Обзор и экспериментальное сравнение методов кластеризации текстов[1]

[1,2] *П. А. Пархоменко <parhomenko @ispras.ru>*
[1,3] *А. А. Григорьев <agrigorev@ispras.ru>*
[1] *Н. А. Астраханцев <astrakhantsev@ispras.ru>*
[1] *Институт системного программирования РАН,*
109004, Россия, г. Москва, ул. А. Солженицына, д. 25
[2] *Московский государственный университет имени М.В. Ломоносова,*
119991, Россия, Москва, Ленинские горы, д. 1
[3] *Национальный исследовательский университет «Высшая школа экономики»*
101000 Россия, Москва, ул. Мясницкая, д.20

Аннотация. Кластеризация текстовых документов применяется во многих приложениях, таких как информационный поиск, исследовательский поиск, определение спама. Этой задаче посвящено множество научных работ, однако в настоящее время остается недостаточно изученным влияние специфики научных статей, в частности принадлежности документов одной предметной области или недоступности полных текстов, на эффективность кластеризации. В данной работе предлагаются обзор и экспериментальное сравнение методов кластеризации текстовых документов в приложении к научным статьям. Исследуются методы, основанные на мешке слов, извлечении терминологии, тематическом моделировании, а также векторном представлении слов (word embedding) и документов, полученном с помощью искусственных нейронных сетей (word2vec, paragraph2vec).

Ключевые слова: кластеризация текстовых документов; мешок слов; извлечение терминологии; тематическое моделирование; векторное представление; искусственные нейронные сети

DOI: 10.15514/ISPRAS-2017-29(2)-6

Для цитирования: Пархоменко П.А., Григорьев А.А, Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, том 29, вып. 2, 2017 г., стр. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6

[1] Эта работа поддержана грантом РФФИ №14-07-00692

1. Введение

Кластеризация текстовых документов, то есть разбиение множества документов на близкие по смыслу подмножества, является фундаментальной задачей обработки текстов. Ее результаты используются как для непосредственного анализа исходного множества документов, так и для информационного поиска [1], определения спама [2], помощи в проведении судебномедицинских экспертиз [3] и социологических исследований [4].

Особого внимания заслуживает кластеризация научных статей. В настоящее время их количество настолько велико, что прочитать все их, даже в одной области знаний, не представляется возможным; более того, возникают серьезные сложности и с самим поиском нужных статей, особенно при отсутствии четкого понимания предметной области или самой информационной потребности.

Возможное решение заключается в навигации на основе кластеров (clustering based navigation) [5] и других методов исследовательского поиска, в которых часто используется кластеризация как один из этапов [6].

К настоящему времени произведено множество обзоров и экспериментальных сравнений методов кластеризации, но в большей их части не рассматриваются современные методы, например векторные представления слов, полученные с помощью нейронных сетей (word embedding), а также не учитывается специфика научных статей, в частности, тот факт, что во многих практических приложениях необходимо кластеризовать статьи, принадлежащие одной предметной области, по более узким направлениям, причем полные тексты статей не всегда доступны.

Данная работа призвана восполнить эти недостатки путем анализа и экспериментального сравнения как классических, так и современных методов кластеризации текстов в приложении к научным статьям. Статья устроена следующим образом. Во втором разделе приводится обзор существующих работ, в том числе других обзоров и экспериментальных сравнений. В следующем разделе описывается методика экспериментальных исследований. Четвертый раздел посвящен результатам экспериментального сравнения и их обсуждению. Далее приводится заключение, подводящее итог статьи и предлагающее направления дальнейшей работы.

2. Существующие обзоры и экспериментальные сравнения

В большинстве обзоров, посвященных методам кластеризации текстов, документы представляются как векторы, в виде мешка слов (bag of words, bow), так что кластеризация документов рассматривается именно как кластеризация bowвекторов.

Так, Andrews и Fox в работе 2007 года [7] описывают способы представления набора документов в виде векторной модели, в том числе различные способы предобработки текстов, а также алгоритмы их кластеризации, такие как

модификации k-means (или метод k-средних), EM-алгоритм и спектральная кластеризация. Так как одним из главных недостатков представления документов в виде мешка слов является высокая размерность и разреженность получаемых векторов, авторы также представляют методы понижения размерности векторного пространства.

Рассматривая разделительные (partitional) алгоритмы кластеризации документов, в частности k-means, более детально, Huang представляет описание и сравнение мер близости между bow-векторами [8]. В статье описаны шесть различных мер близости, между которыми проведено экспериментальное сравнение на алгоритме k-means; лучшие результаты по метрикам чистоты (purity) и энтропии показала кластеризация, использующая в качестве меры близости коэффициент Жаккара (Jaccard coeficient) и коэффициент корреляции Пирсона (Pearson correlation coefficient).

Sathiyakumari и др. [9] также рассматривают кластеризацию документов только применительно к их представлению в виде мешка слов. Они выделяют четыре группы методов кластеризации таких представлений: разделительная кластеризация, иерархическая кластеризация, k-средних и EM-алгоритм, хотя во многих других работах k-means включается в группу разделительных алгоритмов [10, 11].

Как видно в вышеупомянутых работах, кластеризация документов обычно сводится к кластеризации их векторных представлений в виде мешка слов. Кластеризации векторов в общем случае, безотносительно к текстовым документам, также посвящено множество работ [10, 12].

Более широкий спектр возможных векторных представлений документа разбирается в одной из глав книги Mining Text Data [13]. В частности, в ней описываются методы, использующие в качестве признаков документов часто встречающиеся наборы слов, а также методы тематического моделирования. Кроме того, обозреваются подходы к онлайн-кластеризации текстов, использованию графовых методов кластеризации (в случае если между текстами существуют связи) и имеющейся заранее информации для кластеризации на основе алгоритмов частичного обучения (semi-supervised).

В некоторых обзорах авторы отдельно выделяют методы так называемой семантической кластеризации. Saiyad и др. [14] считают определяющим отличием семантической кластеризации от традиционной, основанной на мешке слов, использование семантических отношений между словами документов. Авторы относят к методам семантической кластеризации несколько групп алгоритмов: алгоритмы, основанные на онтологиях, таких как WordNet; алгоритмы, использующие в качестве признаков документа наборы связанных по смыслу слов; а также алгоритмы, основанные на графах концептов или именованных сущностей с семантическими отношениями между ними. Кроме того, к этой группе авторами отнесены алгоритмы, использующие латентное семантическое индексирование, хотя обычно они считаются методами тематического моделирования.

3. Методы кластеризации документов

Как правило, процесс кластеризации текстовых документов можно логически разделить на два основных этапа. На первом этапе текстовые представления документов по определенным правилам переводят в векторные представления, для того чтобы на втором этапе применить к полученным векторам методы кластеризации, основанные на расстоянии между ними.

Ниже будут сначала представлены различные способы отображения документов в векторное пространство, а затем — методы кластеризации векторов. Кроме того, будут описаны возможные варианты предобработки текстов и меры, с помощью которых обычно оценивают эффективность кластеризации.

3.1 Методы на основе bag-of-words

Наиболее простым представлением документов в векторном виде является так называемый мешок слов. В данном случае на основе набора документов строится словарь из всех встречающихся в нем n-грамм, где n меньше или равно какомуто заранее заданному значению. Документ представляется набором признаков, каждому из которых соответствует одна n-грамма из словаря.

BinaryBOW. В простейшем, бинарном, случае данный признак принимает значение 1 в случае, если в документе встречается соответствующая n-грамма, и 0 — иначе.

CountBOW. Предполагая, что значимость появления n-граммы в документе тем больше, чем чаще она в нем появляется, этот метод учитывает количество вхождений n-граммы в документе, помимо самого факта вхождения. Таким образом, каждый признак показывает, сколько раз соответствующая n-грамма появляется в документе.

TF-IDF. Для того чтобы снизить влияние длины текста на его признаки, используется нормализация количества вхождений n-грамм на размер текста. Тогда каждый признак принимает вид частоты n-граммы (term frequency или tf) [15], которая считается как отношение количества вхождений соответствующей n-граммы к общему количеству слов в документе. Поскольку, как правило, значимость появления в документе различных n-грамм различается, применяются различные схемы взвешивания признаков.

Наиболее широко используемая из таких схем — TF-IDF. Она использует предположение о том, что значимость n-граммы прямо пропорциональна частоте ее появления в документе и обратно пропорциональна доле документов в наборе, в которых эта n-грамма встречается. Таким образом, наибольший вес получает n-грамма, часто встречающаяся в одном документе, но не встречающаяся в остальных, а значит — отличающая этот документ от остальных. Признаки документов в этом подходе представляют собой произведение двух величин, частоты n-граммы и обратной частоты документа (inverse document frequency):

$$TF \cdot IDF(t_i, d_j, D) = tf(t_i, d_j) \cdot \log \frac{|D|}{|(d_j \supset t_i)|},$$

где $tf(t_i, d_j)$ — частота n-граммы t_i в документе d_j, D — набор документов, $|(d_j \supset t_i)|$ — все такие документы в наборе, в которых встречается n-грамма t_i .

BM25. Whissel и др. [16] экспериментально показывают, что лучшие результаты в кластеризации текстов демонстрирует другой вариант взвешивания значимости слов: метод BM25. В нем ограничивается значимость частоты n-граммы, а также она не только нормализуется по его размеру, но и ограничивается сверху, что позволяет избежать присваивания слову слишком большого веса [13]. Значение признаков для n-граммы t_i в документе d_j в этом методе рассчитывается по следующей формуле.

$$idf(t_i) \cdot \frac{tf(t_i, d_j) \cdot (k_1 + 1)}{k_1 \cdot (1 - b + b \cdot \frac{|d_j|}{|d_{avg}|}) + tf(t_i, d_j)},$$

где $|d_j|$ — длина данного документа; $|d_{avg}|$ — средняя длина документов в наборе; k_1 и b — свободные параметры.

Стоит отметить, что в методах использующих в качестве признаков n-граммы, как правило учитываются не все из них. Есть несколько способов отбрасывать незначимые n-граммы. Один из них: не учитывать те n-граммы, количество вхождений в наборе документов которых ниже, чем определенный заранее порог. Другой способ: отсортировать все n-граммы в словаре по частоте употребления и учитывать только m первых, где m также задается заранее. Третий способ: не учитывать n-граммы, входящие в слишком большую долю документов из набора, поскольку такие слова, как правило, не несут смысловой нагрузки, позволяющей характеризовать документ. Также можно не учитывать слова, входящие в заранее подготовленный список стоп-слов.

3.2 Bag-of-terms

Голомазов [17] использует термины в качестве признаков документа для кластеризации и классификации документов. В таком подходе значимыми считаются не все n-граммы, а только определенный набор заранее выделенных терминов. При этом, построив матрицу вхождений терминов в документы, с ними можно оперировать точно так же, как с обычными n-граммами в выше описанных методах.

Методы кластеризации с использованием извлеченных терминов полезны в случае документов небольшой длины из узкой предметной области. Например, D. Pinto [18] и др. кластеризуют аннотации научных статей используя оригинальные методы извлечения терминов.

3.3 Тематическое моделирование

Другие способы извлечения признаков из документов часто называют методами тематического моделирования, так как в них каждый результирующий признак можно отнести к определенной теме, представленной в наборе документов. К основным методам тематического моделирования относятся Latent Semantic Analysis (LSA), Nonnegative Matrix Factorization (NMF), Probablistic LSA (pLSA) и Latent Dirichlet Allocation (LDA).

LSA. Метод LSA (также называемый latent semantic indexing, LSI), предложенный Deerwester и др. [19], использует для понижения размерности метод сингулярного разложения (singular value decomposition или SVD). SVD позволяет отобразить данные в новое пространство меньшей размерности, в котором все базисные вектора будут ортогональны, а разброс данных в ортогональной проекции на эти оси — максимальным. В нем изначальная матрица данных X (как правило, TF-IDF в случае кластеризации текстов) раскладывается на 3 следующим образом: $X = U\Sigma V^T$, где U и V — матрицы, состоящие из левых и правых сингулярных векторов матрицы X, а Σ — диагональная матрица, состоящая из сингулярных значений X. Результирующая же матрица, строки которой соответствуют векторам документов, имеет имеет вид $T = XV_l$, где V_l — первые l строк матрицы V, соответствующие l наибольшим сингулярным значениям из Σ.

NMF. Xu и др. [20] предлагают использовать для кластеризации вектора, полученные методом NMF, в котором, как и в LSA, данные отображаются в новое пространство с целью максимизировать разброс по каждой из его осей. Отличия NMF от LSA состоят в том, что в NMF новое пространство может быть не ортогонально, а также принимает только матрицы, в которых все элементы неотрицательны. Как утверждается в оригинальной статье, это позволяет достичь более сильного соответствия между результирующими осями и кластерами документов.

pLSA. В работе Hofmann и др. [21] вводится метод probabilistic latent semantic allocation (pLSA). На вход методу подаются набор слов W, набор документов D, а также количество тем $|T|$ в этом наборе. В результате он генерирует две матрицы. Элементы матрицы $\Phi_{|W|\times|T|}$ соответствуют вероятности того, что определенное слово относится к определенной теме: $\phi_{wt} = p(w|t)$. Элементы $\Theta_{|T|\times|D|}$ соответствуют вероятности того, что определенная тема встречается в документе: $\theta_{td} = p(t|d)$. Эти матрицы строятся с помощью максимизации логарифма правдоподобия следующей функции.

$$L(D, \Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} c(w, d) \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

где $c(w, d)$ — количество вхождений слова w в документе d. Затем в качестве представления документов используются столбцы матрицы Θ.

LDA. Другой метод моделирования документов с помощью тематического моделирования, LDA, предлагается Blei и др. [22]. Они выделяют несколько недостатков pLSA, с которыми справляется их метод. Во-первых, количество параметров pLSA линейно зависит от размера обучающего корпуса. Во-вторых, неясно, как оценивать вероятности документов не из обучающего корпуса. В отличие от pLSA, LDA делает предположения о случайном распределении векторов тем и векторов документов. И для тем, и для документов предполагается, что их вектора порождаются распределением из параметрического семейства распределений Дирихле.

3.4 Word embeddings

В 2013 году Т. Миколов и др. [23] представили модель skipgram (также часто упоминается как word2vec модель). Эта модель, обученная на корпусе текстов, отображает слова в векторное пространство небольшой размерности таким образом, что расстояние между ними тем меньше, чем ближе значения этих слов. Такой эффект достигается с помощью искусственной нейронной сети, натренированной предсказывать по вектору слова его контекст; таким образом слова, появляющиеся в схожих контекстах, отображаются в близкие вектора.

WVAvgPool. С помощью таких векторов можно получить и векторное представление документов: например Xing и др. [24] предлагают строить вектора документов простым усреднением векторов всех слов в этом документе. В экспериментах на задаче классификации текстов данный подход значительно превзошел LDA.

В 2015, через два года после публикации статьи о векторном представлении слов, Le и Mikolov [25] описали два метода векторного представления документов под общим названием Paragraph Vectors. Они используют схожую с word2vec нейросетевую модель, пытаясь по вектору, относящемуся к документу, предсказать встречающиеся в нем слова.

В первом методе, названном Distributed Memory (**PV-DM**), нейросеть по вектору документа и некоторой последовательности векторов слов тренируется предсказывать вектор следующего слова в документе.

Во втором методе, Distributed Bag of Words (**PV-DBOW**), нейросеть обучается предсказывать все слова в документе по его вектору.

Таким образом, основное отличие PV-DM от PV-DBOW состоит в том, что PV-DM учитывает информацию о порядке слов в документе.

3.5 Кластеризация признаков

WordClustering. Slonim и Tishby предлагают использовать в качестве признаков не сами nграммы, а их кластеры [26]. В таком подходе проводится два шага кластеризации: сначала кластеризуются все n-граммы в словаре, затем количество вхождений n-грамм каждого кластера используется в качестве признаков для кластеризации документов. При этом n-граммы представляются в виде столбцов TF-IDF матрицы и могут быть кластеризованы любым методом

кластеризации векторов. В оригинальной статье для кластеризации был использован Information Bottleneck Algorithm.

WVClustering. Вместо представления слов в виде столбцов TF-IDF матрицы можно также использовать вектора, полученные с помощью word2vec. Такой подход используют в своей статье Qimin и др. [27].

3.6 Семантическая кластеризация

Некоторые работы выделяют в отдельную группу методы семантической кластеризации, которые используют семантические отношения между словами для представления документов.

К таким методам относятся, в том числе, алгоритмы, основанные на онтологиях. Например, Hotho и др. [28] используют онтологии, чтобы находить в тексте синонимы и воспринимать их в качестве одного элемента, тем самым сокращая размерность пространства.

Choudhary и Bhattacharyya [29] представляют каждый текст в виде графа, чьи вершины соответствуют словам текста, а ребра — семантическим отношениям между этими словами.

Также к семантической кластеризации относят методы, использующие в качестве признаков документа лексические цепочки — наборы связанных по смыслу слов в тексте [30].

3.7 Методы кластеризации векторов

Можно выделить несколько основных групп методов кластеризации векторов

Методы разделительной кластеризации итеративно переприсваивают объектам метки кластеров пока не будет найдено оптимальное разделение на кластеры в соответствии с определенной функцией близости между объектами. Как правило, количество кластеров в таких методах определяется как параметр заранее и обозначается как k.

В кластеризации документов широко используется метод **k-means**, который изначально случайным образом выбирает центр масс для каждого из k кластеров и присваивает каждому документу метку того кластера, расстояние до центра масс которого от него меньше. А затем, на каждой итерации, алгоритм вычисляет центры масс кластеров и переприсваивает их метки документам до сходимости, то есть неизменности меток всех документов.

В отличие от предыдущего метода, **k-medoids** выбирает в качестве центра масс медианный объект из кластера, таким образом, решая проблему устойчивости к выбросам.

Иерархическая кластеризация подразумевает построение дендрограммы — дерева кластеров, в котором корнем является кластер состоящий из всего набора данных, а дети каждой вершины этого дерева соответствуют разделению этого кластера на подкластеры. Дендрограмма может строиться двумя способами: снизу вверх или сверху вниз.

В первом случае, изначально каждый объект выделен в отдельный кластер, наиболее близкие из которых затем объединяются в один. Такой подход называется **агломеративной кластеризацией**. В обратном подходе — **дивизивной кластеризации** — сначала все объекты объединены в один кластер, который затем рекурсивно разделяется на подкластеры.

Методы кластеризации основанные на плотности (density-based) определяют как кластеры плотно расположенные группы объектов. Один из широко используемых методов этой группы — **DBSCAN** — работает следующим образом. Начиная выполнение на случайном объекте выборки, он определяет, есть ли в окрестности радиуса ϵ этого объекта количество объектов, не меньшее заранее заданного параметра $minSamples$, и, если есть, определяет эту окрестность как кластер; далее все объекты, лежащие в ϵ-окрестности кластера, присваиваются этому кластеру. Это повторяется до тех пор, пока есть непосещенные объекты. Если в итоге объект оказывается не принадлежащим никакому кластеру, он помечается как шум. ϵ также задаётся как внешний параметр метода.

3.8 Меры оценки эффективности

Для оценки эффективности кластеризации традиционно выделяют два типа мер: внешние меры, использующие дополнительную (внешнюю) информацию о настоящем распределении объектов по классам, и внутренние меры, использующие только информацию о самой кластеризации.

Следуя обзору Amigo и др. [31], можно выделить следующие основные группы **внешних мер эффективности**.

Меры, основанные на сопоставлении множеств: Purity, Inverse Purity [32], F-measure. Эти меры основаны на метриках точности и полноты, стандартных для оценки эффективности информационного поиска.

Меры, основанные на подсчете пар: Jaccard Coefficient, Folkes-Mallows Index, Rand Index (RI) [33], Adjusted Rand Index (ARI) [34]. Меры из данной группы основаны на подсчете пар объектов, в зависимости от их попадания в один и тот же класс/кластер или в разные.

Меры, основанные на энтропии: собственно Entropy, а также Class Entropy [35], Variation of Information [33], Mutual Information (MI) [20], Adjusted Mutual Information(AMI) [36], Normalized Mutual Information(NMI) [37], Vmeasure [38]. Меры из данной группы основаны на подсчете пар объектов, в зависимости от их попадания в один и тот же класс/кластер или в разные.

Меры, сочетающие свойства предыдущих групп мер: BCubed Precision [39], BCubed Recall, BCubed F-measure. Эти меры усредняют стандартные метрики точности/полноты/F-меры для каждого объекта; как показали Амиго и др. [31], BCubed F-measure удовлетворяет всем предложенным в этой работе аксиомам, в отличие от остальных мер.

К настоящему времени предложено более 30 **внутренних мер эффективности** [40] и проведено множество их сравнений [40, 41, 42].

В экспериментальном сравнении 30 мер Arbelaitz и др. [40] показывают, что меры Silhouette [43], Davies–Bouldin [44], Calinski–Harabasz [45], обобщенные индексы Dunn [46], индекс COP [47] и SDbw [48] показывают лучшие результаты, чем остальные меры, при этом превосходство меры Silhouette статистически значимо (тест Шаффера с уровнем значимости 10%).

4. Методика экспериментальных исследований

Общая схема работы исследуемых методов состоит из 3 этапов, которые подробно описаны в подразделах 4.1–4.3. Подраздел 4.4 посвящен используемым наборам данных; в последнем подразделе аргументируется выбор мер эффективности.

4.1 Предобработка

Для предварительной обработки входного текста применялась следующая последовательность действий:

1. токенизация: использовалась библиотека NLTK (Natural Language Toolkit)[2] [49];
2. удаление знаков препинания;
3. перевод слов в нижний регистр;
4. удаление стоп-слов: использовались списки стоп-слов из NLTK и Scikit-learn[3] [50];
5. стемминг: использовался стемминг Snowball (Porter2)[4] из библиотеки NLTK.

4.2 Векторизация

Основное отличие исследуемых методов заключалось в способе векторизации текста. Были исследованы следующие методы: BinaryBOW, CountBOW, TermBOW, TF-IDF, BM25, NMF, LDA, WVAvgPool, PV-DM, PV-DBOW, WordClustering, WVClustering.

Метод TermBOW представляет собой модификацию методов CountBOW и TF-IDF, в которых вместо слов рассматриваются термины, найденные с помощью методов CValue, Weirdness, LinkProbability, NovelTopicModel, DomainModel, KeyConceptRelatedness, Voting, PU (см. обзор методов извлечения терминологии [51]). Использовалась реализация библиотеки ATR4S[5] [52].

[2] http://www.nltk.org/
[3] http://scikit-learn.org/
[4] http://snowball.tartarus.org/algorithms/english/stemmer.html
[5] https://github.com/ispras/atr4s

Для реализации методов BinaryBOW, CountBOW, TF-IDF, LDA, NMF, WordClustering, WVClustering, TermBOW использовалась библиотека Scikit-learn.

Для реализации методов WVAvgPool, WVClustering, PV-DM, PV-DBOW использовалась библиотека Gensim[6] [53], предоставляющая методы для тематического моделирования и получения векторных представлений слов (word2vec, doc2vec).

Для WVAvgPool использовалась модель Word2Vec, обученная на текстах английской Википедии (на февраль 2015 года)[7].

Поскольку алгоритм k-means, выбранный в качестве основного метода кластеризации (см. 4.3), основан на Евклидовом расстоянии, которое учитывает длину векторов, и поскольку для кластеризации документов по тематикам их длина не важна [8], все векторы были нормализованы в L2-норме.

4.3 Кластеризация

Были исследованы следующие методы кластеризации: k-means, агломеративная кластеризация, спектральная кластеризация (использовалась реализация библиотеки Scikit-learn). После предварительных экспериментов было решено не проводить исследование алгоритма DBSCAN [54], так как он продемонстрировал слишком высокую чувствительность к выбору параметров (min_samples и eps); кроме того, DBSCAN достаточно много объектов не относит ни к одному из кластеров, помечая их как шум, что затрудняет его сравнение с другими алгоритмами кластеризации.

Количество кластеров k задавалось в качестве параметра для алгоритмов k-means и агломеративной кластеризации. Для каждого исследуемого набора данных фиксировалось множество возможных значений параметра k, которые перебирались в процессе поиска лучшего набора параметров. Остальные параметры алгоритмов кластеризации использовались по умолчанию[8].

В частности, для инициализации центров кластеров в k-means применялся алгоритм k-means++ [55]. Для получения устойчивых результатов совершалось 10 запусков k-means; из 10 полученных кластеризаций выбиралась лучшая (минимизирующая суммарное расстояние всех кластеризуемых объектов до ближайших центров кластеров).

4.4 Наборы данных

Экспериментальное исследование проводилось на наборах данных 20 Newsgroups (20 NG)[9] [56], Krapivin (KR) [57], аннотации из Krapivin (Krapivin-abstracts, KRabs), TREC GEN 2007 (TG2007) [58].

[6] https://radimrehurek.com/gensim/
[7] https://github.com/idio/wiki2vec
[8] Версия Scikit-learn: 0.18.1
[9] http://qwone.com/~jason/20Newsgroups/

Набор данных 20 Newsgroups представляет собой 18846 новостных статей, каждая из которых посвящена одной из 20 тем. Включение этого набора данных в исследование было продиктовано его частым использованием в работах, посвященных кластеризации текстов (например, в [8, 26, 59]).

Набор данных Krapivin состоит из научных статей, посвященных компьютерным наукам (Computer Science). В качестве ожидаемой кластеризации статей были использованы темы верхнего уровня классификации ACM CCS (Computing Classification System)[10], которые проставляются статьям вручную экспертами. Из набора данных Krapivin были взяты статьи, имеющие ровно одну тему верхнего уровня. Таких статей оказалось 1478, количество различных тем: 12.

Помимо основного набора данных Krapivin, также использовался набор данных, состоящий только из аннотаций научных статей. Это было сделано для того, чтобы оценить эффективность работы методов в случае, когда доступны только небольшие части текстов.

Набор данных TREC GEN 2007 состоит из статей, посвященных геномике. Набор данных был составлен для проведения конкурса по извлечению сущностей TREC 2007 Genomics Track[11]. Организаторы использовали корпус научных статей Highwire Press[12], состоящий из 160 000 статей, взятых из 49 журналов, посвященных геномике.

Из этого корпуса было выделено несколько тысяч статей, которые были предоставлены экспертам в данной предметной области для разметки. Разметка заключалась в нахождении в статьях ответов на 36 вопросов, фиксированных организаторами. Положительный ответ на каждый вопрос обозначает наличие в тексте статьи описания некоторой темы. из этого множества статей были удалены те, для которых были даны ответы на два и более вопроса. Оставшиеся статьи относились в отдельный класс в том случае, если экспертами были даны ответы только на один вопрос. Итоговый набор составляет 2325 статей.

4.5 Меры эффективности

В ходе экспериментального исследования были использованы следующие внешние меры эффективности: Adjusted Mutual Information(AMI), Normalized Mutual Information(NMI), Adjusted Rand Index (ARI), V-measure. Мы использовали несколько мер эффективности, поскольку (1) для данной задачи не существует единой общепринятой меры (см. подраздел 3.8) и (2) это позволяет произвести сравнение с исследованиями, описанными в других работах.

При этом в качестве основной меры эффективности была выбрана AMI по следующим причинам. Во-первых, меры, основанные на Mutual Information и

[10] http://dl.acm.org/ccs/ccs.cfm
[11] http://trec.nist.gov/data/t2007_genomics.html
[12] http://home.highwire.org/

Rand Index являются наиболее популярными. Во-вторых, в мерах AMI и ARI вводится поправка на случайность (adjusted for chance [60]: при сравнении случайных кластеризаций эти меры имеют близкое к нулю значение, в то время как значения NMI могут быть сильно больше 0 при большом количестве кластеров). В-третьих, классы в научных статьях обычно несбалансированны, то есть данные представляют собой набор как больших, так и малых классов, а Романо и др. [60] показали, что в таких случаях AMI является предпочтительной мерой по сравнению с ARI.

Также использовались следующие внутренние меры эффективности: Silhouette Coefficient (Silhouette, SC); Calinski-Harabaz Index (CHI). Помимо их популярности, выбор этих мер обусловлен тем, что они хорошо подходят для оптимизации параметров алгоритма k-means, так как основаны на похожих предположениях [61].

Стоит отметить, что меньшему значению меры Calinski-Harabaz Index соответствует большее значение эффективности. Для всех остальных рассматриваемых мер верно обратное: большее значение меры соответствует большей эффективности.

5. Результаты экспериментальных исследований

В данном разделе описаны результаты экспериментальных исследований методов кластеризации текстов.

5.1. Сравнение методов

В таблице 1 представлены максимальные значения меры эффективности AMI для разных наборов данных. Эти данные были получены путем перебора различных параметров методов векторизации и параметра k (количество кластеров) алгоритма k-means, и выбора максимально значения для каждого метода и для каждого набора данных. Представленные значения можно считать потенциальными максимумами для исследуемых методов.

Табл, 1. Максимальное значение AMI (k-means)
Table 1. Maximum value of AMI (k-means)

	20NG	KR	KRabs	TG2007
BinaryBOW	0.2586	0.2402	0.2041	0.3581
CountBOW	0.2957	0.2453	0.1598	0.4018
TermBOW	0.3123	0.2659	0.1266	0.5038
TF-IDF	0.4911	0.2826	0.2705	0.5051
BM25	0.4261	**0.3069**	**0.2824**	0.5291
NMF	0.4438	0.2642	0.262	0.4882
LDA	0.3391	0.2831	0.2237	0.4155

WVAvgPool	0.141	0.174	0.1608	0.2847
PV-DM	0.5901	0.3014	0.2483	**0.56**
PV-DBOW	**0.6735**	0.2773	0.251	0.5026
WordClustering	0.2188	0.2296	0.2059	0.4193
WVClustering	0.1159	0.0875	0.0613	0.1636

В соответствии с таблицей 1, наибольшие значения AMI имеют PV-DBOW, PV-DM, BM25, TF-IDF и NMF.

Можно видеть, что значение функции эффективности AMI на различных наборах при фиксированных методах сильно отличается. К примеру, для всех исследуемых методов значение AMI на наборе данных 20 Newsgroups выше, чем на Krapivin и Krapivin-abstracts. Возможная причина — большинство тем в 20 Newsgroups семантически далеки друг от друга (политика, спорт, автомобили), в то время как все статьи набора данных krapivin и krapivin-abstracts посвящены компьютерным наукам. Также все методы на наборе данных Krapivin-abstracts работают хуже, чем на Krapivin, однако разница не столь велика, что может быть объяснено тем, в аннотациях пытаются кратко изложить суть статьи и используют для этого специфичные термины.

5.2 Визуализация матрицы ошибок

Для визуализации результатов кластеризации принято использовать матрицу ошибок (confusion matrix) — матрицу, каждая строка которой соответствует распределению объектов классов, присутствующих в разметке набора данных, по кластерам, полученным с помощью используемых методов, а каждый столбец — распределению объектов кластера по классам.

Чем интенсивнее цвет прямоугольника на пересечении, тем большее количество объектов класса, соответствующего строке, были отнесены методом в кластер, соответствующий столбцу.

Визуализация матрицы ошибок в случае, когда метод работает идеально (все объекты одного класса и только они относятся в один кластер), представляет собой квадратную матрицу, у которой в каждой строке и в каждом столбце закрашен только один квадрат.

На рисунке 1 изображена матрица ошибок для метода PV-DBOW на наборе данных 20 Newsgroups. Можно видеть, что метод довольно точно определяет классы: практически для каждой строки и для каждого столбца существует только один темный квадрат. В соответствии с матрицей ошибок, большие части 3 и 4 классов (нумерация с 0) были отнесены в один кластер. Это можно объяснить их темами: 3 класс посвящен устройствам IBM (’comp.sys.ibm.pc.hardware’), а 4 класс — устройствам Mac (’comp.sys.mac.hardware’). Похожая ситуация наблюдается с классами 16 и 20 (’soc.religion.christian’ и ’talk.religion.misc’), а также 17 и 19 (’talk.politics.guns’ и ’talk.politics.misc’).

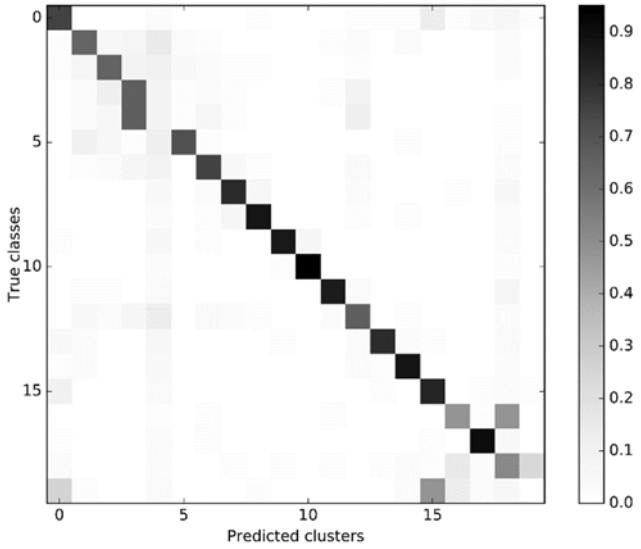

Рис. 1. Матрица ошибок PV-DBOW (k-means) на 20 Newsgroups
Fig. 1. Confusion matrix for PV-DBOW (k-means) on 20 Newsgroups

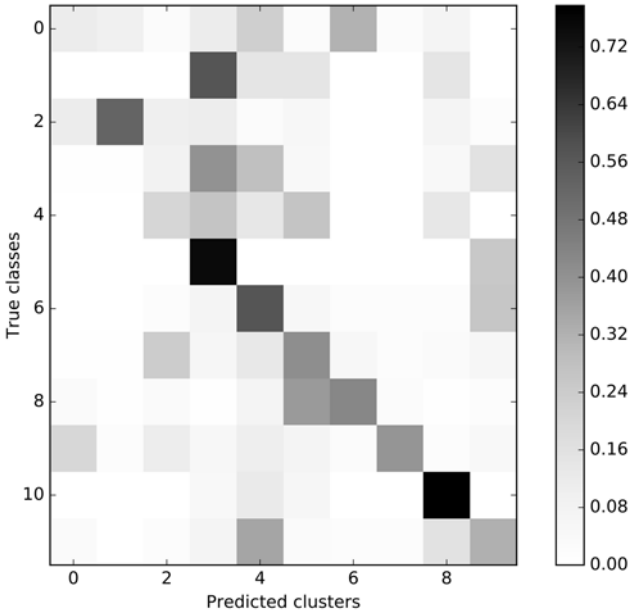

Рис. 2. Матрица ошибок BM25 (k-means) на Krapivin
Fig. 2. Confusion matrix for BM25 (k-means) on Krapivin

На рисунке 2 изображена матрица ошибок для метода PV-DM на наборе данных Krapivin. Несмотря на то, что для этой матрицы отсутствует явная структура, для нее можно выделить некоторые закономерности. Например, можно наблюдать соответствие между классами и кластерами: практически для каждой строки можно выделить столбец, в пересечении с которым содержится темный квадрат. Кластер 3 содержит значительную часть объектов класса 1 (тема верхнего уровня классификации ACM CCS: Social and professional topics) и класса 5 (General and reference); кластеру 5 соответствуют классы 7 (Theory of computation) и 8 (Mathematics of computing).

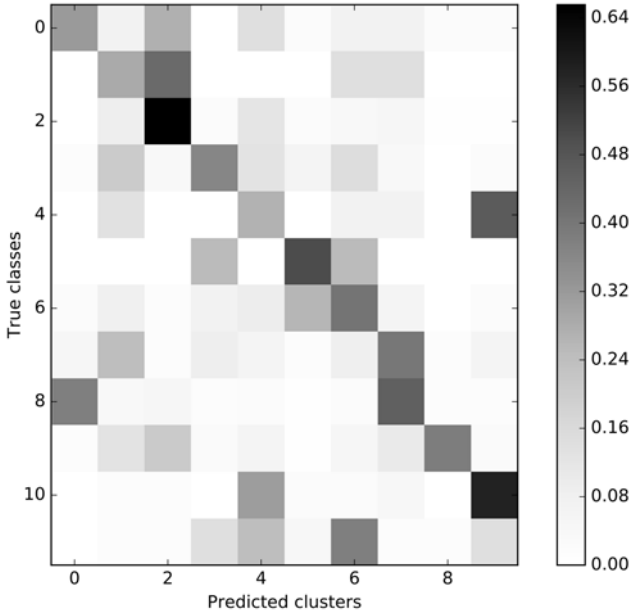

Рис. 3. Матрица ошибок BM25 (k-means) на Krapivin-abstracts
Fig. 3. Fig. 2. Confusion matrix for BM25 (k-means) on Krapivin-abstracts

На рисунке 3 изображена матрица ошибок для метода BM25 на наборе данных Krapivin-abstracts. Аналогично набору данных Krapivin, у матрицы ошибок тяжело выделить ясную структуру.

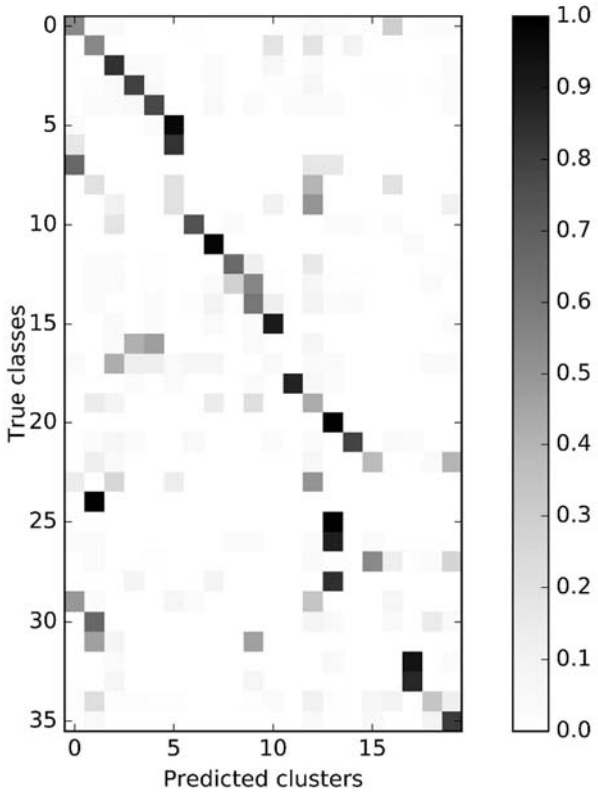

Рис. 4. Матрица ошибок PV-DM (k-means) на TREC GEN 2007
Fig. 4. Confusion matrix for PV-DB (k-means) on TREC GEN 2007

На рисунке 4 изображена матрица ошибок для метода PV-DM на наборе данных TREC GEN 2007. Количество кластеров (20) намного меньше количества классов (36). В связи с этим, в некоторых кластерах содержится большая часть объектов сразу нескольких классов. Значительная часть объектов практически каждого класса была отнесена ровно в 1 кластер. Такое поведение метода может быть объяснено тем, что количество классов слишком большое и некоторые из них содержат тексты, описывающие схожие темы.

5.3 Внутренние меры эффективности

Во многих реальных задачах для исследуемых наборов данных не существует информации о распределении документов по классам, из-за этого возникают трудности с выбором модели и параметров с наибольшей эффективностью.

Одним из подходов для решения этих проблем является оптимизация внутренних мер эффективности (в частности, Silhouette Coefficient и Calinski-Harabaz Index).

Табл. 2. Значение AMI (k-means) при подборе параметров с помощью оптимизации Silhouette
Table 2. AMI values (k-means) when selecting parameters using Silhouette optimization

	20NG	KR	KRabs	TG2007
BinaryBOW	0.0378	0.0343	0.0108	0.0946
CountBOW	0.2114	0.1589	-0.0012	0.0863
TermBOW	0.0692	0.0203	0.0161	0.2878
TF-IDF	0.0451	0.2404	0.0096	0.1565
BM25	0.0765	0.1769	0.1977	0.1241
NMF	0.0217	0.1301	0.006	0.1941
LDA	0.1371	0.2021	0.1561	0.2616
WVAvgPool	0.0706	0.1194	0.1074	0.1657
PV-DM	0.4757	0.2351	0.1774	**0.4716**
PV-DBOW	**0.6551**	**0.2515**	**0.2437**	0.4467
WordClustering	0.0421	0.1333	0.0394	0.1756
WVClustering	0.026	0.0181	0.0415	0.0152

Табл. 3. Значение AMI (k-means) при подборе параметров с помощью оптимизации Calinski-Harabaz Index
Table 3. AMI values (k-means) when selecting parameters using Calinski-Harabaz Index optimization

	20NG	KR	KRabs	TG2007
BinaryBOW	0.2581	0.1687	0.0023	0.3327
CountBOW	0.0355	0.0103	0.0145	0.189
TermBOW	0.2961	0.0723	0.0151	0.02
TF-IDF	0.0264	0.0789	0.0015	0.2267
BM25	0.3197	0.2155	0.1148	0.3886
NMF	0.4365	0.2285	**0.211**	0.39
LDA	0.2195	0.1765	0.1545	0.2651
WVAvgPool	0.141	0.1573	0.1608	0.2671
PV-DM	0.449	**0.2315**	0.1804	**0.4712**
PV-DBOW	**0.5962**	0.2044	0.1987	0.427
WordClustering	0.1991	0.1985	0.1899	0.3971
WVClustering	0.0363	0.0161	0.0496	0.0352

В таблице 2 содержатся значения меры эффективности AMI при подборе параметров с помощью оптимизация Silhouette Coefficient; в таблице 3 — аналогичные значения для Calinski-Harabaz Index.

Исходя из приведенных таблиц можно сделать вывод, что разные методы хорошо оптимизируются разными внутренними мерами эффективности. К примеру, для поиска параметров PVDBOW лучше подходит Silhouette Coefficient, а для BM25 и NMF — Calinski-Harabaz Index.

Для оценки связи внутренних метрик эффективности и внешних была посчитана ранговая корреляция Кендалла[13] [63].

В приложении В находятся таблицы с результатами вычислений ранговой корреляции (21 и 24), таблицы с оптимальными значениями внутренних мер эффективности (таблицы 19 и 22) и таблицы с отношением полученного значения AMI, оптимизированного с помощью внутренних мер эффективности, к максимально возможному значению AMI (таблицы 20 и 23).

Для большинства методов корреляция близка к нулю либо значительно отличается в зависимости от набора данных, однако в некоторых случаях использование внутренних мер эффективности позволяет подобрать параметры метода, при которых значение меры эффективности AMI близко к оптимальному.

5.4 Внешние меры эффективности

Как было отмечено выше, в разделе 4.5, для задачи кластеризации не существует общепринятой внешней меры эффективности: наравне с AMI часто используются Normalized Mutual Information(NMI), Adjusted Rand Index (ARI), Vmeasure и другие.

Для оценки того, существенно ли влияет выбор внешней меры эффективности на определение лучшего метода, также были вычислены значения мер NMI, ARI, V-measure и корреляция Кендалла между ними и AMI, см. приложение А.

На основе проведенных экспериментов можно сделать вывод, что корреляция между AMI и остальными мерами эффективности достаточно высока.

Также можно отметить, что на трех из четырех наборов данных (кроме набора данных Krapivin) все внешние меры эффективности имеют максимальное значение при использовании одних и тех же методов. На наборе данных Krapivin NMI, ARI и V-measure принимают оптимальное значение при использовании метода PV-DM, а NMI — при BM25 (впрочем, разность значений мер для этих методов не превосходит 0.02).

[13] В данной задаче ранговая корреляция предпочтительнее обычной, так как больший интерес представляет относительный порядок методов для двух исследуемых мер; при этом корреляция Кендалла предоставляет более надежную оценку, чем ранговая корреляция Спирмена, особенно в случае небольших размеров выборки [62].

Таким образом можно считать, что выбор внешней меры эффективности не оказывает значительного влияния на определение наиболее эффективного метода.

5.5 Другие методы кластеризации

Помимо k-means также были исследованы агломеративная кластеризация и спектральная кластеризация. В связи с ограниченностью вычислительных ресурсов было принято решение исследовать другие способы кластеризации только для методов, имеющих высокое значение меры AMI для k-means, а именно: PV-DBOW, PV-DM, BM25, TF-IDF и NMF.

Максимальные возможные значения AMI при использовании агломеративной кластеризации содержатся в таблице 4. Аналогичные результаты для спектральной кластеризации содержатся в таблице 5.

Табл. 4. Максимальное значение AMI (агломеративная кластеризация)
Table 4. Maximum values of AMI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.4281	0.2642	0.2264	0.4481
BM25	0.4999	0.2918	**0.2495**	0.5092
NMF	0.2837	0.2334	0.1918	0.4361
PV-DM	0.5089	**0.3024**	0.2204	**0.5112**
PV-DBOW	**0.5883**	0.2928	0.2173	0.5029

Табл. 5. Максимальное значение AMI (спектральная кластеризация)
Table 5. Maximum values of AMI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.4086	0.2455	0.2355	0.4348
BM25	0.4812	0.253	**0.2832**	0.4914
NMF	0.3905	0.2519	0.2323	0.4401
PV-DM	0.5092	**0.2781**	0.2281	**0.5209**
PV-DBOW	**0.6072**	0.2519	0.2316	0.4964

В таблицах 6 и 7 содержатся значения меры эффективности AMI при применении агломеративной кластеризации и при оптимизации мер эффективности Silhouette Coefficient и Calinski-Harabaz Index, соответственно. Аналогичные данные для спектральной кластеризации см. в таблицах 8 и 9. Оптимальные значения внутренних мер эффективности, их коэффициенты корреляции с AMI и отношения значений AMI, полученных с помощью

оптимизации внешних мер, к максимальному значению AMI представлены в таблицах в приложении В.

Табл. 6. Значение AMI (агломеративная кластеризация) при подборе параметров с помощью оптимизации Silhouette Coefficient
Table 6. AMI values (agglomerate clustering) when selecting parameters using Silhouette Coefficient optimization

	20NG	KR	KRabs	TG2007
TF-IDF	0.0385	0.0158	0.0155	0.0911
BM25	0.1516	0.1949	**0.2057**	0.1755
NMF	0.0517	0.0598	0.007	0.2559
PV-DM	0.4126	**0.2751**	0.1481	**0.4638**
PV-DBOW	**0.4836**	0.2353	0.1413	0.4626

Табл. 7. Значение AMI (агломеративная кластеризация) при подборе параметров с помощью оптимизации Calinski-Harabaz Index
Table 7. AMI values (agglomerate clustering) when selecting parameters using Calinski-Harabaz Index optimization

	20NG	KR	KRabs	TG2007
TF-IDF	0.0482	0.0111	0.0129	0.227
BM25	0.4141	**0.266**	**0.1973**	0.3804
NMF	-0.0	0.0003	-0.0005	0.0007
PV-DM	0.3608	0.2076	0.1281	0.4406
PV-DBOW	**0.5256**	0.2226	0.1794	**0.4431**

Табл.8. Значение AMI (спектральная кластеризация) при подборе параметров с помощью оптимизации Silhouette Coefficient
Table 8. AMI values (spectral clustering) when selecting parameters using Silhouette Coefficient optimization

	20NG	KR	KRabs	TG2007
TF-IDF	0.0347	0.05	0.0135	0.0936
BM25	0.0787	0.1704	0.1155	0.2172
NMF	0.0306	0.0814	0.0092	0.2632
PV-DM	0.4176	**0.2411**	**0.1785**	**0.4528**
PV-DBOW	**0.5945**	0.2409	0.1726	0.446

Табл. 9. Значение AMI (спектральная кластеризация) при подборе параметров с помощью оптимизации Calinski-Harabaz Index
Table 9. AMI values (spectral clustering) when selecting parameters using Calinski-Harabaz Index optimization

	20NG	KR	KRabs	TG2007
TF-IDF	0.0247	0.1218	0.0058	0.2122
BM25	0.4383	**0.2303**	**0.2266**	0.388
NMF	-0.0001	0.0017	-0.0014	0.0006
PV-DM	0.2991	0.2225	0.141	**0.4549**
PV-DBOW	**0.5372**	0.2298	0.2041	0.4205

Для каждого набора данных максимальные значения AMI при использовании k-means (табл. 1) выше, чем при использовании и агломеративной кластеризации (табл. 4) и спектральной (табл. 5).

Сравнение значений AMI, полученных с помощью оптимизации внутренних мер эффективности, показывает, что набору данных Krapivin соответствует большее значение AMI при использовании агломеративной кластеризации; для остальных наборов данных предпочтительнее использовать k-means.

Отметим, что агломеративная кластеризация может быть полезной в случае, когда требуется изменять число кластеров и не пересчитывать при этом всю кластеризацию, поскольку агломеративная кластеризация строит дендрограмму для всех объектов и позволяет производить разбиение по разным порогам.

Сравнение значений AMI, вычисленных при оптимизации внутренних мер эффективности, демонстрирует преимущество k-means перед спектральной кластеризацией.

Аналогичные выводы можно сделать при сравнении агломеративной кластеризации и спектральной.

5.6 Время работы

Время работы методов, в случае применения кластеризации k-means, на исследуемых наборах данных описано в таблице 10. Оно было получено путем путем усреднения трех запусков; значения параметров методов соответствовали значениям, при которых максимизируется функция эффективности AMI.

В таблице 11 содержится аналогичная информация для агломеративной кластеризации, в таблице 12 — для спектральной.

Конфигурация вычислительного устройства: Intel Xeon E312xx (8 CPU), 2GHz, 64GB RAM.

Табл. 10. Время работы методов (в секундах) при применении k-means
Table 10. Method running time (in seconds) when using k-means

	20NG	KR	KRabs	TG2007
BinaryBOW	334	130	9	251
CountBOW	342	169	14	273
TermBOW	130	171	13	327
TF-IDF	340	132	11	501
BM25	264	225	5	265
NMF	4443	820	561	1384
LDA	225	169	13	303
WVAvgPool	202	189	8	390
PV-DM	291	1050	24	2038
PV-DBOW	468	668	32	1666
WordClustering	882	808	69	775
WVClustering	566	355	27	916

Табл. 11. Время работы методов (в секундах) при применении агломеративной кластеризации
Table 11. Method running time (in seconds) when using agglomerate clustering

	20NG	KR	KRabs	TG2007
TF-IDF	1094	130	9	261
BM25	1296	266	9	260
NMF	5212	874	594	1373
PV-DM	357	482	22	1134
PV-DBOW	645	687	19	1527

Таблица 12. Время работы методов (в секундах) при применении спектральной кластеризации
Table 12. Method running time (in seconds) when using spectral clustering

	20NG	KR	KRabs	TG2007
TF-IDF	163	121	4	239
BM25	177	123	4	307
NMF	2821	840	535	1291
PV-DM	419	904	20	1953
PV-DBOW	521	382	27	848

6. Заключение

В данной работе рассмотрены и экспериментально исследованы методы кластеризации текстовых документов, в том числе, научных статей. Каждый метод состоял из трех последовательных этапов: предварительная обработка текста (см. раздел 4.1); векторизация предобработанного текста (см. раздел 4.2); кластеризация векторов (см. раздел 4.3).

Экспериментальное исследование показало, что лучшим методом (при условии оптимизации параметров с помощью внутренней меры эффективности) является k-means с векторизацией Paragraph Vectors для всех наборов данных; кроме Krapivin, для которого лучше оказалась агломеративная кластеризация. Стоит отметить, что эффективность разных модификаций метода Paragraph Vectors (DBOW и DM) сильно зависит от набора данных: для новостных текстов и аннотаций научных статей DBOW значительно превосходит DM, в то время как на двух остальных наборах данных DM несколько лучше, чем DBOW. При этом кластеризация аннотаций научных статей оказалась менее эффективной, чем кластеризация полных статей, которая, в свою очередь, показала значительно меньшую эффективность по сравнению с кластеризацией научных статей.

Также в данной работе были исследованы меры эффективности кластеризации: как внешние (Adjusted Mutual Information, Normalized Mutual Information, Adjusted Rand Index, V-measure), так и внутренние (Silhouette Coefficient, Calinski-Harabaz Index). В частности, была посчитана корреляция между внешними мерами эффективности; полученные значения позволяют сделать вывод об относительно высокой взаимозаменяемости этих мер. Внутренние меры показали достаточно высокую эффективность (хотя и не самую высокую стабильность) для задачи оптимизации параметров методов: результаты методов, оптимизированных с помощью меры Silhouette параметры, составили от 82% до 97% от лучших результатов. Кроме того, выяснилось, что для оптимизации разных методов лучше подходят разные внутренние меры эффективности: так, для BM25 выше корреляция у меры Calinski-Harabaz, в то время как для Paragraph Vectors – Silhouette.

Наиболее перспективными направлениями дальнейшей работы представляется улучшение эффективности кластеризации научных статей за счет использования дополнительной информации, такой как граф цитирования и мета-данные статей (авторы, год и место издания), а также проведение экспериментальных исследований на других наборах данных.

Список литературы

[1]. Liu Xiaoyong, Croft W Bruce. Cluster-based retrieval using language models. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2004, pp. 186–193.

[2]. Sasaki Minoru, Shinnou Hiroyuki. Spam detection using text clustering. 2005 International Conference on Cyberworlds (CW'05). IEEE. 2005, pp. 316-319.

[3]. Sergio Decherchi, Simone Tacconi, Judith Redi et al. Text clustering for digital forensics analysis. Computational Intelligence in Security for Information Systems. Springer, 2009, pp. 29–36.
[4]. E Dransfield, G Morrot, J-F Martin et al. The application of a text clustering statisticalanalysis to aid the interpretation of focus group interviews.Food Quality and Preference. 2004. T. 15, № 5, pp. 477–488.
[5]. Bader Aljaber, Nicola Stokes, James Bailey et al. Document clustering of scientific texts using citation contexts. Information Retrieval. 2010. T. 13, № 2, pp. 101–131.
[6]. Marchionini Gary. Exploratory search: from finding to understanding. Communications of the ACM. 2006. T. 49, № 4, pp. 41–46.
[7]. Andrews Nicholas O, Fox Edward A. Recent developments in document clustering: Tech. Rep.: Technical report, Computer Science, Virginia Tech, 2007.
[8]. Huang Anna. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008, pp. 49–56.
[9]. Sathiyakumari K, Manimekalai G, Preamsudha V. A survey on various approaches in document clustering.
[10]. Popat Shraddha K, Emmanuel M. Review and comparative study of clustering techniques.
[11]. Anastasiu David C, Tagarelli Andrea, Karypis George. Document Clustering: The Next Frontier. 2013.
[12]. Aggarwal Charu C, Reddy Chandan K. Data clustering: algorithms and applications. CRC Press, 2013.
[13]. Aggarwal Charu C, Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012.
[14]. Saiyad Nagma Y, Prajapati Harshadkumar B, Dabhi Vipul K. A Survey of Document Clustering using Semantic Approach.
[15]. Salton Gerard, Buckley Christopher. Termweighting approaches in automatic text retrieval. Information processing & management. 1988. T. 24, № 5, pp 513–523.
[16]. Whissell John S, Clarke Charles LA. Improving document clustering using Okapi BM25 feature weighting. Information retrieval. 2011. T. 14, № 5, pp. 466–487.
[17]. Голомазов Д. Д. Методы и средства управления научной информацией с использованием онтологий. Диссертация кандидата физико-математических наук. Москва. 2012.
[18]. Pinto David, Jim´enez-Salazar H´ector, Rosso Paolo. Clustering abstracts of scientific texts using the transition point technique. International Conference on Intelligent Text Processing and Computational Linguistics. Springer. 2006, pp. 536–546.
[19]. Scott Deerwester, Susan T Dumais, George W Furnas et al. Indexing by latent semantic analysis. Journal of the American society for information science. 1990. T. 41, № 6, pp. 391.
[20]. Xu Wei, Liu Xin, Gong Yihong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM. 2003, pp. 267–273.
[21]. Hofmann Thomas. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 1999, pp. 50–57.
[22]. Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation. Journal of machine Learning research. 2003. T. 3, № Jan., pp. 993–1022.

[23]. Tomas Mikolov, Kai Chen, Greg Corrado et al. Efficient estimation of word representationsin vector space. arXiv preprint, arXiv:1301.3781. 2013.
[24]. Chao Xing, Dong Wang, Xuewei Zhang et al. Document classification with distributions ofword vectors. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. IEEE. 2014, pp. 1–5.
[25]. Le Quoc V, Mikolov Tomas. Distributed Representations of Sentences and Documents. ICML. T. 14. 2014, pp. 1188– 1196.
[26]. Slonim Noam, Tishby Naftali. Document clustering using word clusters via the information bottleneck method. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2000, pp. 208–215.
[27]. Cao Qimin, Guo Qiao, Wang Yongliang et al. Text clustering using VSM with feature clusters. Neural Computing and Applications. 2015. T. 26, № 4, pp. 995–1003.
[28]. Hotho Andreas, Maedche Alexander, Staab Steffen. Ontology-based text document clustering.
[29]. Choudhary Bhoopesh, Bhattacharyya Pushpak. Text clustering using semantics. Proceedings of the 11th International World Wide Web Conference. 2002, pp. 1–4.
[30]. Jayarajan Dinakar, Deodhare Dipti, Ravindran B. Lexical Chains as Document Features. Third International Joint Conference on Natural Language Processing. Citeseer. 2008, pp. 111.
[31]. Enrique Amigo´, Julio Gonzalo, Javier Artiles et al. A comparison of extrinsic clustering evaluationmetrics based on formal constraints. Information retrieval. 2009. T. 12, № 4, pp. 461–486.
[32]. Zhao Ying, Karypis George, Du Ding-Zhu. Criterion functions for document clustering: Tech. Rep.: Technical Report, 2005.
[33]. Meil˘a Marina. Comparing clusterings by the variation of information. Learning theory and kernel machines. Springer, 2003, pp. 173–187.
[34]. Hubert Lawrence, Arabie Phipps. Comparing partitions. Journal of classification. 1985. T. 2, № 1, pp. 193–218.
[35]. Bakus J, Hussin MF, Kamel M. A SOM-based document clustering using phrases. Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. IEEE. T. 5. 2002, pp. 2212–2216.
[36]. Vinh Nguyen Xuan, Epps Julien, Bailey James. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. Proceedings of the 26th Annual International Conference on Machine Learning. ACM. 2009, pp. 1073–1080.
[37]. Strehl Alexander, Ghosh Joydeep. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002. T. 3, № Dec., pp. 583–617.
[38]. Rosenberg Andrew, Hirschberg Julia. VMeasure: A Conditional Entropy-Based External Cluster Evaluation Measure. EMNLP-CoNLL. T. 7. 2007, pp. 410–420.
[39]. Bagga Amit, Baldwin Breck. Entity-based cross-document coreferencing using the vector space model. Proceedings of the 17th international conference on Computational linguistics-Volume 1. 1998, pp. 79–85.
[40]. Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza et al. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013. T. 46, № 1, pp. 243–256.
[41]. Yanchi Liu, Zhongmou Li, Hui Xiong et al. Understanding of internal clustering validation measures. 2010 IEEE International Conference on Data Mining. IEEE. 2010, pp. 911–916.

[42]. Er´endira Rend´on, Itzel Abundez, Alejandra Arizmendi et al. Internal versus external cluster validation indexes.. International Journal of computers and communications. 2011. Т. 5, № 1, pp. 27–34.
[43]. Rousseeuw Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987. Т. 20, pp. 53–65.
[44]. Davies David L, Bouldin Donald W. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence. 1979. № 2, pp. 224–227.
[45]. Calin´ski Tadeusz, Harabasz Jerzy. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods. 1974. Т. 3, № 1, pp. 1–27.
[46]. Bezdek James C, Pal Nikhil R. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 1998. Т. 28, № 3, pp. 301–315.
[47]. Ibai Gurrutxaga, In˜aki Albisua, Olatz Arbelaitz et al. SEP/COP: An efficient method to find the bestpartition in hierarchical clustering based on a new cluster validity index.Pattern Recognition. 2010. Т. 43, № 10, pp. 3364–3373.
[48]. Halkidi Maria, Vazirgiannis Michalis. Clustering validity assessment: Finding the optimal partitioning of a data set. Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE. 2001, pp. 187–194.
[49]. Bird Steven. NLTK: the natural language toolkit. Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics. 2006, pp. 69–72.
[50]. Scikit-learn: Machine Learning in Python. F. Pedregosa, G. Varoquaux, A. Gramfort [и др.]. Journal of Machine Learning Research. 2011. Т. 12, pp. 2825–2830.
[51]. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Yu. Methods for automatic term recognition in domain-specific text collections: A survey. Programming and Computer Software. 2015. Т. 41, № 6, pp. 336–349.
[52]. Astrakhantsev Nikita. ATR4S: Toolkit with State-of-the-art Automatic Terms Recognition Methods in Scala. arXiv preprint, arXiv:1611.07804. 2016.
[53]. Reh˚u˘rek R., Sojka P.˘ Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, 2010, pp. 45–50.
[54]. Martin Ester, Hans-Peter Kriegel, J¨org Sander Er´endira Rend´on, Itzel Abundez, Alejandra Arizmendi et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. Т. 96. 1996, pp. 226–231.
[55]. Arthur David, Vassilvitskii Sergei. kmeans++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.
[56]. Lang Ken. Newsweeder: Learning to filter netnews. Proceedings of the 12th international conference on machine learning. 1995, pp. 331–339.
[57]. Krapivin M., Autaeu A., Marchese M. Large dataset for keyphrases extraction. 2009. URL: http://eprints.biblio.unitn.it/1671/1/disi09055krapivin-autayeu-marchese.pdf.
[58]. William Hersh, Aaron Cohen, Lynn Ruslen et al. TREC 2007 Genomics Track Overview. 2007.
[59]. Xie Pengtao, Xing Eric P. Integrating document clustering and topic modeling. arXiv preprint, arXiv:1309.6874. 2013.
[60]. Simone Romano, Nguyen Xuan Vinh, James Bailey et al. Adjusting for Chance Clustering Comparison Measures. arXiv preprint, arXiv:1512.01286. 2015.

[61]. Van Craenendonck Toon, Blockeel Hendrik. Using internal validity measures to compare clustering algorithms. AutoML Workshop at ICML 2015, pp. 1–8.
[62]. Field Andy. Discovering statistics using IBM SPSS statistics. Sage, 2013.
[63]. Kendall Maurice G. A new measure of rank correlation. Biometrika. 1938. Т. 30, № ½, pp. 81–93.

Приложение А. Внешние меры

Табл. 13. Максимальное значение NMI
Table 13. Maximum value of NMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.2652	0.2608	0.2246	0.4288
CountBOW	0.3009	0.2833	0.1964	0.4885
TermBOW	0.327	0.2978	0.1504	0.5418
TF-IDF	0.5136	0.304	0.2963	0.5611
BM25	0.4411	0.3268	**0.3058**	0.5829
NMF	0.4631	0.2878	0.2751	0.5244
LDA	0.3531	0.3162	0.2484	0.4755
WVAvgPool	0.1509	0.1975	0.1922	0.3618
PV-DM	0.5951	**0.338**	0.2701	**0.5938**
PV-DBOW	**0.6816**	0.3099	0.2758	0.5779
WordClustering	0.2274	0.2519	0.2301	0.4742
WVClustering	0.1218	0.1145	0.0918	0.2357

Табл. 14. Максимальное значение ARI
Table 14. Maximum value of ARI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.1318	0.1687	0.1347	0.1701
CountBOW	0.1126	0.1655	0.0921	0.2358
TermBOW	0.1487	0.1868	0.0785	0.3209
TF-IDF	0.292	0.1957	0.2148	0.3043
BM25	0.1707	0.2113	**0.2295**	0.3285
NMF	0.1955	0.1673	0.2063	0.295
LDA	0.1878	0.2209	0.1738	0.2698
WVAvgPool	0.0551	0.1029	0.0762	0.143
PV-DM	0.4572	**0.2211**	0.2098	**0.3418**
PV-DBOW	**0.5677**	0.1993	0.1727	0.2829
WordClustering	0.0757	0.1464	0.1255	0.2183
WVClustering	0.0309	0.0462	0.0457	0.0797

Табл. 15. Максимальное значение V-measure
Table 15. Maximum value of V-measure

	20NG	KR	KRabs	TG2007
BinaryBOW	0.2651	0.2606	0.2244	0.4266
CountBOW	0.3009	0.2817	0.1954	0.4862
TermBOW	0.3268	0.2971	0.1494	0.5416
TF-IDF	0.5132	0.3038	0.2925	0.5593
BM25	0.441	0.3266	**0.3052**	0.5818
NMF	0.4594	0.2876	0.2751	0.5242
LDA	0.3503	0.3145	0.2472	0.4737
WVAvgPool	0.1507	0.1963	0.1917	0.3599
PV-DM	0.5951	**0.3371**	0.2698	**0.5921**
PV-DBOW	**0.6804**	0.3077	0.2739	0.5745
WordClustering	0.2272	0.2513	0.2297	0.4733
WVClustering	0.1217	0.1138	0.0842	0.2353

Табл. 16. Корреляция NMI и AMI
Table 16. Correlation between NMI and AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.8732	0.8936	0.8435	0.8322
CountBOW	0.8146	0.885	0.8795	0.8667
TermBOW	0.8504	0.9248	0.8553	0.8971
TF-IDF	0.873	0.8965	0.8802	0.8675
BM25	0.8543	0.8159	0.8361	0.8414
NMF	0.9888	0.9839	0.9831	0.9683
LDA	0.9272	0.8805	0.8925	0.7455
WVAvgPool	0.9556	0.6429	1.0	0.3778
PV-DM	0.6958	0.8092	0.807	0.7053
PV-DBOW	0.7591	0.7557	0.7281	0.8148
WordClustering	0.9171	0.8467	0.9009	0.8481
WVClustering	0.6904	0.8701	0.7316	0.8282

Табл. 17. Корреляция ARI и AMI
Table 17. Correlation between ARI and AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.8044	0.705	0.6674	0.6839

CountBOW	0.7926	0.82	0.7872	0.6469
TermBOW	0.6179	0.8523	0.6653	0.7201
TF-IDF	0.804	0.7125	0.7862	0.7533
BM25	0.6993	0.5223	0.6292	0.6537
NMF	0.9632	0.9396	0.8919	0.9307
LDA	0.8447	0.7526	0.7281	0.4306
WVAvgPool	0.9556	0.1429	0.5714	0.4667
PV-DM	0.8454	0.6851	0.5785	0.5199
PV-DBOW	0.8409	0.5855	0.6952	0.8459
WordClustering	0.704	0.6146	0.6279	0.603
WVClustering	0.6919	0.5548	0.511	0.6823

Табл. 18. Корреляция V-measure и AMI
Table 18. Correlation between V-measure and AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.9163	0.903	0.885	0.8432
CountBOW	0.8607	0.8959	0.9296	0.8707
TermBOW	0.8792	0.9322	0.8769	0.9196
TF-IDF	0.9114	0.907	0.9222	0.8864
BM25	0.8711	0.8312	0.8498	0.8695
NMF	0.9902	0.9843	0.9833	0.9714
LDA	0.9336	0.8904	0.8986	0.7898
WVAvgPool	1.0	0.7143	1.0	0.3778
PV-DM	0.7241	0.8232	0.8224	0.7255
PV-DBOW	0.7922	0.7702	0.7553	0.823
WordClustering	0.9267	0.8652	0.9079	0.8724
WVClustering	0.8267	0.867	0.8091	0.8192

Приложение B. Внутренние меры

Табл. 19. Максимальное значение Silhouette
Table 19. Maximum value of Silhouette

	20NG	KR	KRabs	TG2007
BinaryBOW	0.0558	0.0358	0.0294	0.0906
CountBOW	-0.2097	-0.0487	0.0262	0.0429
TermBOW	0.7047	0.4581	0.5768	0.3947

TF-IDF	0.1663	0.0432	0.2043	0.1179
BM25	0.0042	0.0305	0.0124	0.0609
NMF	0.1116	0.3627	0.1915	0.4467
LDA	0.3224	0.5262	0.5115	0.5448
WVAvgPool	0.0575	0.086	0.0496	0.1133
PV-DM	0.0287	0.0301	0.0309	0.0612
PV-DBOW	0.0209	0.0194	0.0241	0.0667
WordClustering	0.179	0.187	0.1601	0.1893
WVClustering	0.9943	0.9139	0.9953	0.9953

Табл. 20. Silhouette: доля от лучшего AMI
Table 20. Silhouette: share from the best AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.1462	0.1428	0.0529	0.2642
CountBOW	0.7149	0.6478	0	0.2148
TermBOW	0.2216	0.0763	0.1272	0.5713
TF-IDF	0.0918	0.8507	0.0355	0.3098
BM25	0.1795	0.5764	0.7001	0.2345
NMF	0.0489	0.4924	0.0229	0.3976
LDA	0.4043	0.7139	0.6978	0.6296
WVAvgPool	0.5007	0.6862	0.6679	0.5820
PV-DM	0.8061	0.7800	0.7145	0.8421
PV-DBOW	0.9727	0.9070	0.9709	0.8888
WordClustering	0.1924	0.5806	0.1914	0.4188
WVClustering	0.2243	0.2069	0.6770	0.0929

Табл, 21. Корреляция Silhouette и AMI
Table 21. Correlation between Silhouette and AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	-0.5274	0.1205	-0.0203	-0.3558
CountBOW	**0.5133**	0.4969	0.1228	-0.2095
TermBOW	-0.5763	-0.1145	-0.2712	-0.0788
TF-IDF	-0.048	0.0594	0.0475	-0.0838
BM25	-0.4191	-0.0713	-0.1088	-0.1785
NMF	-0.3425	0.0162	-0.4528	0.0746
LDA	-0.0394	0.3323	**0.1509**	-0.047

WVAvgPool	-0.2889	0.0	-0.2857	-0.1556
PV-DM	0.1476	0.3013	0.0118	0.191
PV-DBOW	-0.1241	**0.5092**	0.0961	**0.3275**
WordClustering	-0.2188	-0.4654	-0.3628	-0.398
WVClustering	-0.3363	-0.1768	-0.1346	-0.4956

Табл. 22. Минимальное значение CHI
Table 22. Minimal values of CHI

	20NG	KR	KRabs	TG2007
BinaryBOW	47.07	4.75	1.79	7.13
CountBOW	0.13	0.79	1.38	13.86
TermBOW	0.23	3.08	5.39	4.8
TF-IDF	0.07	0.76	0.67	4.74
BM25	23.28	4.3	1.92	5.28
NMF	43.9	10.96	6.03	16.05
LDA	733.99	173.67	154.37	162.48
WVAvgPool	47.1	32.21	39.87	27.14
PV-DM	37.66	7.49	10.76	7.67
PV-DBOW	41.41	5.63	9.93	6.52
WordClustering	122.93	24.97	9.79	26.79
WVClustering	3.16	4.08	11.29	1.0

Табл, 23. CHI: доля от лучшего AMI
Table 23. CHI: share from the best AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.9981	0.7023	0.0113	0.9291
CountBOW	0.1201	0.0420	0.0907	0.4704
TermBOW	0.9481	0.2719	0.1193	0.0397
TF-IDF	0.0538	0.2792	0.0055	0.4488
BM25	0.7503	0.7022	0.4065	0.7345
NMF	0.9836	0.8649	0.8053	0.7989
LDA	0.6473	0.6235	0.6907	0.6380
WVAvgPool	1.0000	0.9040	1.0000	0.9382
PV-DM	0.7609	0.7681	0.7265	0.8414
PV-DBOW	0.8852	0.7371	0.7916	0.8496
WordClustering	0.9100	0.8645	0.9223	0.9471
WVClustering	0.3132	0.1840	0.8091	0.2152

Табл. 24. Корреляция CHI и AMI
Table 24. Correlation between CHI and AMI

	20NG	KR	KRabs	TG2007
BinaryBOW	0.4554	0.1761	-0.2598	0.7134
CountBOW	-0.3365	-0.2167	-0.421	0.4362
TermBOW	0.5576	0.5057	0.4136	0.4147
TF-IDF	0.2611	-0.1347	-0.2672	0.394
BM25	0.3607	0.1787	0.0412	0.3957
NMF	0.674	0.6169	0.6563	0.4737
LDA	0.0706	-0.1319	-0.1211	0.1978
WVAvgPool	0.3333	0.0714	0.5	0.3778
PV-DM	-0.1389	0.182	0.0952	0.2801
PV-DBOW	0.2415	-0.1614	0.143	0.1151
WordClustering	0.3194	0.5193	0.5287	0.6116
WVClustering	0.3388	-0.0908	0.1729	0.2168

Табл. 25. Максимальное значение Silhouette (агломеративная кластеризация)
Table 25. Maximum values of Silhouette (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.1527	0.0398	0.2064	0.103
BM25	-0.0043	0.0244	0.0067	0.0832
NMF	0.1168	0.3687	0.2088	0.3005
PV-DM	0.0214	0.0219	0.0204	0.0531
PV-DBOW	0.0105	0.0163	0.0153	0.0581

Табл. 26. Silhouette: доля от лучшего AMI (агломеративная кластеризация)
Table 26. CHI: share from the best AMI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0899	0.0598	0.0685	0.2033
BM25	0.3033	0.6679	**0.8244**	0.3447
NMF	0.1822	0.2562	0.0365	0.5868
PV-DM	0.8108	**0.9097**	0.6720	0.9073
PV-DBOW	**0.8220**	0.8036	0.6503	**0.9199**

Табл. 27. Корреляция Silhouette и AMI (агломеративная кластеризация)
Table 27. Correlation between Silhouette and AMI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.012	0.0991	0.0595	-0.0959
BM25	-0.2995	0.1244	-0.1574	-0.1661
NMF	-0.3646	0.0037	-0.4792	0.0569
PV-DM	**0.0373**	0.3053	0.0781	0.0347
PV-DBOW	0.0331	**0.343**	**0.2689**	**0.3275**

Табл. 28. Минимальное значение CHI (агломеративная кластеризация)
Table 28. Minimal values of CHI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0837	0.7698	1.1986	5.6026
BM25	24.5886	5.4675	2.9236	6.4625
NMF	1.0	0.7676	1.0	1.0
PV-DM	27.4417	6.6093	7.5673	7.3344
PV-DBOW	32.5286	5.1623	7.7801	6.3808

Табл. 29. CHI: доля от лучшего AMI (агломеративная кластеризация)
Table 29. CHI: share from the best AMI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.1126	0.0420	0.0570	0.5066
BM25	0.8284	**0.9116**	0.7908	0.7471
NMF	0	0.0013	0	0.0016
PV-DM	0.7090	0.6865	0.5812	0.8619
PV-DBOW	**0.8934**	0.7602	**0.8256**	**0.8811**

Табл. 30. Корреляция CHI и AMI (агломеративная кластеризация)
Table 30. Correlation between CHI and AMI (agglomerate clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.3629	-0.0076	-0.088	**0.4156**
BM25	**0.4256**	**0.1569**	**0.2492**	0.388
NMF	-0.4744	-0.5008	-0.5042	-0.503
PV-DM	-0.1619	0.0978	-0.1447	0.2549
PV-DBOW	0.3022	-0.0272	-0.1425	0.0557

Табл. 31. Максимальное значение Silhouette (спектральная кластеризация)
Table 31. Maximum values of Silhouette (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.1712	0.0405	0.2113	0.1061
BM25	0.0067	0.0297	0.0111	0.0504
NMF	0.1651	0.3372	0.1822	0.3306
PV-DM	0.0234	0.0265	0.0308	0.0552
PV-DBOW	0.018	0.0084	0.0221	0.0643

Табл. 32. Silhouette: доля от лучшего AMI (спектральная кластеризация)
Table 32. Silhouette: share from the best AMI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0849	0.2037	0.0573	0.2153
BM25	0.1635	0.6735	0.4078	0.4420
NMF	0.0784	0.3231	0.0396	0.5980
PV-DM	0.8201	0.8670	**0.7826**	0.8693
PV-DBOW	**0.9791**	**0.9563**	0.7453	**0.8985**

Табл. 33. Корреляция Silhouette и AMI (спектральная кластеризация)
Table 33. Correlation between Silhouette and AMI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0163	0.0857	0.0577	-0.052
BM25	-0.494	0.0711	-0.4639	-0.0254
NMF	-0.4847	-0.1423	-0.5149	-0.0264
PV-DM	**0.3644**	**0.2689**	**0.2092**	0.1384
PV-DBOW	-0.0532	0.1237	-0.0496	**0.3322**

Табл. 34. Минимальное значение CHI (спектральная кластеризация)
Table 34. Minimal values of CHI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0887	1.4317	0.4466	5.9291
BM25	29.3736	5.8227	3.2461	6.6132
NMF	1.0	0.2563	1.0	1.0
PV-DM	32.3986	7.3227	10.2393	7.943
PV-DBOW	39.4728	5.6117	9.7885	6.8773

Табл. 35. CHI: доля от лучшего AMI (спектральная кластеризация)
Table 35. CHI: share from the best AMI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	0.0605	0.4961	0.0246	0.4880
BM25	**0.9108**	0.9103	0.8001	0.7896
NMF	0	0.0067	0	0.0014
PV-DM	0.5874	0.8001	0.6181	**0.8733**
PV-DBOW	0.8847	**0.9123**	**0.8813**	0.8471

Табл. 36. Корреляция CHI и AMI (спектральная кластеризация)
Table 36. Correlation between CHI and AMI (spectral clustering)

	20NG	KR	KRabs	TG2007
TF-IDF	**0.3472**	-0.0455	-0.0785	0.4048
BM25	0.3287	0.1476	**0.3762**	**0.3665**
NMF	-0.4579	-0.4662	-0.4673	-0.4908
PV-DM	-0.4174	**0.2206**	-0.1289	0.2824
PV-DBOW	0.1277	-0.0259	0.1781	0.0714

A survey and an experimental comparison of methods for text clustering: application to scientific articles

[1,2] *P.A. Parhomenko <parhomenko @ispras.ru>*
[1,3] *A.A. Grigorev <agrigorev@ispras.ru>*
[1] *N.A. Astrakhantsev <astrakhantsev@ispras.ru>*
[1] *Institute for System Programming of the RAS,*
25 Alexander Solzhenitsyn Str., Moscow, 109004, Russian Federation
[2] *Lomonosov Moscow State University,*
GSP-1, Leninskie Gory, Moscow, 119991, Russia
[3] *National Research University Higher School of Economics (HSE)*
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia

Abstract. Text documents clustering is used in many applications such as information retrieval, exploratory search, spam detection. This problem is the subject of many scientific papers, but the specificity of scientific articles in regards to the clustering efficiency remains to be studied insufficiently; in particular, if all documents belong to the same domain or if full texts of articles are unavailable. This paper presents an overview and an experimental comparison of text clustering methods in application to scientific articles. We study methods based on bag of words, terminology extraction, topic modeling, word embedding and document embedding obtained by artificial neural networks (word2vec, paragraph2vec).

Keywords: text documents clustering; bag of words; terminology extraction; topic modeling; word and document embedding; artificial neural networks

DOI: 10.15514/ISPRAS-2017-29(2)-6

For citation: Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles. Trudy ISP RAN/Proc. ISP RAS, 2017, vol. 29, issue 2, pp. 161-200 (in Russian). DOI: 10.15514/ISPRAS-2017-29(2)-6

References

[1]. Liu Xiaoyong, Croft W Bruce. Cluster-based retrieval using language models. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2004, pp. 186–193.
[2]. Sasaki Minoru, Shinnou Hiroyuki. Spam detection using text clustering. 2005 International Conference on Cyberworlds (CW'05). IEEE. 2005, pp. 316-319.
[3]. Text clustering for digital forensics analysis. Sergio Decherchi, Simone Tacconi, Judith Redi [и др.]. Computational Intelligence in Security for Information Systems. Springer, 2009, pp. 29–36.
[4]. The application of a text clustering statisticalanalysis to aid the interpretation of focus group interviews. E Dransfield, G Morrot, J-F Martin [и др.]. Food Quality and Preference. 2004. Т. 15, № 5, pp. 477–488.
[5]. Document clustering of scientific texts using citation contexts. Bader Aljaber, Nicola Stokes, James Bailey [и др.]. Information Retrieval. 2010. Т. 13, № 2, pp. 101–131.
[6]. Marchionini Gary. Exploratory search: from finding to understanding. Communications of the ACM. 2006. Т. 49, № 4, pp. 41–46.
[7]. Andrews Nicholas O, Fox Edward A. Recent developments in document clustering: Tech. Rep.: Technical report, Computer Science, Virginia Tech, 2007.
[8]. Huang Anna. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008, pp. 49–56.
[9]. Sathiyakumari K, Manimekalai G, Preamsudha V. A survey on various approaches in document clustering.
[10]. Popat Shraddha K, Emmanuel M. Review and comparative study of clustering techniques.
[11]. Anastasiu David C, Tagarelli Andrea, Karypis George. Document Clustering: The Next Frontier. 2013.
[12]. Aggarwal Charu C, Reddy Chandan K. Data clustering: algorithms and applications. CRC Press, 2013.
[13]. Aggarwal Charu C, Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012.
[14]. Saiyad Nagma Y, Prajapati Harshadkumar B, Dabhi Vipul K. A Survey of Document Clustering using Semantic Approach.
[15]. Salton Gerard, Buckley Christopher. Termweighting approaches in automatic text retrieval. Information processing & management. 1988. Т. 24, № 5, pp 513–523.
[16]. Whissell John S, Clarke Charles LA. Improving document clustering using Okapi BM25 feature weighting. Information retrieval. 2011. Т. 14, № 5, pp. 466–487.
[17]. Golomazov D.D. Methods and tools for managing scientific information with ontologies. PhD Thesis. Moscow, 2012 (in Russian).

[18]. Pinto David, Jim´enez-Salazar H´ector, Rosso Paolo. Clustering abstracts of scientific texts using the transition point technique. International Conference on Intelligent Text Processing and Computational Linguistics. Springer. 2006, pp. 536–546.
[19]. Scott Deerwester, Susan T Dumais, George W Furnas et al. Indexing by latent semantic analysis. Journal of the American society for information science. 1990. Т. 41, № 6, pp. 391.
[20]. Xu Wei, Liu Xin, Gong Yihong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM. 2003, pp. 267–273.
[21]. Hofmann Thomas. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. ACM. 1999, pp. 50–57.
[22]. Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation. Journal of machine Learning research. 2003. Т. 3, № Jan., pp. 993–1022.
[23]. Tomas Mikolov, Kai Chen, Greg Corrado et al. Efficient estimation of word representationsin vector space. arXiv preprint, arXiv:1301.3781. 2013.
[24]. Chao Xing, Dong Wang, Xuewei Zhang et al. Document classification with distributions ofword vectors. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. IEEE. 2014, pp. 1–5.
[25]. Le Quoc V, Mikolov Tomas. Distributed Representations of Sentences and Documents. ICML. Т. 14. 2014, pp. 1188– 1196.
[26]. Slonim Noam, Tishby Naftali. Document clustering using word clusters via the information bottleneck method. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2000, pp. 208–215.
[27]. Cao Qimin, Guo Qiao, Wang Yongliang et al. Text clustering using VSM with feature clusters. Neural Computing and Applications. 2015. Т. 26, № 4, pp. 995–1003.
[28]. Hotho Andreas, Maedche Alexander, Staab Steffen. Ontology-based text document clustering.
[29]. Choudhary Bhoopesh, Bhattacharyya Pushpak. Text clustering using semantics. Proceedings of the 11th International World Wide Web Conference. 2002, pp. 1–4.
[30]. Jayarajan Dinakar, Deodhare Dipti, Ravindran B. Lexical Chains as Document Features. Third International Joint Conference on Natural Language Processing. Citeseer. 2008, pp. 111.
[31]. Enrique Amigo´, Julio Gonzalo, Javier Artiles et al. A comparison of extrinsic clustering evaluationmetrics based on formal constraints. Information retrieval. 2009. Т. 12, № 4, pp. 461–486.
[32]. Zhao Ying, Karypis George, Du Ding-Zhu. Criterion functions for document clustering: Tech. Rep.: Technical Report, 2005.
[33]. Meil˘a Marina. Comparing clusterings by the variation of information. Learning theory and kernel machines. Springer, 2003, pp. 173–187.
[34]. Hubert Lawrence, Arabie Phipps. Comparing partitions. Journal of classification. 1985. Т. 2, № 1, pp. 193–218.
[35]. Bakus J, Hussin MF, Kamel M. A SOM-based document clustering using phrases. Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. IEEE. Т. 5. 2002, pp. 2212–2216.

[36]. Vinh Nguyen Xuan, Epps Julien, Bailey James. Information theoretic measures for clusterings comparison: is a correction for chance necessary? Proceedings of the 26th Annual International Conference on Machine Learning. ACM. 2009, pp. 1073–1080.
[37]. Strehl Alexander, Ghosh Joydeep. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002. T. 3, № Dec., pp. 583–617.
[38]. Rosenberg Andrew, Hirschberg Julia. VMeasure: A Conditional Entropy-Based External Cluster Evaluation Measure. EMNLP-CoNLL. T. 7, 2007, pp. 410–420.
[39]. Bagga Amit, Baldwin Breck. Entity-based cross-document coreferencing using the vector space model. Proceedings of the 17th international conference on Computational linguistics-Volume 1. 1998, pp. 79–85.
[40]. Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza et al. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013. T. 46, № 1, pp. 243–256.
[41]. Yanchi Liu, Zhongmou Li, Hui Xiong et al. Understanding of internal clustering validationmeasures. 2010 IEEE International Conference on Data Mining. IEEE. 2010, pp. 911–916.
[42]. Er´endira Rend´on, Itzel Abundez, Alejandra Arizmendi et al. Internal versus external cluster validation indexes. International Journal of computers and communications. 2011. T. 5, № 1, pp. 27–34.
[43]. Rousseeuw Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987. T. 20, pp. 53–65.
[44]. Davies David L, Bouldin Donald W. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence. 1979. № 2, pp. 224–227.
[45]. Calin´ski Tadeusz, Harabasz Jerzy. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods. 1974. T. 3, № 1, pp. 1–27.
[46]. Bezdek James C, Pal Nikhil R. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 1998. T. 28, № 3, pp. 301–315.
[47]. Ibai Gurrutxaga, In˜aki Albisua, Olatz Arbelaitz et al. SEP/COP: An efficient method to find the bestpartition in hierarchical clustering based on a new cluster validity index. Pattern Recognition. 2010. T. 43, № 10, pp. 3364–3373.
[48]. Halkidi Maria, Vazirgiannis Michalis. Clustering validity assessment: Finding the optimal partitioning of a data set. Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE. 2001, pp. 187–194.
[49]. Bird Steven. NLTK: the natural language toolkit. Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics. 2006, pp. 69–72.
[50]. Scikit-learn: Machine Learning in Python. F. Pedregosa, G. Varoquaux, A. Gramfort et al. Journal of Machine Learning Research. 2011. T. 12, pp. 2825–2830.
[51]. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Yu. Methods for automatic term recognition in domain-specific text collections: A survey. Programming and Computer Software. 2015. T. 41, № 6, pp. 336–349.
[52]. Astrakhantsev Nikita. ATR4S: Toolkit with State-of-the-art Automatic Terms Recognition Methods in Scala. arXiv preprint, arXiv:1611.07804. 2016.
[53]. Reh˚u˘rek R., Sojka P.˘ Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, 2010, pp. 45–50.

[54]. Martin Ester, Hans-Peter Kriegel, J¨org Sander et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. T. 96. 1996, pp. 226–231.
[55]. Arthur David, Vassilvitskii Sergei. kmeans++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.
[56]. Lang Ken. Newsweeder: Learning to filter netnews. Proceedings of the 12th international conference on machine learning. 1995, pp. 331–339.
[57]. Krapivin M., Autaeu A., Marchese M. Large dataset for keyphrases extraction. 2009. URL: http://eprints.biblio.unitn.it/1671/1/disi09055krapivin-autayeu-marchese.pdf.
[58]. William Hersh, Aaron Cohen, Lynn Ruslen et al. TREC 2007 Genomics Track Overview. 2007.
[59]. Xie Pengtao, Xing Eric P. Integrating document clustering and topic modeling. arXiv preprint, arXiv:1309.6874. 2013.
[60]. Simone Romano, Nguyen Xuan Vinh, James Bailey et al. Adjusting for Chance Clustering Comparison Measures. arXiv preprint, arXiv:1512.01286. 2015.
[61]. Van Craenendonck Toon, Blockeel Hendrik. Using internal validity measures to compare clustering algorithms. AutoML Workshop at ICML 2015, pp. 1–8.
[62]. Field Andy. Discovering statistics using IBM SPSS statistics. Sage, 2013.
[63]. Kendall Maurice G. A new measure of rank correlation. Biometrika. 1938. T. 30, № ½, pp. 81–93.