КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 5 С. 1185–1197
DOI: 10.20537/2076-7633-2022-14-5-1185-1197

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.8

Семантическая структуризация текстовых документов на основе паттернов сущностей естественного языка

Н. А. Игнатьев[a], У. Ю. Тулиев[b]

Национальный университет Узбекистана,
Ташкент, Узбекистан

E-mail: [a] n_ignatev@rambler.ru, [b] u.tuliyev@mail.ru

Получено 30.03.2022, после доработки — 07.06.2022.
Принято к публикации 08.06.2022.

Рассматривается технология создания паттернов из слов (понятий) естественного языка по текстовым данным в модели «мешок слов». Паттерны применяются для снижения размерности исходного пространства в описании документов и поиска семантически связанных слов по темам. Процесс снижения размерности реализуется через формирование по паттернам латентных признаков. Исследуется многообразие структур отношений документов для разбиения их на темы в латентном пространстве.

Считается, что заданное множество документов (объектов) разделено на два непересекающихся класса, для анализа которых необходимо использовать общий словарь. Принадлежность слов к общему словарю изначально неизвестна. Объекты классов рассматриваются в ситуации оппозиции друг к другу. Количественные параметры оппозиционности определяются через значения устойчивости каждого признака и обобщенные оценки объектов по непересекающимся наборам признаков.

Для вычисления устойчивости используются разбиения значений признаков на непересекающиеся интервалы, оптимальные границы которых определяются по специальному критерию. Максимум устойчивости достигается при условии, что в границах каждого интервала содержатся значения одного из двух классов.

Состав признаков в наборах (паттернах из слов) формируется из упорядоченной по значениям устойчивости последовательности. Процесс формирования паттернов и латентных признаков на их основе реализуется по правилам иерархической агломеративной группировки.

Набор латентных признаков используется для кластерного анализа документов по метрическим алгоритмам группировки. В процессе анализа применяется коэффициент контентной аутентичности на основе данных о принадлежности документов к классам. Коэффициент является численной характеристикой доминирования представителей классов в группах.

Для разбиения документов на темы предложено использовать объединение групп по отношению их центров. В качестве закономерностей по каждой теме рассматривается упорядоченная по частоте встречаемости последовательность слов из общего словаря.

Приводятся результаты вычислительного эксперимента на коллекциях авторефератов научных диссертаций. Сформированы последовательности слов из общего словаря по четырем темам.

Ключевые слова: тематическое моделирование, иерархическая агломеративная группировка, онтология, общий словарь, контентная аутентичность

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 5 P. 1185–1197
DOI: 10.20537/2076-7633-2022-14-5-1185-1197

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.8

Semantic structuring of text documents based on patterns of natural language entities

N. A. Ignatev[a], U. Yu. Tuliyev[b]

National university of Uzbekistan,
Tashkent, Uzbekistan

E-mail: [a] n_ignatev@rambler.ru, [b] u.tuliyev@mail.ru

Received 30.03.2022, after completion — 07.06.2022.
Accepted for publication 08.06.2022.

The technology of creating patterns from natural language words (concepts) based on text data in the bag of words model is considered. Patterns are used to reduce the dimension of the original space in the description of documents and search for semantically related words by topic. The process of dimensionality reduction is implemented through the formation of patterns of latent features. The variety of structures of document relations is investigated in order to divide them into themes in the latent space.

It is considered that a given set of documents (objects) is divided into two non-overlapping classes, for the analysis of which it is necessary to use a common dictionary. The belonging of words to a common vocabulary is initially unknown. Class objects are considered as opposition to each other. Quantitative parameters of oppositionality are determined through the values of the stability of each feature and generalized assessments of objects according to non-overlapping sets of features.

To calculate the stability, the feature values are divided into non-intersecting intervals, the optimal boundaries of which are determined by a special criterion. The maximum stability is achieved under the condition that the boundaries of each interval contain values of one of the two classes.

The composition of features in sets (patterns of words) is formed from a sequence ordered by stability values. The process of formation of patterns and latent features based on them is implemented according to the rules of hierarchical agglomerative grouping.

A set of latent features is used for cluster analysis of documents using metric grouping algorithms. The analysis applies the coefficient of content authenticity based on the data on the belonging of documents to classes. The coefficient is a numerical characteristic of the dominance of class representatives in groups.

To divide documents into topics, it is proposed to use the union of groups in relation to their centers. As patterns for each topic, a sequence of words ordered by frequency of occurrence from a common dictionary is considered.

The results of a computational experiment on collections of abstracts of scientific dissertations are presented. Sequences of words from the general dictionary on 4 topics are formed.

Keywords: topic modeling, hierarchical agglomerative grouping, ontology, general vocabulary, content authenticity

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 5, pp. 1185–1197 (Russian).

Введение

Анализ многообразия структур отношений в коллекциях документов на естественном языке (ЕЯ) применяется для добычи новых знаний при тематическом моделировании. Например, в качестве новых знаний при моделировании рассматривается наличие семантической связанности слов, лежащих в основе разбиения документов на темы, интерпретируемые векторные представления слов для решения проблемы синонимии и полисемии.

Чаще всего при тематическом моделировании задают описание документов некоторым дискретным распределением вероятностей на множестве тем, а темы — дискретным распределением вероятностей на множестве термов. Решение задач при таком описании рассматривается через построение вероятностной тематической модели (Probabilistic Topic Model, PTM), в которой любой текст отображается в вектор вероятностей тем. Каждый элемент вектора показывает долю соответствующей темы в тексте, а семантика темы описывается частотным словарем слов ЕЯ.

Методы машинного обучения на основе векторных моделей с успехом применяются для извлечения знаний из текстовых документов в различных предметных областях. При векторном представлении в виде частотных распределений слов в контекстах считается, что семантически близкие слова имеют близкие векторы. В качестве варианта для тестирования моделей рассматривается ручное формирование наборов пар слов с экспертными оценками близости.

Суть методики сравнительного тематического анализа (СТА), описанной в [Краснов, Дементов, Шварцман, 2020], состоит в том, чтобы для группировки использовать мягкую кластеризацию текстов с обучением тематической модели на общем словаре и объединенной коллекции документов. Результативность сравнения двух коллекций предложено измерять с помощью метрики «коэффициент контентной аутентичности». При вычислении этого коэффициента используется сумма модулей отклонений от равномерного распределения тематик, деленная на число тематик. Максимальное значение коэффициента равно единице, когда каждая из выявленных тематик относится только к одной из сравниваемых коллекций.

При вероятностном тематическом моделировании [Vorontsov, Potapenko, 2015] возможности для интерпретации экспертами содержания тем не ограничиваются значениями семантической близости слов. Оптимизация разбиения документов на темы достигается за счет использования дополнительных критериев-регуляризаторов. Для интерпретации результатов разбиения анализируется список наиболее вероятных слов, характерных для каждой темы. Как правило, практический интерес представляют содержимое списка из предметных тем и практически полное игнорирование списка из фоновых тем.

На интерпретацию смысла слов из словарей по разным предметным областям требуется привлечение для анализа специализированных ресурсов – онтологий и введение определенных ограничений. Построение онтологий соответствующих предметных областей [Дорофеюк, Покровская, Чернявский, 2018] является распространенным способом формализации разных областей знаний. Как правило, в состав онтологий входят объекты, понятия, атрибуты и отношения. Процесс формализации начинается с формирования понятий, т. е. выявления структуры совокупности (классов) объектов предметной области. Существуют условные разделения на виды онтологий. К числу наиболее распространенных из них относятся:

- словарь — список однозначных терминов;
- глоссарий — словарь многозначных терминов с указанием их значений;
- тезаурус — глоссарий с заданной системой семантических связей.

В информационных системах онтология [Городецкий, Тушканова, 2018], представленная в виде метамодели данных и знаний, рассматривается как посредник для общения пользователей и программ семантических приложений. Семантическая интерпретация результатов вычислений на данных ЕЯ для пользователя предусматривает их трансформацию в машинно-понимаемую форму и обратно.

Распределение данных в задачах тематического моделирования в общем-то считается неизвестным. Наличие выбросов может сильно повлиять на результаты кластерного анализа. В целях снижения влияния выбросов предлагается использовать достижение робастной устойчивости через группировку данных по результатам нелинейных преобразований признаков [Ignatiev, 2021].

К числу одного из отличий методов семантической кластеризации от традиционных [Пархоменко, Григорьев, Астраханцев, 2017] относится использование в качестве признаков документа наборов связанных по смыслу слов. Под паттерном в данной работе понимается набор сущностей ЕЯ для вычисления обобщенных оценок объектов (документов) в качестве значений латентного признака для кластерного анализа. При формировании наборов используются значения функции принадлежности объектов классов к непересекающимся интервалам по каждому признаку (частоте встречаемости терма в документах). Выбор оптимальных границ интервалов производится по специальному критерию [Згуральская, 2018].

Одной из целей введения паттернов является разработка новых методов тематического моделирования для исследования семантической связанности слов при представлении текстовых данных через модель «мешок слов». Заранее не известно, какое число тем задать модели, так как это число является свободным параметром [Краснов, Дементов, Шварцман, 2020]. Поскольку решаемые задачи являются некорректными, количество тем, количество паттернов и их состав определяются на основе дополнительных условий. Условия устанавливают: вхождение слов в общий словарь, количество слов в общем словаре, разбиение документов на темы.

Описание документов через латентные признаки позволяет:

- уменьшить размерность признакового пространства;
- реализовать обучение тематической модели на общем словаре коллекций из двух классов документов;
- использовать традиционные метрические алгоритмы для кластерного анализа документов;
- обосновать выбор количества тем по результатам кластерного анализа;
- определять наборы семантически связанных слов по темам.

Коллекция документов для машинного обучения при разбиении ее на два непересекающихся класса может быть представлена:

- объединением двух непересекающихся подмножеств, каждое из которых содержит документы из одной либо нескольких предметных областей;
- наборами статей из двух периодических изданий, каждое из которых представляет отдельный класс.

Кластеризация документов обычно сводится к кластеризации их векторных представлений и в общем случае является безотносительной к самим документам. При анализе структуры описаний документов метрическими алгоритмами приходится решать проблему проклятия размерности. Требуется вводить различные ограничения для снижения размерности признакового пространства при поиске семантической связанности слов в документах, разбитых на темы.

Существует много способов связи (группировки) документов при поиске схожей тематики. Для установления близости «схожести» рассматривают отношения по таким показателям, как ссылки, гиперссылки, общие авторы, цитирование, совместное упоминание и т. д. Разбиение на классы множества документов расширяет возможности использования различных форм логических закономерностей при исследовании структуры отношений на этом множестве.

Метрические алгоритмы в данной работе применяются для разбиения множества документов на непересекающиеся подмножества (темы) по определяемому набору латентных признаков. Считается, что объединение паттернов слов для формирования латентных признаков представляет общий словарь коллекции из двух непересекающихся классов документов. Наличие классификации позволяет определять оптимальное количество тем и аутентичность (похожесть) документов в них по словам из общего словаря.

Процедура формирования общего словаря определяет порядок отбора информативных наборов признаков (слов) по нескольким критериям [Петровский, Лобанов, 2014]. Основными этапами реализации процедуры являются:

- вычисление устойчивости признаков (слов) и формирование упорядоченной последовательности на их основе;
- выбор паттернов из упорядоченной последовательности по правилам иерархической агломеративной группировки.

Постановка задачи

Дана выборка объектов $E_0 = \{S_1, \ldots, S_m\}$, разделенная на два непересекающихся подмножества (класса) K_1, K_2. Объекты E_0 описываются набором признаков $X(n) = (x_1, \ldots, x_n)$, значения которых представляют частоты встречаемости слов в документах. Определены следующие процедуры:

- для вычисления набора латентных признаков $Y(t) = (y_1, \ldots, y_t)$ по непересекающимся подмножествам $X(n_1), \ldots, X(n_t)$, $t \geqslant 1$, $\forall c \in \{1, \ldots, t\}$, $n_c > 1$, $n_1 + \ldots + n_t \leqslant n$;
- для разбиения объектов E_0 на заданное число непересекающихся по набору $Y(\sigma) \subset Y(t)$ (групп) $G_1, \ldots, G_p$.

Требуется:

- оценить качество разбиения $F(p, Y(\sigma))$ на p групп по набору латентных признаков $Y(\sigma) \subset \subset Y(t)$;
- выделить количество тем документов по множеству групп $G_1, \ldots, G_p$;
- определить последовательности из слов по каждой теме.

Для решения поставленной задачи применяется кластеризация сырых признаков (частотных представлений слов) и объектов (документов). Процесс формирования латентного признакового пространства и отбора множества слов в общий словарь для двух классов документов рассчитан на использование нескольких критериев.

Формирование наборов латентных признаков связано:

- с нелинейным преобразованием значений сырых признаков в унифицированное представление в {1, 2};
- с вычислением значений устойчивости по каждому признаку;

- с формированием непересекающихся паттернов сырых признаков по упорядоченной последовательности значений их устойчивости;
- с вычислением латентных признаков по каждому паттерну.

Фундаментальными свойствами рассматриваемых ниже критериев разбиения значений признаков на интервалы являются инвариантность их результатов к масштабам измерений и малая вариабельность устойчивости на разных выборках из генеральной совокупности. Доказательство наличия таких свойств приводится в [Игнатьев, Рахимова, Лолаев, 2021]. Результаты разбиения на интервалы используются для преобразования значений признаков к градациям в номинальной шкале и вычисления меры компактности по латентным признакам.

Определяются условия объединения паттернов в общий словарь и формирования по ним набора латентных признаков для разбиения документов на группы. Определение количества тем документов рассматривается как процесс объединения групп по заданному отношению.

Разбиение значений признаков на интервалы

Для анализа многообразия отношений значений количественных (сырых и латентных) признаков на числовой оси и вычисления функции принадлежности объектов к классам предлагается использовать два критерия для поиска оптимальных границ непересекающихся интервалов. Поиску экстремумов этих критериев предшествует упорядочение значений признаков по неубыванию.

Пусть для значений признака $x_c \in X(n)$ в описании объектов $K_1 \cup K_2$ построена упорядоченная по неубыванию последовательность

$$r_1, \ldots, r_j, \ldots, r_m. \tag{1}$$

В качестве границ двух непересекающихся интервалов $[\pi_1; \pi_2]$, $(\pi_2; \pi_3]$, определяемых по (1), используются $\pi_1 = r_1$, $\pi_2 = r_j$, $1 < j < m$, $\pi_3 = r_m$. Интервалы $[\pi_1; \pi_2]$ и $(\pi_2; \pi_3]$ идентифицируются соответственно как первый и второй. Вес признака у объектов классов по (1) вычисляется как максимум произведения внутриклассового сходства и межклассового различия по критерию из [Игнатьев, Рахимова, Лолаев, 2021]

$$\left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} \left(u_i^d - 1\right) u_i^d}{\sum\limits_{i=1}^{2} |K_i| \left(|K_i| - 1\right)} \right) \left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} u_i^d \left(|K_{3-i}| - u_{3-i}^d\right)}{2|K_1| \cdot |K_2|} \right) \to \max_{\pi_1 < \pi_2 < \pi_3}, \tag{2}$$

где u_i^d $\left(u_{3-i}^d\right)$ — количество значений признака x_c у объектов из класса K_i (K_{3-i}) в d-м интервале. Множество допустимых значений критерия (2) принадлежит (0; 1] и используется для оценки объектов классов на числовой оси. Если в каждом интервале содержатся все значения признака объектов из одного класса, то его вес равен 1.

Граница между классами (порог) для количественного (сырого или латентного) признака вычисляется как

$$\Gamma = \frac{\pi_2 + b}{2}, \tag{3}$$

где b — ближайшее к π_2 значение из интервала $(\pi_2; \pi_3]$, определяемого по (2). При вычислении порога по (3) не делается никаких предположений о природе среды данных. Значение (2) интерпретируется как мера компактности объектов выборки из двух классов на числовой оси.

В данной работе эта мера применяется для оценки паттернов из слов по значениям латентных признаков, формируемым по частоте встречаемости слов в документах.

Определим ограничения на использование альтернативного (2) критерия, число непересекающихся интервалов для реализации которого изначально неизвестно. Пусть q ($2 \leqslant q \leqslant \min\limits_{i=1,2} |K_i|$) — число отличных друг от друга значений признака по (1) и $a_{j1}, \ldots, a_{jq}$ — их количество в классе K_j, $j = 1, 2$. В случае равенства

$$\frac{a_{11}}{a_{21}} = \ldots = \frac{a_{1q}}{a_{2q}} = \frac{|K_1|}{|K_2|} \tag{4}$$

не существует интервалов, в границах которых частота встречаемости значений признака у объектов из класса K_t будет больше, чем частота встречаемости у объектов из класса K_{3-t}, $t = 1, 2$.

При отсутствии ограничения (4) для разбиения (1) на множество из p_c ($p_c \geqslant 2$) непересекающихся интервалов $\left\{[r_u; r_v]^i\right\}$, $1 \leqslant u, u \leqslant v \leqslant h$, $i = 1, \ldots, p_c$, предлагается использовать критерий из [Ignatiev, 2021]. Значения в границах интервала $[r_u; r_v]^i$ при анализе данных рассматривается как градация номинального признака. Считается, что множество чисел, идентифицирующих p_c градаций номинального признака, всегда можно взаимно однозначно отобразить в множество $\{1, \ldots, p_c\}$.

Пусть $d_{tc}(u, v)$, $d_{3-t,c}(u, v)$ — количество представителей классов K_t, K_{3-t} в интервале $[r_u; r_v]^i$, $i \in \{1, \ldots, p_c\}$. Для рекурсивной процедуры выбора значений r_u, r_v используется критерий

$$\left| \frac{d_{tc}(u, v)}{|K_t|} - \frac{d_{3-t,c}(u, v)}{|K_{3-t}|} \right| \to \max . \tag{5}$$

Границы первого интервала $[r_u; r_v]^1$ на последовательности (1) вычисляются по максимуму критерия (5). Аналогичным образом определяются границы для $[r_u; r_v]^\tau$, $\tau > 1$, на значениях (1), не вошедших в $[r_u; r_v]^1, \ldots, [r_u; r_v]^{\tau-1}$. Критерием останова процедуры служит покрытие всех значений (1) непересекающимися интервалами.

Нелинейные преобразования признаков и вычисление обобщенных оценок объектов

Суть нелинейных преобразований признаков сводится к замене их исходных значений на значения функции принадлежности объектов к классам. В целях унификации обозначений вместо $d_{tc}(u, v)$, $t = 1, 2$, для интервала $[r_u; r_v]^\mu$ по $x_c \in X(n)$ будем использовать $d_{tc}(\mu)$. Значение функции принадлежности $f_c(\mu)$ к классу K_1 по интервалу $[r_u; r_v]^\mu$ (градации $\mu \in \{1, \ldots, p_c\}$) вычисляется как

$$f_c(\gamma) = \frac{\frac{d_{1c}(\gamma)}{T_{1c}}}{\frac{d_{1c}(\gamma)}{T_{1c}} + \frac{d_{2c}(\gamma)}{T_{2c}}}. \tag{6}$$

Очевидно, что отношение порядка между градациями номинальных признаков не существует. Замена градаций признака на значения функции принадлежности объектов к классу K_1 по (6) при $p_c > 2$ рассматривается как нелинейное преобразование, при котором определен порядок следования на числовой оси.

Множеству из p_c, $2 \leqslant p_c < m$, допустимых значений (градаций) признака $x_c \in X(n)$ можно поставить в соответствие числа $1, 2, \ldots, p_c$. При вычислении функции принадлежности $f_c(\mu)$ к классу K_1 по градации $\mu \in \{1, 2, \ldots, p_c\}$ в качестве $d_{1c}(\mu)$ ($d_{2c}(\mu)$) используется число объектов класса K_1 (K_2) со значением μ.

Обозначим через Z, $Z \subset X(n)$, множество признаков, для которых выполнены условия (4), и через $D = \{i \mid x_i \in X(n) \setminus Z\}$ — множество индексов признаков, которые можно использовать для нелинейных преобразований.

Граница между объектами классов по (6) для $x_c \in X(n) \setminus Z$ определяется как

$$G_c = \frac{q_1 + q_2}{2}, \tag{7}$$

где $q_2 = \max\{f_c(\mu) \mid 0{,}5 - f_c(\mu) > 0,\ \mu = 1, \ldots, p_c\}$, $q_1 = \min\{f_c(\mu) \mid 1 - f_c(\mu) < 0{,}5,\ \mu = 1, \ldots, p_c\}$. Вес признака $x_c \in X(n) \setminus Z$ у объектов можно вычислить по (2) либо через градации из {1, 2} в номинальной шкале. При вычислении значения градации a_{ic}, $c \in D$ для объекта $S_i = \{x_{iu}\}_{u \in D}$ с использованием (7) используется проверка условия $x_{ic} \in [r_u; r_v]^{\mu}$. Проверка условия необходима для выбора значений функции принадлежности $f_c(\mu)$ для вычисления a_{ic} как

$$a_{ic} = \begin{cases} 1, & f_c(\mu) < G_c, \\ 2, & f_c(\mu) > G_c. \end{cases}$$

Обозначим через g^j_{1c}, g^j_{2c} количество значений градации $j \in \{1, 2\}$ признака $x_c \in X(n) \setminus Z$ в описании объектов соответственно класса K_1 и K_2. Межклассовое различие по признаку x_c определяется как величина

$$\lambda_c = 1 - \frac{\sum\limits_{j=1}^{2} g^j_{1c} g^j_{2c}}{|K_1| \cdot |K_2|}. \tag{8}$$

Степень однородности (мера внутриклассового сходства) β_c значений градаций признака по классам K_1, K_2 вычисляется по формуле

$$\beta_c = \frac{\sum\limits_{j=1}^{2} g^j_{1c}\left(g^j_{1c} - 1\right) + g^j_{2c}\left(g^j_{2c} - 1\right)}{|K_1|\left(|K_1| - 1\right) + |K_2|\left(|K_2| - 1\right)}. \tag{9}$$

С помощью (8), (9) вес признака $x_c \in X(n) \setminus Z$ в номинальной шкале аналогично (2) определяется как произведение внутриклассового сходства и межклассового различия:

$$w_c = \beta_c \lambda_c. \tag{10}$$

Множество допустимых значений весов признаков, вычисляемых по (10), принадлежит интервалу (0; 1].

Для вычисления обобщенных оценок объектов [Дорофеюк, Покровская, Чернявский, 2018] на E_0 используются вклады градаций признаков. Вклад градации $j \in \{1, 2\}$ признака $x_c \in X(n) \setminus Z$ определяется как

$$\eta_c(j) = w_c \left(\frac{\alpha^1_{cj}}{|K_1|} - \frac{\alpha^2_{cj}}{|K_2|} \right), \tag{11}$$

где α^1_{cj}, α^2_{cj} — количество значений градации j признака x_c соответственно в классах K_1 и K_2, w_c — вес признака x_c по (10). Обобщенная оценка объекта $S_r \in E_0$ по описанию в номинальной шкале измерений $S_r = \{a_{ri}\}_{i \in D}$ на наборе $X(n) \setminus Z$ и вкладам (11) вычисляется как

$$R(S_r) = \sum_{i \in D} \eta_i(a_{ri}). \tag{12}$$

Базовым понятием для формирования информативных наборов признаков является устойчивость признака. Для вычисления устойчивости используются значения функции принадлежности. Пусть в описании объекта $S_r \in K_1 \cup K_2$ исходные значения признаков из $X(n) \setminus Z$ заменены на значения функции принадлежности $S_r = \{b_{ri}\}_{i \in D}$ по (6). Устойчивость признака $x_c \in X(n) \setminus Z$ вычисляется как

$$\varphi(c) = \frac{1}{m} \sum_{r=1}^{m} \begin{cases} b_{rc}, & b_{rc} > 0{,}5, \\ 1 - b_{rc}, & b_{rc} < 0{,}5. \end{cases} \tag{13}$$

Множество допустимых значений (13) принадлежат (0,5; 1]. Устойчивость $\varphi(c) = 1$, если по границе (7) объекты без ошибок разделяются на классы K_1 и K_2. Упорядоченность слов по (13) в общем-то не зависит от размерности пространства для описания коллекции документов.

Вычисление обобщенных оценок по (12) на разных наборах сырых признаков используется при выборе латентного пространства для описания документов. Как правило, методы формирования наборов определяются исходя из целей решаемых задач.

Вычисление коэффициента контентной аутентичности

Смысл вычисления коэффициента контентной аутентичности заключается в поиске оптимального числа кластеров, определяющих разбиение документов на темы в латентном признаковом пространстве. Для формирования латентного пространства и общего словаря на его основе предлагается:

- упорядочить признаки с индексами из D по неубыванию их значений устойчивости (13) как
$$\varphi(\varepsilon_1), \ldots, \varphi(\varepsilon_j), \ldots, \varphi(\varepsilon_{dim}), \quad \varepsilon_j \in D, \quad dim = |D|; \tag{14}$$
- сформировать множество латентных признаков $Y(t) = (y_1, \ldots, y_t)$ по (14);
- выделить $Y(\sigma) \subset Y(t)$, $\sigma \leqslant t$, и множество связанных по D с $Y(\sigma)$ сырых признаков в качестве общего словаря.

Количество слов для общего словаря из (14) является, в общем-то, свободным параметром. Этот параметр может задаваться по эвристическим соображениям либо по результатам иерархической агломеративной группировки при формировании набора латентных признаков.

Обозначим через $TYPLAM$ множество индексов сырых признаков, включенных в состав группы алгоритмом иерархической группировки, $lugat$ — ограничение на количество слов в общем словаре, $guruh$ — количество латентных признаков. Реализация алгоритма по шагам будет следующей.

Шаг 1. $j = 0$. $guruh = 0$.

Шаг 2. Вычислить $j = j + 1$. $crit = 10$. $u = \varepsilon_j$. $TYPLAM = \{u\}$. $guruh = guruh + 1$.

Цикл по $t \in \{1, \ldots, m\}$ $R(S_t) = \eta_u(a_{tu})$. Конец **цикла**.

Шаг 3. $u = \varepsilon_{j+1}$. **Цикл** по $t \in \{1, \ldots, m\}$ $b_t = R(S_t) + \eta_u(a_{tu})$. Конец **цикла**.

$M_1 = \sum\limits_{S_t \in K_1} b_t$. $M_2 = \sum\limits_{S_t \in K_2} b_t$. $M_1 = \frac{M_1}{|K_1|}$. $M_2 = \frac{M_2}{|K_2|}$. $\theta = 0$. $\gamma = 0$.

Цикл по $t \in \{1, \ldots, m\}$. Если $S_t \in K_1$, то $\theta = \theta + |b_t - M_1|$, $\gamma = \gamma + |b_t - M_2|$. Иначе $\theta = \theta + |b_t - M_2|$, $\gamma = \gamma + |b_t - M_1|$. Конец **цикла**.

Шаг 4. Если $\frac{\theta}{\gamma} < crit$, то $crit = \frac{\theta}{\gamma}$, $TYPLAM = TYPLAM \cup \{u\}$, $j = j + 1$, идти 3.

Шаг 5. Вывод $\{R(S_t)\}_{t \in \{1, \ldots, m\}}$, $TYPLAM$.

Шаг 6. Если $j < lugat$, то идти 2; Иначе вывод $guruh$.

Шаг 7. Конец.

Множество значений $\{R(S_t)\}_{t\in\{1,\ldots,m\}}$, полученное на шаге 4 алгоритма, формируют описания объектов K_1, K_2 по набору $Y(guruh) = (y_1, \ldots, y_{guruh})$ в латентном признаковом пространстве. Каждому $y_i \in Y(guruh)$ соответствует набор $X(n_1) \subset X(n) \setminus Z$. Рекомендуемым ограничением на мощность словаря является $lugat \leqslant 200$. Анализ процесса формирования набора латентных признаков по (14) можно производить по значениям (2). Как правило, при включении новых признаков в набор наблюдается тенденция снижения их значений по (2).

Оценивать набор латентных признаков предлагается по результатам разбиения множества объектов (документов) на заданное число непересекающихся групп. Пусть $G_1, \ldots, G_h$, $i = 1, \ldots, h$, $h \geqslant 2$, — разбиение на непересекающиеся группы по набору латентных признаков $Y(guruh)$. Для каждой группы G_i определим значение функции принадлежности объектов к классу K_1 по G_i как $\lambda_i(K_1) = \frac{d_{i1}}{|G_i|}$, где d_{i1} — число объектов класса K_1 в G_i. Контентная аутентичность документов из E_0 при разбиении их на h групп будет вычисляться как

$$F(h, Y(guruh)) = \frac{1}{m}\sum_{j=1}^{h}\begin{cases}|G_j|\lambda_j(K_1), & \lambda_j(K_1) > 0{,}5;\\ |G_j|(1-\lambda_j(K_1)), & \lambda_j(K_1) < 0{,}5.\end{cases} \tag{15}$$

Целесообразность использования (15) для вычисления контентной аутентичности по общему словарю можно проверить следующим образом. Нужно сформировать два набора латентных признаков по правилам агломеративной группировки на (14) при выборе на первом шаге $j = 0$ и $j > 0$. Максимальное значение (15) при $j = 0$ будет указывать на корректность использования упорядоченности признаков по значениям их устойчивости. В качестве первого шага для определения оптимального количества тем документов предлагается использовать разбиение их на разное число непересекающихся групп с последующим сравнением результатов разбиения по (15).

Будем считать, что известно разбиение объектов на группы $G_1, \ldots, G_h$ и представители какого класса, K_1 или K_2, доминируют в каждой из них. Обозначим через $T = (M_1, \ldots, M_h)$ множество центров групп (объектов), представленное как объединение двух непересекающихся подмножеств — Q_1 и Q_2, $T = Q_1 \cup Q_2$, с доминированием в группах представителей соответственно класса K_1 и K_2. Для определения оптимального числа подмножеств центров групп (далее — тем документов) предлагается использовать отношение связанности объектов, описанное в [Ignatyev, 2018] для задачи распознавания с непересекающимися классами.

Пусть $T_c \subset T$ — множество граничных объектов на T,

$$T_c = \left\{M_i \in Q_u \mid \rho(M, M_i) = \min_{M_j\in Q_u,\, M\in Q_{3-u}} \rho(M, M_j)\right\},$$
$$r_i = \min_{M_i\in Q_u,\, M\in Q_{3-u}} \rho(M, M_i), \quad r_j = \min_{M_j\in Q_u,\, M\in Q_{3-u}} \rho(M, M_j)$$

по евклидовой метрике. Объекты M_i, $M_j \in Q_u$ связаны по $M \in Q_u \cap T_c$, если $\rho(M_i, M) < r_i$ и $\rho(M_j, M) < r_j$. Отношение связанности гарантирует единственность разбиения множества T на непересекающиеся группы $\Psi_1, \ldots, \Psi_\alpha$, $\alpha \leqslant h$.

Отношение связанности по множеству центров групп T является одним из способов разбиения документов на темы по алгоритму из [Ignatyev, 2018]. Для отбора наиболее значимых слов по каждой теме предлагается ранжировать их по частоте встречаемости. Разбиение на темы позволяет выделять наборы слов, характерных для определенного класса документов, вероятность использования которых в другом классе относительно невысока.

Вычислительный эксперимент

Вычислительный эксперимент проводился на текстовых данных из 12 предметных областей, представленных множеством авторефератов ВАК Республики Узбекистан [Тулиев, Лолаев, 2021]. Для описания использовалось 457 слов из объединения 12 словарей по предметным областям, составленным экспертным путем. Востребованность предметных словарей была в использовании их для тестирования результатов эксперимента при формировании общих словарей по коллекциям из 1530 документов.

Цель эксперимента заключается в демонстрации возможностей методики тематического моделирования на текстовых данных, представленных как «мешок слов». Результаты эксперимента представлены на коллекции авторефератов по предметам «физика» и «математика» (класс K_1) и остальным 10 предметам (класс K_2).

В [Тулиев, Лолаев, 2021] было показано, что из 12 предметов в коллекции авторефератов семантически наиболее связанными являются научные работы по предметам «физика» и «математика». Причинно-следственное исследование такой связи проводится с использованием описанных в данной работе методов.

Для построения общего словаря на упорядоченной по (13) последовательности из 457 признаков использовался описанный выше алгоритм иерархической агломеративной группировки. При ограничении $lugat = 180$ и $j = 0$ на первом шаге алгоритма в общий словарь по классам K_1 и K_2 включены 178 слов, которые были разделены на 30 непересекающихся паттернов.

С помощью коэффициента контентной аутентичности (15) приведем аргументы в пользу выбора латентного пространства с использованием упорядочения по устойчивости (13) сырых признаков, начиная с $j = 0$. Смысл аргументации заключается в доказательстве снижения показателя (15) при выборе $j > 0$ на первом шаге. При $j = 160$ при ограничении $lugat = 180$ на число слов в общем словаре по алгоритму иерархической агломеративной группировки сформирован набор из 24 латентных признаков.

Наборы из 30 (по числу паттернов) и 24 латентных признаков использовались для разбиения выборки документов на группы алгоритмом k-means из библиотеки языка Python [Pedregosa et al., 2011]. Результаты вычисления коэффициента контентной аутентичности при разном числе групп объектов приводятся в табл. 1.

Таблица 1. Коэффициенты контентной аутентичности при разном числе групп

Количество групп	Коэффициент контентной аутентичности (15) по наборам признаков	
	30	24
2	0,96078	0,87647
3	0,96797	0,87647
4	0,96666	0,87647
5	0,96143	0,87647
11	0,96274	0,88104
33	0,97189	0,88627
38	0,97058	0,88693

Как видно из табл. 1, при разном числе групп значения контентной аутентичности (15) очень близки. Результаты вычисления контентной аутентичности подтверждают положительный эффект от выбора 30 латентных признаков относительно 24 соответственно при $j = 0$ и $j = 160$ на первом шаге иерархического агломеративного алгоритма.

Рассмотрим 33 центра групп, сформированных по набору из 30 латентных признаков, в 7 из которых доминируют представители K_1, в 26 — K_2. По отношению связанности объектов центры групп инициируют разделение документов на четыре области (темы). Документы с доминированием K_1 — на три и K_2 — на одну тему. Последовательности из 15 слов по каждой теме показаны в табл. 2.

Таблица 2. Последовательности слов по четырем темам

Класс K_1			Класс K_2
Тема 1	Тема 2	Тема 3	Тема 4
препарат	препарат	коэффициент	ресурс
генетика	генетика	динамика	обучение
концентрация	концентрация	спектр	техника
клетка	фермент	фото	энергия
оксид	клетка	ядро	механизм
микроб	микроб	энергия	спектр
синтез	оксид	электрон	температура
атом	синтез	температура	финансы
фермент	атом	модель	банк
запас	бактерия	механизм	функция
бактерия	белок	проводник	национальный
плазма	статистика	атом	динамика
область	динамика	эксперимент	коэффициент
производная	комплекс	фермент	электрон
эксперимент	фонд	кристалл	кислота

Последовательность слов по каждой теме (см. табл. 2) получена из объединения словарей по 12 предметам и в определенной степени отражает специфику научных работ, разделенных на два непересекающихся класса. Для решения проблемы лексической многозначности ЕЯ необходимо использовать предметно-ориентированные онтологии.

В качестве примера рассмотрим автореферат диссертационной работы «Процессы фрагментации ядер кислорода во взаимодействиях с протонами при 3,25 А ГэВ/с и механизмы образования протонов в π^-,p,α,C(C)- и p(^{16}O,^{20}NE)-соударениях при 3–300 ГэВ». При принятии решения о теме работы использовались:

- близость описания документа по 30 латентным признакам к центру 33-й группы;
- доминирование в группе представителей из класса K_1 (физика и математика) и принадлежность группы к теме 3 по отношению связанности объектов.

Заключение

Предложена методика формирования общих словарей и наборов семантически связанных последовательностей слов по темам с использованием разбиения коллекций документов на два непересекающихся класса. При реализации методики последовательно применялись группировка слов для формирования латентного признакового пространства и группировка документов в этом пространстве. Дано обоснование выбора числа слов обобщенного словаря и числа тем для представления документов.

Список литературы (References)

Городецкий В. И., Тушканова О. Н. Семантические технологии для семантических приложений. Часть 1. Базовые компоненты семантических технологий // Искусственный интеллект и принятие решений. — 2018. — № 4. — С. 61–71.

Gorodetskii V. I., Tushkanova O. N. Semanticheskie tekhnologii dlya semanticheskikh prilozhenii. Chast' 1. Bazovye komponenty semanticheskikh tekhnologii [Semantic technologies for semantic applications. Part 1. Basic components of semantic technologies] // Artificial intelligence and decision making. — 2018. — No. 4. — P. 61–71 (in Russian).

Дорофеюк А. А., Покровская И. В., Чернявский А. Л. Структуризация объектов нечисловой природы // Информационные технологии и вычислительные системы. — 2018. — № 1. — С. 16–21.

Dorofeyuk A. A., Pokrovskaya I. V., Chernyavskii A. L. Strukturizatsiya ob"ektov nechislovoi prirody [Data structuring for nonnumeric objects] // Information Technologies and Computing Systems. — 2018. — No. 1. — P. 16–21 (in Russian).

Згуральская Е. Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей // Известия Самарского научного центра Российской академии наук. — 2018. — Т. 20, № 4 (3). — С. 451–455.

Zgural'skaya E. N. Ustoichivost' razbieniya dannykh na intervaly v zadachakh raspoznavaniya i poisk skrytykh zakonomernostei [Sustainability of dividing data in intervals in the problems of recognition and searching for hidden laws] // Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk. — 2018. — Vol. 20, No. 4 (3). — P. 451–455 (in Russian).

Игнатьев Н. А., Рахимова М. А., Лолаев М. Я. Особенности отбора информативных наборов признаков на данных с пропусками // Проблемы вычислительной и прикладной математики. — 2021. — № 6/1 (37). — С. 113–122.

Ignat'ev N. A., Rakhimova M. A., Lolaev M. Ya. Osobennosti otbora informativnykh naborov priznakov na dannykh s propuskami [Specifications of selection informative sets of features on data with missing] // Problemy vychislitel'noi i prikladnoi matematiki. — 2021. — No. 6/1 (37). — P. 113–122 (in Russian).

Краснов Ф. В., Дементов А. В., Шварцман М. Е. Использование тематических моделей для парного сравнения коллекций научных статей // Информ. и ее примен. — 2020. — Т. 14, № 3. — С. 129–135.

Krasnov F. V., Dimentov A. V., Shvartsman M. E. Ispol'zovanie tematicheskikh modelei dlya parnogo sravneniya kollektsii nauchnykh statei [Using Topic Models for Pairwise Comparison of Collections of Scientific Papers] // Informatics and Applications. — 2020. — Vol. 14, No. 3. — P. 129–135 (in Russian).

Пархоменко П. А., Григорьев А. А., Астраханцев Н. А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. — 2017. — Т. 29, вып. 2. — С. 161–200.

Parkhomenko P. A., Grigor'ev A. A., Astrakhantsev N. A. Obzor i eksperimental'noe sravnenie metodov klasterizatsii tekstov [A survey and an experimental comparison of methods for text clustering: application to scientific articles] // Trudy ISP RAN. — 2017. — Vol. 29, No. 2. — P. 161–200 (in Russian).

Петровский А. Б., Лобанов В. Н. Многокритериальный выбор в пространстве признаков большой размерности: мультиметодная технология ПАКС-М // Искусственный интеллект и принятие решений. — 2014. — № 3. — С. 92–104.

Petrovsky A. B., Lobanov V. N. Mnogokriterial'nyi vybor v prostranstve priznakov bol'shoi razmernosti: mul'timetodnaya tekhnologiya PAKS-M [Multiple Criteria Choice in the Attribute Space of Large Dimension: Multimethod Technology PAKS-M] // Artificial Intelligence and Decision Making. — 2014. — No. 3. — P. 92–104 (in Russian).

Тулиев У. Ю., Лолаев М. Я. О формировании пространства для описания тематических документов // Вестник РГГУ. Сер. Информатика. Информационная безопасность. Математика. — 2021. — № 1. — С. 35–50. — DOI: 10.28995/2686-679X-2021-1-35-50

Tuliev U. Yu., Lolaev M. Ya. O formirovanii prostranstva dlya opisaniya tematicheskikh dokumentov [On the formation of space for the description of thematic documents] // Vestnik RGGU. Ser. Informatika. Informatsionnaya bezopasnost'. Matematika. — 2021. — No. 1. — P. 35–50. — DOI: 10.28995/2686-679X-2021-1-35-50 (in Russian).

Ignatiev N. A. On nonlinear transformations of features based on the functions of objects belonging to classes // Pattern Recognition and Image Analysis. — 2021. — Vol. 31, No. 2. — P. 197–204.

Ignatyev N. A. Structure choice for relations between objects in metric classification algorithms // Pattern Recognition and Image Analysis. — 2018. — Vol. 28, No. 4. — P. 590–597.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: machine learning in Python // Journal of machine learning research. — 2011. — Vol. 12. — P. 2825–2830.

Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // Machine Learning Journal. — 2015. — Vol. 101. — P. 303–323.