Нелинейное ранжирование результатов разведочного информационного поиска.*

Мамонов К. Р.[1]***, Воронцов К. В.***[1]***, Еремеев М. А.***[1]
mamonov.kr@phystech.ru, vokov@forecsys.ru, maks5507@yandex.ru
[1] Московский физико-технический институт, Москва, Россия

Имея коллекцию документов, пользователю порой очень сложно в них разобраться. Существует множество подходов для поиска среди этих документов, но их недостаточно, когда пользователь хочет получить доступ к соответствующим документам в некотором логическом порядке, например, для учебных целей. В данной работе описан алгоритм ранжирования документов от простого к сложному, от общего к частному, то есть в том порядке, в котором пользователю будет легче разбираться в новой для него тематической области. Данный подход даёт пользователю абсолютно новый способ потребления контента.

Ключевые слова: *граф чтения, тематическое моделирование, информационный поиск.*

1 Введение

В связи с последними научными достижениями, особенно развитием интернета, многократно возрос объем текстовой информации, которую приходится обрабатывать человеку. Все более актуальной становится проблема ранжирования информации для более понятной и быстрой её обратоки и понимания.

Предлагается алгоритм построения графа чтения по статьям Википедии [5] с применением методов тематичесского моделирования.

Тематические модели широко используются на практике для решения задач ранжирования документов. Одним из продвинутых инструментов в тематическом моделировании является библиотека BigARTM [4]. Она предоставляет широкий выбор для настройки модели, используя обширный класс регуляризаторов.

2 Постановка задачи

Текущий подход к построению подобных графов описан в работе [3]. Сначала проводится предобработка текстов — удаляются стоп-слова и другой шум, все слова приводятся к начальной форме и строится матрица документ-слово. Затем данная матрица переводится LDA алгоритмом [1] в матрицу документ-тема и происходит колибровка тематической модели. после этого из полученной матрицы строится частичный порядок документов коллекции и дерево чтения в целом.

Узким местом являются тематические модели. Мы предлагаем применить для них алгоритм BigARTM [4], вместо LDA [1], построить мультимодальную тематическую модель [2] и использовать в качестве меры общности не энтропию, а долю терминов с узким распределением $p(w|t)$.

Был взят набор текстовых коллекций Википедии $\mathfrak{D} = \{D_i\}_{i=1}^{N}$, где каждая коллекция D состоит из документов d, словари коллекциий текстов W, состоящих из термов w, и множества тем T, состоящих из тем t.

*Задачу поставил: Воронцов К. В. Консультант: Еремеев М. А.

Распределение вида $p(t|x)$ будем называть тематикой объекта x. Можно говорить о тематике документа $p(t|d)$, терма $p(t|w)$, терма в документе $p(t|d,w)$.

Целью нелинейного ранжирования является построение частичного порядка на коллекции документов D, в частности, это может быть совокупность деревьев (лес документов). Для этого строится тематичесская модель и для каждого документа находится вероятность того, что документ i принадлежит теме j, так получается матрица $\theta_{ij} = p(t|d)$. По ней можно для каждого документа посчитать меру общности документа $g(d_i) = \sum_m -$ $-F_{im}log(F_{im})$ и меру пересеченния двух документов $o(d_i, d_j) = \frac{F_i \cdot F_j}{|F_i^2|+|F_j^2|-F_i \cdot F_j}$, которые отражают сложность документов и позволяют построить матрицу смежности A требуемого графа.

Критерием качества S нашего алгоритма будем считать величину обратную к среднеквадратичному отклонению между двумя матрицами смежности графов чтения. Эталонные графы с матрицей смежности $\hat{A}$ построим из категорий Википедии. Цель данной работы состоит в решении следующей задачи:

$$\sum_{i,j=1}^{n}(A_{ij} - \hat{A_{ij}})^2 \to \min_{\mathfrak{D}} \tag{1}$$

Литература

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] A.O. Ianina and K.V. Vorontsov. Multimodal topic modeling for exploratory search in collective blog. *Machine Learning and Data Analysis*, 2(2):173–186, 2016.

[3] Georgia Koutrika, Lei Liu, and Steven J. Simske. Generating reading orders over document collections. In Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman, editors, *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 507–518. IEEE Computer Society, 2015.

[4] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[5] WikiPedia, a web-based, free-content encyclopedia. http://www.wikipedia.org.