Лекция 11

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ТЕОРИЯ ЛИТЕРАТУРЫ: АНТОЛОГИЯ, ТОМ IV

Цифровые гуманитарные науки: теоретизация исследовательской практики

Цель: ознакомиться с эссе Теда Андервуда «Цифровые гуманитарные науки: теоретизация исследовательской практики» и возможностями использования современных цифровых технологий в литературоведческом анализе.

Ключевые слова: цифровой, технологии, текст, поиск, анализ.

В данной лекции разговор пойдет об эссе Теда Андервуда «Цифровые гуманитарные науки: теоретизация исследовательской практики».

Цифровые исследования предлагают литературоведам возможность по-новому обосновывать свои наблюдения и гипотезы посредством использования данных оцифрованных текстов, утверждает Тед Андервуд. Теперь ученые могут более эффективно определять использование тех или иных слов или словосочетаний в различных текстах, изменение с течением времени частоты их употребления, осуществлять сравнительный анализ между различными текстами и обнаруживать общность текстов, т. е. все то, что сложно было бы реализовать, используя нецифровые, выполненные на бумаге при помощи карандаша, исследования. В своем эссе Тед Андервуд демонстрирует, как может работать метод поиска и обработки материалов при использовании цифровых технологий.

Маленькая революция

Гуманитарии готовятся к разговору о методах цифровых исследований, которые будут интересны по многим причинам – не в последнюю очередь потому, что такие исследования очень запоздали, считает Т. Андервуд. Алгоритмическая разработка крупных электронных баз данных в течение двух десятилетий занимает центральное место в гуманитарных науках. Мы называем эту практику «поиском», но «поиск» – обманчиво скромное название для сложной технологии, которая стала играть доказательную роль в науке. Многие из функций, которые нам кажутся новыми в области интеллектуального анализа, незаметно прижились в нашей дисциплине, ведь гуманитарии начали использовать полнотекстовый поиск в 1990-х годах. Хотя интеллектуальный анализ данных широко представлен как новая технология, которая сейчас импортируется в гуманитарные науки, Т. Андервуд утверждает, что ее лучше понимать как философский дискурс, который может помочь гуманитариям более тщательно и предметно думать о существующих методах алгоритмического исследования.

Во-первых, что значит сказать, что поиск выполняет «доказательную роль в науке»? Появление парадокса здесь отчасти вызвано словом «поиск», которое размывает границы между различными технологиями. Библиографический поиск может быть немного большим, чем просто помощь для памяти – например, если ученый осуществляет поиск по известному ему названию. Полнотекстовый поиск выглядит похожим: мы можем вводить поисковые запросы в том же поле, где мы бы ввели заголовок. Но основная технология и методы ее применения различны.

На практике полнотекстовый поиск часто является логической смысловой операцией для набора документов, которые могут или не могут существовать. Например, если Т. Андервуд предполагает, что слово «румянец» является показателем нравственного сознания в поэзии XIX века, он обращается к базе данных первичных источников и ищет стихи, содержащие как слово «румянец», так и слово «сознание». Если он находит достаточно примеров, то излагает эти данные в статье. Если нет, он обычно продолжает свои попытки. Возможно, сочетание слов «румянец» и «стыд» будет более результативным?

Здесь поиск – это не просто помощь в быстроте обработки материалов; это аналог эксперимента – хотя, конечно, есть кое-что немного сомнительное в экспериментах, которые повторяются до тех пор, пока не дадут желаемый результат. Поисковые термины, которые выбрал Т. Андервуд, кодируют подразумеваемую гипотезу о литературном значении символа или слова, и гипотеза подтверждается, когда исследователь получает достаточное количество удачных попыток. Возможно, в статье, которую в результате напишет ученый, будет ссылка только на некоторые из этих источников, потому что, возможно, для раскрытия данной проблемы

нет необходимости в «больших данных». Но на самом деле исследователь использовал алгоритмы для изучения большого набора данных, и в процессе поиска был сформирован способ подачи материала или подтверждены его интуитивные предположения относительно репрезентативности источников.

Внутренняя математика полнотекстового поиска также имеет больше общего с интеллектуальным анализом данных, чем с библиографическим поиском. Если, приводит пример Т. Андервуд, он осуществляет поиск по названию «Моби Дик», результаты будут легко сканироваться. Но в полнотекстовом поиске часто бывает много совпадений, чтобы пользователь мог увидеть их все. Вместо этого алгоритм должен сортировать их в соответствии с некоторой степенью релевантности.

Словом, полнотекстовый поиск не является поисковым средством, аналогичным карточному каталогу. Это – название для большого семейства алгоритмов, которые гуманитарии использовали в течение нескольких десятилетий для проверки гипотез и сортировки документов, подтверждающих их гипотезу. Простые формы полнотекстового поиска стали доступны уже в 1970-х годах (Лексис был ранним примером), но базы данных CD-ROM исторических источников не были широко распространены до 1990-х годов. Даже сегодня технология, возможно, не проникла в исторические дисциплины так же основательно, как в случае с литературоведением, поскольку историки больше полагаются на неопубликованные источники. Однако одно недавнее исследование дает основание предполагать, что гуманитарии по целому ряду дисциплин в значительной степени опираются на поисковые системы и используют их для исследований способами, которые не сильно отличаются от того, как их применяет широкая публика. Научные последствия поисковых методов трудно оценить, так как ученые склонны скрывать описание процесса своего открытия в опубликованной работе. Но, будучи в числе тех, кто начинал писать диссертацию непосредственно перед появлением полнотекстовых баз данных, Т. Андервуд помнит, что, как казалось, заканчивал её в другом мире.

Самым очевидным результатом новой технологии было то, многие литературоведы в 90-х годах обнаружили, что они пишут о более широком диапазоне первичных источников. Т. Андервуд предполагает, что и определяемые для разработки научные проблемы также изменились, поскольку теперь они решаются благодаря использованию возможностей полнотекстового поиска. Например, до 1990 года узко определенные темы представляли для Т. Андервуда сложность: не было предметного каталога библиотеки Конгресса для «описания работы как «энергии» в британской романтической эпохе». Полнотекстовый поиск сделал эту тему для исследования до смешного легкой.

Т. Андервуд вспоминает, что ему даже стало неловко. Правила исследовательской игры, похоже, изменились таким образом, что было невозможно проиграть. В конце концов, сколько источников вам необходимо для определения важности темы? Двадцать? Когда поисковые запросы были ограничены сетями предыдущих ссылок и упоминаний, это была высокая цифра. Но в базе данных, содержащей миллионы предложений, полнотекстовый поиск может отобразить двадцать примеров чего угодно. Даже в то время было ясно, что это может усилить смещение, предвзятость аргументации гипотезы.

«Подводные камни»

Оглядываясь назад, Т. Андервуд понимает, что недооценил масштаб проблемы. Это правда, что полнотекстовый поиск может подтвердить почти любой тезис, с которым вы работаете, но это может быть не самая опасная проблема. Проблема посерьезнее заключается в том, что источники сортировки в соответствии с вашим запросом также отфильтровывают все альтернативные тезисы, с которыми вы не работаете. Поиск – это форма интеллектуального анализа данных, но странно сфокусированная форма, которая показывает только то, что вы уже знаете и чего ожидаете. Это ограничение было бы проблемой в любой области, но особенно острой в исторических исследованиях, поскольку другие периоды не всегда организуют свои знания способами, которые мы находим наглядными. Наши догадки о поисковых терминах могут проецировать современные нам ассоциации и оставлять не раскрытыми незнакомые модели мышления.

Гуманитарии не тратили много времени на обсуждение этой проблемы в 1990-х годах, потому что поисковые системы обычно были нашим единственным доступом к крупным цифровым базам. Но в последние годы практика исследований диверсифицирована, и герменевтические ограничения поиска становятся очевидными.

Ученые в области компьютерных наук, подполя интеллектуального анализа данных и машинного обучения специализировались на проблеме извлечения искомой информации из большой подборки данных. Они разработали ряд альтернатив поиску, основанных на более осознанном, философски строгом учете интерпретации. Т. Андервуд признает, что слова «философски строгий учет интерпретации» могут быть настоящим откровением. Гуманитарии склонны думать об информатике как об инструментальном, а не философском дискурсе. Основополагающий язык интеллектуального анализа данных – байесовская статистика – это способ рассуждения об интерпретации, который может помочь нам приблизиться к крупным базам текстов более принципиальным способом. В частности, байесовская статистика подчеркивает герменевтическую спираль, которая хорошо знакома гуманитариям, поскольку мы подходим к каждому вопросу с некоторыми предыдущими предположениями (называемыми «предыдущими вероятностями»), а также с конкретными видами неопределенности. Когда мы сталкиваемся с новыми доказательствами, наша интерпретация сразу формируется существующими предположениями и способна их перестроить. Этот герменевтический цикл во многом интуитивен, когда речь идет об одном тексте; задача интеллектуального анализа данных заключается в том, чтобы объяснить, как он может работать с таким огромным набором текстов, который не может быть изучен одним читателем. Все стратегии сопоставления данных будут делать некоторые предположения о шаблонах, которые мы ожидаем найти. Но некоторые стратегии также способны выявить доказательства, которые оспаривают, опровергают предшествующие предположения.

Например, привычка литературоведов использовать поиск по ключевым словам для поиска пересечений тем (например, «румянец/стыд» или «работа/энергия») неявно основана на предположении, что совпадение слов выявит связь между их значениями. Это предположение связано с моделью смысла, называемой лингвистами «гипотезой распределения», которая гласит, что значение слова связано с его распределением по контекстам. Это может быть не идеальная модель, но она оказалась полезной в информатике, а также в литературном исследовании, и если мы хотим продолжать использовать ее в качестве эвристики, есть более гибкие способы ее использования, чем многократные попытки угадывания отдельных пар слов. Алгоритмы, основанные на предположениях распределения, могут отображать язык, который был на практике связан с любым термином за данный период. Например, слово, наиболее часто ассоциируемое с «румянцем» в коллекции текстов из 4 тыс. 8-сот 20-ти сборников XVIII и XIX столетий, не «стыд», а «бесхитростный» – деталь, которая могла бы интересным образом усложнить предположения ученых о нравственном сознании, если они используют стратегию поисков, достаточно гибкую, чтобы выявить подобные детали. Такие стратегии сопоставления не заменят поиск по ключевым словам для всех целей. Когда вы уже знаете, что ищете, поисковая система является подходящим инструментом. Но в исторической науке бывают случаи, когда мы не знаем, что ищем, хотя и думаем, что знаем.

На самом деле, признает Т. Андервуд, будет поспешным предположить, что тема, которую он изучает, может быть связана с одним словом, вроде «румянец». Возможно, другой термин, который он не может угадать, был более важным в этот период или, возможно, социальные явления, имеющие отношение к его вопросу, формируются на пересечении многих разных терминов. Если нам нужна более открытая стратегия, мы можем сопоставить печатную запись, позволяя алгоритму организовать слова коллекции в кластеры терминов, которые имеют тенденцию встречаться в одних и тех же контекстах. Эта стратегия (известная как «моделирование темы») способна выявлять дискурсивные шаблоны, которые исследователь и не искал.

Поскольку моделирование темы позволяет слову «принадлежать» к более чем одной «теме», оно может выявлять закономерности ассоциаций, которые меняются во времени.

Наблюдения, проводимые более глубоко, могут привести к интересным результатам.

Мэтью Джокерс использовал тематическое моделирование для составления романов XIX века; Роберт К. Нельсон использовал его, чтобы соотнести тематические акценты в газетах эпохи Гражданской войны в связи с меняющимися обстоятельствами военных действий. Эндрю Голдстоун и Тед Андервуд использовали эту технику для определения увеличения и уменьшения числа различных критических терминов в литературоведении XX века.

Вместо подробного разговора о тематическом моделировании Т. Андервуд предлагает рассмотреть, как инновации такого рода вызывают запоздалый разговор об алгоритмическом исследовании в целом. Тематическое моделирование будет и должно стать предметом дискуссии, и это следовало сделать 20 лет назад.

Исследователи не могут позволить себе рассматривать алгоритмы как закрытые и таинственные «черные ящики». Если мы будем использовать алгоритмы в наших исследованиях, мы должны взломать их и выяснить, как они работают. К счастью, тематическое моделирование не является запатентованным, как многие алгоритмы при веб-поиске. Алгоритмы тематического моделирования являются общедоступными, и гуманитарии смогли изменить их в соответствии со своими целями.

Чтобы понять интерпретационные ограничения алгоритма, нужно понять его математическую основу. Например, в наиболее распространенной форме тематического моделирования количество разработанных тем является одним из исходных предположений, вносимых в процесс моделирования. Как следствие, алгоритм не может дать достаточно веских и однозначных ответов о единстве любого дискурса или о его границах. Всегда можно моделировать одну и ту же коллекцию с большим или меньшим количеством тем, чтобы разбить или разделить результаты по-разному. С другой стороны, алгоритм достаточно хорош в выявлении закономерностей ассоциаций, которые мы могли бы не заметить и проигнорировать.

Использование алгоритмов для исследования вызывает ряд интересных, но незнакомых философских вопросов. Гуманитариям больше нравится применять количественные методы, особенно когда те можно представить в их привычной роли в качестве инструментов проверки на поздних этапах исследований. Использование алгоритма в качестве источника исходных данных похоже на вытягивание кролика из шляпы (несмотря на то, что мы делали подобное с поисковыми системами в течение нескольких десятилетий).

Например, в недавнем выпуске «Публикации Ассоциации современного языка» Алан Лю задает вопрос о том, стремятся ли моделисты (те, кто использует метод моделирования) к цели «интерпретации табула раса – инициирование интерпретации посредством обнаружения явлений без выдвижения гипотезы». Если бы это было так, это создало бы реальный философский тупик. И, конечно же, в журналах можно найти технофильные «рапсодии», которые говорят о том, что мы зашли в такой тупик: завершающая фаза, где «данные» окончательно вытесняют всю «теорию».

Однако эти «рапсодии» недостаточно хорошо информированы о статистических моделях, используемых при анализе данных. Это не тот случай, когда тематическое моделирование (или любой другой алгоритм интеллектуальной обработки данных) претендует на то, чтобы быть действительно «без гипотез». Модель – это абстракция, созданная людьми, и специалисты в области теории вычислительных машин и систем уже давно это признают. Байесовские вероятностные модели, распространенные в настоящее время в этой дисциплине, особенно тщательны в определении исходных интерпретационных предположений.

Исследователь, который хочет вписать тематическую модель в коллекцию документов, должен начать с указания, например, количества тем, которые он ожидает найти, и степень размытости, которую он ожидает от этих тем. В процессе моделирования компьютер не генерирует идеи из ничего; его расчеты скорее являются способом гармонизации этих исходных предположений человека со сложными доказательствами, представленными самими документами (иначе говоря, компьютер помогает нам «подогнать» модель к доказательствам). Этот способ исследования может быть более открытым, чем поиск по ключевым словам, поскольку предположения о степени размытости более гибки, чем конкретное предположение о том, что, скажем, румянец символизирует стыд. Но процесс интерпретации все еще формируется и инициируется предположениями человека.

Гуманитарии анализируют большие массивы данных. Проблема в том, говорит Т. Андервуд, что мы используем алгоритмы поиска, которые мы никогда не обсуждали, не делали их проблемой для размышления, и, возможно, используем их слишком проецирующим методом. Хотя статистический язык информатики может показаться чуждым нашей дисциплинарной традиции, Т. Андервуд считает, что гуманитариям нужно понять этот язык, чтобы разработать исследовательские методы, которые позволят нам работать с большими объемами данных, оставаясь верными нашим собственным герменевтическим принципам.

Этот новый для гуманитариев вид междисциплинарного разговора позволит нам узнать много полезного. Но также и мы можем внести свой вклад. Т. Андервуд предполагает, что у точных дисциплин есть своя полезная версия герменевтической теории, но и она не без недостатков. Трудность моделирования исторических изменений, например, не совсем понятна вне гуманитарных наук. Ученые, которые пытаются смоделировать печатные документы за значительный промежуток времени, часто делают предположения о непрерывности, которую гуманитарии признают ограниченной. Таким образом, появляется редкая возможность для подлинно продуктивного обмена научной методологией и гуманитарной теорией по этой и многим другим темам.

Вопросы для закрепления темы:

1. В чем принципиальное отличие традиционной и цифровой обработки художественного текста?
2. Какие новые возможности открывают перед литературоведческой наукой цифровые технологии?
3. В чем достоинства и недостатки метода цифровой обработки большого корпуса текстов?

Литература:

1. Coenen Frans Data Mining: Past, Present, and Future// Knowledge Engineering Review. – 2011. – № 26.
2. Bovens Luc, Hartmann Stephan. Bayesian Epistemology. – Oxford, 2003.
3. Kruschke K. John. Doing Bayesian Data Analysis. – Burlington, MA, 2011.
4. Sahlgren Magnus. The Distributional Hypothesis//Rivista di Linguistica. – № 20.
5. Turney D. Peter, Pantel Patrick. From Frequency to Meaning: Vector Space Models of Semantics// Journal of Artificial Intelligence Research. – 2010. – № 37.
6. Underwood Ted. Mapping Mutable Genres in Structurally Complex Volumes: International Conference on Big Data-2013//arXiv.org (Cornell University Library, http://arxiv.org/abs/1309.3323).
7. Jockers L. Matthew. Macroanalysis: Digital Methods and Literary History. – Urbana, 2013.
8. Nelson K. Robert. Mining the Dispatch, Digital Scholarship Lab// University of Richmond http://dsl.richmond.edu/dispatch/.
9. Goldstone Andrew, Underwood Ted. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us//New Literary History (forthcoming) https://www.ideals.illinois.edu/handle/2142/49323.