# история тематического моделирования. «чтение издалека» и тематическое моделирование в исторических исследованиях

***Рене Брауэр***

## *Университет Гетеборга (Швеция), кафедра наук о земле*

***Матс Фридлунд***

## *Университет Аалто, Финляндия,* кафедра промышленного дизайна и производства

����.fr�������@�����o.fi

Тематическое моделирование �ТМ�� — это метод, применяемый в рам- ках “цифровой истории�, который представляет собой методологию работы с оцифрованными источниками. Такой метод, возможно, наиболее близок к реализации идеи “чтения издалека� боль�их объемов текста, предложенной историком литературы Франко Мореттти. Вдохновленный концепцией “чте- ния издалека� метод ТМ применяется в исторических исследованиях с начала 2000-х. Настоящее исследование предлагает обзор применения современного передового метода ТМ в исторических исследованиях.

ТМ сначала разрабатывалось при обработке естественного языка и для ма-

�инного обучения в 1990-х для ���� ������ �возможные переводы — добыча данных, интеллектуальный анализ данных, «просев� информации���. Пре- имущество метода заключалось в возможности охватить значительно боль-

�е материала, чем при традиционных способах обработки. Одной из первых применялась модель Латентного размещения Дирихле, позволяющая реализо- вывать поисковые функции ТМ, количественное подтверждение интуитивных предположений или обобщение боль�ого корпуса текстов. ТМ имеет дело со многими конкурирующими теориями и гипотезами, и является активной об- ластью исследования в компьютерных науках.

Обзор исторических текстов берет за отправную точку первую статью, свя- занную с этой темой, опубликованную в рецензируемом издании в 2006, а в качестве конечной точки — опубликованную в 2013 исследовательскую моно- графию. Всего рассмотрено 23 исторических исследования, затрагивающих ТМ. Чтобы представить поле исследования в общем виде, научные работы были рассмотрены с помощью количественного подхода дистанционного чте- ния и проанализированы с точки зрения уровня образовательной подготовки авторов, их пола, академического статуса, географического расположения уч- реждения, в котором они заняты; также с точки зрения типа корпуса текстов, языка, хронологии и географичекой принадлежности. Результаты показали, что боль�инство исследователей — мужчины, находящиеся в начале акаде- мической карьеры, преимущественно занятые в университетах США; значи- тельное количество рассмотренных текстов представляют собой неклассичес- кие электронные виды публикаций. Не смотря на применение в исторических исследованиях, ТМ скорее рассматривается как технологический метод, а

боль�инство авторов, пи�у�их на эту тему, имеют техническое образование. Изучаемые корпусы текстов преимущественно на англйском языке, касаются американских или глобальных тем, концентрируются на новей�ей истории. Таким образом, ТМ представляется скорее как зарождающаяся, чем устано- вив�аяся историческая методология.

**Историзирование тематических моделей. Дистанционное чтение текстов по тематическому моделированию в рамках исторических исследований**

***Рене Брауэр***

*Гетеборгский университет, Швеция, кафедра Земли наук*

***Матс Фридлунд***

*Университет Аалто (Финляндия), кафедра инженерного проектирования и производства, Профессор*

[mats.fridlund@aalto.fi](mailto:mats.fridlund@aalto.fi)

### Введение: оцифровка инструментария историка

В ремесле историка происходят практические и методологические изменения в форме «цифровой истории». (Weller 2013) Это не первый случай, когда компьютерные методы рассматриваются как имеющие потенциал революционизировать исторические исследования. В течение в поздно 1960-е а также рано 1970-е в подниматься из клиометрия а также 'количественный история' вел в пределах история к что а Добрый из культура война сломанный вне в в профессия и шквал напряженных конференц-панелей, публичных споров и круглых столов с субтитрами, такими как «Муза и ее врачи» и «Новая и старая история». Эта культурная война противопоставила «новую» историю, на которую в значительной степени повлияли теория и методология социальных наук, более традиционным практикам нарративных историков. «Новые» историки использовали компьютеры для выполнения расчетов и соединений, которых раньше никогда не было, и их результаты временами поражали воображение. Головокруженные успехом, может быть, просто увлеченные до самоуверенности, эти историки пила маленький цель в кто угодно предложение сопротивление к их результаты или же их техники. На вызов на конференции несколько историков ответили не более чем математическим уравнением . (Томас 2004, 56)

Несмотря на успехи этой «клиометрической революции», ей так и не удалось революционизировать исторические исследования в большом масштабе, но вместо этого она добавила ценный инструмент в руки историка. инструмент коробка. Будь то в непрерывный 'цифровой история' является собирается к быть а революция или не просто очередное пополнение в арсенале историка, пока рано говорить, но тем не менее стоит пытающийся к видит это Текущий статус. К делать это мы в это бумага находятся мы собираемся проанализировать, возможно, самый центральный — и, безусловно, самый актуальный — из новых методологических инструментов в наборе инструментов цифрового историка в виде «тематического моделирование».

Тематическое моделирование является выдающимся методологическим примером подхода историка литературы Франко Моретти к «дистанционному чтению» ( литературной) истории, который он описал как « при котором «история быстро станет совсем другой , чем она есть сейчас: она станет «второстепенной». рука': лоскутное одеяло из чужих исследований, *без единого прямого текстуального прочтения* ». (Moretti 2000: 57, курсив в оригинале, см. также Moretti 2013) . чем в данные что Можно быть получил из 'Закрыть чтение' из тексты, зависит от на чтение и анализ агрегированных «метаданных» текстов : заголовков, имен авторов, лет публикации, принадлежности а также ключевые слова. Другая срок за далекий чтение является 'не чтение' (Мюллер 2007) с это связь к далекий чтение а также метаданные объяснил в в следующий путь:

Пока были книги, книг было больше, чем вы могли прочитать. В жизни профессионала или ученого чтение в сильном смысле «внимательное чтение» почти наверняка отходит на второй план по сравнению с выяснением того, что есть в книге, без фактического чтения всего или даже какой-либо ее части. Для этого существуют вековые методы, некоторые из которых более респектабельны, чем другие, и они включают беглый просмотр текста или просмотр его на глаз, чтение библиографии или следование тому, что кто-то другой говорит или пишет об этом. Зная как к "не читать" является просто в качестве важный в качестве зная как к читать. А

временный отвечать к в вопрос какие метаданные находятся хороший за, тогда, мощь сказать

что метаданные позволяют сжимать не только один текст, но в достаточно обширной среде они позволяют сжимать произвольно большие наборы текстов. И если вы используете методы визуализации — все более мощный цифровой инструмент — эти сжатые представления могут отображаться, как если бы они были местами на какой-то карте. Точно так же, как пустое пространство в книге с хорошим макетом отображает рельеф страниц и ориентирует читателей до того, как они на самом деле «прочитают», так и метаданные, «разложенные» правильным образом, могут предоставить читателям одновременный обзор многих книг и прямой доступ к ним. их внимание на области, где было бы полезно прочитать внимательно. Таково обещание «дистанционного чтения» Франко Моретти. (Мюллер, 2007 г.)

Этот исследование дайте ан обзор из в история из тема моделирование в пределах цифровой гуманитарных наук и изучить его применение в цифровой истории, а также возможные будущие методологические расширения. Мы будут также анализировать это использует в условия из различный исторический и методологический параметры: цели из расследования, какие исторический периоды Это имеет был применен к, языки, количество из темы, виды из тексты, а также виды из публикации.

**Тематическое моделирование как информатика: значение, приложения и потенциал** Тематическое моделирование (ТМ) обычно представляет собой некую форму компьютерного инструмента обработки текста, который можно использовать для постулирования сложных скрытых структур, ответственных за набор наблюдений, что позволяет использовать статистический вывод. восстановить эту структуру. Такой подход особенно полезен при работе с текстом, где наблюдаемые данные (слова) явно предназначены для передачи скрытой структуры.

(их значение). (Гриффитс и Стейверс, 2004 г., стр. 5228).

Проще говоря, *тематическая модель — это компьютерная программа, которая из текста генерирует «темы» или «темы»: строки слов, которые должны указывать на адреса тем в тексте.* Основная идея состоит в том, что слова, которые «тесно» группируются, имеют общую значимую связь, т. е. «тему», «тему» или «мотив» текста, которые в непрофессиональных терминах можно понимать как «важный» или «значительный» ключ. слова общей темы.

Огромным преимуществом TM является то, что он позволяет анализировать гораздо большие объемы данных по сравнению с традиционными подходами, позволяя использовать новые способы интеллектуального анализа данных. За пример Это бы быть практически невозможно с использованием традиционный методы к обобщить все публикации журнала *Science за* 1990–1999 гг., составляя корпус из 57 миллионов слов (Blei & Lafferty 2007). Поэтому структурирование текстового материала данных, независимо от размера, вероятно, представляет собой главное преимущество ТМ. Кроме того *, ТМ может функционировать как инструмент поиска,* намного превосходящий традиционный поиск по одному слову (Mimno 2012). Поскольку ТМ потенциально идентифицирует темы в текстах, их можно искать в корпусе, превращая его в функцию поиска. И , наконец, *ТМ может служить количественной проверкой интуиции* . Поскольку ТМ идентифицирует наиболее важные «темы» текста, ее можно использовать в качестве индикаторов того, какие темы затрагиваются (и, что более интересно, не затрагиваются) в тексте. например, сельский разработка

программный документ ЕС провозгласил себя коренным отходом от прежней политики усилия, по включая качественный из жизнь аспекты среди другие. Однако, Только а несколько идентифицированных темы раздали с эти новый вопросы, по сравнению к в традиционный сельскохозяйственная направленность. Итак, этот «перерыв» кажется в первую очередь риторическим (Brauer & Dymitrow 2013).

Сегодняшнее тематическое моделирование основано на развитии так называемого «скрытого семантического анализа» (LSA) в рамках обработки естественного языка и машинного обучения в 1990-х годах (Deerwester et al 1990). Версия тематического моделирования, наиболее часто используемая историками, - это «скрытое распределение Дирихле» (LDA), разработанное в начале 2000-х годов группой исследователей под руководством Дэвида Блея и представленное в 2003 году. Алгоритм LDA сначала удаляет так называемые «стоп-слова». из текста, например a, an, the, там, под, which и т. д., которые имеют только реляционное значение. Это ускоряет обработку и фильтрация за «значимый» темы. затем в алгоритм предполагает что каждый Документ представляет собой «набор слов», в котором совместно встречающиеся слова имеют какое-то общее значение, и на основе среднего статистического значения (например, выборки Гиббса) строятся темы. В LDA существует множество различных допущений, но три основных допущения (Blei 2012) — это следующий:

* порядок слов в анализируемом тексте не имеющий отношения
* порядок документов из анализируемого корпуса не имеющий отношения
* номер темы указан ранее известен

Эти находятся довольно смелый предположения, Однако, Это кажется что даже несмотря на это LDA может определять значимые темы (Mimno 2012). Другим алгоритмом является коррелированная тематическая модель (CTM), которая является дальнейшим развитием подхода LSA (Blei & Лафферти 2007) а также что пытается к адрес в проблема из имея к предполагать в количество тем перед анализом. CTM, в отличие от LDA, не предполагает, что темы не связаны между собой, и пытается построить «корреляции» между отдельными темами (отсюда и название). Преимущество CTM заключается в том, что количество тем не нужно указывать заранее, так как они являются результатом корреляции. Более техническая сторона исследований ТМ заключается в постоянном совершенствовании задействованных алгоритмов (Асунсьон и др. , 2011 г.; Бэйли и др. , 2011 г.; Дауд, 2012 г.; Хух и Фейнберг, 2012 г.; Цзяньпин и др. 2012).

ТМ программного обеспечения использовал по в большинство из исследователи является в на основе LDA Машинное обучение за Язык Инструментарий (МОЛОТОК) развитый по исследователи в в Университет Массачусетс-Амхерст. МОЛОТОК работает через а команда линия интерфейс, изготовление Это а в некотором роде устрашающий за люди просто получающий начал с ТМ. МОЛОТОК требует определения двух параметров, прежде чем он сможет обнаруживать темы в корпусе: количество тем и размер документа ( *фрагмента* ) в корпусе. (Джокерс 2013, 133-34) Однако, там является В данный момент нет обычно согласовано на стандартный какие эти параметры должен быть. А правило из большой палец' предложенный по Дэйвид Мимно является 100 темы с фрагментами документа по 1000 слов (Mimno in Jockers 2013, 134). Однако это должно быть адаптировано к каждому отдельному анализируемому корпусу на основе «наилучшего соответствия» конкретной ситуации; поэтому он представляет собой постоянные усилия по совершенствованию. Также разрабатываются другие реализации LDA, такие как приложение Paper Machines (Johnson-Roberson 2012).

Кроме того, еще один проблема активно работал на является найти в Лучший возможный способ, которым исследователи интерпретируют значение тем. Чанг и др. (2009) обсуждают различные статистические решения проблемы, связанной с интерпретацией тем людьми; Джокерс (2013) помогает его интерпретации, визуализируя темы в стиле, похожем на облака слов; в то время как Heuser & Le-Khac (2012), среди прочего, определяют тему в сочетании с исходным текстом, выделяя ключевые слова на документ

страницы. В любом случае это представляет собой область исследований, как в попытке определить, так и в наилучшем визуализировать в темы (Блей 2012). возможный использовать из тема моделирование в качестве а Функция поиска также является актуальной исследовательской задачей. Мимно (2012) разработал метод, в котором Это является возможный к идентифицировать темы в пределах один корпус а также поиск за их в другой . Расширение на это идея, Это становится возможный к использовать ТМ над несколько языки, используя похожий трупы в другой языки обращение в такой же проблема (например а Статья в Википедии о том же термине, Mimno et al. 2013). Другие изучаемые области — это расширение ТМ от слов до других представлений, таких как изображения, секвенирование генов или научные сетевые структуры (Li et. al. 2010; Chen et. al. 2012; Ding 2011). Последний но нет наименее находятся большой усилия преданный к улучшение в ТМ пользователь интерфейс, сделать тематическое моделирование более удобным для пользователя (Blei 2012).

### Тема моделирование в качестве история: историки обработка, моделирование, а также анализ темы

Это исследование появления тематического моделирования в исторических исследованиях принимает в качестве конечных точек первую публикацию рецензируемой журнальной статьи историка, использующего тематическое моделирование, в 2006 г. и публикацию в 2013 г. первой академической исследовательской монографии историка, использующего тематическое моделирование. тематическое моделирование. Обсуждаемые здесь исследования — это те исторические исследования, которые мы обнаружили с 2006 по 2012 год включительно.

Первая рецензируемая научная статья историка — более раннее историческое исследование (Griffiths & Steyvers, 2004) было написано двумя учеными-когнитивистами — имела название «Вероятностная декомпозиция темы американской газеты восемнадцатого века» и была написана историком Шэрон Блок вместе с ученым-компьютерщиком Дэвидом Ньюманом и опубликованы в *Журнале Американского общества информационных наук. а также Технологии (ДЖАСИСТ)* в 2006. К дайте ан пример из а также а Чувствовать за что может повлечь за собой моделирование исторической темы, мы в дальнейшем дадим более внимательное прочтение этого новаторского исследования, чем мы в последующем дадим другое исследования.

В статье применено тематическое моделирование для анализа содержания крупных американских колониальных газета *Пенсильвания Газета* между 1728-1800 гг. состоящий из около 80 000 текстов в виде статей и рекламных объявлений. ТМ использовалась в экспериментальных целях для проверки возможности ее применения для структурирования содержания газеты . Они обнаруженный что самый идентифицированный темы мы тривиальный — представляющий общие лингвистически структуры или атрибуты отдельных аспектов, или просто шум — допуская, что в интерпретация был сильно помог по а историк привычный с в предмет Дело в том, что разницу между «тривиальным» и «интересным» иногда не так-то просто определить. Анализируя типы рекламы с течением времени, они могли отображать относительные тенденции с течением времени. Например, через возникновение и последующую гибель из их ТКАНЬ тема (включая слова как; 'шелк', 'хлопок', 'то же самое', 'белый', «черный», «лен» и т. д.), они смогли укрепить предыдущее предположение о подъеме и последующем падении суконной промышленности Пенсильвании. Аналогичные тенденции могут быть установлены для расширения правительства, религии и преступности. Их вывод заключался в том, что ТМ может предоставить количественную меру для этих первоначальных более качественных исторических данных. интуиция из в период. Их вывод был что в главный преимущество заключалась в количестве документов, которые можно было охватить, по сравнению с более традиционными методами и в возможности использования ТМ в качестве количественного индикатора более широких общих тенденций.

В их статья Блокировать & Новый человек также заявил что там было "был а огромный увеличивать

в в количество из исторический начальный источники доступный онлайн. Тем не менее там имеет был маленький

проделанная работа по обработке, моделированию или анализу этих недавно доступных корпусов». (Новый человек & Блокировать 2006 г., 753) Что был в ситуация тогда а также что является Все еще в ситуации, о чем свидетельствует литературовед Мэтью Джокерс, который в 2013 году в первой исторической исследовательской монографии с использованием тематического моделирования « *Макроанализ: цифровые методы и история литературы* » (University of Illinois Press, 2013) сетует на отсутствие научных работ в области цифровых технологий. гуманитарные науки:

Безусловно , литературоведы воспользовались преимуществами оцифрованного текстового материала, но это использование было в основном в области поиска, извлечения и доступа. У нас есть нет пока что видимый в масштабирование из наш научный вопросы в соответствие с в массовое масштабирование цифрового контента, который сейчас хранится в цифровых библиотеках двадцать первого века. В это Google Книги эпоха, мы Можно брать за предоставленный что немного цифровой версия из в текст, который нам нужен, будет доступен где-то в Интернете, но мы еще не полностью сформулировали или изучили способы, которыми эти массивные корпуса предлагают новые возможности для исследований и новые способы осмысления нашего литературного предмета. (Джокерс 2013, 16-17)

Наше исследование подтверждает эту точку зрения в отношении тематического моделирования, поскольку в течение 2006–2012 годов мы обнаружили всего около двадцати исторических исследований, использующих тематическое моделирование, из которых подавляющее большинство либо остановилось на набросках возможных применений, либо исследовало метод, а не использовало его для ответа на конкретные исторические вопросы.

### Дистанционное чтение по историческому тематическому моделированию

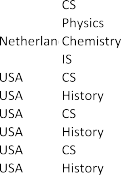
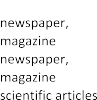
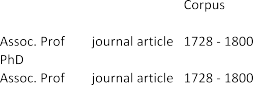
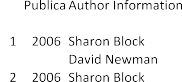
Тексты, использующие историческое тематическое моделирование, включенные в это исследование, могут показаться традиционным историкам несколько ненадежными, поскольку они выходят за рамки стандартных академических текстов. Следуя замечанию Тони Веллер о том, что «традиционные формы публикации в истории не подходят для быстро меняющихся дискурсов цифровой эпохи, — продемонстрировано по в факт что самый чистый цифровой история тексты иметь тенденцию к быть в в форма веб-сайтов, блогов, онлайн-статей и журналов, а не традиционного исторического выхода монографии» (Weller 2013, 4), мы также включили такие тексты, а также материалы конференций, если эти тексты содержат исторические исследования тематического моделирования.

тексты мы найденный по первый добыча в по в настоящее время канонический тексты из исторический исследования литературы по тематическому моделированию — например, Newman & Block 2006, Block 2006, Blevins 2010, Mimno 2012, Nelson 2011 , на которые все ссылались в текстах с использованием исторического тематического моделирования — для авторов, статей, ссылок и цитат, связанных с этими исследованиями. Этот был последовал по идет поиск через Google, Google Ученый а также Книги Google с такими ключевыми словами, как «тематические модели», «тематическое моделирование» в сочетании с «история» и «исторический», а затем авторы, статьи, ссылки и цитаты, которые были связаны с исследованиями, найденными с помощью этого, были отслежены до найти дополнительные тексты. Мы ограничились текстами в Английский.

Найденные исследования затем просматривались, чтобы понять, действительно ли они с использованием тема моделирование в Любые главный путь ведущий к исследования Только упоминание тему моделирования попутно разбирать вне.

В результате было получено 23 текста с использованием тематического моделирования, показанных в таблице 1, а также включенных в список литературы, отмеченный звездочкой (\*). Эти тексты были проанализированы в режиме отдаленного или не очень близкого чтения, поскольку мы в первую очередь анализировали не детали моделирования темы в исследованиях, а скорее более крупные закономерности. по теме моделирование пользователи а также использовать. Несмотря на то что мы намереваться к посвящать а расширенный исследование делать внимательное (r) прочтение использования тематических моделей в исторических исследованиях мы уже сейчас можем констатировать что в большинство из исторический исследования находятся прежде всего исследовательский или же предполагаемый в



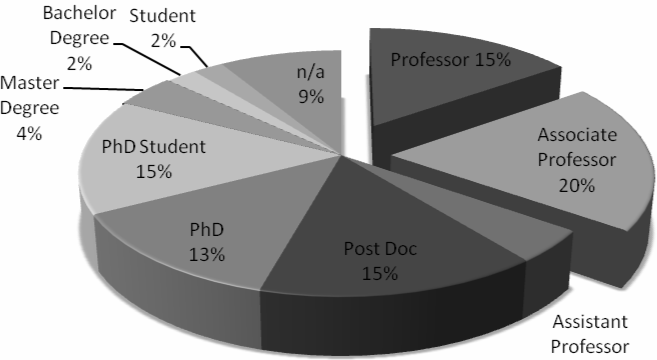


**Стол 1.** Тексты с исторический исследования 2006–2012 гг. с использованием тема моделирование. Полный ссылки можно найти в библиографии. CS означает информатику, а IS — информатику.

что Oни находятся сосредоточенный на развивающийся, тестирование или же оценка ТМ в качестве а исторический метод, а не использовать его для решения независимой исторической проблемы, что во многом соответствует обсуждаемому жалобу Джокерса выше.

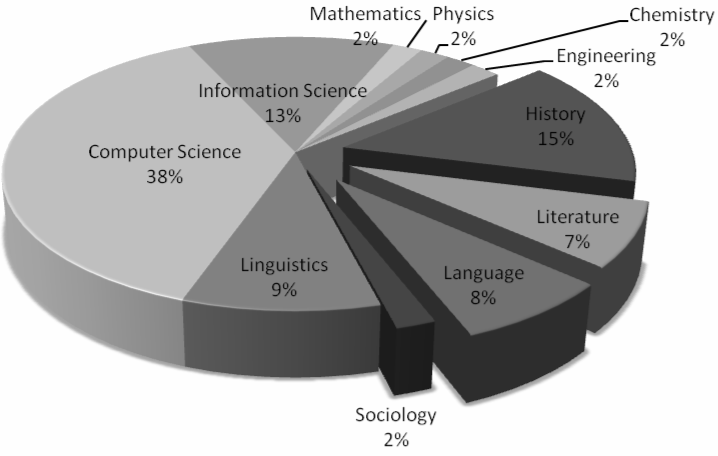
Авторы текстов и корпусов характеризовались по нескольким параметрам: академической подготовке авторов, полу, званию и стране академического учреждения; тип корпуса, язык, хронология и географическая направленность анализируемого корпуса. Далее мы приводим представление наших результатов как в виде подведения итогов обсуждения результатов, так и в виде диаграмм, которые также обсуждаются. Как и в большинстве исследований с использованием тематических моделей, многие результаты не являются неожиданными для тех, кто следит за развитием этой области.

Два таких неудивительных факта об *авторах текстов* заключаются в том, что подавляющее большинство авторов — мужчины — и только два автора с узнаваемыми женскими именами; что подавляющее большинство (92%) находятся в американских (американских и канадских) академических учреждениях, а остальные являются отдельными исследователями, находящимися в Нидерландах, Швеция, Япония а также Китай. В некотором роде возможно более удивительный является что, судя по академическому стажу авторов, это кажется игрой молодых людей, по крайней мере, 56% нештатных младших исследователей, из которых почти четверть студенты или аспиранты, и около трети авторов являются штатными или доцентами. (Рис. 1)



**Рис. 1.** Академическое образование авторов в таблице 1

Кроме того, еще одним интересным открытием является то, что оказывается, что это поле, которое очень много Все еще существование технологии управляемый в что Только 30% из в авторы имеют а дисциплинарная подготовка по гуманитарным наукам (история, литература и языки) и почти 60% приходится на технические и естественные науки (рис. 2). Это не считая 9% из авторы из лингвистика кто мог быть из либо это гуманист или же технический стороны, хотя создается впечатление, что большинство из них можно прочно отнести к техническому лагерю. Наконец, один интересный вывод заключается в том, что такая относительно большая часть (13%) текстов, использующих тематическое моделирование, составляют нестандартные академические публикации, такие как блоги и веб-сайты.



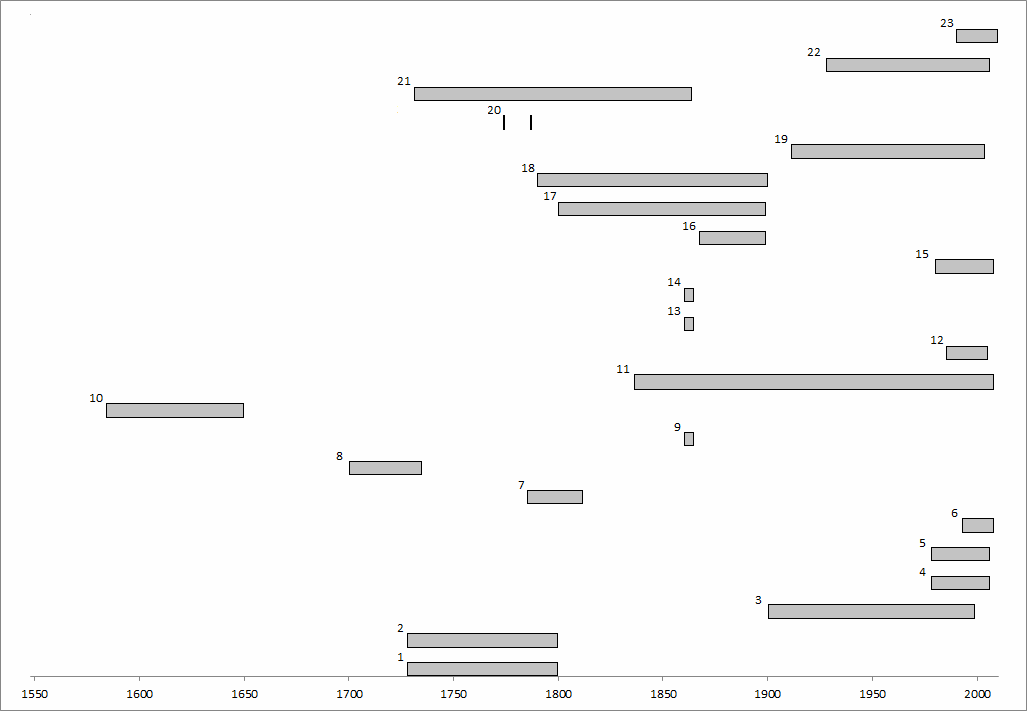
**Рис. 2:** Академическое образование авторов в таблице 1

Что касается корпусов текстов, то *и* здесь есть ожидаемые результаты. Во-первых, как сказано выше, большинство использует LDA в своей реализации MALLET и что большинство корпусов (74%) написаны на английском языке, за ним следует немецкий (13%) и, что еще более удивительно, 13% на нескольких языках. В соответствии с этим географический области в корпуса относится к находятся прежде всего в США (62%) но интересно то, что значительная часть (12%) имеют глобальный охват. Хронологический период каждого корпуса варьируется от 2 до 134 лет, но большинство (55%) составляют от 2 до 30 лет. Один из самых интересных результаты является что Это является так современный сфокусированный. другой корпуса тексты обложек между 1564–2010 гг. (Рис. 3) с а фокус на в около подарок с почти а третий запуск после 1977. Этот является также отражение в какие Добрый из корпус что является учился с два самый большой части (70%) существование научный статьи а также газеты а также более традиционный исторический материал такой в качестве романы а также рукописный тексты изготовление вверх в меньшинство.

### Заключительное обсуждение: история переоснащения?

Этот имеют состоял в а первый пытаться в направлении систематически оценка в современное состояние использования тематического моделирования в рамках прогнозируемой цифровой революции исторических исследований. В исследовании применялся дистанционный подход к чтению корпуса. из 23 тексты состоящий из исторический исследования из тема моделирование 2006-2012 гг. Хотя более внимательное прочтение использования тематического моделирования в корпусе можно оставить для будущего исследования, исследование показало, что очень немногие глубокие и исчерпывающие исторические исследования из тема моделирование Можно быть найденный. метод является В данный момент очень много ан эмерджентный метод в своем младенчество.

С методологической точки зрения тематическое моделирование достигло некоторой стабильности в что там находятся прежде всего один метод (ЛДА) а также один реализация (МОЛОТОК) то есть использовал по в большинство из пользователи. Однако, несмотря на это там находятся а большой интерес как среди компьютер ученые а также историки в развитие новый варианты а также приложение-



**Рис. 3.** Временное покрытие корпусов, где номера относятся к их идентификаторам в табл. 1.

ции (такой в качестве Бумага Машины) за тема моделирование. ТМ также показывает большой потенциал в использовании в качестве функции поиска и метода индексации. Вероятно, лучшее текущее применение ТМ — это ее применение для количественной проверки интуиции. То, что большая часть — хотя и не вся — работа, проделанная при разработке ТМ, выполняется людьми из компьютерных дисциплин, неудивительно, что может быть более удивительным то, что они также в большинстве своем разрабатывают ее для исторических целей. исследования.

Когда дело доходит до исторического обзора, многие результаты довольно неудивительны и ожидаемы, например, доминирование США и Англии. Что менее ожидаемо, так это доминирование технологий и младших исследователей. Исторические исследования с использованием тематического моделирования во многом следуют модели из естественных наук в том смысле, что пока это скорее игра молодых людей и ученых-компьютерщиков, чем игра авторитетных историков. Эта относительная нехватка опытных гуманитариев, вероятно, может в значительной степени объяснить, почему так много исследований сосредоточено на ближайшем настоящем и на развитии методов. В отличие от естественных и технических наук, в гуманитарных науках новые критические взгляды и вопросы обычно считаются плодами опытных ученых. Возможно, чего в тематическом моделировании не хватает больше, чем более сложных моделей, так это опыта, позволяющего задавать новые неожиданные вопросы. вопросы.

В настоящее время тематическое моделирование в основном разрабатывается и изучается, а не использованный в качестве а надежный исторический метод. А также несмотря на то что представляющий ан интересная и многообещающая методология для исторических приложений все еще остается решением в поисках идеальной проблемы, чтобы доказать свою ценность для историков. Или, может быть , лучше, это технология поиска исторического приложения-убийцы, которое превратит его в необходимый острый передовой инструмент в работе историка. ящик для инструментов.

### Библиография

\* **Андерсон, А., Макфарланд, Д. & Джурафский, Д.** (2012). В направлении а Вычислительная история ACL: 1980–2008 гг., *Материалы специального семинара ACL-2012 по повторному открытию 50 лет открытий,* 13-21.

\* **Эндрю, П. и Элджи-Хьюитт, М.** (2012). Эффект Вертера I: Топологически Гёте *.* В *дистанционных чтениях / описательных поворотах: топологии немецкой культуры в долгом девятнадцатом веке* . Мэтт Эрлин и Линн Тэтлок, изд. Рочестер, Нью-Йорк: Camden House, готовится к печати, 24 стр.

**Асунсьон, А., Смит, П., и Веллинг, М.** (2011). Асинхронная распределенная оценка тематических моделей для анализа документов, *Статистическая методология* , 8:1, 3-17.

**Бэйли, М., Карман, М. и Крестани, Ф.** (2011). Модель скрытой темы с несколькими коллекциями для федеративного поиска, Поиск *информации* , 14:4, 390-412.

**Блей, Д.М. (** 2012). Вероятностные тематические модели. *Сообщения ACM* , 55(4), 77-84.

\* **Блей, Д. М., & Лафферти, Джей Ди** (2007). А коррелированный тема модель из *Наука* , *Анналы прикладной статистики* , 1:1, 17-35.

**Блей, Д.М. и Лафферти, Д.Д.** (2007). Коррелированная тематическая модель *науки* , *Анналы прикладной статистики* , 1:2, 634-642.

\* **Блевинс, Д.** (2010). Тематическое моделирование Дневник Марты Баллард , *история: мысли о науке и истории в эпоху цифровых технологий ,* блог доступен по адресу: historying.org/01/04/2010/topic-modeling-martha-ballards-diary/ , по состоянию на 29 июля 2013 г.

\* **Блок, С. и Ньюман, DJ** (2011). Что, где, когда и иногда почему: интеллектуальный анализ данных. Двадцать лет истории женщин. Тезисы, *Journal of Women 's History* , 23:1, 81–109.

\* **Блок, С** . (2006). Делает Более с Оцифровка: Ан введение к тема моделирование ранних американских источников, *Common-place* : *The Interactive Journal of Early American Life* 6:2, [www.common-place.org/vol-06/no-02/tales/,](http://www.common-place.org/vol-06/no-02/tales/) доступ 2013-07-29.

**Брауэр, Р. и Димитров, М.** (2013). Цифровое моделирование регионального развития в Европе: новый методологический подход к анализу политики, *Человек-Город-Природа* [готовится к печати].

**Чанг, Дж., Бойд-Грабер, Дж., Ван, С., Герриш, С. & Блей, Д. М.** (2009). Чтение чайных листьев: как люди интерпретируют тематические модели, доступно на: www.umiacs . umd.edu/~jbg/docs/nips2009-rtl.pdf, доступ 2013-07-23.

**Чен, ИКС., Ху, ИКС., Лим, TY, Шен, ИКС., Парк, ЕК, & Розен, ГЛ** (2012). Использование функциональной и таксономической структуры геномных данных с помощью вероятностного тематического моделирования, *IEEE/АКМ сделки на вычислительный биология а также биоинформатика* , 9:4, 980-994.

**Дауд, А.** (2012). Использование моделирования темы времени для динамического повторения на основе семантики.

нахождение интереса к поиску, *Системы, основанные на знаниях* , 26, 154-163.

**Дирвестер, С., Дюме, С.Т., Фернас, Г.В., Ландауэр , Т.К. и Харшман , Р.** (1990). Индексирование с помощью скрытого семантического анализа, *Журнал Американского общества информационных наук и технологий* , 41:6, 391-407.

**Дин, Ю. (** 2011). Научное сотрудничество и одобрение: сетевой анализ

соавторство и сети цитирования, *Journal of Informetrics* , 5:1, 187–203.

\* **Эрлин, М. (** 2012). Переосмысление истории немецкой литературы 1731–1864 гг.: статистический подход, доступно по адресу: hdw.artsci.wustl.edu/node/42, по состоянию на 15 августа 2013 г. **Гриффитс, Т.Л., Стейверс, М. (** 2004). Нахождение научных тем, *Труды*

*Национальная академия наук Соединенных Штатов Америки* , 101:1, 5228-5235.

\* **Холл, Д., Джурафски, Д., Мэннинг, CD (** 2008). Изучение истории идей с использованием тематических моделей, *Материалы конференции 2008 г. по эмпирическим методам обработки естественного языка* , 363–371.

\* **Холл, DLW** (2008). *Отслеживание эволюции науки* , диплом бакалавра с отличием, Стэнфордский университет, 58 стр.

\* **Хе, К., Чен, Б., Пей, Дж., Цю, Б., Митра, П., Джайлз К.Л. (** 2009). Обнаружение эволюции темы в научной литературе: как могут помочь цитаты?, *Конференция ACM по управлению информацией и знаниями,* 957-966.

\* **Хойзер, Р., Ле-Хак, Л. (** 2012). Количественная литературная история 2958 девятнадцатого века британский Романы: Семантический Когорта Метод, *Брошюры из в Стэнфордская литературная лаборатория 4* , 66 стр.

**Ха, С., Файнберг, С.Э. (** 2012). Дискриминативное тематическое моделирование, основанное на множественном обучении, *Операции по обнаружению знаний на основе данных* , 5:4, 1-25.

**Цзяньпин, З., Цзянцзяо, Д., Вэньцзюнь, С., & Ченгронг, В.** (2012). Темы моделирование на основе выборочного распределения Ципфа, *Expert Systems With Applications* , 39:7, 6541-6546.

**Джокерс, М.Л.** (2013). *Макроанализ: цифровые методы и история литературы* . Университет Иллинойса Press.

\* **Джокерс М.Л., Мимно Д. (** 2012). Значимые темы в литературе XIX века, *Публикации факультета — кафедра английского языка* . Бумага 105. 24 стр.

**Джонсон-Роберсон, К. (** 2012). Бумагоделательные машины, *блог metaLAB (at) Harvard* , доступен по адресу metalab.harvard.edu/2012/07/paper-machines/, по состоянию на 12 июля 2012 г.

\* **Кушкулей, С. (** 2012). Анализ тенденций в Harper's Bazaar, *Workshop on Computational Linguistics for Literature* , 84–87.

**Ли, Л.-Дж., Ван, К., Лим, Ю., Блей, Д., и Фей-Фей, Л. (** 2010). Построение и использование *семантивизуальной* иерархии изображений. В *2010 г. на конференции IEEE по компьютерному зрению и распознаванию образов (CVPR)* , 3336 — 3343.

\* **Мимно, Д. (** 2012). Вычислительная историография: интеллектуальный анализ данных в классических журналах века, *Journal on Computing and Cultural Heritage* , 5: 1, 1–19.

**Мимно, Д., Уоллах, Х.М., Нарадовски, Дж., Смит, Д.А., и МакКаллум, А. (** 2009). Многоязычный Тема Модели, *Труды из в 2009 г. Конференция на Эмпирические методы обработки естественного языка* , 880-889.

**Моретти, Ф.** (2000), Догадки о мировой литературе, *New Left Review* NS, 1:1, 54–68.

**Моретти, Ф.** (2013), *Дистанционное чтение* , Лондон: Verso.

**Мюллер, М.** (2007), «Примечания к руководству пользователя Monk», *The MONK Project Wiki,* apps.lis.illinois.edu/wiki/display/MONK/Notes+towards+a+user+manual+of+M onk, 6 мая, 2007.

\* **Маллен, Л. (** 2011). Разговорный словарь: анализ текста восточных виндзорских разговорных отношений, доклад, представленный на ежегодном собрании Американского общества церковной истории, Бостон, 9 января 2011 г., доступен по адресу: lin-colnmullen.com/downloads/docs/Mullen. Vocabulary-of-Conversion.pdf, по состоянию на 29 июля 2013 г.

\* **Нельсон Р. К. (** 2011). Из Монстры, Мужчины — А также Тема Моделирование, *Новый Йорк Разъединение времен Блог* , доступный в: mindator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/?\_r=0 получен доступ 2013-07-29.

\* **Нельсон, РК (** 2011). Веб -сайт *Mining the* Dispatch, доступный по адресу: dsl.richmond. edu/dispatch/pages/home по состоянию на 29 июля 2013 г.

\* **Ньюман, Д.Дж., и Блок, С. (** 2006). Вероятностная тематическая декомпозиция восемнадцатого века американский Газета, *Журнал из в американский Общество за Информатика и технологии* , 57:6, 753-767.

\* **Ридделл, АБ** (2012). Как к Читать 22 198 Журнал Статьи: Изучение в История Немецкие исследования с использованием тематических моделей, статья, представленная на семинаре USP , «Большие данные и гуманитарные науки», вторник, 9 октября , Программа университетских ученых, Университет Дьюка , доступна по адресу: ariddell.org/static/how-to-read -n-articles.pdf, по состоянию на 29.07.2013, 24 . стр.

\* **Темплтон, ТС, Коричневый, Т., Баттачария, С., & Бойд-Грабер, Дж. (** 2012). Добыча в Отправлять под Надзор: С использованием несчастный случай Считает к Руководство Темы из Richmond Daily Dispatch Corpus, *Чикагский коллоквиум по цифровым гуманитарным наукам и компьютерам Наука* , доступный в: [www.umiacs.umd.edu/~jbg/docs/](http://www.umiacs.umd.edu/~jbg/docs/) Доступ к файлу slda\_civil\_war.pdf 2013-07-29.

**Томас, II, WG** . (2004). Компьютеры и историческое воображение, в книге «Компаньон *по цифровым гуманитарным наукам* », Сьюзан Шрайбман, Рэй Сименс и Джон Ансворт, ред., Оксфорд: Блэквелл, 56–68.

**Веллер, Т.** (2013). Введение: история в цифровую эпоху, *«История в цифровую эпоху* », Тони Веллер, изд., Лондон: Routledge, 1–19.

\* **Виттек П. и Равенек В. (** 2011). Supporting the Exploration of a Corpus of Scholarly Correspondences 17th Century by Topic Modeling, *SDH 2011 Supporting Digital Humanities: Answering the Unskable* , доступно по адресу : [www.clarin.nl/sites/de-fail/files/sdh2011-wittek-ravenek.pdf](http://www.clarin.nl/sites/de-) , по состоянию на 14 августа 2013 г., 6 стр.

\* **Ян, Т., Торгет, Эй Джей, Михалча, Р. (** 2011). Тема Моделирование на Исторический Газеты, *Материалы 5-го семинара ACL-HLT по языковым технологиям для культурного наследия, социальных и гуманитарных наук. Ассоциация компьютерной лингвистики* , 96–104

\***Yeung, C.A., Jatowt, A. (**2012). Studying How the Past is Remembered: To- wards Computational History through Large Scale Text Mining, CIKM’11, available at: [www.dl.kuis.kyoto-u.ac.jp/~adam/cikm11a.pdf,](http://www.dl.kuis.kyoto-u.ac.jp/~adam/cikm11a.pdf) accessed 2013-07-29.

# digital humanities-проекты по истории и культуре раннего нового времени и интеллектуальные сообщества современ- ности: на двух концах «галактики гуттенберга»1

## ***Владимир Макаров*** *Казанский (Приволжский) Федеральный университет (Россия),* доцент кафедры английского языка, кандидат филологических наук

�����@v��k�rov.����

Две черты, два «приема избегания� роднят современные исследования и проекты в области «цифровых гуманитарных наук� ��������� ����������, DH�� с другими современными практиками гуманитарного исследования, например, с «новым историзмом�. Во-первых, DH избегают единого определения, со-

1 Статья написана в рамках проекта РГНФ «Информационно-исследовательская база дан- ных «Современники Шекспира: электронное научное издание�« �грант № 11-04-12064 в��.