Сравнительный анализ точности методов визуализации структуры коллекции текстов

Ф.В.Краснов

***Аннотация*— Визуализация многомерных данных является важнейшим этапом исследований данных. Зачастую от плоского вида данных «на глазок» принимаются решения по дальнейшим этапам исследования. Высокая наглядность и убедительность представления на плоскости многомерных векторов с сохранением расстояний успешно использована в моделях дистрибутивной семантики (Word2Vec, GloVe, NaVec). С другой стороны, неточность двухмерной проекции может привести к тому, что время будет потрачено на поиск несуществующих многомерных структур. Автор поставил задачу оценить точность методов уменьшения размерности со следующими ограничениями: многомерность возникает в результате векторного представления текстовых документов, уменьшение размерности нацелено на визуализацию на плоскости. В многочисленных методах уменьшения размерности не выделяют отдельного класса подходов именно для визуализации. Для измерения точности был выбран подход с использованием размеченных данных и количественной оценки сохранения разметки при уменьшении размерности. Автор исследовал 12 методов понижения размерности на двух размеченных наборах данных с русскими и английскими тестами. С помощью, метрики Коэффициента силуэта был определен наиболее точный метод визуализации для текстовых данных — UMAP с расстоянием Хеллингера в качестве метрики.**

***Ключевые слова*— обучение многообразиям, машинное обучение, визуализация данных, коллекции текстов, коэффициент силуэта.**

I. Введение

Огромное количество информации скапливается в многочисленных текстовых базах, хранящихся в личных ПК, локальных и корпоративных сетях. И объем этой информации стремительно увеличивается. Чтение объемных текстов и поиск в гигантских массивах текстовых данных малоэффективны, поэтому становятся все более востребованными решения по извлечению информации из неструктурированных данных (текстов). Извлечение информации — это задача автоматического построения структурированных данных из машиночитаемых документов. К машиночитаемым документам могут относится как слабо структурированные (научные статьи, нормативные акты, техническая документация), так и сильно структурированные (счет-фактуры, медицинские карты, договора).

Методика извлечения информации для коллекций текстов строится на терм-документной матрице (ТДМ). В ТДМ строки соответствуют документам из коллекции, а столбцы соответствуют термам. Существуют различные способы для определения значения каждого элемента матрицы. Одним и способов является TF-IDF. Основными особенностями ТДМ являются их высокая размерность и разряженность. Эффективным представлением для ТДМ является разряженная матрица (sparse matrix), значительная часть элементов которой состоит из нулей. Для работы с ТДМ разработан специальный математический аппарат, позволяющий выполнять основные операции линейной алгебры, такие как умножение и инвертирование, в разряженном формате.

Для коллекции из N документов с M уникальными термами размерность ТДМ будет NxM. Элементы ТДМ v_{ij} будут принадлежать пространству $v_{ij} \in R^{NxM}$. Разместить такую матрицу в оперативной памяти для выполнения матричного разложения или поиска собственных векторов даже для средних по размеру коллекций документов не возможно. При степени разряженности $\mu = 99\%$ все преобразования в оперативной памяти выполняются только над (1- μ) * M * N элементами, что приводит к уменьшению хранимой информации на два порядка.

Все виды интеллектуального анализа текстов имеют этап направленный на визуализацию данных. Наборы данных большой размерности может быть очень трудно визуализировать. В то время как данные в двух или трех измерениях могут быть построены для отображения внутренней структуры данных, эквивалентные многомерные графики гораздо менее интуитивно понятны. Чтобы облегчить визуализацию структуры набора данных, необходимо каким-то образом уменьшить размерность.

Самый простой способ добиться этого уменьшения размерности - сделать случайную проекцию данных. Хотя это позволяет в некоторой степени визуализировать структуру данных, случайность выбора оставляет желать лучшего. При случайной проекции наиболее интересная структура данных, вероятно, будет потеряна.

Для решения этой проблемы был разработан ряд контролируемых и неконтролируемых структур

Статья получена 9 апреля 2021.

Ф.В.Краснов, к.т.н., Директор департамента семантических систем, NAUMEN R&D, 620028, г.Екатеринбург, ул. Татищева, 49А, БЦ «Татищевский», каб. 432 (e-mail: fkrasnov@naumen.ru), https://orcid.org/0000-0002-9881-7371.

снижения линейной размерности, таких как анализ главных компонентов (PCA), независимый компонентный анализ, линейный дискриминантный анализ и другие. Эти алгоритмы определяют конкретные рубрики для выбора «интересной» линейной проекции данных. Эти методы могут быть мощными, но часто упускают важную нелинейную структуру данных.

Обучение многообразиям (Manifold Learning) можно рассматривать как попытку обобщить линейные структуры, такие как PCA, чтобы они были чувствительны к нелинейной структуре данных. Хотя существуют варианты обучения с учителем, типичная проблема обучения многообразию не учитывается: она изучает многомерную структуру данных из самих данных без использования заранее определенных классификаций.

Особенности ТДМ представляют дополнительную сложность для всех без исключения методов уменьшения размерности для визуализации. Но и сами методы вносят дополнительную степень свободы за счет собственных гипер параметров.

В настоящем исследовании автор поставил задачу сравнения различных методов понижения размерности для визуализации структуры коллекции текстов следующим образом: путем сравнения метрик качества выделяемых кластеров найти лучшую функцию понижения размерности для визуализации текстовых коллекций.

Математическая формулировка задачи данного исследования выглядит следующим образом:

$$Clustering\ quality\ metric \xrightarrow{f} max,$$

где $f: v_{ij} \in R_+^{N \cdot M} \to R^{N \cdot \widehat{M}}$ при $\widehat{M} = 2$ и ТДМ = $\{v_{ij}\}$.

Так как в ТДМ присутствуют числа от 0 до 1, то пространство R можно сузить до положительных чисел R_+. Но из-за высокой разряженности ТДМ вектора документов v_i не являются плотными распределениями вероятностей. Это накладывает ограничения на возможные для использования методы уменьшения размерности.

Статья состоит из введения, обзора методик, описания эксперимента и заключения.

II.Методика

В соответствии с задачей данного исследования необходимо найти какой метод уменьшения размерности дает наилучший результат при работе с ТДМ для визуализации.

Методы уменьшения размерности делятся на линейные и нелинейные. К линейным относятся PCA и его вариации, SVD, Neighborhood Components Analysis (NCA) и Non-negative matrix factorization (NMF). К нелинейным методам уменьшения размерности относят t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Multi-dimensional Scaling (MDS), Isomap, Spectral Embeding (SE), Linear Discriminant Analysis (LDA) и Modified Locally-Linear Embedding (MLLE). Все вышеперечисленные методы могут быть применены к анализу ТДМ. С другой стороны есть методы, которые работают только по принципу обучения с учителем (LDA, NCA), есть методы, которые работают только по принципу обучения без учителя (PCA, SVD, NMF, t-SNE, MDS, SE, MLLE) и есть методы, которые поддерживают оба варианта обучения (UMAP).

Большинство методов уменьшения размерности имеют собственные гипер параметры, которые значительно влияют на получаемую двухмерную проекцию $R^{N \cdot \widehat{M}}$. Примером такого влияния согласно [1] служит параметр перплексия для t-SNE. В зависимости от значений перплексии от 10 до 100 вид $R^{N \cdot \widehat{M}}$ изменяется драматически.

Одним из современных методов, показывающих результаты визуализации лучше чем t-SNE является UMAP [2]. UMAP часто лучше сохраняет некоторые аспекты глобальной структуры данных, чем большинство реализаций t-SNE. Это означает, что он часто может обеспечить лучшую «общую картину» ваших данных, а также сохранить отношения между соседями.

В работе [4] представлены den-SNE и densMAP — инструменты визуализации с сохранением плотности, основанные на t-SNE и UMAP, соответственно, предназначенные для более точной визуальной интерпретации многомерных данных в различных научных областях.

Отдельно стоит отметить метод Totally Random Trees Embedding (RTE), который кодирует данные по индексам листьев, на которых заканчивается вектор документа. Затем этот индекс кодируется, что приводит к высоко размерному разреженному двоичному кодированию. Поскольку соседние точки данных с большей вероятностью лежат в одном листе дерева, преобразование выполняет неявную непараметрическую оценку плотности.

Большинство исследований [1,2] предлагает оценивать картинки визуально. Такой подход может внести субъективность в оценку в отличии от оценки с помощью метрик качества кластеров в наборе данных. Методика, предлагаемая автором, состоит в том, чтобы создать набор данных из различных коллекций текстов. При этом сохранив для каждого собственные метки классов. При таком подходе мы знаем истинное разбиение (ground truth) и можем оценить отклонения.

Для исследования выбрана метрика Коэффициент силуэта (silhouette score) [3], так как она имеет в своей формуле нормировку, позволяющую сравнивать пересечения кластеров для различных методов уменьшения размерности. Лучшее значение равное единице один (1), а худшее - минус единице (-1). Значения около 0 указывают на перекрывающиеся кластеры. Отрицательные значения обычно указывают

на то, что образец был назначен не тому кластеру, поскольку другой кластер более похож.
Коэффициент силуэта является примером такой оценки, где более высокий показатель коэффициента силуэта относится к модели с лучше определенными кластерами. Коэффициент силуэта определяется для каждого документа d_i и состоит из двух составляющих:

a: Среднее расстояние между d_i и всеми другими d_j того же класса:

$$a(i) = \frac{1}{|C_i|-1}\sum_{j \in C_i, i \neq j} d(i,j) \quad (1)$$

b: Среднее расстояние между d_i и всеми другими d_j в следующем ближайшем кластере.

$$b(i) = min_{(k \neq i)} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (2)$$

Тогда коэффициент силуэта s_i для одного документа d_i определяется как:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (3)$$

Коэффициент силуэта (S) для коллекции документов вычисляется как среднее значение коэффициента силуэта для каждого документа.

III. Описание эксперимента

Для проведения эксперимента были созданы два набора данных: коллекции текстов на русском и на английском языках. Для создания коллекции текстов на английском был взят за основу набор данных из [5], но выбраны наиболее контрастные по смыслу группы — alt.atheism, comp.graphics, misc.forsale, sci.med. Для создания коллекции текстов на русском языке использован тот же принцип смешения различных коллекций контрастных по смыслу. Взяты фрагмент из НКРЯ, тексты ГОСТов по тематике ИТ, тексты ГОСТов по тематике железные дороги [8] и корпус ТАЙГА [6] в равных пропорциях. Оба набора данных сбалансированы по количеству документов в кластерах.

Далее к обоим наборам данных применены методы уменьшения размерности с учителем: LDA, NCA и UMAP. В результате автор получил базовую точность выделения кластеров из наборов данных. Метрикой для оценки точности выделения кластеров послужил Коэффициент силуэта. Результаты этого этапа исследования приведены в Таблице 1.

Таблица 1 Значения Коэффициента силуэта для коллекции для обучения с учителем

Метод			LDA	NCA	UMAP
Набор данных на русском	**ТДМ (COUNTS)**	**COSINE**	**0,977**	0,507	**0,977**
		EUCLEDIAN	**0,932**	0,357	0,898
	ТДМ (TF-IDF)	**COSINE**	0,617	0,812	**0,987**
		EUCLEDIAN	0,540	0,763	**0,914**
Набор данных на английском	**ТДМ (COUNTS)**	**COSINE**	**0,966**	0,827	**0,935**
		EUCLEDIAN	**0,921**	0,255	0,879
	ТДМ (TF-IDF)	**COSINE**	**0,946**	0,635	0,781
		EUCLEDIAN	**0,931**	0,427	**0,905**

При построении ТДМ было использовано два подхода: счетчики слов (COUNTS) и веса TF-IDF с нормализацией L_2. Для вычисления расстояния $d(i,j)$ в формулах 1 и 2 использованы две метрики расстояния: косинусная (COSINE) и Евклидова (EUCLEDIAN).

Из Таблицы 1 следует, что кластеры в наборах данных выделяются после понижения размерности для визуализации с высокой точностью. Методы LDA и UMAP дают наименьшую ошибку (2%-4%) при использовании косинусного расстояния в Коэффициенте силуэта для обоих наборов данных. В русскоязычном наборе данных точность выделения выше на 4%-8%. Метод NCA показал достаточно низкие значения точности, но отсутствие гипер параметров не позволяет сделать точную настройку. Метод LDA так же не имеет гипер параметров. А UMAP использован с настройками гипер параметров по умолчанию, то есть без оптимизации.

На следующем этапе эксперимента проведено измерение качества понижения размерности для методов работающих без учителя.

Таблица 2. Значения Коэффициента силуэта для коллекции на английском (обучение без учителя)

Метод	Набор данных на английском			
	ТДМ (COUNTS)		ТДМ (TF-IDF)	
	COSINE	EUCLEDIAN	COSINE	EUCLEDIAN
Isomap	0,209	0,192	0,394	0,3
MDS	-0,026	-0,08	-0,017	-0,038
MLLE	-0,377	-0,471	-0,301	-0,12
RTE	-0,333	-0,121	-0,163	-0,078
SE	-0,021	-0,016	**0,566**	0,411
SRP	-0,018	-0,1	-0,019	-0,035
SVD	-0,153	-0,209	0,244	0,093
t-SNE p100	-0,037	-0,002	**0,533**	0,371
t-SNE p50	0,036	0,047	**0,502**	0,352
t-SNE p30	0,079	0,087	0,431	0,309
UMAP cosine	-0,005	0,273	0,118	0,314
UMAP hellinger	0,138	0,366	0,103	0,385

Из Таблицы 2 видно, что точность для методов визуализации без учителя существенно ниже, чем в случае обучения с учителем для английского корпуса текстов. Но в большинстве случаев на практике методы обучения без учителя — это единственная возможность визуализировать структуру данных коллекции текстов.

Наилучшую точность S = 0,566 для методов визуализации без учителя показал метод Spectral Embedding (SE) для ТДМ на основе TF-IDF с косинусным расстоянием в качестве метрики для формул (1,2). На рисунке 1 приведено изображение полученных кластеров.

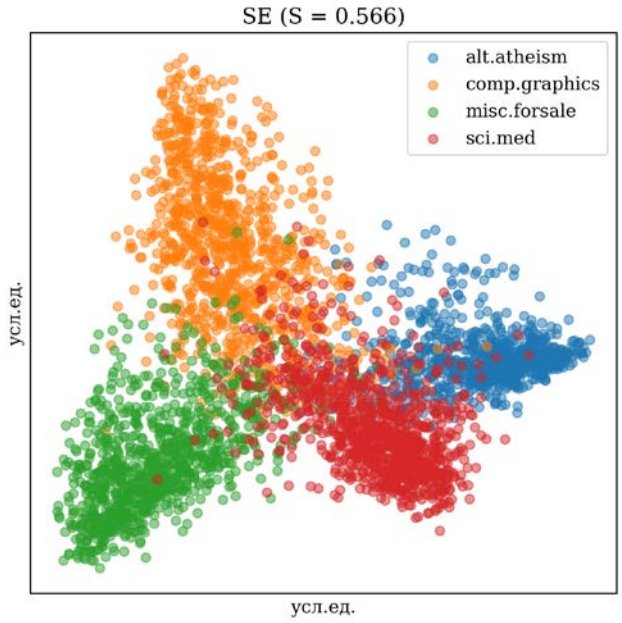

Рисунок 1. Визуализация с помощью Spectral Embedding.

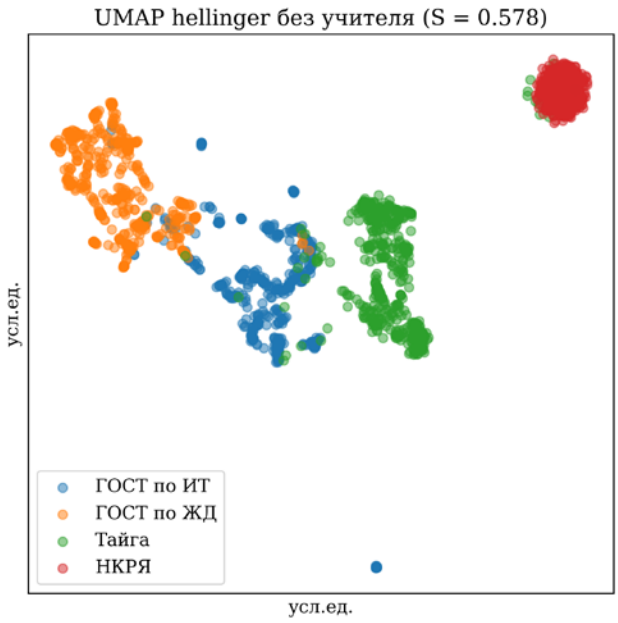

Рисунок 2. Визуализация с помощью UMAP

Для коллекции текстов на русском языке точность визуализации несколько выше, чем на английском. Сразу несколько методов показали значения точности больше 0,5 .

Таблица 3. Значения Коэффициента силуэта для коллекции на русском (обучение без учителя)

Метод	Набор данных на русском			
	ТДМ (COUNTS)		ТДМ (TFIDF)	
	COSINE	EUCLEDIAN	COSINE	EUCLEDIAN
Isomap	0,425	0,433	0,453	**0,485**
MDS	0,203	0,166	-0,057	0,013
MLLE	0,045	0,140	0,053	0,466
RTE	0,094	0,159	0,099	0,154
SE	**0,479**	0,412	0,429	0,391
SRP	0,013	-0,074	-0,099	-0,023
SVD	0,257	0,336	**0,515**	**0,564**
t-SNE p100	0,145	0,297	0,305	0,329
t-SNE p50	0,461	0,394	**0,504**	0,352
t-SNE p30	0,435	0,371	0,198	0,325
UMAP cosine	0,082	0,453	0,039	0,500
UMAP hellinger	0,357	**0,574**	0,251	**0,578**

Но самую высокую точность S=0,578 показал метод UMAP гипер параметром metric='hellinger'. На рисунке 2 показана визуализация кластеров для этого случая.

Отметим, что опция densMAP [3] для UMAP, которая была использована в отдельном эксперименте, показала точность S=0,565. Это означает, что заявленный эффект сохранения локальной плотности не проявился при использовании в качестве метрики коэффициента силуэта.

Несмотря на то, что метод UMAP показал наилучшую точность для русской коллекции как при обучении с учителем, так и без обучения с учителем, сложность использования этого метода состоит в значительном количестве гипер параметров. Кроме того, низкая точность UMAP для английской коллекции не позволяет назвать этот метод универсальным.

IV. Заключение

В работе рассмотрены методы уменьшения размерности и произведено сравнение их точности для визуализации многомерных данных. В качестве многомерных данных автор использовал ТДМ на основании счетчиков слов и весов TF-IDF для коллекций текстов на русском и английском.

Автор показал, что при обучении с учителем кластеры в наборах данных выделяются после понижения размерности для визуализации с высокой (>95%) точностью. Методы LDA и UMAP при обучении с учителем дают наименьшую ошибку (2%-4%) при использовании косинусного расстояния в Коэффициенте силуэта для обоих наборов данных. В русскоязычном наборе данных точность выделения кластеров выше на 4%-8%, чем в англоязычном, что объясняется более высокой информационной энтропией русского языка.

Методы обучения без учителя для визуализации показывают ошибку не менее 40%. Наиболее точный результат визуализации (S=0.578) продемонстрировал метод UMAP, использующий расстояние Хеллингера.

Важным выводом данного исследования стала не самая точная (S=0.5), но стабильная для обоих наборов

данных визуализация, показанная методом t-SNE.

Методы Isomap, MLLE, RTS и MDS не показали высокую точность визуализации.

Задача понижения размерности применительно к тестам не достаточно изучена с точки зрения точности и объяснимости [7].

Библиография

[1] Maaten, L. V. D. and Geoffrey E. Hinton. “Visualizing Data using t-SNE.” Journal of Machine Learning Research 9 (2008): 2579-2605.

[2] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

[3] Peter J. Rousseeuw . “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics (1987) 20: 53–65. doi:10.1016/0377-0427(87)90125-7

[4] Narayan, A., Berger, B., & Cho, H. “Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability.” bioRxiv (2020).

[5] Lang, Ken. "Newsweeder: Learning to filter netnews." Machine Learning Proceedings 1995. Morgan Kaufmann, 1995. 331-339.

[6] Shavrina T., Shapovalova O. “To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser”. In Proc. of “CORPORA2017”, International Conference , Saint-Petersbourg, (2017).

[7] Краснов Ф.В., Смазневич И.С. Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов // Вычислительные технологии. 2020. Т. 25. № 5. С. 107-123.

[8] Краснов Ф. В., Баскакова Е. Н., Смазневич И. С. 2021. Принцип построения корпуса нормативно-технических документов. PREPRINTS.RU. https://doi.org/10.24108/preprints-3112181.

Comparative Analysis of the Accuracy of Methods for Visualizing the Structure of a Text Collection

F.V. Krasnov

Abstract— Visualization of multidimensional data is the most important stage of data research. Often, decisions on the further stages of the study are made from the flat view of the data based on "rough proportions". High visibility and persuasiveness of representation on the plane of multidimensional vectors with the preservation of distances is used in models of distributive semantics (Word2Vec, GloVe, NaVec) successfully. On the other hand, the inaccuracy of the two-dimensional projection can lead to time being spent searching for non-existent multidimensional structures. The author set the task to evaluate the accuracy of dimensionality reduction methods with the following limitations: multi-dimensionality arises as a result of vector representation of text documents, dimensionality reduction is aimed at visualization on the plane. In numerous methods of dimension reduction, there is no separate class of approaches specifically for visualization. To measure the accuracy, an approach was chosen using marked-up data and quantifying the preservation of the markup while reducing the dimension. The author investigated 12 methods of reducing the dimension on two labeled data sets in Russian and English. Using the Silhouette Coefficient metric, the most accurate visualization method for text data was determined as UMAP with the Hellinger distance as the metric.

Keywords - diversity learning, machine learning, data visualization, text collections, silhouette coefficient.

REFERENCES

[1] Maaten, L. V. D. and Geoffrey E. Hinton. “Visualizing Data using t-SNE.” Journal of Machine Learning Research 9 (2008): 2579-2605.

[2] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

[3] Peter J. Rousseeuw . “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics (1987) 20: 53–65. doi:10.1016/0377-0427(87)90125-7

[4] Narayan, A., Berger, B., & Cho, H. “Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability.” bioRxiv (2020).

[5] Lang, Ken. "Newsweeder: Learning to filter netnews." Machine Learning Proceedings 1995. Morgan Kaufmann, 1995. 331-339.

[6] Shavrina T., Shapovalova O. “To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser”. In Proc. of “CORPORA2017”, International Conference , Saint-Petersbourg, (2017).

[7] Krasnov F.V., Smaznevich I.S. The explicability factor of the algorithm in the problems of searching for the similarity of text documents // Computational technologies. 2020. V. 25. № 5. P. 107-123

[8] Krasnov F.V., Baskakova E.N., Smaznevich I.S. 2021. The principle of constructing a corpus of normative and technical documents. PREPRINTS.RU. https://doi.org/10.24108/preprints-3112181.

F.V. Krasnov, Ph.D., Director of the Department of Semantic Systems, NAUMEN R&D, 620028, Yekaterinburg, st. Tatishcheva, 49A, Business Center "Tatishchevsky", office. 432 (e-mail: fkrasnov@naumen.ru), https://orcid.org/0000-0002-9881-7371.