ПРИКЛАДНЫЕ АСПЕКТЫ МАТЕМАТИЧЕСКОГО ПРИМЕНЕНИЯ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА К КЛАССИФИКАЦИИ НАУЧНЫХ ТЕКСТОВ

APPLIED ASPECTS OF MATHEMATICAL APPLICATION OF LATENT-SEMANTIC ANALYSIS TO CLASSIFICATION OF SCIENTIFIC TEXTS

A. Ismukanova

Summary. The classification of scientific texts in the Russian and Kazakh languages, by means of assignment of a universal decimal code (UDC) is an actual problem nowadays. The problem of the classification of scientific texts is easily solved for the English language due to its simplicity of morphology and syntax. On this way there is a number of unresolved tasks for the Russian language and practically the usage of analogical reception for the Kazakh language is not investigated. For the Russian language several researches of applicability of different approaches were conducted.

New technologies for the LSA model could represent a important advance of the assessment of scientific texts.

Keywords: latent semantic analysis (LSA), machine learning (ML), classification.

Исмуканова Айгерим Наурызбаевна
Аспирант, Омский государственный университет им. Ф. М. Достоевского, г. Омск
aigera_ismukan@mail.ru

Аннотация. Классификация научных текстов на русском и казахском языках, посредством присвоения им универсального десятичного кода (УДК) является актуальной задачей. Задача классификации научных текстов прекрасно решается для английского языка в силу простоты морфологии и синтаксиса этого языка. На данном пути имеется ряд нерешенных задач для русского языка и практически не исследовано использование аналогичных приемов для казахского языка. Для русского языка проводились несколько исследований применимости разных подходов.

Новые технологии, для модели LSA (латентного-семантического анализа) могли представлять важное усовершенствование в исследовании оценки научных текстов.

Ключевые слова: латентный семантический анализ (ЛСА), машинное обучение (МА), классификация, матрица.

Модель Latent Semantic Analysis (ЛСА) (Landauer & Dumais, 1997) является теорией значения независимо от структуризации текста, направлено на уровень восприятия информации с большим набором языковых переплетений [3, с. 211–240]. С целью достижения поставленных задач латентно семантический анализ разделяет на 2 (два) вида значения, а именно в дистрибутивных образцах лингвистического выражения присутствуют простые выражения (например, слова) и в рамках более сложных выражениях (например, предложения и параграфы) рассматривает в более обширных образцах языка.

Прежде чем обратиться к применению ЛСА для определения текста казахского и русского языков, мы представляем явный вычислительный алгоритм, используемый в ЛСА, чтобы изучить семантические представления и вывести ассоциации среди слов.

Исходные данные латентно-семантического анализа применялись для обучения системы (слова, словосочетания, термины) классификации научных текстов. ЛСА основан на использовании разложения вещественной матрицы по сингулярным значениям или SVD — разложения (SVD — Singular Value Decomposition).

С помощью него любую матрицу можно разложить во множество ортогональных матриц, линейная комбинация которых является достаточно точным приближением к исходной матрице.

Работы по усовершенствованию и адаптации к различным задачам латентного семантического анализа (ЛСА) ведутся давно. Суть метода достаточна проста. В начале на вход алгоритму поступает набор текстов, который преобразуется в матрицу частоты встречаемости слов в этих текстах. Номер строки соответствует слову, а номер столбца тексту. С помощью алгоритма сингулярного разложения (SVD) у полученной матрицы понижается ранг. Это позволяет отбросить зависимости слов и выделить так называемое семантическое ядро. Затем на основе полученной матрицы с пониженным рангом вычисляются коэффициенты корреляции между текстами. В одной из первых работ [3, с. 211–240] эмпирически определен порог, по которому можно сгруппировать тексты по схожей тематике. Если часть из этих текстов

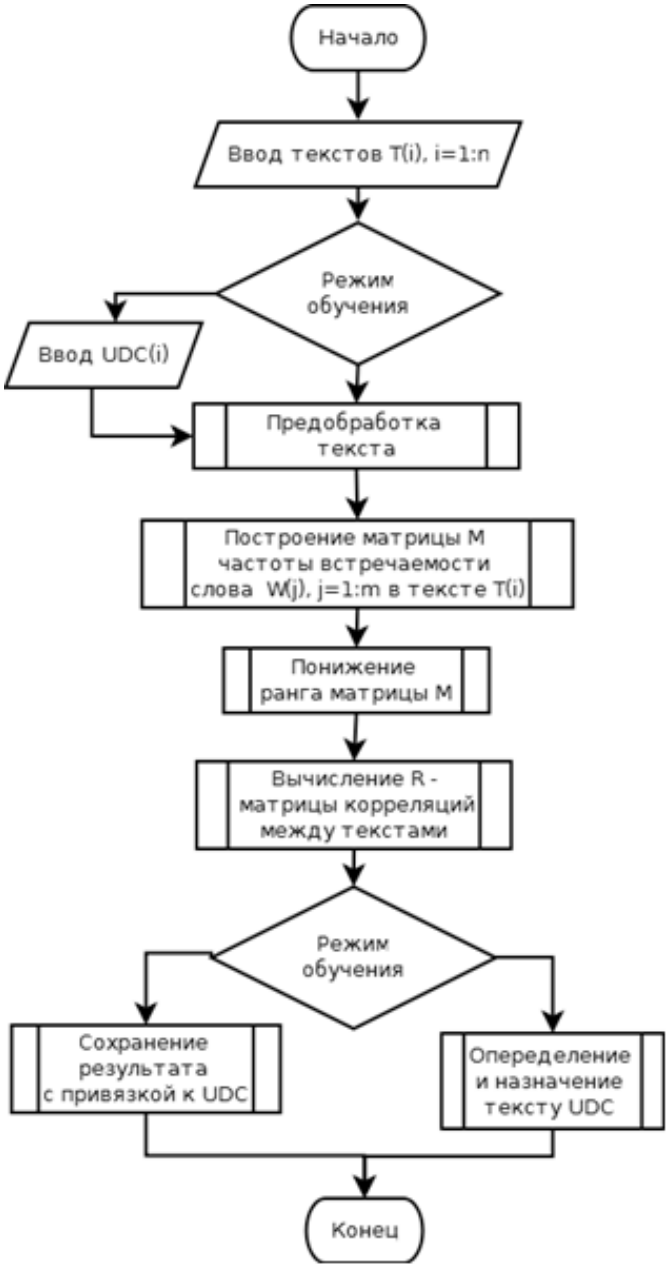

Рис. 1. Блок схема алгоритма

уже имеет универсальный десятичный код (УДК), конечно, правильно выставленный автором или редактором, то этот код или близкий к нему будет и у всей группы. Это позволит вычислить УДК в автоматическом режиме.

Описанная схема представлена на рис. 1.

Для алгоритма берем в документе в качестве слова — матрицу, m*n, в которой каждый вход a_{ij} является местной частотой данного i слова в данном документе j. Это неопределенное количество совместной встречаемости документа слова сначала нужно преобразовать, чтобы загрузить каждое слово согласно тому, насколько нормативным это находится в определении значения документа. Можно через сингулярное разложение (SVD) уменьшить преобразование матрицы [4, с.2276–2283].

Таким образом, прежде, чем вывести семантические отношения, алгоритм преобразовывает данные (Landauer &Dumais, 1997). Во-первых, чтобы приблизить темп роста простого изучения, каждый вход a_{ij} преобразован от его местной частоты до его истинного веса, где для каждого слова i в документе j соответствует [1, с 4–11]:

$$weight_{ij}^{loc} = \log\left(freq_{ij}^{loc} + 1\right) \quad (1)$$

Затем, глобальный вес каждого слова, которое включает информацию теоретической энтропии слова через документы, вычисляется:

$$weight_{i}^{glob} = \frac{1+\sum_{j=1}^{n} p_{ij} * \log\left(p_{ij}\right)}{\log\left(n\right)} \quad (2)$$

В уравнении для $weight^{glob}_{i}$ количество p_{ij} определено как местная частота слова, которое я разделила на глобальную частоту того слова через все документы j

$$p_{ij} = \frac{freg_{ij}^{loc}}{\sum_{j=1}^{n} freg_{ij}^{loc}} \quad (3)$$

Взвешенная ценность каждого термина — таким образом, имеет вес, разделенный на его глобальный вес,

$$weight_{ij}^{term} = \frac{weight_{ij}^{loc}}{weight_{i}^{glob}} \quad (4)$$

Заметьте, что, если слово будет несколько раз встречаться в документе, то его вес будет относительно большим, так как это непосредственно связано с местной частотой. Кроме того, если слово будет часто встречатся в нескольких документах, то его глобальный вес будет высок, поскольку это непосредственно связано с энтропией слова. Так как вес термина непосредственно связан с местным весом, но обратно пропорционально связан с глобальным весом, из этого следует, что слово будет высоко нагружено, если это будет часто встречаться в документе относительно других слов в документе, но нечасто среди всех документов относительно его частоты в данном документе. В действительности надбавка термина уменьшает важность слов, присутствие которых в документе неинформативно к определению значения того документа. Ассоциации между пунктами лучше создают их информативностью, а не просто их совместной встречаемостью [1, с 4–11].

Факторизация измерения и сокращение

Матрица документа слова M подвергся нагружающему терминному преобразованию, это может быть учтено, используя сингулярное разложение. Теорема линейной алгебры — это любая m*n матрица M, чьи записи — действительные числа, может анализироваться в три матрицы T, ∑, D^T, следующим образом:

$$M = T, \sum, D^T \quad (5)$$

В вышеупомянутом разложении T — m*m матрица, и D^T — n * n матрица; у обоих из которых есть ортогональные колонки. Колонки матрицы называют ортогональными (orthonormal), если каждый вектор колонки **v** в матрице является вектором единицы (т.е., v _v = 1) и каждые две отличительных колонки v, ортогональны друг другу (т.е., v _ u = 0). Матрица имеет следующую форму:

$$= \begin{bmatrix} D & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (6)$$

где D — диагональная матрица, которая не превышает меньший из m или n, размеров ортогональной матрицы M. D называют диагональной матрицей, потому что ее единственные отличные от нуля записи находятся на одной из ее главных диагоналей. Ценности на главной диагонали D после AP — сложение SVD называют исключительными ценностями M, и им заказывают от самого большого до наименьшего количества вдоль главной диагонали D. Исключительные ценности представляют размеры значения для слов, и когда D содержит все исключительные ценности M, оригинальная матрица M может быть точно восстановлена, умножив эти три матрицы T, Σ, и D^T.

Представления и общие черты в уменьшенном пространстве

Размерность пространства семантических представлений может быть уменьшена, если заменив некоторые исключительные ценности на 0. В соответствии с соглашением, исключительные ценности обнулены — из наименьшего количества к самому большому. Если бы мы должны были выбрать s самое большое количество исключительных ценностей, таких, как s — меньше, чем количество исключительного размера, которые первоначально следуют из SVD, то мы могли построить матрицу, которая является s размерным приближением к M с наименьшим количеством ошибки,

$$M_s = T*\Sigma_s* D^T \quad (7)$$

Количество размеров, в которых можно восстановить матрицу совместной встречаемости документа слова, свободно определено пользователем, в качестве параметра внешним к другим работам алгоритма.

Представления слов и документов могут аналогично быть получены, умножив их соответствующие разложения уменьшенной пространственной исключительной матрицей, _s. Таким образом, представления слов в s-пространстве дают

$$T_s = T\Sigma_s \quad (8)$$

и уменьшенные представления документов

$$D_s^T = \sum_s D^T \quad (9)$$

Общие черты между двумя векторами, v_1 и v_2, что каждый предоставляет слово или документ, вычисленным при помощи косинусом угла, между v_1 и v_2,

$$\cos(\theta) = \frac{V_1 * V_2}{\|V_1\| * \|V_2\|} \quad (10)$$

где для вектора v, ||v || длина того вектора и

$$\|V\| = \sqrt{V * V} \quad (11)$$

Обратите внимание, что подобие между двумя векторами v_1 и v_2 непосредственно связано с косинусом угла между другими векторами. Если косинус угла между двумя векторами равняется 1, то эти два вектора синонимичны. Общие черты между двумя векторами слова v_1 и v_2 в уменьшенном пространстве могут быть собраны в единственную матрицу:

$$M_s M_s^T = T_s D_s^T (T_s D_s^T)^T = T_s D_s^T D_{sS}^T T^T = \quad (12)$$

$$= T_{sS}^T T^T = T_s T_s^T$$

От этой матрицы косинус угла между двумя векторами слова v_1 и v_2 может быть вычислена, разделяясь на общие части документа. Документ в уменьшенном пространстве аналогично вычисляются:

$$M_s^T M_s = D_s D_s^T \quad (13)$$

косинус угла между двумя векторами документа d_i может быть вычислен для соответствующего входа на i_{th} ряде и j_{th} колонке.

Для общих черт документа слова в уменьшенном пространстве вход в i_{th} ряду и j_{th} колонка разделены на соответственно слово — и вектор документа, который был переведен на пространство промежуточного звена, таким образом:

$$q'_i = q_i \sqrt{} \quad (14)$$

$$d'_j = \sqrt{d_j} \quad (15)$$

LSA может также использоваться, чтобы определить подобие слов или документов, которые использовались, чтобы определить пространство понятия с документами, относящимся к нему. Чтобы сделать такое сравнение, сначала необходимо преобразовать документ в псевдодокумент уменьшенного пространства,

$$query = q^T U_s^{-1} \quad (16)$$

где q^T- нагруженный термином вопрос. Как только вопрос был преобразован в соответствующее пространство, общие черты слова могут быть вычислены через меры по косинусу, как описано выше (Martin & Berry, 2007) [2, 35–55].

Применяя данный математический метод можно классифицировать различные научные тексты присваивая УДК к документам. В программном корпусе для классификации научных текстов преимущественно используются методы машинного обучения, где в качестве признаков традиционно применяются лексемы.

Данная работа подготовлена на основе выполненных научных работ опубликованных в авторитетных журналах или в трудах международных конференций.

ЛИТЕРАТУРА

1. Martin, D. & Berry, M. 2007. Mathematical Foundations Behind Latent Semantic Analysis. In T. Landauer, D. McNamara, S. Dennis, W. Kintsch (eds.), Lawrence Erlbaum Associates, Inc.: Mahwah, NJ.pp. 35–55.
2. Исмуканова А.Н., Лавров Д. Н.//IV Международной научной конференции «Математическое и компьютерное моделирование», ОмГУ им. Ф. М. Достоевского 11 ноября 2016 г.// Алгоритм классификации научных текстов методом латентного семантического анализа. с. 72–74.
3. Landauer T.K., Dumais S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge // Psychological Review. 1997. 104. PP. 211–240.
4. Ju R. et al. An Efficient Method for Document Categoriza-tion Based on Word2vec and Latent Semantic Analysis. 2015 IEEE Intern. Conf. on Comp. and Inform. Technology; Ubiquitous Com-puting and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. Liverpool, UK, 2015, pp. 2276–2283.

Журнал «Современная наука: актуальные проблемы теории и практики»