



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

"МИРЭА - Российский технологический университет"

РТУ МИРЭА

Институт информационных технологий (ИИТ)

Кафедра практической и прикладной информатики (ППИ)

Доклад

Дисциплина: «Анализ и концептуальное моделирование систем»

Тема: «Модели информационного поиска в массиве документов»

Студент группы ИКБО-04-20 Хан А.А.

(подпись)

Преподаватель Черняускас В. В.

(подпись)

Доклад
представлен

«__» _____ 202__ г.

Москва 2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	3
МОДЕЛЬ ПОИСКА	4
БУЛЕВСКАЯ МОДЕЛЬ	5
ВЕКТОРНАЯ МОДЕЛЬ.....	6
ВЕРОЯТНОСТНАЯ МОДЕЛЬ.....	8
СЕТИ ВЫВОДА	11
ЗАКЛЮЧЕНИЕ	13
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	14

ВВЕДЕНИЕ

В связи с бурным ростом различных источников информации все большее значение приобретает проблема выделения нужных данных из огромного числа документов. Часто вопрос стоит уже не в том, чтобы найти документ или статью, где упоминается интересующая человека информация, а в выборе из сотен и тысяч подобных источников наиболее подходящий и наиболее полно освещающий нужный вопрос документов. Решить эту проблему помогают модели информационного поиска.

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.

МОДЕЛЬ ПОИСКА

Одним из ключевых понятий, характеризующих выбор того или иного метода анализа текстовой информации, а также реализацию конкретного варианта поиска, является модель поиска.

Модель поиска – это сочетание следующих аспектов информационного поиска:

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.

Вариации этих аспектов определяют большое количество всевозможных реализаций систем текстового поиска. Рассмотрим наиболее популярные в настоящее время.

БУЛЕВСКАЯ МОДЕЛЬ

Рассмотрим некоторый словарь $T = \{t_1, \dots, t_n\}$, где t_i — 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: $D \subseteq T$, иначе говоря $D \in \{0,1\}^n$: 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, « $t_5 \text{ OR } t_7 \text{ AND NOT } t_{12}$ », 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.

Такая модель иногда используется во внутренних корпоративных системах поиска, базах данных. Основным недостатком булевой модели является крайняя жесткость и непригодность для ранжирования. 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

Имеется словарь из термов, как в булевой 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

$$M_{ij} = TF_{ij} \cdot IDF_i,$$

где TF_{ij} (Term Frequency, частота термина) — относительная доля слова i в документе j ; IDF_i (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» M_{ij} . Первый сомножитель показывает, насколько данное слово подходит данному документу.

Для примера рассмотрим слово «Пьерс и произведение Л. Н. Толстого «Война и мир». Слово «Пьерс» 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: M_{ij} — степень соответствия слова i документу j .

Каждый документ представляется в этой матрице в виде столбца (j фиксировано, i меняется).

Для того чтобы подсчитать меру релевантности, представим сначала запрос в виде вектора с координатами 0 или 1: $Q = \langle t_3 \text{ AND } t_5 \rangle = \{0, 0, 1, 0, 1, 0, \dots, 0\}$.

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

Мерой релевантности $R(Q, D_j)$ будем считать косинус угла между вектором запроса Q и документом D_j . Для того, чтобы подсчитать это число возьмем скалярное произведение векторов Q и D_j :

$$R(Q, D_j) = \cos \alpha = \frac{Q D_j}{|Q| |D_j|}.$$

Нормализация необходима для того, чтобы уравнивать 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.

Если обобщить, то в основе модели лежит принцип вероятностного ранжирования (Probabilistic Ranking Principle, PRP). Этот принцип заключается в следующем - наивысшая общая эффективность поиска достигается в случае, когда результирующие документы ранжируются по убыванию вероятности их релевантности запросу. Сначала 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. Например, эта оценка может быть вычислена, в соответствии с теоремой Байеса, по некоторой функции вероятностей вхождения термов данного документа в релевантные и нерелевантные документы. С помощью запроса определяется вероятность вхождения заданного термина в релевантные документы, а по полной коллекции документов определяется вероятность вхождения этого термина в нерелевантные документы.

Документом будем считать множество слов без учета частоты встречаемости слова в документе. Можно также представить множество в виде

обычного булевского вектора $D = \{d_1, \dots, d_n\}$, где n — количество всех термов, а d_i 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 Q_k у нас имеются распределения вероятностей на всех документах «быть релевантным» и «быть нерелевантным» запросу Q_k . Обозначается это соответственно, как $P(R|Q_k, D)$ и $P(\bar{R}|Q_k, D)$. Тогда функцией соответствия будем считать отношение двух этих величин:

$$\frac{P(R|Q_k, D)}{P(\bar{R}|Q_k, D)}.$$

Теперь вспомним теорему Байеса:

$$P(a|b) = P(b|a) \frac{P(a)}{P(b)}$$

- $P(a)$ — априорная вероятность гипотезы a ;
- $P(b)$ — вероятность наступления события b ;
- $P(a|b)$ — вероятность гипотезы a при наступлении события b (апостериорная вероятность);
- $P(b|a)$ — вероятность наступления события b при истинности гипотезы a .

Применим ее для числителя и знаменателя дроби, стоящей в функции соответствия:

$$P(R|Q_k, D) = \frac{P(D|R, Q_k)P(R|Q_k)}{P(D|Q_k)}$$

$$P(\bar{R}|Q_k, D) = \frac{P(D|\bar{R}, Q_k)P(\bar{R}|Q_k)}{P(D|Q_k)}$$

$$\frac{P(R|Q_k, D)}{P(\bar{R}|Q_k, D)} = \frac{P(D|R, Q_k)P(R|Q_k)}{P(D|Q_k)} \frac{P(D|Q_k)}{P(D|\bar{R}, Q_k)P(\bar{R}|Q_k)} = \frac{P(R|Q_k) P(D|R, Q_k)}{P(\bar{R}|Q_k) P(D|\bar{R}, Q_k)}.$$

Заметим, что первый множитель $\frac{P(R|Q_k, D)}{P(\bar{R}|Q_k, D)}$ одинаков для всех документов, так как в нем не фигурирует D , и мы его дальше можем не рассматривать. Предполагая независимость всех слов (это очень сильное, и на практике неверное

предположение), второй сомножитель можно представить в виде произведения:

$$\frac{P(D|R, Q_k)}{P(D|\bar{R}, Q_k)} = \prod_{i=1}^n \frac{P(x_i = d_i|R, Q_k)}{P(x_i = d_i|\bar{R}, Q_k)},$$

где x_i — случайный документ, а d_i — число; $P(x_i = d_i|R, Q_k)$ — вероятность того, что i -й терм будет одновременно присутствовать или отсутствовать у случайного документа, релевантного нашему запросу, так же, как и в документе D . В произведении $\prod_{i=1}^n P(x_i = d_i|\bar{R}, Q_k)$ будут именно те вероятности, которые описывают сам документ D , и оно будет равно $P(D|R, Q_k)$ в предположении независимости всех слов.

СЕТИ ВЫВОДА

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.

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.

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.

ЗАКЛЮЧЕНИЕ

В зависимости от постановки задачи поиска могут применяться любые из рассмотренных моделей. На мой взгляд для большинства задач поиска стоит использовать векторные модели, так как они позволяют не просто искать документы, но и осуществлять их ранжирование в зависимости с соответствием запросу пользователя.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

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>