УДК [004.89:004.85:004.421:004.58]+519.688

АДАПТИВНЫЙ ПОИСК ПО ЛОГИЧЕСКИМ ВЫРАЖЕНИЯМ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

Шерстнев А.С., студент GeekBrains [1],
Савенко А.Г., старший преподаватель кафедры ИСиТ [2]

ООО «АйтиРексГрупп Бел» [1],
Институт информационных технологий Белорусского государственного университета информатики и радиоэлектроники [2],
г. Минск, Республика Беларусь

Савенко А.Г. – магистр техн. наук

Аннотация. В работе представлены исследования моделей и алгоритмов поиска информации в информационно-поисковых системах. Проведен анализ существующих проблем и подходов к поиску информации. Разработаны и описаны: модель организации данных, модель подготовки и обработки данных, алгоритм токенизации данных информационной системы, адаптивный поисковый алгоритм по запросу, состоящему из логических выражений. Проведен анализ эффективности предложенных моделей и алгоритмов. Разработана информационно-поисковая система.

Ключевые слова. Адаптивный поиск, поисковые алгоритмы, модель организации данных, графовая база данных, поисковый запрос, логические выражения, релевантность, машинное обучение, информационные системы.

Введение. Поиск информации является важной задачей практически любой автоматизированной системы, связанной с обработкой текстовой информации. В настоящее время уже существует множество методов поиска, систематизированных следующим образом: методы на основе ключевых слов; методы на основе метрик (Дамерау, Левенштейна, Жаккара и др.); ассоциативные методы (фильтр Блюма и методы нечеткого хеширования Корнблюма и Чарикара); методы последовательного поиска (Бойера-Мура); поисковые деревья и др.

Критериями качества поиска в информационно-поисковых системах (ИПС) могут являться точность, полнота поиска, выпадение и F-мера, а также быстродействие поисковых алгоритмов. К сожалению, понятие степени соответствия результатов поиска и их полнота, т.е. релевантность, является субъективным понятием, и зависит от конкретного человека, оценивающего полученные результаты. Также следует отметить, что дополнительным критерием, влияющим на качество и эффективность работы поисковых алгоритмов, является гибкость и удобство формулирования поискового запроса. Если для оптимизации существующих алгоритмов поиска в основном используются различного рода упорядочивание данных (от сортировки данных, до сложных индексирующих систем), то для задания критериев самого поиска, практически нет никаких рекомендаций или оптимизаций. В основном гибкость формирования поисковых запросов определяется неточным совпадением символов или последовательностей символов при поиске подстроки в строке [1, 2]. Кроме того, следует отметить, что различные поисковые алгоритмы по-разному воспринимают пробел (как разделитель между ключевыми словами): некоторые заменяют его логическим оператором «И», некоторые – логическим оператором «ИЛИ», другие же вовсе не используют логических операторов [3].

Поисковые алгоритмы использующие логические операторы (AND – логическое умножение, OR – логическое сложение, NOT – логическое отрицание) и их комбинации относятся к булевой модели поиска. Такая модель формирования поискового запроса позволяет получать более точные и, соответственно, релевантные результаты. Однако булева модель также имеет и ряд недостатков [3]. Основными недостатками существующих ИПС является их проприетарность и невозможность изменения поведения поиска под конкретную предметную область, ограниченность по возможностям использования логических операторов и описанные выше недостатки булевой модели поиска.

Основная часть. С точки зрения информационной модели ИПС, необходимо реализовать сложный поиск на основе логических выражений. В данном случае целесообразно хранить данные информационной системы в виде связей между сущностями, при комбинировании которых можно добиться нужного критерия поиска. Сами сущности должны описывать как можно меньший объем информации для того, чтобы обеспечивать меньшую гранулярность и, следовательно, предоставлять более точные результаты. Связи должны быть односторонними и направленными, а их количество должно быть минимальным.

Пользователь ИПС может не обладать необходимыми компетенциями для формирования точного поискового запроса и понимания предметной области запроса. Отсюда возникает необходимость реализации адаптивного поиска, с той точки зрения, что модель организации данных должна «обучаться» и накапливать также данные, отсутствующие в ИПС, но необходимые для получения релевантного результата. Модель организации данных ИПС целесообразно реализовать в виде графа с послойной организацией данных.

Пример такой организации данных в разрезе одного поискового документа представлен на рисунке 1.

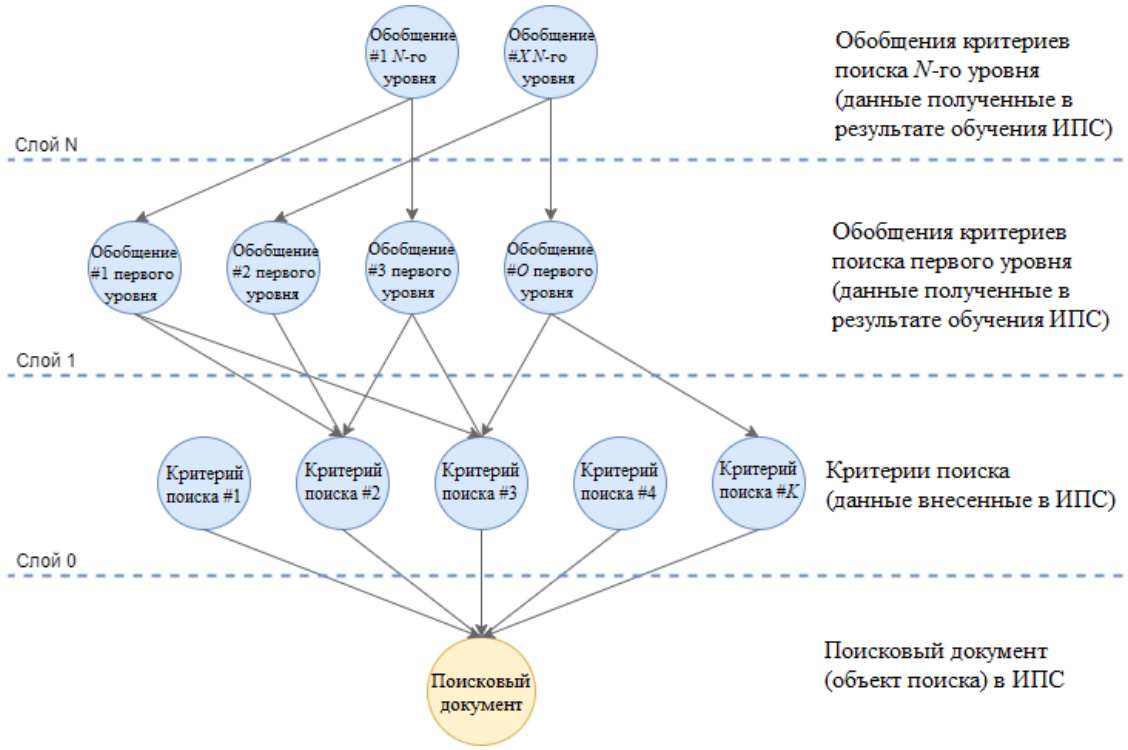

Рисунок 1 – Пример графовой модели организации данных в разрезе одного поискового документа

Как видно из рисунка 1 на самом нижем уровне находится узел с информацией о поисковом документе. Данный узел связан с узлами критериев поиска односторонней связью. В свою очередь узлы критериев связанны со следующим слоем обобщений данных в более широкие группы понятий. Данное обобщение позволит алгоритму поиска находить документы по более широким понятиям их критериев. Так же на одном условном слое связи между узлами запрещены, чтобы предотвратить цикличность при поиске. Таким образом, послойность организации данных и направление связей между слоями, отсутствие связей между узлами одного слоя предотвращают пересечение лишних поисковых документов и при использовании слоев обобщений позволяют сделать поиск адаптивным, а результаты релевантными. Поиск осуществляется в направлении от верхнего слоя (*N*) к нижнему, а результаты поиска также будут содержать информацию одного верхнего слоя (нулевого).

Рассмотрим пример использования предложенной модели в информационно-поисковой системе библиотеки. Поисковым документом может являться книга с определённым названием. Тогда критериями поиска могут быть фамилия автора, год издания, издательство, количество страниц и т. д. Эти данные изначально заносятся и хранятся в базе данных. Обобщениями, полученными при обучении информационно-поисковой системы, например, для такого критерия как год издания могут быть следующие: обобщение первого уровня – «период второй мировой войны», обобщением второго уровня – «советский период» и т. д. Таким образом, не зная определенного года издания, но зная, что книга написана автором в советском союзе или во время второй мировой войны в результате запроса «Купала AND советский период» мы получим релевантные результаты поиска.

Помимо задачи непосредственно поиска данных, первостепенным является их подготовка, обработка и загрузка в графовую базу данных. Входными данными будут поисковые документы и список критериев их поиска. Они должны быть максимально конкретными для того, чтобы алгоритму обучения базы данных (формирования обобщений критериев поиска) было проще выявить нужные обобщения и классы, к которым данные критерии относятся. Выходными данными являются сами критерии поиска, а также список обобщений, к которым данные критерии относятся. Для формирования таких выходных данных, на первом этапе необходимо дополнить каждый входной критерий каким-либо его определением или характеристикой на естественном языке. На втором этапе полученное из стороннего источника данных описание на естественном языке необходимо разбить на отдельные части и выбрать из них ключевые лексемы (токенизировать).

Для машинного обучения базы данных информационно-поисковой системы можно использовать библиотеку обработки естественных текстов NLTK [4]. Базовая модель подготовки данных для ИПС представлена на рисунке 2.

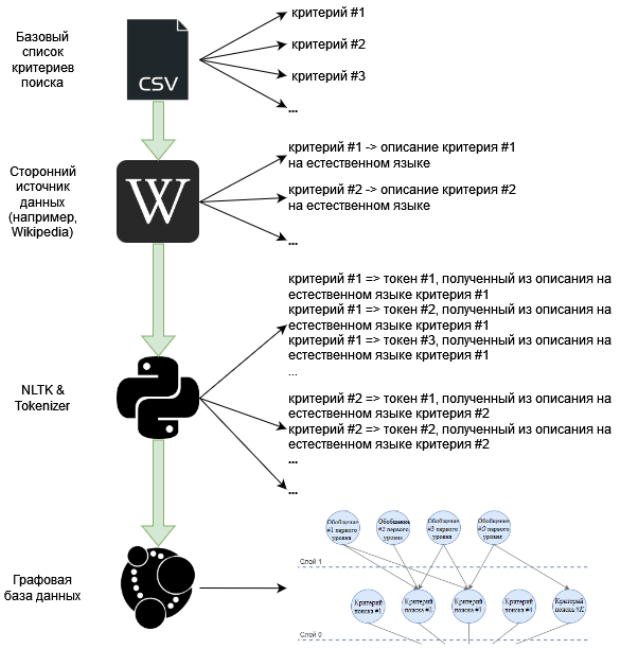

Рисунок 2 – Базовая модель подготовки данных для ИПС

Для получения токенов (ключевых лексем) из описания критериев поиска на естественном языке и связанных с каждым критерием поиска необходимо выполнить следующую последовательность действий (алгоритм токенизации данных):

— шаг 1. Входящее описание разбивается на именные группы (словосочетания). Выбираются словосочетания, в которых имя существительное является вершиной, то есть главным словом, определяющим характеристику всей составляющей;

— шаг 2. Выполнить цикл по всем получившимся именным группа и каждую разбить на слова;

— шаг 3. Из полученных слов исключить все «стоп-слова», все знаки пунктуации;

— шаг 4. На заключительном шаге формируется очищенная именная группа, которая и попадет в базу данных как одно из обобщений.

Поисковой запрос может включать в себя основные логические операции, такие, как И, ИЛИ, НЕ, а также операцию группировки с приоритетом. Процесс поиска представляет собой обход графа.

Однако, прежде чем делать обход по графу необходимо преобразовать входящее выражение в дизъюнктивную нормальную форму (ДНФ). В качестве логических литералов выступают критерии поиска. ДНФ позволит преобразовать любое входящее логическое выражение в вид дизъюнкции конъюнкций литералов, что позволит произвести минимизацию логического выражения и ускорить алгоритм поиска (за счет уменьшения количества логических литералов) и даст возможность разбить алгоритм на три этапа:

— преобразование исходного выражения в ДНФ. Для этого необходимо закодировать критерии поиска в логический литерал (например, (критерий 1 ∧ критерий 2) ∨ (критерий 3 ∧ критерий 2));

— поиск всех документов, которые должны обладать несколькими критериями одновременно, при этом если критерии находится на более высоком уровне агрегации, то вначале необходимо найти все критерии на самом низком слое связанные с текущим. То есть выполняется цикл по всем конъюнктивным группам и формирование запросов для обхода графа с учетом включения всех критериев поиска;

— объединение всех результатов предыдущего этапа и удаление встречающихся дубликатов [5].

Алгоритм ранжирования результатов поиска. Для решения проблемы ранжирования результатов поискового запроса предлагается использовать алгоритм подсказок, который предоставит пользователю дополнительную информацию (основываясь на полученных результатах поиска) для ранжирования после первой итерации поиска. Это позволит пользователю при последующем уточнении поискового запроса использовать уже полученную информацию, например выбрать чаще встречающиеся в результате критерии поиска и таким образом получить результат в некотором ранжированном виде.

Для реализации данной идеи в предложенной модели после каждого поискового запроса можно формировать список «подсказок», сформированный из узлов непосредственно связанных с узлами результатов поиска. После чего выполнить подсчет количества каждого из узлов и уже выдать результат пользователю в ранжированном виде.

Предложенный механизм «подсказок» был проверен в ходе эксплуатации на примере ИПС предприятия, реализующей поиск сотрудников по их компетенциям. В результате работы при

поисковом запросе, имеющем вид «language AND database», будут получены два сотрудника, как и ожидалось – на рисунке 3 изображены желтыми кругами. Дальше алгоритм выберет все компетенции, которые связанны с сотрудниками – на рисунке 3 изображены зелеными кругами.

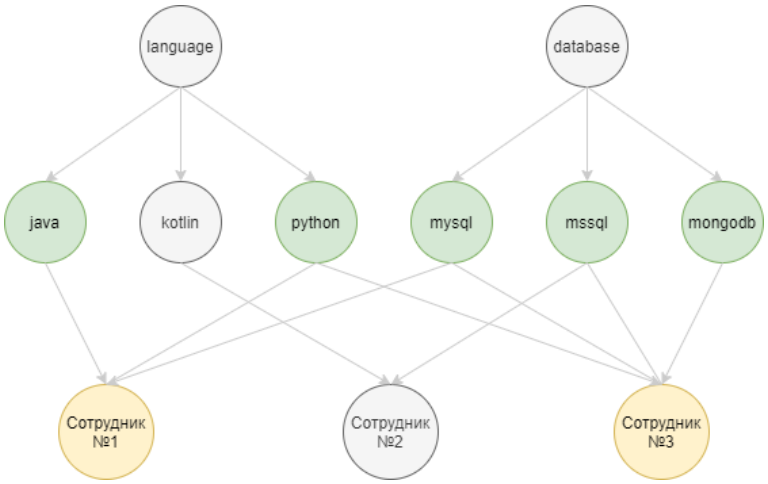

Рисунок 3 – Пример графа результатов поиска

Далее после подсчета количества всех компетенций (критериев поиска) алгоритм выдаст результаты подсчета вершин критериев поиска. Затем любую из этих вершин (компетенций), предложенную как подсказку, пользователь может использовать для уточнения поискового запроса и соответственно ранжирования искомых сотрудников.

Заключение. Благодаря предложенной модели ранжирование результатов можно осуществлять даже с учетом значимости критериев поиска, явно не заданных в поисковом запросе.

Важной задачей при разработке модели адаптивного поиска является ее быстродействие, т. к. необходимо обрабатывать большой объем данных за приемлемое время ожидания пользователя. Для оценки предложенных моделей был произведен анализ быстродействия, давший приемлемые результаты при числе слоев до тридцати [6].

Список использованных источников:

1 Зуенко А.А. Поисковые запросы на основе операций с логическими векторами / А. А. Зуенко, А. А. Алмаматов // Труды Кольского научного центра РАН. – 2013. – Выпуск № 5 (18). – с.119-124.

2 Шоркин, А. П. Методы и алгоритмы информационного поиска на неточное соответствие / А. П. Шоркин // Доклады БГУИР. - 2011. - № 2 (56). - С. 13 - 15.

3 Касекеева, А. Б. Исследование методов информационного поиска / А. Б. Касекеева // BIG DATA and Advanced Analytics : сборник материалов V Международной научно-практической конференции, Минск, 13–14 марта 2019 г. Ч.1 / БГУИР; редкол. : В. А. Богуш [и др.]. – Минск, 2019. – С. 324 – 330.

4 Документация по библиотеке NLTK [Электронный ресурс]. Режим доступа: https://nltk.org.

5 Savenko, A. G. Model and algorithm for adaptive search by logical expressions / Savenko A. G., Sherstnev A. S. // Информационные технологии и системы 2021 = Information Technologies and Systems 2021: материалы международной научной конференции, Минск, 24 ноября 2021 г. / БГУИР; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2021. – С. 109–110.

6 Савенко, А. Г. Модели и алгоритмы для адаптивного поиска в информационно-поисковых системах / Савенко А. Г., Шерстнев А. С. // Веснік сувязі. – 2022. – №1. – С. 47–53.

UDC [004.89:004.85:004.421:004.58]+519.688

SEARCH BY LOGICAL EXPRESSIONS IN INFORMATION SEARCH SYSTEMS

Sherstnev A.S. [1], Savenko A.G. [2]

ITRex Group Bel LLC [1],
Institute of Information Technologies of the Belarusian State University of Informatics and Radioelectronics [2],
Minsk, Republic of Belarus

Savenko A.G. – Senior Lecturer, Master of Engineering Sciences

Annotation. The paper presents studies of models and algorithms for information retrieval in information retrieval systems, an analysis of existing problems and approaches to information retrieval is carried out. As a result of the study, a data organization model, a data preparation and processing model, an information system data tokenization algorithm, an adaptive search algorithm for a query consisting of logical expressions were created, an analysis of the effectiveness of the developed models and algorithms was carried out, an information retrieval system (web application) was developed. with embedded models and algorithms.

Keywords. Adaptive search, search algorithms, data organization model, graph database, search query, logical expressions, relevance, machine learning, information systems.