DOI: 10.36535/0203-6460-2021-01-3

Что имеет большее значение? Сравнение влияния концептуальных и документальных отношений в тематических моделях*

Сильвия ТЕРРАНЬИ (Silvia TERRAGNI),

Элизабетта ФЕРСИНИ (Elisabetta FERSINI),

Энца МЕССИНА (Enza MESSINA)
Миланский университет Бикокка,
г. Милан, Италия

Дебора НОЦЦА (Debora NOZZA)
Университет Боккони,
г. Милан, Италия

Тематические модели широко используются для обнаружения скрытых тем в коллекциях документов. В статье предполагается изучить роль двух разных типов реляционной информации, т.е. документальных отношений и концептуальных отношений. Хотя использование сети документов значительно улучшает когерентность темы, введение понятий и их отношений не оказывает влияния на результаты как качественно, так и количественно.

ВВЕДЕНИЕ

Тематические модели являются набором порождающих вероятностных моделей, предназначенных для обнаружения тематической информации (или тем) в неструктурированном массиве документов. Эти модели, включая известное латентное размещение Дирихле (Latent Dirichlet Allocation - LDA) [1], обычно рассматривают текст как уникальный источник информации и основаны на предположении, что тексты являются независимыми и одинаково распределенными. Тем не менее, в некоторых случаях реального мира документы часто характеризуются соответствующей реляционной структурой: научные статьи можно связать через библиографические ссылки, сетевые страницы могут представлять гиперссылки между собой, а пользователи в социальных сетях могут быть друзьями. Одним из первых подходов, подробно моделирующих отношения между документами, является Реляционная тематическая модель (Relational Topic Model - RTM) [2], основанная на предположении, что связанные документы вероятнее всего затрагивают одни и те же темы.

Традиционные тематические модели также предполагают, что тематическое распределение слова не зависит от других скрытых тем, принимая во внимание распределение темы в документе. Тем не менее, предшествующая работа обосновывает, что введение дополнительного знания об отношениях между словами улучшает когерентность обнаруженных тем [3, 4, 5]. Такой тип отношений широко рассматривается относительно понятия синоним, но это не всегда происходит в реальном сценарии из-за двусмысленности слов. Таким образом в соответствии с этим предположением важно принимать во внимание сам концепт, выходящий за рамки слова, наравне с самим словом, так как это позволит ассоциировать одну и ту же

* Перевод Terragni S., Nozza D., Fersini E., Messina E. Which matters most? Comparing the impact of concept and document relationships in topic models//Proceedings of the First Workshop on Insights from Negative Results in NLP. — 2020. —P. 32 – 40. —https://www.aclweb.org/anthology/2020/insights-1.5.pdf

тему со словами, которые действительно близки, но не являются синонимами. Например, представляется возможным осознать, что слово «engine», ассоциируемое с понятием «search engine», находится далеко от слова «motor», но близко к слову «information retrieval». Ряд работ изучает использование названного объекта в тематических моделях [6, 7, 8]. но ни одна из них не анализирует эту проблему на реляционных установках.

Вклад. В этой статье изучается роль двух типов реляционной информации: (1) концептуальные отношения между словами и названными объектами, полученные путем векторных представлений слов, и (2) отношения на уровне документа, извлеченные из сети документов. Влияние этих двух типов реляционной информации оценивается с помощью рассмотрения традиционных тематических моделей и введения двух новых тематических моделей с ограничением объектов. Исходный код можно посмотреть в следующей ссылке: https://github.com/MIND-Lab/EC-RTM.

СВЯЗАННЫЕ РАБОТЫ

Латентное размещение Дирихле (LDA) [1] – порождающая вероятностная модель, описывающая массив документов через набор тем *K*, полностью рассматриваемых как распределения слов в фиксированном словаре. Согласно размещению Дирихле, предполагается, что документ состоит из сбора тем. Слова образуются в соответствии с темой, обозначенной этим сбором. Латентное размещение Дирихле может быть расширено за счет рассмотрения различных типов реляционной информации.

Реляционные тематические модели на уровне слов уменьшают принятие независимости слов в документе или теме. Они могут грубо подразделяться на модели, кодирующие порядок слов [9-13] и синтаксические зависимости [14, 15]. а также модели, объединяющие семантические отношения или отношения знания предметной области [16,17, 4, 3]. Позже растущий интерес к векторным представлениям слов привел к объединению отношений, возникающих из векторных представлений слов [18-24].

Реляционные тематические модели на уровне документов предполагают, что два связанных документа вероятнее всего имеют схожие тематические распределения. Реляционная тематическая модель и ее расширения [25-29] основаны на LDA и моделируют каждую связь как бинарную переменную, принимая во внимание существование связи между парой документов. Другие подходы включают регуляризационные тематические модели [30, 31], которые дополняют целевую функцию модели проблемой регуляризации нейронной сети, полиноминальной регрессией Дирихле [32] и ее расширениями [33, 34], объединяющими связи путем их просмотра как атрибута для каждого документа. Перспективная парадигма использует нейронный вариационный вывод для логического вывода тем [35-37]. Нейронная реляционная тематическая модель (Neural Relational Topic Model - NRTM) [38] основана на пакетном вариационном автокодировщике (Stacked Variational AutoEncoder - SVAE) для вывода тем и предсказания связей с использованием многоуровневого перцептрона.

ТЕМАТИЧЕСКИЕ МОДЕЛИ С ОГРАНИЧЕНИЕМ ОБЪЕКТА

Мы предлагаем латентное размещение Дирихле с ограничением объекта (Entity Constrained Latent Dirichlet Allocation, EC - LDA) и реляционные тематические модели с ограничением объекта (Entity Constrained Relational Topic Models, EC - RTM), два класса моделей, нацеленных на объединение связей объект-объект и объект-слово в традиционных тематических моделях. Следуя авторам [3, 26] мы ограничиваем совместное распределение LDA и RTM через использование потенциальных функций, которые моделируют взаимосвязи объект-объект и/или объект-слово. Эти потенциальные функции могут быть вынесены за скобки из совместного распределения и последующие могут быть получены с использованием ослабленного сэмплирования по Гиббсу для логического вывода. Помимо EC - LDA, реляционные тематические модели с органичением объекта (EC - RTM) также предполагают, что два связанных документа вероятнее всего будут обсуждать одни и те же темы. О совместном распределении предложенных моделей см. в **ПРИЛОЖЕНИИ.** Для дальнейшего ознакомления с моделями с ограничением объекта адресуем читателя к авторам [3;26].

Определяем словарь E, содержащий уникальные названные объекты массива, и словарь W, содержащий уникальные слова. Получаем словарь Γ как объединение словарей слов и уникальных названных объектов. Отношения между ними обозначим как множество знания L и каждый предмет знания $l \in L$ объединяется функцией вероятности $f_l(z,u)$, которая представляет реально значимую оценку распределения скрытой темы z слова или реализацию названного объекта u.

Получаем знание L с использованием метода Skip-Gram [39]. С учетом тренировочного массива для векторного представления слов, содержащего большое, но конечное множество Λ, модель для векторного представления слов может быть выражена функцией отображения $C' : \Gamma \mapsto \mathbb{R}^t$. Для каждой реализации $u \in \Gamma$ определяем *ограниченное* множество L_u^m, содержащее слова и названные объекты, которые вероятнее всего объединены одними и теми же темами *и.* Множество L_u^m пределяется как:

$$L_u^m = \left\{ v \in \Gamma \middle| sim\left(C'(u), C(v)\right) > \in_m \right\} \qquad (1),$$

где *sim* – сходство по косинусу двух векторов, а $\in_m$ – заданный порог. Мы также определим *неограниченное* множество L_u^c, содержащее слова и названные объекты, которые вероятнее всего не объединены одними и теми же темами *и.* Множество L_u^c определяется как:

$$L_u^c = \left\{ v \in \Gamma \middle| sim\left(C'(u), C(v)\right) > \in_c \right\} \qquad (2),$$

где $\in_c$ – заданный порог.

Примером ограниченного множества названного объекта *"Artificial neural network"* может быть{*Artificial neuron, ANN, perceptron*}, содержащее названные объекты, которые вероятнее всего принадлежат одной и той же теме. Аналогично, примером неограниченного

множества названного объекта "*Artificial neural network*" может быть {*Olympic games, Athlete*}, которое отмечает названные объекты, относящиеся к спорту, а не к Машинному обучению.

Функция вероятности объект-объект (Entity-Entity, EE)

Мы выделяем функцию вероятности объект-объект, которая моделирует отношения между названными объектами. Пусть $N_{ze'}$ будет максимумом между 1 и тематическими подсчетами, т,е, числом встречаемости e', приписанным к теме z. Тогда функция $f_l(z,u)$ будет выглядеть следующим образом:

$$f_l(z,u)=\begin{cases}\sum\limits_{\substack{e'\in L_u^m\\ e'\in E}}\log N_{ze'}+\sum\limits_{\substack{e'\in L_u^c\\ e'\in E}}\log\dfrac{1}{N_{ze'}}, & \textit{если } u\in E\\ 0 & \textit{иначе}\end{cases} \quad (3)$$

Эта функция увеличивает вероятность того, что объект *и* будет приписан к тем же темам, что и объекты, принадлежащие L_u^m. Точно также функция вероятности уменьшает возможность того, что названные объекты будут взяты из одинаковых тем, что и объекты множества L_u^c.

Модели, которые могут кодировать функцию вероятности объект-объект (EE), будут относиться к латентному размещению Дирихле с органичением объекта (EC-LDA) и реляционным тематическим моделям с ограничением объекта (EC-RTM).

Функция вероятности объект - слово (Entity-Word, EW)

Допустим $N_{zw'}$ – максимум от 1 до тематических подсчетов, т,е, подсчетов слова w', принадлежащего теме z. Следующая функция вероятности касается отношений объектов и реализаций слов:

$$f_l(z,u)=\begin{cases}\sum\limits_{\substack{w'\in L_u^m\\ w'\in W}}\log N_{zw'}+\sum\limits_{\substack{w'\in L_w^c\\ w'\in W}}\log\dfrac{1}{N_{zw'}}, & \textit{если } u\in E\\ \sum\limits_{\substack{e'\in L_u^m\\ e'\in E}}\log N_{ze'}+\sum\limits_{\substack{e'\in L_u^c\\ e'\in E}}\log\dfrac{1}{N_{ze'}}, & \textit{если } u\in W\end{cases} \quad (4)$$

Функция вероятности моделирует следующие случаи:

• Если *и* – названный объект, тогда мы рассматриваем только слова, которые содержатся в ограниченном и неограниченном множествах *и*, т.е. L_u^m и L_u^c;

• Если *и* – слово, тогда мы рассматриваем только названные объекты, содержащиеся в ограниченном и неограниченном множествах *и*, т.е. L_u^m и L_u^c.

Эти модели, кодирующие отношения объект - слово, называются латентными распределениями Дирихле с ограничением объекта (EC-LDA) и реляционными тематическими моделями с ограничением объекта (EC-RTM).

ЭКСПЕРИМЕНТАЛЬНАЯ УСТАНОВКА

Массивы данных. Экспериментальное изучение было проведено на двух реляционных исходных массивах данных: (1) *Cora-ML* [40], сеть цитирований на множестве статей Машинного обучения [41] и (2) *WebKB* (www.cs.cmu.edu/~WebKB/ILP-data.html), сетевой массив данных, собранный из 4 разных университетов, в котором ссылки являются гиперссылками. Табл. 1 сообщает основную статистику данных массивов.

Предварительная обработка. Идентификация названных в тексте объектов, как правило, осуществляется через серию методов, касающихся задачи распознания названного объекта [42-44]. Признаются один раз названные объекты, следующий шаг – связать их с недвойственными понятиями, такими, как, например, ресурсы в Базе знания. Этот процесс известен как задача связывания названного объекта [45-49].

В данной статье используем средство DBPedia Spotlight [50] (доверие= 0,5 и поддержка =0,0), чтобы идентифицировать названные в тексте объекты и связать их с единицами DBPedia. Мы добавили приставку «NE/» к каждому идентифицированному объекту для отделения его от слов. К тексту применили общую предварительную обработку. Рассматривали только ограниченные множества, которые извлекались из Wikipedia2VEc [51]. Подробности о гиперпараметрах и предварительной обработке см. в **ПРИЛОЖЕНИИ**.

Сравниваемые модели. Сравнили предложенные модели (т.е., EC-LDA-EE, EC-LDA-EW, и EC-RTM-EE, EC-RTM-EW) с важными актуальными подходами, т. е. латентным размещением Дирихле [1], реляционной тематической моделью [2], пакетным вариационным автокодировщиком и нейронной реляционной тематической моделью [38].

Показатели. Мы используем методы *KL-U*, *KL-V* и *KL-B*, чтобы измерить семантическую важность и идентифицировать ненужные и маловажные темы [53]. Также путем вычисления разнообразия тем измеряем, насколько разными являются темы по отношению друг к другу [54]. Наконец, рассматриваем два показателя тематической когерентности, т. е. NPMI [55] и C_V [56], которые измеряют, как много топ-10 слов темы связаны друг с другом, Оценки подсчитываются с использованием средства Palmetto* и Википедии** в качестве библиографического фонда.

Таблица 1

Статистика исходных данных массивов

Массивы данных	#Документы	#Ссылки	Тип документа	Тип ссылки
Cora-ML	2 807	5 278	Название + реферат	Цитирование
WebKB	877	1 608	Сетевая страница	Гиперссылка

* http://www.github.com/dice-group/Palmetto

** Данные англоязычной Википедии по состоянию на 23 марта 2019 г.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Количественные результаты. Табл. 2 и 3 показывают действие моделей с точки зрения всех рассмотренных оценок относительно увеличивающегося числа тем в массивах данных*. Результаты отражают, что модели, рассматривающие реляционную информацию вообще, получают более высокую эффективность, чем нереляционные модели. Другими словами, введение понятия ограничения в моделях ECRTM-EE и EC-RTM-EW, кажется, не вносит значительных улучшений в отношение RTM. Это может быть мотивировано тем фактом, что множества с органичением, дополнительно включенные в модели EC-RTM, уже охвачены в распределении слово-тема, полученном RTM.

Разные поведения можно наблюдать для оценок C_v, для которых NRTM и SVAE получают значительно более высокую эффективность. Эта противоположная тенденция по отношению к другим оценкам тем, может быть объяснена фактом, что C_v поощряет присутствие редких слов, даже если они содержатся в ненужных темах, по утверждению авторов [56]**.

Качественные результаты. В табл. 4 отражены топ-10 слов для массива Cora-ML, связанные с примером темы «Genetic Programming» для моделей EC-RTM-EE, EC-RTM-EW, LDA, RTM, SVAE и NRTM. Чтобы анализировать, может ли аннотация названного объекта внести вклад в способность интерпретировать тему, сообщаем слова из LDA и RTM (относящиеся как к LDA* и RTM*), касающиеся массива Cora-ML, содержащего только слова. Как ожидается от количественных результатов, темы, извлеченные с помощью предложенных моделей, незначительно отличаются от RTM*, в дальнейшем демонстрируя гипотезу, что наложенные ограничения уже были охвачены оригинальной моделью.

Таблица 2

Выполнение на массиве Cora-ML с числом тем, равным 10, 30 и 50

	KL-U			*KL-V*			*KL-B*			TD			*NPMI*			Cv		
	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50
LDA	1,855	1,572	1,259	1,226	1,231	1,059	0,052	0,119	0,168	0,816	0,736	0,654	0,098	0,080	0,071	0,399	0,389	0,386
RTM	2,001	2,046	1,820	1,357	1,563	1,460	0,095	**0,207**	**0,283**	0,814	**0,747**	0,666	0,099	0,082	0,071	0,348	0,391	0,392
EC-LDA-EE	1,845	1,520	1,375	1,225	1,238	1,066	0,052	0,119	0,167	0,814	0,742	0,659	0,098	0,079	0,069	0,397	0,390	0,389
EC-LDA-EW	1,800	1,518	1,381	1,230	1,236	1,065	0,052	0,119	0,168	0,817	0,740	0,660	0,094	0,079	0,070	0,395	0,389	0,387
EC-RTM-EE	2,033	**2,082**	**1,849**	**1,362**	1,564	**1,472**	0,095	0,205	0,280	0,817	0,747	**0,675**	**0,099**	0,081	0,071	0,402	0,394	0,392
EC-RTM-EW	**2,079**	1,990	1,643	1,361	**1,565**	1,470	**0,096**	0,206	0,282	0,820	0,746	0,671	0,098	**0,082**	**0,072**	0,340	0,392	0,392
SVAE										**0,893**	0,694	0,577	-0,099	-0,095	-0,096	**0,456**	**0,456**	**0,453**
NRTM										0,857	0,525	0,381	-0,083	-0,082	-0,082	0,442	0,447	0,446

Таблица 3

Выполнение на массиве WebKB с числом тем, равным 10, 30 и 50

	KL-U			*KL-V*			*KL-B*			TD			*NPMI*			Cv		
	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50
LDA	1,695	1,256	1,130	1,054	0,943	0,775	0,069	0,142	0,199	0,761	0,617	0,538	0,039	0,040	0,030	0,378	0,379	0,379
RTM	**1,986**	1,795	1,430	**1,202**	1,239	1,109	**0,119**	0,225	**0,303**	0,760	0,608	0,532	0,043	0,043	0,036	0,377	0,380	0,380
EC-LDA-EE	1,643	1,289	1,061	1,055	0,948	0,780	0,069	0,143	0,200	0,769	0,623	0,542	0,043	0,041	0,033	0,379	0,380	0,381
EC-LDA-EW	1,736	1,345	1,075	1,062	0,981	0,784	0,069	0,138	0,198	0,764	**0,651**	**0,547**	0,042	0,038	0,033	0,376	0,381	0,382
EC-RTM-EE	1,867	**1,944**	1,468	1,199	1,246	1,119	0,118	**0,226**	0,303	0,760	0,612	0,536	**0,048**	**0,043**	**0,039**	0,377	0,382	0,381
EC-RTM-EW	1,979	1,786	**1,646**	1,199	**1,294**	**1,127**	0,117	0,217	0,302	0,759	0,639	0,543	0,045	0,042	0,036	0,377	0,382	0,384
SVAE										**0,829**	0,563	0,454	-0,116	-0,110	-0,112	**0,460**	0,450	0,452
NRTM										0,734	0,360	0,283	-0,114	-0,117	-0,119	0,454	**0,455**	**0,458**

Таблица 4

Тема «Genetic Programming» для массива Cora-ML

Модели	**Топ-10 слов**
LDA*	problem genetic algorithms problems programming search optimization fitness population space
RTM*	genetic control programming fitness reinforcement population algorithms paper environment behavior
EC-RTM-EE	NE/Genetic_programming programs NE/Genetic_algorithm population fitness genetic evolutionary program NE/Evolution strategies
EC-RTM-EW	NE/Genetic-programming NE/Genetic-algorithm population fitness genetic evolutionary NE/Evolution encoding operator operators
SVAE	koza NE/Multidisciplinary-design-Optimization splice bitsback NE/Genetic_programming fitness orientation NE/Ploidy NE/Exon coded
NRTM	genetic reactive NE/Genetic_programming NE/Case casebased neuroevolution ssa NE/Genetic_algorithm coevolutionary problemsolving

*Вычисление показателей KL не практично для SVAE и NRTM, поскольку они не моделируют распределения – слово-и документ-тема.

** https://bit.ly/3jApSAC

Качественные рассмотрения можно сделать относительно изучения нового моделирования документов типа объект-уровень. Хотя это представление приведет к темам, содержащим явные понятия (например, «NE/ Genetic Programming»), темы, полученные RTM*, кажется, должны одинаково интерпретироваться, поскольку они могут идентифицировать названные объекты в форме отдельных слов (например, «genetic», «programming», «algorithm»). Более того, различие в представлении становится очевидным только тогда, когда названные объекты содержат два и более слов (например, «NE/ Evolution» и «evolution» эквивалентны). Польза применения методов NEEL для распознавания названных объектов в темах может пригодиться для автоматического обеспечения связей с KB (таких как Wikipedia) при затратах на вычисление обнаружения названных объектов. Помимо этого, предложенная новая функция вероятности дает возможность пользователям искусственно манипулировать моделью, чтобы получить объяснения по назначениям тем или ограниченным объектам в одной и той же теме на основе знания области людьми.

Что касается SVAE и NRTM, то их темы, кажется, трудно интерпретировать с точки зрения качества, подтверждающейся результатами количественной оценки.

ЗАКЛЮЧЕНИЕ

Предлагаем два класса тематических моделей с ограничением объекта для объединения различных типов реляционной информации. Результаты демонстрируют, что модели, изучающие отношения документ-уровень достигают улучшения относительно их нереляционных аналогов. Другими словами, концкптуальные отношения незначительно улучшает либо когерентность темы, либо интерпретабельность. В качестве дальнейшей работы планируем изучать полиреляционные тематические модели, извлекающие другие отношения из данных, и рассмотреть метод контекстуального кодирования для представления объекта также и в многоязычных установках [57, 58].

ПРИЛОЖЕНИЕ

1. Предварительная обработка

Мы набрали строчными буквами текст, удалили английские стоп-слова, слова, встречающиеся менее 10 раз, и отфильтровали документы, содержащие менее 2 слов. Подробности по составлению словаря приводятся в табл. 5.

2. Гиперпараметры

Каждый эксперимент с заданным набором параметров повторялся 100 раз и измерения эффективности усредняются по ряду выборок.

Гиперпараметры α и β устанавливаются равными 50/K и 0,1 соответственно (как сообщается в [52]) для всех рассматриваемых моделей. Все сравниваемые модели проверяются на 1 500 итерациях по Гиббсу.

По нашей оценке, мы рассматриваем только ограниченные отношения, которые могут генерироваться объектами и словами. Чтобы выбрать наиболее подходящее значение для порога $\in_m$, мы изучили эффективность тематической когерентности наших моделей, варьируя значением параметра. Значения всех моделей с функциями вероятности EE и EW составляют 0,8 и 0,7 соответственно для массива данных Cora-ML и 0,6, и 0,6 для WebKB.

3. Совместные распределения предложенных моделей

Ради полноты приводим совместное распределение предложенных моделей. Латентное распределение с органичением объекта Дирихле определяет следующую вероятность распределения:

$$P\left(u,z,\theta,\Phi\middle|\alpha,\beta,L\right)\propto \tag{5а}$$

$$\prod_{d=1}^{D}p\left(\theta_d|\alpha\right)\prod_{n=1}^{N_d}p\left(u_{nd}|\Phi_{z_{nd}}\right)p\left(z_{nd}|\theta_d\right) \tag{5б}$$

$$\prod_{k}^{K}p\left(\Phi_k|\beta\right)\cdot\xi\left(z,L\right) \tag{5в},$$

где

D означает множество документов,

N_d – длина документа d,

K означает фиксированное число тем,

u означает множество слов и реализаций названных объектов ,

ξ представляет множество распределений тем,

θ представляет распределение документ-тема,

Φ означает распределение темы-слова,

α и β являются гиперпараметрами Дирихле, связанными с θ и Φ

$$\xi\left(z,L\right)=\prod_{z\in Z}\exp f_l\left(z,u\right)\ .$$

Таблица 5

Резюме словарей для критериев массивов данных до и после фазы предварительной обработки

	Обработанный массив			Необработанный массив
	#уникальные объекты	# уникальные слова	# уникальные объекты и слова	#уникальные слова
Cora	384	2 675	3 059	3 012
WebKB	355	1 874	2 299	2 247

Аналогично, совместная вероятность распределения реляционных тематических моделей с ограничением объекта определяется следующим образом:

$$P\left(u,z,y,\theta,\Phi\middle|\alpha,\beta,\eta,\nu,L\right)\propto \quad (6а)$$

$$\prod_{d=1}^{D}p\left(\theta_d\middle|\alpha\right)\prod_{n=1}^{N_d}p\left(u_{nd}\middle|\Phi_{z_{nd}}\right)p\left(z_{nd}\middle|\theta_d\right) \quad (6б)$$

$$\prod_{k}^{K}p\left(\Phi_k\middle|\beta\right)\prod_{\substack{d,d'\in D\\ d'\neq d}}\psi_\sigma\left(y_{d,d'}\middle|z_d,s_{d'}\eta,\nu\right)\cdot\xi\left(z,L\right) \quad (6в),$$

где ψ_σ - функция вероятности связи, определяемая как $\psi_\sigma\left(y=1\right)=\sigma\left(\eta^T\left(\overline{z}_d{}^{\circ}\overline{z}_{d'}\right)+\nu\right)$, σ – сигмовидная функция и $\overline{z}_d=\frac{1}{N_d}\sum_n z_{nd}$. Эта функция связи моделирует каждую попарную бинарную переменную, касающуюся связей как логистическую регрессию (со скрытыми совместными вариантами), которая задает параметры совместных коэффициентов η и пересечения ν.

4. Инфраструктура вычислений

Эксперименты проводились на трех общих компьютерах с использованием центрального процессора. Модели могут быть выполнены с помощью базовой инфраструктуры. Два компьютера имеют 8 Гб оперативной памяти и еще один - 16 Гб оперативной памяти.

ЛИТЕРАТУРА

1. *Blei D. M., Ng A, Y., Jordan M. I.* Latent Dirichlet allocation// Journal of Machine Learning Research. — 2003.— Vol. 3. —P. 993–1022.

2. *Chang J., Blei D. M.* Relational topic models for document networks// Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009. — 2009. — P. 81–88.

3. *Yang Y., Downey D., Boyd-Graber J. L.* Efficient methods for incorporating knowledge into topic models// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. — 2015. —P. 308–317.

4. *Chen Z., Mukherjee A., Liu B., Hsu M., Castellanos M., Ghosh R.* Discovering coherent topics using general knowledge// Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013.— 2013. — P. 209–218.

5. *Chen Z., Mukherjee A., Liu B., Hsu M., Castellanos M., Ghosh R.* Leveraging multi-domain prior knowledge in topic models// Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI). —2013. — P. 2071–2077.

6. *Kim H., Sun Y., Hockenmaier J., Han J.* ETM: Entity topic models for mining documents associated with entities// Proceedings of the 12th IEEE International Conference on Data Mining, ICDM 2012. — 2012. — P. 349–358,

7. *Wang Q., Song D., Li X.* Incorporating entity correlation knowledge into topic modeling// Proceedings of the IEEE International Conference on Big Knowledge, ICBK 2017, Hefei, China, August 9-10, 2017. — 2017. — P. 254–258.

8. *Allahyari M., Kochut K.* Discovering coherent topics with entity topic models// Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016. — 2016. — P. 26–33.

9. *Wang X., McCallum A., Wei X.* Topical n-grams: Phrase and topic discovery, with an application to information retrieval// Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA. —2007. — P. 697–702.

10. *Gruber A., Weiss Y., Rosen-Zvi M.* Hidden topic markov models// Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007. — 2007. — P. 163–170.

11. *Lindsey R. V., Headden W., Stipicevic M.* A phrase-discovering topic model using hierarchial pitmat-yor processes// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL 2012, July 12-14, 2012, Jeju Island, Korea. — 2012. — P. 214–222.

12. *Fei G., Chen Z., Liu B.* Review topic discovery with phrases using the polyaúrn model// Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014, August 23-29, 2014, Dublin, Ireland. — 2014. — P. 667–676.

13. *Wallach H. M.* Topic modeling: Beyond bag-of-words// Proceedings of the 23rd International Conference on Machine Learning, (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006. — 2006. — P. 977–984.

14. *Griffiths T. L., Steyvers M., Blei D. M., Tenenbaum J. B.* Integrating topics and syntax//Advances in Neural Information Processing Systems 17. — [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]. — 2004. — P. 537–544.

15. *Boyd-Graber J. L., Blei D. M.* Syntactic topic models// Advances in Neural Information Processing Systems 21 // Proceedings of the 22nd Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008. — 2008. — P. 185–192.

16. *Andrzejewski D., Zhu X., Craven M.* Incorporating domain knowledge into topic modeling via Dirichlet forest priors// Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009. — 2009. —P. 25–32,

17. *Andrzejewski D., Zhu X., Craven M., Recht B.* A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic// IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. —2011. — P. 1171–1177.

18. *Petterson J., Smola A. J., Caetano T. S., Buntine W. L., Narayanamurthy S. M.* Word features for latent Dirichlet allocation// Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010// Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada. — 2010. — P. 1921–1929.

19. *Zhao H., Du L., Buntine W. L.* A word embeddings informed focused topic model// Proceedings of The 9th Asian Conference on Machine Learning, ACML 2017, Seoul, Korea, November 15-17, 2017. — 2017. — P. 423–438.

20. *Das R., Zaheer M., Dyer C.* Gaussian LDA for topic models with word embeddings// Proceedings of the 53rd

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. — 2015. — P. 795–804.

21. *Nguyen D. Q., Billingsley R., Du L., Johnson M.* Improving topic models with latent feature word representations // Transactions of the Association for Computational Linguistics, — 2015. — Vol. 3. — P. 299–313.

22. *Li C., Wang H., Zhang Z., Sun A., Ma Z.* Topic modeling for short texts with auxiliary word embeddings// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016. — 2016. — P. 165–174.

23. *Batmanghelich K., Saeedi A., Narasimhan K., Gershman S.* Nonparametric spherical topic modeling with word embeddings// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. — 2016.

24. *Nozza D., Fersini E., Messina E.* Unsupervised irony detection: A probabilistic model with word embeddings// International Conference on Knowledge Discovery and Information Retrieval, volume 2. — P. 68–76. — SCITEPRESS, 2016.

25. *Chen N., Zhu J., Xia F., Zhang B.* Generalized relational topic models with data augmentation// Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI 2013, Beijing, China, August 3-9, 2013. — 2013. — P. 1273–1279.

26. *Terragni S., Fersini E., Messina E.* Constrained relational topic models// Information Sciences, — 2020. — Vol. 512. — P. 581 – 594.

27. *Zhang A., Zhu J., Zhang B.* Sparse relational topic models for document networks// Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I. —2013. — P. 670–685.

28. *Yang W., Boyd-Graber J. L., Resnik P.* Birds of a feather linked together: A discriminative topic model using link-based priors// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. — 2015.— P. 261–266.

29. *Yang W., Boyd-Graber J. L., Resnik P.* A discriminative topic model using document network structure// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.— 2016.

30. *He Y., Wang C., Jiang C.* Modeling document networks with tree-averaged copula regularization// Proceedings of the 10th ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017. — 2017. — P. 691–699.

31. *Mei Q., Cai D., Zhang D., Zhai C.* Topic modeling with network regularization// Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008. —2008. — P. 101–110.

32. *Mimno D. M., McCallum A.* Topic models conditioned on arbitrary features with dirichlet-multinomial regression// UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008. — 2008. — P. 411–418.

33. *Hefny A., Gordon G., Sycara K.* Random walk features for network-aware topic models// NIPS 2013 Workshop on Frontiers of Network Analysis, volume 6. — 2013.

34. *Wahabzada M., Xu Z., Kersting K.* Topic models conditioned on relations// Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III. —2010. — P. 402–417.

35. *Miao Y., Yu L., Blunsom P.* Neural variational inference for text processing// Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, volume 48 of JMLR Workshop and Conference Proceedings, pp. 1727–1736. — JMLR.org., 2016.

36. *Bianchi F., Terragni S., Hovy D.* Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. —[arXiv preprint arXiv:2004.03974. — 2020].

37. *Bianchi F., Terragni S., Hovy D., Nozza D., Fersini E.* Cross-lingual contextualized topic models with zero-shot learning, — [arXiv preprint arXiv:2004.07737.— 2020].

38. *Bai H., Chen Z., Lyu M. R., King I., Xu Z.* Neural relational topic models for scientific article analysis// Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018. — 2018. — P. 27–36.

39. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* Distributed representations of words and phrases and their compositionality// Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013// Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. — 2013. —P. 3111–3119.

40. *McCallum A., Corrada-Emmanuel A., Wang X.* Topic and role discovery in social networks// Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05, Edinburgh, Scotland, UK, July 30 – August 5, 2005, — 2005, — P. 786–791.

41. *Sen P., Namata G., Bilgic M., Getoor L., Gallagher B., Eliassi-Rad T.* Collective classification in network data// AI Magazine, — 2008, — Vol. 29, No. 3. — P. 93–106.

42. *Fersini E., Messina E., Felici G., Roth D.* Soft-constrained inference for Named Entity Recognition// Information Processing & Management, — 2014. — Vol. 50, No. 5, — P. 807–819.

43. *Ritter A., Clark S., Mausam, Etzioni O.* Named entity recognition in tweets: An experimental study// Proc, of the 2011 Conference on Empirical Methods in Natural Language Processing. —2011. — P. 1524–1534.

44. *Li J., Sun A., Han J., Li C.* A survey on deep learning for named entity recognition// IEEE Transactions on Knowledge and Data Engineering. — 2020.

45. *Cucerzan S.* Large-scale named entity disambiguation based on Wikipedia data// Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2007. — P. 708–716.

46. *Dredze M., McNamee P., Rao D., Gerber A., Finin T.* Entity disambiguation for knowledge base population// Proc, of the 23rd International Conference on Computational Linguistics. —2010. — P. 277–285.

47. *Basile P., Caputo A., Semeraro G., Narducci F.* UNIBA: Exploiting a distributional semantic model for disambiguating and linking entities in tweets// Proc. of the 5th Workshop on Making Sense of Microposts co-located with the

24th International World Wide Web Conference, volume 1395, page 62. — 2015.

48. *Cecchini F. M., Fersini E., Manchanda P., Messina E., Nozza D., Palmonari M., Sas C.* UNIMIB@NEEL-IT: Named entity recognition and linking of Italian Tweets// Proc. of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 1749. — 2016.

49. *Nozza D., Sas C., Fersini E., Messina E.* Word embeddings for unsupervised named entity linking// International Conference on Knowledge Science, Engineering and Management. — P. 115–132. — Springer, 2019.

50. *Mendes P. N., Jakob M., García-Silva A., Bizer C.* Dbpedia spotlight: Shedding light on the web of documents// Proceedings of the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011, ACM International Conference Proceeding Series, — P. 1–8. — ACM, 2011.

51. *Yamada I., Asai A., Shindo H., Takeda H., Takefuji Y.* Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia, — [arXiv preprint arXiv:1812.06280.— 2018].

52. *Griffiths T. L., Steyvers M.* Finding scientific topics// Proceedings of the National Academy of Sciences, 101(Suppl, 1), — 2004. — P. 5228–5235,

53. *AlSumait L., Barbara D., Gentle J., Domeniconi C.* Topic significance ranking of LDA generative models// Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009. —2009. — P. 67–82.

54. *Dieng A. B., Ruiz F. J. R., Blei D. M.* Topic modeling in embedding spaces. — [CoRR, abs/1907.04907. — 2019].

55. *Aletras N., Stevenson M.* Evaluating topic coherence using distributional semantics// Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany. — P. 13–22. — The Association for Computer Linguistics, 2013.

56. *Röder M., Both A., Hinneburg A.* Exploring the space of topic coherence measures// Proceedings of the 8th ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015. — 2015. — P. 399–408,

57. *Devlin J., Chang M.-L., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers). — P. 4171–4186. —Association for Computational Linguistics, 2019.

58. *Nozza D., Bianchi F., Hovy D.* What the [mask]? making sense of language-specific bert models. —[arXiv preprint arXiv:2003.02912. — 2020].