Иерархические тематические модели

Ефимова Ирина

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. К. В. Воронцов

6 ноября 2016

Тематическое моделирование

- **Тема** – специальная терминология предметной области; набор терминов (слов или словосочетаний), совместно встречающихся в документах.
- **Тематическая модель** – модель коллекции текстовых документов, которая определяет к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.
- **Тематическое моделирование** – построение тематической модели.

Задачи:

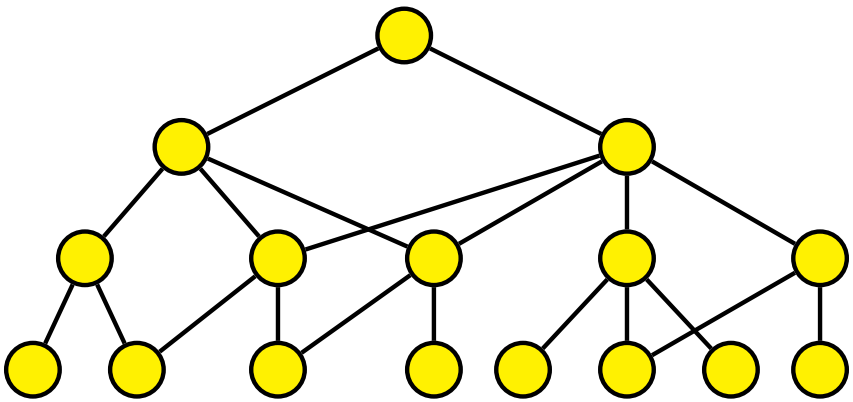
- Выявить тематическую структуру коллекции документов
- Найти сжатое тематическое описание каждого документа

Иерархия

Тематическая иерархия – многодольный граф тем с увеличивающимся числом тем на каждом уровне.

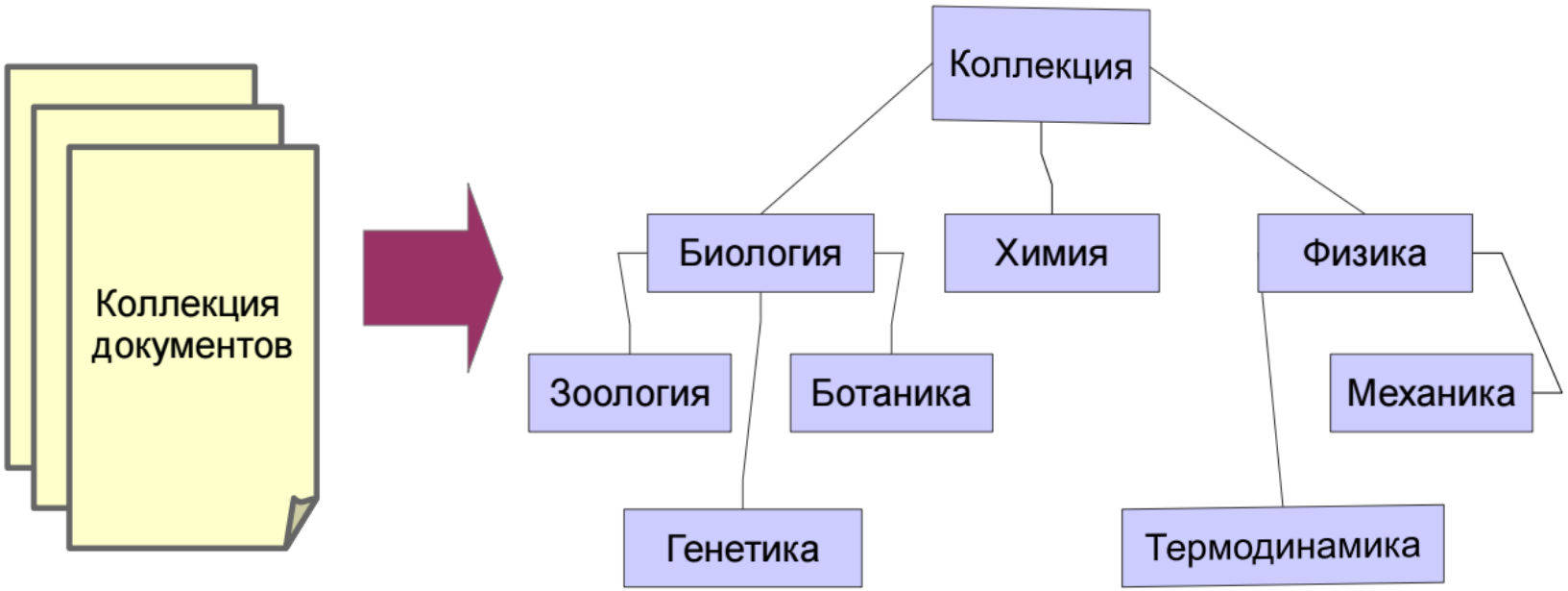
Преимущество: упрощает навигацию по коллекции документов.

Задача: построить тематическую иерархию тем коллекции документов.



Пример задачи

На основе тематических иерархий строятся системы представления знаний, удобные для систематизации и интерактивной визуализации больших текстовых коллекций: новостные статьи, научные публикации, бизнес-документы.



Литература

- Blei D. M., Jordan M., Tenenbaum J. Hierarchical Topic Models and the Nested Chinese Restaurant Process. NIPS, 2003.
- Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research), November 2011. Vol. 12, Pp.12749–2775.
- Pujara J., Skomorch P. Large-Scale Hierarchical Topic Models // NIPS Workshop on Big Learning, 2012.
- Wang C., Liu X. Song Y., Han J. Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach // Proc. 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15) // Association for Computing Machinery (ACM), August 2015.

Задача построения тематических моделей

Дано: W – словарь терминов,
D – коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$,
n_{dw} – частота термина w в документе d,
n_d – длина документа d.

Найти: параметры модели $p(w|d) = \sum_{s \in S} \phi_{ws}\theta_{sd}$:

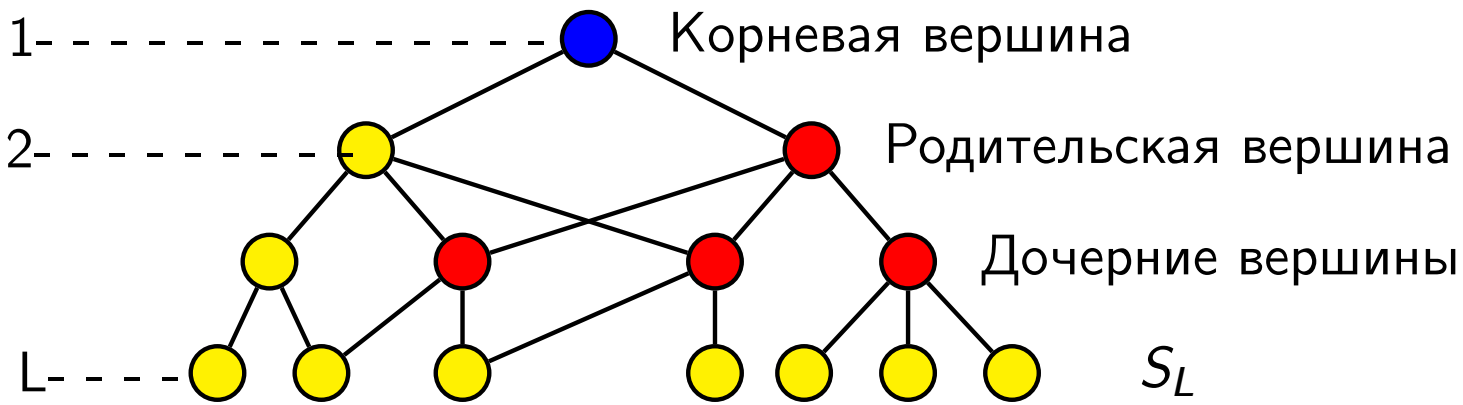
$\phi_{ws} = p(w|s)$ – вероятности терминов w в каждой теме s,
$\theta_{sd} = p(s|d)$ – вероятности тем s в каждом документе d.

Предположения:

- Порядок слов в документе не важен;
- Порядок документов в коллекции на важен;
- Гипотеза условной независимости $p(w|d,s) = p(w|s)$.

Задача построения иерархических тематических моделей

- $1, ..., L$ – уровни иерархии;
- $S_1, ..., S_L$ – множество тем уровней иерархий.
- Плоские тематические модели: темы формируют одно можество S.

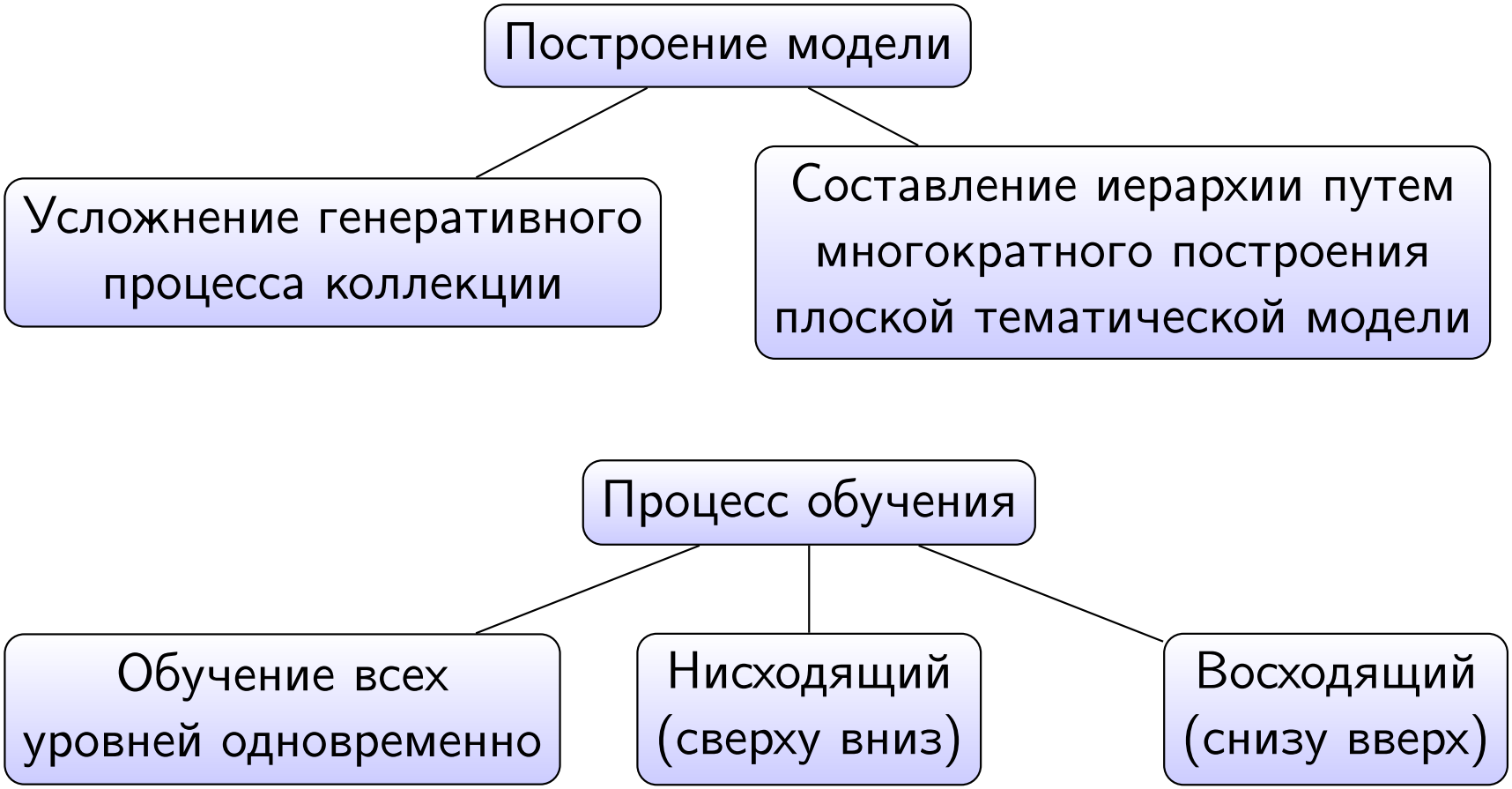


Построить иерархию для коллекции D.

Критерии/Требования

- **Требование полноты описания** – каждая тема должна описываться релевантными терминами и документами;
- **Структурное требование** – иерархия должна быть представлена в виде разреженного многодольного графа с увеличивающимся количеством тем на каждом уровне;
- **Требование масштабируемости** – обучение модели должно требовать малого числа проходов по коллекции;
- **Требование расслоения документа** – возможность оценить, какая доля документа сконцентрирована на каждом уровне иерархии (вектор пропорций).

Подходы к решению



LDA – Latent Dirichlet Allocation

Процесс порождения документа d

Вход: гиперпараметры априорных распределений α, η, ξ;

Выход: документ d;

1: сгенерировать распределение $\phi_s \sim Dir(\eta)$ для всех $s \in S$;
2: сгенерировать длину документа $n_d \sim Poisson(\xi)$;
3: сгенерировать распределение над множеством тем $\theta_d \sim Dir(\alpha)$;
4: **для всех** $i = 1, ..., n_d$ сгенерировать слово w_i в d
5: $\quad t \sim Mult(\theta_d)$;
6: $\quad w_i \sim Mult(\phi_s)$.

Для обучения модели используют вариационный вывод или семплирование Гиббса.

hLDA – hierarchicial LDA

nCRP (nested Chinese Restaurant Process) – задает априорное распределение на путь документа от корня к листу.

Процесс выбора подтемы s_{l+1} для темы s_l документа d при условии, что для $d_{pr} \in D_{pr} \subset D$ пути уже построены, $s_{l+1,i} \in S_{l+1}$:

$$p(s_{l+1,i}|s(D_{pr})) \sim \#\{d_{pr} \in D_{pr} : s_{l+1}(d_{pr}) = s_{l+1,i}\};$$
$$p(s_{l+1,new}|s(D_{pr})) \sim \gamma.$$

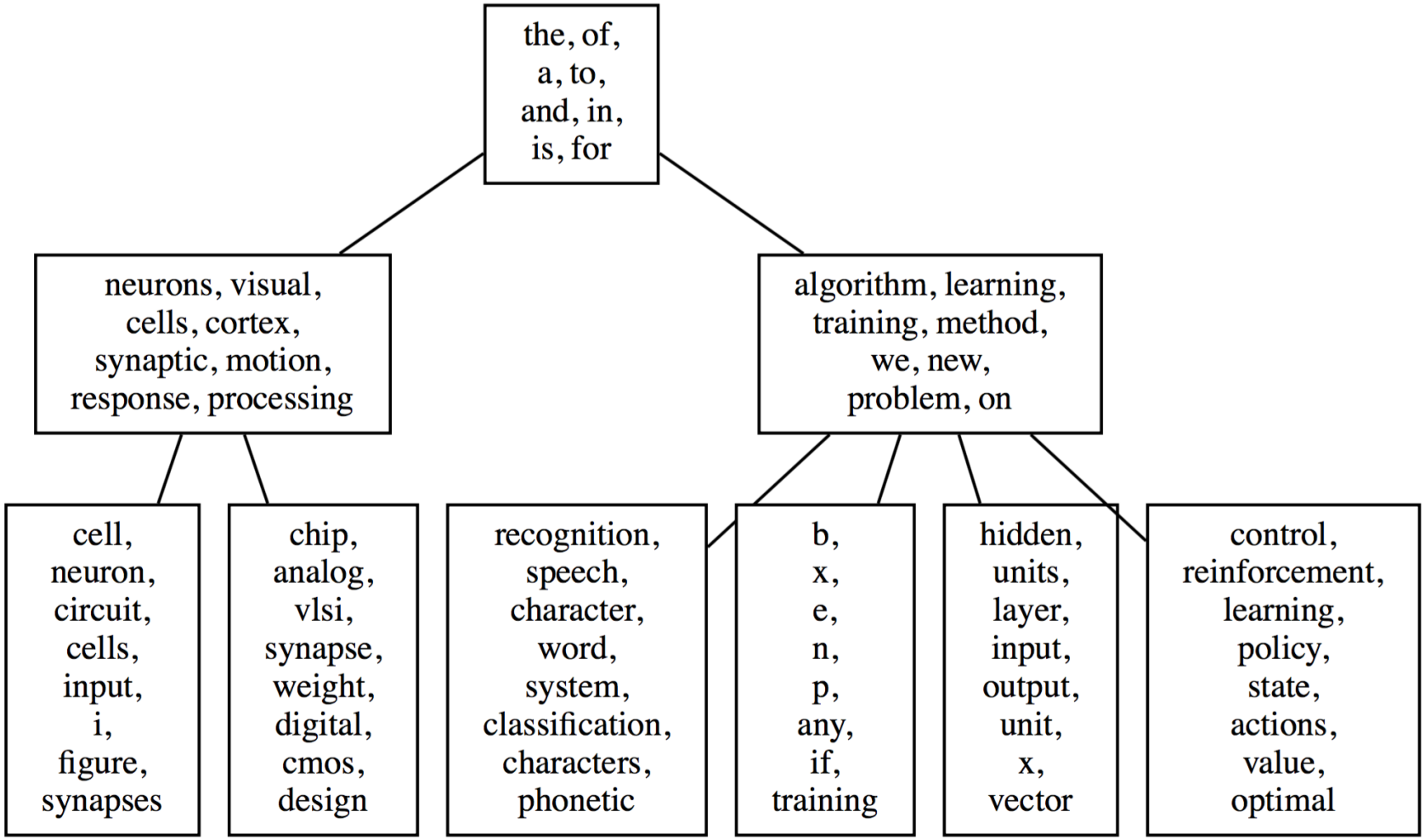
hLDA

Процесс порождения документа d

1: корень дерева – s_1;
2: **для всех** $l = 2, ..., L$
3: выбрать подтему s_{l+1} с помощью nCRP;
4: сгенерировать распределение на пути $(s_1, s_2, ..., s_L)$, $\theta_d \sim Dir(\alpha)$;
5: **для всех** $i = 1, ..., n_d$
6: $l \sim Mult(\theta_d)$;
7: $w_i \sim Mult(\phi_{s_l})$.

Для обучения используют семплирование Гиббса.

hLDA


the, of, a, to, and, in, is, for
neurons, visual, cells, cortex, synaptic, motion, response, processing
algorithm, learning, training, method, we, new, problem, on
cell, neuron, circuit, cells, input, i, figure, synapses
chip, analog, vlsi, synapse, weight, digital, cmos, design
recognition, speech, character, word, system, classification, characters, phonetic
b, x, e, n, p, any, if, training
hidden, units, layer, input, output, unit, x, vector
control, reinforcement, learning, policy, state, actions, value, optimal

Анализ hLDA

- Усложнение генеративного процесса коллекции;
- Обучение всех уровней одновременно;
- Иерархия – дерево;
- **Преимущество**: автоматическое определение количества тем в каждой вершине $\Rightarrow$ новый документ может добавить в дерево новые темы;
- **Недостатки**:
  - глубина дерева L фиксирована;
  - каждому документу соответствует один путь в дереве.

HDP – Hierarchical Dirichlet Processes

Алгоритм порождения документа d.

Вход: $\alpha, \beta, \gamma, \eta$;
Выход: Документ d.

1: Сгенерировать $H \sim Dir(\eta)$;
2: Сгенерировать $G_0 \sim DP(\gamma, H)$;
3: Задать длину документа n_d;
4: Сгенерировать $\theta_d \sim DP(\alpha, G_0)$;
5: **для всех** $i = 1, ..., n_d$
6: $\quad s \sim Mult(\theta_d)$;
7: $\quad$ **если** тема s является новой **то**
8: $\quad\quad$ Сгенерировать вектор темы $\phi_s \sim Dir(\beta)$
9: $\quad w_i \sim Mult(\phi_s)$.

---

hHDP – topic hierarchies of HDP

Алгоритм **hvHDP** – hierarchical vocabulary clustering

Вход: параметры HDP;
Выход: иерархия тем.

1: Построить модель HDP: $F = \Phi_{cur}\Theta_{cur}$, S_{cur} – множество тем;
2: **пока** $|S_{cur}| > 1$
3: Построить модель *HDP* на Φ_{cur}: $\Phi_{cur} = \Phi_{new}\Theta_{new}$, S_{new} – новое множество тем;
4: $\Phi_{cur} = \Phi_{new}$, $S_{cur} = S_{new}$.

---

hHDP – topic hierarchies of Hierarchical Dirichlet Processes

Алгоритм **htHDP** – hierarchical topic clustering

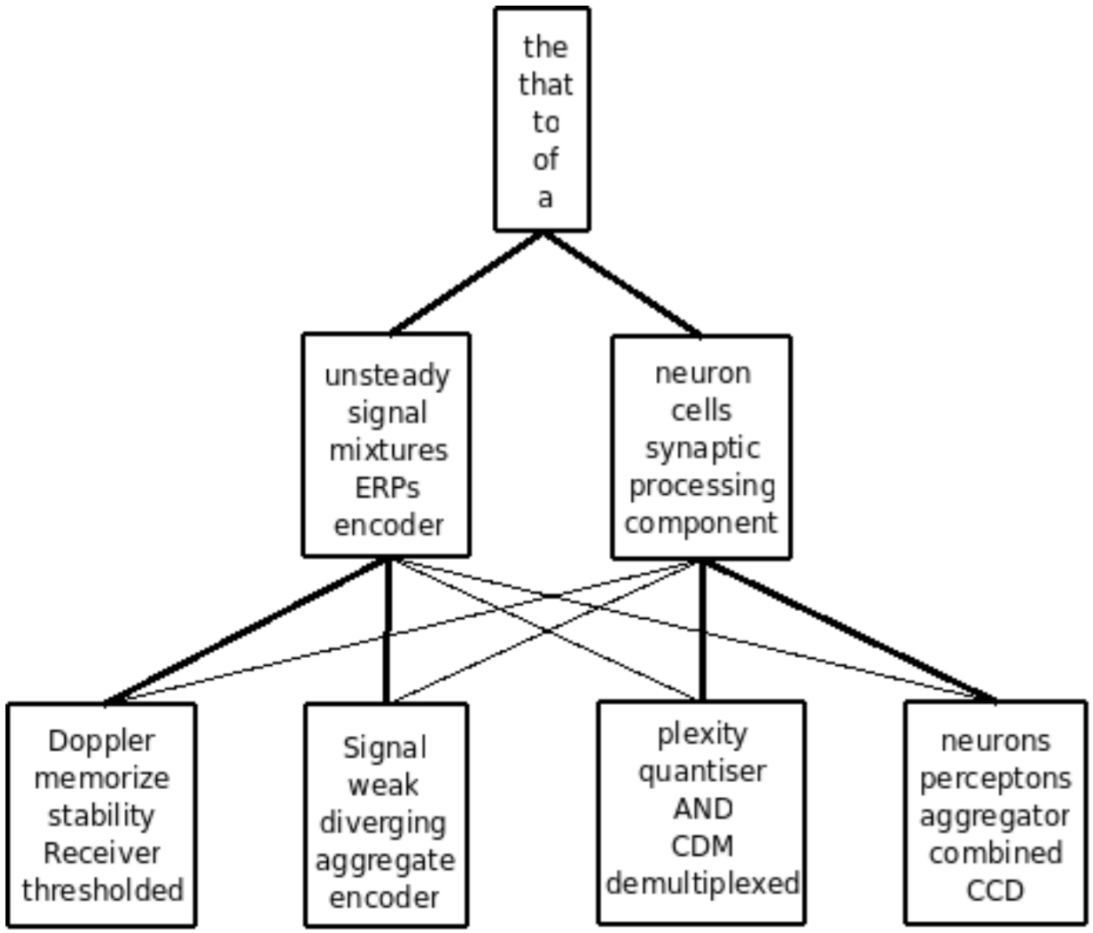
Вход: параметры HDP;
Выход: иерархия тем.

1: Построить модель HDP: $F = \Phi_{cur}\Theta_{cur}$, S_{cur} – множество тем;
2: **пока** $|S_{cur}| > 1$
3: Построить модель *HDP* на Θ_{cur}: $\Theta_{cur} = \Phi_{new}\Theta_{new}$, S_{new} – новое множество тем;
4: $\Theta_{cur} = \Theta_{new}$, $S_{cur} = S_{new}$.

Для обучения модели используют семплирование Гиббса, вариационный вывод.

hHDP

hvHDP


the
that
to
of
a
unsteady
signal
mixtures
ERPs
encoder
neuron
cells
synaptic
processing
component
Doppler
memorize
stability
Receiver
thresholded
Signal
weak
diverging
aggregate
encoder
plexity
quantiser
AND
CDM
demultiplexed
neurons
perceptons
aggregator
combined
CCD

Анализ hHDP

- Многократное построение плоской тематической модели HDP;
- Восходящий тип обучения;
- Иерархия – многодольный граф (возможно множественное наследование тем) $\Rightarrow$ каждому документу может соответствовать несколько несколько путей в графе;
- **Преимущество**: автоматическое определение количества тем в каждой вершине и количества уровней (HDP);
- **Недостатки**: описание тем либо только терминами, либо только документами.

splitLDA

Цель: добиться высокой *масштабируемости*.

Требование *масштабируемости*: обучение модели должно требовать малого числа проходов по коллекции. В противном случае область применимости метода ограничится небольшими коллекциями.

Mr.LDA – LDA, основанный на вариационном выводе и реализованный в рамках MapReduce на Hadoop.

- На шаге Map производится оптимизация параметров, связанных с документами;
- На шаге Reduce — с темами.

splitLDA

Вход: D, параметры $|S|$, L;
Выход: иерархия тем;

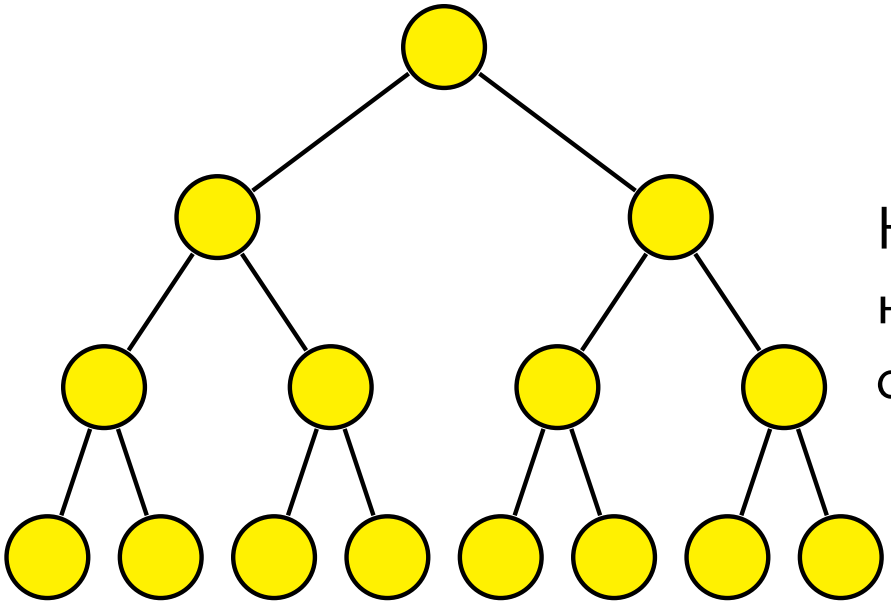
1: построить Mr.LDA на D, получить $\theta_{1:D}$, $\beta_{1:|S|}$;
2: $D = \bigcup_{i \in |S|} D_{1i}$;
3: **для всех** $i \in |S|$
4: LEARN-NODE($|S|, D_i, 1, L$);

LEARN-NODE($|S|, D, l, L$)

1: **если** $l < L$ **то**
2: построить Mr.LDA на D, получить $\theta_{1:D}$, $\beta_{1:|S|}$;
3: $D = \bigcup_{i \in |S|} D_{1i}$;
4: **для всех** $i \in |S|$
5: LEARN-NODE($|S|, D_i, l+1, L$);

Нет зависимости между надтемами и темами, темами одного уровня.

splitLDA



$|S| = 2,\ L = 3$

Нет зависимости между надтемами и темами, темами одного уровня.

Анализ splitLDA

- Многократное построение плоской тематической модели Mr.LDA;
- Нисходящий процесс обучения;
- Иерархия – дерево;
- **Преимущество**: высокая масштабируемость;
- **Недостаток**: в каждой вершине необходимо явно задавать количество тем и гиперпараметры Mr.LDA, задающие априорные распределения.

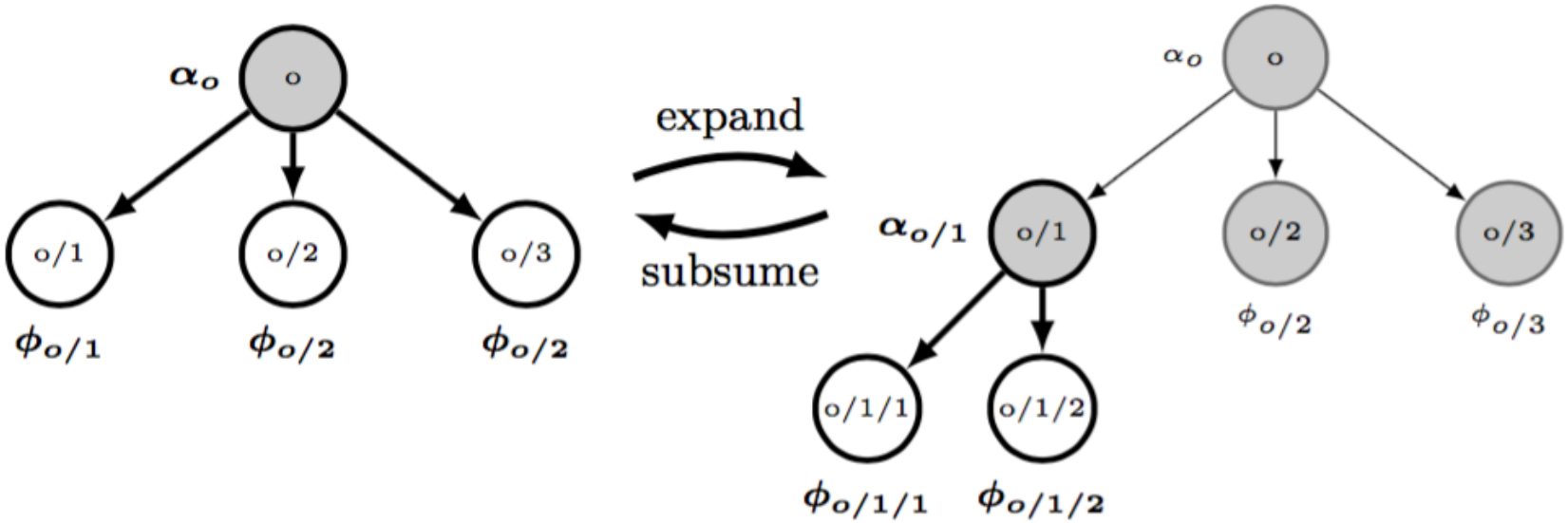
STROD – Scalable Recursive Orthogonal Decomposition

Процесс генерации документа d:

1: **для всех** l = 1,...,L-1
2: **для всех** $s_l \in S_l$
3: сгенерировать распределение над подтемами данной вершины $\theta_{s_l} \sim Dir(\alpha_{s_l})$;
4: **для всех** $i = 1, ..., n_d$
5: начать с корневой вершины s_1;
6: **для всех** $l = 1, ..., L-1$
7: выбрать подтему $s_{l+1} \sim Mult(\theta_{s_l})$;
8: сгенерировать $w \sim Mult(\phi_{s_L})$.

При обучении модели используется **метод моментов**, что позволяет строить дерево рекурсивным способом.

STROD



Параметры модели оцениваются с помощью тензорных разложений, что обеспечивает робастность алгоритма.

Анализ STROD

- Многократное построение плоской тематической модели;
- Нисходящий процесс обучения;
- Иерархия – дерево;
- **Преимущество**: масштабируемый, **устойчивый, интерактивный** алгоритм.

Сравнение

Критерий	hLDA	hHDP	splitLDA	STROD
Полнота	+	-	+	-
Структура	-	+	-	-
Масштабируемость	-	-	+	+
Расслоение документа	+	-	-	-

Заключение

- Сделан обзор основных методов построения иерархических тематических моделей: hLDA, hHDP, splitLDA, STROD;
- Проведен сравнительный анализ методов.