Тематическое моделирование (часть 2)

К. В. Воронцов
vokov@forecsys.ru

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

ШАД Яндекс • 3 ноября 2015

Содержание

Постановка задачи

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} — сколько раз термин w встретился в документе d

n_d — длина документа d

Найти: параметры модели $\frac{n_{dw}}{n_d} \approx p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения

Она *некорректно поставлена*, т. к. её решение не единственно:

$$\underset{W\times D}{\left(\frac{n_{dw}}{n_d}\right)} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta} = (\Phi S)(S^{-1}\Theta) = \underset{W\times T}{\Phi'} \cdot \underset{T\times D}{\Theta'}$$

для невырожденных $\underset{T\times T}{S}$ таких, что Φ', Θ' тоже стохастические.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации логарифма правдоподобия:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

$$\text{E-шаг:}\quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt}\right), \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td}\right), \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw} \end{cases}$$

где $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

LDA — Latent Dirichlet Allocation [Blei, 2003]

Максимизация апостериорной вероятности ($\beta_w, \alpha_t > -1$):

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\ln \text{ правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \beta_w\right), & n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \alpha_t\right), & n_{td} = \sum_{w\in d} n_{dw} p_{tdw} \end{cases}$$

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация ln правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right), & n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right), & n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

Комбинирование регуляризованных тематических моделей

Максимизация ln правдоподобия с n регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \phi_{wt}}\right), & n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \theta_{td}}\right), & n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw} \end{cases}$$

Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы ln правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$p_{tdw} = \underset{t \in T}{\operatorname{norm}} \left(\phi_{wt} \theta_{td} \right)$$

M-шаг:
$$\phi_{wt} = \underset{w \in W^m}{\operatorname{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw}$$

Регуляризатор для классификации и категоризации текстов

Цель: построить тематическую модель классификации.

Y — множество классов;

$n_{dy} = [\text{документ } d \text{ относится к классу } y]$ — обучающие данные;

$p(y|d) = \sum_{t \in T} \phi_{yt}\theta_{td}$ — линейная модель классификации.

Регуляризатор — правдоподобие модальности классов:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{y \in Y} n_{dy} \ln \sum_{t \in T} \phi_{yt}\theta_{td} \ \to \ \max,$$

это тематическая модель с двумя модальностями, W и Y.

TM превосходит SVM в случае несбалансированных классов.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор для задач регрессии

Цель: построить тематическую модель регрессии.

$y_d \in \mathbb{R}$ для всех документов $d \in D$ — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — линейная модель регрессии, $v \in \mathbb{R}^{|T|}$.

Регуляризатор — среднеквадратичная ошибка (МНК):

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2 \to \max$$

Подставляем, получаем формулы M-шага:

$$\theta_{td} = \operatorname*{norm}_t \left(n_{td} + \tau \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.

Сглаживание, разреживание и частичное обучение

Цель: обобщить LDA для разреживания и частичного обучения, взяв параметры $\beta_{wt}, \alpha_{td} \in \mathbb{R}$ вместо $\beta_w, \alpha_t > -1$:

$$R(\Phi, \Theta) = \sum_{t \in B} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in B} \alpha_{td} \ln \theta_{td} \to \max.$$

Подставляем, получаем формулы M-шага LDA, для всех $t \in B$:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_{wt}\big), \qquad \theta_{td} = \underset{w}{\text{norm}}\big(n_{td} + \alpha_{td}\big),$$

$\beta_{wt} > 0$ — сглаживание, термин w в «белом списке» темы t,
$\beta_{wt} < 0$ — разреживание, термин w в «чёрном списке» темы t,
$\alpha_{td} > 0$ — сглаживание, тема t в «белом списке» документа d,
$\alpha_{td} < 0$ — разреживание, тема t в «чёрном списке» документа d.

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Volume 101, Issue 1 (2015), Pp. 303-323.

Регуляризатор декоррелирования тем

Цель: усилить различность тем, выделить в каждой теме лексическое ядро, отличающее её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} = \underset{w}{\text{norm}} \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор для максимизации когерентности тем

Цель: сгруппировать в одни темы те термины $u, w \in W$, которые часто встречающиеся рядом.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\phi_{wt} = \underset{w}{\text{norm}}\left(n_{wt} + \tau \sum_{u\in W\setminus w} C_{uw} n_{ut}\right).$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор для учёта связей между документами

Цель: улучшить темы, используя ссылки или цитирования (если документы ссылаются друг на друга, то их темы близки):

n_{dc} — число ссылок из d на c.

Максимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \operatorname{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} = \operatorname*{norm}_t \left(n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc} \right).$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор для темпоральных тематических моделей

Цель: найти событийные и перманентные темы.

Y — моменты времени (например, годы публикаций),

$y(d)$ — метка времени документа d,

$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Разреживание тем $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ в каждый момент y:

$$R_1(\Theta) = \tau_1 \sum_{y \in Y} \mathrm{KL}\left(\tfrac{1}{|T|} \| p(t|y)\right) \to \max.$$

Сглаживание тем $p(y|t)$ в соседние моменты y, $y-1$:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left| p(y|t) - p(y-1|t) \right| \to \max.$$

Регуляризатор для сокращения числа тем

Цель: исключить незначимые темы.

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя KL-дивергенцию между $p(t)$ и равномерным распределением:

$$R(\Theta) = \tau \sum_{t \in S} \mathrm{KL}\left(\tfrac{1}{|T|} \| p(t)\right) \to \max.$$

Подставляем, получаем формулу M-шага:

$$\theta_{td} = \underset{t}{\mathrm{norm}}\left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right).$$

Эффект: строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

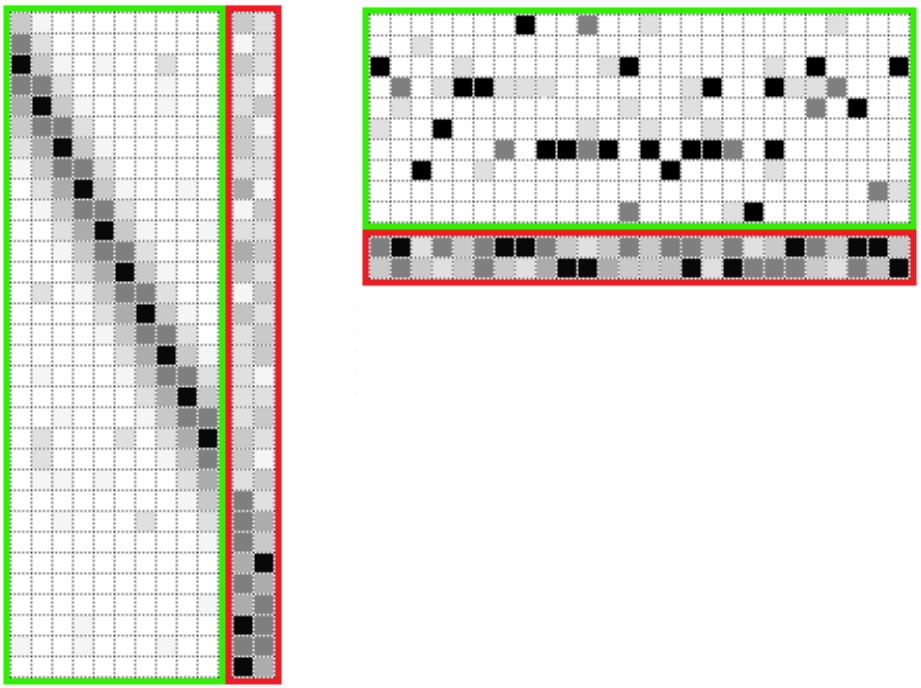
Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // SLDS 2015.

Требования интерпретируемости и гипотезы о структуре тем

Предметные темы S содержат термины предметной области, $p(w|t)$ разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, $p(w|t)$ и $p(t|d)$ не разреженные в этих темах

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Разреживание + Сглаживание + Декорреляция

M-шаг при комбинировании 6 регуляризаторов:

$$\phi_{wt} = \underset{w}{\text{norm}}\Big(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_3 \phi_{wt} \underbrace{\sum_{s \in S \setminus t} \phi_{ws}}_{\text{декорреляция}} \Big)$$

$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\ \text{малых тем}}} \Big)$$

Траектория регуляризации (*regularization path*) в пространстве $\tau = (\tau_1, \ldots, \tau_6)$ подбирается экспериментальным путём.

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

Стандартная методика оценивания моделей языка

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}\ln p(w|d)}{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}}\right),$$

$d = d' \sqcup d''$ — случайное разбиение тестового документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D;
параметры θ_{td} оцениваются по первой половине d';
перплексия вычисляется по второй половине d''.

Интерпретации перплексии:
1) $\mathcal{P}(D') \to |d''|$ при $n \to \infty$, если слова равновероятны;
2) насколько хорошо мы предсказываем слова в документах (чем меньше перплексия, тем лучше).

Проверка гипотезы условной независимости

Гипотеза условной независимости: $p(w|d,t) = p(w|t)$.

Отклонение распределения $p(w|d,t) = \frac{n_{dwt}}{n_{dt}}$ от $p(w|t) = \frac{n_{wt}}{n_t}$:

$$S_{dt} = \mathrm{KL}_w\Big(p(w|d,t) \,\Big\|\, p(w|t)\Big) = \sum_{w \in d} \frac{n_{dwt}}{n_{dt}} \ln \frac{n_{dwt} n_t}{n_{dt} n_{wt}},$$

где $n_{dwt} = n_{dw} p(t|d,w)$ — результат E-шага.

Среднее по документам отклонение темы t от общей по коллекции:

$$S_t = \sum_{d \in D} p(d|t) S_{dt} = \sum_{d \in D} \frac{n_{dt}}{n_t} S_{dt} =$$

$$= \mathrm{KL}_{d,w}\Big(p(d,w|t) \,\Big\|\, p(d|t) \cdot p(w|t)\Big) = \sum_{d,w} \frac{n_{dwt}}{n_t} \ln \frac{n_{dwt} n_t}{n_{dt} n_{wt}}.$$

David Mimno, David Blei. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011.

Оценка интерпретируемости темы: когерентность

Когерентность темы t — средняя поточечная взаимная информация топ-слов темы (pointwise mutual information, PMI):

$$\text{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \text{PMI}(w_i, w_j),$$

где w_i — i-й термин в порядке убывания ϕ_{wt}, $k = 10$.

$\text{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация.

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

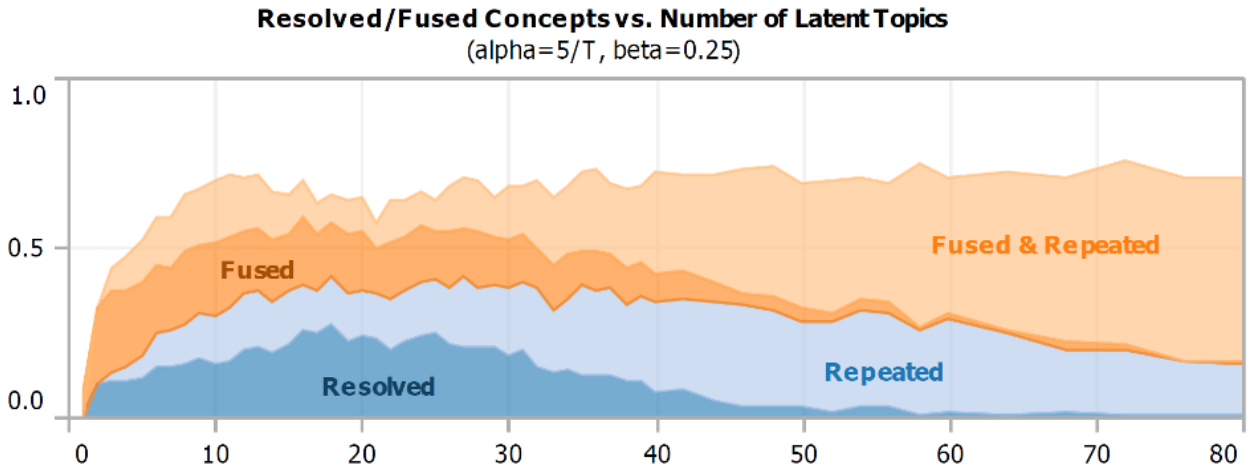
Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Оценки разреженности темы

- Разреженность:
  - доля нулевых элементов в Φ
  - доля нулевых элементов в Θ
- Характеристики интерпретируемости тем:
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - чистота темы: $\sum_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - доля фоновых слов: $\frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t|d,w)$
  - доля нетематичных документов: $\frac{1}{|D|} \sum_{d \in D} \Big[\sum_{t \in B} p(t|d) > 0.95 \Big]$
  - доля нетематичных терминов: $\frac{1}{|W|} \sum_{w \in W} \Big[\sum_{t \in B} p(t|w) > 0.95 \Big]$

Внешние критерии

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество категоризации документов
- Экспертное оценивание тем *методом интрузий*
- Точность соответствия тем заданным *концептам*
  (число ненайденных и расщеплённых тем и концептов)



Chuang J., Gupta S., Manning C., Heer J. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment // ICML-2013.

Интерактивный анализ качества тематической модели

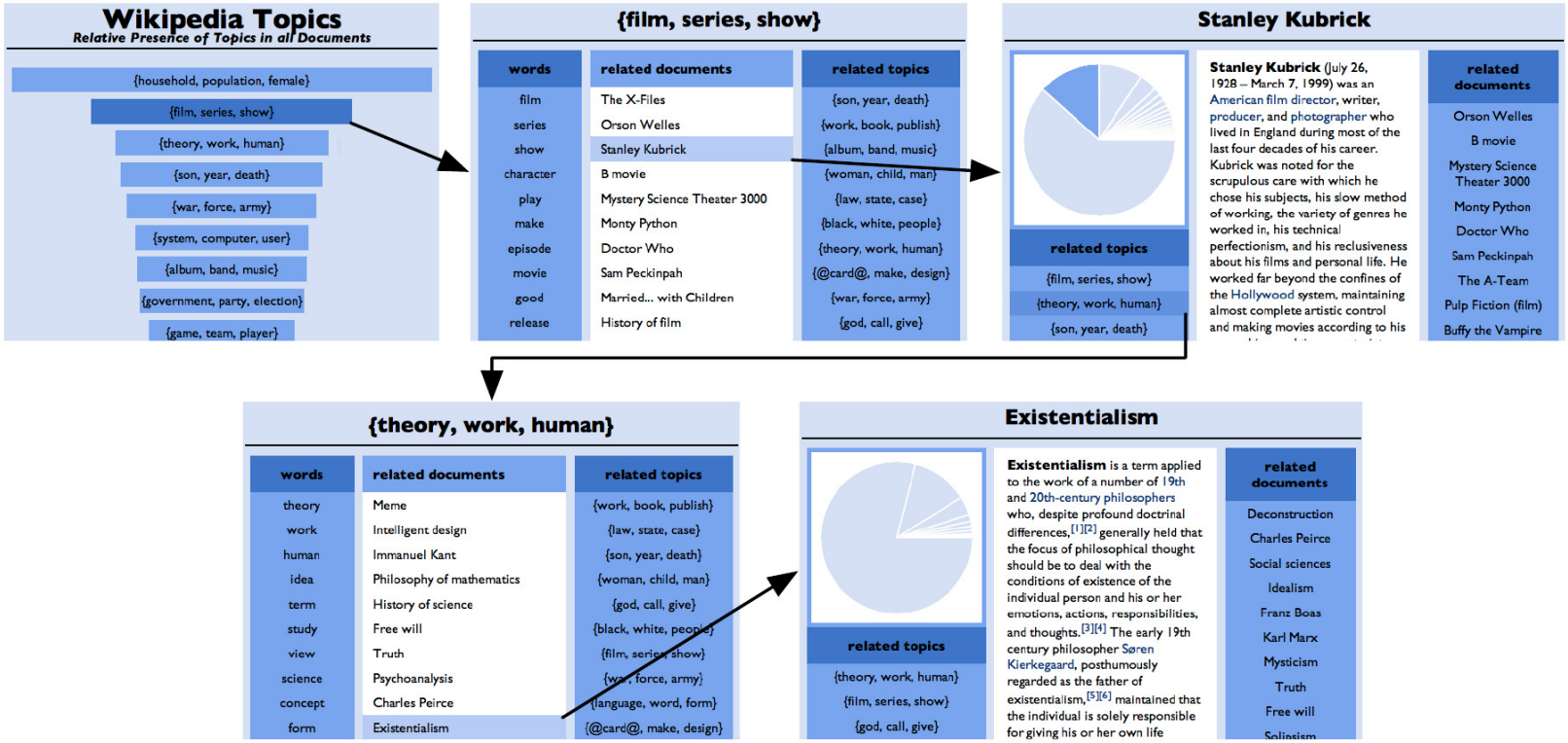
Процесс интерактивного улучшения тем:

- построить тематическую модель;
- для каждой интерпретируемой темы t вывести списки:
  — термины, ранжированные по $p(t|w)$, $p(w|t)$;
  — документы, ранжированные по $p(t|d)$, $p(d|t)$;
  — темы, ранжированные по сходству $p(w|t)$;
- отметить «белые» и «чёрные» элементы списков;
- внести разметку в регуляризаторы частичного обучения;
- именовать найденные новые интерпретируемые темы;
- добавить их к старым темам;
- перестроить модель и снова разметить темы;

См. также *Hu Y., Boyd-Graber J., Satinoff B.* Interactive topic modeling // HLT-2011. ACL. pp. 248–257.

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

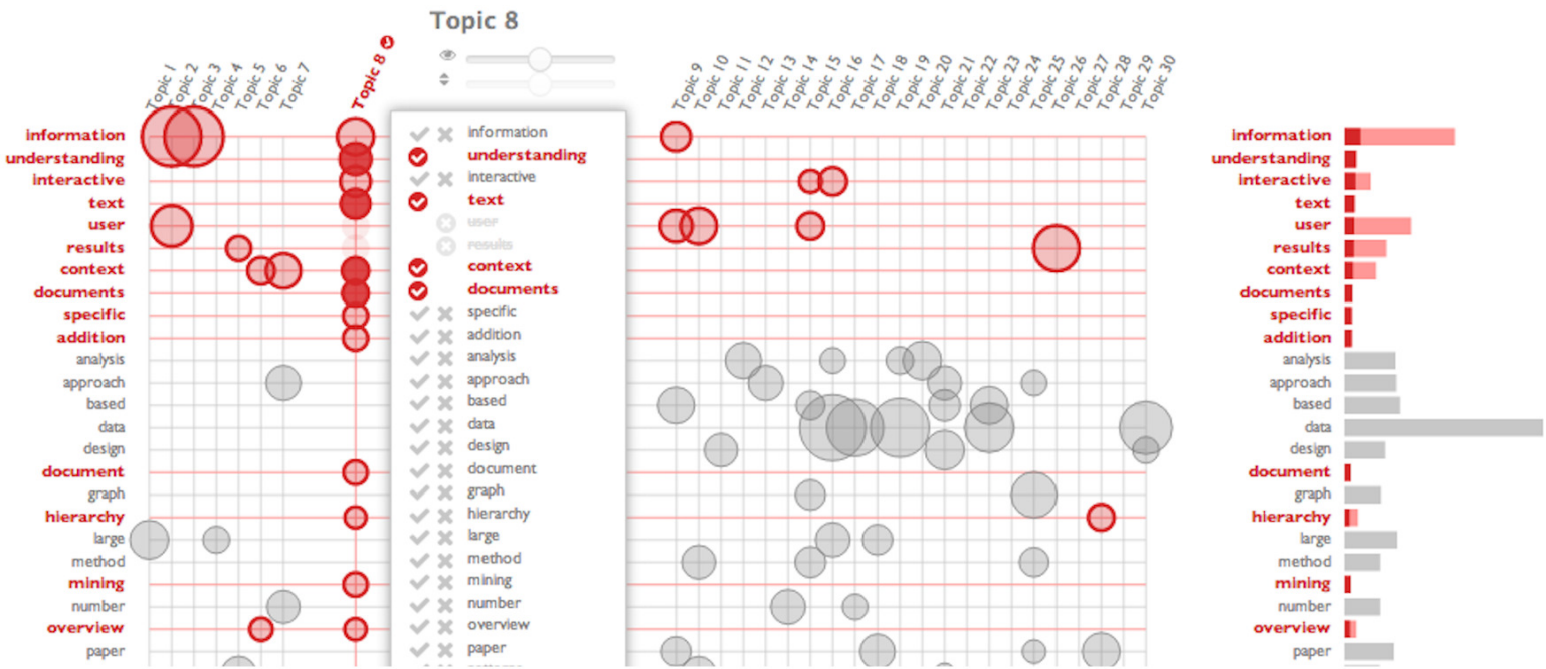


https://github.com/ajbc/tmv

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

Система Termite

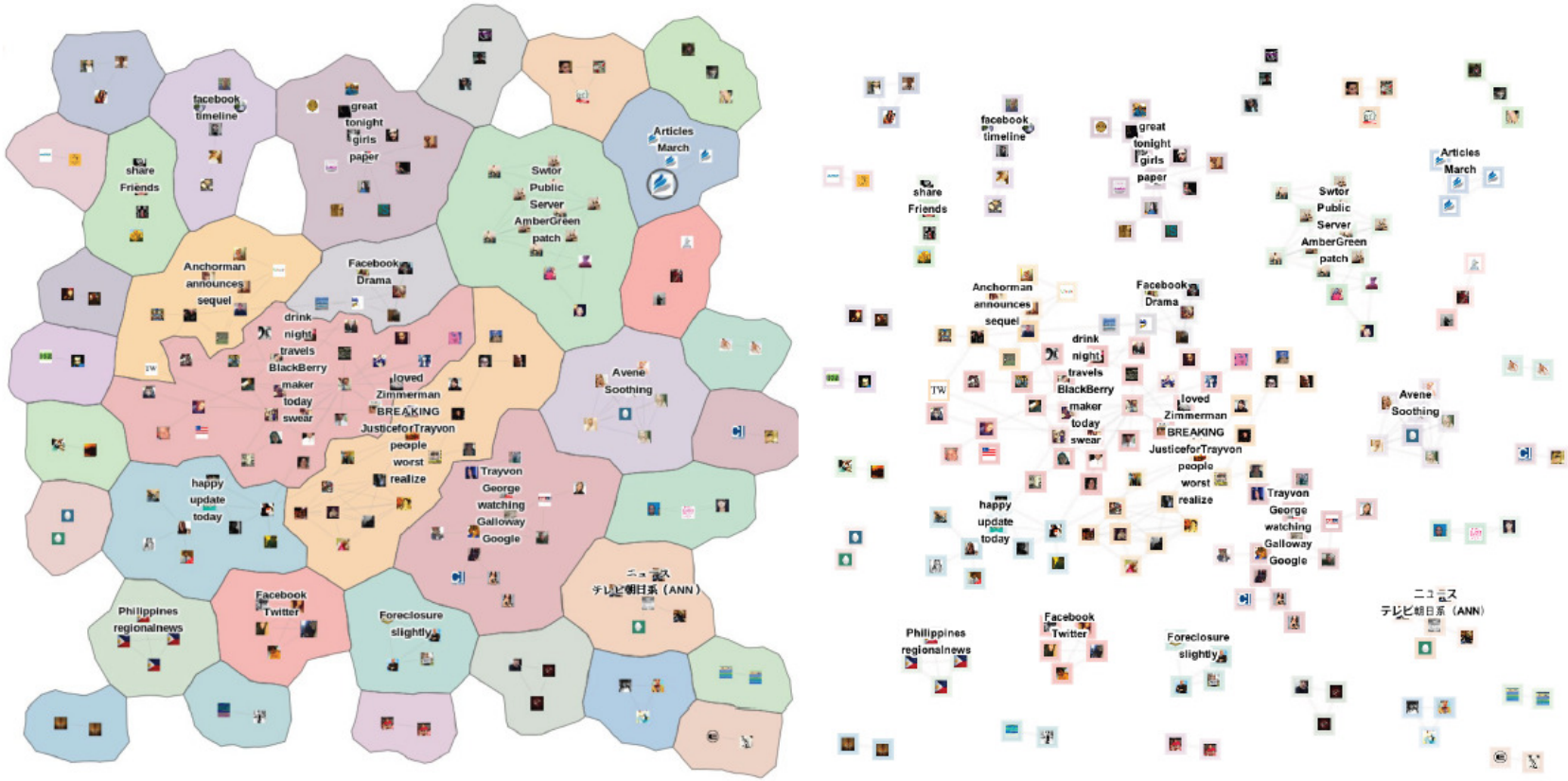
Интерактивная визуализация матрицы Φ и сравнение тем:



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

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models // International Working Conference on Advanced Visual Interfaces, 2012. ACM. pp. 74–77.

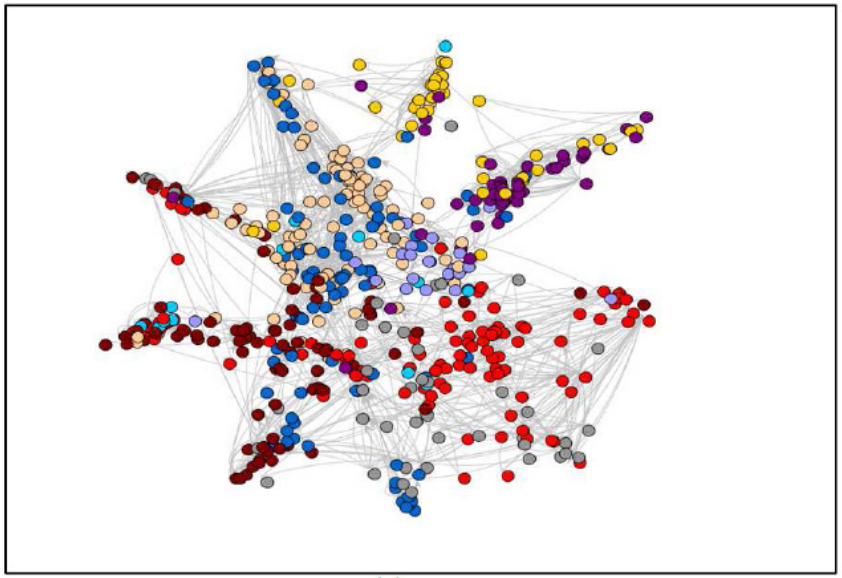
Дорожная карта: кластеризация релевантных документов

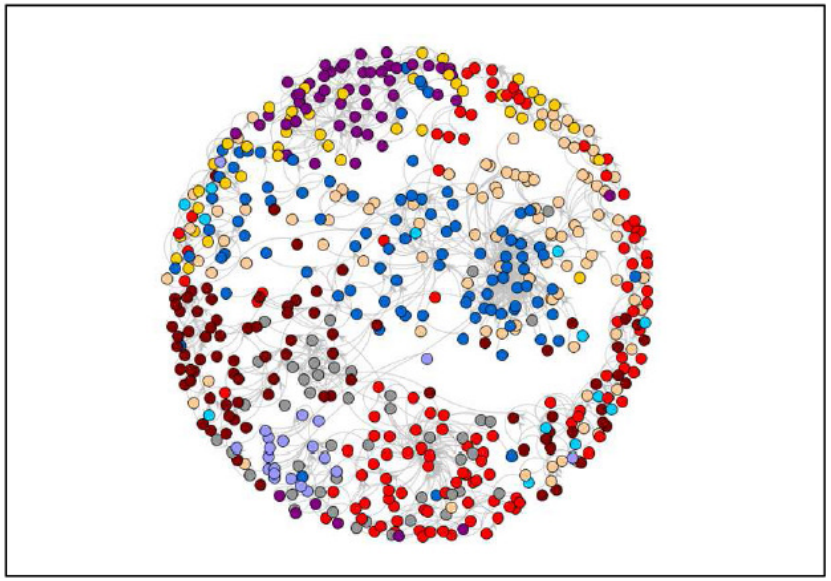


«A map metaphor visualization (left) seems more appealing than a plain graph layout (right), and clusters seem easier to identify.»

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

Дорожная карта: кластеризация релевантных документов

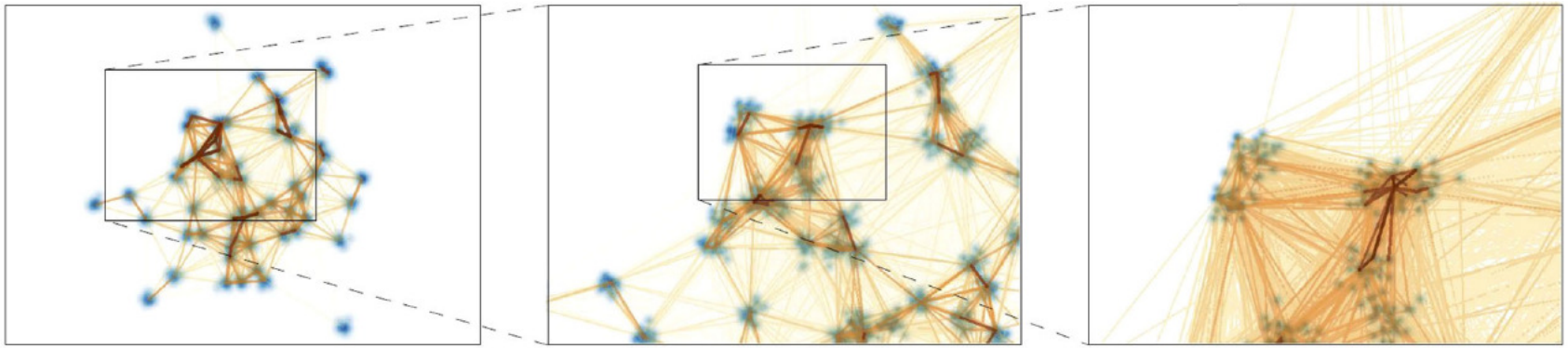




- Точки — это документы (или их фрагменты)
- Кластеры — это группы тематически схожих документов
- Форму облака точек можно настраивать

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

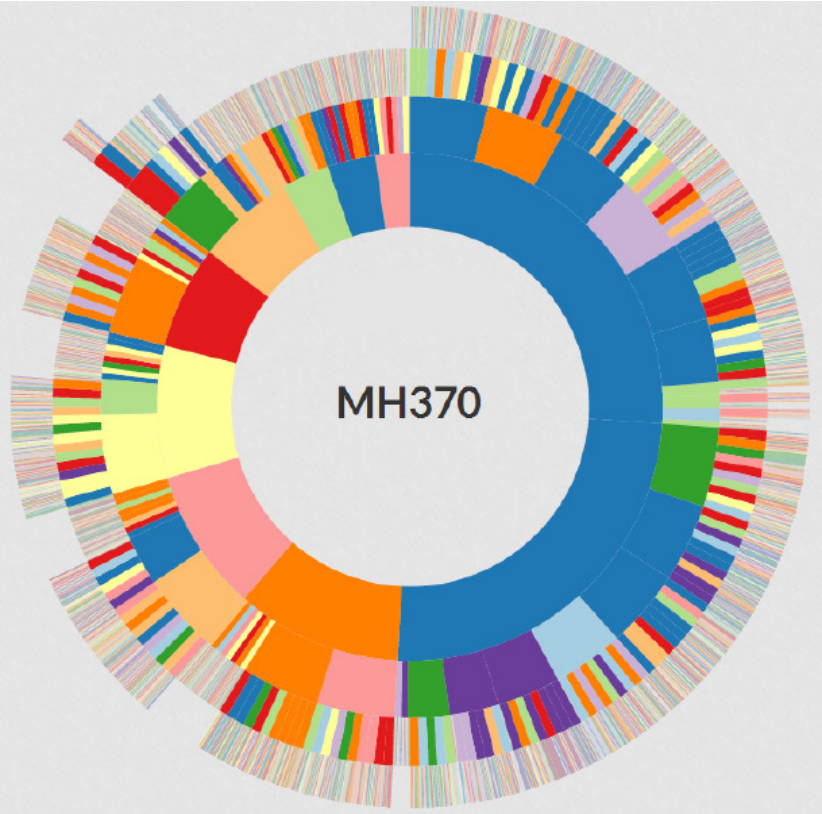
Дорожная карта: кластеризация релевантных документов



- Кластеры
  - кластеров
    - кластеров
      - кластеров...

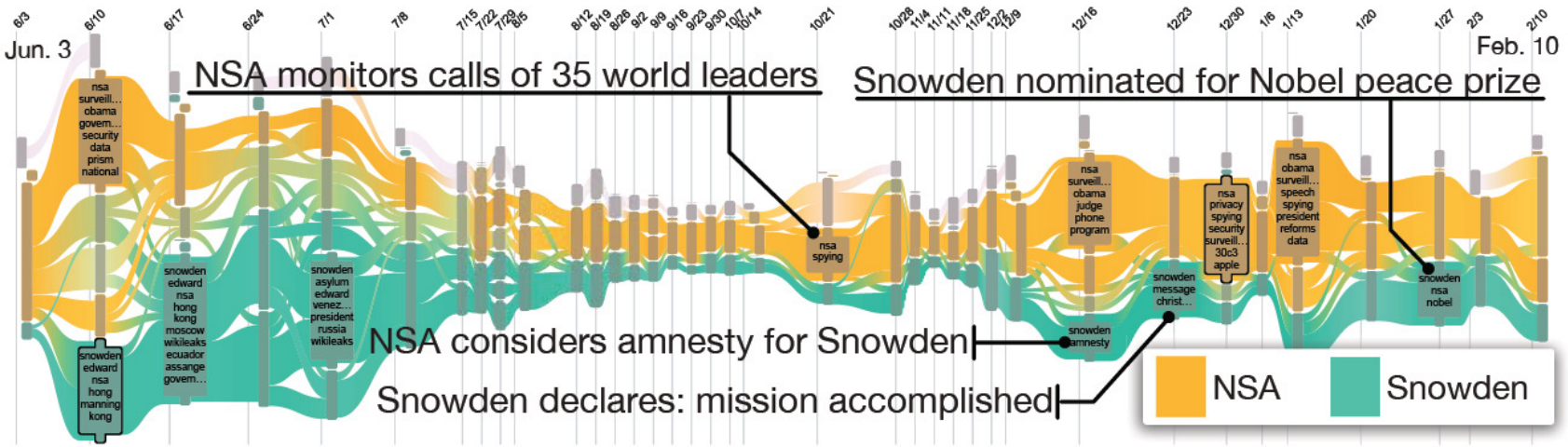
M.Zinsmaier, *U.Brandes*, *O.Deussen*, *H.Strobelt*. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

Тематическая иерархия: структура предметной области



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

Динамика тем: эволюция предметной области

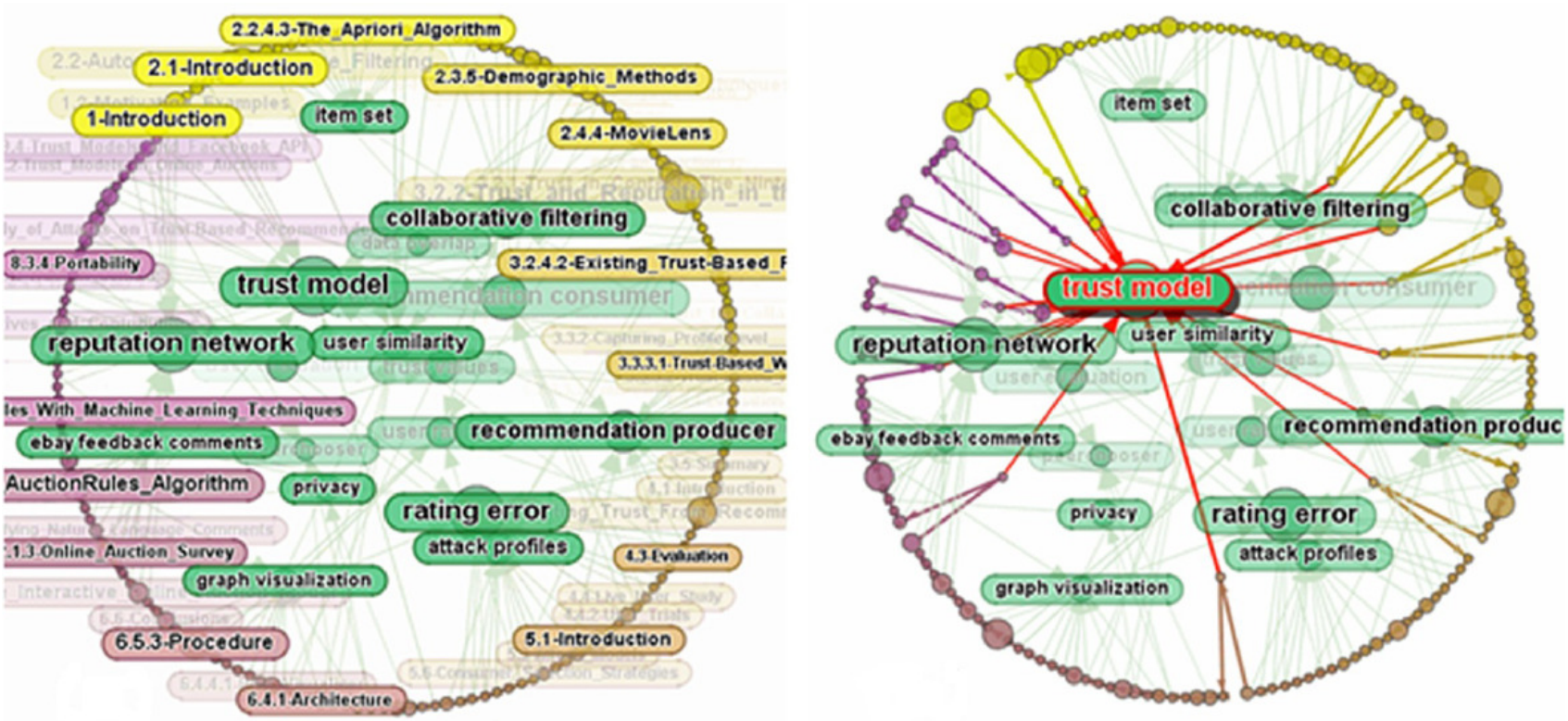


Эволюция иерархии тем. Коллекция Prism (2013/06/03 – 2014/02/09)

- эксперт выбирает сечение тематической иерархии,
- затем отмечает события в интерактивном режиме,
- и генерирует отчёт.

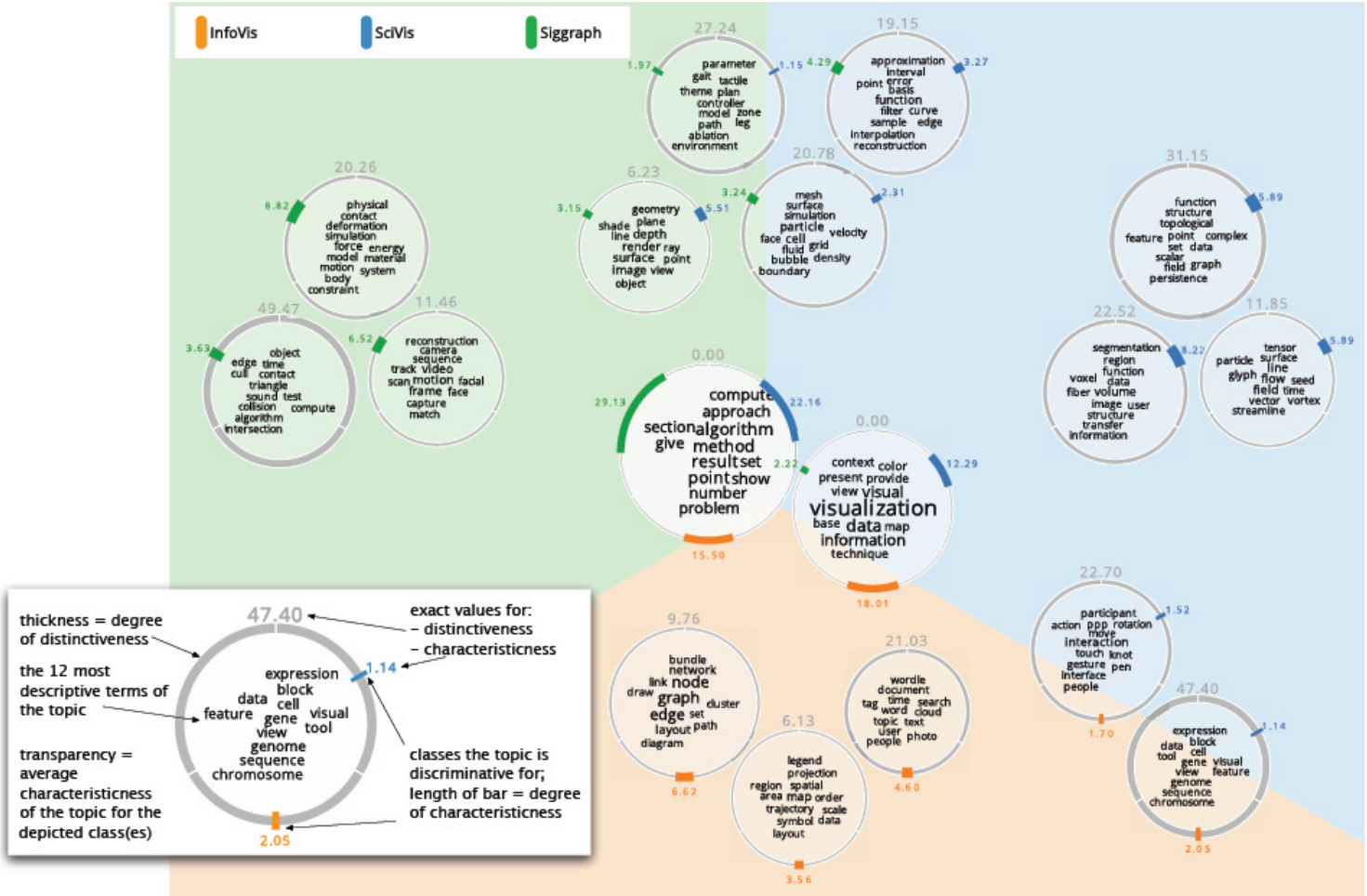
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

Тематическая сегментация документа запроса



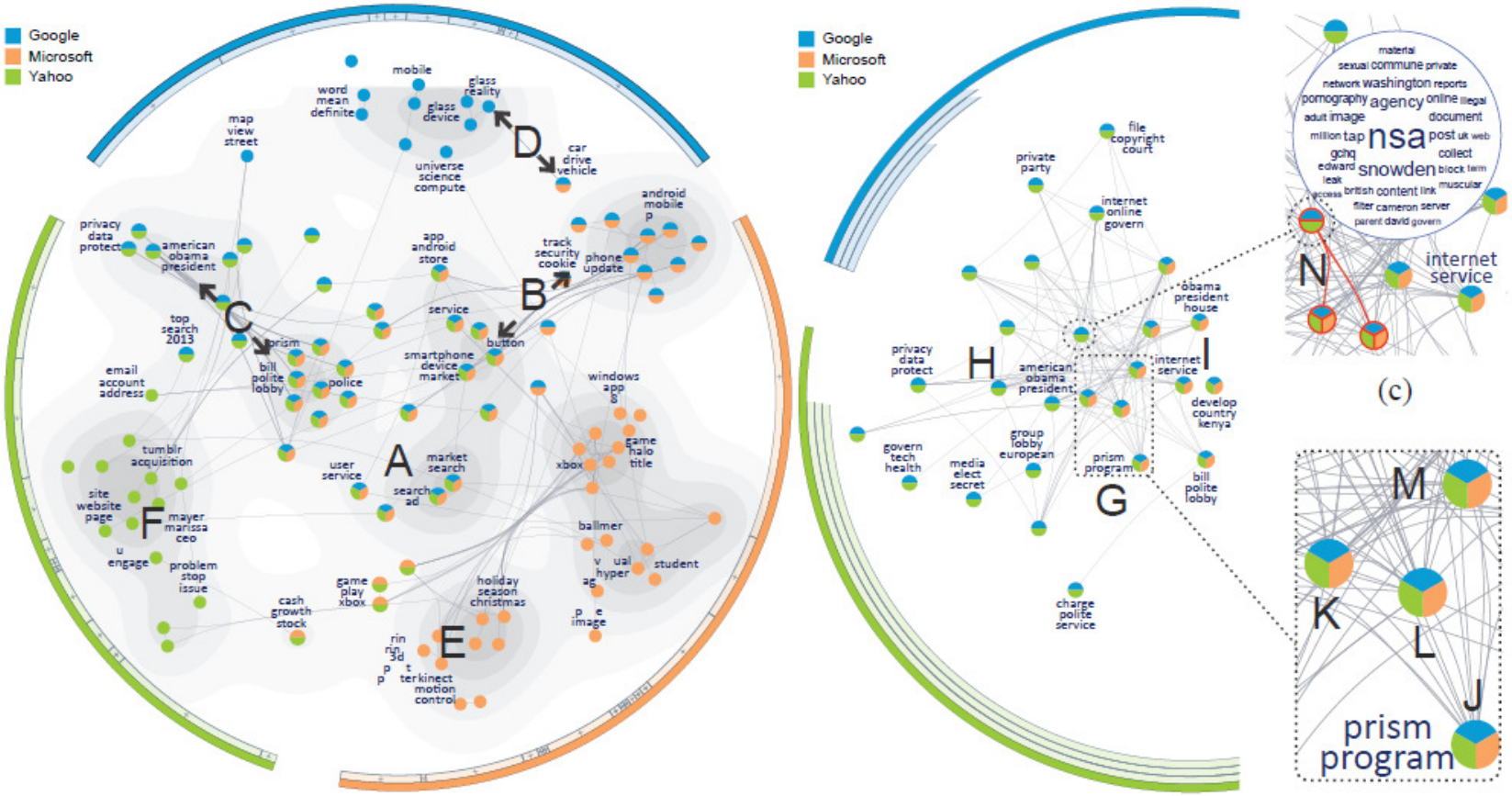
Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

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



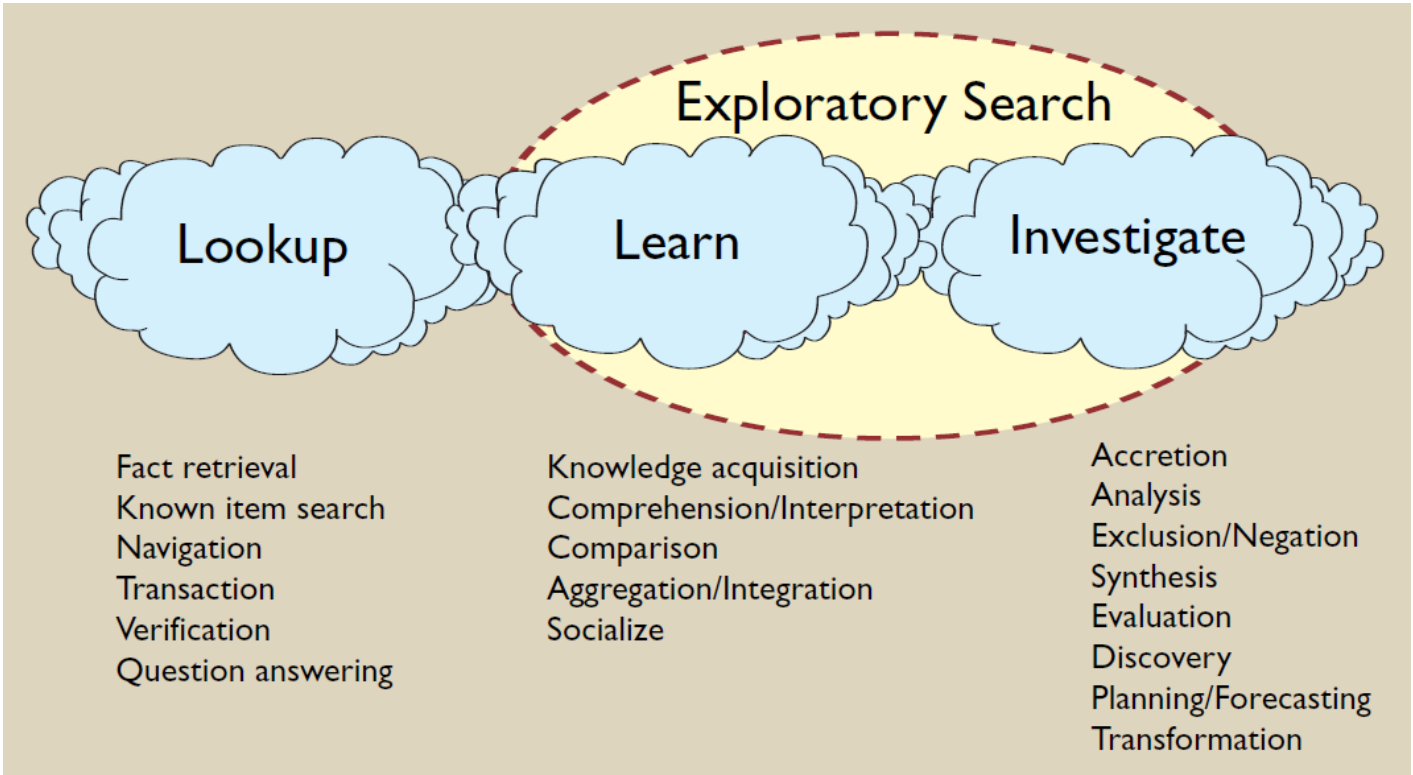
Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, Baining Guo. TopicPanorama: a full picture of relevant topics. IEEE Symp. on Visual Analytics Science and Technology. 2014.

http://textvis.lnu.se

Интерактивный обзор 220 средств визуализации текстов

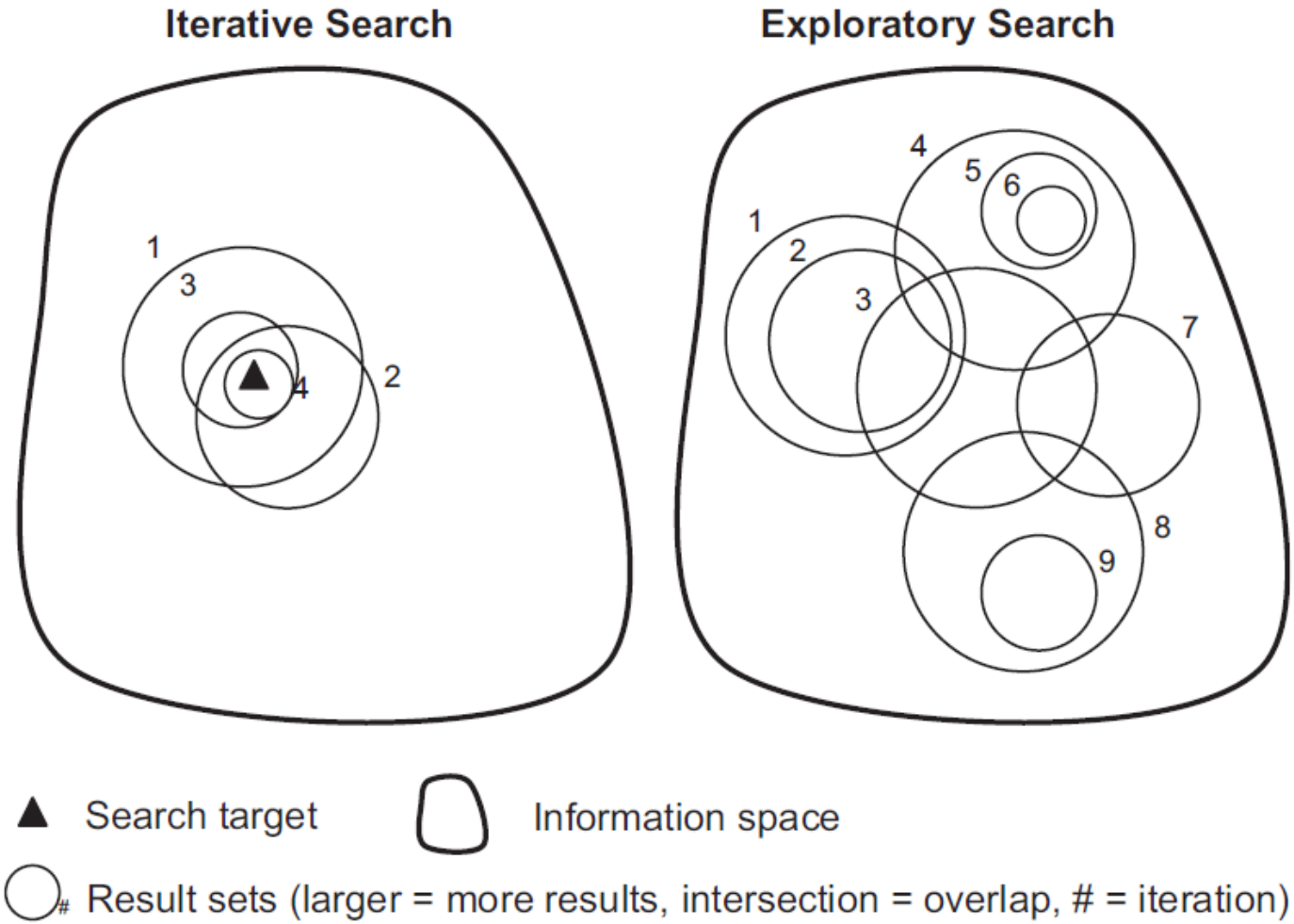
Разведочный поиск как инструмент самообразования

- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

От поиска “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.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

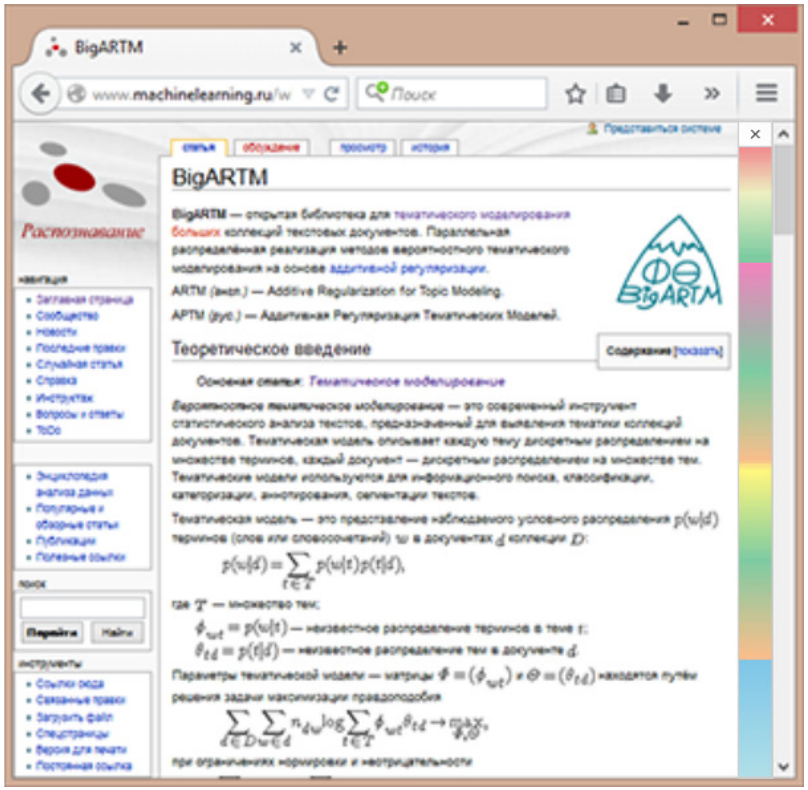
- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

1. имея любой текст под рукой, в любом приложении,
2. хотим получить картину содержащихся в нём тем-подтем,
3. и «дорожную карту» предметной области в целом

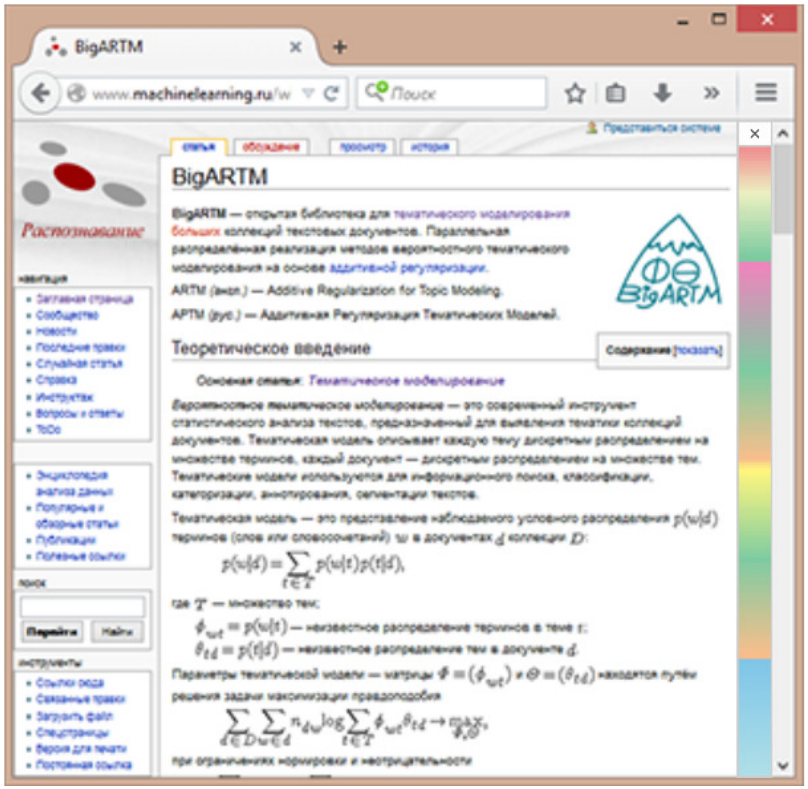
Разведочный поиск: прототип интерфейса

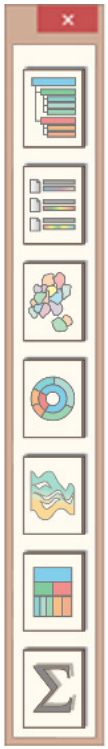
Радужная полоса напоминает, что знания всегда под рукой



Разведочный поиск: прототип интерфейса

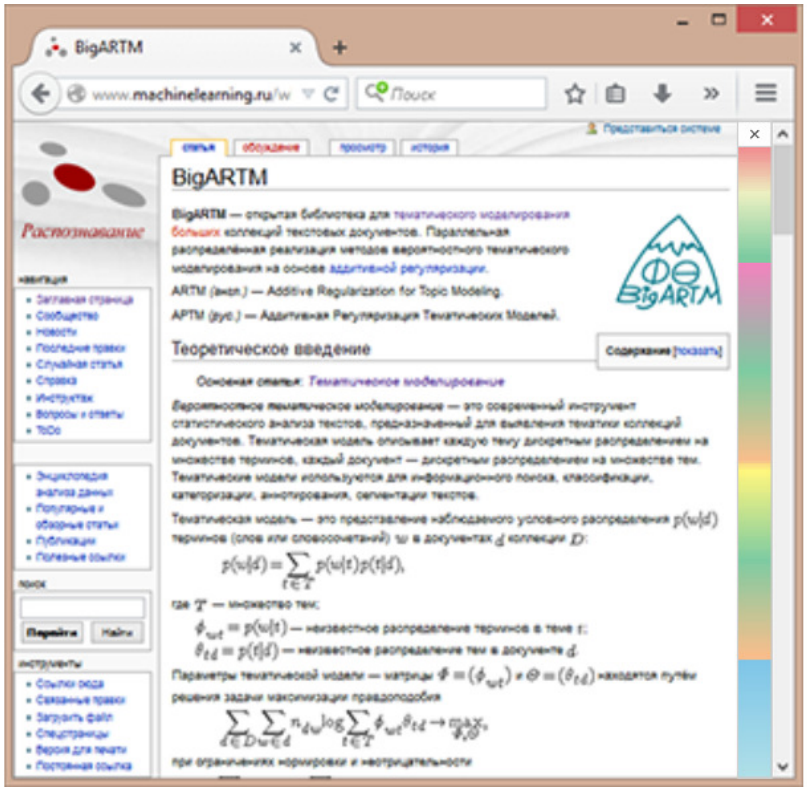
Клик по радужной полосе — тематический поисковый запрос

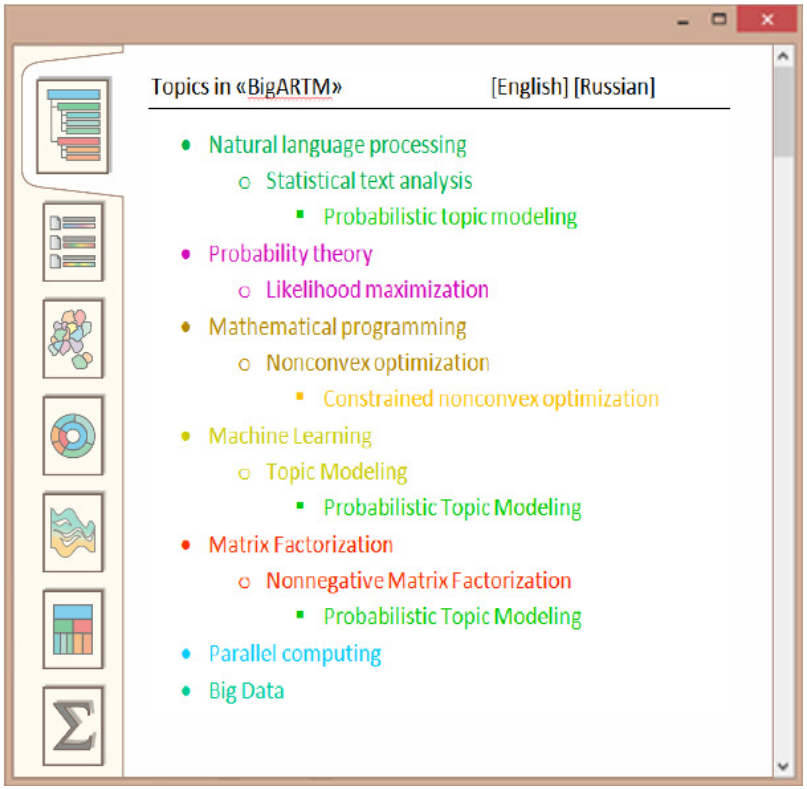




Разведочный поиск: прототип интерфейса

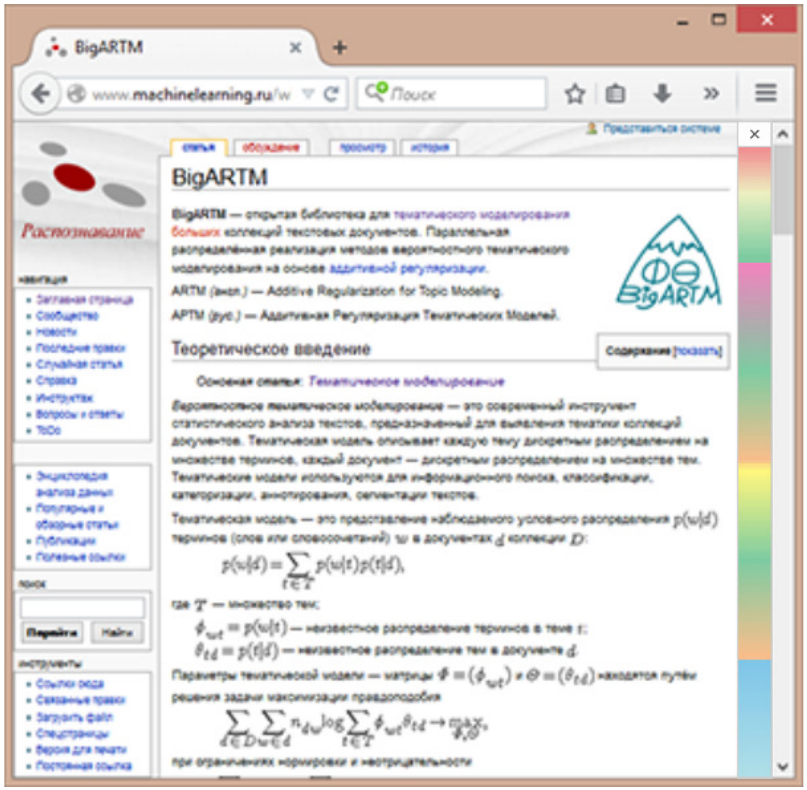
Темы-подтемы выбранного фрагмента текста

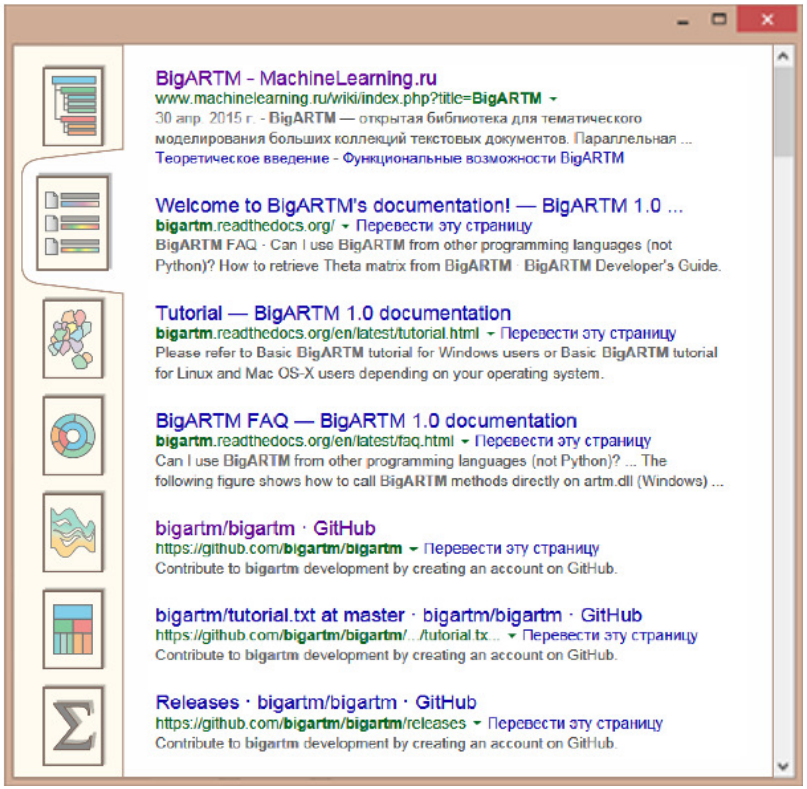




Разведочный поиск: прототип интерфейса

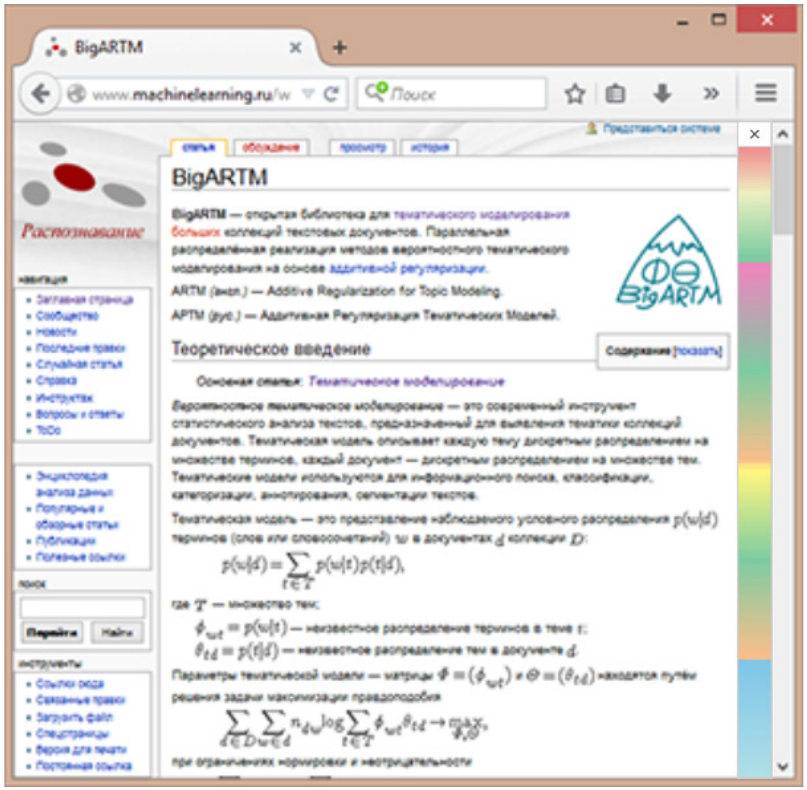
Документы и иные объекты, ранжированные по релевантности

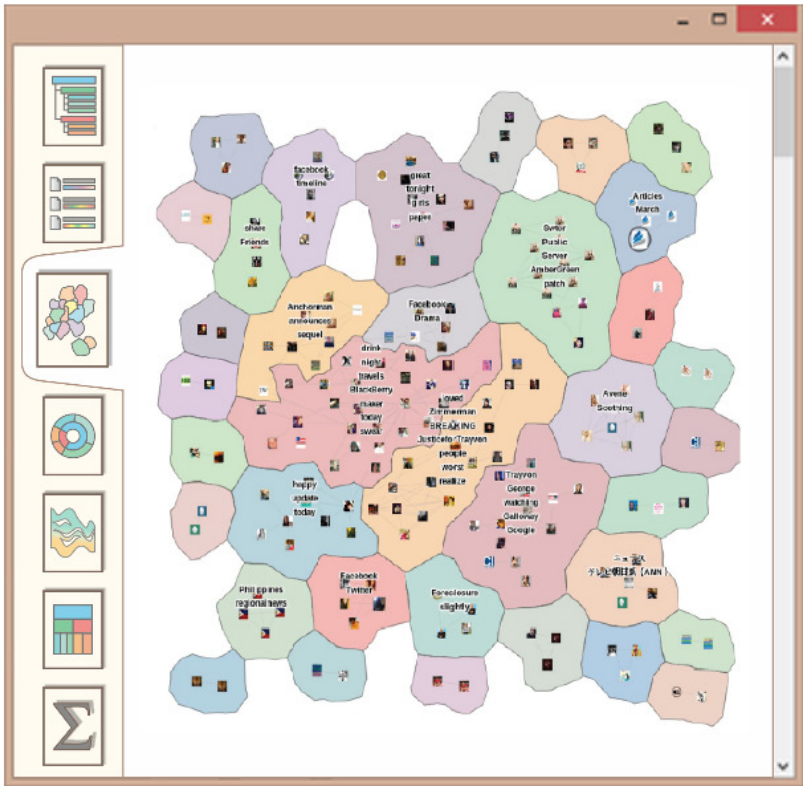




Разведочный поиск: прототип интерфейса

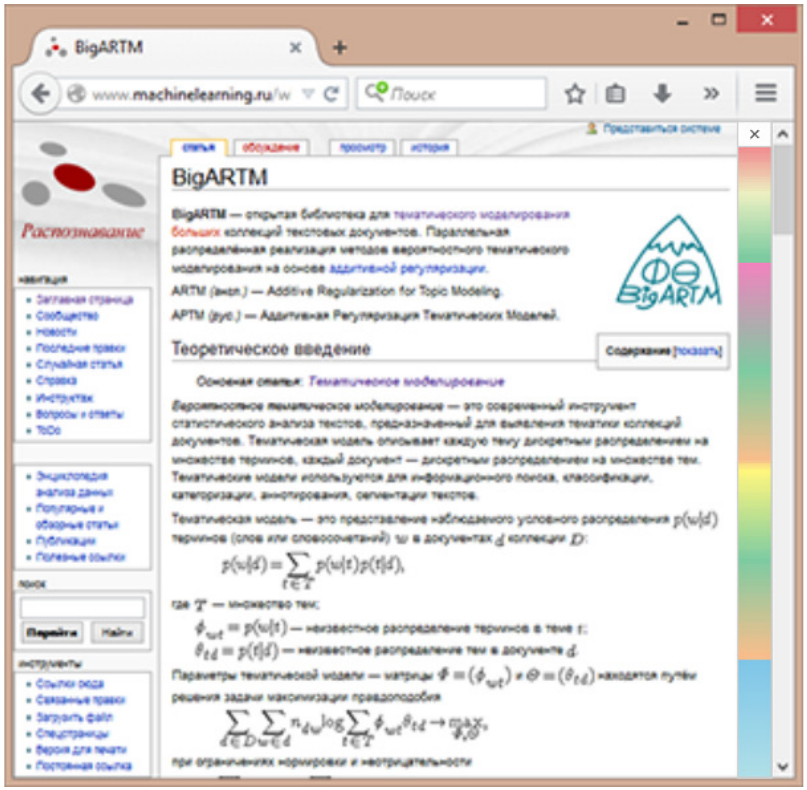
Дорожная карта: кластеризация релевантных документов

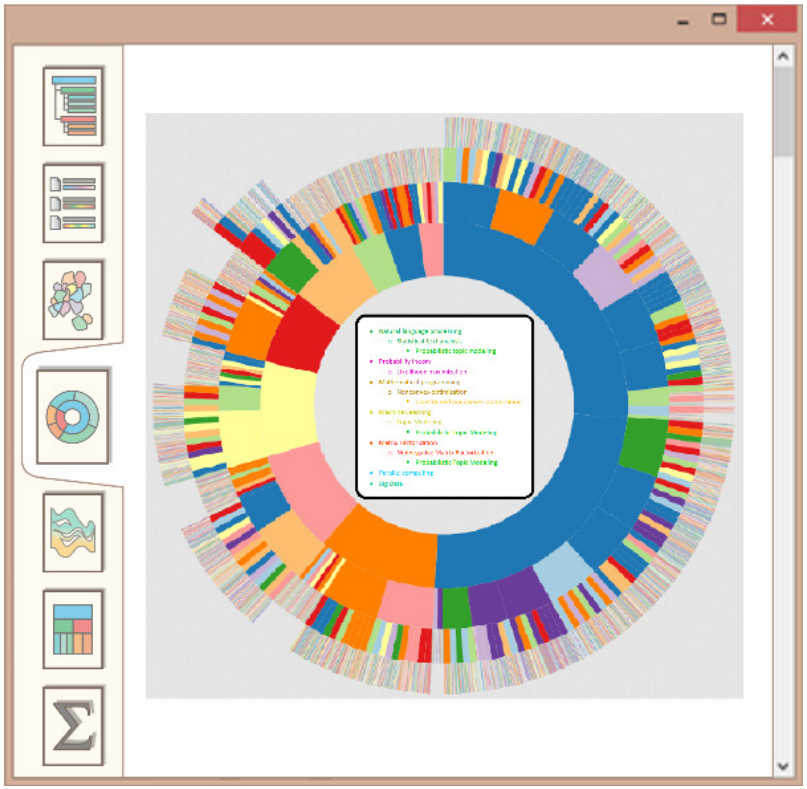




Разведочный поиск: прототип интерфейса

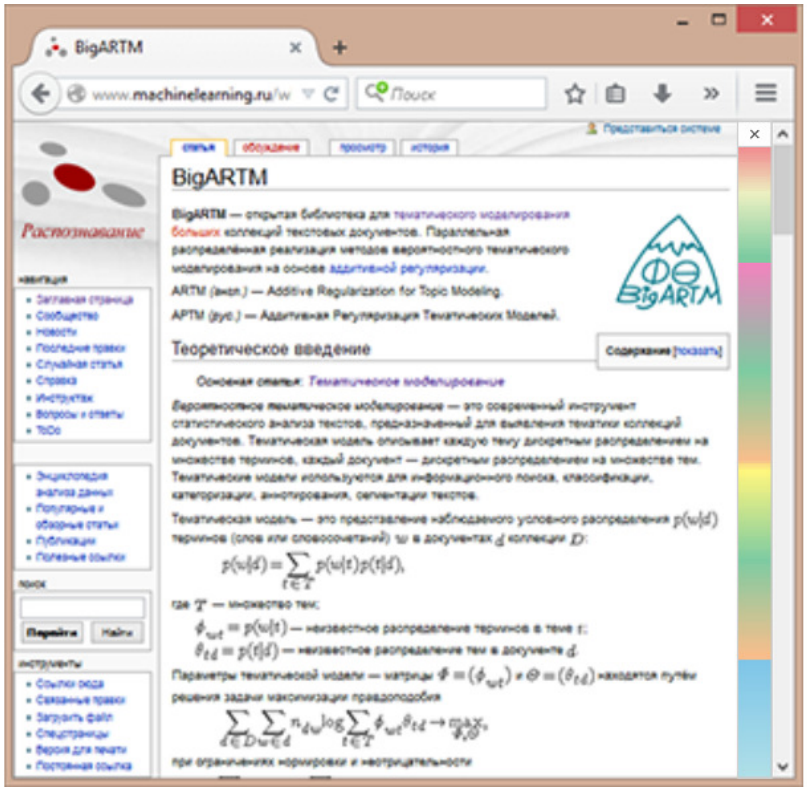
Тематическая иерархия: структура предметной области

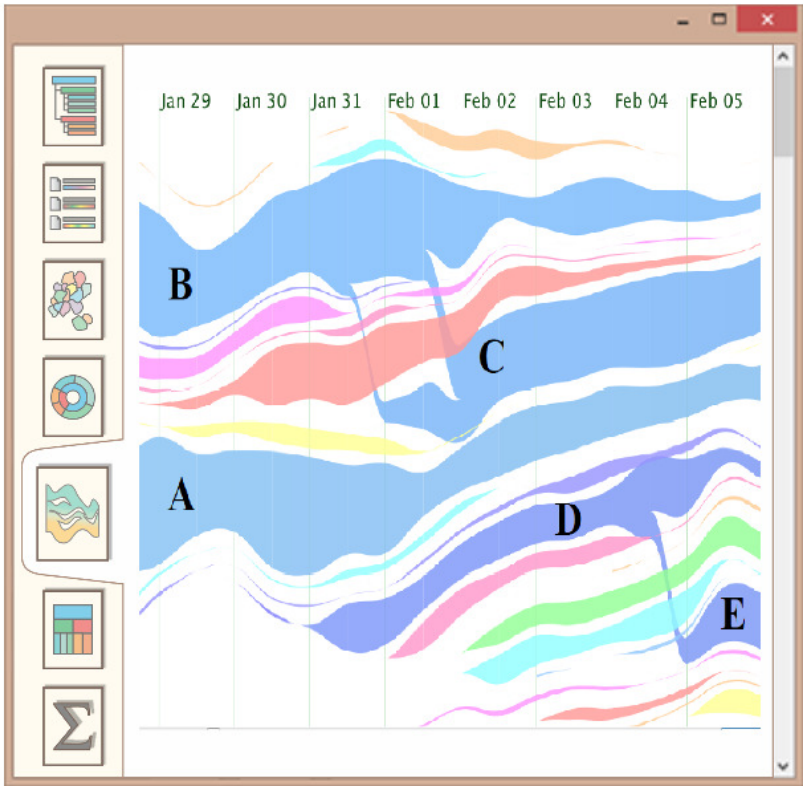




Разведочный поиск: прототип интерфейса

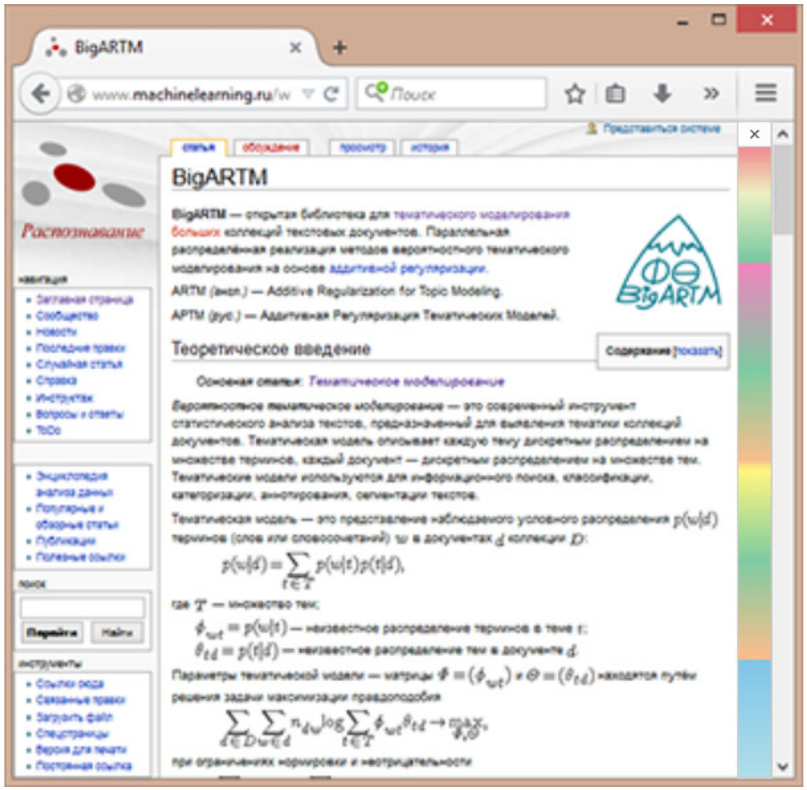
Динамика тем: эволюция предметной области

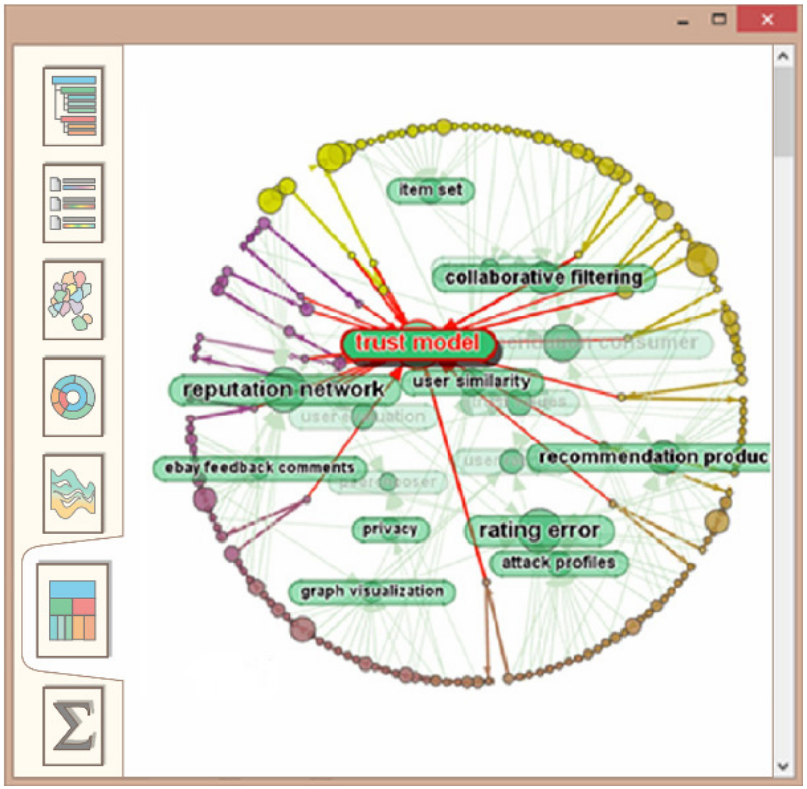




Разведочный поиск: прототип интерфейса

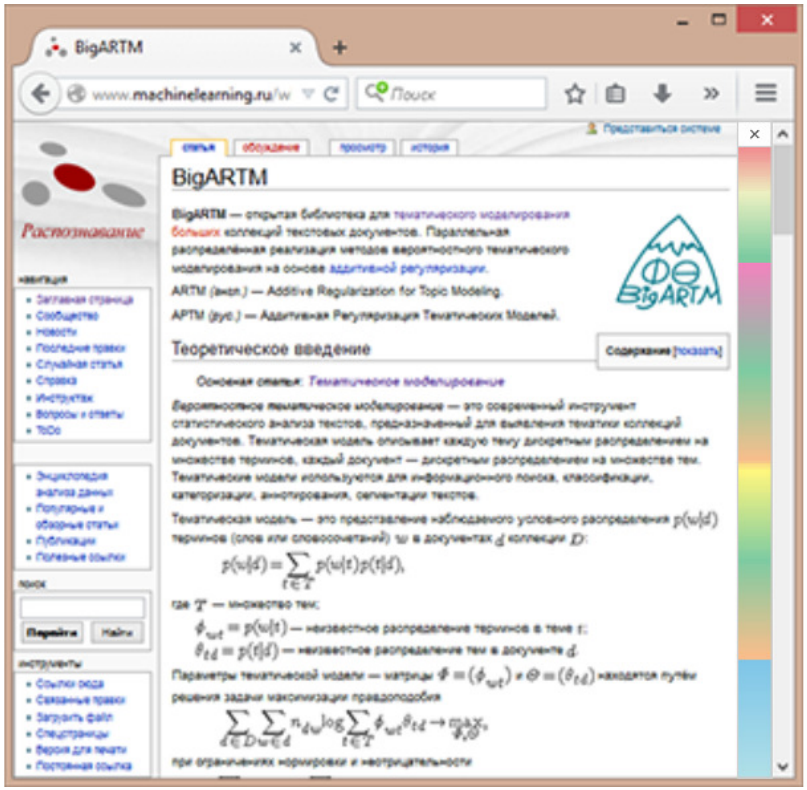
Тематическая сегментация документа запроса

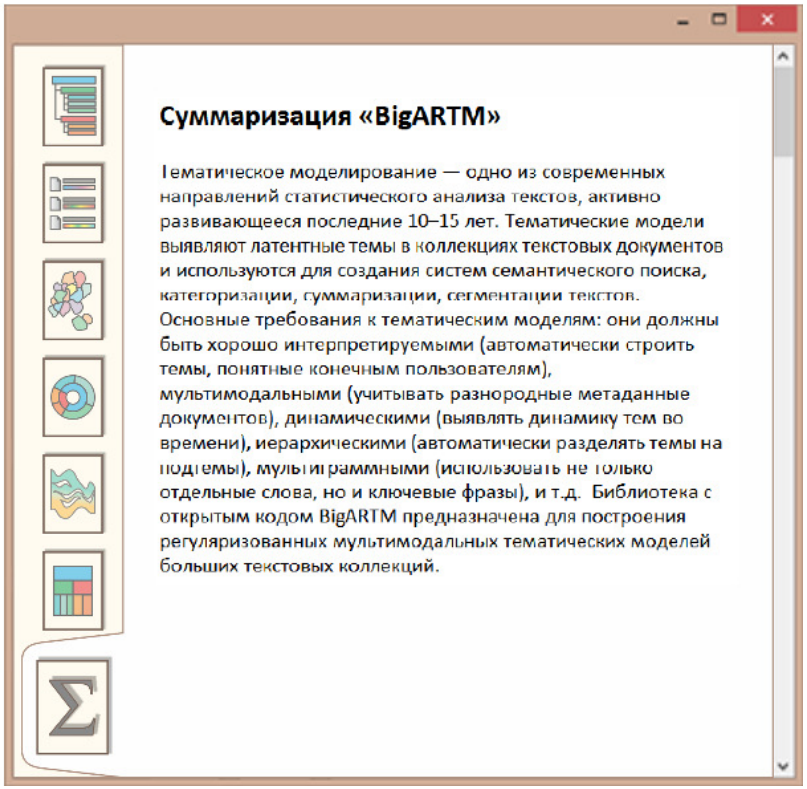




Разведочный поиск: прототип интерфейса

Суммаризация документа запроса





Резюме

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Тематическое моделирование — некорректно поставленная задача стохастического матричного разложения
- Стандартные методы — PLSA и LDA
- Аддитивная регуляризация (ARTM) — общий подход к построению комбинированных и многофункциональных тематических моделей.
- Взгляд на тематическое моделирование как на задачу многокритериальной оптимизации: регуляризаторы разнообразны, метрики качества тоже разнообразны.
- Разведочный информационный поиск (Exploratory Search) — одно из перспективных приложений тематического моделирования