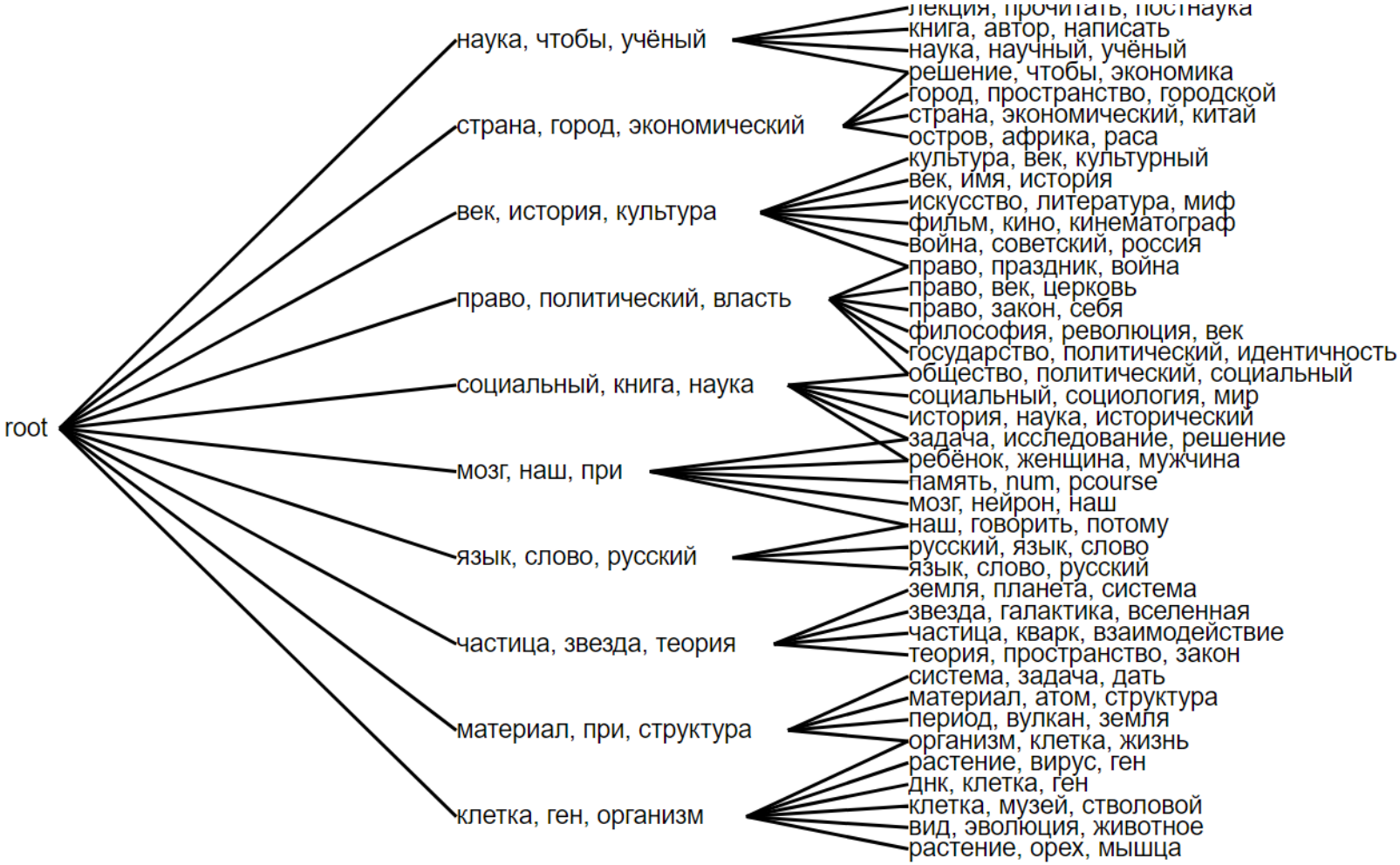
Пример спектра (коллекция postnauka.ru)

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

Пример спектра (коллекция lenta.ru)

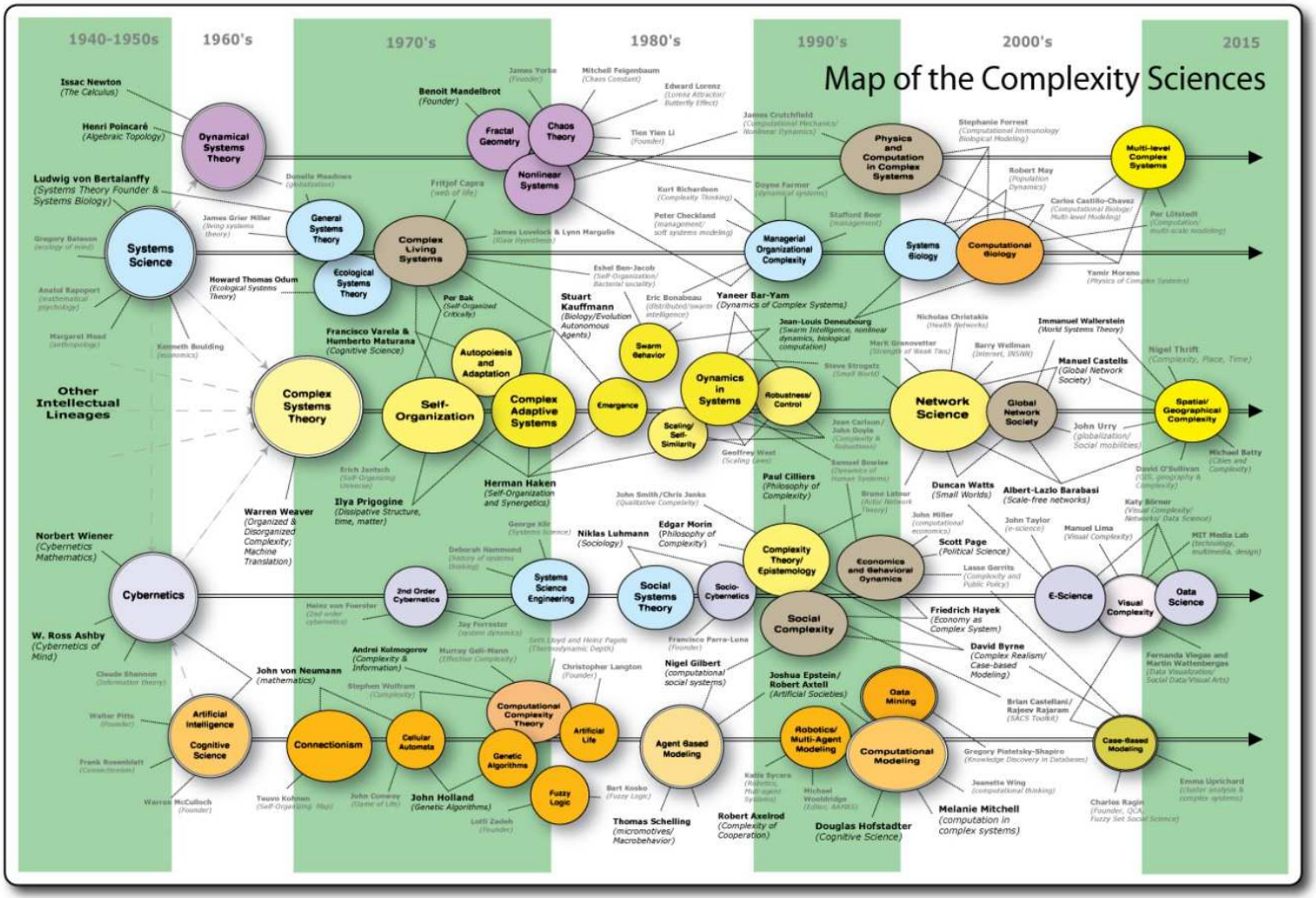
1. спортсмен, допинг, олимпиада, рио, де, россия, проба, жанейро, wada, олимпийский_игра, соревнование
2. команда, матч, счёт, клуб, победа, чемпионат, турнир, минута, футболист, встреча, летний, футбол
3. евро, евровидение, страна, россия, конкурс, франция, болельщик, англия, украина, футбол, певец
4. пройти, мероприятие, россия, акция, фестиваль, москва, фильм, участник, картина, театр, музей
5. фильм, сериал, продукт, актёр, компания, продукция, процент, россия, книга, товар, картина, сезон
6. россия, москва, турист, процент, россиянин, страна, отель, рейс, путешественник, город, тысяча
7. процент, доллар, рубль, нефть, цена, россия, баррель, страна, уровень, вырасти, рынок, рост
8. компания, миллиард_рубль, процент, миллиард_доллар, россия, сумма, миллион_доллар, банк, банка
9. закон, законопроект, документ, реклама, использование, деятельность, поправка, внести, организация
10. россия, страна, керченский_пролив, российский, боинг, работа, чайка, ряд, гражданин, аэропорт
11. партия, кандидат, журналист, праймериза, выбор, единый_россия, госдума, выборы
12. россия, украина, крым, решение, киев, депутат, вопрос, отношение, страна, мнение, право, москва
13. россия, страна, турция, сша, ес, евросоюз, москва, санкция, отношение, украина, вопрос, государство
14. россия, сирия, исламский_государство, сша, нато, иго, запретить, террорист, страна, боевик
15. ракета, путин, россия, запуск, глава_государство, союз, спутник, президент
16. учёный, клетка, исследование, исследователь, ген, университет, оказаться, процент, помощь, организм
17. земля, животное, учёный, животный, тысяча, звезда, планета, обнаружить, кошка, территория, жизнь
18. самолёт, километр, машина, борт, пассажир, вертолёт, погибнуть, лайнер, пилот, час, район, яхта
19. полицейский, полиция, мужчина, задержать, автомобиль, улица, москва, пострадать, life
20. статья, убийство, задержать, суд, отношение, ук_рф, подозревать, следствие, обвинять, трамп, часть
21. ребёнок, женщина, мужчина, летний, дом, сын, семья, мальчик, жена, полиция, дочь, школа, врач
22. видео, youtube, ролик, фото, фотография, канал, снимка, auto, instagram, девушка, страница, группа
23. facebook, пользователь, интернет, страница, twitter, пост, написать, соцсеть, вконтакте, аккаунт
24. устройство, смартфон, компания, мотоциклист, игра, байкер, видео, миллион_доллар, робот, молодая
25. бренд, модель, компания, обувь, основать, одежда, релиз, коллекция, редакция, часы, поступить

Иерархический спектр (коллекция postnauka.ru)



Иерархический спектр (коллекция lenta.ru)

Пример карты предметной области, построенной вручную



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

Тематическая карта (концепт)

- **Интерпретация осей:** время–темы или сложность–темы
- **Иерархичность:** темы делятся на подтемы
- **Спектр тем:** гуманитарные → естественные → точные
- **Интерактивность:** реализация мантры Шнейдермана
- **Суммаризация:** на карте любого масштаба много текста

Карты с интерпретируемыми осями

Точки соответствуют статьям подборки / поисковой выдачи

Числовые признаки документов, которые могут откладываться по осям XY-графика, задавать цвет, размер или стиль точек:

- дата-время публикации
- номер основной темы в спектре
- цитируемость данной статьи
- цитируемость авторов статьи
- когнитивная сложность текста (см. далее)
- узость/обзорность — по энтропии распределения $p(t|d)$
- центральность — число тематически схожих статей
- новизна — по распределениям $p(\text{время}|t)$, $p(\text{время}|w)$
- хайповость — по упоминаниям в социальных сетях

Оценивание когнитивной сложности текста

Основные предположения:

- *уровни языка*: фонетический, морфологический, лексический, синтаксический, дискурсивный
- на уровне i текст может быть представлен в виде последовательности *токенов* алфавита A_i
- *сложность текста* на уровне i — это доля токенов, имеющих аномально высокую частоту
- частота токена *аномально высокая*, если она превышает 95%-ю квантиль его частоты в референтном корпусе
- *референтный корпус* — тексты, не являющиеся сложными для выбранной читательской аудитории (в частности, тема)

M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

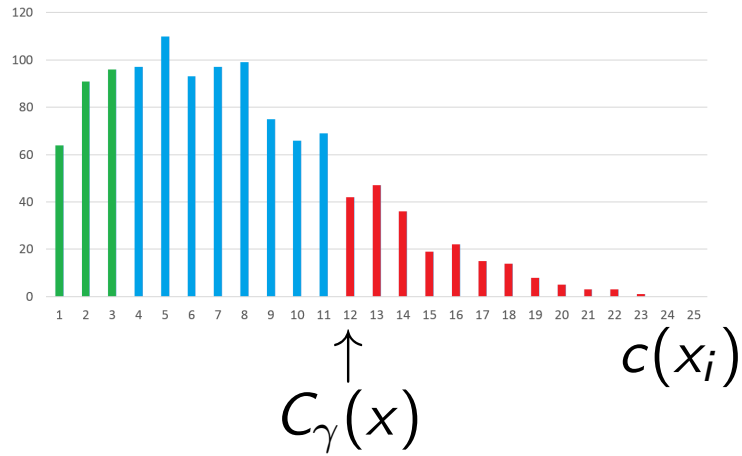
Квантильный подход к оцениванию сложности текста

$x_1, \ldots, x_{n_d}$ — последовательность токенов текста d;
$c(x_i) = \left(\bar{r}(x_i) - r_i\right)$ — оценка сложности токена x_i, где
r_i — расстояние от токена x_i до его предыдущего вхождения,
$\bar{r}(x)$ — среднее r_i токена x в *референтном корпусе*.

Оценка сложности текста — доля сложных токенов в тексте:

$$C(d) = \frac{1}{n_d} \sum_{i=1}^{n_d} [c(x_i) > C_\gamma(x_i)],$$

$C_\gamma(x)$ — γ-квантиль распределения сложности токена x в *референтном корпусе* несложных текстов



А.А.Биркин. Природа речи. М.: Ликбез, 2009.

Обучаемая линейная модель когнитивной сложности текста

Пусть $C_k(d), k = 1, \ldots, K$ — различные оценки сложности.

Линейная агрегированная оценка сложности с параметрами α_k:

$$C(d, \alpha) = \sum_{k=1}^{K} \alpha_k C_k(d), \quad \alpha_k \geqslant 0.$$

Данные экспертного сравнения пар документов:
$d \prec d'$ — документ d' сложнее документа.

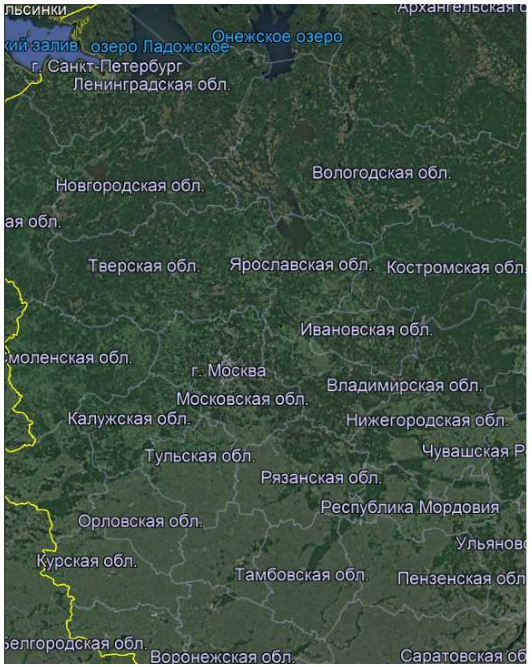
Критерий обучения агрегированной оценки:

$$\sum_{d \prec d'} \mathscr{L}\big(C(d', \alpha) - C(d, \alpha)\big) \to \min_{\alpha},$$

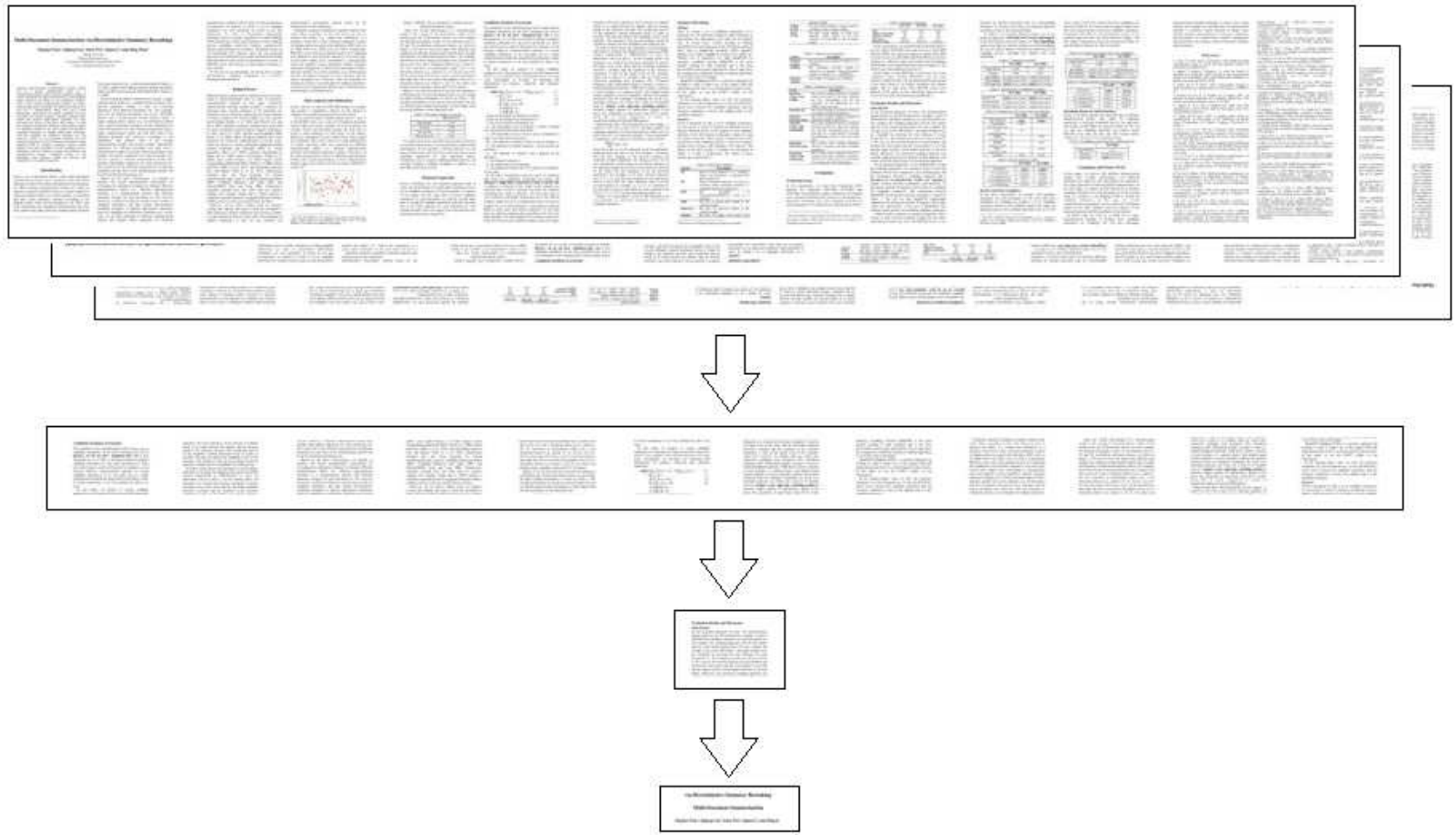
где $\mathscr{L}(M)$ — гладкая невозрастающая функция отступа M.

Аналогия с геоинформационными системами

Если представить, что вся Земля плотно покрыта текстами, написанными человечеством, то с любой высоты мы увидим суммаризацию множества текстов, попадающих в поле зрения.



Детальность суммаризации определяется бюджетом места



Вместо резюме

Эволюция информационного поиска:

- книги, библиотеки — долго искать, долго понимать
- поисковые машины — быстро искать, долго понимать
- визуальный поиск — быстро искать, быстро понимать