Tsyganova S.V., Strijov V.V. The construction of hierarchical thematic models for document collection // Applied Informatics, 2013, 1 : 109-115.

УДК 519.256

Построение иерархических тематических моделей коллекции документов.[1]

С. В. Цыганова, студентка Московского физито-технического института,
В. В. Стрижов, к.ф.-м.н., доц., н.с. Вычислительного центра РАН

Аннотация. Данная работа посвящена выявлению тематик коллекции текстов и их иерархической структуры. Поставлена задача построения иерархической тематической модели коллекции документов. Для решения поставленной задачи предлагается использование вероятностных тематических моделей. Особое внимание уделяется иерархическим тематическим моделям и, в частности, обсуждению свойств алгоритмов PLSA и LDA. Особенность построения иерархической модели заключается в переходе от понятия «мешка слов» к «мешку документов» в реализации плоских алгоритмов кластеризации. Работа алгоритмов иллюстрируется на текстах тезисов конференции EURO-2012 и на синтетических данных.

Ключевые слова: тематическая модель, иерархические модели, сэмплирование Гиббса, латентный семантический анализ

Введение

Работа посвящена построению иерархических тематических моделей коллекции документов – неупорядоченного набора текстов, тематику которых можно определить. Тематическая модель коллекции документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Тематическое моделирование используется для автоматического определения темы или тем документа при создании электронных библиотек, а также для автоматического поиска документов, посвященных заданной тематике.

Научная задача, решаемая непосредственно авторами в рамках заданной статьи, состоит в разработке и верификации алгоритма построения иерархической тематической модели. Такая модель требуется при создании электронных библиотек с инрархической тематикой, определенной библиографическими стандартами.

Как альтернатива классическим алгоритмам кластеризации [1, 2, 3, 4], основанных на вычислении функции расстояния между документами, в 1999 году Томасом Хоффманом [5] был предложен вероятностный латентный семантический анализ (алгоритм PLSA), основанный на принципе максимизации правдоподобия. Позже, в 2003 году, Дэвидом Блеем [6, 7] был предложен

[1] Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

усовершенствованный метод латентного размещения Дирихле (алгоритм LDA). Данные алгоритмы, в отличие от алгоритмов кластеризации, предполагают, что каждый документ относится к нескольким темам одновременно с некоторыми вероятностями (так называемая мягкая кластеризация) и находят эти неизвестные вероятности. Для ясности изложения предлагаемого алгоритма в работе приведены краткие описания алгортимов PLSA и LDA в авторской интерпретации.

Для работы с текстовыми документами необходимо привести их к каноническому виду для учета различных форм одного и того же слова[8]. Принято несколько базовых предположений относительно коллекции документов:

1) каждый документ — так называемый «мешок слов», т.е. неважно, в каком порядке слова находятся в документе, важна только частота употребления каждого из них;

2) каждое слово соответствует нескольким темам;

3) распределение слов в коллекции связано не с документами, а только с темами (предположение условной независимости);

4) каждый документ относится к нескольким темам с соответствующими вероятностями, т.е. каждый документ описывается набором тем;

5) сами темы и их количество задано;

6) количество тем много меньше объема словаря терминов и числа документов.

Будем считать, что количество тем задано. Сначала получим «плоскую» кластеризацию, т.е. просто разобьем документы на темы с соответствующими вероятностями. Затем построим иерархическую кластеризацию, т.е. выявим иерархию найденных тем. Основой перехода от плоской кластеризации к иерархической будет переход от понятия «мешка слов» к «мешку тем».

1 Постановка задачи

Пусть задана коллекция документов — вектор $\mathsf{D} = \{d_1,...,d_D\}$, каждый документ которого состоит из слов словаря $\mathsf{W} = \{w_1,...,w_W\}$. Представим наши данные матрицей «документ-термин», каждый элемент которой обозначает, сколько раз слово w встечалось в документе d:

$$[n_{dw}] = \begin{bmatrix} n_{11} & \dots & n_{1W} \\ \dots & \dots & \dots \\ n_{D1} & \dots & n_{DW} \end{bmatrix}.$$

Пусть на коллекции определено T тем $\mathsf{T} = \{t_1,\dots,t_T\}$. Введем следующие матрицы.

1. Матрица «документ-тема», каждый элемент которой обозначает число появлений темы t в документе d:

$$[n_{dt}] = \begin{bmatrix} n_{11} & \dots & n_{1T} \\ \dots & \dots & \dots \\ n_{D1} & \dots & n_{DT} \end{bmatrix}.$$

2. Матрица «слово-тема», каждый элемент которой обозначает, сколько раз слово w было отнесено к теме t:

$$[n_{wt}] = \begin{bmatrix} n_{11} & \dots & n_{1T} \\ \dots & \dots & \dots \\ n_{W1} & \dots & n_{WT} \end{bmatrix}.$$

Рассмотрим на декартовом произведении $\mathsf{D} \times \mathsf{W} \times \mathsf{T}$ дискретные распределения $p(d,w,t)$, в котором переменная t, соответствующая теме, не задана явно. Согласно формуле полной вероятности, совместная вероятность появления документа и слова выражается следующим образом:

$$p(d,w) = \sum_{t \in \mathsf{T}} p(d)\,p(t \mid d)\,p(w \mid d,t) \qquad (1)$$

Согласно предположению 3) относительно коллекции документов,

$$p(w \mid d,t) = p(w \mid t).$$

Тогда формулу (1) можно переписать следующим образом:

$$p(d,w) = p(d)\,p(w \mid d) = \sum_{t \in \mathsf{T}} p(d)\,p(t \mid d)\,p(w \mid t). \qquad (2)$$

Поделим (2) на $p(d)$:

$$p(w \mid d) = \sum_{t \in \mathsf{T}} p(t \mid d)\,p(w \mid t). \qquad (3)$$

Требуется построить тематическую модель, т.е. найти

$$p(w \mid t) \text{ для всех } t \in \mathsf{T} \quad \text{и} \quad p(t \mid d) \text{ для всех } d \in \mathsf{D}. \qquad (4)$$

Обозначив $p(w \mid t) = \varphi_{wt}$, а $p(t \mid d) = \theta_{dt}$, сведем постановку задачи к нахождению матриц Θ и Φ, где:

$$\boldsymbol{\Theta} = [\theta_{dt}] = \begin{bmatrix} \theta_{11} & \dots & \theta_{1T} \\ \dots & \dots & \dots \\ \theta_{D1} & \dots & \theta_{DT} \end{bmatrix}, \qquad \boldsymbol{\Phi} = [\varphi_{wt}] = \begin{bmatrix} \varphi_{11} & \dots & \varphi_{1T} \\ \dots & \dots & \dots \\ \varphi_{W1} & \dots & \varphi_{WT} \end{bmatrix}.$$

2 Алгоритм вероятностного латентного семантического анализа

Максимизируя правдоподобие выборки, найдем вероятности того, что слово w и документ d принадлежит теме t.

Используя (2), запишем логарифм функционала правдоподобия плотности

распределения выборки для функции вероятности совместного появления документа d и слова w в выборке D следующим образом:

$$\mathsf{L} = \sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw} \ln p(d,w) = \sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw} \ln \sum_{t\in \mathsf{T}} p(w\,|\,t)\,p(t\,|\,d)\underbrace{p(d)}_{const} \to \max_{p(w|t),\,p(t|d)} \qquad (5)$$

при ограничениях нормировки вероятностей:

$$\sum_{t\in \mathsf{T}} p(t) = 1; \qquad \sum_{d\in \mathsf{D}} p(t\,|\,d) = 1; \qquad \sum_{w\in \mathsf{W}} p(w\,|\,t) = 1;$$

Для решения задачи (5) используем итерационный *EM*-алгоритм, состоящий из двух шагов:

E-шаг. Вычисление скрытых переменных $p(t\,|\,d,w)$. При фиксированных параметрах $p(t\,|\,d)$ и $p(w\,|\,t)$ вычислим условные вероятности тем $p(t\,|\,d,w)$, используя формулу Байеса:

$$p(t\,|\,d,w) = \frac{p(t)\,p(t\,|\,d)\,p(w\,|\,t)}{p(d,w)}. \qquad (6)$$

M-шаг. Приближенное решение задачи максимизации правдоподобия и обновление параметров $p(t)$, $p(t\,|\,d)$ и $p(w\,|\,t)$. Находя точки максимума правдоподобия L (5) на итерации, приблизим параметры $p(w\,|\,t)$, $p(t\,|\,d)$ к тем величинам, которым они равны при максимальном правдоподобии L. Как видно из постановки задачи (4), приближаемые параметры $p(t)$, $p(t\,|\,d)$ и $p(w\,|\,t)$ являются искомыми распределениями:

$$p(t) = \frac{\sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw} p(t\,|\,d,w)}{\sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw}}, \qquad (7)$$

$$p(t\,|\,d) = \frac{\sum_{w\in \mathsf{W}} n_{dw} p(t\,|\,d,w)}{\sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw} p(t\,|\,d,w)} \equiv \theta_{dt}, \qquad (8)$$

$$p(w\,|\,t) = \frac{\sum_{d\in \mathsf{D}} n_{dw} p(t\,|\,d,w)}{\sum_{d\in \mathsf{D}}\sum_{w\in \mathsf{W}} n_{dw} p(t\,|\,d,w)} \equiv \varphi_{wt}. \qquad (9)$$

Итерации *EM*-алгоритма повторяются до тех пор, пока $p(t\,|\,d)$ и $p(w\,|\,t)$ не стабилизируются. Вывод формул (7), (8) и (9) описан ниже.

Для пояснения смысла формул, используемых на М-шаге алгоритма, кратко приведем их вывод . Запишем лагранжиан функции правдоподобия:

$$L = \sum_{d\in D}\sum_{w\in W} n_{dw} \ln p(d,w) - \nu\left(\sum_{t\in T} p(t) - 1\right) - \sum_{t\in T}\lambda_t\left(\sum_{d\in D} p(t\mid d) - 1\right) - \sum_{t\in T}\mu_t\left(\sum_{w\in W} p(w\mid t) - 1\right).$$

и будем искать точки максимума. Для этого продифференцируем лагранжиан L по нашим неизвестным переменным $p(t)$, $p(t\mid d)$, $p(w\mid t)$ и приравняем производную к нулю.

1. Продифференцируем лагранжиан по $p(t)$.

$$\frac{\partial L}{\partial p(t)} = \sum_{d\in D}\sum_{w\in W} n_{dw}\frac{p(t\mid d)p(w\mid t)}{p(d,w)} - \nu = 0.$$

Затем, домножив на $p(t)$ правую и левую части, выразим ν:

$$\sum_{d\in D}\sum_{w\in W} n_{dw}\frac{p(t\mid d)p(w\mid t)p(t)}{p(d,w)} = \nu p(t) \Rightarrow \nu = \sum_{d\in D}\sum_{w\in W} n_{dw}.$$

Теперь подставив формулу (6), выразим $p(t)$ и получим формулу (7).

2. Теперь продифференцируем лагранжиан по $p(t\mid d)$ и поступим аналогично первому пункту, домножив обе части полученного равенства на $p(d\mid t)$ и подставив (6). Получим формулу (8):

$$\frac{\partial L}{\partial p(t\mid d)} = \sum_{w\in W} n_{dw}\frac{p(t)p(w\mid t)}{p(d,w)} - \lambda_t = 0;$$

$$\sum_{w\in W} n_{dw}\frac{p(t\mid d)p(w\mid t)p(t)}{p(d,w)} = \lambda_t p(t\mid d) \Rightarrow \lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw} p(t\mid d,w);$$

3. И, наконец, продифференцируем лагранжиан по $p(w\mid t)$ и вновь выразим $p(w\mid t)$ аналогично 1 и 2 пунктам. Получим формулу (9):

$$\frac{\partial L}{\partial p(w\mid t)} = \sum_{d\in D} n_{dw}\frac{p(t)p(t\mid d)}{p(d,w)} - \mu_t = 0;$$

$$\sum_{d\in D} n_{dw}\frac{p(t\mid d)p(w\mid t)p(t)}{p(d,w)} = \mu_t p(w\mid t) \Rightarrow \mu_t = \sum_{d\in D}\sum_{w\in W} n_{dw} p(t\mid d,w);$$

3 Алгоритм латентного размещения Дирихле

Предполагается, что порождение тем происходит до порождения документов, и каждая тема задается некоторым распределением вероятности $p(w\mid t)$ на словаре W. Каждый документ является случайным набором тем.

Представим два последних сомножителя формулы (3) в виде случайных векторов:

$$p(d,w) = \sum_{t\in T} p(d)\underbrace{p(t\mid d)}_{\theta_{td}}\underbrace{p(w\mid t)}_{\varphi_{wt}}$$

Считаем, что $\boldsymbol{\theta}_{\mathbf{d}}$ и $\boldsymbol{\varphi}_{\mathbf{w}}$ - случайные векторы из распределения Дирихле:

$$\boldsymbol{\theta}_{\mathbf{d}} = (\theta_{1d},\ldots,\theta_{Td}) \in \mathrm{Dir}(\theta,\alpha),$$

$$\varphi_w = (\varphi_{1d},...,\varphi_{Td}) \in \mathrm{Dir}(\varphi,\beta).$$

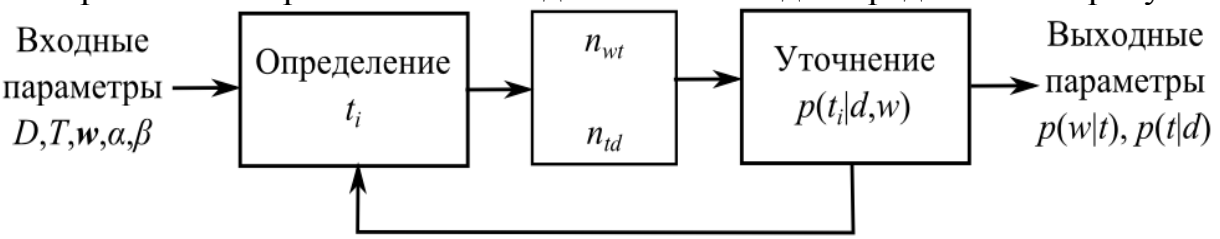
Заметим следующее свойство коллекции документов: каждая тема характеризуется некоторым небольшим множеством терминов, а каждый документ характеризуется небольшим набором тем. Таким образом, коллекция документов представляет собой некоторую «разреженную» структуру. Используем для описания разреженной структуры распределение Дирихле с параметрами α или β, которые управляют степенью разреженности.

Вернемся к формулировке задачи (4). Вероятность появления темы t в коллекции записывается следующим образом:

$$p(t \mid d,w) = \frac{p(d,w,t)}{p(d,w)} = \frac{p(d)}{p(d,w)}\, p(w \mid t)\, p(t \mid d) \propto \frac{p(w,t)\,p(d,t)}{p(t)\,p(d)} \propto \frac{n_{wt}+\beta}{\sum_{v=1}^{W} n_{vt} + W\beta}\, \frac{n_{dt}+\alpha}{\sum_{z=1}^{T} n_{dz} + T\alpha}.$$

Эта формула называется формулой обновления параметров в сэмплировании Гиббса. Подробный вывод формулы можно посмотреть в [9, 10, 11].

Алгоритм сэмплирования Гиббса для LDA наглядно представим на рисунке:



Алгоритм выполняется следующим образом: Выбирается тема t_i из распределения $p(t \mid d,w)$, фиксируются все остальные темы и подсчитывается значения n_{wt} и n_{td}. Затем уточняется распределение $p(t \mid d,w)$ и вновь выбирается тема t_j из этого распределения. Все шаги алгоритма повторяются в цикле, пока распределение не сойдется. Алгоритм на псевдокоде можно найти в [12].

4 Критерии качества тематических моделей

Степень неопределенности (perplexity) равна логарифму правдоподобия:

$$\mathrm{Perplexity} = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td}\right).$$

Точность (precision) и полнота (recall). Эти оценки используются для алгоритмов, выполняющих классификацию. Точностью относительно темы $Y = \{t_1,...,t_T\}$ называется количество правильно отнесенных документов к теме Y по отношению к общему числу документов, отнесенных алгоритмом к теме Y.

Обозначим t_e тему, которые определили эксперты как доминирующую для данного документа, а t_a — аналогичную тему, определенную алгоритмом:

$$\mathrm{Precision} = P_Y = \frac{\#\{d \in D : t_e = t_a = Y\}}{\#\{d \in D : t_a = Y\}}, \qquad (10)$$

знак # означает мощность множества.

Определим полноту относительно темы Y как долю правильно классифицированных документов среди общего числа документов из темы Y:

$$\text{Recall} = R_Y = \frac{\#\{d \in D : t_e = t_a = Y\}}{\#\{d \in D : t_e = Y\}}.$$

Когда коллекция содержит T тем, усредним значения P_Y и R_Y по всем темам. Для обобщенного показателя, включающего в себя полноту и точность, используют F-меру:

$$F = \frac{2PR}{P+R}. \qquad (11)$$

Вообще по отношению к алгоритмам PLSA и LDA, которые сопоставляют каждому документу вектор долей всех тем в данном документе, рассматривать точность и полноту нецелесообразно. Наша дальнейшая задача — построить иерархию тем и сравнить её с экспертной иерархией, а для этого в дальнейшем будем считать доминирующую тему документа единственной темой этого документа.

5 Вычислительный эксперимент

Проверим правильность работы описанных выше алгоритмов сначала на тестовой коллекции. Возьмем три народные сказки — «Теремок», «Колобок» и «Репка», в каждой из которых есть повторяющиеся части. Эти части будем считать отдельным документом. Каждую сказку будем считать отдельной темой. Словарь и каждый из документов можно посмотреть в приложенных к статье файлах [13].

Тестовая коллекция содержит 17 документов, 119 терминов в словаре и три темы. Оба алгоритма не допустили ни одной ошибки при кластеризации, т.к. тестовая коллекция содержит небольшое число документов с заданными тематиками.

Проанализируем работу алгоритмов PLSA и LDA на реальных данных — коллекции тезисов конференции EURO-2012. Для иллюстрации работы обоих алгоритмов построим следующие графики, см. рис. 1, 2. По оси OX отложим номера тем, которые сопоставили эксперты каждому документу, а по оси OY номера тем, которые сопоставил этим документам алгоритм. Каждый документ соответствует некоторой точке. Несколько документов показано в виде круга, радиус которого зависит от их количества. На рис. 1 круг наибольшего радиуса соответствует шести документам. В случае отсутствия ошибок кластеризации получим диагональ, состоящую из кругов. Ошибки кластеризации показаны как точки вне диагонали.

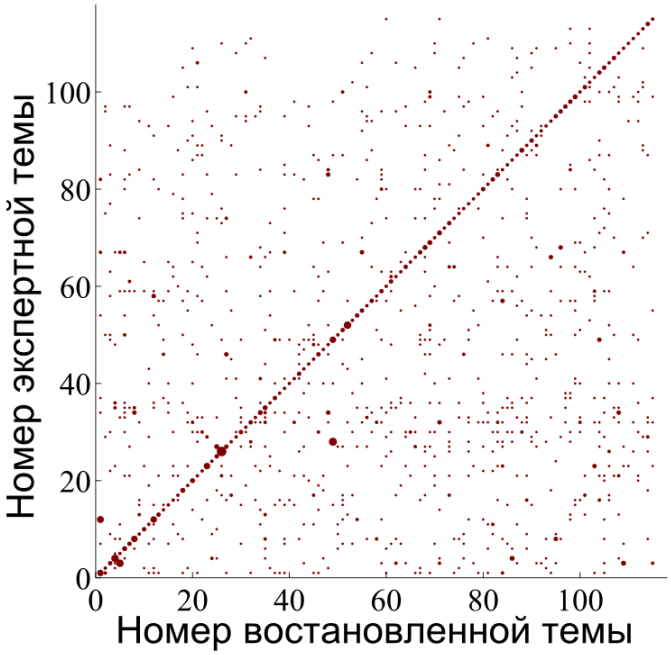


Рис. 1: Сравнение номеров тем, определенных алгоритмом PLSA и экспертами.

В алгоритме PLSA на диагонали лежит 324 документа, а в алгоритме LDA — 400 документов из 1341.

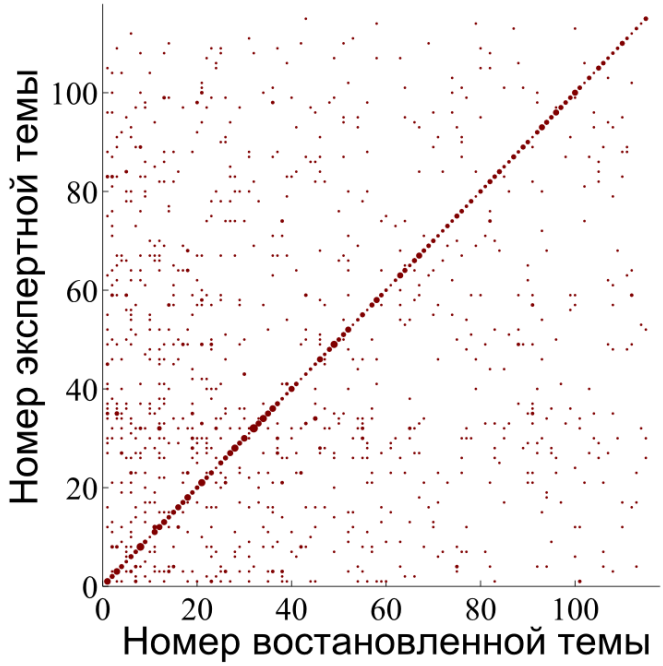
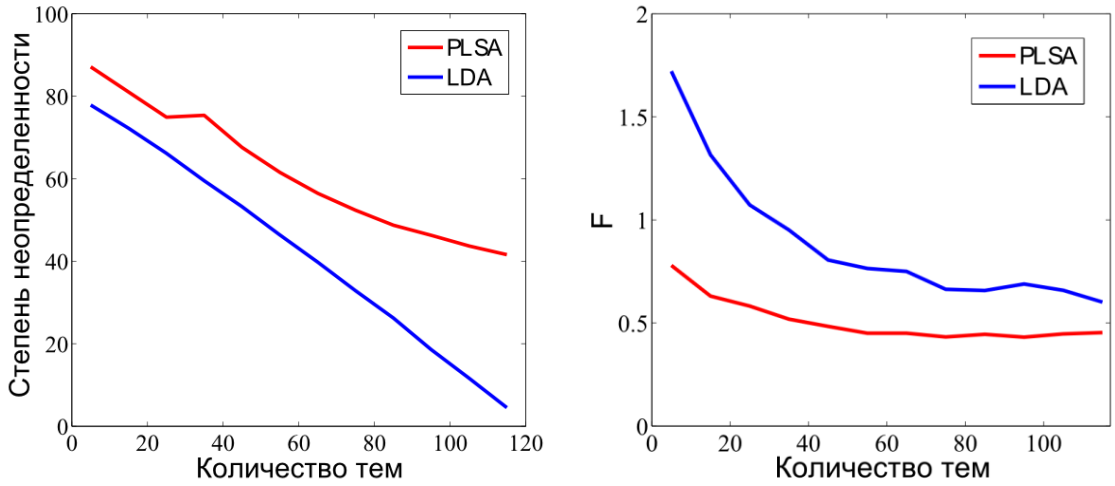


Рис. 2: Сравнение номеров тем, определенных алгоритмом LDA и экспертами.

Сравним работу алгоритмов PLSA и LDA по степени неопределенности

(10). На рис. 3а показан график зависимости степени неопределенности коллекции документов (отложена по оси ординат) от количества тем для алгоритмов PLSA и LDA (отложено по оси абсцисс).



[а] [б]

Рис. 3: Сравнение алгоритмов PLSA и LDA.

Как видно из рис. 3а, правдоподобие выборки алгоритма PLSA при любом количестве тем выше, чем для алгоритма LDA. В данном случае это связано с тем, что начальные приближения для работы алгоритма PLSA были получены с использованием экспертной кластеризации (см. описание алгоритма). Алгоритм LDA не использует никаких экспертных оценок. Таким образом, в графике нас интересует только общая тенденция — правдоподобие падает с ростом количества тем.

Сравним алгоритмы по полноте (11). На рис. 3б показан график зависимости величины F от количества тем. Как видно из рис. 3б, алгоритм LDA кластеризует документы лучше, чем алгоритм PLSA при любом количестве тем.

Обсуждение алгоритмов. Итак, проанализировав результаты кластеризации при различном количестве тем, можно сделать следующие выводы. Во-первых, с ростом числа тем правдоподобие коллекции документов падает. Во-вторых, алгоритм LDA проводит лучшую кластеризацию, чем PLSA, не используя никаких предварительных экспертных оценок. Известны некоторые дополнительные недостатки PLSA по сравнению с LDA: трудоемкость $D \times W$ и чувствительность к начальному приближению.

Иерархическая кластеризация. Для построения простейшей иерархии будем использовать те же алгоритмы, что и для плоской кластеризации, только со следующими предположениями. Теперь, когда каждое слово относится к какой-нибудь теме с наибольшей вероятностью, не будем различать слова, относящиеся к одной и той же теме, т.е. перейдем от понятия «мешка слов» к понятию «мешок тем». Добавим ещё к нашим данным два стихотворения про снег и попробуем

разбить наши документы на два направления (интуитивно этими направлениями являются: 1) стихи про снег и 2) сказки про животных, деда и бабу). Результат показан на рис. 4.

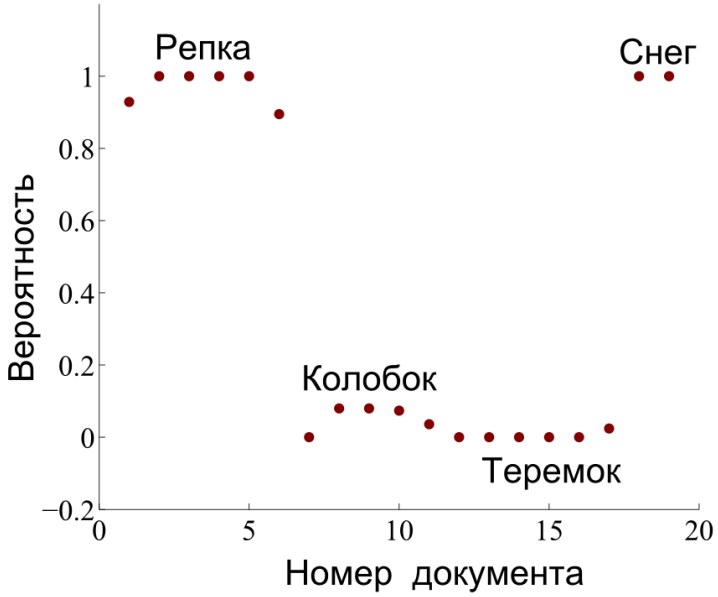
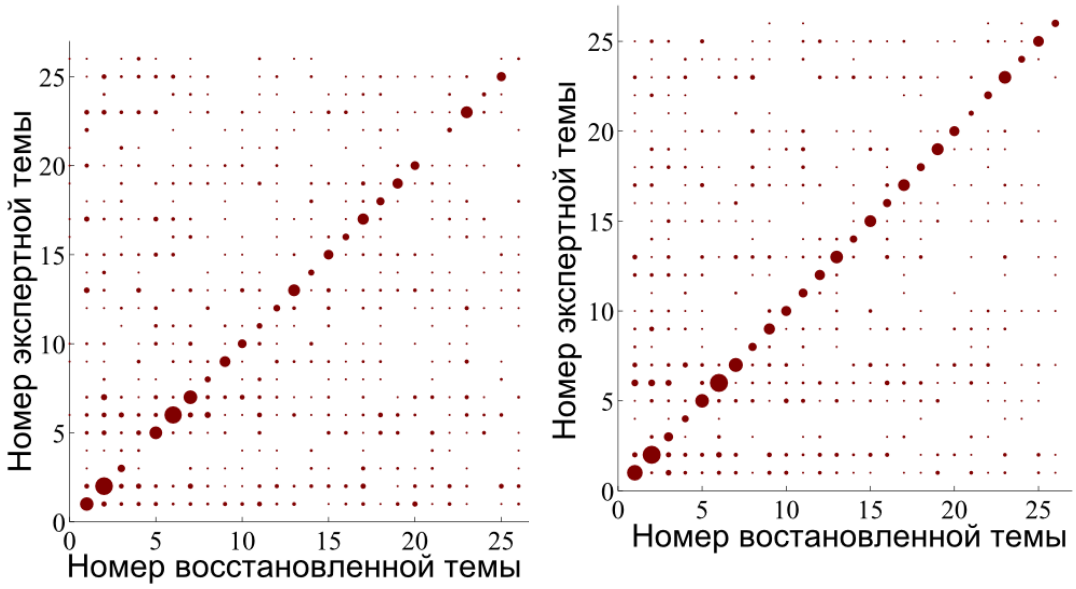


Рис. 4: Сравнение номеров тем, определенных алгоритмом LDA и экспертами.

По оси абсцисс указан номера документов . По оси ординат указана вероятность того, что документ принадлежит заданной теме. Заметим, чтово-первых, все сказки имеют общих персонажей и, соответственно, во многом схожие действия последних. Таким образом, словари сказок довольно схожи. Во-вторых, словари сказки «Репка» и стихов про снег пересекаются по очень малому числу терминов (2-3 из 119).

Согласно экспертной кластеризации, все сказки относятся к одному направлению, а стихотворения — к другому. Из рис. 4 видно, что экспертная и алгоритмическая кластеризации не совпадают. Учитывая свойства коллекции документов, сделаем вывод о том, что алгоритмом ко второму направлению были отнесены документы, имеющие небольшое количество общих слов. Это означает, что некорректное задание числа направлений влечет неадекватную кластеризацию в случае жесткой модели кластеризации.

Посмотрим, что получится применительно к тезисам конференции Euro-2012.



[а] [б]

Рис. 5: Результаты работы алгоритмов на коллекции EURO-2012.

Рис. 5, построенный аналогично рис. 2 показывает результаты инрерхического тематического моделирования тезизов конференции EURO-2012. Рис. 5а показывает результаты, полученные алгоритмом PLSA. На диагонали лежит 463 документаиз 1341. Рис. 5б показывает результаты, полученные алгоритмом LDA. На диагонали лежит 580 документов. Это означает достаточно высокое соответствие полученной тематической модели ожиданиям экспертов.

6 Заключение

В работе были исследованы алгоритмы PLSA и LDA для кластеризации коллекции документов по темам. Было выяснено, что алгоритм LDA имеет существенные преимущества перед алгоритмом PLSA. На основании плоской кластеризации документов был предложен метод построения иерархии документов при помощи перехода от понятия «мешка слов» к «мешку тем». Получены результаты построения иерархической кластеризации для тезисов конференции EURO-2012. Полученная кластеризация в значительной мере соответствует ожиданиям экспертов.

Авторы выражают благодарность А. А. Адуенко за подготовку данных для проведения вычислительного эксперимента.

Список литературы

[1] *Hartigan J. A., Wong M. A.* Algorithm as 136: A k-means clustering algorithm // Applied statistics, 1978. Vol. 28. P. 100–108.

[2] *Pal N. R., Bezdek J. C.* On cluster validity for the fuzzy c-means model // IEEE Transactions on Fuzzy Systems, 1995. Vol. 3(3). P. 370–379.

[3] *Tibshirani R., Hastie T.* Discriminative adaptive nearest neighbor classification // IEEE transactions on pattern analysis and machine intelligence, 1996. Vol. 18, No. 6. P. 606–616

[4] *Peng J., Gunopulos D., Domenciconi C.* An adaptive metric machine for pattern classification // Advances in Neural Information Processing Systems 13. MIT Press, 2000. P. 458 –464.

[5] *Hofmann T.* Probabilistic latent semantic analysis // Proceedings of the 22nd annual interanational ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999. P. 50–57.

[6] *Blei D. M.* Introdaction to Probabilistic Topic Models. http://www.cs.princeton.edu/ blei/papers/Blei2011.pdf (20.12.2012).

[7] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. Vol. 3, P. 993-1022.

[8] *Адуенко А., Кузьмин А., Стрижов В.* Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия Тульского государственного университета. Естественные науки, 2012. Вып. 4. Стр. 119–131.

[9] *Кропотов Д. А.* Тематические модели для коллекций текстов. http://www.machinelearning.ru/wiki/images/8/82/BMMO11_14.pdf (20.12.2012).

[10] *Steyvers M., Griffiths T.* Probabilistic Topic Models // Latent SemanticAnalysis: A Road to Meaning. Berlin: Laurence Erlbaum, 2007.

[11] *Steyvers M., Griffiths T.* Finding Sientiffic topics // Proceedings of the National Academy of Science, 2003. Vol. 101, P. 5228–5235.

[12] *Воронцов К. В.* Вероятностные тематические модели. http://www.machinelearning.ru/wiki/ (20.12.2012).

[13] *Цыганова С. В.* Тестовая выборка для построения иерахической тематической модели. https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/Tsyganova2012Topic Ierarhy/ (20.12.2012).

Построение иерархических тематических моделей коллекции документов.[2]

С. В. Цыганова, В. В. Стрижов

Аннотация. Данная работа посвящена выявлению тематик коллекции текстов и их иерархической структуры. Поставлена задача построения иерархической тематической модели коллекции документов. Для решения поставленной задачи предлагается использование вероятностных тематических моделей. Особое внимание уделяется иерархическим тематическим моделям и, в частности, обсуждению свойств алгоритмов PLSA и LDA. Особенность построения иерархической модели заключается в переходе от понятия «мешка слов» к «мешку документов» в реализации плоских алгоритмов кластеризации. Работа алгоритмов иллюстрируется на текстах тезисов конференции EURO-2012 и на синтетических данных.

Ключевые слова: тематическая модель, иерархические модели, сэмплирование Гиббса, латентный семантический анализ

The construction of hierarchical thematic models for documents' collection.

Tsyganova S.V., Strijov V. V.

Abstract. This work is devoted to detection themes of documents' collection and to their hierarchical structure. The main task is to construct hierarchical thematic model for documents' collection. To solve this task it's suggested to use probabilistic topic models. The main attention is paid to hierarchical thematic models and, particulary, to discuss the properties of PLSA and LDA algorythms. The peculiarity of construction of hierarchical model is the crossing from the conception of «bag of words» to conception of «bag of themes». The work is illustrate on theses of Euro-2012 conference and on synthetic data.

Keywords: thematic models, hierarchical models, Gibbs sampling, latent semantic analysis.

[2] Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.