УДК 025.4.03

Аспекты применения в информационном поиске математических моделей

Ерохин В.В.
ФГАОУ ВО «Московский государственный институт международных отношений (университет) Министерства иностранных дел РФ»

Аннотация: Рассматриваются модели информационного поиска: теоретико-множественные модели поиска информации; векторная модель информационного поиска; вероятностная модель поиска информации. Анализ трех математических моделей извлечения информации показывает, что, несмотря на их низкую результативность, теоретико-множественные модели наиболее часто применяются из-за простоты реализации, в отличие от вероятностных моделей, которые требуют описательных процессов сложных параметров. Векторные модели являются наиболее популярными, поскольку их эффективность на практике выше.

Ключевые слова: информация, информационный поиск, математические модели.

Информационный поиск является одним из важных направлений деятельности коммерческих фирм. Эффективность функционирования коммерческих фирм зависит от качества их информационной деятельности.

Информационный поиск в коммерческих фирмах определяется как извлечение из документальных и информационных носителей новых знаний о объектах, субъектах и технологиях предметной деятельности фирмы, а также о событиях и фактах, имеющих отношение к инфраструктуре и области социально-экономической среды.

По мере повышения объемов баз данных возникает проблема организации результативного поиска информации, которую сотрудник коммерческой фирмы за минимальное время может найти в соответствии с содержанием запроса информационного поиска. Запросом информационного поиска является содержащее требования на выдачу информации, либо какое-либо входное сообщение. Результатом информационного поиска может быть совокупность электронных документов или электронных ресурсов, соответствующих информационной надобности пользователя. Информационной надобностью является информация, которую пользователь желает получить [1, 2].

Классический информационный поиск ориентирован на структурированные данные. Однако в большинстве случаев информация представлена не как структурированные данные, а как простые текстовые электронные документы или базы

данных (БД). При этом в БД, хранящих структурированные данные, имеются поля, отображающие массивы текстовой информации. Характеристиками такой неструктурированной информацией является по каким-либо формальным признакам неупорядоченность. При работе с такими данными проявляются сложности, особенно при проведении информационного поиска.

Проблема поиска информации в текстовых полях данных - это междисциплинарная область науки, называемая поиском информации, основанная на достижениях информатики и лингвистики. Ключевой практической проблемой поиска информации является суждение о степени совпадения информации, содержащейся в информационном электронном документе, с запросом информации.

Запрос информации содержит естественное выражение информации, которая состоит из нескольких слов на естественном языке. При поиске информации необходимо отыскать различные информационные документы, которые найдут все варианты написания, содержащиеся в информационном запросе [3].

Чтобы эффективно выполнять поиск информации, текстовое содержимое информационного запроса и документа должно быть лингвистически обработано, поскольку большинство слов в русском языке представлены в тексте в виде текстовых форм. Слова русского языка содержат десятки различных грамматических описаний. Поэтому необходимо поставить слова на общую морфологическую основу. Текст в естественном языке содержит следующие категории слов смыслового значения: союзы, предлоги и наречия. Речевая обработка включает в себя синтаксический и морфологический анализ генерируемой естественным языком текстовой информации. Эта обработка является необходимым шагом, который предшествует построению математических моделей для извлечения информации [4].

Существует три основных класса, которые можно отличить от всех моделей получения информации [5].

1. Теоретико-множественные модели поиска информации.

В этих моделях поиска информации используется теория множеств.

Например, логическая модель поиска – это модель поиска информации, в которой вы можете обрабатывать любой запрос в форме логического выражения AND, OR и NOT. Документ рассматривается как серия слов. Если этот термин существует в документе, соответствующая переменная принимает значение «True» (логический блок и документ считается релевантным) или «False» (логический нуль и документ не имеет значения). Модель обработки логического запроса информации имеет несколько

недостатков:

• Многие электронные документы могут выводиться для конкретного запроса, и пользователь должен добавить дополнительные условия для запроса, чтобы снизить результирующую выборку. Поиск осуществляется путем проб и ошибок. Поэтому запрос отображается в сложной логической формуле, которая требует от пользователя не только знания предметной области, но и знания правил для создания формул.

• Полученная выборка не может быть оценена по релевантности, так как имеются только два значения релевантности («релевантные» и «нерелевантные»).

Преимущества булевой модели включают предсказуемость результатов информационного поиска.

Современные логические модели поиска информации также содержат операторы близости для элементов запроса информации. Это может быть либо количество промежуточных слов между элементами запроса, выявленными в электронном документе, либо указание местоположения выявленных элементов запроса.

Иным примером является модель нечетких множеств, где допускается частичное членство элемента в множестве. Требования к информации в этой модели представлены так же, как и логическая модель, но логические операции переопределяются для учета возможности неполного элемента, принадлежащего множеству. Согласование электронных документов с запросом определяется так же, как и логическая модель. Эта модель также практически не позволяет ранжировать результаты поиска.

Ключевым критерием для получения информации является степень соответствия информационным потребностям пользователя, то есть релевантность, которая является субъективной концепцией, и степень зависит от лица, оценивающего результаты поискового запроса.

Наиболее применяемым методом в оценке релевантности в теоретико-множественных моделях для получения информации является метод TF-IDF. Этот метод применяется в большинстве информационно-поисковых систем. Сущность метода – определение значимости текстовой информации в информационном документе для слова или фразы слов в информационном запросе. Вес слова пропорционален количеству слов, используемых в тексте, и обратно пропорционально частоте использования слова в других документальных БД.

Структура формулы TF-IDF.

TF (term frequency – частота слова) – отношение количества вхождений какого-либо слова к общему количеству в электронном документе слов. Таким образом, можно оценить важность слова l_i в пределах рассматриваемого текста m_i.

$$TF = m_i(\sum m_k)^{-1},$$

где m_i – количество анализируемых употреблений слова; m_k – общее количество употреблений слов.

IDF (inverse document frequency – обратная частота документа) – частотная инверсия, с которой определенное (заданное, требуемое) слово в электронных документальных БД встречается. Это снижает весовые оценки наиболее часто употребляемых слов.

$$IDF = \log|N| - \log|n_i \supset l_i|,$$

где $|N|$ – общее количество электронных документов; $|n_i \supset l_i|$ – количество электронных документов, в которых встречается слово l_i, (если $m_i \neq 0$).

Результат метода TF-IDF определяется произведением двух сомножителей: TF и IDF. Наибольшую весовую оценку в методе TF-IDF получают слова с максимальной частотой в границах анализируемого электронного документа и с пониженной частотой употреблений в иных электронных документах.

В случаях, когда информационная поисковая система возвращает совокупность электронных документов, соответствующих запросу, для оценивания эффективности информационного поиска применяют два базисных статистических показателя: точность и полноту.

Точность информационного поиска характеризуется долей релевантных электронных документов в количественном объёме найденных документов. Формула по определению точности информационного поиска (P) имеет вид:

$$P = N_r / N_{all},$$

где N_r – количество выявленных релевантных электронных документов, N_{all} – количество выявленных электронных документов.

Полнотой информационного поиска (V) называется доля выявленных релевантных электронных документов среди всех релевантных электронных документов. Формула по расчету полноты информационного поиска имеет следующий вид:

$$V = N_r / N_{r.all},$$

где $N_{r.all}$ – количество найденных и не найденных релевантных электронных

документов.

Точность и полнота поиска информации имеют противоречивый характер. Полноту возможно увеличить до единицы в случаях возвращения всех электронных документов на все запросы. Точность как правило уменьшается. Чтобы отыскать баланс между полнотой и точностью в информационном поиске необходимо использовать их среднее гармоническое взвешенное (H).

$$H = PV[xV + (1 - x)P]^{-1},$$

где x – переменная баланса между полнотой и точностью, $x \in [0,1]$.

При осуществлении поиска информации:

$x = 0,5$ – одинаковые весовые оценки присвоены полноте и точности;

$x > 0,5$ – точность является предпочтительной;

$0 \leq x < 0,5$ – полнота является предпочтительной.

При $x = 0,5$ формула по расчёту среднего гармонического взвешенного имеет вид:

$$H = 2PV[V + P]^{-1}.$$

2. Векторная модель информационного поиска.

При применении векторной модели информационного поиска все слова электронного документа должны иметь одинаковую значимость. При этом требуется вычислить весовые оценки соответствия электронных документов запросу, которые должны рассчитываться на базе входящих в информационный запрос слов. Для расчета значимости слов в информационном запросе применяется частота употреблений слова в электронном документе, т.е. чем наиболее часто наблюдается слово в запросном документе, тем значительнее его весовая оценка. При этом слова в электронном документе могут располагать различной значимостью. Можно сделать вывод, что чем наиболее часто в совокупности документов используется требуемое слово, тем меньше при ранжировании результатов информационного поиска его значимость.

Электронный документ в векторной модели поиска информации представляется в виде набора векторов с координатами слов информационного запроса в зависимости от частоты вхождения в электронный документ:

$$W_{ji} = lf_{ji} \times inf_{ji}, \quad j = 1 \ldots H,$$

где lf_{ji} – частота слова j в документе i; inf_{ji}, – частота встречаемости слова j по всем электронным документам.

Векторная модель информационного поиска имеет преимущество в виде результирующей выборки, которая по релевантности может быть упорядочена.

Недостатком этой модели является то, что в тексте электронного документа используется совокупность независимых слов.

3. Вероятностная модель поиска информации.

Вероятностная модель имеет возможность учитывать весовые оценки и взаимосвязи разнообразных слов информационного запроса и электронных документов. Электронные документы и запросы к ним анализируются подобно векторной модели поиска информации. При этом вводятся дополнительно два параметра: вероятность нерелевантности P_{jnr} электронного документа j и вероятность релевантности P_{jr}. Устанавливаются дополнительные параметры R_{nr} и R_{nor} – потери, характеризующие соответственно получение нерелевантных электронных документов и неполучение – релевантных. Функция релевантности имеет вид:

$F(j) = P_{jr} / P_{jnr} – R_{nor} / R_{nr}$.

Значение функции $F(j)$ характеризует нерелевантность или релевантность электронного документа. Функция $F(j)$ может быть рассчитана только во взаимосвязи с другими параметрами.

Вероятностная модель не дает значительных результатов для роста результативности поиска информации из-за сложности получения верных значений требуемых параметров [4, 5].

Вследствие анализа трех математических моделей извлечения информации следует, что, несмотря на их низкую результативность, теоретико-множественные модели наиболее часто применяются из-за простоты реализации, в отличие от вероятностных моделей, которые требуют описательных процессов сложных параметров. Векторные модели являются наиболее популярными, поскольку их эффективность на практике выше.

Выбор математической модели извлечения информации при решении поставленных задач зависит от: требований к точности и полноте извлечения информации; требования к созданию результатов поиска информации; насколько результативна математическая модель для проведения поиска; квалификационных и компетентностных показателей пользователей.

Список литературы:

1. Аверченков, В.И. Системы организационного управления / В.И. Аверченков, В.В. Ерохин. – Брянск: БГТУ, 2006. – 208 с.

2. Аверченков, В.И. Системы организационного управления / В.И. Аверченков, В.В. Ерохин. – Брянск: БГТУ, 2012. – 208 с.
3. Ерохин, В.В. Инновационная деятельность в повышении конкурентоспособности продукции / В.В. Ерохин // Вестник образовательного консорциума среднерусский университет. Информационные технологии. – 2015. – №6. – С. 25-28.
4. Ерохин, В.В. Анализ финансового рынка с использованием информационных систем / В.В. Ерохин, Е.В. Елисеева, А.М. Хлопяников // Результаты социально-экономических и междисциплинарных научных исследований XXI века– Самара: Поволжская научная корпорация, 2016. – С.62-74.
5. Ерохин, В.В. Безопасность информационных систем / В.В. Ерохин, Д.А. Погонышева, И.Г. Степченко. – Москва: ФЛИНТА, Наука, 2015. – 184 с.
6. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Московский университет, 2011. – 512 с.