РТУ МИРЭА

Модели информационного поиска в массиве документов

ИКБО-04-20 Хан А.А.

Что такое информационный поиск

ИНФОРМАЦИОННЫЙ ПОИСК – НАПРАВЛЕНИЕ ИССЛЕДОВАНИЙ, ИЗУЧАЮЩЕЕ ВОПРОСЫ ПОИСКА ДОКУМЕНТОВ, ОБРАБОТКИ РЕЗУЛЬТАТОВ ПОИСКА, А ТАКЖЕ ЦЕЛЫЙ РЯД СМЕЖНЫХ ВОПРОСОВ: МОДЕЛИРОВАНИЯ, КЛАССИФИКАЦИИ, КЛАСТЕРИЗАЦИИ, ФИЛЬТРАЦИИ ДОКУМЕНТОВ, ПРОЕКТИРОВАНИЯ АРХИТЕКТУР ПОИСКОВЫХ СИСТЕМ И ПОЛЬЗОВАТЕЛЬСКИХ ИНТЕРФЕЙСОВ, ЯЗЫКИ ЗАПРОСОВ И Т.Д.

Документ – это содержательно законченная единица информации, представленная на каком-либо естественном языке, которая идентифицируется уникальным образом.
Документ – это порция информации, которой оперируют информационно-поисковые системы.
Информационно-поисковая система – это комплекс программных средств, обеспечивающих избирательный отбор по заданным признакам документов, хранимых в электронном представлении.

Аспекты моделей информационного поиска

Модель информационного поиска имеет три ключевых аспекта.

1. Формат представления документа.

Под документом мы будем понимать некий объект, содержащий информацию в зафиксированном виде. Документы могут содержать тексты на естественном или формализованном языке, изображения, звуковую информацию и т.д.

2. Формат представления запроса.

Под запросом мы понимаем формализованный способ выражения информационных потребностей пользователя системы. Для этого используется язык поисковых запросов, синтаксис которых варьируется от системы к системе.

3. Функция соответствия документа запросу.

Степень соответствия запроса и найденного документа (релевантность) — субъективное понятие, поскольку результаты поиска, уместные для одного пользователя, могут быть неуместными для другого.

Модели информационного поиска

КАЖДУЮ РАССМОТРИМ ПО ОТДЕЛЬНОСТИ

Булевская модель

ГДЕ ИСПОЛЬЗУЕТСЯ

Такая модель иногда используется во внутренних корпоративных системах поиска, базах данных.

НЕДОСТАТКИ

Основным недостатком булевской модели является крайняя жесткость и непригодность для ранжирования.

Как работает модель

Предположим, у нас есть некоторый словарь T = {t1, . . . , tn}, где ti – термы.

Термами могут быть слова, какие-то бессмысленные комбинации цифр, букв (почтовые индексы, телефонные номера и т.д.). Некоторые группы слов также считаются одним термом. Термы – не то же самое, что и слова.

Например, Яндекс все падежи одного существительного может считать одним термом.

Документ – это некоторое подмножество словаря, набор термов: D ⊂ T, иначе говоря D ∈ {0, 1}n: на k-й позиции вектора стоит единица в том случае, когда k-е слово из словаря принадлежит документу, и ноль, если слово не принадлежит ему.

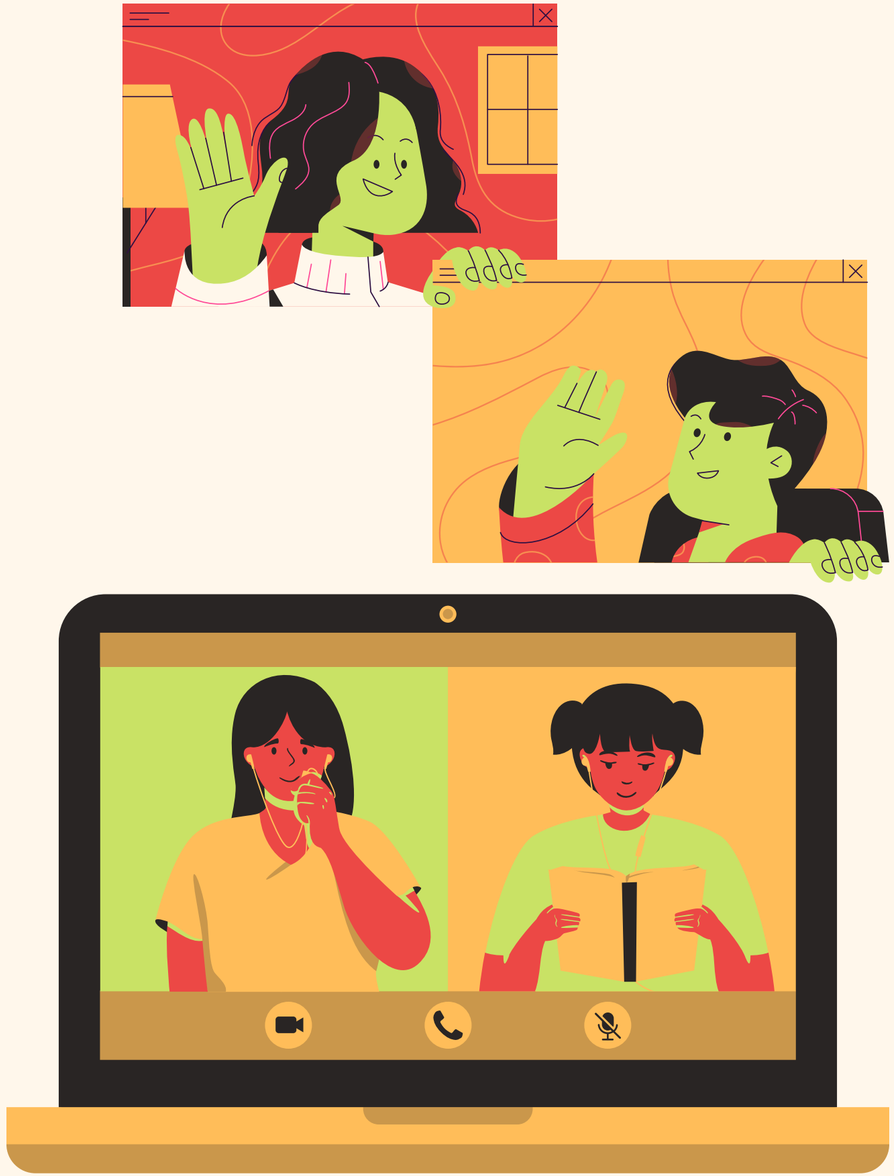
Запрос – булевская формула, например «t5 OR t7 AND NOT t12», что означает, что необходимо найти документы, которые включают пятый или седьмой термы, но не включают двенадцатый.

Если формула выполнена на некотором документе, то будем считать, что этот документ соответствует запросу.

Еще один момент!

Если слово, указанное в запросе, присутствует в документе, то он считается найденным, в противном случае – не найденным.

Не будет найден и документ, в котором встречаются только синонимы слова, указанного в запросе, в случае, когда само слово в документе не встречается



Векторная модель

У НАС ИМЕЕТСЯ СЛОВАРЬ ИЗ ТЕРМОВ, КАК В БУЛЕВСКОЙ МОДЕЛИ

Каждый документ представляется мультимножеством слов. Мультимножество – неупорядоченная коллекция, аналогичная множеству, но допускающая наличие в коллекции одновременно двух и более одинаковых значений. Каждый терм – это координата векторного пространства, говорящая о том, насколько «сильно» он входит в документ. Таким образом, каждый документ – это набор из n чисел.

Определим матрицу M по формуле

$$MIJ = TFIJ \cdot IDFI$$

Где TFij (Term Frequency, частота терма) – относительная доля слова i в документе j; IDFi (Inversed Document Frequency) – величина, обратная количеству документов, содержащих слово i. Другими словами, это количество всех документов, поделенное на количество документов, которые содержат слово i.

Разберемся, в чем состоит «физический смысл» Mij . Первый сомножитель показывает, насколько данное слово подходит данному документу.

Для этого разберем пример.

Давайте разберем пример

РАССМОТРИМ СЛОВО «ПЬЕР» И ПРОИЗВЕДЕНИЕ Л. Н. ТОЛСТОГО «ВОЙНА И МИР».

Слово «Пьер» окажется достаточно часто встречающимся словом, и первый сомножитель (доля слова «Пьер» среди всех слов романа) будет велик. Теперь посмотрим на второй сомножитель. Его величина зависит от того, является ли слово общеупотребительным или редким: чем более редким окажется слово, тем больше будет сомножитель. За счет этого слово «Пьер» для романа будет более значимо, чем, например, слово «дворянин», даже если они встречались в тексте одинаковое число раз. Таким образом, в двух словах можно сказать так: Mij – степень соответствия слова i документу j. Каждый документ представляется в этой матрице в виде столбца (j фиксировано, i меняется).

Но это еще не все!

Для того чтобы подсчитать меру релевантности, представим сначала запрос в виде вектора с координатами 0 или 1: Q = «t3 AND t5»= {0, 0, 1, 0, 1, 0, . . . , 0}.

Каждый документ – набор таких координат: много нулевых координат (это те термы, которые не встречаются) и несколько ненулевых координат.

Мерой релевантности R(Q, Dj) будем считать косинус угла между вектором запроса Q и документом Dj .

Нормализация необходима для того, чтобы уравнять веса документов с разным количеством слов.

Вероятностная модель

В 1977 году Робертсон (Robertson) и Спарк-Джоунз (Sparck-Jones) обосновали и реализовали вероятностную модель.

Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов, выбранных пользователем или полученных автоматически при каком-нибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости термов в релевантном наборе и в остальной части коллекции.

Документом будем считать множество слов без учета частоты встречаемости слова в документе. Можно также представить множество в виде обычного булевского вектора D = {d1, . . . , dn}, где n – количество всех термов, а di может принимать значения из множества {0, 1}. Запросом будем считать множество слов.

Соответствие документа запросу будем строить следующим образом: представим себе, что для каждого фиксированного запроса Qk у нас имеются распределения вероятностей на всех документах «быть релевантным» и «быть нерелевантным» запросу Qk. Обозначается это соответственно как P(R|Qk, D) и P(R|Qk, D).

Если обобщить

Наивысшая общая эффективности поиска достигается в случае, когда результирующие документы ранжируются по убыванию вероятности их релевантности запросу. Сначала для каждого для каждого документа оценивается вероятность того, что он релевантен запросу, а затем по этим оценкам выполняется ранжирование документов.

Существуют различные способы получения этих оценок, а также дополнительные предположения и гипотезы на основе априорных сведений относительно документов коллекции, которые и определяют конкретную реализацию вероятностной модели поиска.

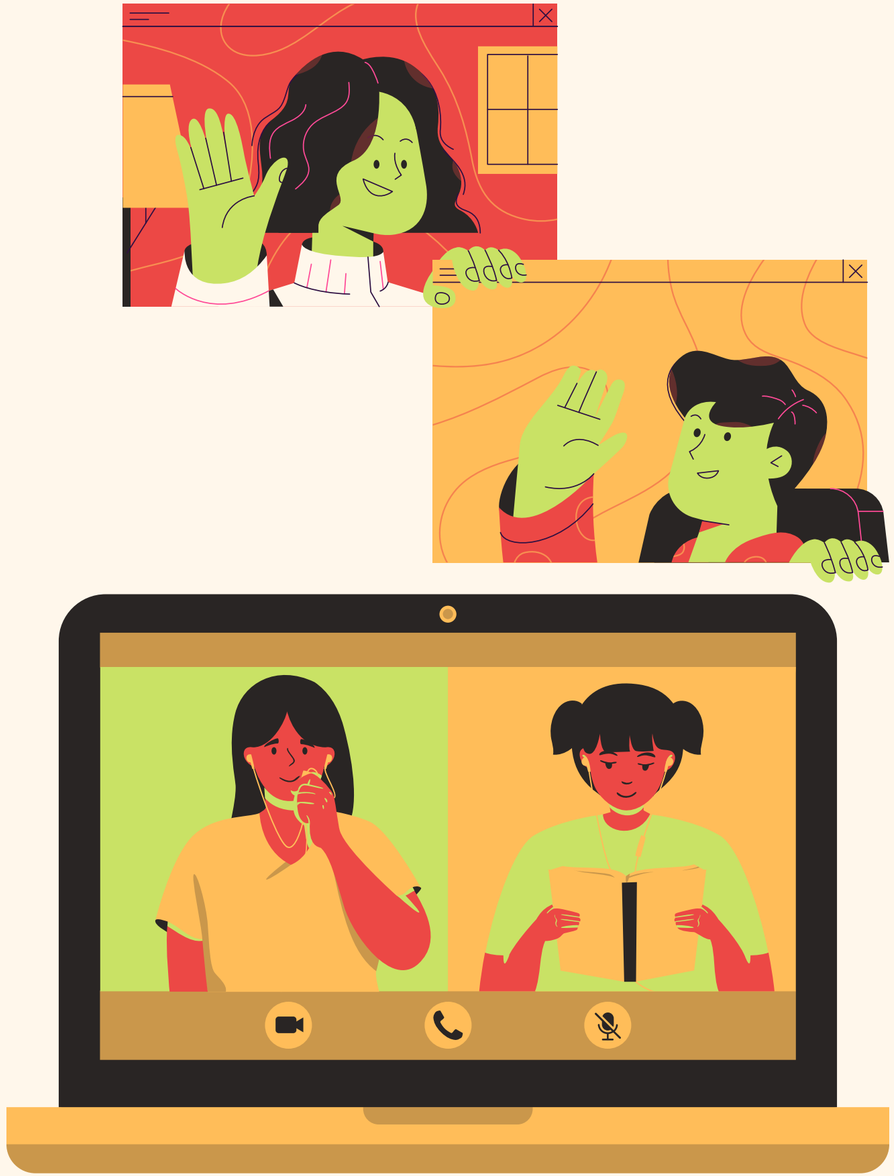
С помощью запроса определяется вероятность вхождения заданного терма в релевантные документы, а по полной коллекции документов определяется вероятность вхождения этого терма в нерелевантные документы.

Сети вывода

ПОХОЖИ НА ВЕРОЯТНОСТНУЮ МОДЕЛЬ

Так же, как и вероятностные модели, сети вывода основаны на принципе вероятностного ранжирования результирующих документов поиска. Главное их отличие от вероятностных моделей заключается в том, что используется оценка не вероятности релевантности документа запросу, а вероятности того, что он удовлетворяет информационным потребностям пользователя.

В рамках данной модели процесс поиска документов описывается как процесс рассуждений в условиях неопределенности. В процессе такого рассуждения оценивается вероятность того, что информационные потребности пользователя, выраженные с помощью одного или нескольких запросов, удовлетворены.



На чем основан метод

Сеть вывода основана на Байесовской сети, которая включает узлы четырех видов. Узлами первого вида являются документы коллекции, изученные пользователем в процессе поиска. Узлами второго вида являются термы, которыми описывается содержание документов, Узлами третьего вида являются запросы, состоящие из термов, которыми описывается содержание документов. Узел четвертого типа в сети только один, и он соответствует информационным потребностям пользователя, которые не известны поисковой системе. Все узлы первого и второго вида формируются заранее для заданной коллекции. Узлы третьего вида и их связи с узлами термов, описывающих документы, и узлом информационных потребностей формируются для каждого конкретного запроса.

В конце мы получаем

После того, как сеть построена, осуществляется оценка документов коллекции. Это реализуется распространением по сети оценки вероятности узла конкретного документа. Результатом распространения является вычисление вероятности узла информационных потребностей. При этом оценка для каждого документа строится независимо от оценок других документов, с учетом матриц описывающих связи между узлами документов и узлами термов, узлами термов и узлами запросов. Процесс оценки повторяется для каждого документа, затем они ранжируются на основе вычисленных оценок вероятности узла информационных потребностей.

Информационные источники

[1] Sergey Brin and Larry Page. The Anatomy of a Search Engine http://www-db.stanford.edu/pub/papers/google.pdf

[2] Илья Сегалович. Как работают поисковые системы http://company.yandex.ru/articles/article10.html

[3] Amy Langville and Carl Meyer. Deeper Inside PageRank http://meyer.math.ncsu.edu/Meyer/PS_Files/DeeperInsidePR.pdf

[4] Norbert Fuhr. Probabilistic Models in Information Retrieval http://www.is.informatik.uni-duisburg.de/bib/fulltext/ir/Fuhr:92.pdf

Спасибо за внимание!