Мультимодальные тематические модели для разведочного поиска в коллективном блоге*

А. О. Янина[1,2]***, К. В. Воронцов***[1,2]
yanina-n@yandex-team.ru; vokov@forecsys.ru
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]Яндекс, Россия, г. Москва, ул. Льва Толстого, 16

Разведочный информационный поиск нацелен на приобретение и систематизацию профессиональных знаний в отличие от поисковых систем, отвечающих на короткие запросы массовых пользователей. Для него характерно отсутствие как точной формулировки запроса, так и единственного правильного ответа. В данной работе предлагается технология тематического разведочного поиска. Рассматривается задача поиска тематически близких документов по текстовому запросу произвольной длины. Применение аддитивной регуляризации тематических моделей (ARTM — additive regularization for topic modeling) позволяет комбинировать требования различности тем и разреженности векторных тематических представлений документов, а также учитывать дополнительные данные об авторах и категориях документов. Для построения тематических моделей используется библиотека с открытым кодом BigARTM. Предлагается методика оценивания точности и полноты тематического поиска на основе оценок асессоров. Эксперименты на данных коллективного блога habraharb.ru показывают, что качество тематического поиска сравнимо с качеством асессорского поиска и даже несколько превосходит его по критерию полноты, при этом асессоры тратят в среднем по 30 мин на каждый тематический запрос, тогда как тематическая поисковая система выдает результат практически мгновенно.

Ключевые слова: *информационный поиск; разведочный поиск; тематическое моделирование; аддитивная регуляризация тематических моделей; BigARTM*

DOI: 10.21469/22233792.2.2.04

1 Введение

Современные поисковые системы отвечают на короткие четко сформулированные запросы массового пользователя. Исследовательский или *разведочный поиск* (exploratory search) — это относительно новая парадигма в информационном поиске, нацеленная на самообразование, приобретение и систематизацию знаний [1,2]. Потенциальные пользователи разведочного поиска — исследователи, преподаватели, студенты, специалисты различных профессий, работа которых связана с накоплением и анализом информации. Переход к обществу, основанному на знаниях, приводит к расширению информационных потребностей людей и необходимости создания принципиально новых инструментов поиска.

Основной особенностью разведочного поиска является отсутствие точной формулировки запроса и отсутствие единственного ответа. Когда пользователь плохо ориентируется в терминологии или слабо представляет себе структуру предметной области, его первейшей информационной потребностью становится получение «дорожной карты» предметной области, определение наиболее важных тем, систематизация и визуализация релевантной информации по этим темам. В этих случаях трудно или вообще невозможно сформулировать запрос в виде короткой текстовой строки. Проще наметить направление поиска,

*Работа выполнена при финансовой поддержке РФФИ, проекты 16-37-00498, 14-07-00847 и 14-07-00908.

задав в качестве запроса большой фрагмент текста, документ или подборку документов. Целью разведочного поиска является получение ответов на сложные вопросы: «какие темы представлены в тексте запроса», «что читать в первую очередь по этим темам», «что находится на стыке этих тем со смежными областями», «какова структура данной предметной области», «как она развивалась во времени», «каковы последние достижения», «где находятся основные центры компетентности», «кто является экспертом по данной теме» и т. д. Пользователь обычной поисковой системы вынужден итеративно переформулировать свои короткие запросы, расширяя зону поиска по мере усвоения терминологии предметной области, периодически пересматривая и систематизируя результаты поиска. Это требует больших затрат времени и высокой квалификации. При отсутствии инструмента для получения «общей картины» всегда остается сомнение, что какой-то важный аспект изучаемой проблемы так и не был найден. Если образно представить итеративный поиск как блуждание по лабиринту знаний, то разведочный поиск — это средство автоматического построения карты любой части этого лабиринта.

Исследования разведочного поиска можно условно разделить на несколько направлений: изучение поведения пользователей обычных поисковых систем, разработка системных архитектур, сценариев и средств визуализации для разведочного поиска, развитие и применение методов кластеризации, категоризации и семантического анализа текстов.

Отдельным направлением работ в области разведочного поиска является создание размеченных коллекций для оценивания качества поиска [3–5]. В [3] методы оценивания качества разведочного поиска делятся на две большие группы: user-centered и system-centered. Подходы, учитывающие пользовательское поведение в процессе поиска, являются наиболее сложными и ресурсоемкими, но позволяют более точно оценивать качество поиска. Например, в [5] с помощью машинного обучения строится предсказательная модель действий пользователя в ходе разведочного поиска. Признаки для обучения классификатора генерируются из данных об информационной потребности пользователя и о полноте доступной информации по заданной теме. Информационная потребность пользователя определяется по числу запросов, длине каждого запроса, числу слов в запросе, энтропии запроса. Покрытие определяется тремя признаками: числом посещенных веб-страниц; числом страниц, на которых пользователь провел больше 30 с; числом сохраненных страниц. Кроме того, учитываются такие признаки, как «эффективность пользователя» и «интерпретируемость запроса». Далее по этим признакам настраивается предсказательная модель близости пользователя к требуемому результату поиска. Качество разведочного поиска оценивается по шкале от 1 до 4.

В данной работе рассматриваются методы *тематического разведочного поиска*. В их основе лежат следующие предположения: (1) в коллекции текстов, написанных на естественном языке, можно выделить относительно небольшое число тем, много меньшее числа слов и числа документов; (2) каждая тема представляется своим лексиконом — семантически однородным частотным словарем слов и выражений; (3) семантика каждого документа представляется частотным списком тем. *Вероятностные тематические модели* (probabilistic topic models) формализуют эти предположения, представляя каждую тему дискретным распределением вероятностей на множестве слов, а каждый документ — дискретным распределением вероятностей на множестве тем [6–8].

Векторные семантические описания позволяют решать перечисленные выше задачи тематического поиска. Одна из основных — задача поиска тематически близких документов. Системы полнотекстового поиска основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [9]. Поисковая система ищет

документы, содержащие все слова запроса, поэтому по длинному запросу, скорее всего, ничего не будет найдено. Тематическая поисковая система обходит эту проблему. Она строит тематическую модель коллекции документов, преобразуя каждый документ в относительно короткий частотный список тем. Текст запроса, каким бы длинным он ни был, также преобразуется в короткий список тем. Таким образом, для поиска документов схожей тематики применимы те же механизмы индексирования, поиска и ранжирования, только в роли слов выступают темы.

В тематическом разведочном поиске нет итеративного переформулирования запросов, поэтому нет необходимости в сложных методиках оценивания поведения пользователей. В данной работе для измерения качества поиска используются обычные критерии точности и полноты на основе оценок асессоров и предлагается методика формирования выборки запросов для тематического поиска.

В литературе по разведочному поиску тематическое моделирование стали использовать относительно недавно [10–13], а многие обзоры о нем вообще не упоминают [14–19]. В недавней статье [13] важными преимуществами тематических моделей называются гибкость, возможности визуализации и навигации. В то же время, в качестве недостатков отмечаются проблемы с интерпретируемостью тем, трудности с модификацией тематической модели при поступлении новых документов и высокая вычислительная сложность. Эти проблемы относятся к устаревшим методам и успешно решены в последние годы: десятки новых моделей разработаны для улучшения интерпретируемости; онлайновые алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [20–22]. С другой стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации моделей [23, 24]. Однако эти исследования пока не привели к созданию общедоступных систем разведочного поиска. Тенденция к сближению этих двух научных направлений наметилась лишь в последние годы.

В системах разведочного поиска к тематическим моделям предъявляется нетривиальная совокупность требований. Они должны автоматически строить существенно различающиеся и хорошо интерпретируемые темы; определять оптимальное число тем или иерархически разбивать темы на подтемы; учитывать не только отдельные слова, но и тематически значимые словосочетания; учитывать разнородные метаданные документов: авторство, время, категории, теги. Этим требованиям по отдельности удовлетворяют различные байесовские тематические модели [8, 25]. Однако комбинирование моделей в байесовском подходе наталкивается на значительные технические трудности. Поэтому в данной работе используется небайесовский многокритериальный подход *аддитивной регуляризации тематических моделей*, ARTM [26].

Для комбинирования моделей, разнородных требований и источников данных в ARTM ставится задача оптимизации взвешенной суммы критериев правдоподобия и регуляризаторов. Независимо от выбранного сочетания регуляризаторов, задача решается одним и тем же регуляризованным EM (expectation-maximization) алгоритмом. Это позволило сочетать модульную технологию тематического моделирования с высокоэффективным параллельным онлайновым EM-алгоритмом в библиотеке с открытым кодом BigARTM (`bigartm.org`) [27]. В предшествующих работах было показано, что ARTM позволяет улучшать интерпретируемость тем одновременно с разреживанием модели и выделением слов общей лексики [28, 29], отбрасывать зависимые и неинформативные темы [30], обрабатывать разнородные документы, содержащие наряду со словами токены различных модальностей [22], использовать словари ключевых слов для выделения узко специализирован-

ных тем, в частности для изучения межнациональных отношений по данным социальных сетей [31].

В данной работе предлагается мультимодальная регуляризованная тематическая модель для разведочного поиска. Для построения модели используется комбинация регуляризаторов, встроенных в библиотеку BigARTM. Предлагается методика оценивания точности и полноты тематического поиска на основе оценок асессоров. С помощью данной методики обосновывается выбор числа тем и дополнительных модальностей.

2 Вероятностное тематическое моделирование

Пусть D — конечное множество (коллекция) текстовых документов, T — конечное множество тем, M — конечное множество модальностей. Каждой модальности $m \in M$ соответствует словарь — конечное множество токенов W_m. Примерами модальностей являются слова, биграммы, теги, категории, авторы, метки времени. Обозначим через W объединение непересекающихся множеств W_m по всем модальностям $m \in M$. Каждый документ $d \in D$ представляет собой последовательность токенов $w_1, \ldots, w_{n_d}$ из W, где n_d — длина документа. Принимая «гипотезу мешка слов», будем считать, что последовательность токенов не важна, и учитывать только число вхождений n_{dw} токена w в документ d.

Вероятностная тематическая модель описывает условную вероятность появления токенов w в документе $d \in D$ как вероятностную смесь распределений $\varphi_{wt} = p(w|t)$ токенов в темах $t \in T$ с коэффициентами $\theta_{td} = p(t|d)$, зависящими от документов:

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \quad w \in W_m,\ d \in D.$$

Матрицы $\Phi = (\varphi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$ будем использовать для обозначения параметров тематической модели.

В вероятностном латентном семантическом анализе (PLSA — probabilistic latent semantic analysis) используется единственная модальность терминов (как правило, отдельных слов) и ставится задача максимизации логарифма правдоподобия модели $p(w|d)$ при ограничениях неотрицательности и нормировки столбцов матриц Φ и Θ [6].

В аддитивной регуляризации тематических моделей (ARTM) критерий логарифма правдоподобия вводится для каждой модальности и максимизируется их взвешенная сумма [22,29]. В общем случае данная задача имеет бесконечно много решений, поэтому к этой сумме добавляются дополнительные критерии-регуляризаторы R_i:

$$\sum_{m \in M} \frac{\tau_m}{n_m} \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

где $n_m = \sum_{d \in D} \sum_{w \in W_m} n_{dw}$ — нормировочный множитель для балансировки модальностей. Задача оптимизации решается с помощью регуляризованного EM-алгоритма [22,29]. Веса модальностей τ_m и коэффициенты регуляризации τ_i подбираются в эксперименте.

Регуляризатор сглаживания вводит в модель требование, чтобы распределения φ_{wt} и θ_{td} были похожи на заданные распределения β_w и α_t соответственно [28]:

$$R(\Phi, \Theta) = \beta \sum_{m \in M} \sum_{t \in T} \sum_{w \in W^m} \beta_w \ln \varphi_{wt} + \alpha \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Регуляризатор разреживания имеет такой же вид, но коэффициенты регуляризации β и α отрицательны, что способствует появлению нулевых элементов в распределениях φ_{wt} и θ_{td}. Эксперименты показывают, что в тематических моделях возможно сильное разреживание матриц Φ и Θ, до 90%–98%, практически без потери качества модели [29].

Разреживание матрицы Θ, сохраняющее остальные критерии качества модели, важно для тематического поиска, так как это позволяет находить более компактные семантические представления документов и запросов.

Регуляризатор декоррелирования минимизирует ковариации между вектор-столбцами матрицы Φ, повышая различность тем и улучшая интерпретируемость модели [32]:

$$R(\Phi) = -\tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \to \max .$$

Этот регуляризатор имеет побочный эффект разреживания матрицы Φ, поэтому отдельный регуляризатор разреживания для Φ совместно с ним можно не применять.

3 Выбор стратегии регуляризации

Эксперименты проводились на коллекции из 132 157 статей коллективного блога `habrahabr.ru`. Кроме основной модальности терминов (52 354 слова) использовались следующие модальности: 524 авторов статей, 10 000 комментаторов (авторов комментариев к статьям), 2546 тегов, 123 хаба (категории).

Терминами считались слова длиной больше двух букв. Из числа терминов были исключены слова общей лексики — 5% самых высокочастотных слов в коллекции. Предварительная обработка текстов включала в себя удаление пунктуации, приведение слов к нижнему регистру, замену буквы «ё» на букву «е», лемматизацию при помощи морфологического анализатора pymorphy2.

Всего на Хабрахабре 693 509 пользователей, но из них большая часть только читает и комментирует статьи, не размещая собственные статьи в блоге. Поэтому в качестве комментаторов были выбраны 10 000 активных пользователей следующих трех категорий: авторы хотя бы одной статьи; авторы не менее десяти комментариев с лайками других пользователей; пользователи из групп «старожилы», «звезды», «легенды» и «авторы», составляющие ядро аудитории Хабрахабра.

Тематическая модель строилась с помощью библиотеки BigARTM. Столбцы матрицы Φ инициализировались по умолчанию случайными распределениями, столбцы матрицы Θ — равномерными. В каждую тематическую модель были включены три регуляризатора: декоррелирование распределений терминов в темах (с коэффициентом τ), разреживание распределений тем в документах (с коэффициентом α), сглаживание распределений терминов в темах (с коэффициентом β). Регуляризаторы добавлялись в модель в указанном порядке один за другим. При добавлении каждого регуляризатора его коэффициент регуляризации выбирался из заданной сетки значений по нескольким критериям качества. Для каждого значения коэффициента регуляризации производилось 8 итераций EM-алгоритма. Из всех значений выбиралось то, при котором улучшался хотя бы один из критериев без существенного ухудшения остальных. При этом коэффициенты предыдущих регуляризаторов не изменялись.

Для оценивания модели использовались следующие критерии качества [29]: перплексия, разреженность распределений тем в документах, разреженность распределений токенов в темах для каждой из пяти модальностей — термины, авторы, комментаторы, теги, хабы. Под разреженностью понимается доля нулевых элементов в матрице распределений.

На рис. 1 показаны зависимости перплексии и разреженности от числа итераций при различных значениях коэффициентов регуляризации. В результате была выбрана совокупность коэффициентов регуляризации $\tau = 10^8$, $\alpha = -1{,}5$, $\beta = 0{,}5$. Жирной кривой выделена наилучшая траектория регуляризации.

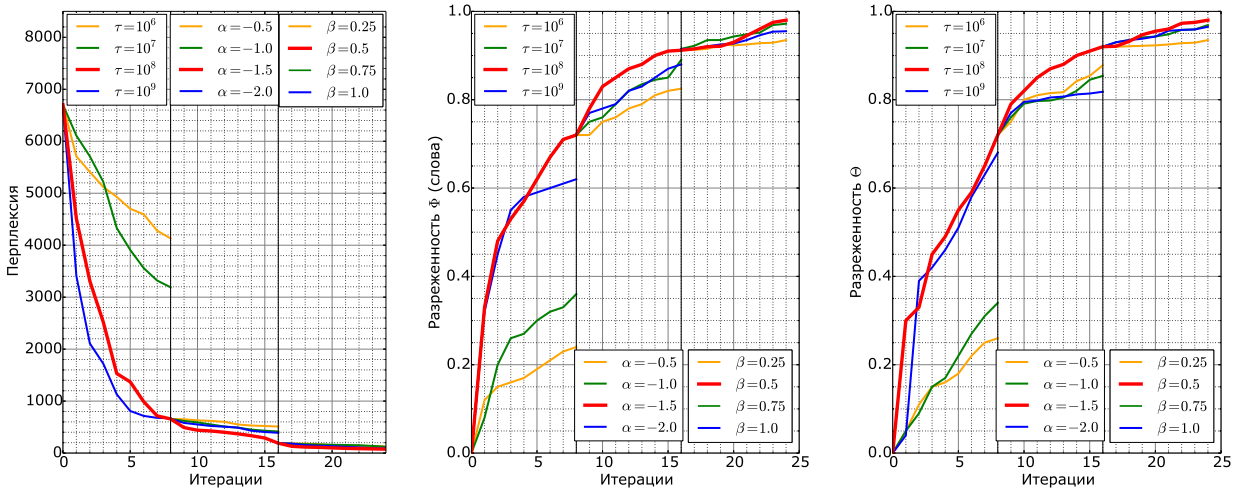


Рис. 1 Зависимости перплексии и разреженности матриц Θ и Φ (только по модальности терминов) от числа итераций и коэффициентов регуляризации.

Веса модальностей τ_m также подбирались по сетке методом проб и ошибок, по критериям перплексии, разреженности и качества тематического поиска (см. ниже). В итоге были подобраны следующие значения τ_m: 1,0 для терминов; 0,5 для авторов; 0,75 для комментаторов; 15,0 для тегов; 10,0 для хабов.

4 Разведочный тематический поиск

Допустим, тематическая модель коллекции уже построена, имеется матрица Φ распределений терминов в темах и текст запроса $q = (w_1, \ldots, w_{n_q})$. Построим для него распределение $\theta_{tq} = p(t|q)$, запустив тематическое моделирование документа q при фиксированной матрице Φ (библиотека BigARTM поддерживает такой режим запуска). Отранжируем документы коллекции $d \in D$ по убыванию косинусной меры близости к запросу q:

$$\text{cosine\_similarity}(q, d) = \frac{\sum_t \theta_{tq}\theta_{td}}{\left(\sum_t \theta_{tq}^2\right)^{1/2}\left(\sum_t \theta_{td}^2\right)^{1/2}}.$$

В качестве выдачи тематического поиска возьмем k документов с векторами θ_d, самыми близкими к вектору запроса θ_q. Число k является параметром поискового механизма или процедуры оценивания качества поиска.

Для оценивания качества тематического поиска предлагается следующая методика, основанная на асессорских оценках релевантности.

Шаг 1. Организатор тестирования составляет множество запросов, соответствующих тематике коллекции. Запросы могут формироваться либо из фрагментов документов коллекции, либо из сторонних текстов — это два альтернативных варианта методики. Каждый запрос должен быть достаточно кратким, чтобы асессор мог быстро понять его смысл, в то же время достаточно полным, чтобы между асессорами не возникало расхождений в его интерпретации. Примерный объем текста запроса — одна страница формата А4. Запрос может иметь заголовок. Текст запроса должен быть информативнее заголовка в том смысле, что обычные поисковые системы не должны давать удовлетворительно полного ответа по тексту заголовка. Вместе с коллекцией запросов асессорам может сообщаться модельная ситуация поиска. Например, в случае новостной коллекции запросом

может быть текст одного или нескольких новостных сообщений, а релевантными документами — тексты новостей, необходимые для восстановления цепочки связанных событий. В случае коллективного блога Хабрахабр запросом может быть несколько текстовых фрагментов, отобранных из внешних источников, а релевантными документами — все статьи Хабрахабра по соответствующей тематике.

Шаг 2. Асессорам раздаются запросы и инструкция, объясняющая поисковое задание и модельную ситуацию поиска. По каждому запросу асессор должен найти в коллекции как можно больше релевантных документов и предоставить список найденных документов. Он может пользоваться любыми доступными ему средствами поиска. Также замеряется время, потраченное асессором на обработку запроса. Число асессоров m, обработавших каждый запрос, является параметром методики. Чем больше m, тем объективнее будут оценки полноты поиска.

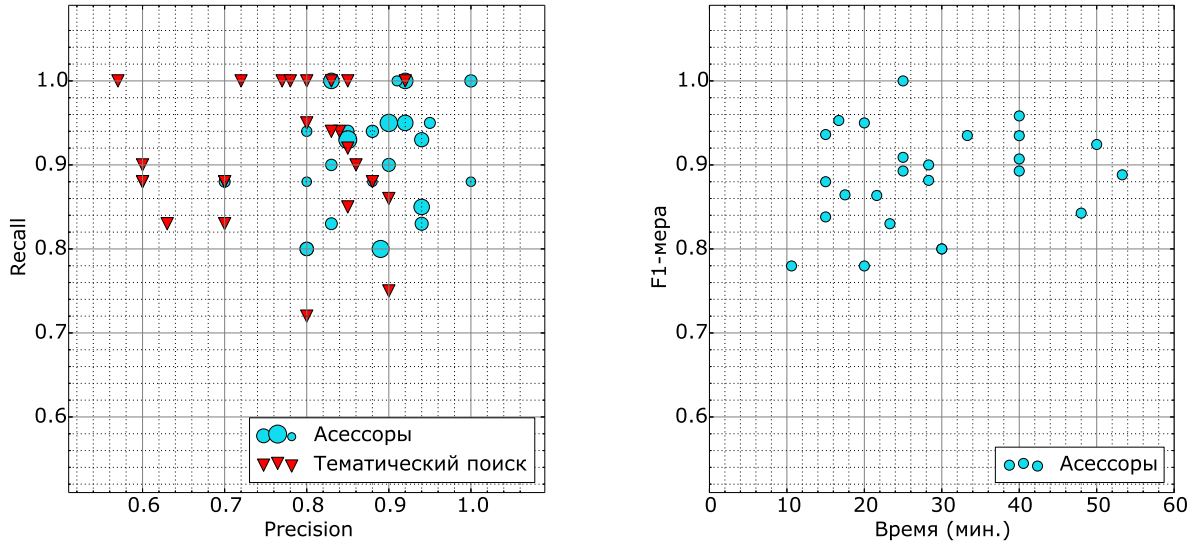
Шаг 3. На том же запросе асессору дается второе задание — разметить результаты тематического поиска. Для каждого пункта поисковой выдачи асессор ставит оценку релевантности в бинарной или порядковой шкале. Документ считается релевантным запросу, если хотя бы один асессор нашел этот документ или если этот документ был найден тематическим поиском и хотя бы n из m асессоров отметили его как релевантный. Число n также является параметром методики.

Для каждого запроса определим две меры качества поиска: *точность* Precision@k — доля релевантных документов среди первых k найденных; *полнота* Recall@k — доля k первых найденных релевантных документов среди всех релевантных. Для измерения качества тематического поиска точность и полнота усредняются по всем запросам. При измерении качества асессорского поиска точность и полнота усредняются еще и по асессорам. Агрегированная оценка качества поиска F_1-*мера* определяется как среднее гармоническое точности P и полноты R: $F_1 = (P + R)/(2PR)$.

Описанная методика была применена к коллекции Хабрахабра. Для эксперимента были составлены 25 запросов по тематике коллективного блога путем копирования текстовых фрагментов из различных внешних источников. Тексты запросов не превышали одной страницы формата A4. Примеры заголовков запросов представлены в табл. 1. Полный список использованных запросов и инструкцию для асессоров можно найти на странице русскоязычного вики-ресурса MachineLearning.ru «Оценивание качества разведочного поиска (эксперимент)». Каждый запрос обрабатывался $m = 3$ асессорами. Результат тематического поиска считался релевантным, если хотя бы $n = 2$ асессора отметили его как релевантный.

Таблица 1 Заголовки запросов для разведочного поиска

Алгоритмы раскраски графов	IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Self-driving Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure



(*а*) точность и полнота поиска

(*б*) время и F_1-мера (асессоры)

Рис. 2 Качество разведочного поиска по 25 запросам для асессоров и тематического поиска

Результаты эксперимента показаны на рис. 2. Точки на графиках соответствуют запросам. На графике 2, *а* сравниваются точность и полнота поиска, выполненного асессорами, и тематического разведочного поиска. Видно, что точность в среднем выше у асессорского поиска, а полнота — у тематического. Полнота тематического поиска принимала наивысшее значение 1,0 для 8 из 25 запросов. Размер точек пропорционален времени, в среднем затраченному асессорами на обработку данного запроса. График 2, *б* показывает, что нет прямой зависимости между временем, затраченным асессором, и качеством поиска. В среднем на обработку одного запроса асессоры тратили около 30 мин.

Таким образом, тематический поиск позволяет находить релевантные документы полнее и значительно быстрее, чем это делают асессоры, ценой незначительного ухудшения точности (появления нерелевантных документов в результатах поиска).

5 Выбор тематической модели по критерию качества поиска

Множества релевантных документов, найденные асессорами для каждого запроса, позволяют оценивать точность и полноту тематического поиска для новых тематических моделей. Появляется возможность сравнивать тематические модели по критериям качества поиска. Были проведены два таких эксперимента, их результаты сведены в табл. 2. В случаях, когда с помощью новых моделей алгоритм тематического поиска находил новые релевантные документы, расширяли множества релевантных документов и пересчитывали оценки точности и полноты для всех моделей.

В первом эксперименте сравнивались мультимодальные модели с различными сочетаниями модальностей (термины, авторы, комментаторы, теги, хабы) и с числом тем $|T| = 200$. Совместное использование всех модальностей значительно улучшает полноту и точность поиска. Основной вклад вносят модальности терминов и тегов. Модели без терминов, а также унимодальная модель, учитывающая только термины, показывают заметно худшие результаты.

Во втором эксперименте сравнивались модели с числом тем $|T| = 100$, 200, 300, 400 и 500. Оптимальное качество поиска достигается при 200 темах, дальнейшее увеличение числа тем не ведет к повышению точности и полноты. Таким образом, качество поиска

Таблица 2 Сравнение асессорского и тематического поиска по критериям Precision@k и Recall@k: для моделей с разными сочетаниями модальностей (Слова, Комментаторы, Теги, Хабы) при числе тем $|T| = 200$ и для моделей со всеми пятью модальностями и с разным числом тем $|T|$

Критерий	Асессоры	Модальности						Число тем				
		С	К	ТХ	СТ	СХ	СТХ	100	200	300	400	500
Precision@5	0,82	0,63	0,54	0,59	0,74	0,73	0,73	0,61	**0,74**	0,71	0,69	0,59
Precision@10	0,87	0,67	0,56	0,58	0,77	0,74	0,75	0,65	**0,77**	0,72	0,67	0,61
Precision@15	0,86	0,65	0,53	0,55	0,67	0,67	0,68	0,67	**0,68**	0,67	0,65	0,62
Precision@20	0,85	0,64	0,53	0,54	0,66	0,67	0,68	0,64	**0,68**	0,67	0,64	0,60
Recall@5	0,78	0,77	0,63	0,69	0,82	0,81	0,82	0,62	**0,82**	0,80	0,72	0,63
Recall@10	0,84	0,79	0,64	0,71	0,88	0,82	0,87	0,63	**0,88**	0,81	0,75	0,64
Recall@15	0,88	0,82	0,67	0,73	0,90	0,84	0,89	0,67	**0,90**	0,82	0,77	0,67
Recall@20	0,88	0,85	0,68	0,74	0,91	0,85	0,89	0,69	**0,91**	0,85	0,77	0,68

является эффективным внешним критерием для определения числа тем, в то время как внутренние критерии, такие как перплексия, не позволяют судить о числе тем в коллекции [30].

6 Заключение

Конечной целью разведочного информационного поиска является интенсификация и автоматизация процессов приобретения и систематизации знаний людьми. Тематическое моделирование рассматривается как одна из его ключевых технологий. В данной работе исследуется тематический поиск по длинным текстовым запросам на примере коллекции статей коллективного блога Хабрахабр.

Тематическая модель строится с помощью библиотеки с открытым кодом BigARTM, которая позволяет оптимизировать одновременно несколько критериев качества и находить сжатые векторные тематические представления статей и запросов. Для подбора коэффициентов регуляризации использована «жадная» стратегия последовательного добавления регуляризаторов в тематическую модель. Тематический поиск реализуется путем сравнения тематических векторов запроса и статей по косинусной мере близости.

Для оценивания качества поиска разработана специальная коллекция запросов — заданий разведочного поиска, которые сначала выполняются людьми (асессорами), затем системой тематического поиска, затем релевантность найденных ею документов снова оценивается асессорами. Данная методика позволяет, единожды сделав разметку результатов поиска, многократно вычислять оценки качества тематического поиска для различных тематических моделей и механизмов поиска.

На данных Хабрахабра показано, что тематический поиск находит релевантные документы полнее и намного быстрее, чем это делают асессоры. Подбор тематической модели по критериям точности и полноты поиска показал, что использование категоризации статей по тегам и хабам улучшает качество поиска существеннее, чем использование метаинформации об авторах статей и комментариев.

Литература

[1] *Marchionini G.* Exploratory search: From finding to understanding // Commun. ACM, 2006. Vol. 49. No. 4. P. 41–46.

[2] *White R. W., Roth R. A.* Exploratory search: Beyond the query-response paradigm. — Morgan and Claypool Publs., 2009. 98 p.

[3] *Kraaij W., Post W.* Task based evaluation of exploratory search systems // SIGIR Workshop on Evaluating Exploratory Search Systems Proceedings. — ACM, 2006. P. 24–27.

[4] *Potthast M., Hagen M., Völske M., Stein B.* Exploratory search missions for TREC topics // EuroHCIR, 2013. Vol. 1033. P. 7–10.

[5] *Shah C., Hendahewa C., Gonzalez-Ibanez R.* Rain or shine? Forecasting search process performance in exploratory search tasks // J. Assoc. Inform. Sci. Technol., 2016. Vol. 67. No. 7. P. 1607–1623.

[6] *Hofmann T.* Probabilistic latent semantic indexing // 22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings. — New York, NY, USA: ACM, 1999. P. 50–57.

[7] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // J. Machine Learn. Res., 2003. Vol. 3. P. 993–1022.

[8] *Blei D. M.* Probabilistic topic models // Commun. ACM, 2012. Vol. 55. No. 4. P. 77–84.

[9] *Manning C. D., Raghavan P., Schütze H.* Introduction to information retrieval. — New York, NY, USA: Cambridge University Press, 2008. 504 p.

[10] *Scherer M., von Landesberger T., Schreck T.* Topic modeling for search and exploration in multivariate research data repositories // Research and Advanced Technology for Digital Libraries: Conference (International) on Theory and Practice of Digital Libraries Proceedings / Eds. T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, C. J. Farrugia. — Berlin–Heidelberg: Springer, 2013. P. 370–373.

[11] *Grant C. E., George C. P., Kanjilal V., Nirkhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AIAA Press, 2015. P. 43–48.

[12] *Rönnqvist S.* Exploratory topic modeling with distributional semantics // 14th Symposium (International) on Advances in Intelligent Data Analysis Proceedings / Eds. E. Fromont, T. De Bie, M. van Leeuwen. — Saint Etienne, France: Springer International Publs., 2015. P. 241–252.

[13] *Veas E. E., di Sciascio C.* Interactive topic analysis with visual analytics and recommender systems // 2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, Joint Conference (International) on Artificial Intelligence. — Aachen, Germany: CEUR-WS.org, 2015.

[14] *Feldman S. E.* The answer machine // Synthesis Lectures Inform. Concepts Retrieval Services, 2012. Vol. 4. P. 1–137.

[15] *Rahman M.* Search engines going beyond keyword search: A survey // Int. J. Comput. Appl., 2013. Vol. 75. No. 17. P. 1–8.

[16] *Singh R., Hsu Y.-W., Moon N.* Multiple perspective interactive search: A paradigm for exploratory search and information retrieval on the Web // Multimedia Tools Appl., 2013. Vol. 62. No. 2. P. 507–543.

[17] *Jiang T.* Exploratory search: A critical analysis of the theoretical foundations, system features, and research trends // Library and information sciences: Trends and research / Eds. C. Chen, R. Larsen. — Berlin–Heidelberg: Springer, 2014. P. 79–103.

[18] *Marie N., Gandon F.* Survey of linked data based exploration systems // 3rd Workshop (International) on Intelligent Exploration of Semantic Data co-located with the 13th Semantic Web Conference (International) Proceedings. — Riva del Garda, Italy, 2014.

[19] *Jacksi K., Dimililer N., Zeebaree S. R. M.* A survey of exploratory search systems based on LOD resources // 5th Conference (International) on Computing and Informatics Proceedings. — Malaysia: School of Computing, University Utara, 2015. P. 501–509.

[20] *Mimno D., Hoffman M., Blei D.* Sparse stochastic inference for latent Dirichlet allocation // 29th Conference (International) on Machine Learning Proceedings / Eds. J. Langford J. Pineau. — New York, NY, USA: Omnipress, 2012. P. 1599–1606.

[21] *Bassiou N., Kotropoulos C.* Online PLSA: Batch updating techniques including out-of-vocabulary words // IEEE Trans. Neural Networks Learning Syst., 2014. Vol. 25. No. 11. P. 1953–1966.

[22] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-Bayesian additive regularization for multimodal topic modeling of large collections // Workshop on Topic Models: Post-Processing and Applications Proceedings. — New York, NY, USA: ACM, 2015. P. 29–37.

[23] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Adv. Inform. Retrieval, 2009. Vol. 5478. P. 29–41.

[24] *Andrzejewski D., Buttler D.* Latent topic feedback for information retrieval // 17th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2011. P. 600–608.

[25] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: A survey // Frontiers Comput. Sci. China, 2010. Vol. 4. No. 2. P. 280–301.

[26] *Vorontsov K. V.* Additive regularization for topic models of text collections // Dokl. Math., 2014. Vol. 89. No. 3. P. 301–304.

[27] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // 5th Conference (International) on Analysis of Images, Social Networks and Texts, 2016.

[28] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis images, social networks and texts / Eds. D. I. Ignatov, M. Yu. Khachay, M. Y. Panchenko, *et al.* — Communications in computer and information science ser. — Springer International Publs., 2014. Vol. 436. P. 29–46.

[29] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications, 2015. Vol. 101. No. 1. P. 303–323.

[30] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive regularization of topic models for topic selection and sparse factorization // 3rd Symposium (International) on Learning and Data Sciences Proceedings / Ed. A. Gammerman. — U.K.: University of London, 2015. P. 193–202.

[31] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // 15th Mexican Conference (International) on Artificial Intelligence Proceedings, 2016.

[32] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // 7th Symposium (International) Chinese Spoken Language Processing Proceedings, 2010. P. 224–228.

Поступила в редакцию 02.09.2016

Multimodal topic modeling for exploratory search in collective blog*

***A. O. Ianina*[1,2] *and K. V. Vorontsov*[1,2]**

yanina-n@yandex-team.ru; vokov@forecsys.ru

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

[2]Yandex LLC, 16 Leo Tolstoy Str., Moscow, Russia

Exploratory Search is a new paradigm in information retrieval focused on the acquisition and systematization of knowledge by professionals, unlike major Web search engines that answer short text queries of mass users. An exploratory search engine has been developed based on probabilistic topic modeling for seeking information thematically relevant to the long text queries. *Additive Regularization for Topic Modeling* (ARTM) was used to combine many requirements such as sparsity, diversity, and interpretability of topics and to incorporate heterogeneous modalities such as authors, tags, and categories into the model. The parallelized online implementation of ARTM was used in open source library *BigARTM* (`bigartm.org`). The thematic search is implemented by maximizing cosine similarity between query and document both represented by their sparse distributions over topics. The authors evaluate precision and recall of the thematic search by a two-step procedure. First, human assessors perform exploratory search tasks manually using any available search utilities (it takes them about 30 min per task in average). Second, they evaluate the relevance of search results found by the present thematic search engine for the same tasks. The experiments on the collection of 132 000 articles from `habrahabr.ru` collective blog showed that thematic search provided comparable precision and better recall, also reducing search time from half an hour to seconds. With data labeled by assessors, the optimal number of topics was determined and it was shown that the joint use of all modalities (authors of articles, authors of comments, tags, and hub categories) significantly improves the search quality.

Keywords: *information retrieval; exploratory search; topic modeling; additive regularization for topic modeling; BigARTM*

DOI: 10.21469/22233792.2.2.04

References

[1] Marchionini, G. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49(4):41–46.

[2] White, R. W., and R. A. Roth. 2009. *Exploratory search: Beyond the query-response paradigm*. Morgan and Claypool Publs. 98 p.

[3] Kraaij, W., and W. Post. 2006. Task based evaluation of exploratory search systems. *SIGIR Workshop on Evaluating Exploratory Search Systems Proceedings*. ACM. 24–27.

[4] Potthast, M., M. Hagen, M. Völske, and B. Stein. 2013. Exploratory search missions for TREC topics. *EuroHCIR* 1033:7–10.

[5] Shah, C., C. Hendahewa, and R. Gonzalez-Ibanez. 2016. Rain or shine? Forecasting search process performance in exploratory search tasks. *J. Assoc. Inform. Sci. Technol.* 67(7):1607–1623.

*The research was supported by the Russian Foundation for Basic Research, grants 16-37-00498, 14-07-00847, and 14-07-00908.

[6] Hofmann, T. 1999. Probabilistic latent semantic indexing. *22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings*. New York, NY: ACM. 50–57.

[7] Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

[8] Blei, D. M. 2012. Probabilistic topic models. *Commun. ACM* 55(4):77–84.

[9] Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. New York, NY: Cambridge University Press. 504 p.

[10] Scherer, M., T. von Landesberger, and T. Schreck. 2013. Topic modeling for search and exploration in multivariate research data repositories. *Research and Advanced Technology for Digital Libraries: Conference (International) on Theory and Practice of Digital Libraries Proceedings*. Eds. T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia. Berlin–Heidelberg: Springer. 370–373.

[11] Grant, C. E., C. P. George, V. Kanjilal, S. Nirkhiwale, J. N. Wilson, and D. Z. Wang. 2015. A topic-based search, visualization, and exploration system. *FLAIRS Conference*. AAAI Press. 43–48.

[12] Rönnqvist, S. 2015. Exploratory topic modeling with distributional semantics. *14th International Symposium on Advances in Intelligent Data Analysis Proceedings*. Eds. E. Fromont, T. De Bie, and M. van Leeuwen. Saint Etienne, France: Springer International Publs. 241–252.

[13] Veas, E. E., and C. di Sciascio. 2015. Interactive topic analysis with visual analytics and recommender systems. *2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, Joint Conference (International) on Artificial Intelligence*. Aachen, Germany: CEUR-WS.org.

[14] Feldman, S. E. 2012. The answer machine. *Synthesis Lectures Inform. Concepts Retrieval Services* 4(3):1–137.

[15] Rahman, M. 2013. Search engines going beyond keyword search: A survey. *Int. J. Comput. Appl.* 75(17):1–8.

[16] Singh, R., Y.-W. Hsu, and N. Moon. 2013. Multiple perspective interactive search: A paradigm for exploratory search and information retrieval on the Web. *Multimedia Tools Appl.* 62(2):507–543.

[17] Jiang, T. 2014. Exploratory search: A critical analysis of the theoretical foundations, system features, and research trends. *Library and information sciences: Trends and research*. Eds. C. Chen and R. Larsen. Berlin–Heidelberg: Springer. 79–103.

[18] Marie, N., and F. Gandon. 2014. Survey of linked data based exploration systems. *3rd Workshop (International) on Intelligent Exploration of Semantic Data co-located with the 13th Semantic Web Conference (International) Proceedings*. Riva del Garda, Italy.

[19] Jacksi, K., N. Dimililer, and S. R. M. Zeebaree. 2015. A survey of exploratory search systems based on LOD resources. *5th Conference (International) on Computing and Informatics Proceedings*. Malaysia: School of Computing, Universiti Utara. 501–509.

[20] Mimno, D., M. Hoffman, and D. Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. *29th Conference (International) on Machine Learning Proceedings*. Eds. J. Langford and J. Pineau. New York, NY: Omnipress. 1599–1606.

[21] Bassiou, N., and C. Kotropoulos. 2014. Online PLSA: Batch updating techniques including out-of-vocabulary words. *IEEE Trans. Neural Networks Learning Systems* 25(11):1953–1966.

[22] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. 2015. Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Workshop on Topic Models: Post-Processing and Applications Proceedings*. New York, NY: ACM. 29–37.

[23] Yi, X., and J. Allan. 2009. A comparative study of utilizing topic models for information retrieval. *Advances in information retrieval*. Berlin–Heidelberg: Springer. 5478:29–41.

[24] Andrzejewski, D., and D. Buttler. 2011. Latent topic feedback for information retrieval. *17th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. 600–608.

[25] Daud, A., J. Li, L. Zhou, and F. Muhammad. 2010. Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers Comput. Sci. China* 4(2):280–301.

[26] Vorontsov, K. V. 2014. Additive regularization for topic models of text collections. *Dokl. Math.* 89(3):301–304.

[27] Frei, O., and M. Apishev. 2016. Parallel non-blocking deterministic algorithm for online topic modeling. *5th Conference (International) on Analysis of Images, Social Networks and Texts*.

[28] Vorontsov, K. V., and A. A. Potapenko. 2014. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *Analysis images, social networks and texts*. Eds. D. I. Ignatov, M. Yu. Khachay, M. Y. Panchenko, *et al.* Communications in computer and information science ser. Springer International Publs. 436:29–46.

[29] Vorontsov, K. V., and A. A. Potapenko. 2015. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications* 101(1):303–323.

[30] Vorontsov, K. V., A. A. Potapenko, and A. V. Plavin. 2015. Additive regularization of topic models for topic selection and sparse factorization. Ed. A. Gammerman. *3rd Symposium (International) on Learning and Data Sciences Proceedings*. University of London, U.K. Switzerland: Springer International Publs. 193–202.

[31] Apishev, M., S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov. 2016. Additive regularization for topic modeling in sociological studies of user-generated text content. *15th Mexican Conference (International) on Artificial Intelligence Proceedings*.

[32] Tan, Y., and Z. Ou. 2010. Topic-weak-correlated latent Dirichlet allocation. *7th Symposium (International) Chinese Spoken Language Processing Proceedings*. 224–228.

Received September 2, 2016