МЕЖДУНАРОДНЫЙ ФОРУМ ПО ИНФОРМАЦИИ

2021 **Том 46** **№ 1**

Выходит ежеквартально **Основан в 1975 г.**

СОДЕРЖАНИЕ

Москва **Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН)**

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

DOI: 10.36535/0203-6460-2021-01-1

Библиометрическое отображение исследования по подготовке пользователя для безопасного применения информационных систем*

Дамьян ФУЙС
(Damjan FUJS)
Люблянский университет,
г. Любляна, Словения

Симон ВРХОВЕЦ
(Simon VRHOVEC)
Мариборский университет,
г. Марибор, Словения

Дамьян ВАВПОТИЧ
(Damjan VAVPOTIČ)
Люблянский университет,
г. Любляна, Словения

Информационные системы повсеместно распространены в организациях всех размеров. Для их безопасного применения пользователи должны быть тщательно подготовлены соответствующим образом. В связи с распространенностью информационных систем число научных публикаций о подготовке пользователей для безопасного использования информационных систем из года в год растет. Чтобы преодолеть проблему ручного труда при обзоре такого объема знания и идти в ногу с исследовательскими тенденциями, было проведено библиометрическое отображение в виде карт исследования по подготовке пользователей для безопасного применения информационных систем. Общее число документов, равное 1955 единицам, опубликованных в период 1991-2019 гг., взято из библиографической базы данных Web of Science 21 ноября 2019 г. Авторы с топовой продуктивностью, организации, страны и области исследования были идентифицированы с помощью встроенного в Web of Science средства для анализа результатов. Кроме того, осуществлено отображение в виде карт ключевых слов (КС) на основе программного обеспечения VOSviewer. Анализ сетевой работы и входящих в нее карт КС обнаружил шесть кластеров: Здравоохранение, Принятие Технологии, Управление, Информационная Безопасность, Технические Решения и Физическая Безопасность. Результаты данного анализа предполагают для проведения в будущем привлекательные исследовательские направления, такие как подготовка в сфере информационной безопасности в здравоохранении и индивидуальная подготовка пользователя как альтернатива подходу «одна форма для всех».

* Перевод Fujs D., Vrhovec C., Vavpotič D. Bibliometric mapping of research on user training for secure use of information systems//Journal of Universal Computer Science. —2020. — Vol. 26, No. 7 —P.764 -782. — https://www.researchgate.net/profile/Damjan_Fujs/publication/344163093_Bibliometric_Mapping_of_Research_on_User_Training_for_Secure_Use_of_Information_Systems/links/5f575962458515e96d3911d3/Bibliometric-Mapping-of-Research-on-User-Training-for-Secure-Use-of-Information-Systems.pdf

ВВЕДЕНИЕ

Люди вовлечены или должны быть вовлечены в образование с раннего возраста, поскольку обучение является естественным для хода человеческого развития. Исследование образования уходит корнями к древнегреческому философу Платону, который интересовался фундаментальными вопросами образования: кто и как должен получать образование [1]? Ответ на первый вопрос кажется вполне простым – каждый должен получить какое-то образование. Это также справедливо в отношении кибербезопасности и особенно безопасности информационных систем. Пользователи информационных систем должны быть соответствующим образом подготовлены для их безопасного использования [2]. Организации стремятся научить своих сотрудников избегать киберугроз и защищать интересы организаций [3]. Однако подходы образования по схеме «одна форма подготовки для всех» не могут соответствовать всем ситуациям, а некоторые подходы будут больше, чем другие, подходить в определенных ситуациях [4]. Например, подходы могут рассматривать различия уровня знания пользователей информационных систем, относящегося к кибербезопасности [5, 6, 7]. Такие подходы способны увеличить эффективность подготовки и снизить вероятность или масштаб сопротивления относительно подготовки [8].

В последние годы было проведено много разнообразных обзоров литературы по кибербезопасности и областей исследования в сферах образования. Например, в работе [9] рассматривалось использование качественных подходов в кибербезопасности, которые включали исследование образования по безопасности и подготовке, а в работе [10] изучались компетенции информатики сферы здравоохранения, решающие для образования в информационной технологии. Однако оказывается, что здесь существует пробел в исследовании, поскольку ни один из этих обзоров литературы подробно не фокусировался на образовании в сфере кибербезопасности, а также на более детальной подготовке по безопасному использованию информационных систем. Традиционные обзоры литературы обычно используют чтение релевантных статей с добавлением к ним элемента исследовательского суждения. Эта субъективность может быть снижена анализом КС (т. е. библиометрическим отображением на картах), поскольку он опирается на автоматический качественный анализ с заранее определенным алгоритмом [11]. Недавно библиометрия стала притягательной силой для множества научных дисциплин (например, науки сферы здравоохранении [12], туризм [13] и вычислительная наука [14]).

Чтобы рассмотреть представленный научный пробел, воспользуемся библиометрией и определим направления в исследовании по подготовке пользователя к безопасному использованию информационных систем; мы составили библиометрические карты [15], которые позволяют определить научные направления и выявить наиболее заметные исследовательские вклады. Этот обзор литературы может помочь исследователям и практикам в области кибербезопасности сфокусироваться на соответствующих направлениях в исследовании относительно подготовки пользователя для безопасного применения информационных систем и найти такие, где можно продвигаться за рамки существующего положения дел. В целях достижения этого данная статья изучает следующие вопросы исследования:

Вопрос исследования 1: Кто является наиболее продуктивными авторами, каковы страны, организации и научные области, связанные с исследованием по обучению пользователей для безопасного применения информационных систем?

Вопрос исследования 2: Какие КС наиболее часто появляются в исследовании по подготовке пользователя для безопасного применения информационных систем?

Вопрос исследования 3: Какие КС появлялись раньше, а какие позже в исследовании по подготовке пользователя для безопасного использования информационных систем?

ТЕОРЕТИЧЕСКАЯ ОСНОВА

Подготовка пользователя для безопасного применения информационных систем

Информационные системы могут быть определены как сущность, состоящая из пользователей, выполняющих относящиеся к информации задачи, и разнообразных информационных технологий, использующихся для реализации этих задач [16]. Следовательно, информационные системы включают множество компонентов, таких как персональные компьютеры, социальные сети, банкоматы, смартфоны для делового и личного пользования и т.д. Уже некоторые ранние исследования в сфере информационных систем концентрировались на подготовке пользователей информационных систем (например, компьютерные инструкции) [17, 18]. Более новые исследования фокусировались на пользователях информационных систем, внедряющих меры безопасности [19, 20]. Недавно распространение исследований по образованию в сфере кибербезопасности расширилось благодаря помощи инновационных обучающих подходов, таких как правила по безопасности [21], персональная подготовка [22] и растущая реальность [23].

В течение последнего десятилетия значение обучения пользователей безопасному использованию информационных систем кажется растет [24]. Неотвечающая требованиям подготовка пользователей для безопасного использования информационных систем в организациях касается, с одной стороны, как пользователей со слабым знанием безопасного использования информационных систем, так и персонала, не имеющего навыков обучения, с другой. Кибернетическая безопасность рассматривается, как правило, в области отделений информационной технологии (IT). Персонал таких отделений обычно хорошо осведомлен о киберугрозах и контрмерах с технологической точки зрения. Однако информационно-технологический персонал часто не имеет навыков, необходимых для подготовки пользователей информационных систем, и по заведенной практике пользователи не применяют необходимые меры кибербезопасности, поскольку они считают, что кибербезопасность входит в сферу ответственности отделов IT. Даже при наличии самостоятельных отделов по информационной безопасности это представляет проблему, так как образование в сфере кибербезопасности для пользователей информационных систем часто остается без внимания из-за отсутствия персонала по кибербезопасности [25]. Эти проблемы, кажется, проявляются в большом масштабе. Например, почти половина организаций в Великобри-

тании сообщала, что их проблемы по кибербезопасности были связаны с отсутствием навыков у своих сотрудников [26]. Следовательно, решающим может быть то, что все сотрудники организации, а не только персонал, занятый информационными технологиями, должны быть в достаточной степени подготовлены для безопасного использования информационных систем. Также решающим вопросом может быть развитие организационной культуры, где кибербезопасность считается для каждого ответственностью, а не привилегией.

Библиометрия и связанные подходы

Библиометрия уходит своими корнями в статистику и библиографию и может быть описана как количественное библиографическое исследование литературы (например, ассоциации между публикациями и их ссылками) [27]. Библиометрия позволяет анализировать выбранные темы, возвращаясь на многие десятилетия назад (например, 30 лет [28] и 50 лет [29], следовательно, охватывая значительный объем данных и получая глубокое проникновение в эволюцию исследуемой темы).

Существуют две другие метрики измерения направлений публикации, имеющие отношение к терминам: наукометрия и информетрия [27]. *Наукометрия* часто используется в исследованиях, относящихся к безопасности информационных систем, применяющих качественные метрики научной деятельности (например, импакт-фактор журнала, h-индекс журнала, квартиль журнала, год публикации, ссылки), и может быть использована для определения влияния авторов [27]. Например, недавнее наукометрическое исследование обнаружило, что научные публикации с более длинными рефератами и публикации с большим числом ссылок получают высокое число ссылок в области исследования информационной безопасности [30].

Информетрия – самый широко используемый метод в вычислительной науке, поскольку он касается не только библиографической информации [27], но и включает разнообразные метрики, которые фокусируются на информационной продуктивности [31]. Информетрия затрагивает не только научные метрики, она также применима во множестве областей, где может анализироваться информация [32]. Разнообразные научные и другие БД обеспечивают информетрию. Однако есть проблемы с последовательностью метрик в разных БД. Например, Google Scholar обеспечивает иной подсчет ссылок для публикаций, чем Web of Science. Чтобы рассмотреть эту проблему имеется ряд появившихся в литературе решений, таких как использование технологии «быстрых статей» (smart papers) и блокчейна, позволяющей децентрализованную публикацию и вычисление на основе информетрии [32].

Также возникают новые подходы, такие как *взаимное посредничество публикаций (intermediacy of publications)*. Эти подходы дают возможность проводить сравнение между старыми и более поздними публикациями и могут помочь осуществить мониторинг эволюции научного знания на основе сети ссылок [33]. Поскольку научная продуктивность растет, то средства и методы, позволяющие проводить эффективный анализ массы данных, приобретут большую важность [34].

МЕТОД

Применяемая для исследования методология приведена на рис. 1 и подробно представлена в следующих подразделах.

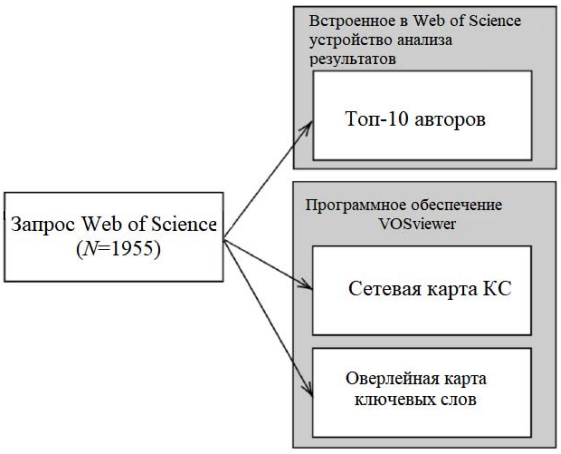


Рис. 1. Сбор данных и анализ

Сбор данных

В целях получения наиболее релевантных статей по подготовке потребителя к безопасному использованию информационных систем в анализ были включены следующие темы для просмотра в библиографических БД: информационные системы, подготовка и образование в сфере безопасности. Оба термина – education (образование) и training (подготовка) являются релевантными, так как пользователи могут либо самостоятельно обучаться, либо руководствуясь определенной подготовкой. Приведенные выше темы использовались для формирования запроса, который применялся для поиска релевантных документов в БД Web of Science (https://apps.webofknowledge.com/):

TOPIC: (*information systems* AND (*security training* OR *security education*))

Выбранные темы намеренно были очень широкими, чтобы увеличить исчерпывающую картину поисковой области. Затем результаты уточнялись по типу документов: *ARTICLE* (*СТАТЬЯ*) or *PROCEEDINGS PAPER* (*ТРУДЫ*). Это позволило осуществить поиск журнальных статей и материалов конференций, связанных с образованием в области безопасности информационных систем. Библиографические документы, общее количество равно 1955 (N=1955), были получены 21 ноября 2019 г.

Анализ данных

Первое, библиографические документы анализировались с помощью встроенного устройства Web of Science для идентификации топ-10 внесших свой вклад авторов, организаций, стран и областей исследования. Отчеты по результатам анализа включают общее число публикаций (N), долю всех публикаций, включенных в обзор, которую они представляют (%), наиболее цитируемую публикацию, не исключая самоцитирование (Ref_N), авторов самой цитируемой публикации и число ссылок для самой цитируемой публикации (TC_R). Публикации могут перекрываться между разными авторами, странами и областями исследования. Например, оба автора А и Б могут иметь одинаковое число идентифици-

рованных публикаций (например, 7 публикаций). Это означает, что они не являются соавторами любой из них и все публикации разные (т.е. 7+7=14 публикаций), они написали их совместно и все публикации перекрываются (нааример, 7 публикаций) или они написали в соавторстве несколько публикаций (например,что-то между 8 и 13 публикациями).

Второе, библиографические документы анализировались с помощью разработанного программного обеспечения VOSViewer (версия 1.6.13) [35]. Были созданы два разных библиографических наглядных представления, а именно, сетевая и оверлейная карты КС. *Сетевая карта ключевых слов* (*Network map of keywords*) отражает корреляции между КС и включает создание графов, где каждое КС визуально представлено узлом, размер которого пропорционален числу публикаций, где узлами показывают родственные КС, т.е. КС, которые обычно встречаются вместе в публикации [35]. *Оверлейная карта ключевых слов* (*Overlay map of keywords*) включает измерение времени в сетевой граф с помощью указания ранних и поздних появлений КС в публикациях. Был проведен кластерный анализ 331 (из 7310) слова, появившегося по меньшей мере 5 раз в изучаемых публикациях для обеих карт КС. Визуальное отображение тематических областей на основе совместной встречаемости КС [35] позволяет проводить как качественный, так и количественный анализ и является полезным средством, облегчающим идентификацию релевантных областей исследования, хотя для детального анализа определенной области исследования необходимо проведение систематического обзора литературы.

РЕЗУЛЬТАТЫ

Данный раздел прежде всего представляет анализ топовых авторов, организаций, стран и доминирующих научных областей в исследовании по подготовке пользователей для безопасного использования информационных систем. Далее приводятся и анализируются карты сети и перекрытия КС.

Топовые авторы

В соответствии с табл. 1 все внесшие свой вклад топовые авторы опубликовали схожее число публикаций. Тем не менее, мы можем разделить их на три группы, а именно: авторы с 7, 6 и 5 публикациями. Кроме того, разумно рассмотреть число ссылок, поскольку они могут быть наводящим на мысль показателем качества автора в дополнение к числу публикаций. Выделяются два автора с публикациями, имеющими более высокое число ссылок – Чэнь Л. – 55 ссылок [41] и Ван С.– 49 ссылок [38]. Обе статьи появились сравнительно недавно, но несмотря на это имеют высокое число ссылок, дополнительно показывающих, что обе они хорошего качества, независимо от высокой доли самоцитирований (25,4% и 26,5% соответственно).

Табл. 2 отражает топовые организации. Система Калифорнийского университета (University of California System) кажется самой продуктивной организацией, а Университетская система шт. Джорджия (University System of Georgia) – самой влиятельной среди организаций в соответствии с числом ссылок на наиболее цитируемую публикацию. Ни один из указанных в табл. 2 авторов не выглядит топовым автором, это показывает, что самые продуктивные ученые совсем необязательно выходят из самых продуктивных научных организаций.

Как видно из табл. 3, США имеют наибольшее число публикаций, за ними сравнительно близко идет Китай. К этим двум странам можно добавить Индию, которая единственная из далее приведенных стран имеет более 100 публикаций. Индия среди указанных топовых стран является страной с самой цитируемой публикацией, за ней идут США и Южная Корея.

Таблица 1

Топ-10 авторов

Автор	N	%	Ref_N	Авторы Ref_N	TC_R
Tugnait JK	7	0,35	Tugnait [36]	*Tugnait JK*	34
Kim J	7	0,35	Park [37]	Park HE, *Kim J*, Park YS	8
Wang C	6	0,30	Wang [38]	Wang HM, *Wang C* , Ng DWK	49
Li X	6	0,30	Wu [39]	Wu Y, Weng J, Tang Z, *Li X*	9
Du Q	6	0,30	Xu [40]	Xu D, Ren P, Wang Y, *Du Q*, Sun L	2
Ren P	6	0,30	[Xu [39]	Xu D, *Ren P* , Wang Y, Du Q, Sun L	2
Sun L	6	0,30	Xu [40]	Xu D, Ren P, Wang Y, Du Q, *Sun L*	2
Wang Y	6	0,30	Xu [40]	Xu D, Ren P, *Wang Y*, Du Q, Sun L	2
Chen L	5	0,25	Liu [41]	Liu X, Lu R, Ma J, *Chen L*, Qin B	55
Chen W	5	0,25	Hsu [42]	Hsu J, Liu D, Yiu YM, Zhao HT, Chen ZR, Li J, *Chen W*	21

Примечание: В квадратных скобках приводится порядковый номер автора в «Литературе».

Таблица 2

Топ-10 организаций

Организация	N	%	Ref_N	TC_R
University of California System	28	1,43	Gottlieb et al. [43]	58
Chinese Academy of Sciences	23	1,17	Peng et al. [44]	132
State University of Florida	22	1,12	Biros et al. [45]	40
University of Texas System	18	0,92	Siponen et al. [46]	129

Организация	N	%	Ref_N	TC_R
Bejing Jiaotong University	17	0,86	Zhu et al. [47]	7
Penn State University	16	0,81	D'Arcy et al. [4]	335
United States Department of Defense	16	0,81	Biros et al. [45]	40
University System of Georgia	16	0,81	Straub and Welke [48]	399
Bejing University of Posts	14	0,71	Peng et al. [44]	132
Telecommunications				
University of London	13	0,71	Perera et al.[49]	95

Таблица 3

Топ-10 стран

Страна	N	%	Ref_N	TC_R
США	486	24,84	Straub and Welke [48]	399
Китай	326	16,66	Yuan et al.[50]	171
Индия	116	5,93	Subashini and Kavitha [51]	958
Англия	88	4,49	Willison and Warkentin [52]	134
Австралия	83	4,24	Minasny et al. [53]	147
Россия	72	3,68	Klimova et al. [54]	26
Германия	67	3,42	Baumgart [55]	85
Южная Корея	63	3,22	D'Arcy et al. [56]	335
Канада	59	3,01	Stern et al. [57]	109
Испания	51	2,60	Fernaondez-Alemoan et al. [58]	190

Таблица 4

Топ-10 областей исследования

Область исследования	N	%	Ref_N	TC_R
Вычислительная наука	830	42,4	Subashini and Kavitha [52]	958
Инжиниринг	568	29,0	Ming et al. [59]	157
Образование и исследование образования	219	11,2	Einterz et al.[60]	165
Телекоммуникации	179	9,2	Yuan et al. [50]	171
Экономика бизнеса	121	6,2	Straub and Welke [48]	399
Науки здравоохранения и услуги	86	4,4	Wu et al. [61]	242
Информатика и библиотековедение	79	4,0	Straub and Welke, [48]	399
Медицинская информатика	72	3,7	Wu et al. [61]	242
Здравоохранение, окружающая среда, профессиональная защита	59	3,0	Stern et al. [57]	109
Общественные науки, другие темы	55	2,77	D'Arcy and Hovav [4]	44

Табл. 4 отражает самые доминирующие области исследования. Области вычислительной науки и инжиниринга доминируют в соответствии с числом публикаций, поскольку они охватывают более двух третьих всех публикаций.

Отображение ключевых слов и кластерный анализ

В целях более глубокого понимания и идентификации «горячих» точек исследования библиографические данные представлены визуально. Более связанные КС располагаются ближе друг к другу, что означает: между ними существуют только незначительные различия и они имеют более высокую совместную встречаемость. На рис. 2 показаны шесть идентифицированных основных кластеров*: Healthcare (Здравоохранение), Technology Adoption (Принятие Технологии), Management (Управление), Information Security (Информационная Безопасность), Technical Solutions (Технические Решения) и Physical Security (Физическая Безопасность).

Самыми известными КС в исследовании относительно подготовки пользователя для безопасного применения информационных систем являются: information security (информационная безопасность), management (управление), awareness (осознанность), machine learning (машинное обучение), intrusion detection (интрузивное детектирование), design (дизайн), network security (безопасность сети) и cyber security (кибербезопасность). Карта также предполагает наличие двух различных полюсов. Левый полюс главным образом представляет темы, относящиеся к человеку, а правый – к технологии.

* Чтобы различать названия кластеров и КС, названия кластеров будут приводиться с заглавной буквы. Для рисунков 2 и 3 это разграничение не свойственно. Рисунки приводятся так, как даны у авторов данной работы. — (*прим. ред.*)

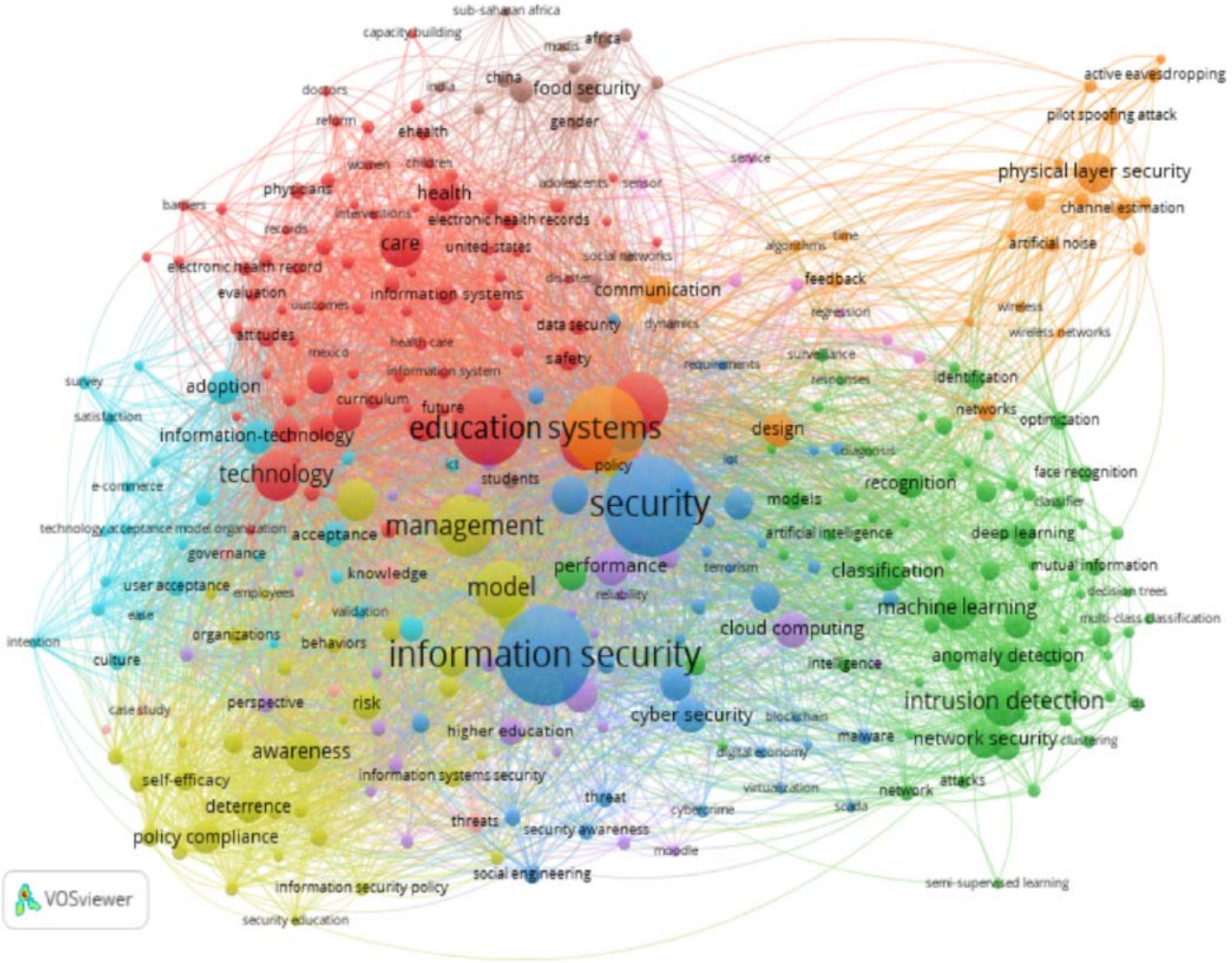


Рис. 2. Изображение сети ключевых слов с выделенными особым образом кластерами

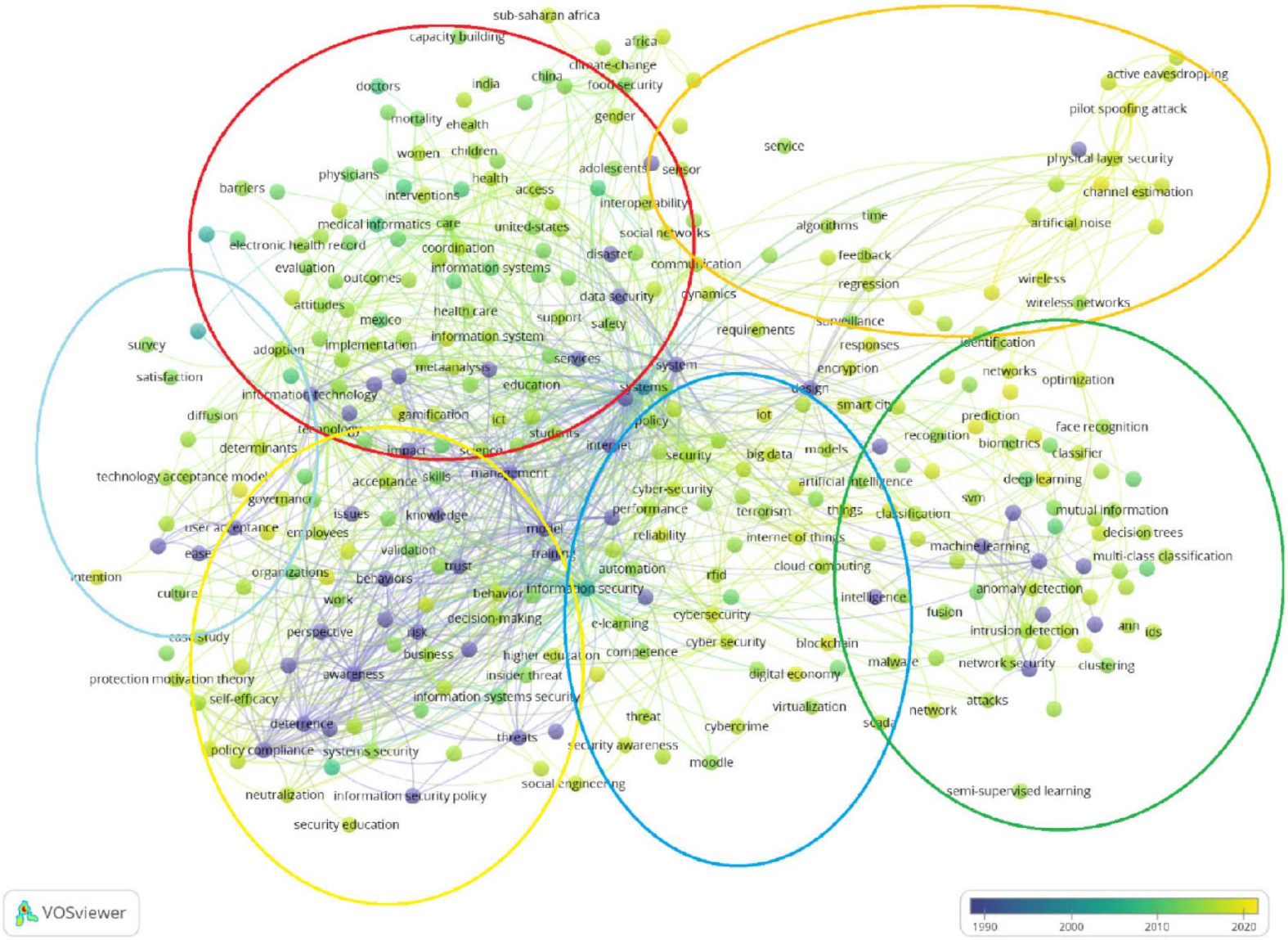


Рис. 3. Оверлейная карта ключевых слов с помеченными кластерами

Однако кажется, что некоторые темы расположены вопреки ожиданию на противоположном полюсе, например, technology (технология) располагается на левом полюсе, представляющем темы, относящиеся к человеку, возможно это из-за того, что это фраза с широким значением (например, как в Принятии Технологии) и поэтому может быть тесно связана с темами из обоих полюсов. Также интересно, что education (образование) помещено в кластер Healthcare (Здравоохранение). Это означает важность защиты чувствительных персональных данных в области здравоохранения. Следовательно, пользователи информационных систем должны быть хорошо образованы в сфере информационной безопасности и защиты данных в этом секторе.

Ключевое слово education (образование) сильно связано с information security (информационной безопасностью) в кластере Information Security (Информационная Безопасность), adoption и acceptance (принятие) – в кластере Technology Adoption (принятие технологии), a security policy compliance (согласие с политикой безопасности), management (управление) и awareness (осознанность) – в кластере Management (Управление). В кластере Technical Solutions (Технические Решения) education (образование) связано с classification (классификацией) и attacks (атаками) и с system (системой), time (временем) и с design (дизайном) – в кластере Physical Security (Физическая Безопасность). Ключевое слово security (безопасность), кажется, в преобладающей степени относится к технологическому полюсу и в меньшей степени - к социологическому. Это предполагает, что в будущем больший акцент необходимо делать на социотехнологическое исследование. Представленные в табл. 4 данные также это поддерживают, поскольку публикации в социологических науках, кажется, получают меньшее число ссылок, демонстрируя более низкую привлекательность темы.

Далее, в анализ были включены временные измерения. Рис. 3 представляет оверлейную карту КС, где отмечены шесть идентифицированных кластеров КС. Чтобы показать временное измерение этих терминов используется тоновая палитра, от более темного до светлого. Темный тон означает, что КС появилось уже в 1990-х гг., а светлый – что оно появилось только недавно. Большинство старых ключевых слов находится в кластерах Management (Управление) и Technical Solutions (Технические Решения). Узлы недавно появившихся КС, таких как security education (образование в сфере безопасности), social engineering (социальный инжиниринг), security awareness (осознанность безопасности), culture (культура), intention (намерение), identification (идентификация) и responses (реакция), находятся значительно дальше от центра, указывая, что они менее связаны.

Из обеих карт КС можно установить, что большинство КС в кластере Healthcare (Здравоохранение) появилось сравнительно недавно, что они включают КС education (образование) и что они сильно связаны с кластером Information Security (Информационная Безопасность). Это ясно указывает на важность образования по информационной безопасности в сфере здравоохранения. В основном это, возможно, произошло благодаря недавним усилиям информатизации в секторе здравоохранения, где значительная доля работников не была достаточно подготовлена в отношении сферы информационной безопасности или там просто отсутствовала мотивация придерживаться политики по информационной безопасности. Например, первая обязанность медиков – улучшение условий по сохранению здоровья своих пациентов, тогда как что-либо еще часто имеет вторичный характер, несмотря на отношение с очень чувствительными медицинскими данными. Кроме того, работники здравоохранения, как правило, менее подготовлены, чем работники других секторов с более длинной историей информатизации [62]. Также стоит отметить, что относящееся к информационной безопасности знание может в значительной степени варьироваться среди работников в одной и той же организации [63]. Важность индивидуальной подготовки пользователя дополнительно акцентируется КС в кластере Management (Управление). Ключевые слова, такие как culture (культура), behavior (поведение), awareness (осознанность) и satisfaction (удовлетворение), показывают, что принятие пользователем мер безопасности зависит не только от его знания и компетенций, но также и от других относящихся к человеку аспектов. Эти аспекты значительно влияют на то, будут ли пользователи воспринимать подготовку в качестве соответствующего и стоящего усилия [64]. Кажется, что этот вид визуализации указывает на недавно появившееся направление в исследовании. Комбинации новых терминов предполагают – индивидуальное обучение станет перспективным будущим направлением исследования. Исследователи могут идти по пути принятия подготовки пользователя для безопасного применения информационных систем на уровне знания и компетенции отдельного работника. Даже более важным может быть осуществление такой подготовки в секторах, часто имеющих дело с высокочувствительными данными, такими как здравоохранение. Этот вид исследования уже появляется (например, [6, 5, 22]), так как подходы подготовки пользователя по типу «одна форма для всех» не могут быть оптимальными [4].

Кластер Technical Solutions (Технические Решения) также является кластером с рядом последних КС, которые хорошо связаны с КС в кластере Information Security (Информационная Безопасность). Это подразумевает потенциал определенных технологий для значительного вклада в безопасное использование информационных систем. Например, deep learning (глубокое обучение), intrusion detection (интрузивное детектирование), artificial intelligence (искусственный интеллект) и КС, относящиеся к продвинутым статистическим подходам – все это облегчает и продвигает в сторону предотвращения киберугрозы, определения и получения ответа. Однако возможности для будущего исследования лежат не только в этих областях, но также и в подготовке пользователя для безопасного использования информационных систем. Общепринято, что education (образование) связано только с КС authentication (аутентификация), classification (классификация) и attacks (атаки) до тех пор, пока не появляются какие-либо заслуживающие внимания ассоциации с artificial intellingence (искусственным интеллектом, machine learning (машинным обучением) или deep learning (глубоким обучением). Это указывает на пробел исследования и возможность обратиться к нему в будущем. Однако связи между узлами могут интерпретироваться в обоих направлениях. Кроме того, это означает, что существует необходимость в исследовании того, как готовить пользователей относи-

тельно использования продвинутых технологических решений. Следовательно, важность подготовки пользователя может возрасти в связи с широким принятием больших данных, умных городов (smart cities), интернета вещей и других появляющихся продвинутых технологий. Кроме того, проникновение этих технологий в будущие информационные системы вызовет даже большую потребность в индивидуальной подготовке пользователя относительно их безопасного применения.

ОБСУЖДЕНИЕ

Теоретическое и практическое применение

В данной статье проводится изучение положения дел в исследовании по подготовке пользователей для безопасного использования информационных систем. Это комплексное исследование с рядом входящих научных областей, таких как вычислительная наука, образование, экономика, управление и т.д. В статье приводится несколько теоретических и практических выводов на основе проделанного исследования. Первое – анализ КС показывает некоторые интересные области для рассмотрения в будущем, особенно информационную безопасность в определенных контекстах, таких как здравоохранение. Кроме того, заслуживает внимания тот факт, что более высокое совпадение и корреляция КС автоматически не означает качество публикаций, т.е. количество не переходит в качество само по себе, однако это может помочь области исследования постепенно эволюционировать и в итоге дойти до полного развития [65]. Второе – даны таблицы под общим названием «топ-10». Таблицы идентифицируют участвующих (т.е. авторы, организации, страны и области) в исследовании по подготовке пользователей для безопасного использования информационных систем. Эти таблицы могут стать отправной точкой для будущих исследований, касающихся качества работ в изучаемых областях, дополняя результаты описанного в данной статье анализа. Третье – на основе КС было обнаружено, что образование или обучение может быть внедрено как на уровне человеческой деятельности [7], так и на уровне технологии (например, машинное обучение). В обоих случаях присутствует человеческий элемент – или как человек, который обучается, или как человек, который создает обучающую машину. Четвертое – результаты проведенного исследования показывают, что в будущем потребуется сделать больший акцент на индивидуальную подготовку в целях безопасного использования информационных систем.

Ограничения и будущая работа

В статье указываются некоторые ограничения, на которые читатель должен обратить внимание. Во-первых, следует отметить, что данные были взяты из библиографической БД Web of Science, включающей самые влиятельные потоки публикаций с наивысшими стандартами [13]. Поиск работ в будущем в других библиографических БД, таких как Scopus, ACM DL и IEEE Xplore, может быть выгодным, поскольку ими часто пользуются исследователи из сферы безопасности. Во-вторых, библиографическая БД Web of Science предлагает организациям разнообразные подписки. Даже если один и тот же поисковый запрос выполняется в тех же самых указателях Web of Science, поиск дает разные результаты в разных учреждениях, если их подписки различаются. Поисковый запрос был выполнен в массиве Web of Science Core Collection, который включает следующие указатели: SCI-EXPANDED (1900 г. – настоящее время), SSCI (1900 г. – настоящее время), A&HCI (1975 г. – настоящее время), CPCI-S (2011 г. – настоящее время), SPCI-SSH (2011 г. -настоящее время), BKCI-S (2011 г. – настоящее время), BKCI-SSH (2011 г. – настоящее время), ESCI (2015 г. – настоящее время), CCR-EXPANDED (2011 г. – настоящее время) и IC (2011 г. – настоящее время). Это означает, что ряд докладов конференций, опубликованных в период 1991-2010 гг. не вошел в данное исследование. В-третьих, визуальное отображение науки в виде карт не может заменить систематические обзоры литературы [66], однако оно предлагает альтернативный анализ и дает возможность посмотреть на тенденции исследований. Такие анализы динамичны, т. е. со временем они могут изменяться. Это может считаться как ограничением, так и направлением для будущих изучений в определенное время.

ЛИТЕРАТУРА

1. *Noddings N.* Philosophy of Education//Encyclopedia of the Social and Cultural Foundations of Education, pp. 1–156. — SAGE Publications, Inc., 2455 TellerRoad, Thousand Oaks California 91320 United States, 2012.

2. *Choi S., Martins J. T., Bernik I.* Information security: Listening to the perspective of organisational insiders// Journal of Information Science. — 2018. — Vol. 44, No. 6. — P. 752–767.

3. *Aldawood H., Skinner G.* Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues//Future Internet. — 2019. — Vol. 11, No. (3). — P. 73.

4. *D'Arcy J., Hovav A.* Does one size fit all? Examining the differential effects of IS security countermeasures// Journal of Business Ethics. — 2009. — Vol. 89, No. (S1). — P. 59–71.

5. *Vasileiou I., Furnell S.* Enhancing security education recognising threshold concepts and other influencing factors// ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 398–403, Funchal, Madeira, Portugal. — 2018.

6. *Friesel A., Ward A., Welzer T., Poboroniuc M., Mrozek Z.* Building a shared understanding of the skills and competences in order to respond to the current global technical challenges//2014 IEEE Global Engineering Education Conference (EDUCON), pp. 676–679, Istanbul. — IEEE, 2014.

7. *Vavpotič D., Zvanut B., Trobec I.* A comparative evaluation of e-learning and traditional pedagogical process elements// Educational Technology and Society. — 2013. — Vol. 16, No. 3. — P. 76–87.

8. *Vrhovec S. L., Hovelja T., Vavpotič D., Krisper M.* Diagnosing organizational risks in software projects: Stakeholder resistance// International Journal of Project Management. — 2015. — Vol. 33, No. 6. — P. 1262–1273.

9. *Fujs D., Mihelič A., Vrhovec S. L. R.* The power of interpretation// Proceedings of the 14th International Conference on Availability, Reliability and Security - ARES '19, pp. 1–10, New York, New York, USA. — ACM Press, 2019.

10. *Kokol P., Saranto K., Blažun Vošner H.* eHealth and health informatics competences: A systemic analysis of

literature production based on bibliometrics// Kybernetes. — 2018. — Vol. 47, No. 5. — P. 1018–1030.

11. *Fergnani A.* Mapping futures studies scholarship from 1968 to present: A bibliometric review of thematic clusters, research trends, and research gaps// Futures. — 2019. — Vol. 105 (September 2018). — P. 104–123.

12. *Holman D., Lynch R., Reeves A.* How do health behaviour interventions take account of social context? A literature trend and co-citation analysis// Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine. — 2018. — Vol. 22, No. 4. — P. 389–410.

13. *Garrigos-Simon F., Narangajavana-Kaosiri Y., Lengua-Lengua I.* Tourism and sustainability: A bibliometric and visualization analysis// Sustainability. — 2018. — Vol. 10, No. 6. — P. 1976.

14. *Blanco-Mesa, F., Léon-Castro E., Merigó J. M.* A bibliometric analysis of aggregation operators//Applied Soft Computing. — 2019. — Vol. 81. — P. 105-488.

15. *van Eck N. J., Waltman L.* VOSviewer Manual. Technical Report September, Universiteit Leiden, CWTS Meaningful metrics. — 2019.

16. *Vavpotič D., Vasilecas O.* Selecting a methodology for business information systems development: Decision model and tool support// Computer Science and Information Systems. — 2012. — Vol. 9, No. 1. — P. 135–164.

17. *Meliopoulos A. P. S., Cokkinides G. J., Contaxis G. C.* Computer aided instruction of power system security control functions// IEEE Transactions on Power Systems. — 1987. — Vol. 2, No. 1. — P. 232–238.

18. *Chowdhury B., Clark D.* COPERITE computer-aided tool for power engineering research, instruction, training and education// IEEE Transactions on Power Systems. — 1992. — Vol. 7, No. 4. — P. 1565–1570.

19. *Sasse M. A., Brostoff S., Weirich D.* Transforming the 'weakest link' - A human/computer interaction approach to usable and effective security// BT Technology Journal. — 2001. — Vol. 19, No. 3. — P. 122–131.

20. *Stanton J. M., Stam K. R., Mastrangelo P., Jolton J.* Analysis of end user security behaviors// Computers & Security. — 2019. — Vol. 24, No. 2. — P. 124–133.

21. *Cone B. D., Irvine C. E., Thompson M. F., Nguyen T. D.* A video game for cyber security training and awareness// Computers and Security. — 2007. — Vol. 26, No. 1. — P. 63–72.

22. *Vasileiou I., Furnell S.* Personalising Security Education: Factors Influencing Individual Awareness and AC / P. Mori P., S. Furnell, and O. Camp (ed.)// Information Systems Security and Privacy: 4th International Conference, ICISSP 2018, pp. 315–321, Funchal - Madeira, Portugal. — Springer, 2018.

23. *Logofatu B., Visan A.* New trends in the educational area. Case study regarding the usability of google apps tools within the department for distance learning// The 11th International Scientific Conference eLearning and Software for Education, pp. 526–531, Bucharest.— 2015.

24. *Švábenský V., Vykopal J., Čeleda P.* What are cybersecurity education papers about? A systematic literature review of SIGCSE and ITiCSE Conferences// The 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20). — 2020.

25. *(ISC)2.* Strategies for building and growing strong cybersecurity teams. Technical report, (ISC)2. — 2019.

26. *Furnell S., Fischer P., Finch A.* Can't get the staff? The growing need for cyber-security skills// Computer Fraud & Security. — 2017. — Vol. 2017, No.2. —P. 5–10.

27. *Hood W. W., Wilson C. S.* The literature of bibliometrics, scientometrics,and informetrics //Scientometrics. — 2001. — Vol. 52, No. 2. — P. 291–314.

28. *López-Robles J., Otegi-Olaso J., Porto Gómez, I., Cobo M.* 30 years of intelligence models in management and business: A bibliometric review// International Journal of Information Management. — 2019. — Vol. 48(January). — P. 22–38.

29. *Iqbal W., Javed R. T., Qadir J., Mian A. N., Tyson G., Hassan S. U., Crowcroft J.* Five decades of the ACM Special Interest Group on Data Communications (SIGCOMM): A bibliometric perspective// Computer Communication Review. — 2019. — Vol. 49, No. 5. — P. 29–37.

30. *Wendzel S., Lévy-Bencheton, C., Caviglione L.* Not all areas are equal: analysis of citations in information security research// Scientometrics. — 2020. — Vol. 122, No. 1. — P. 267–286.

31. *Sengupta I. N.* Bibliometrics, informetrics, scientometrics and librametrics: An overview// Libri. — 1992. — Vol. 42, No. 2. — P. 75–98.

32. *Hoffman M. R., Ibáñez, L.-D., Simperl, E.* Scholarly publishing on the blockchain – from smart papers to smart informetrics// Data Science. — 2019. — Vol. 2, No. (1-2). — P. 291–310.

33. *Šubelj L., Waltman L., Traag V., van Eck N. J.* Intermediacy of publications// Royal Society Open Science. — 2020. — Vol. 7, No. (1). — P. 190-207:1–16.

34. *Markscheffel B., Kretschmer H., Pichappan P.* Report of 14 th International Conference on Webometrics, Informetrics and Scientometrics (WIS) & 19th COLLNET Meeting 05 to 08 December 2018, University of Macau, Macau// COLLNET Journal of Scientometrics and Information Management. — 2019. — Vol. 13, No. 1. — P. 3–6.

35. *van Eck N. J., Waltman L.* Software survey: VOSviewer, a computer program for bibliometric mapping Scientometrics. — 2010. — Vol. 84, No. 2. — P. 523–538.

36. *Tugnait J. K.* Self-contamination for detection of pilot contamination attack in multiple antenna systems// IEEE Wireless Communications Letters. — 2015. — Vol. 4, No. 5. — P. 525–528.

37. *Park E. H., Kim J., Park Y. S.* The role of information security learning and individual factors in disclosing patients' health information// Computers & Security. — 2017. — Vol. 65. — P. 64–76.

38. *Wang H.-M., Wang C., Ng D. W. K.* Artificial Noise Assisted Secure Transmission Under Training and Feedback// IEEE Transactions on Signal Processing. — 2015. — Vol. 63, No. 23. — P. 6285–6298.

39. *Wu Y., Weng J., Tang Z., Li X., Deng R. H.* Vulnerabilities, Attacks, and Countermeasures in Balise-Based Train Control Systems// IEEE Transactions on Intelligent Transportation Systems. — 2017. — Vol. 18, No. 4. — P. 814–823.

40. *Xu D., Ren P., Wang Y., Du Q., Sun L.* ICASBDC: A channel estimation and identification mechanism for MISO-OFDM systems under pilot spoofing attack//2017 IEEE International Conference on Communications (ICC), pp. 1–6. — IEEE, 2017.

41. *Liu X., Lu R., Ma J., Chen L., Qin B.* Privacy-preserving patient-centric clinical decision support system on Naive Bayesian classification// IEEE Journal of

Biomedical and Health Informatics. — 2016. — Vol. 20, No. 2. — P. 655–668.

42. *Hsu J., Liu D., Yu Y. M., Zhao H. T., Chen Z. R., Li J., Chen W.* The Top Chinese Mobile Health Apps: A Systematic Investigation// Journal of Medical Internet Research. — 2016. — Vol. 18, No. 8 :e222.

43. *Gottlieb L. M., Tirozzi K. J., Manchanda R., Burns A. R., Sandel M. T.* Moving electronic medical records upstream// American Journal of Preventive Medicine. — 2015. — Vol. 48, No. 2. — P. 215–218.

44. *Peng M., Sun Y., Li X., Mao Z., Wang C.* Recent advances in Cloud radio access networks: System architectures, keytechniques, and open issues// IEEE Communications Surveys & Tutorials. — 2017. — Vol. 18, No. 3. — P. 2282–2308.

45. *Biros D. P., George J. F., Zmud R. W.* Inducing sensitivity to deception in order to improve decision making performance: A field study// MIS Quarterly. — 2002. — Vol. 26, No. 2. — P. 119.

46. *Siponen M., Adam Mahmood M., Pahnila S.* Employees' adherence to information security policies: An exploratory field study// Information & Management. — 2014. — Vol. 51, No. 2. — P. 217–224.

47. *Zhu L., Yu F. R., Tang T., Ning B.* An Integrated Train–Ground Communication System Using Wireless Network Virtualization: Security and Quality of Service Provisioning// IEEE Transactions on Vehicular Technology. — 2017. — Vol. 65, No. 12. — P. 9607–9616.

48. *Straub D. W., Welke R. J.* Coping with systems risk: Security planning models for management decision making// MIS Quarterly. — 1998. — Vol. 22, No. 4. — P. 441.

49. *Perera G., Broadbent M., Callard F., Chang C.-K., Downs J., Dutta R., Fernandes A., Hayes R. D., Henderson M., Jackson R., Jewell A., Kadra G., Little R., Pritchard M., Shetty H., Tulloch A., Stewart R.* Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource// BMJ Open. — 2016. — Vol. 6, No. 3:e008721.

50. *Yuan C., Sun X., Lv R.* Fingerprint liveness detection based on multi-scale LPQ and PCA// China Communications. — 2016. — Vol. 13, No. 7. — P. 60–65.

51. *Subashini S., Kavitha V.* A survey on security issues in service delivery models of cloud computing// Journal of Network and Computer Applications. — 2011. — Vol. 34, No. 1. — P. 1–11.

52. *Willison R., Warkentin, M.* Beyond Deterrence: An expanded view of employee computer abuse// MIS Quarterly. — 2013. — Vol. 37, No. 1. — P. 1–20.

53. *Minasny B., McBratney A. B., Malone B. P., Wheeler I.* Digital mapping of soil carbon//Advances in Agronomy. volume 118, pages 1–47. — Elsevier, 2013.

54. *Klimova A., Rondeau E., Andersson K., Porras J., Rybin A., Zaslavsky A.* An international Master's program in green ICT as a contribution to sustainable development// Journal of Cleaner Production. — 2016. — Vol. 135. — P. 223–239.

55. *Baumgart D. C.* Personal digital assistants in health care: Experienced clinicians in the palm of your hand?// The Lancet. — 2005. — Vol. 366, No. 9492. — P. 1210–1222.

56. *D'Arcy J., Hovav A., Galletta D.* User awareness of security countermeasures and its impact on Information Systems misuse: A Deterrence approach// Information Systems Research. — 2009. — Vol. 20, No. 1. — P. 79–98.

57. *Stern N. J., Hiett K. L., Alfredsson G. A., Kristinsson K. G., Reirsen J., Haedardottir H., Briem H., Gunnarsson E., Georgsson F., Lowman R., Berndtson E., Lammerding A. M., Paoli G. M., Musgrove M. T.* Campylobacter spp. in Icelandic poultry operations and human disease//Epidemiology and Infection. — 2003. — Vol. 130, No. 1. — P. 23–32.

58. *Fernández-Alemán J. L., Señor I. C., Lozoya P. á. O., Toval A.* Security and privacy in electronic health records: A systematic literature review// Journal of Biomedical Informatics. — 2013. — Vol. 46, No. 3. — P. 541–562.

59. *Ming J., Hazen T. J., Glass J. R., Reynolds D. A.* Robust speaker recognition in noisy conditions// IEEE Transactions on Audio, Speech and Language Processing. — 2017. — Vol. 15, No. 5. — P. 1711–1723.

60. *Einterz R. M., Kimaiyo S., Mengech H. N., Khwa-Otsyula B. O., Esamai F., Quigley F., Mamlin J. J.* Responding to the HIV pandemic: The power of an Academic Medical partnership// Academic Medicine. — 2007. — Vol. 82, No.8. — P. 812–818.

61. *Wu J.-H., Wang S.-C., Lin L.-M.* Mobile computing acceptance factors in the healthcare industry: A structural equation model// International Journal of Medical Informatic. — 2007. — Vol. 76, No. 1. — P. 66–77.

62. *Vrhovec S., Markelj B.* Relating mobile device use and adherence to information security policy with data breach consequences in hospitals// Journal of Universal Computer Science. — 2018. — Vol. 24, No. 5. — P. 634–645.

63. *van Niekerk J.* Establishing an information security culture in organizations: An outcomes based education approach. Dissertation, Nelson Mandela Metropolitan University. — 2005.

64. *Dincelli E., Goel S.* Research design for study of cultural and societal influence on online privacy behavior//Proceedings of 2015 IPIP 8.11/11.13 Dewald Roode Information Security Research Workshop, pp. 1–18, Newark, Delaware. — 2015.

65. *Hicks D., Wouters P., Waltman L., de Rijcke S., Rafols I.* Bibliometrics: The Leiden Manifesto for research metrics// Nature. — 2015. — Vol. 520, No. (7548). — P. 429–431.

66. *Hallinger P.* Science mapping the knowledge base on educational leadership and management in Africa, 1960–2018//School Leadership & Management. — 2019. — Vol. 39, No. 5. — P. 537–560.

DOI: 10.36535/0203-6460-2021-01-2

Новый подход к тематическому моделированию: от пространства документов к пространству терминов*

Магнус САЛЬГРЕН
(Magnus SAHLGREN)

Исследовательские институты Швеции,
г. Стокгольм, Швеция

В статье рассматривается проблема опоры на документы как базовое понятие для определения взаимодействий термина в стандартных тематических моделях. В качестве альтернативы этой практики мы переформулируем распределения тем в латентные факторы в пространстве сходства терминов. Поясняется идея использования ряда стандартных векторных представлений слов путем построения очень широких окон контекстов. Пространства векторных представлений трансформируются в редкие пространства сходства, а темы извлекаются стандартным способом, перенося факторизацию на пространство заметно меньшего размера. Используется ряд разных способов факторизации и оцениваются различные модели с применением широкого спектра оценочных показателей, включая ранее опубликованные измерения когерентности, а также новые измерения, которые, предположительно, лучше отвечают применениям тематических моделей в реальном мире. Результаты однозначно отражают, что в большинстве случаев модели на основе терминов превосходят стандартные модели на основе документа.

ВВЕДЕНИЕ

Тематические модели часто используются в сценариях реального текстового анализа как способы эффективного изучения данных. Типичным для таких сценариев является обращение к тематической модели со стандартными параметрами данных, а также извлечение некоторого фиксированного числа *n* тем и некоторого фиксированного числа *m* слов по теме, и затем следует интерпретация, составление выводов и окончательного списка терминов. Общим выбором как для *n*, так и для *m* служит примерно 10 единиц. Это подразумевает, что аналитику нужно просмотреть только примерно 100 терминов вместо чтения собрания текстов, содержащего вероятно сотни тысяч или даже миллионов встречающихся слов. С точки зрения эффективности это является ценным средством контент-анализа.

Тематические модели извлекают темы путем раскрытия (латентных) взаимодействий между терминами в пространстве документа. Эта методология, очевидно, предполагает, что данные существуют в четких и постоянных границах документов, в лучшем случае даже справедливо распределение числа слов на документ. К сожалению, это предположение редко встречается в реальных сценариях, где данные существуют в потоках, в группах с нечеткими границами документов или с очень большим разнообразием длины документов. Чтобы осуществить такой сценарий желательно использовать модель, нечувствительную к форматированию вводимых данных. В этой статье описывается и оценивается подобный подход, который представляет в виде векторов процесс тематического моделирования полностью в пространстве терминов. Это делает модель менее чувствительной к форматированию документа и в результате даже более точной.

* Перевод Sahlgren M. Rethinking topic modelling: From document-space to term-document//Findings of the Association for Computational Linguistics: EMNLP 2020, November 16-20. — 2020. —P.2250 -2259. —https://www.aclweb.org/anthology/2020.findings-emnlp.204.pdf

Первоначально работа была мотивирована практическим использованием тематических моделей в анализе сценариев реального мира. В подобных применениях – общих, в частности, для общественных наук, а также сферы безопасности и защиты – аналитик беспокоится только о терминах верхних рангов в окончательном списке терминов. Поэтому мы дополнительно вводим ряд показателей оценки для тематических моделей, которые могут лучше отвечать практическому рассмотрению, чем обычно используемые известные, присущие цели (и в своем большинстве теоретические) измерения оценки. Также предоставляем оценку, считающую тематическое моделирование сценарием аннотации документа и использующую подготовленные вручную аннотации в качестве золотого стандарта. Наши результаты во всех показателях оценки явно демонстрируют, что подходы на основе терминов намного превосходят тематические модели на основе стандартного документа.

ТЕМАТИЧЕСКИЕ МОДЕЛИ НА ОСНОВЕ ДОКУМЕНТА

Тематические модели – семейство методов на основе латентной переменной, предназначенных определять интересные модели во встречаемостях терминов по всему документу. Большинство тематических моделей берут за отправную точку стандартную модель векторного пространства (Vector space model, VSM, т.е. матрицу термин-документ, взвешенную некоторой подходящей схемой взвешивания терминов, такой как TF-IDF). Затем это пространство термин-документ факторизируется на низкоразмерное представление, в котором размеры интерпретируются как темы. Это позволяет и терминам, и документам быть описанными как распределения относительно тем, и соответственно темам – как распределениям относительно терминов и документов. Выбор метода факторизации является основным выбором разработки, когда он касается тематического моделирования. Общие подходы включают разложение по сингулярным числам (Singular Value Decomposition – SVD) [1], факторизацию неотрицательных матриц (Non-negative Matrix Factorization, NMF) [2], латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) [3] и более современные сложные многопараметричекие программы [4, 5].

Несмотря на выбор метода факторизации, все тематические на основе документа модели опираются на предположение, что латентные взаимосвязи между терминами полагаются на тематическое разнообразие в пространстве документов. Это предположение ясно предстает в виде порождаемой истории, переданной с помощью моделей, таких как pLSA [6] или LDA, которая словесно описывает субъективный выбор (ряда) тем для беседы, и для каждой темы выбирается ряд репрезентативных терминов. Эта история несет в себе интуитивный смысл, но отметим, что понятие документа является полностью случайным по отношению к истории; оно входит в историю только как единица текста, являющегося результатом субъекта. Мы утверждаем, что понятие документа является необязательным ограничением для тематических моделей, лимитирующих применение такого рода моделей к данным с собственным форматированием, и что тематические взаимосвязи терминов можно гораздо лучше смоделировать непосредственно в пространстве терминов.

ОТ ПРОСТРАНСТВА ДОКУМЕНТОВ К ПРОСТРАНСТВУ ТЕРМИНОВ

Таким образом мы предлагаем полностью сфокусироваться на пространстве терминов и совершенно отказаться от зависимости относительно понятия документов. Вместо того, чтобы строить векторы терминов для каждого *документа* в данных (т.е. стандартную VSM), мы построим векторные представления слов для всех *терминов* в данных из больших контекстных окон протяженностью примерно в 50 токенов*. Использование таких широких контекстных окон гарантирует, что векторные представления имеют возможность кодирования более широкой, и таким образом, более тематической, контекстуальной информации.

Существует много способов построения векторных представлений слов. В эту статью включены четыре различных подхода:

- Матрица совместной встречаемости (Co-occurrence matrix, COOC), стандартная матрица термин-термин, которая взвешивает подсчеты встречаемости, собранные внутри скользящего контекстного окна, с помощью положительной точечной взаимной информации [7];
- Случайное индексирование (Random Indexing, RI), возрастающая случайная перспективная технология, аккумулирующая векторные представления слов путем сложения векторов случайного индексирования для всех слов в их контекстах [8];
- Word2Vec (W2V), упрощенная многопараметрическая программа, распознающая векторные представления с использованием языка моделирования реальности [9];
- Doc2Vec (D2V), упрощенная многопараметрическая программа, использующая ту же самую архитектуру, что и Word2Vec, но которая умеет предсказывать идентификаторы документов, а не слова [10].

Эти методы имеют свои относительные достоинства и недостатки. Подход COOC прост и направлен, но размерность векторных представлений эквивалентна размеру словаря, который может стать неиспользуемым на больших множествах. RI решает проблему размерности, так как оно использует векторы фиксированного размера, но с дополнительным шумом. Word2Vec широко признана эффективной и точной, но требует дополнительного числа проверочных данных. С другой стороны, Doc2Vec первоначально разрабатывалась для применения к обработке документов, что делает ее интересным кандидатом для более тематически ориентированного применения, такого, какое представлено здесь.

Для каждого окончательного векторного представления слов вычисляем матрицу сходства, содержащую парные сходства между всеми векторами терминов в пространстве векторных представлений. Упрощаем матрицу сходства, удалив из нее менее ценные объекты, создающие для нас разреженное пространство сходства для работы с ним. Это выгодно с вычислительной точки зрения, а также устраняет шум от представлений. Чтобы извлечь темы из пространства сходства, нужно применить любой тип алгоритма, определяющий кластеры

* Размер контекстного окна, безусловно, является параметром, который можно смоделировать и оптимизировать под определенные данные и сценарий анализа. Берем по умолчанию 50 токенов (реализаций) в этих экспериментах и признаем, что применение иного параметра может привести к другим результатам.

или латентные переменные*. В нашем случае выбран ряд методов простой факторизации, включающих:

- Разложение по сингулярным числам (SVD) [11],
- Факторизация неотрицательных матриц (NFM) [2],
- Изучение словаря (Dictionary Learning. DL) [12].

Для каждого из этих методов факторизации извлекаем *n* компоненты (где *n* по умолчанию составляет 10) тем же самым способом, что и для стандартной тематической модели. В экспериментах, приведенных в разделах «Документы против терминов», «Методы факторизации» и «Измерение темы», все результаты представляют собой среднее относительно 10 просмотров разных методов факторизации.

СВЯЗАННЫЕ РАБОТЫ

Есть ряд предыдущих исследований, изучающих использование представлений на основе термина в тематическом моделировании. Одним из примеров является работа авторов [13], которые также основывают свое решение на матрице термин-термин, но их матрица термин-термин является не матрицей встречаемости, а матрицей *корреляции*, созданной из стандартной матрицы термин-документ. Поэтому их модель все еще полагается на данные, форматированные вручную в когерентную документальную структуру. Наоборот, рассматриваемые нами модели не накладывают каких-либо ограничений на форматирование данных, в то же время принимая более строгое определение тематических взаимосвязей в форме расширения контекстных окон, (в нашем случае) – 50 терминов, которое, как правило, значительно меньше и таким образом более точно, чем весь документ.

Другой пример – работа автора [14], который кластеризует векторные представления слов (созданные с помощью Word2Vec), используя Гауссову смесь распределений (Gaussian Mixture Model, GMM). Это предшествующая работа, которая описанными в ней подходами близка к рассматриваемым нами, но есть ряд важных различий. Изучаем диапазон способов векторных представлений слов, используем более широкое контекстуальное окно (50 терминов, а не 11-17) и применяем ряд методов стандартной факторизации вместо GMM, чтобы извлечь кластеры терминов. Несмотря на эти различия, мы считаем работу автора [14] важным стимулом нашего исследования.

Еще одним стимулом для нашей работы является исследование авторов [15], которое объединяет матрицу сходства слов со стандартной моделью NFM на основе документа. Матрица сходства слов строится при помощи модели Skipgram из Word2Vec и используется как дополнительный термин в алгоритме блочного покоординатного спуска, применяемом в решении NFM. Данный подход, удачно названный Факторизацией неотрициальных матриц с поддержкой семантики (Semantics Assisted NFM, SeaNMF), первоначально разработан для данных из коротких документов, в их случае размер контекстных окон, используемых в векторных представлениях Skipgram, равняется длине документов в данных. Авторы [15] утверждают, что разреженность матрицы сходства слов очень выгодна для эффективности этой модели, следовательно, то же самое преимущество применяется и к нашему случаю. Наиболее важное различие между моделью SeaNMF и рассматриваемыми нами подходами состоит в том, что последние полагаются *только* на матрицу сходства слов и таким образом совсем не используют матрицу термин-документ.

В отличие от предыдущих исследований сфокусируемся на общей идее использования векторных представлений слов, а не на SVM как основе для тематического моделирования и сравним ряд разных методов факторизации. Также используем более широкий ряд методов оценки и введем новый ряд мер, лучше соответствующих практическому применению тематических моделей.

МЕТОДЫ ОЦЕНКИ

Поскольку (наибольшее) практическое использование тематических моделей фокусируется только на окончательных списках терминов, это должно стать центром внимания нашей оценки. Проблема здесь состоит в том, что определение качества списков терминов может быть известной субъективной задачей, сравнимой (в шутку) с гаданием на кофейной гуще [16]. Были попытки прийти к более объективным измерениям оценки в тематических моделях, которые, как правило, принимают форму использования разных форм измерения внутренней информации, таких как энтропия, дилемма или когерентность [17-20]. Однако такие теоретические измерения информации не всегда коррелируют с семантической интерпретабельностью, как отмечают авторы [16], и если это так, то неясно, почему *семантическая* интерпретабельность должна коррелировать с *тематической* когерентностью.

В свете этих трудностей несколько заметным представляется то, что аннотации золотого стандарта темы, как правило, не используются в качестве стандартного показателя оценки тематических моделей. Безусловно, вероятнее всего мы не сможем найти такие аннотации для отдельных терминов или списков терминов, но можем найти их на уровне текста. Даже если тематическое моделирование существенно отличается от категоризации или кластеризации текста, то все еще можно использовать категории текста как цели в оценке тематических моделей с учетом того, что категории являются тематическими по сути. То есть, если мы можем найти массив текстов, в котором текст вручную маркируется по одной теме или более, то можно просто сравнить эту золотую стандартную тематическую задачу с тем, что создано тематической моделью.

Один из простых способов сделать это (сравнение), который также дает картину того, как человек-аналитик может использовать результат тематической модели в практическом анализе сценария, состоит в том, чтобы собрать все документы, охваченные каждой темой (т.е. в которой встречается один или более терминов в списке тем), и затем подсчитать перекрытие между этим множеством документов и множеством документов, названных по тематическим категориям в золотом стандарте. Это приведет к доле перекрытия между тематической моделью и золотым стандартом. Утверждаем, что это простой и перспективный способ оценить тематические модели, которые непосредственно отображают применимость на практике. В табл. 2 - 5 назовем этот показатель «Истиной».

* Наши первоначальные эксперименты включали стандартные методы кластеризации, такие как k-means, агломеративную кластеризацию и методы на основе плотности, но при использовании кластеризации не было обнаружено никакого значительного улучшения по сравнению с факторизацией.

Человек-аналитик может также интересоваться и другими приведенными ниже факторами:

• Перекрытие: сколько тем перекрывается? Определим это количество как долю идентичных терминов в темах; чем их меньше, тем, предположительно, это лучше с точки зрения аналитика.

• Охват: сколько данных охватывают темы? Аналитик может отдать предпочтение такому решению – определим это количество как долю текстов, содержащих термины в темах; подходит ли это к большим или малым охватам зависит от сценария анализа.

• Уникальность: как часто термины из разных тем совместно встречаются в одном и том же документе? Если мы хотим низкий охват данных (т.е. небольшие и фокусные темы), мы вероятно будем страдать от более высокой уникальности тем, тогда как преследуя цель большого охвата данных, нам надо ожидать более низкой уникальности тем.

• Отделение: насколько векторные представления отличаются между темами? Это измеряется как среднее различие между сходством по косинусу терминов *внутри* темы и сходством по косинусу терминов *между* темами.

• Время: сколько времени занимает процесс факторизации (т.е. заключить тему в) пространства сходства? Для воспроизводства и сравнения используем функции факторизации из scikit-learn (https://scikit-learn.org/) по возможности с установкой по умолчанию.

Также включаем измерения когерентности UCI [18] и UMASS [19] в качестве сравнения. Эти меры суммируют значения PMI (точечной взаимной информации, Pointwise mutual information – PMI) всех пар слов в темах; мера UMASS подсчитывает совместные встречаемости внутри целого документа, тогда как мера UCI определяет границы совместной встречаемости слов внутри скользящего окна:

$$UMASS\left(w_i, w_j, \in\right) = \log \frac{D\left(w_i, w_j\right) + \in}{D\left(w_j\right)} \qquad (1)$$

$$UCI\left(w_i, w_j, \in\right) = \log \frac{p\left(w_i, w_j\right) + \in}{p\left(w_i\right) p\left(w_j\right)} \qquad (2)$$

Придерживаясь мнения авторов [21], устанавливаем $\in < 1$, в нашем случае $\in = 0{,}001$.

ЭКСПЕРИМЕНТЫ

Следующие эксперименты используют ряд разных массивов данных, взятых из двух разных источников. Первым источником данных является Swedish news, собранный вручную и аннотированый по темам людьми-экспертами. Этот массив данных хорошо соответствует сценарию анализа реального мира. Однако поскольку источник Swedish news – относительно небольшой и публично недоступен*, также создаем ряд искусственно аннотированных массивов данных на английском языке на основе англоязычной Википедии (English Wikipedia). Различные массивы данных подробно представлены в табл. 1 и в последующих разделах статьи.

Все эксперименты в статье обрабатываются на машине Intel Xeon E5-2620 2,40 Гц центрального процессора и 192 Гб оперативной памяти. Все техники факторизации осуществляются на стандартных установках и применениях в версии 0.20.0 метода scikit-learn. Используем применения метода Gensim (https://radimrehurek.com/gensim/) из Word2Vec и Doc2Vec со стандартными параметрами установки и внутренние применения Python из COOC и RI. Применение RI доступно на сайте: https://ghetto.sics.se/mange/ri.

Данные на основе Swedish News

Шведский массив данных состоит из новостных статей, собранных из ведущих шведских газет (Svenska dagbladet, Dagens nyheter, Aftonbladet, Expressen) авторами [22]. Каждая новостная статья аннотировалась вручную по нескольким разным категориям экспертами отделения журналистики, медиа и коммуникации Университета г. Гетеборга. Была использована категория HuvudÄmne (по-английски *main topic*, основная тема) как золотой стандарт названия, так как она явно представляет основную тему новостной статьи. На практике полезная тематическая модель должна быть способной минимально идентифицировать эти 34 различные основные темы из данных. Данные содержат 895 новостных статей с общим числом токенов – 366 456. Средняя длина документов составляет около 400 терминов, с очень высокой вариативностью. Для данных шведского массива игнорируются термины с частотой меньше 5.

Таблица 1

Массивы данных, используемые в экспериментах

Данные	#Тексты	#Токены	#Типы	#Темы	Минимальная частота
Swedish News	895	366 456	33 358	34	5
English Wikipedia	100 000	14 784 214	269 741	40 109	10
English Wikipedia (небольшие темы)	213 656	30 873 801	273 056	125 397	20
English Wikipedia (средние темы)	112 653	16 316 965	173 509	11 194	10
English Wikipedia (большие темы)	1 273	196 378	13 398	20	5

*Данные можно получить, контактируя с авторами исследования Swedish news, [22].

Таблица 2

Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
VSM	NMF	-7,72	58,73	0,17	1,00	0,11	0,21	0,18	123,20
	LDA	**-7,24**	53,37	0,36	1,00	0,11	0,17	0,19	202,44
COOC	NMF	-9,92	95,50	0,08	1,00	0,28	0,35	0,28	356,54
RI	NMF	-8,46	145,74	0,03	0,74	0,91	0,31	**0,34**	361,56
W2V	NMF	-10,41	**159,60**	0,00	0,26	0,92	0,37	0,31	468,50
D2V	NMF	-15,54	146,11	0,00	0,55	0,79	**0,45**	0,23	477,92

Примечание: Результаты для массива данных Swedish News по различным векторным представлениям (VSM, COOC, RI, Word2Vec, Doc2Vec) относительно 7 разных показателей оценки, включая оценки когерентности UMASS и UCI, тематическое перекрытие, тематический охват, уникальность, отделение и перекрытие с истиной. Также приводится время (сек.) обработки для каждой факторизации. Все оценки представляют собой среднее относительно 10 реализаций.

Таблица 3

Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
	NMF	-9,07	56,44	0,16	1,00	0,10	0,13	0,00	14,462,20
VSM	LDA	-2,31	58,55	0,34	1,00	0,12	0,15	0,01	12,399,78
COOC	NMF	1,62	152,08	0,00	1,00	0,95	0,27	0,03	8,895,03
RI	NMF	**11,55**	178,86	0,00	0,74	0,95	0,11	**0,06**	10,169,69
W2V	NMF	-5,67	141,64	0,00	0,57	0,86	**0,50**	0,01	12,098,17
D2V	NMF	9,63	**185,88**	0,00	0,14	0,97	0,44	0,02	9,571,91

Примечание: Результаты для массива данных English Wikipedia по различным векторным представлениям (VSM, COOC, RI, Word2Vec, Doc2Vec) относительно 7 разных показателей оценки, включая оценки когерентности UMASS и UCI, тематическое перекрытие, тематический охват, уникальность, отделение и перекрытие с истиной. Также приводится время (сек.) обработки для каждой факторизации. Все оценки представляют собой среднее относительно 10 реализаций.

Данные на основе Wikipedia

Так как массив данных Swedish News сравнительно небольшой и публично недоступен, также включаем ряд более крупных массивов данных на основе случайных выборок статей из English Wikipedia. Выборки созданы случайным выбором текста параграфов из Wikipedia и использованием названий векторных представлений Wikipedia как названия темы для текста. Два примера таких тем – «Изменение климата в Финляндии» и «Майк Тайсон против Мишеля Спинкса». Также используем вероятностную выборочную стратегию, создающую в среднем 20 текстовых выборок по теме, со стандартным отклонением, равным примерно 10, и минимальным числом выборок – около 5.

Как видно в табл. 1, создано 4 разных массива данных на основе этой стратегии[*]. Первый содержит 100 000 текстов с общим числом 14 784 119 униграмм терминов. Средняя длина документа составляет примерно 150 терминов со стандартным отклонением около 50 (самый длинный документ содержит примерно 1 тыс. терминов, а самый короткий – 50). Чтобы иметь возможность изучать влияние размера темы на тематические модели, создаем три разных массива данных с разнообразным числом текстов на тему. Создаем данные для небольших, среднего размера и больших тем, в которых маленькими темами являются такие, которые имеют 5 и меньше текстов по теме, большие темы – 50 и более текстов на тему, а попадающие между ними считаются среднего размера. Это приводит к 125 397 небольшим темам, содержащим 30 873 748 токенов, 11 194 темам среднего размера с наличием 16 316 954 токенов и 20 большим темам, содержащим 196 378 токенов. Используем минимальную частоту порога (10 встречаемостей) для английских данных, исключение составляют данные небольших тем, где мы вместо этого устанавливаем минимальную частоту порога – 20 встречаемостей, и данные больших тем, в которых используем в качестве порога 5 встречаемостей.

Документы против терминов

В первой группе экспериментов сравниваем тематические модели на основе документа с моделями на основе термина. Включаем две разных модели на основе документа – NMF и LDA,[*] обе применяются к стандартной VSM с помощью взвешивания TF-IDF. Сравниваем эти основные модели с четырьмя разными моделями на основе термина, которые используют NMF как метод факторизации[**]; стандартная матрица совместной встречаемости взвешивается с помощью PPMI (COOC), Random Indexing (RI), Word2Vec (W2V) и Doc2Vec (D2V).

Табл. 2 отражает результаты данных массива Swedish News. Базовые модели на основе документа получают более высокие оценки измерения UMASS на основе документа, но значительно более низкие оценки изме-

[*] Массивы данныхWikipedia можно скачать через ссылку: https://bit.ly/33hhyiQ

[*] Используем NMF и LDA, поскольку они являются наиболее общими методами факторизации, применяемыми в стандартных тематических моделях.

[**]Используем здесь NMF, поскольку она относительно устойчива. Сравнение разных методов факторизации для моделей на основе термина представлено в табл. 4.

рения UCI на основе слова. Модели на основе документа имеют более высокое перекрытие между темами и они также охватывают больше данных, но за счет меньшего числа уникальных тематических задач. Модели на основе термина имеют более высокое среднее отделение между терминами внутри, а не вдоль тем, и они больше соответствуют тематическому распределению вручную; лучшей моделью относительно перекрытия с истинными названиями является RI, которое охватывает 34 % золотого стандарта.

Табл. 3 представляет результаты данных английского массива. В этом случае отмечаем, что модели на основе терминов значительно превосходят модели на основе документа не только по измерению UCI, но и по измерению UMASS, за исключением W2V, которая имеет более низкую оценку, чем основная модель VSM+LDA. Снова отметим, что модели на основе документов имеют более высокий охват тем там, где модели на основе терминов совсем не имеют ни одного перекрытия для данных английского массива. Отметим также, что модели на основе документов стремятся охватить больше данных, чем модели на основе терминов, и что модели на основе терминов имеют более уникальные тематические распределения. Модели на основе терминов также имеют более высокое среднее отделение между терминами внутри, а не вдоль тем, и они также стремятся лучше соответствовать аннотациям золотого стандарта – н отмечается очень низкое перекрытие для всех моделей по данным английского массива; наилучшей моделью в этом случае также является случайное индексирование, которое имеет перекрытие с составленными людьми аннотациями только на уровне 6 %.

МЕТОДЫ ФАКТОРИЗАЦИИ

Обращаемся к эффектам использования различных методов факторизации для разных представлений. Табл. 2 и 3 показывают, что разница между NMF и LDA для модели на основе документа заметно ощутима для более крупного английского массива данных, в котором LDA действует несколько лучше, чем NMF. Для небольшого шведского массива данных существенной разницы нет.

Табл. 4 демонстрирует эффекты использования различных методов факторизации с применением моделей на основе терминов. Включаем три разных метода факторизации для двух разных векторных представлений (COOC и W2V) в этих результатах. Заметим, что NMF ведет к лучшим результатам для обоих векторных представлений с участием данных шведского массива, но эти результаты являются более разнородными для данных английского массива. Что касается как векторных представлений COOC, так и W2V, то изучение словаря приводит к наилучшим измерениям UMASS, UCI. SVD ведет к наилучшему разделению внутри и вдоль тем для векторных представлений COOC, а NMF – для векторных представлений W2V. Изучение словаря (DL) приводит к лучшему перекрытию аннотаций тем с участием людей для COOC, тогда как нет различия в перекрытии между разными методами факторизации для векторных представлений W2V. SVD является самым быстрым методом с участием применения технологии scikit-learn.

Измерение темы

Поскольку темы, как правило, поступают в разных размерах, то представляется уместным задать вопрос, как различные модели обрабатывают разные по размеру темы. Как указывается в разделе «Данные на основе Wikipedia», мы используем три массива данных с темами разного размера; небольшие темы, охватывающие по меньшей мере 5 текстов каждая, большие темы – 50 текстов каждая, и темы среднего размера, охватывающие от 5 до 50 текстов каждая. Табл. 5 отражает результаты модели LDA на основе документа, векторные представления W2V на основе NMF, а также векторные представления RI на основе SVD. Включаем RI в этот пример, так как оно очень хорошо выполняется на небольших и среднего размера темах.

Таблица 4

Шведский массив									
Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
COOC	NMF	**-9,08**	**121,04**	0,00	1,00	0,72	**0,31**	**0,34**	204,13
	SVD	-11,24	101,69	0,03	1,00	0,30	0,28	0,30	**100,34**
	DL	-13,60	117,97	0,05	0,95	0,57	0,26	0,24	222,05
W2V	NMF	**-8,80**	**150,04**	0,00	0,28	0,97	**0,46**	**0,47**	244,76
	SVD	-9,03	139,70	0,00	1,00	0,88	0,38	0,29	**98,82**
	DL	-12,83	149,31	0,00	0,40	0,88	0,37	0,37	221,82
Английский массив									
Векторное представление	Model	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
COOC	NMF	2,16	152,68	0,00	1,00	0,93	0,26	0,02	8,167,09
	SVD	-4,89	117,62	0,02	1,00	0,31	**0,30**	0,01	**3,529,35**
	DL	**11,41**	**162,32**	0,07	0,55	0,95	0,19	**0,05**	3,818,89
W2V	NMF	-7,51	137,2	0,00	0,60	0,85	**0,49**	0,01	8,892,46
	SVD	-2,35	150,28	0,00	0,82	0,76	0,39	0,01	**4,965,02**
	DL	**1,13**	**162,53**	0,00	0,44	0,88	0,36	0,01	5,973,62

Примечание: Результаты использования различных методов факторизации (NMF, SVD и Dictionary Learning) для векторных представлений COOC и Word2Vec в массивах данных Swedish News (вверху) и English Wikipedia (внизу). Время обработки приводится в сек. (с участием применений в методе scikit-learn), и все оценки представляют собой среднее относительно 10 реализаций.

Таблица 5

Измерение темы	Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина
	VSM	LDA	4,93	102,12	0,13	1,00	0,18	0,27	0,00
Небольшое	W2V	NMF	0,63	161,66	0,00	0,17	0,92	**0,53**	0,00
	RI	SVD	**16,24**	**188,14**	0,00	0,38	0,93	0,09	**0,01**
	VSM	LDA	4,76	107,27	0,10	1,00	0,20	0,27	0,04
Среднее	W2V	NMF	2,84	168,03	0,00	0,25	0,89	**0,49**	0,01
	RI	SVD	**15,29**	**186,31**	0,00	0,31	0,85	0,11	**0,06**
	VSM	LDA	-10,77	90,33	0,06	1,00	0,23	0,28	0,31
Большое	W2V	NMF	**-9,93**	**136,60**	0,00	0,30	0,97	**0,54**	**0,88**
	RI	SVD	-10,15	104,26	0,01	1,00	0,56	0,28	0,41

Примечание: Эффективность модели LDA на основе документа, W2V на основе NMF и RI на основе SVD для данных с темами разного размера. Как и в табл. 2 - 4, все оценки представляют собой среднее относительно 10 реализаций.

Наиболее примечательным в этих результатах табл. 5 является то, что ни одна из моделей не работает хорошо относительно перекрытия с названиями золотого стандарта на небольших и среднего размера темах. Векторные представления RI с факторизацией SVD получают удивительно высокие оценки UMASS и UCI, и это единственная модель с каким-либо видимым перекрытием с истинными названиями для небольших тем (едва заметное 1% перекрытие), а также имеет наибольшее перекрытие для тем среднего размера (6%). Что касается больших тем, то все модели работают значительно лучше относительно перекрытия с истинными названиями; модель на основе документа имеет перекрытие в 31%, RI – в 41%, а W2V – очень высокое перекрытие в 88%.

ОБСУЖДЕНИЕ

Как очевидно из описанных в этой статье экспериментов, различные тематические модели имеет разные свойства, и правильный выбор тематической модели зависит от определенной информационной потребности сценария конкретного анализа. Даже если модели на основе терминов в целом превосходят стандартные модели на основе документа по всем данным и показателям, используемым в этой статье, все еще могут иметь место ситуации, в которых модель на основе документа будет более подходящей в использовании. Такой сценарий может быть в случае, если аналитику требуется решение с большим охватом данных; модели на основе документа стремятся привести к более высокому охвату данных, но здесь возможно перекрытие между темами, и распределение темы (подсчитывается как встречаемость тематических терминов в документах) является менее уникальным в сравнении с моделями на основе терминов.

С другой стороны, модели на основе терминов дают больше уникальных тем с меньшим перекрытием и лучшим отделением между темами. Модели на основе терминов также достигают более высоких значений во всех показателях оценки (UMASS и UCI когерентность, представление отделения и перекрытие с истинными названиями) с исключением в данных массива Swedish News, в котором модели на основе документа ведут к более высокой когерентности UMASS. В целом различие между моделями на основе документа и на основе терминов ниже при рассмотрении измерения UMASS, чем при изучении измерения UCI, которое может быть объяснено тем фактом, что первые используют документы в качестве единиц подсчета совместной встречаемости, а вторые – слова.

Заметим, имеется большое противоречие между реализациями, что затрудняет предоставление какого-либо определенного вывода относительно выбора оптимальной разработки тематической модели на основе терминов. Определенные методы факторизации кажутся более подходящими к определенным представлениям и определенным данным. В целом NMF, кажется, работает лучше в этих экспериментах для большинства векторных представлений слов в массиве данных Swedish News, а изучение словаря (DL) лучше работает в этих экспериментах для данных массива English Wikipedia. С другой стороны, если темы небольшие, то SVD, вероятно, работает лучше, в частности, для векторных представлений RI.

Что касается различных типов векторных представлений слов, то отмечаем, что модель COOC, как правило, приводит к самому высокому охвату данных, за ней следует RI, которое также стремится иметь наилучшее перекрытие с аннотациями золотого стандарта, сделанными людьми, за исключением случая больших тем, когда Word2Vec значительно лучше. Заметим, что и Word2Vec, и Doc2Vec имеют высокое среднее отделение терминов в отличие от тем, но добавление документальной информации в Doc2Vec не кажется полезным для вывода о теме.

Подчеркнем, что данные, используемые в этих экспериментах, содержат только одну тему в документе, тогда как многие другие сценарии тематического моделирования оперируют многими темами в документе. Мы не считаем, что это ограничение должно иметь какое-то влияние на общий характер наших результатов, так как модели на основе терминов в высшей степени применимы к сценариям со многими темами. Предложенное сравнение золотого стандарта также непосредственно применимо к данным со многими темами.

ЗАКЛЮЧЕНИЕ

Статья демонстрирует большую полезность просмотра вывода о теме в тематических моделях в целях выявления (скрытых) латентных факторов в пространстве терминов, чем в пространстве документов. Предлагается простая модель на основе терминов, использующая

стандартные векторные представления слов с участием методов стандартной факторизации. Несмотря на их простоту, такие модели на основе терминов превосходят все тестируемые модели на основе документа по всем показателям оценки, используемым в статье. Также предлагается задача тематической категоризации, применяющая тематические аннотации золотого стандарта, а также ряд других показателей, которые могут лучше отвечать анализу сценариев реального мира, чем тип внутренних измерений, широко используемый в литературе по тематическим моделям. Использование этих дополнительных измерений стимулирует нас характеризовать различные свойства тематических моделей, а также сделать сознательный выбор разработки тематической модели для определенных информационных потребностей.

Наши эксперименты демонстрируют, что оптимальная модель вероятнее всего должна быть ориентированной на данные и конкретную задачу, и что оптимальный выбор определенных представлений и методов факторизации очевидно будет различаться от случая к случаю. Тем не менее, в качестве надежной основы предлагаем использовать Word2Vec представления с участием факторизации NMF.

Делаем вывод, что модели на основе терминов являются конкурентными, если не превосходящими, в сравнении с традиционными моделями на основе документа, с рядом дополнительных выгод, включая независимость документального форматирования и относительную устойчивость к размеру темы. Хотя изучаемые в этой статье модели превосходят модели на основе документов по всем показателям, считаем наш подход на основе терминов простой основной моделью с большими возможностями в улучшении.

Благодарность. Данная работа была выполнена при частичной поддержке Шведского научного агентства FOI и Шведского научного совета (грант 2017-02429, Лингвистические изучения обществ). Автор выражает признательность Магнусу Роселлю (Шведское научное агентство FOI), Йоханнесу Йохансонну (Отделение журналистики, медиа и коммуникаций, Университет г. Гетеборга), а также Стефану Дальбергу (Отделение гуманитарных и общественных наук, Шведский университет, расположенный в центре Швеции) за участие в дискуссиях по статье. Автор особо благодарит Бенгта Йохансонна (Отделение журналистики, медиа и коммуникаций, Университет г. Гетеборга) за предоставление доступа к аннотированному массиву данных Swedish News.

ЛИТЕРАТУРА

1. *Deerwester S., Dumais S. T., Furnas G. T., Landauer T. K., Harshman R.* Indexing by latent semantic analysis// Journal of the American Society for Information Science. — 1990. — Vol. 41, No. 6. — P. 391–407.

2. *Lee D.D., Seung H.S.* Algorithms for non-negative matrix factorization/ T. K. Leen, T. G. Dietterich, and V. Tresp (eds.), Advances in Neural Information Processing Systems 13, pages 556–562. — MIT Press, 2001.

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

4. *Cao Z., Li S., Liu Y., Li W., Ji H.* A novel neural topic model and its supervised extension// AAAI Conference on Artificial Intelligence. — 2015.

5. *Miao Y., Grefenstette E., Blunsom P.* Discovering discrete latent topics with neural variational inference//Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 2410–2419. — JMLR.org., 2017.

6. *Hofmann T.* Probabilistic latent semantic analysis// Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, page 289–296, San Francisco, CA, USA. — Morgan Kaufmann Publishers Inc., 1999.

7. *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings// Transactions of the Association for Computational Linguistics. — 2015. — Vol. 3. — P. 211–225.

8. *Sahlgren M.* An introduction to random indexing// Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE), Copenhagen, Denmark. — 2005.

9. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space// Proceedings of International Conference on Learning Representations (ICLR). —2013.

10. *Le Q., Mikolov T.* Distributed representations of sentences and documents//Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32, ICML'14, page II–1188–II–1196. — JMLR.org., 2014.

11. *Golub G. H., Van Loan C. F.* Matrix Computations, third edition. — The Johns Hopkins University Press, 1996.

12. *Mairal J., Bach F., Ponce J., Sapiro G.* Online dictionary learning for sparse coding// Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 689–696, New York, NY, USA. — Association for Computing Machinery, 2009.

13. *Arora S., Ge R., Halpern Y., Mimno D., Moitra A., Sontag D., Wu Y., Zhu M.* A practical algorithm for topic modeling with provable guarantees//Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13), pages II–280–II–288. — JMLR.org., 2013.

14. *Sridhar V.K. R.* Unsupervised topic modeling for short texts using distributed representations of words// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 192–200. — Association for Computational Linguistics, 2015.

15. *Shi T., Kang K., Choo J., Reddy C. K.* Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations //Proceedings of the 2018 World Wide Web Conference (WWW'18), pages 1105–1114, Republic and Canton of Geneva, Switzerland. — International World Wide Web Conferences Steering Committee, 2018.

16. *Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. M.* Reading tea leaves: How humans interpret topic models // Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09, pages 288–296, USA. — Curran Associates Inc., 2015.

17. *Wallach H. M., Murray I., Salakhutdinov R., Mimno D.* Evaluation methods for topic models//Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), pages 1105–1112, New York, NY, USA. — ACM, 2009.

18. *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries//Proceedings of

the 10th Annual Joint Conference on Digital Libraries, JCDL '10, page 215–224, New York, NY, USA. — Association for Computing Machinery, 2010.

19. *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models//Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, page 262–272, USA. — Association for Computational Linguistics, 2011.

20. *Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D.* Exploring topic coherence over many models and many topics// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL '12, pages 952–961, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2012.

21. *Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D.* Exploring topic coherence over many models and many topics//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952–961, Jeju Island, Korea. — Association for Computational Linguistics, 2012.

22. *Johansson B., Strömbäck J.* Kampen om mediebilden: nyhetsjournalistik i valrörelsen 2018.— Institutet för Mediestudier, 2019.

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].

DOI: 10.36535/0203-6460-2021-01-4

Сходство документов на основе аспекта на примере научных статей*

Мальте ОСТЕНДОРФ
(Malte OSTENDORFF),

ГЕОРГ РЕМ
(Georg REHM)

Немецкий научно-иссследовательский центр по искусственному интеллекту, г. Берлин, Германия

Терри РУАС
(**Terry RUAS**)

Вуппертальский университет, г. Вупперталь, Германия

Тилль БЛЮМЕ
(**Till BLUME**)

Кильский университет, г. Киль, Германия

Бела ГИПП
(**Bela GIPP**)

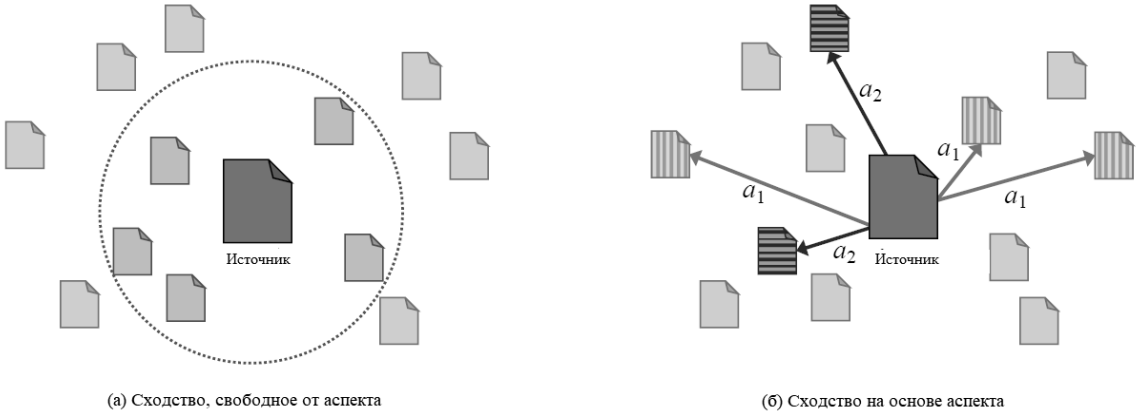
Университет г. Констанц, г. Констанц, Германия

Традиционные измерения сходства документов обеспечивают крупномодульное разграничение между схожими и несхожими документами. Обычно эти измерения не рассматривают в каких аспектах два документа являются схожими. Это ограничивает степень структурирования прикладных задач, таких как рекомендательные системы, которые полагаются на сходство документов. В статье понятие сходства расширяется аспектом информации через выполнение задачи классификации пар документов. Оценивается сходство документов на основе аспекта на примере научных публикаций. Ссылки в статьях отражают сходство по аспекту, например, часть названия, в котором встречается ссылка, выполняет функции категории для пары цитирующей и цитируемой статьи. Использовался ряд вариаций моделей Transformer, таких как ROBERTa, ELECTRA, XLNet и BERT, и они сравнивались с ведущей моделью LSTM. Наши эксперименты проводились на двух недавно созданных наборах данных, подсчитывающих 172 073 научные статьи из собраний ACL Anthology и CORD-19. Относительно выполнения результаты определяют в качестве лучшей систему SciBERT. Качественное исследование обосновывает наши количественные результаты. Выводы стимулируют проведение дальнейших исследований сходства документов на основе аспекта и разработку рекомендательных систем на основе оценки технологий. Наборы данных, коды и подготовленные модели являются публично доступными.

ВВЕДЕНИЕ

Рекомендательные системы (РС) помогают ученым в поиске релевантных статей для их работы. Когда обратная связь от пользователя осуществляется редко или недоступна, то применяются подходы на основе контента и соответствующие измерения сходства документов [1]. Рекомендательные системы советуют документ-кандидат в зависимости от его сходства или несходства по отношению к документу-источнику. Эта крупномодульная оценка сходства (похож или непохож) отрицает многие фасеты, способные сделать два документа схожими. Относительно общего понятия сходства авторы работ [2, 3] даже утверждают, что сходство является плохо определяемым понятием, если нельзя утверждать к какому аспекту относится сходство. В РС для научных статей сходство часто связано с множеством фасетов представленного исследования, например, метод, полученные данные [4]. Учитывая, что сходство документов может дифференцировать аспекты исследования, есть возможность получить определенные смоделированные рекомендации. Например, разрешается рекомендовать статью со схожими методами, но различными полученными данными.

* Перевод Ostendorff M., Ruas T., Blume T., Gipp B., Rehm G. Aspect-based document similarity for research papers. — https://arxiv.org/pdf/2010.06395.pdf



(а) Сходство, свободное от аспекта

(б) Сходство на основе аспекта

Рис. 1. Большинство РС полагается на измерения сходства между источником и *k* большинтсвом схожих целевых документов (а). Это отрицает аспекты, по которым два и более документов могут быть схожими. В сходстве документов на основе аспекта (б) документы объединены в соответствии с их внутренними, связанными с ними аспектами (a_1 и a_2).

Таким образом рекомендательная система позволяет облегчать обнаружение аналогий в научной литературе [5]. Описываем соответствующее сходство множественного аспекта в научных статьях как *сходство документов на основе аспекта.* На рис. 1 изображено сходство на основе аспекта в отличие от сходства, свободного от аспекта (традиционное). Следуя примеру научной статьи, аспект a_1 относится к полученным данным, а аспект a_2 – к методу (рис. 1б).

В предыдущей работе [6] мы предлагаем вывести аспект сходства документов, формулируя проблему как многоклассовую классификацию пар документов. В данной статье мы расширяем нашу предыдущую работу до многокатегорийного сценария и сосредотачиваемся на научной литературе, а не на общей (статьи в Wikipedia). Подобно авторам работ [7, 8] используем ссылки как учебные сигналы. Вместо использования ссылок для бинарной классификации (т.е. схожий или несхожий документ) мы включаем название раздела, в котором фигурирует ссылка, как категорию для пары документов. Названия разделов в ссылках описывают сходство по аспекту цитирующего и цитируемого документа. Наши наборы данных берутся из ACL Anthology [9] и CORD-19 [10].

В итоге наш вклад состоит в следующем: (1) расширение традиционного сходства документов до основанного на аспекте в задаче многокатегорийной многоклассовой классификации документа; (2) показ того, что сходство документов на основе аспекта хорошо подходит для научных статей; (3) оценка шести моделей Transformer и основной для задачи классификации пар документов; (4) публикация наших кодов источников, подготовленных моделей и двух наборов данных из областей компьютерной лингвистики и биомедицины для усиления дальнейшего исследования.

СВЯЗАННЫЕ РАБОТЫ

Далее обсуждается работа авторов по сходству текста, рекомендации и применения Transformer.

Авторы [3] обсуждают понятие сходства как часто плохо определяемого в литературе и используемого в качестве «совокупности терминов, охватывающих довольно разные явления». Эти авторы [3] также формализуют то, чем является сходство текста, и предполагают, что контент, структура и стиль являются основными измерениями, присущими тексту. Что касается рекомендации литературы, то информация о контенте и пользователе является самым распространенным измерением для рассмотрения [1].

Ряд авторов [5] изучает сходство документа на основе аспекта как задачу сегментации, а не классификации. Они (авторы) делят аннотации совместных работ и работ по вычислениям на четыре класса в зависимости от их научного аспекта: описание, цель, метод и полученные данные. Сходство по косинусу, вычисленное на сегменте презентаций, позволяет провести поиск сходных статей по отдельному аспекту. Авторы работы [4] применяют тот же подход сегментации к массиву CORD-19 [10]. Совместная работа авторов [11] придерживается аналогичного подхода к рекомендациям ссылок. Эти авторы классифицируют разделы по дискурсам фасетов и строят векторы документа для каждого фасета. Однако сегментация является сверхоптимальной альтернативой, так как она нарушает когерентность документов. Что касается классификации пар документов, то сходство на основе аспекта происходит без нарушений когерентности документа.

Наши эксперименты изучают языковые модели Transformer [12]. Модели BERT [13], ROBERTa [14], XLNet [15] и ELECTRA [16] улучшают многие задачи NLP (Natural Language Processing - Обработка естественного языка), например, вывод естественного языка [17, 18] и семантическое сходство текстов [19]. Авторы работы [20] показывают, как модели BERT можно объединять в сеть Siamese [21] для создания векторных представлений, подходящих для сравнения друг с другом при помощи сходства по косинусу. Ряд исследований [22, 23] анализирует модели BERT, чтобы классифицировать одиночные документы в соответствии с восприятием или темой. Исследования авторов [8, 24] рассматривают предназначенные для определенной области модели Transformer по отношению к задачам NLP в научных документах.

Более того, ученые [8] являются первыми, кто использовал модели Transformer для кодирования названий и аннотаций статей в целях создания рекомендаций. Авторы работы [25] используют модели BERT для рекомендательных систем, но кодируют только названия статей. Иные недавние рекомендательные системы полагаются на другие технологии, такие как анализ совместного цитирования, TF-IDF или Paragraph Vectors [26, 27].

В предыдущей работе [6] мы моделируем сходство на основе аспекта как задачу многоклассовой классификации пар документов. Используем края интеллектуального графа из Wikidata как аспект информации сходства статей в Wikipedia. Применяемое определение задачи допускает только монокатегорийнную классификацию. Для научных статей это определение не вполне подходит. Две статьи могут быть схожи во многих аспектах. Соответственно мы ставим задачу многоклассовой классификации и расширяем ее до многокатегорийной.

Для наших экспериментов мы воспользуемся ссылками и разделом названий, в которых встречаются ссылки, как категориями классификации. Авторы [28] показывают соответствующий подход в контексте связывания объектов. Они утверждают, что во многих ситуациях ссылка на объект предлагает относительно крупномодульную семантическую информацию. Чтобы учитывать различные аспекты, в которых упоминается объект, эти авторы [28] связывают объекты не только с их соответвующими статьями в Wikipedia, но и с разделами, представляющими различные аспекты.

Что касается сходства на уровне сегмента и парной многоклассовой монокатегорийной классификации, то первоначальные подходы, изучающие сходство на основе аспекта, являются доступными. В частности, модели Transformer, кажется, обещают успешное решение задач относительно сходства, классификации и иных соответствующих вопросов.

ЭКСПЕРИМЕНТЫ

Представляем нашу методологию (рис. 2) классификации сходства научных статей на основе аспекта.

Наборы данных

Генерация аннотированных людьми данных для рекомендации научных статей затратна и ограничена небольшим количеством [1]. Набор данных небольшого размера мешает внедрению обучающих алгоритмов. Чтобы избежать проблемы нехватки данных, ученые полагаются на ссылки как на основу истины, т.е. когда ссылка существует между двумя статьями, обе статьи считаются схожими [7;8]. Либо ссылка существует, либо не соответствует категории в бинарной классификации. Для создания сходства на основе аспекта мы переносим эту идею в проблему многокатегорийной многоклассовой классификации. В качестве основы истины адаптируем название раздела, в котором ссылка из статьи A (источник) на B (цель) встречается как название класса (рис. 2a). Эта классификация является многоклассовой из-за множества названий разделов, а также многокатегорийной, так как статья A может цитироваться в нескольких разделах. Например, статья A, цитирующая B, в разделе Введение (Introduction) и Обсуждение (Discussion) должна соотноситься с одной выборкой набора данных.

ACL Anthology. Принимаем библиографический массив ACL Anthology [9] как набор данных. Он содержит 22 878 научных статей по вычислительной лингвистике. Помимо полных текстов массив ACL Anthology предоставляет дополнительные данные по цитированию. Ссылки аннотируются с помощью названия раздела, в котором расположены маркеры ссылок. Эта информация востребована в наших экспериментах.

CORD-19. Открытый набор научных данных по COVID-19 (CORD-19) – собрание статей по COVID-19 и относящихся к коронавирусу исследованиям из нескольких биомедицинских цифровых библиотек [10]. Ссылки и метаданные всех статей CORD-19 стандартизированы в соответствии с регулярной обработкой авторов [29]. Ссылки в CORD-19 также аннотируются с помощью названий разделов.

Предварительная обработка данных

Рассматривая ACL Anthology и CORD-19, получаем два набора данных для парной многокатегорийной многоклассовой классификации. Названия разделов из ссылок, т.е. названия классов, представлены в табл. 1. Нормализуем названия разделов (lowercase, letters-only, singular to plural) и разложим составные разделы на много простых – Conclusion and Future Work (Заключение и Дальнейшая работа) на Conclusion; Future Work (Заключение; Дальнейшая работа). Сделаем запрос в прикладной программный интерфейс DBLP [30] и Semantic Scholar [29] для соотнесения ссылок и извлечем недостающую информацию из статей, например, рефератов. Также удалим необоснованные статьи без текста или дублирующиеся. Разделим оба набора данных ACL Anthology и CORD-19 на десять классов в соотвествии с их числом выборок, посредством которых первые девять содержат наиболее популярные названия разделов, а в десятую (Прочее) сгруппируем оставшиеся. Даже если решение на основе наших десяти классов может отрицать вариации названий разделов в литературе, наша модель все еще дублирует ряд определенных авторами [4 и 5] аспектов исследования. Итоговое распределение классов является несбалансированным, но отражает истинную природу совокупностей, как показыает табл. 5. Подлинники для воспроизведения массивов данных доступны при наличии нашего кода источника.

Негативная выборка

Помимо десяти положительных классов (табл. 1) введем класс, названный *None*, который будет выступать в роли отрицательного оппонента наших положительных выборок в той же пропорции [31]. Пара документов класса *None* являются случайно выбранными и непохожими. Случайная пара статей является негативной выборкой, когда статьи не сущестуют как положительная пара, не являются совместно цитируемыми, не объединены авторами и не опубликованы в одном и том же номере. Получаем 24 275 негативных выборок для набора данных ACL Anthology и 33 083 для набора данных CORD-19.Эти выборки позволят различать модели между схожими и несхожими документами.

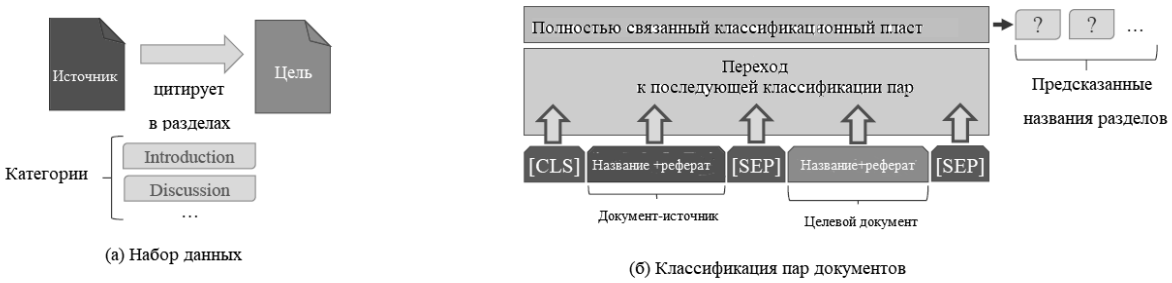


Рис. 2. Использование названий разделов из ссылок как категории для пар документов.
Эти разделы определяют аспекты сходства. Модель Transformer с названиями и рефератами в качестве входных данных используется для классификации.

Таблица 1

Распределение названий классов, извлеченных из названий цитирующих разделов в двух наборах данных

Название класса	**Подсчет**	**Название класса**	**Подсчет**		**Название класса**	**Подсчет**	**Название класса**	**Подсчет**
Введение	16 279	Заключение	1 158		Введение	15 108	Описание	454
Связанные работы	12 600	Обсуждение	1 132		Обсуждение	13 258	Материалы	420
Эксперимент	4 025	Оценка	971		Заключение	1 003	Вирус	218
Описание	1 365	Методы	719		Результаты	910	Дальнейшая работа	171
результаты	1 181	*Прочее*	22 249		Методы	523	*Прочее*	43 154
(а) ACL Anthology					(б) CORD-19			

Примечание: Верхние девять разделов-классов приводятся в убывающем порядке, оставшиеся группируются как Прочее.

Системы

Сфокусируемся на последовательной классификации пар с использованием моделей на основе архитектуры Transformer [12]. Такие модели на основе Transformer часто используются в задачах сходства текстов [7, 20]. Более того, авторы [6] обнаружили неоригинальные модели Transformer, т.е. BERT [13], XLNet [15], сети vanilla Siamese [21] и традиционные векторные представления слов (например Glo Ve [32], Paragraph Vectors [33]) в задаче классификации пар документов. Следовательно, исключаем сети Siamese и предварительные опытные модели векторных представлений слов из наших экспериментов. Вместо этого изучаем шесть вариаций Transformer и дополнительную основу для сравнения. Названия и рефераты пар научных статей, используемые как входные данные в модель, посредством которых символ [SEP] – Source Evaluation Panel, группа оценки источника – разделяет источник и целевую статью (рис. 2б). Данная процедура основана на нашей предыдущей работе [6]. В наших экспериментах мы не используем полные тексты, так как многие статьи не доступны бесплатно и отобранные модели Transformer накладывают жесткое ограничение в 512 символов.

Основа LSTM. В качестве основы используем бинаправленную LSTM [34]. Чтобы получить представления пар документов, введем названия и рефераты двух документов в LSTM, посредством которой статьи отделяются особым разделителем символов. Используем токенайзер (лексический анализатор) библиотеки SpaCy [35] и векторы слов из библиотеки fastText [36]. Векторы слов предварительно проверяются на аннотациях наборов данных ACL Anthology и CORD-19.

BERT, Covid-BERT & SciBERT. BERT- нейронная языковая модель на основе архитектуры Transformer [13]. Признано, что модели BERT предварительно проверяются на большом текстовом массиве без пересмотра. Две предварительно опытные цели – восстановление замаксированных средств идентификации пользователя (т.е моделирование языка масок) и последующее NSP (Next sentence prediction). После предварительного опыта модели BERT хорошо мотивированы для определенных задач, такие как сходство предложений [20] или классификация документов [23]. Некоторые модели BERT, предварительно проверенные на различных совокупностях, публично доступны. Для наших экспериментов мы оцениваем три вариации BERT: (1) модель BERT от авторов [13], проверенная на English Wikipedia и the BooksCorpus [37]. (2) SciBERT [24], вариация BERT, предназначенная для научной литературы, которая предварительно обработана на научных статьях по вычислительной технике и биомедицине; (3) Covid-BERT [38] – оригинальная модель BERT от авторов [13], но хорошо настроенная на CORD-19.

BioBERT [39] – другая модель BERT, специализирующаяся на биомедицинской области. Но мы исключаем BioBERT из наших экспериментов, так как SciBERT

превосходит ее в биомедицинских задачах [24]. Также опускаем вариации BERT от авторов [8], поскольку они используют цитирование на протяжении предварительно обработанной рискованной утечки данных в наш тестовый набор. Все три модели – BERT, SciBERT и Covid-BERT пожохи по своей структуре, за исключением набора, используемого на протяжении подготовки языковой модели.

RoBERTa. Авторы [14] предложили RoBERTa, которая является моделью BERT, предназначенной для более крупных массивов, более длительного времени проверки, и убирает задачу NSP из своей цели. Более того, RoBERTa использует дополнительные совокупности для предварительной обработки, главным образом такие как Common Crawl News [40], OpenWebText [42] и STORIES [42].

XLNet. В отличие от BERT модель XLNet [15] является не автокодировщиком, а авторегрессивной языковой моделью. XLNet не применяет NSP. Мы используем опубликованную ее авторами модель, XLNet, которая предварительно обработана на совокупностях – Wikipedia, BooksCorpus [37], Giga5 [43], ClueWeb 2012-B [44] и Common Crawl [45].

ELECTRA. ELECTRA [16] должна дополнительно маскировать языковое моделирование предварительно обработанной целью, заключающейся в выявлении перемещенных средств идентификации пользователя во входящей последовательности. Для этой цели авторы [16] используют генератор, который перемещает средства идентификации и дискриминатор сети, выявляющий перемещения. Генератор и дискриминатор – модели Transformer. ELECTRA не прибегает к задаче NSP. Для наших экспериментов применяется дискриминатор модели ELECTRA. Предварительно обученный дискриминатор модели ELECTRA заранее обрабатывается на тех же данных, что и BERT.

Гиперпараметры и применение. Мы выбираем нужные гиперпараметры LSTM в соответствии с полученными авторами [46] следующими данными: 10 периодов для подготовки, размер группы b=8, скорость изучения $\eta = 1^{-5}$, два уровня LSTM со 100 скрытыми размерами, внимание и выдача с вероятностью d=0,1. Тогда как основа LSTM использует vanilla PyTorch, все методы на основе Transformer применяются с использованием the Huggingface API [47]. Каждая модель Transformer используется в ее версии BASE. Гиперпараметры для хорошо мотивированной Transformer строятся с помощью работы авторов [13]: четыре подготовленных периода, скорость изучения $\eta = 2^{-5}$, размер группы b=8, и оптимизатор Adam с $\varepsilon = 1^{-8}$. Проводим оценку в стратифицированной k-кратной перекрестной проверке с k=4 (т.е. класс распределения остается идентичным для каждого повторения). Это приводит в среднем к 54 618, 75 /18 206, 2^5 подготовленных/опытных образцов для набора данных ACL Anthology и 74 436/ 24 812 – для набора данных CORD-19. Код источника, массивы данных и подготовленые модели публично доступны* Мы предоставляем Google Colab для испытания подготовленных моделей на любых статьях из Semantic Scholar**.

РЕЗУЛЬТАТЫ

Наши результаты разделены на три части: полная оценка, оценка категорий классов и количественная оценка***.

Полная оценка

Подробные результаты нашей количественной оценки представлены в табл. 2. Мы проводим оценку как 4- кратную перекрестную проверку на основе наших наборов данных. Сообщаем микро- и макро- среднее для полноты, точности и значения F1, чтобы принять во внимание несбалансированное распределение по категориям и классам (см. раздел «Наборы данных»).

Таблица 2

Полное значение F1 (со стандартным отклонением), полнота и точность для макро- и микро- среднего 7 методов для набора данных ACL Anthology и CORD-19

	ACL Anthology						CORD-19					
Массив данных	macro avg			micro avg			macro avg			micro avg		
	F1 (std)	P	R	F1(std)	P	R	F1 (std)	P	R	F1 (std)	P	R
^{LSTM}baseline	.063 ±.001	.069	.058	.290 ±.004	**.761**	.179	.128 ±.001	.137	.121	.579 ±.005	.758	.469
BERT	.256 ±.002	.317	.238	.641 ±.002	.719	.578	.387 ±.011	**.619**	.357	.822 ±.002	.840	.806
Covid-BERT	.270 ±.006	.404	.253	.648 ±.005	.715	.592	.394 ±.010	.578	.364	.818 ±.001	.836	.802
SciBERT	**.326 ±.005**	**.458**	**.303**	**.678 ±.002**	.725	**.637**	**.439 ±.010**	.560	**.401**	**.833 ±.003**	**.846**	**.820**
RoBERTa	.250 ±.003	.285	.232	.626 ±.003	.703	.564	.332 ±.008	.473	.316	.820 ±.001	.840	.801
XLNet	.263 ±.011	.372	.250	.645 ±.011	.705	.595	.362 ±.025	.523	.345	.817 ±.002	.832	.804
ELECTRA	.245 ±.005	.287	.228	.616 ±.021	.693	.554	.280 ±.001	.306	.276	.820 ±.002	.840	.801

Примечание:SciBERT выдает лучшие результаты для обоих наборов данных.

* GitHub repository: https://github.com/malteos/aspect-document-similarity

** https://colab.research.google.com/github/malteos/aspect-document-similarity/blob/master/demo.ipynb

*** Оценка по категориям и количественная оценки перестают использовать один из двух наборов данных из-за пространственных ограничений, но доступны на GitHub.

С учетом полных оценок SciBERT является лучшим методом с 0,326 макро-F1 и 0,678 микро-F1 в наборе данных ACL Anthology и с 0,439 макро-F1 и 0,833 микро-F1 в наборе данных CORD-19. Все модели Transformer по всем показателям превосходят LSTM за исключением микро-точности в наборе ACL Anthology. Этот разрыв между макро- и микро- средними результатами существует из-за несоответствия категорий классов (см. раздел «Оценка категорий классов»). BERT, SciBERT и Covid-BERT в среднем лучше выполняются на ACL Anthology и CORD-19 при сравнении основы и других моделей на основе Transformer. Что касается набора данных ACL Anthology, то методы ранжируются одинаково для макро- и микро-. SciBERT представляет более высокие оценки с большим отрывом ото всех, за ней следуют Covid-BERT, XLNet и BERT. Менее эффективными являются RoBERTa (0,626 микро-F1) и ELECTRA (0,616 микро-F1). С точки зрения макро-среднего методы представляют одинаковые средние значения для массивов CORD-19 и ACL Anthology за исключением BERT, превышающего XLNet. Только для микро-среднего в массиве CORD-19 результат отличается, т. е. ELECTRA и RoBERTa достигают более высоких оценок F1, чем Covid-BERT и XLNet. Даже если Covid-BERT лучше мотивирован на наборе CORD-19, его эффективность содержит 0,818 микро-F1.

Оценка категорий классов

Делим оба набора данных ACL Anthology и CORD-19 на 11 категорий классов с положительными и отрицательными примерами (разделы «Предварительная обработка данных» и «Негативная выборка»). Каждый класс представляет различный раздел, в котором статья получает ссылку. Раздел указывает на то, в каких аспектах две статьи являются схожими. Эти аспекты могут также быть двусмысленными, затрудняя задачу классификации названий. Следующий раздел изучает эффективность классификации относительно различных категорий классов. Табл. 3 представляет оценку F1, полноту и точность набора SciBERT для всех 11 категорий. Дополнительно мы включаем полные результаты для единичных и многокатегорийных выборок (т.е. 2 и ≥ 3). Оставшиеся методы из табл. 2 представляют более низкие, но пропорционально схожие значения*.

Категория None имеет самую высокую с большим отрывом оценку F1 (0,942 для набора ACL Anthology и 0,980 для набора CORD-19). Категория Other показывает вторую лучшую оценку F1, которая в сценарии классификации сходства (похожий-непохожий) может быть интерпретирована как противоположный класс по отношению к категории None. Оставшиеся положительные категории показывают более низкие оценки, а также более низкий ряд (число) выборок. Поскольку мы проводим 4-кратную перекрестную проверку, сооотношение подготовленного и опытного образцов составляет 75/25. В наборе CORD-19 10 788 (категория Other) опытных образцов существуют относительно 3 777 образцов (категория Introduction), которая является наиболее общим названием раздела (табл. 1). Пока более низкое число опытных образцов необязательно коррелирует с низкой точностью. В наборе ACL Anthology категория Related Work (3 150 опытных образцов) создает более высокие оценки по сравнению с категорией Introduction (4 069 образцов) с оценкой F1 в 0,638 при наличии только 113 образцов. Рзультаты в табл. 3 отражают воздействие категорий классов на общую эффективность. Шесть категорий (набор ACL Anthology – Conclusion, Discussion, Evaluation и Methods; набор CORD-19 – Future Work и Virus) имеют оценки F1 от 0 до 0,05. Различие в числе образцов и трудности в раскрытии латентной информации с точки зрения аспектов способствуют снижению точности в некоторых категориях. Даже для экспертов области местоположение того, где одна статья цитирует другую, например, в Introduction или Experiment, не является тривиальным для прогнозирования.

Таблица 3

Результаты SciBERTотносительно наборов данных ACL Anthology и CORD-19 по категории классов, числу доступных опытных образцов (ограниченное множество), оценке F1 (со стандартным отклонением), полноте (R) и точности (P).

ACL Anthology					CORD-19				
Категория	**Опытные образцы**	**F1 (Std)**	**P**	**R**	**Категория**	**Опытные образцы**	**F1 (Std)**	**P**	**R**
Background	341	0.436 ± 0.045	0.651	0.329	Background	113	0.617 ± 0.042	0.655	0.588
Conclusion	289	0.000 ± 0.000	0.000	0.000	Conclusion	250	0.274±0.039	0.563	0.182
Discussion	283	0.000 ± 0.000	0.000	0.000	Discussion	3314	0.636 ± 0.008	0.641	0.631
Evaluation	242	0.008 ± 0.007	0.396	0.004	Future work	42	0.032 ± 0.064	0.150	0.018
Experiment	1006	0.360 ± 0.008	0.491	0.284	Introduction	3777	0.644 ± 0.004	0.669	0.620
Introduction	4069	0.527 ± 0.005	0.576	0.486	Materials	105	0.241 ± 0.038	0.552	0.157
Methods	179	0.014 ± 0.028	0.208	0.007	Methods	130	0.205 ± 0.030	0.519	0.130
Related work	3150	0.638 ± 0.012	0.660	0.617	Results	227	0.322 ± 0.021	0.558	0.227
Results	295	0.015 ± 0.011	0.475	0.008	Virus	54	0.000 ± 0.000	0.000	0.000
Other	5562	0.645 ± 0.005	0.646	0.645	Other	10788	0.876 ± 0.002	0.872	0.879
None	6068	0.942 ± 0.002	0.934	0.951	*None*	8270	0.979 ± 0.001	0.980	0.977
1 label	15652	0.721 ± 0.002	0.717	0.726	1 label	22885	0.860 ± 0.003	0.844	0.876
2 labels	1968	0.540 ± 0.003	0.738	0.425	2 labels	1632	0.656 ± 0.004	0.849	0.535
> 3 labels	585	0.492 ± 0.015	0.857	0.345	> 3 labels	295	0.590 ± 0.010	0.925	0.433

* Подробные данные по оставшимся методам доступны вмсете с подгтовленными моделями в нашем хранилище GitHub.

Таблица 4

Матирца рассеяния из выбранного множества категорий для SciBERT на наборе данных CORD-19

Основополагающая истина		**Предсказания**														
Разделы	**Выборка**	N	**B**	**C**	**D**	**I**	**O**	**R**	**C,O**	**D,I**	**D,O**	**D,R**	**I,O**	**O,R**	**D,I,O**	**D,O,R**
C,D	21	-	-	-	1	6	7	-	-	1	-	-	1	-	-	-
C,O	79	-	-	2	1	2	58	-	13	-	-	-	3	-	-	-
D,I	**459**	**1**	-	-	**163**	**146**	**17**	-	-	**103**	**7**	**2**	**9**	-	**10**	-
D,O	351	1	2	-	102	30	120	1	-	15	59	1	4	1	4	-
D,R	65	1	-	-	6	10	10	-	-	1	3	28	-	-	-	1
I,O	453	2	1	-	15	114	215	1	-	12	16	1	62	-	9	-
D,I,O	142	1	1	-	28	31	11	-	-	33	8	-	12	-	14	-
D,O,R	23	-	-	-	5	-	7	-	-	-	5	2	-	1	-	1

Примечание: (N=None, C=Conclusion, O=Other, D=Discussion, I=Introduction, R=Results). Например (**жирное выделение**), 459 опытных образцов приписаны к Discussion и *I*ntroduction (D, I), из которых 103 являются правильно класифицированными. Оставшиеся образцы в большинстве случаев классифицированы как монокатегорийные, т.е. либо Discussion (163), либо Introduction (146).

Нижние ряды в табл. 3 иллюстрируют эффект множества категорий. Значения F1 снижаются в обоих массивах по мере роста числа категорий. Это происходит из-за снижающейся полноты. Точность растет с увеличением категорий. Табл. 4 демонстрирует долю распределения многокатегорийных опытных образцов в наборе CORD-19 и соответствующих предсказаниях SciBERT (этот список лимитивирован из-за пространственных ограничений). Когда представлены две и более категорий, SciBERT часто правильно предсказывает одну из категорий, но не другие. Например, две категории из Discussion и Introduction имеют правильными только 22% опытных образцов. Пока SciBERT правильно предсказывает для оставшихся образцов одну из двух категорий, т.е. либо *Duscussion* (35%) или Introduction (31%). Мы считаем сравнимыми разультаты для других множеств категорий, таких как Discussion (D), Introduction (I) и **Others** (O).

Качественная оценка

Чтобы обосновать наши количественные полученные данные, мы качественно оцениваем предсказание от SciBERT для набора ACL Anthology. Для каждого примера в табл. 5 SciBERT предсказывает, цитирует ли источник целевую статью и в каком разделе должна встретиться ссылка. Вручную изучаем предсказания относительно их правильности.

Первым примером авторов [48] и коллектива авторов [49] является правильное предсказание. С учетом основополагающей истины этим аспектом является Other (ссылка встречается в разделе, называемом «Результаты по данным тестирования»). Мы оцениваем Introduction как верятное обоснованное предсказание, поскольку работа авторов [48] подчиняется общей, описываемой авторами [49] задаче. Поэтому можно цитировать ее во введении. Все предсказания в примере 2 - правильны. По сравнению с другими примерами мы считаем пример 2 простым случаем, так как обе статьи упоминают свою тему (т.е. сегментацию запроса, [50, 51]) в названии и в первом предложении реферата (намек на категорию «Introduction»). Оба реферата примера 2 также относятся к «взаимной информации и EM алгоритма оптимизации» как к их методам. В примере 3 ряд авторов [52, 53] не обменивается ни одной ссылкой. Следовательно, пара статьей приписывается к категории None согласно данным основополагающей истины, даже если они, как правило, связаны. Фамилии авторов [52, 53] относятся к машинному переводу с китайского языка. Пока мы не согласны с предсказанием нашей модели Experiment, так как две статьи проводят различные эксперименты, делая Experiment необоснованным предсказанием. Предсказания примера 4 правильные. Коллектив авторов [54, 1992 г.] публикуется до работы авторов [55, 2007 г.] и поэтому ссылки не существует. Тем не менее две статьи охватывают близкую тему. Таким образом, можно ожидать ссылку на авторов [54] в работе авторов [55] в разделе введение, как предсказала SciBERT. Наша модель находит это семантическое сходство, учитывая их латентную информацию по теме. Примеры 5-6 прдставляют две пары, в которрых None была правильно предсказана в соответствии с основополагающей истиной. Ряд работ из примера 6, как правило, не связан друг с другом тематически, как уже предполагают их названия. Тем не менее, авторы [56] и авторский коллектив [57] в примере 5 объединены темой разрешения многозначности. Таким образом, мы должны согласиться с предсказанием положительной категории.

Кратко, качественная оценка не противоречит нашим количественным полученным данным. SciBERT различает документы на более высоком уровне и классифицирует, какие аспекты делают их схожими. В дополнение к традиционному сходству документов предсказания на основе аспекта позволяют оценить, как две статьи относятся друг к другу на уровне семантики. Например, являются ли схожими две статьи в аспектах Introduction или Experiment, представляется ценной информацией, особенно в обзорах литературы.

ОБСУЖДЕНИЕ

В наших экспериментах SciBERT превосходит все другие методы в парной классификации документов. Мы наблюдаем, что внутри области предварительная обработка и цель NSP часто ведут к более высоким оценкам F1. Переход общих языковых моделей к определенной области, как правило, снижает эффективность в наших экспериментах. Возможным объяснением этого

является уже определенный словарь в массивах ACL Anthology или CORD-19. Ряд работ авторов [24, 39] также исследует переход обучения между областями со схожими результатами. Covid-BERT кажется исключением, так как он выдает более низкие результаты (micro-F1), чем BERT в наборее CORD-19, даже если Covid-BERT был хорошо мотивированным на набор CORD-19. Мы наблюдаем языковую модель хорошо мотивировпнную на Covid-BERT, что не гарантирует более высокую эффективность по сравнению с предварительной обработкой из случая в SciBERT. Тем не менее авторы Covid-BERT предоставляют слишком мало информации, чтобы дать собственное объяснение ее эффективности. Отдельно от предварительной обработки внутри области цель NSP имеет положительное влияние на модели. Все системы на основе BERT, использующие NSP, превосходят модели, которые исключают NSP (XLNet, RoBERTa и ELECTRA). Мы приписываем положительный эффект от NSP ее сходству с нашей задачей, поскольку обе являются следствием задач парной классификации. Табл. 2 и 3 показывают вариацию между названиями и обоими наборами данных. Большее число подготовленных опытных образцов в CORD-19 (36%) могут способствовать более высокой эффективности в сравнении с набором ACL Anthology. Несбалансированное распределение классов и различные проблемы категорий вынуждают эффективность различаться между категориями классов. Высокие оценки F1 свыше 0,9 для негативных выборок ожидаемы, поскольку категория None является неотъемлемым сходством свободным от аспекта или проблемы предсказания цитирования. Модели Transformer показаны с целью хорошего выполнения этих двух проблем [8, 20]. Помимо несбалансированного распределения подготовленных опытных образцов мы приписываем различия между положительными категориями их двусмысленности и другим проблемам, свойственным категориям классов. Авторы часто расходятся по вопросу именования их разделов (например, Results, Evaluation), таким образом усиливается проблема наименования разных аспектов статьи. Это также способствует высокому числу выборок категории Other. Некоторые разделы, также ориентированные на контент, более уникальны, чем другие. Раздел Introduction, как правило, содержит контент, отличающийся от раздела Results. Различие в содержании позволяет некоторым разделам и соответствующим категориям классов легче, чем другим, различаться и предсказывать. Предполагаем слабую эффективность для Future Work из-за нехватки или отсутствия информации в названиях или аннотациях.

Нашей основной исследовательской целью в этой статье является изучение методов, способных объединять аспект информации в традиционной классификации сходства-различия. В связи с этим мы считаем результаты перспективными. В частности, оценка micro-F1 0,86 из SciBERT для набора CORD-19 является вдохновляющей. Наша количественная оценка указывает на то, что предсказания SciBERT могут правильно идентифицировать схожие аспекты двух научных статей. В целях подтверждения, если наше первое показание обобщается, то требуется проведение большого качественного исследования. Более того, мы наблюдаем, что категории классов с небольшими подготовленными данными действуют слабо. Например, Conclusion и Discussion имеют нулевую оценку F1 для набора ACL Anthology, тогда как для большего набора данных CORD-19 Duscussion выдает 0,636 F1. Мы ожидаем, что большие подготовленные данные приведут к более правильным прогнозам.

Таблица 5

Примеры категорий пар научных статей (источник и цель), определнных в соответствии со ссылками и предсказанных с участием SciBERT

	Статья-источник	Статья-цель	Ссылка	Предсказание
1	UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures (Bar et al., 2012)	SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity (Agirre et al., 2012)	Other	Introduction x
2	Query segmentation based on eigenspace similarity (Zhang et al., 2009)	Unsupervised query segmentation using generative language models and wikipedia (Tan and Peng, 2008)	Introduction, Experiment	Introduction ✓, Experiment ✓
3	Transition-Based Parsing of the Chinese Treebank using a Global Discriminative Model (Zhang and Clark, 2009)	Enhancing Statistical Machine Translation with Character Alignment (Xi et al., 2012)	None	Experiment x
4	Experiments in evaluating interactive spoken language systems (Polifroni et al., 1992)	Evaluating information presentation strategies for spoken recommendations (Winterboer and Moore, 2007)	None	Introduction x, Other x
5	Similarity-based Word Sense Disambiguation (Karov and Edelman, 1998)	Targeted disambiguation of ad-hoc, homogeneous sets of named entities (Wang etal., 2012)	None	None ✓
6	SciSumm: A Multi-Document Summarization System for Scientific Articles (Agarwal et al., 2011)	Improving question-answering with linking dialogues (Gandhe et al., 2006)	None	None ✓

Примечание: Основанные на ограниченном множестве правильные предсказания отмечены ✓, необоснованные – x.

ЗАКЛЮЧЕНИЕ

В этой статье мы применяем к научным статьям парную многокатегорийную, многоклассовую классификацию документов, чтобы вычислить оценку сходства документа на основе аспекта. Обрабатываем названия разделов как аспекты цитирования статей и названий, соответственно встречающихся в этих разделах. Изучаемые модели обучены для предсказания цитирований и соответствующих категорий, основанных на названии статьи и ее реферате. Мы оцениваем модели Transformer BERT, Covid-BERT, SCiBERT, ELECTRA, RoBERTa и XLNet и основу LSTM относительно двух научных наборов, т.е. ACL Anthology и CORD-19. В целом SciBERT в наших экспериментах работает лучше. Несмотря на сложность задачи, SciBERT предсказала сходство документов на основе аспекта с оценкой F1 свыше 0,83. Эффективность SciBERT стимулирует дальнейшее исследование в этом направлении. Кажется обоснованным включить задачу сходства документов на основе аспекта в качестве новой цели предварительной обработки в архитектуре Transformer. Эта новая цель могла бы быть интегрирована похожим способом как двойственная цель предсказания цитирования, предложенная авторами [8]. В качестве дальнейшей работы планируем интегрировать сходство документов на основе аспекта в рекомендательную систему. Таким образом, стимулируя большее исследование пользователей, чтобы подтвердить наши первые выводы относительно того, что сходство документов на основе аспекта действительно помогает пользователям находить более релевантные рекомендации. Однако наш расширенный эмпирический анализ уже демонстрирует, что модели Transformer хорошо подходят для правильного вычисления сходства документов на основе аспекта на примере научных статей.

Благодарность. Хотим выразить благодарность всем рецензентам и Кристофу Альту за их рекомендации и ценную обратную связь. Представленное в этой статье исследование профинансировано Немецким федеральным министерством образования и исследований через проект QURATOR (Unternehmen Region, Wachstumskern, no. 03WKDAIA).

ЛИТЕРАТУРА

1. *Beel J., Gipp B., Langer S., Breitinger C.* Research-paper recommender systems: A literature survey//International Journal on Digital Libraries. — 2016. — Vol. 17, No. 4. — P. 305–338.

2. *Goodman N.* Seven strictures on similarity. Problems and projects. — 1972.

3. *Bar D., Zesch T., Gurevych I.* A Reflective View on Text Similarity//International Conference Recent Advances in Natural Language Processing (RANLP), pp. 515–520. — 2011.

4. *Huang T.-H. K., Huang C.-Y., Ding C. -Y. C., Yen-Chia Hsu Y.-C., Giles C L.* CODA-19: Reliably annotating research aspects on 10,000+ CORD-19 abstracts using a non-expert crowd. — [arXiv:2005.02367.] — 2020.

5. *Chan J., Chang J. C., Hope T., Shahaf D., Kittur A.* SOLVENT: A mixed initiative system for finding analogies between research papers// Proceedings of the ACM on Human-Computer Interaction, 2(CSCW):1–21, nov.— 2011.

6. *Ostendorff M., Ruas T., Schubotz M., Rehm G., Gipp B.* Pairwise multi-class document classification for semantic relations between Wikipedia articles//Proceedings of the 2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL'20). — 2020.

7. *Jiang J.-Y., Zhang M., Li C., Bendersky M., Golbandi N., Najork M.* Semantic text matching for long-form documents // The World Wide Web Conference on - WWW '19, pages 795–806, New York, New York, USA. — ACM Press, 2019.

8. *Cohan A., Feldman S., Beltagy I., Downey D., Weld D. S.* SPECTER: Document-level representation learning using citation-informed Transformers// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). — 2020.

9. *Steven Bird S., Dale R., Dorr B. J., Gibson B., Joseph M. T., Kan M. Y., Lee D., Powley B., Radev D. R., Tan Y. F.* The ACL Anthology reference corpus: A reference dataset for bibliographic research in Computational Linguistics// Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, pp. 1755–1759.— 2008.

10. *Wang L. L., Lo K., Chandrasekhar Y., Reas R., Yang J., Eide D., Funk K., Kinney R., Liu Z., Merrill W., Mooney P., Murdick D., Rishi D., Sheehan J., Shen Z., Stilson B.,Wade A. D., Wang K., Wilhelm C., Xie B., Raymond D., Weld D. S., Etzioni O., Kohlmeier S.* CORD-19: The Covid-19 open research dataset. — 2020. — [arXiv:2004.10706.]

11 *Kobayashi Y., Shimbo M., Matsumoto Y.* Citation recommendation using distributed representation of discourse facets in scientific articles// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 243–251, New York, NY, USA, may.— ACM, 2018.

12. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.* Attention is all you need// Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, Jun. — 2017.

14. *Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V.* RoBERTa: A robustly optimized BERT pretraining approach. — 2019. — [arXiv:1907.11692.]

15. *Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V.* XLNet: Generalized autoregressive pretraining for language understanding// Advances in Neural Informati on Processing Systems 32, pp. 5754–5764. — 2019.

16. *Clark K., Luong M.-T., Le Q. V., Manning C. D.* ELECTRA: Pre-training text encoders as discriminators rather than generators//International Conference on Learning Representations, pp.1–18. — 2020.

17. *Bowman S. R., Angeli G., Potts C., Manning C. D.* A large annotated corpus for learning natural language inference// Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 632–642. — 2015.

18. *Williams A., Nangia N., Bowman S.* A broad-coverage challenge corpus for sentence understanding through inference. — [arXiv:1704.05426, pages 1112–1122]. — 2018.

19. *Daniel Cer D., Diab M., Agirre E., Lopez-Gazpio I., Specia L.* SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation// Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), volume 371, pp. 1–14, Stroudsburg, PA, USA.— Association for Computational Linguistics, 2017.

20. *Reimers N., Gurevych I.* Sentence-BERT: Sentence embeddings using Siamese BERT-networks//The 2019 Con-

ference on Empirical Methods in Natural Language Processing (EMNLP 2019). — 2019.

21. *Bromley J., Bentz J. W., L. Bottou L., Guyon I., Lecun Y., Moore C., Sackinger E., Shah R.* Signature verification using a Siamese time delay neural network//International Journal of Pattern Recognition and Artificial Intelligence. — 1993. — Vol. 7, No.4.

22. *Adhikari A., Ram A., Tang R., Lin J., Cheriton D. R.* DocBERT: BERT for Document Classification. — 2019. — [arXiv:1904.08398v1].

23. *Ostendorff M., Bourgonje P., Berger M., Moreno-Schneider J., Rehm G., Gipp B.* Enriching BERT with knowledge graph embeddings for document classification// Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), pp. 305–312, Erlangen, Germany. — German Society for Computational Linguistics & Language Technology, 2019.

24. *Beltagy I., Lo K., Cohan A.* SciBERT: A pretrained language model for scientific text//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3613–3618, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2019.

25. *Hassan H. A. M., Giuseppe Sansonetti G., Gasparetti F., Micarelli A., Beel J.* BERT, ELMo, USE and InferSent sentence encoders: The panacea for research-paper recommendation?// CEUR Workshop Proceedings, volume 2431, pp. 6–10. — 2019.

26. *Kanakia A., Shen Z., Eide D., Wang K.* A scalable hybrid research paper recommender system for Microsoft Academic// The World Wide Web Conference on - WWW '19, pp. 2893–2899, New York, New York, USA. — ACM Press, 2019.

27. *Collins A., Beel J.* Document embeddings vs. keyphrases vs. terms: An online evaluation in digital library recommender systems// ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 130–133. — 2019.

28. *Nanni F., Ponzetto S. P., Dietz L.* Entity-aspect linking: Providing fine-grained semantics of entities in context// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 49–58, New York, NY, USA, may. — ACM, 2018.

29. *Lo K., Wang L. L., Neumann M., Kinney R., Weld D. S.* S2ORC: The Semantic Scholar open research corpus. — 2019. —[arXiv:1911.02782].

30. *Ley M.* DBLP: Some lessons learned// Proceedings of the VLDB Endowment. — 2009. — Vol. 2, No. 2. — P. 1493–1500, aug.

31. *Mikolov T., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality// Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, pp. 3111–3119. — 2013.

32. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. — 2014.

33. *Le Q. V., Mikolov T.* Distributed representations of sentences and documents// Proceedings of the 31st International Conference on Machine Learning. — 2014. — Vol. 32. — P. 1188–1196.

34. *Hochreiter S., Schmidhuber J.* Long short-term memory//Neural Computation. — 1997. — Vol. 9, No. 8. — P. 1735–1780, nov.

35. *Honnibal M., Montani I.* SpaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. — 2020. — [в печати].

36. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching word vectors with subword information// Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

37. *Zhu Y., Kiros R., Zemel R., Salakhutdinov R., Urtasun R., Torralba A., Fidler S.* Aligning books and movies: Towards story-like visual explanations by watching movies and reading books//Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter:19–27.— 2015.

38. *Chan B.* CORD-19 BERT Model. — 2020. — https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/discussion/138250.

39. *Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining// Bioinformatics. — 2018. — P.1–8, sep.

40. *Nagel S.* 2016. Common Crawl News. — 2016. — http://commoncrawl.org/2016/10/news-dataset-available/.

41. *Gokaslan A., Cohen V.* Openwebtext corpus. — 2019. — https://skylion007.github.io/OpenWebTextCorpus/.

42. *Trinh T. H., Le Q. V.* A simple method for commonsense reasoning. — 2018. — [arXiv:1806.02847].

43. *Parker R., Graff D., Kong J., Chen K., Maeda K.* English gigaword fifth edition. — 2011. — https://catalog.ldc.upenn.edu/LDC2011T07.

44. *Callan J., Hoy M., Changkuk Yoo C., Zhao L.* Clueweb09 data set. — 2009. — https://lemurproject.org/clueweb09/.

45. *Elbaz G.* Common Crawl. — 2007. — http://commoncrawl.org

46. *Reimers N., Gurevych I.* Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 338–348, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2017.

47. *Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Brew J.* HuggingFace's Transformers: State-of-the-art natural language processing. — 2019. — [arXiv:1910.03771, oct.].

48. *Bar D., Biemann C., Gurevych I., Zesch T.* UKP: Computing semantic textual simi- larity by combining multiple content similarity measures// 1st Joint Conference on Lexical and Computational Semantics (SEM 2012), Vol. 2, pp. 435–440. — 2012.

49. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A.* SemEval-2012 task 6: A pilot on semantic textual similarity - Google search// Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 385–393. — 2012.

50. *Zhang C., Sun N., Hu X., Huang T., Chua T.S.* Query segmentation based on eigenspace similarity// ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf., pp. 185–188.— 2009.

51. *Tan B., Peng F.* Unsupervised query segmentation using generative language models and Wikipedia// Proceed-

ing of the 17th international conference on World Wide Web - WWW '08, p. 347, New York, New York, USA. — ACM Press, 2008.

52. *Zhang Y., Clark S.* Transition-based parsing of the Chinese treebank using a global discriminative model// Proceedings of the 11th International Conference on Parsing Technologies - IWPT '09, p. 162, Morristown, NJ, USA. — Association for Computational Linguistics, 2009.

53. *Xi N., Tang C., Dai X., Huang S., Chen J.* Enhancing statistical machine translation with character alignment// 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2(July). — 2012. — P. 285–290.

54. *Polifroni J., Hirschman L., Seneff S., Zue V.* Experiments in evaluating interactive spoken language systems // Proceedings of the workshop on Speech and Natural Language - HLT '91, p. 28, Morristown, NJ, USA. — Association for Computational Linguistics, 1992.

55. *Winterboer A., Moore J. D.* Evaluating information presentation strategies for spoken recommendations// RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems, pages 157–160. — 2007.

56. *Agarwal N., Reddy R. S., Gvr K., Rose C. P.* SciSumm: A multi-document ´summarization system for scientific articles// 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of Student Session (ACL HLT 2011), pp. 115–120. — 2011.

57. *Gandhe S., Gordon A. S., Traum D.* Improving question-answering with linking dialogues// International Conference on Intelligent User Interfaces, Proceedings IUI.— 2006. — P. 369–371.— 2006.

58. *Karov E., Edelman S.* Similarity-based word sense disambiguation// Computational Linguistics. — 1998. — Vol. 24, No. 1.

59. *Wang C., Chakrabarti K., Cheng T., Chaudhuri S.* Targeted disambiguation of ad-hoc, homogeneous sets of named entities// WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web, pp. 719–728. — 2012.

DOI: 10.36535/0203-6460-2021-01-5

Отчет о восьмой национальной конференции Института наукометрии по «Научной коммуникации и наукометрии»*

Восьмая национальная конференция Института наукометрии (Institute of Scientometrics – IOS) под названием «Научная коммуникация и наукометрия» была организована Продвинутым центром лечения, исследования и образования в области рака (Advanced Centre for Treatment, Research and Education in Cancer – ACTREC) и проходила в г. Нави Мумбаи с 22 по 23 ноября 2019 г.

В двухдневной национальной конференции и предшествующим ее работе семинаре «Научная интеграция в исследовании: методы и технологии» приняло участие свыше 105 делегатов, включая основного докладчика, приглашенных докладчиков, экспертов и партнеров по индустрии и делегатов.

Конференцию торжественно открыл главный гость, проф. П. С. Васудев Рао, почетный вице-председатель национального института им. Хоми Бхабха (г. Мумбаи). Другими почетными гостями были проф. Раджеш Диксит, директор CCE, проф. С. Л. Сангам, учредитель отделения библиотековедения и информатики (Department of Library and Information Science – DLIS) Университета шт. Карнатака и основатель IOS, а также д-р Сатиш Муннолли, ответственный секретарь IOS-2019 и библиотекарь ACTREC.

Д-р Сатиш Муннолли представил обзор и генезис темы конференции «Научная коммуникация и наукометрия». Проф. С. Л. Сангам кратко изложил историю IOS и его деятельности. Проф. Раджеш Диксит в своем приветственном обращении сослался на тот факт, что слишком большое количество информации становится проблемой для ученых при отборе и использовании правильного типа информации. Он подчеркнул необходимость методов фильтрации информации. Далее он указал на необходимость обучения молодых специалистов написанию научных статей и представлению авторства в статьях для создания качественного научного результата.

Проф. П. С. Васудев Рао сделал исторический отчет по научной коммуникации и наукометрии и предостерег, что слишком большое доверие цитированию плохо для роста научной продукции. Более того, он остановился на исследовании по выявлению альтернативных моделей для оценки научного результата. Вместе с другими гостями он участвовал в выпуске материалов конференции.

Проф. Каннан Моудагалай, преподаватель отделения химического проектирования и координатор Национальной виртуальной библиотеки Индии (National Virtual Library of India – NVLI), Индийский институт технологий (Indian Institute of Technology – IIT), г. Мумбаи, выступил с основным докладом по теме «Публичные библиотеки как центры обучения через проектирование социальных медиа». В своей речи он уделил внимание публичным библиотекам в качестве центров обучения и проинформировал об учебных пособиях на региональных и иностранных языках, разработанных по NPTEL (National Programme on Technology Enhanced Learning – Национальная программа по совершенствованию обучения в области технологий) в целях обучения слушателей из Индии и других развивающихся стран мира. Он представил подробный отчет о NVLI и ее целях, связанных с цифровизацией и предоставлением доступа к материалам по культуре Индии путем участия музеев, архивов и публичных библиотек в этой программе. Он также упомянул об обучающих программах в системе KOHA для библиотекарей, предлагаемых через NVLI, и разработку каталога публичных библиотек в Индии. Далее он сказал, что под эгидой Национальной миссии библиотек NVLI хотела бы стать партнером публичных библиотек в обучении их персонала использованию программного обеспечение для библиотек, такого как KOHA и DSpace.

Первую сессию первого дня работы конференции возглавляли д-р С. К. Саванур, учредитель DLIS в колледже Джоши-Бедекар, г. Тхана, и д-р Сатиш Канамади, библиотекарь Института социальных наук Тата, г. Мумбаи. На сессии было два приглашенных выступления и пять докладов, представленных делегатами. Д-р. Г. Махеш, старший ведущий ученый SCIR-NISCAIR, г. Нью-

* Перевод Pujar S. M. A report. Eighth conference of Insitute of Scientometrics on "Scholarly Communication and Scientometrics"//Annals of Library and Information Studies. — 2020. — Vol. 67, March 2020. — P. 70-73. — http://nopr.niscair.res.in/jinfo/alis/ALIS%2067(1)%20(Report).pdf

Дели, выступил с приглашенной речью по теме «Индия в меняющемся ландшафте научной коммуникации». Он обсудил историческую перспективу журналов, изменяющую сценарий журнальных публикаций в различных формах, таких как журналы открытого доступа, журналы данных, видео-журналы, вики-журналы, перекрывающие журналы, зеркальные журналы и т.п. Основным моментом его речи было влияние индийских журналов открытого доступа, в особенности журналов NISCAIR, он указал, что ссылки на статьи, опубликованные в этих журналах, не увеличиваются и одновременно снижаются доходы. Он также затронул такие темы, как незаконные действия сетевых сайтов, исследовательские методы и их адаптация индийскими учеными, альтернативные метрики, управление данными в науке, плагиат, отзывы статей и Plan S.

Д-р Медха Джоши, NCG-TMC, г. Мумбаи, обратилась к собравшимся с приглашенной речью на тему «Научные коммуникации в медицинских науках: направления и вызовы». Она начала с кризиса продолжающихся изданий 1990-х гг. и представила обзор научных коммуникаций в медицинских науках от печатной до цифровой эры, завершая выступление рецензируемыми коллегами журналами, неформальными средствами коммуникации, такими как блог посты, соцмедиа и каналы распространения вспомогательных данных, такие как патологические слайды, презентации, видео и сырые данные и т. п. Она отметила важность связи научных данных, верификации и визуализации. Также обсудила журналы со специальным контентом, такие как ситуационные исследования, видео и системные обзоры. Д-р Джоши также указала на усиливающуюся роль сайтов социальных сетей в медицинских научных коммуникациях.

Презентации докладов открылись д-ром Сароджа, представившей доклад «Инновационные стратегии библиотечных и информационных центров в меняющемся ландшафте научной коммуникации». Она дополнила более раннюю систему и ввела новые модели научной коммуникации, изменения в процессе рецензирования и моделях по подписке. Также она упомянула о переходных соглашениях, включающих соглашение между издательствами и библиотеками относительно чтения и публикации.

Д-р С. К. Саванур в своей приглашенной речи «Наукометрия и читабельность» обсудил возможности чтения как со стороны читателя, так и автора с точки зрения языка, точности, внешнего вмешательства, обучения, грамматики и т.д., и связь между наукометрией и читабельностью. Д-р А. Редди представил доклад «Научное сотрудничество и сети в научных коммуникациях», в котором он обсудил сотрудничества авторов, причины и типы сотрудничеств и выделил роль сетей научной коммуникации, таких как ResearchGate. Д-р Ш. М. Пуджар, приглашенный докладчик, в своем выступлении «Ссылки в системе Google Scholar» выделил ссылки в Google Scholar, охват, преимущества, поиск информации о ссылках и другие научные показатели, использующие бесплатное онлайн средство «Публикуйся или Погибнешь».

Г-н Ганеш Сурвасе представил доклад на тему «Публикации и тенденции цитирования в научном журнале *Nature*» и обсудил модель цитирования в данном журнале, а также сравнил тенденции цитирования между журналами закрытого и открытого доступа.

Вторая сессия открылась под председательством д-ра Б. С. Кадемани, бывшего научного сотрудника "G", BARC, г. Мумбаи. Данная сессия включала один приглашенный доклад и две презентации. Д-р Сатиш Канамади в своем приглашенном докладе «Научное влияние, оценка и инновации» уточнил совместное обучение, аккредитацию, MOOCs и исследование и выделил научное влияние с точки зрения академических, экономических и общественных выгод. Д-р Канамади рассмотрел роль агентств по оценке, таких как QS, Times Higher Education, Leiden rankings, NIRF и т.д. Он упомянул, что исследование и ссылки составляют 50% весового коэффициента при ранжировании такими агентствами по оценке.

Г-жа Мугдха Раул представила доклад «Средства научной коммуникации». Она кратко охарактеризовала несколько онлайн средств и их помощь в улучшении влияния научной коммуникации и поделилась сведениями о новых моделей издательства, таких как имидж банк и визуальные онлайн журналы. Она также подчеркнула роль специалистов сферы библиотековедения и информатики в распространении знаний об этих источниках и средствах усовершенствования исследования. Г-н Рангараджан выступил с докладом «Наукометирический анализ – научный институциональный результат». Он обсудил базы данных цитирования, научные показатели и анализ ссылок публикаций IGCARs с точки зрения авторства, предметных областей, журналов, сотрудничества, h-индекса и т.п. за десятилетний период (2009-2018 гг.).

Третью сессию этого дня возглавил д-р Медха Джоши. В этой сессии было заслушано два приглашенных доклада и одна презентация. Д-р Мурари Тапасви, OSD, Университет Гоа, в своем докладе рассмотрел «Авторское сотрудничество и ссылки: на примере индийских авторов». Он обсудил ранжирование университета Гоа индийскими и международными рейтинговыми агентствами, такими как QS, Times Higher Education и т. д. Далее он уточнил научное сотрудничество Университета Гоа с другими организациями и странами в анализе публикаций университета и проинформировал, что предложенная инициатива способна увеличить число публикаций в индексируемых Scopus журналах.

Д-р Б.С. Кадемани представил доклад «Наукометрические портреты ученых». Он конкретизировал научные коммуникации и поставил ряд вопросов: Почему ученые должны проводить исследования? Зачем оценивать исследование? Какова цель подобной оценки? и т.п. Далее он обсудил важность оценки для разработчиков политики и финансирующих агентств в целях оценки эффективности индивидуумов, учреждений или стран. Он дальше объединил несколько исследований, проведенных в BARC по различным учреждениям, странам, нобелевским лауреатам и подчеркнул, что индивидуумы являются источниками идей, а следовательно, предпочтительных исследований отдельных ученых, которые широко известны как «наукометрический портрет».

Доклад г-на Нилеша Ханде на тему «Вебометрическое исследование сетевых сайтов университетов: исследование с особым акцентом на шт. Махараштра» фокусировался на обзоре вебометрии и обсуждал ранжиро-

вание университетов шт. Махараштра. Г-жа Рупали Кумбхар представила доклад – «Компоненты ИКТ в расписании университетов MLISc в шт. Карнатака».

Второй день работы конференции открылся сессией под председательством проф. С. Л. Сангама, где было заявлено два приглашенных доклада и четыре презентации. Проф. П. В. Коннур, основатель DLIS в университете Рани Ченнамма, район Белагави, и председатель академии LIS, г. Бангалор, говорил на тему «Препятствия в индийских исследованиях». Он обсудил трансформацию исследования от прошлого к настоящему и ее воздействие на общество, ученых и экономику. Он сравнил результаты публикации в Индии с другими странами и подчеркнул, что сила Индии заключается в молодых людских ресурсах и демографических дивидендах, но нехватка навыков, приводящая к слабому трудоустройству выпускников, является большой проблемой Индии.

Г-н Н. В. Сатхьянарьяна, председатель и управляющий директор, «Imformatics Publishing», г. Бангалор, изложил свой взгляд на «Открытый доступ: глобальный сценарий и локальные вызовы». В своей речи он обсудил типы моделей открытого доступа, его прогресс, рост и препятствия. Далее он изложил свои идеи относительно движения открытого доступа, его преобладания и сохранения. Он подчеркнул, что бесплатное повторное использование – проблема на уровне Будапештской декларации по открытому доступу. Он упомянул, что большая часть открытого доступа все еще контролируется издательствами и сегодня множество издателей разрешают зеленый открытый доступ в форме самоархивирования авторами. Такие модели, как ResearchGate и Plan S, могут нарушать модель открытого доступа, что, по его мнению, способствует переходу от потребления к снабжению. Г-н Н. В. Сатхьянарьяна представил обзор процесса роста исследований в Индии, нехватки научных журналов и издательской индустрии. В заключении он выделил потребность в ощутимом изменении проблемы открытого доступа со стороны заинтересованных лиц.

В своей презентации г-жа Свапнали Патил рассказала о «Способах обнаружения статей открытого доступа: анализ». Она затронула проблемы информационного взрыва, агрегации поисковых машин, сложной поисковой процедуры с использованием таких средств, как Open DOAR, и роль средств обнаружения – кнопка Google Scholar, Lazy Scholar, Kopernio и Unpaywall в эффективном поиске полных текстов статей.

Г-жа Швета Патхак представила доклад «Массовые открытые онлайн курсы: их влияние на образование и профессию в области библиотековедения и информатики». Презентация г-жи Гаримы Гуджрал была на тему «Появление научных публикаций открытого доступа в Индийском институте технологий, г. Мумбаи: наукометрическое исследование по реализации Plan S». Г-жа Сунита Пуджар выступила с докладом: «Plan S: возможность или вызов для открытого доступа», в котором она дополнила Plan S и его применения для ученых, издателей, дисциплин и глобального движения в южную часть страны.

Пятая техническая сессия началась под председательством д-ра П. В. Коннура. На ней было представлено три приглашенных доклада и одна презентация. Д-р С.Л. Сангам выступил с докладом «Наукометрический подход к литературе по сельскому хозяйству относительно ВВП». Он уточнил связь между числом публикаций и экономическими показателями (ВВП). Д-р Сангам указал на то, что глобальные результаты сельского хозяйства отражают сильную корреляцию между экономическими показателями и публикациями и уточнил научные публикации Индии по сельскому хозяйству в соответствии с экономическими показателями.

Д-р Правеен Гавали, ученый, Индийский институт геомагнетизма, г. Нави-Мумбаи, в своей презентации – «Наукометрия: проблемы оценки доверия» описал историческую перспективу науки и вкладов ученых в рост науки и исследований. Он упомянул, что экспоненциальный рост науки произошел между 1800 и 1950 гг., приведя к многим моделям авторства, которые усложняют процесс идентификации. Он дополнил наукометрический анализ идентификацией размеров науки и ее применением в оценке опубликованной литературы.

Г-н Шиварам Говда, технический служащий, CSIR-NAL, г. Бангалор, представил приглашенный доклад на тему «Коррелируют ли показатели использования е-ресурсов с ростом публикаций SCI». Он выделил, как библиотечные показатели могут помочь в измерении успеха библиотечных служб, оправдывая расходы, сохранность фондов, окупаемость инвестиций и свидетельства на основе данных, получаемых с помощью принятия решений. Кроме того, он представил ситуационное исследование VTU, исследование консорциума района Белагави, подробно изложив найденную корреляцию между показателями скачиваний и публикациями SCI.

Г-жа Приянка Бозе представила доклад о «Роли социальной сети в научном исследовании», в котором обсудила использование социальных сетей в научном исследовании, преимущественно в общественных науках.

Шестая сессия состоялась под председательством д-ра Ш. М. Пуджар, главного библиотекаря, IGIDR, г. Мумбаи. На сессии было представлено два приглашенных выступления и четыре доклада. Д-р С. К. Саванур выступил на тему «Закон Ципфа в управлении». Он уточнил применение данного закона к управлению различными городами и штатами в Индии.

Д-р В. М. Банкапур, адъюнкт-профессор и председатель, DLIS, Университет Рани Ченнамма, район Белагави, говорил на тему «Исследование устаревания литературы по сельскому хозяйству». Он обсудил анализ цитирований в литературе по сельскому хозяйству преимущественно на основе диссертаций на соискание докторской степени, представленных в Университет сельскохозяйственных наук, г. Дхарвад, чтобы определить скорость устаревания.

Среди презентаций г-н Пракаш Кумбар представил доклад на тему «Наукометрическое измерение исследований в химии». Он описал наукометрический анализ литературы по химии с акцентом на Университет г. Майсур. Г-н Е. Р. Пракаш рассказал о «мегасовместных публикациях и их влиянии на наукометрические профили научных организаций».

Д-р С. Б. Патил сделала доклад на тему «Количественный анализ докторских исследований в Университете шт. Карнатака, г. Дхарвад». Г-н Кави Уппрети выступил с докладом «Библиометрический анализ и визуализация литературы по плагиату».

В целях признания и поощрения вклада молодых ученых награды за лучшие доклады были отданы трем докладчикам презентаций. Награда включала сертификат и денежную премию. Первая награда была отдана г-ну Е. Р. Пракашу и г-ну Ганешу Сурвасе за их доклад на тему «Мегасовместные публикации и их влияние на наукометрические профили научных организаций». Вторая – отдана г-ну Нилешу Ханде за доклад «Вебометрическое исследование сетевых сайтов университетов: исследование с особым акцентом на шт. Махараштра»; третья награда присуждена за доклад г-жи Рупали Кумбхар и д-ра П. Г. Тадасад по теме «Компоненты ИКТ в расписании университетов MLISc в шт. Карнатака».

Проф. П. В. Коннур был главным гостем прощальной церемонии, на которой председательствовал проф. С.Л. Сангам. Д-р Ш. М. Пуджар, генеральный докладчик конференции, представил подробный отчет о конференции. Д-р Сатиш Муннолли, руководитель конференции высказал признательность организаторам и участникам конференции.

Шампрасад М. Пуджар,

главный библиотекарь,
Институт развития исследований
им. Индиры Ганди, г. Мумбаи

Приглашаем российских и зарубежных авторов к сотрудничеству
в журнале «Международный форум по информации».
Оригинальные статьи и другие материалы (рецензии, письма)
можно присылать на русском или английском языке
по почтовому адресу, указанному в «Памятке для авторов»
или по электронной почте: mfi@viniti.ru.

Ответственный за выпуск *Л. В. Кобзева*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.
Подписано в печать 05.04.2021 г.
Бумага офсетная. Формат 60x84 1/8. Гарн. литер. Печать цифровая
Усл. печ. л 6,00 Уч.-изд. л. 6,46 Тираж 33 экз.

Адрес редакции: 125190, Россия, г. Москва, ул. Усиевича, д. 20.
Тел. (499) 155-44-95

Издательство ВИНИТИ РАН. 125190, г. Москва, ул. Усиевича, д. 20.
Тел. (499) 152-08-10, (499) 155-42-85, (499) 151-78-61

ДЛЯ ЗАМЕТОК

ДЛЯ ЗАМЕТОК