Kuzmin A.A., Strijov V.V. Validation of thematic models for document collections // Software Engineering, 2013, 4 : 16-20.

УДК 519.256

Проверка адекватности тематических моделей коллекции документов[1]

А. А. Кузьмин, В. В. Стрижов

Аннотация. Рассматривается коллекция документов с экспертной тематической моделью. Для проверки адекватности экспертной модели предлагается построить алгоритмическую модель путем иерархической кластеризации коллекции текстов агломеративным и дивизимным способами. Определяется степень несоответствия экспертной модели и предлагаемой. В работе сравнивается качество моделей, полученных с помощью агломеративного и дивизимного алгоритмов. Визуализируются отличия полученной модели от экспертной.

Ключевые слова: коллекция документов, тематические модели, иерархические модели, кластеризация.

1 Введение

Перед программным коммитетом конференции с большим числом участников встает задача проверки корректности построения иерархической модели тезисов конференции. В силу большого числа экспертов, субъективности экспертной кластеризации и отсутствия эталонной модели, оценить качество экспертной иерархической модели сложно. Поэтому предлагается построить иерархическую модель коллекции тезисов, основыванную на их терминологическом сходстве, и сравнить результат с экспертной моделью. Для построения модели предлагается использовать метрический алгоритм кластеризации.

Существует два основных типа алгоритмов построения иерархической системы кластеров [1]. Дивизимные (нисходящие) алгоритмы на каждом шаге

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

разбивают имеющиеся кластеры на более мелкие. Агломеративные (восходящие) алгоритмы, на каждом шаге объединяют имеющиеся кластеры в более крупные.

В данной работе для построения тематических моделей используются и дивизимные и агломеративные алгоритмы. При построении иерархической модели проводится «жесткая» кластеризация, согласно которой объект принадлежит только к одному из кластеров, так как каждый документ может принадлежать только одной теме, что соответствует правилам проведения конференции. Для построения предлагается использовать Требуется построить иерархическую модель, сохранив ее схожесть с экспертной, и выделить набор документов, которые попадают в кластеры, отличающиеся от экспертных. Функционал сходства моделей задан числом различий между построенной моделью и моделью, заданной экспертно.

Для построения модели предлагается использовать методы иерархической кластеризации. Для кластеризации множества документов вводится функция расстояния [2]. Каждому документу ставится в соответствие метрический вектор, содержащий информацию о словарном составе этого документа. Для того, чтобы оставить в документе только те слова, которые несут информацию о его сходстве и отличии от других документов, предварительно проводится предобработка документов. Слова приводятся к начальной лексической форме [3], удаляются знаки препинания. Исключаются слова, встречающиеся малое количество раз, а также слова, встречающиеся в большинстве документов. Для этого используется критерий tf · idf (англ. tf — term frequency, idf — inverse document frequency) [4], а также словарь стоп-слов. Затем из оставшихся слов составляется словарь, а документы представляются в виде «мешков слов» [5].

После отсева неинформативных слов из словаря и документов, каждому документу ставится в соответствие булевый вектор. В работе [3] сравниваются способы построения вектора — описания документа.

Кластеризация текстов при построении каждого уровня иерархической

модели может проводиться как с помощью метрических алгоритмов кластеризации, например, K-means [6], FOREL [7], C-means [8], STOLP [9], FRiS-STOLP [10], BoostML, DANN [11] и другие [12, 13], так и с помощью вероятностных методов [14], например с помощью вероятностного латентного семантического анализа (англ. PLSA — probabilistic latent semantic analysis) [5], или латентного размещения Дирихле (англ. LDA — latent Dirichlet allocation) [2]. В данной работе применяется метрический алгоритм кластеризации, подробно описанный в [3].

Результаты работы дивизимного и агломеративного алгоритма иллюстрируются примером тематической иерархической кластеризации тезисов конференции «European Conference on Operational Research, EURO–2012». Предложенная модель сравнивается с экспертной.

2 Постановка задачи

Пусть $W = \{w_1, \ldots, w_n\}$ — заданное множество слов (словарь), где n — количество слов в словаре. Документом d из коллекции D назовем неупорядоченное множество слов из W, $d = \{w_j\}$, где $j \in \{1, \ldots, n\}$.

Поставим в соответствие каждому документу d его описание — булевый вектор $\mathbf{x}$ размерности n следующим образом. Если слово w_j из словаря W встретилось в документе d_s хотя бы раз, то $x_{s,j} = 1$, иначе $x_{s,j} = 0$. Получим матрицу $\mathbf{X}$ «объект-признак», где каждая строка $\mathbf{x}_s = [x_{s,1}, \ldots, x_{s,n}]$ — признаковое описание документа d_s.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \ldots & x_{1,n} \\ \ldots & \ldots & \ldots \\ x_{|D|,1} & \ldots & x_{|D|,n} \end{pmatrix}. \qquad (1)$$

В качестве функции расстояния ρ между документами возьмем

евклидову метрику:

$$\rho(\mathbf{x},\mathbf{x}') = \sqrt{\sum_{j=1}^{n}\left(x_j - x'_j\right)^2}. \qquad (2)$$

Представим экспертную иерархическую модель в виде дерева (см. рис. 1). Глубину дерева обозначим h. Уровнем l иерархии назовем множество всех узлов дерева, находящихся на глубине l. Каждый внутренний узел дерева обозначим $c_{i,l}$, где l — уровень, к которому принадлежит данный узел, а i — номер этого узла среди узлов на данной глубине. Документы d являются листами этого дерева и имеют уровень h. Под кластером $c_{i,l}$ будем понимать множество дочерних элементов данного узла, $c_{i,l} = \{c_{i,l+1}\}$ для некоторых i. Будем говорить, что документ d_s принадлежит кластеру $c_{i,l}$, если путь до данного документа от вершины проходит через узел $c_{i,l}$.

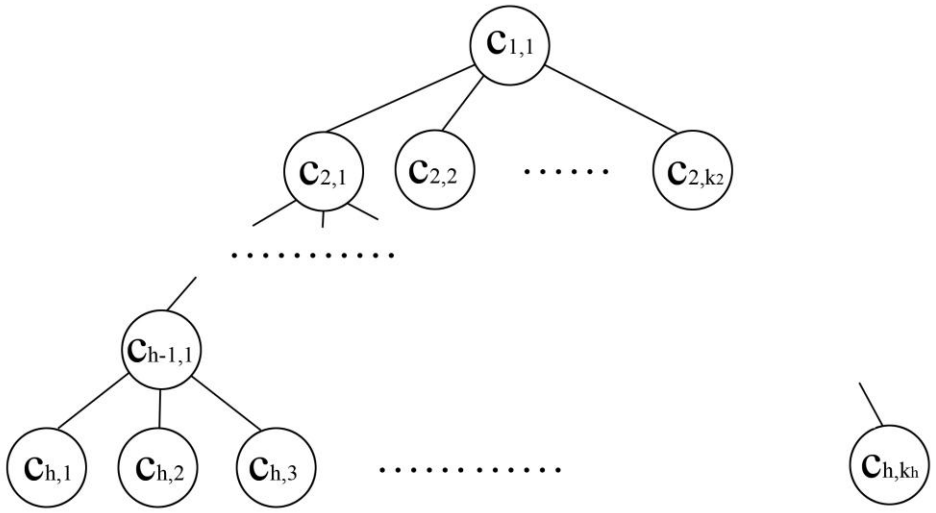


Рис. 1: Иерархическое представление тематической модели.

Зададим путь от корня $c_{1,1}$ до документа d_s следующим образом: $\mathbf{g}^s = (c^s_{i_1,1}, c^s_{i_2,2}, \ldots, c^s_{i_h,h})$, где $c^s_{i_l,l}$ — порядковый номер узла уровня l через который

проходит путь. Символом $\mathbf{g}^s$ обозначим путь, построенный по экспертной модели, и $\hat{\mathbf{g}}_l^s$ — путь, построенный по алгоритмической модели. Обозначим $\boldsymbol{\mu}(\mathbf{g}_l^s)$ — координаты центра кластера, соответствующего документу s на уровне l иерархической модели. Зададим функционал сходства экспертной и построенной иерархических моделей:

$$\mathsf{L} = -\sum_{s=1}^{|D|}\sum_{l=1}^{h}[\mathbf{g}_l^s \neq \hat{\mathbf{g}}_l^s]\cdot\rho(\boldsymbol{\mu}(\mathbf{g}_l^s),\boldsymbol{\mu}(\hat{\mathbf{g}}_l^s)) \to \min, \qquad (3)$$

где индикаторная функция $[\mathbf{g}_l^s \neq \hat{\mathbf{g}}_l^s]$ равна 1, если все элементы векторов $\mathbf{g}_l^s$ и $\hat{\mathbf{g}}_l^s$ совпадают. В противном случае функция равна 0. Оценивая степень схожести кластеров, основываясь на расстоянии ρ между их центрами $\boldsymbol{\mu}$, будем штрафовать алгоритмическую модель тем сильнее, чем больше расстояние между центром кластера $\boldsymbol{\mu}(\hat{\mathbf{g}}^s)$, к которому алгоритм отнес документ, и центром кластера $\boldsymbol{\mu}(\mathbf{g}^s)$, к которому эксперт отнес документ. Задача построения иерархической модели сводится к задаче минимизации функционала (3).

3 Описание алгоритма

Опишем алгоритм тематической кластеризации для случая единственного уровня иерархии тематической модели. Алгоритм итеративно находит центры кластеров $\boldsymbol{\mu}$, после чего относит документ $\mathbf{x}_s$ к тому или иному кластеру исходя из принципа «ближайшего соседа». Зафиксируем число кластеров C равное числу тем в экспертной модели. Центр $\boldsymbol{\mu}_y$ кластера с номером y вычислим по формуле (4).

$$\hat{\boldsymbol{\mu}}_y = \frac{\sum_{s=1}^{|D|}[y_s = y]\cdot\mathbf{x}_s}{\sum_{s=1}^{|D|}[y_s = y]}. \qquad (4)$$

Координаты вычисленных кластеров обозначим $\hat{\boldsymbol{\mu}}_y$. В качестве начального приближения положений центров кластеров выберем центры экспертных

кластеров $\hat{\boldsymbol{\mu}}_y = \boldsymbol{\mu}_y$, $y \in \{1,\ldots,C\}$. Отнесем каждый документ к кластеру:

$$\hat{y}_s = \underset{y\in\{1,\ldots,C\}}{\operatorname{argmin}} \rho(\mathbf{x}_s, \boldsymbol{\mu}_y), \quad s = 1,\ldots,|D|. \tag{5}$$

Каждая итерация алгоритма кластеризации состоит из двух шагов. На первом шаге координаты центров кластеров $\boldsymbol{\mu}_y$ пересчитываются по формуле (4). В качестве функции ρ расстояния между документами используется евклидова метрика (2). На втором шаге для каждого документа $\mathbf{x}_s$ с номером s решается задача (5), после чего документ $\mathbf{x}_s$ относится к кластеру с номером $\hat{y}_s$. Алгоритм останавливается, когда координаты $\boldsymbol{\mu}_y$ на очередной итерации не поменяются после пересчета (4) для всех $y \in \{1,\ldots,C\}$.

Для построения иерархической модели предлагается использовать два альтернативных алгоритма: дивизимный и агломеративный, а затем сравнить полученные результаты с помощью функционала (3).

Обозначим $\boldsymbol{\mu}(c_{i,l})$ — вектор координат центра кластера, соответствующего узлу дерева $c_{i,l}$ (см. рис. 1). Через $\boldsymbol{\mu}(c_{i,l})$ будем обозначать вектор координат центра кластера $c_{i,l}$, полученный из экспертной модели. Значения $\boldsymbol{\mu}(c_{i,l})$ можно найти, используя в формуле (4) в качестве множества $\{\mathbf{x}_s\}$, $s \in \{1,\ldots,|D|\}$ множество центров кластеров $\boldsymbol{\mu}(c_{i,l-1})$, $i \in \{1,\ldots,k_{l+1}\}$, а y_s будет определяться по формуле (5), где множество y будет множеством центров кластеров уровня l. На уровне $l = h$ объектами будут описания $\mathbf{x}_s$ документов d_s, чьи векторы координат есть строки матрицы $\mathbf{X}$ (1).

Агломеративный алгоритм. Алгоритм строит уровни иерархической модели, обходя дерево снизу вверх. На шаге l центры кластеров $\{\hat{\boldsymbol{\mu}}_{h-l+1}\}$ уровня $h-l$ рассматриваются в качестве объектов. К ним применяется описанный выше алгоритм. В качестве начального приближения центров кластеров $\hat{\boldsymbol{\mu}}_y$ берутся

координаты центров экспертных $\boldsymbol{\mu}_y$ кластеров уровня $h-l$. Таким образом на шаге l мы строим уровень иерархии $h-l$.

Дивизимный алгоритм. Алгоритм строит уровни иерархической модели, идя по дереву сверху вниз. На шаге l для каждого кластера $c_{i,l}$, $i \in \{1 \ldots k_l\}$ применяется описанный выше алгоритм использующий в качестве объектов множество $\{\mathbf{x}_s \mid y_s = i\}$, где y_s находится из условия (5). В качестве начального приближения положения центров кластеров берутся центры экспертных кластеров уровня $l+1$, лежащих внутри кластера $c_{i,l}$, $\hat{\boldsymbol{\mu}}_{m,h-l+1}$ $\{m \mid c_{m,l} \in c_{i,l}\}$. В итоге, каждый кластер $c_{i,l}$ делится на более мелкие кластеры:

$$c_{i,l} = \coprod_{m \in c_{i,l}} c_{m,l+1}.$$

На первом шаге в роли кластеров выступает единственный кластер $c_{1,1}$, состоящий из всех документов коллекции.

4 Вычислительный эксперимент

Построим иерархические тематические модели для коллекции тезисов конференции «European Conference on Operational Research, EURO–2012» с помощью агломеративного и дивизимного алгоритмов. Данная коллекция из 1342 документов иерархически рабивается на области (area) и направления (stream), как показано на рис. 2. Сравним полученные модели с экспертной моделью. Более адекватной считается алгоритмическая модель, которая имеет меньшее значение функционала ошибки (3).

Найдем число несоответствий между моделями: 1) количество документов, которые относятся экспертной алгоритмической моделью к различным областям, 2) количество документов,которые относятся экспертной алгоритмической моделью к различным направлениям и 3) количество документов, у которые относятся моделями к разным областям и разным направлениям. Полученные результаты для обоих алгоритмов приводятся в таблице 1.

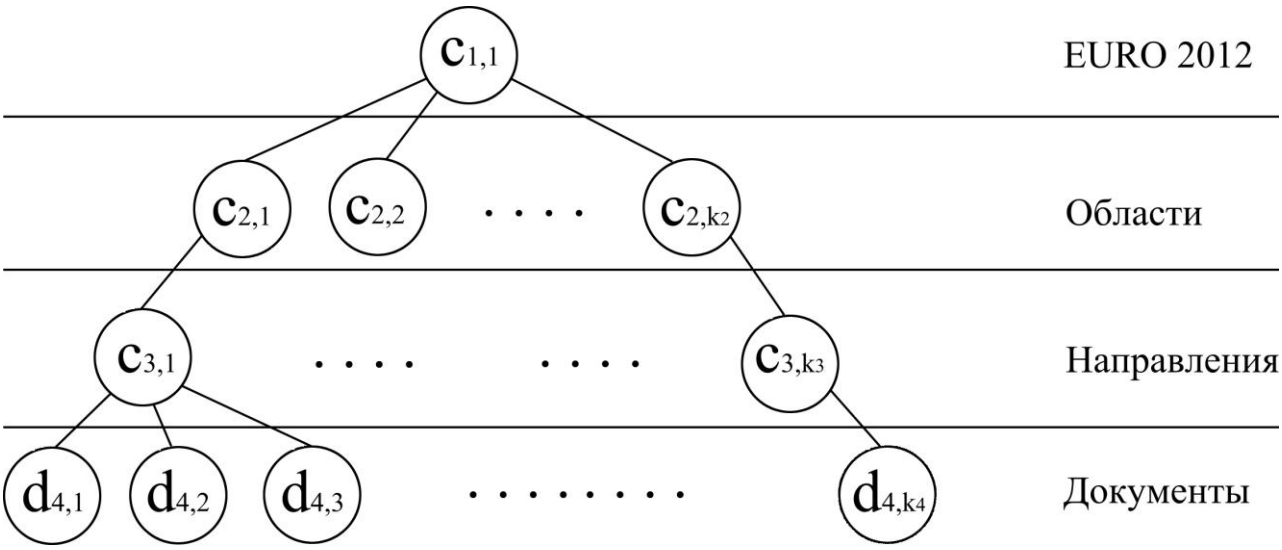


Рис. 2: Пример иерархического представления тематической модели.

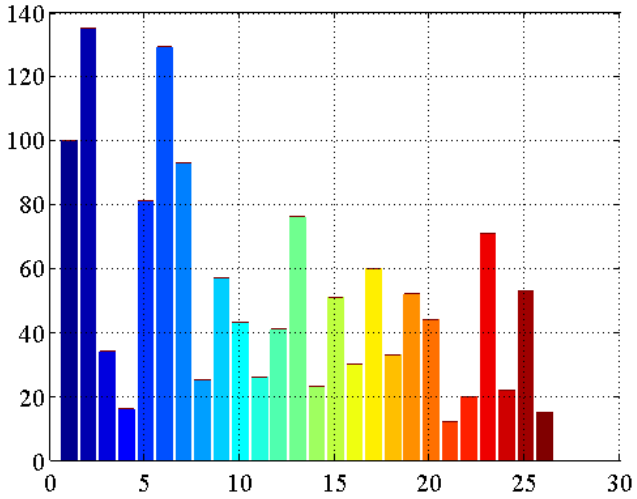
На рис. (3) показано распределение документов по областям. Количество столбцов гистограммы совпадает с количеством кластеров уровня «область». Каждому документу присваивается кодовый цвет области, к которой он отнесен. Высота части столбца одного цвета показывает количество документов. Рис. 3а построен по экспертной кластеризации, поэтому каждый столбец содержит документы только одного цвета. Этот цвет одновременно является цветом области (кластера) с номером данного столбца. Результат работы дивизимного алгоритма изображен на рис. 3б, а результат работы агломеративного алгоритма на рис. 3в.

Таблица 1: Значение функции ошибки для разных способов построения иерархической модели

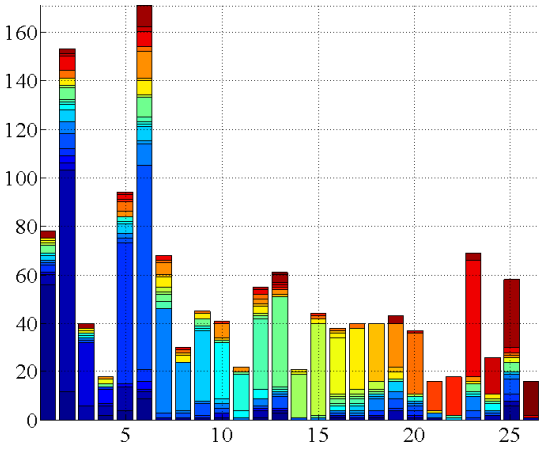
Вид алгоритма	Количество несоответствий на уровне	Количество несоответствий на уровне	Количество несоответствий в обоих	Значение функционала (3)

	«область»	«направление»	уровнях	
Дивизимный	554	555	550	1700
Агломеративный	342	208	182	500

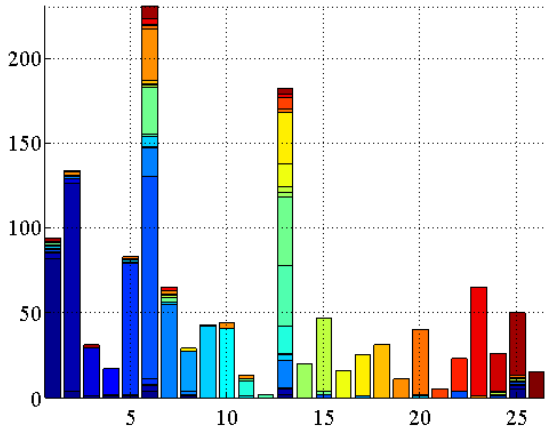
Номера областей (столбцов гистограмм) во всех трех графиках фиксированы. Алгоритические модели перераспределяют документы по темам, вследствие чего каждый столбец содержит документы различных кодовых цветов. Чем выше высота сегмента фиксированного цвета в столбце, тем больше документов из соответствующей области там содержится. Рис. 3в показывает, что в результате использования агломеративного алгоритма, в каждый столбец гистограммы попадают документы из небольшого числа других столбцов. При кластеризации в столбцы попадают не отдельные документы из других столбцов, а целые кластеры уровня «направление». Таким образом изменение модели носит более систематический характер и отличия алгоритмической модели от экспертной легче интерпритировать. Но с другой стороны, при использовании агломеративного алгоритма появляются два кластера (6–ой и 13–ый), к которым вместо части документов из других областей были отнесены сразу целые направления, в результате чего количество документов в этих областях стало отличаться от экспертного почти в два раза.



(а) Экспертная классификация.



(б) Дивизимный алгоритм.



(в) Агломеративный алгоритм.

Рис. 3: Распределение документов по темам для экспертной классификации и для построенных кластеризаций.

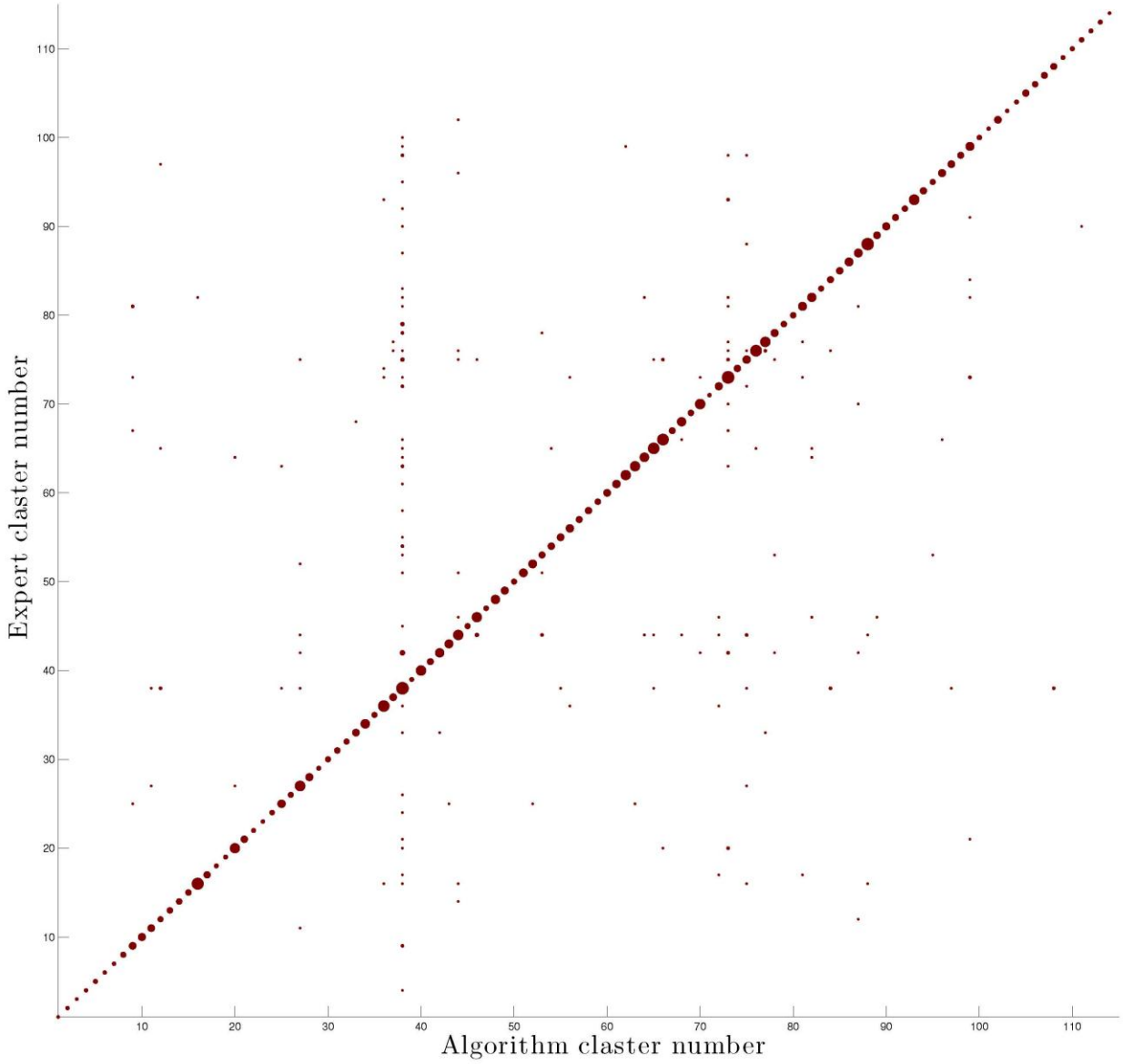


Рис. 4: Перераспределение документов по направлениям для модели, построенной агломеративным алгоритмом.

Перераспределение документов по кластерам уровня «направление» после применения агломеративного алгоритма показаны на рис. 4. Для каждого документа координатой по оси абсцисс является номер направления, к которому отнес алгоритм данный документ, по оси ординат — его экспертное направление. Радиус круга определяется количеством документов, попавших в эту точку.

Наибольшему кругу соответствует количество документов, равное 28. Заметим, что основное число документов попадает на диагональ. Таким образом, у большинства документов номер экспертного направления совпадает с алгоритмическим, что соответствует нашему требованию к схожести построенной метрической модели с экспертной.

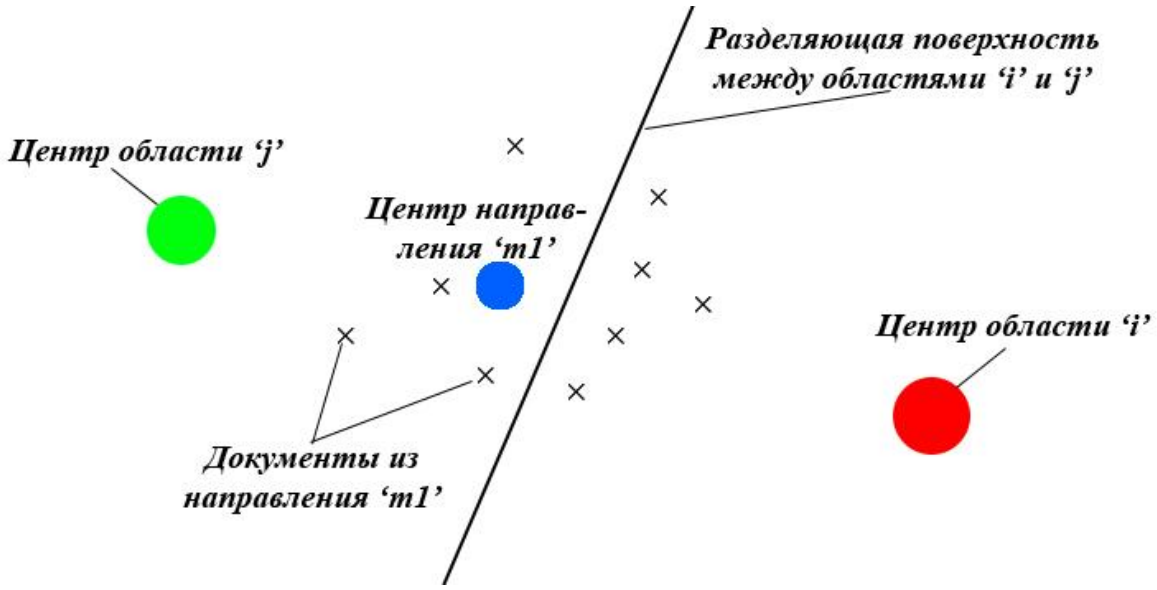


Рис. 5: Направление, попавшее близко к разделяющей поверхности между областями.

При сравнении с экспертной моделью результата работы дивизимного алгоритма, видно, что почти у всех документов у которых не совпала область, не совпало и направление 0. Документ с номером s, попавший в i-ю область, уже не может попасть в направление из j-ой области, даже если с метрической точки зрения, центр этого направления ближе к документу s, чем все центры направлений из i-й области (см. рис. 2). Поэтому документы из тех направлений, центры которых попадают на разделяющую поверхность между различными областями, относятся к разным областям (см. рис. 5). Но так как направление

может быть отнесено только к одной области, то на всех остальных документах, попавших в другую область будет допущена ошибка.

Таким образом, результат получается лучше, если применять агломеративный алгоритм. В этом случае сперва учитывается терминологическое сходство документов при объединении их в направления. Затем каждое направление относится к области, учитывая уже усредненную информацию о терминологии всех документов, попавших в данное направление.

При создании тематической модели с использованием информации не только об экспертном соответствии документов областям, но еще и об экспертном соответствии документов направлениям, позволяет улучшить сходство алгоритмической модели с моделью с экспертно построенными областями на 15% в сравнении с количеством ошибок, допущенных алгоритмом из [3], использующим лишь информацию об экспертной кластеризации на области.

5 Заключение

В данной статье исследовалась проблема проверки адекватности тематической модели. Экспертная модель сравнивалась с моделью, построенной метрическими алгоритмами кластеризации. Были предложены различные метрические способы построения моделей, схожих с экспертной. Наилучшим из предложенных способов является использование агломеративного алгоритма построения иерархической тематической модели. На практике при построении экспертных тематических моделей результаты алгоритмической кластеризации используются для выявления возможных противоречий в действиях экспертов.

Список литературы

[1] Воронцов К.В., Потапенко А.А. Робастные разреженные вероятностные тематические модели // Интеллектуализация обработки информации. Доклады 9-й международной конференции, 2012 С. 605-608.

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

[3] *Кузьмин А. А., Адуенко А. А., Стрижов В. В.* Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия Тульского государственного университета, Естественные науки, 2013.

[4] *Blei D. M., Lafferty J. D.* Topic Models. // Text Mining: Classifcation, Clustering, and Applications. Chapman & Hall/CRC Press, 2009.

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

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

[7] *Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С.* Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985.

[8] *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.

[9] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999, С. 270.

[10] *Борисова И. А.* Использование fris-функции для построения решающего правила и выбора признаков (задача комбинированного типа dx) // Новосибирск. Знания. Онтологии. Теории. Материалы Всероссийской Конференции, 2007. Т. 1. С. 37–44.

[11] *Tibshirani R., Hastie T.* Discriminative adaptive nearest neighbor classification. // IEEE transactions on pattern analysis and machine intelligence, VOL. 18, NO. 6, JUNE 1996.

[12] *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.

[13] *Zagoruiko N. G.* Methods of recognition based on the function of rival similarity. // Pattern recognition and image analysis, 2008. Vol. 18(1). P. 1–6.

[14] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models. // Frontiers of computer science in China, 2010. Vol. 4(2). P. 280–301.

[15] *Manning C. D., Raghavan P., Schütze H.* Introduction to information retrieval. // Cambridge: Camdridge University Press, 2008.

[16] *Loohach R., Garg K.* Effect of distance functions on simple k-means clustering problem // International Journal of Computer Applications, 2012. Vol. 49. No. 6 P. 7–9.

A. A. Kuzmin Moscow Institute of Physics and Technology

V. V. Strijov, Computing Center of the Russian Academy of Sciences

Validation of the thematic models for document collections

Consider a collection of documents with expert thematic model. To verify the adequacy of the expert model build an algorithmic model by hierarchical clustering text collections. The agglomerative and divisive clustering methods are investigated. The algorithmic model error in comparison to the expert model is estimated. The differences between expert model and algorithmic model are visualized.

Key words: document collection, thematic model, hierarchical model, clustering.