Сравнение тематического и простого поиска для находения тематически близких статей

Отчет по заданию спецкурса "Вероятностные тематические модели"

Выполнила: Кибитова Валерия, 517 группа

Задача поиска тематически близких статей состоит в нахождении статей, которые тематически близки к заданной статье. Для решения данной задачи можно применять как простой поиск так и тематическое моделирование.

Рассмотрим решение данной задачи при помощи простого поиска. Для этого введем определение TF-IDF меры, пусть d_j – документ из корпуса D, w_i – слово, n_i – число слов w_i в документе d, n – общее число слов в документе d, D – число документов в корпусе, тогда

$$\text{TF-IDF}(t, d_j) = \frac{n_i}{n} \log \frac{|D|}{|d : t_i \in d, d \in D|}.$$

Для того, чтобы найти тематически близкие документы к заданному, необходимо посчитать вектор TF-IDF всех слов (после лемматизации данных слов) для каждого документа, и затем найти документы, TF-IDF векторы которых находятся ближе всего к TF-IDF вектору заданного документа по некоторой метрике.

Рассмотрим решение данной задачи при помощи тематического моделирования [1]. Тематическая модель — это представление наблюдаемого условного распределения $p(w|d)$ терминов (слов или словосочетаний) w в документах d коллекции D:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где T — множество тем; $\phi_{wt} = p(w|t)$ — неизвестное распределение терминов в теме t; $\theta_{td} = p(t|d)$ — неизвестное распределение тем в документе d. При наличии такой модели, находить близкие документы можно сравнивая векторы распределений $p(t|d)$, используя некоторые метрики.

Для сравнения двух методов, описанных выше была использована коллекция данных ПостНауки, состоящая из 3446 документов на различные темы. Для того, чтобы сравнить векторы документов x и y использовались следующие метрики

[1] http://www.machinelearning.ru/wiki/index.php?title=BigARTM

близости: косинусная мера –

$$\cos(x, y) = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sqrt{\sum\limits_{i=1}^{n} (x_i)^2}\sqrt{\sum\limits_{i=1}^{n} (y_i)^2}},$$

расстояние хелингера –

$$H(x, y) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{x_i} - \sqrt{y_i})},$$

расстояние Дженсона-Шеннона –

$$\mathrm{JSD}(x \parallel y) = \frac{1}{2}D(x \parallel m) + \frac{1}{2}D(y \parallel m),$$

$$\text{где } m = \frac{1}{2}(x + y), D(x\|y) = \sum_i x_i \log \frac{x_i}{y_i}.$$

Так как при сравнении распределений существует так называемая проблема "проклятия размерности то можно исследовать зависимость качества поиска от обнуления хвостов распределений, то есть: $p_i = 0, \text{если } p_i < \frac{1}{|T|}$, где |T| – число тем.

Документы будем считать тематически близкими, если в результате ручного просмотра человек может проинтерпретировать их как документы, которые имеют некоторое множество общих тем (основная тематика документов является схожей). Под качеством тематического поиска будем понимать следующее: пусть n – это число документов, для которых ищутся ближайшие, k – это число документов, которые просматриваются при выдаче результатов поиска, k_i – число документов тематически близких к заданному документу i, тогда качество будет определяться как

$$\frac{\sum_{i=1}^{n} k_i}{nk}.$$

Для построения тематических моделей использовалась библиотека BigARTM. Так как решение задачи тематического моделирования неединственно и неустойчиво, то для решения этой проблемы принято использовать регуляризацию — накладывать дополнительные ограничения на искомое решение. Подход APTM основан на идее многокритериальной регуляризации. Он позволяет строить модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора — оптимизационного критерия, зависящего от параметров модели. Взвешенная сума всех таких критериев максимизируется совместно с основным критерием правдоподобия. В настоящий момент в BigARTM реализованы следующие регуляризаторы:

- artm.SmoothSparsePhiRegulazer – предназначен для сглаживания и разреживания матрицы Ф

- artm.SmoothSparseThetaRegularizer – предназначен для сглаживания и разреживания матрицы Θ
- artm.DecorrelatorPhiRegularizer – декоррелятор Φ
- artm.SpecifiedSparsePhiRegularizer – разреживание матрицы Φ с заданной величиной
- artm.ImproveCoherencePhiRegularizer – повышение когерентности
- artm.SmoothPtdwRegularizer – сглаживание распределений p_{tdw}
- artm.TopicSelectionThetaRegularizer – разреживание распределения $p(t)$ для отбора тем

Для того, чтобы улучшить критерии качества тематической модели в настоящий момент подобраны следующие траектории регуляризации, а именно:

- разреживание необходимо включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- сокращение числа тем включать постепенно, при этом не совмещая с декорреляцией на одной итерации

Так как оптимального числа тем, для задачи тематического моделирования не существует, то можно исследовать зависимость качества тематического поиска от числа тем, при этом набор регуляризаторов и их параметры подбирались по вышеуказанной стратегии индивидуально для каждого числа тем таким образом, чтобы получившиеся темы были интерпретируемыми. В данном случае для исследования качества $k = 5$, $n = 150$, при этом эти 150 документов были выбраны случайно из исходной коллекции, и в последствии были использованы для всех экспериментов. На рис.1 приведены графики зависимости качества от числа тем, при использовании различных метрик. Суффикс _z стоящий рядом с метрикой обозначает, что хвосты исходного распределения были обнулены.

Из графика представленного на рис. 1 следует, что существует некоторая зависимость между числом тем, качеством поиска при использовании любой из рассмотренных метрик. Так, в данном случае наивысшее качество достигается при 115 темах, после чего оно начинает довольно резко падать. Возможно, это можно обьяснить тем, что при сильном увеличении числа тем, начинают появляться много мелких неважных тем, которые мешают оценить тематическую схожесть двух документов. Кроме того, из графика видно, что косинусная мера близости является наихудшим выбором для оценивания сходства двух документов при использовании тематической модели. В данном случае максимальное качество удалось достичь при использовании расстояния хелингера без обнуления хвостов

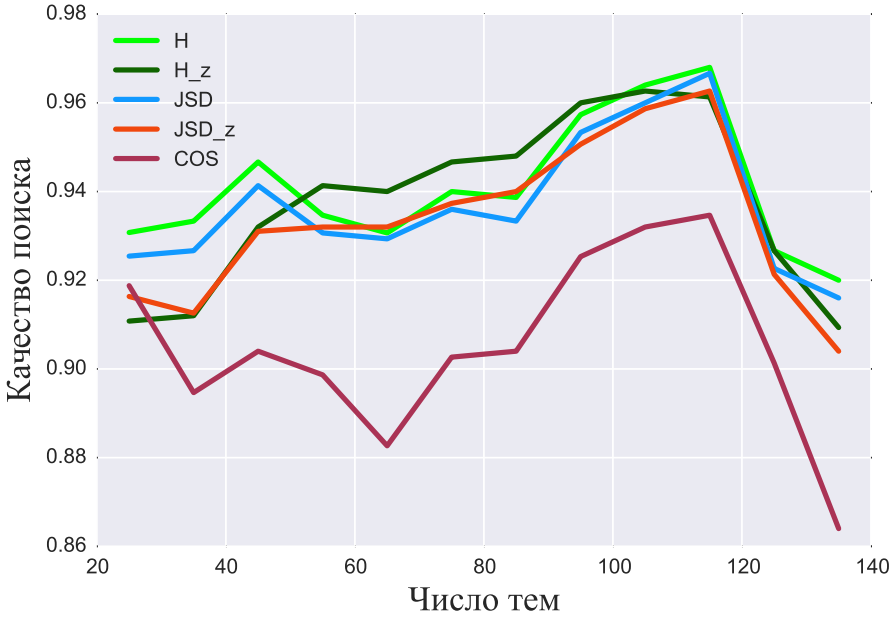

Рис. 1: Зависимость качества поиска от числа тем заданных при тематическом моделировании

распределений, и практически аналогичный результат был получен при использовании расстояния Дженсона-Шеннона.

Для сравнения качества поиска при использовании тематического моделирования по сравнению с простым поиском возьмем модель качество которой было наилучшим. Для сравнения двух подходов рассмотрим качество поиска при различном числе просматриваемых документов, которое обозначено k. На рис. 2 представлены графики зависимости качества поиска при использовании тематической модели от k для различных метрик. На рис. 3 представлены графики зависимости качества простого поиска от k для различных метрик.

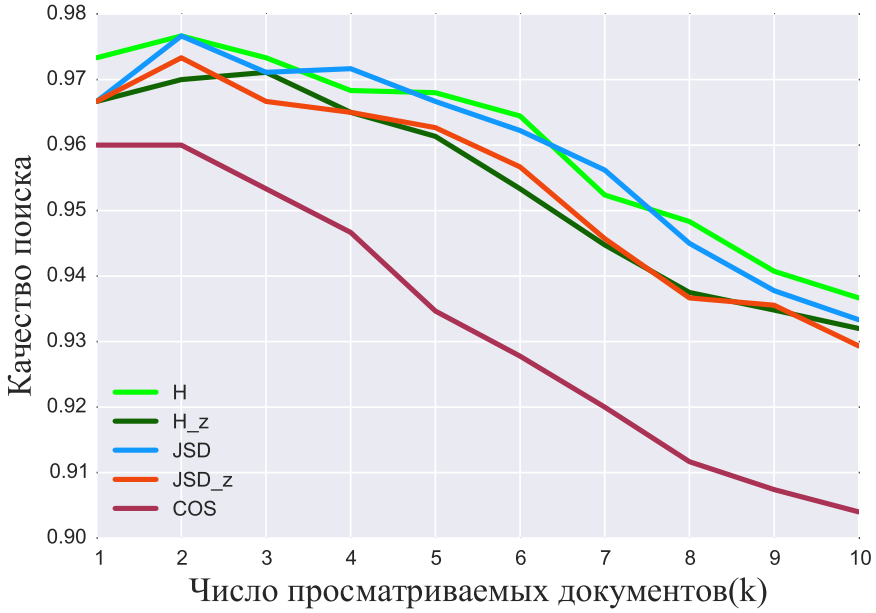

Рис. 2: Зависимость качества поиска от k при тематическом моделировании

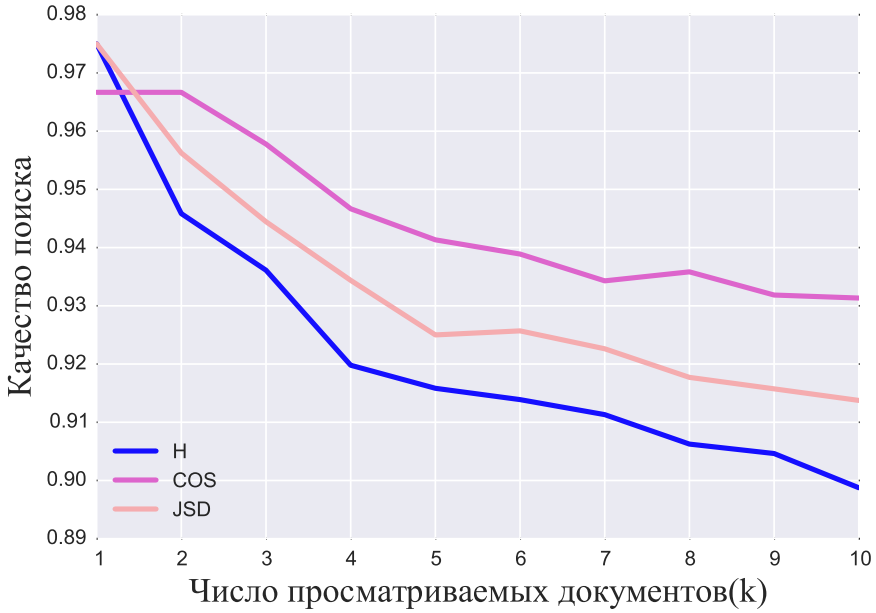

Рис. 3: Зависимость качества поиска от k при простом поиске

Примеры запросов и ответов при простом и тематическом поиске ($k = 5$, с целью однозначной идентификации текста и сокращения текста приведены имена

файлов из коллекции postnauka_clean, рядом с ответом стоит метка, 1 – релевантных файл по отношению к запросу, 0 – нерелевантный) :

Запрос: Одежды на древних статуэтках Кавказа

Ответы простого поиска:

1 Святилища Кавказа

1 Открытие древних статуэток Кавказа

1 FAQ: Открытие древних статуэток Кавказа

1 Автограф | "Древняя бронзовая антропоморфная пластика Кавказа"

1 5 книг об археологии Кавказа

Ответы тематического поиска:

1 Открытие древних статуэток Кавказа

1 Святилища Кавказа

1 FAQ: Святилища Кавказа

1 FAQ: Открытие древних статуэток Кавказа

1 Главы | Антропоморфная пластика, происходящая из святилищ Кавказа

Запрос: Что читать: "Происхождение человека. Эволюция"Элис Робертс

Ответы простого поиска:

1 Антропогенез

1 5 книг об антропогенезе

1 Становление специфики приматов

1 Курс "Происхождение человека: что мы знаем о наших предках"

1 Перспективы: Антропогенез как область знания

Ответы тематического поиска:

1 5 книг об антропогенезе

1 5 книг об эволюционной биологии

1 5 книг о теории эволюции

1 По шагам | Физическая антропология

1 5 книг по расоведению

Запрос: Культурный капитал

Ответы простого поиска:

1 Культурный капитал (другой ответ)

0 Роль хищника в эволюционном процессе

0 «Разномыслие, как ни странно, началось при Сталине. . . »

0 Главы | Европейский интеллектуал: попытка апологии субъективности

1 FAQ: Изобретение культурной политики

Ответы тематического поиска:

1 Главы | Природа эмоционального интеллекта. Когда умный глупеет

1 Школьное образование и умственные способности

1 Главы | Эволюция творческого мышления

1 Тест подсознательных ассоциаций

1 Культурный капитал

Запрос: Прикладная геоурбанистика

Ответы простого поиска:

1 Город настоящего будущего
0 Как написать диссертацию
0 Фрейм-анализ публичной лекции
0 Опросы общественного мнения — между наукой и политикой
1 Формирование понятия «урбанистика»
Ответы тематического поиска:
1 Город настоящего будущего
1 Креативный класс и креативный город
1 Грамматика городского пространства
1 Москва: город, агломерация, жители, система управления
1 Городская среда В.Л.Глазычева
Запрос:ПостБлог#2: рождение ПостНауки, рекуррентные платежи и легендарная салфетка
Ответы простого поиска:
1 ПостБлог#1: Академия, серия книг и реформация страницы Donate
1 Конференция: Главный редактор ПостНауки Ивар Максутов
0 Жизнь после Хиггса
0 Перспективы: Общая микробиология в XXI веке
1 Поддержи ПостНауку!
Ответы тематического поиска:
1 Поддержи ПостНауку!
1 РВК и ПостНаука: Навстречу инновациям
1 Все, что вы хотели знать о будущем
1 Добро пожаловать!
1 ScienceHub

Как видно из графиков изображенных на рис. 2 и 3, качество поиска падает с числом k, что является вполне обоснованным так как не для каждого запроса может присутствовать необходимое число релевантных документов. Для простого поиска в зависимости от k метрикой, используя которую можно достичь наивысшего качества поиска, является либо расстояние Дженсона-Шеннона, либо косинусное расстояние. Для поиска, который использует тематическую модель, в зависимости от k метрикой, используя которую можно достичь наивысшего качества поиска, является либо расстояние Дженсона-Шеннона без обнуления концов распределения, либо расстояние Хелингера без обнуления концов распределения.

Для наглядности на рисунке 4 представлены графики, которые показывают различия в качестве простого и тематического поиска при различных k. Префиксом TF обозначены график для простого поиска, префиксом T для поиска, который использует тематическую модель. На данном рисунке представлены только результаты использования только тех метрик, которые показали наилучшее качество при различных k. Из данного рисунка видно, что качество тематического поиска почти при всех k, лучше чем простого, или совсем незначительно хуже (при $k = 1$), чем простого. Это можно обьяснить тем, что простой поиск лучше

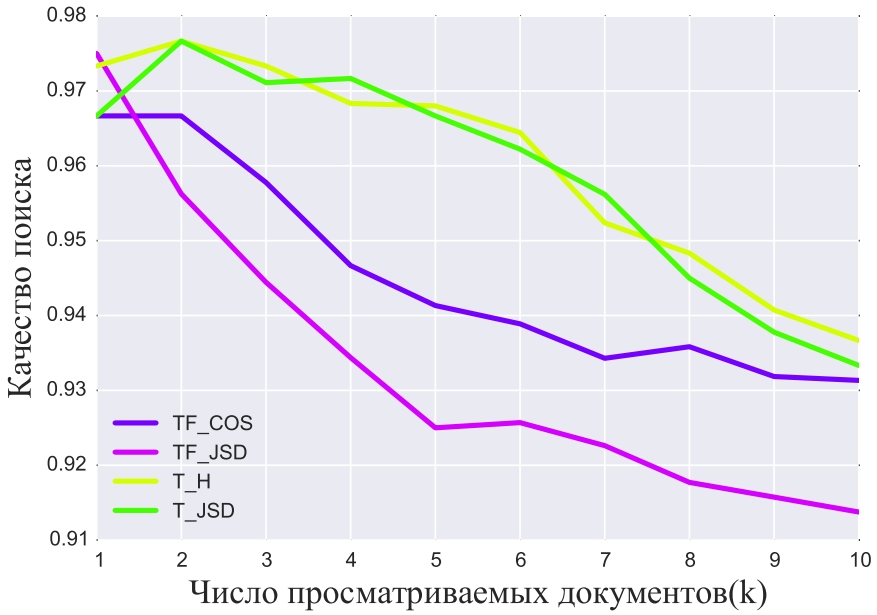

Рис. 4: Сравнение качества тематического и простого поиска при различных k

находит самый ближайший документ, лексика которого соответствует лексике документа запроса, но хуже находит больше документов. Так как не всякий документ имеет похожий тематически похожий документ, лексика которого совпадает с заданным. Тогда как поиск, использующий тематическую модель, позволяет находить документы, лексика которых несколько отличается от заданного, при этом документы будут иметь общие темы. Кроме того, следует отметить что вычисления расстояний при тематическом поиске происходят существенно быстрее, так как размерность векторов сокращается с десятков тысяч (при TF-IDF) до числа нескольких сотен (в данном случае до 125).