

Особенности определения отношений в тезаурусе в области программирования

И. А. Коломойцева, С. С. Бердюкова

Донецкий национальный технический университет
кафедра Программной инженерии им. Л.П. Фельдмана
E-mail: bolatiger@mail.ru

Аннотация:

В статье выполнен анализ существующих модели информационного поиска. Рассмотрены особенности формирования тезауруса в области программирования. Обоснована необходимость использования такого тезауруса в информационном поиске. Описаны структура информационно-поисковой системы с использованием тезауруса и особенности отношений между понятиями в тезаурусе («ниже-выше», «часть-целое», ассоциации), позволяющие его отнести к формальным онтологиям.

Введение

Объём электронных документов, с которыми сталкивается человек, постоянно растёт. И задача поиска информации приобретает всё большую актуальность.

Под поиском информации понимается поиск в коллекции документов, которые являются наиболее релевантными по отношению к произвольным информационным потребностям, выражаемым (представленным) при помощи однократных запросов пользователей [1]. При этом информационная потребность – это тема, о которой пользователь хочет знать больше (следует её отличать от информационного запроса).

На сегодняшний день существуют модели информационного поиска, такие как, например, булев поиск, поиск с помощью векторно-пространственного представления, вероятностный поиск, языковые модели, в том числе, и большие языковые модели. Большие языковые модели (Large Language Models, LLM) успешно решают большинство задач информационного поиска, но обладают одним существенным недостатком. Для построения LLM нужен большой качественный датасет и большие вычислительные мощности [2]. При этом, в существующих моделях LLM редко учитываются особенности конкретных предметных областей. Решение, которое предлагается в этом случае, предусматривает донстройку (расширение) существующей базовой модели LLM на датасете, состоящем из текстов, относящихся к выбранной предметной области. Но при таком подходе возникает зависимость от базовой модели LLM [3, 4].

Автором статьи предлагается другой подход к улучшению характеристик информационного поиска (полноты и в некоторых случаях точности) – использование

такого онтологического ресурса как информационно-поисковый тезаурус. Тезаурус – это словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемыми понятиями, концептами или дескрипторами, и в которых явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями [5]. С помощью тезауруса можно расширять за счёт свойств синонимии, ассоциации и зависимости терминов запроса. Способы расширения запроса описаны в [6].

Целью данной статьи является описание отношений в тезаурусе в области программирования.

Структура информационно-поисковой системы с использованием тезауруса

Информационный поиск – это процесс поиска в большой коллекции некоего неструктурированного материала (обычно – документа), удовлетворяющего информационные потребности [1].

Основными характеристиками информационного поиска являются точность и полнота [1].

Точность – доля релевантных документов среди найденных.

Полнота – доля найденных документов среди всех релевантных.

Показатель, позволяющий найти баланс между точностью и полнотой, называется F-мерой.

Найти нужную информацию пользователь может с помощью информационно-поисковой системы (ИПС). Информационно-поисковая система – это компьютерная система, предоставляющая пользователю возможность доступа к интересующей его информации из

некоторой коллекции документов, расположенной локально или в компьютерной сети, например, сети Интернет.

Общая структура информационно-поисковой системы включает в себя получение запроса от пользователя, предобработку этого запроса, отправку его в модуль поиска, получение списка документов, предположительно удовлетворяющих этому запросу, и ранжирование этого списка. Предобработка запроса нужна для того, чтобы термины запроса привести к тому же виду, что и термины документа. В этом случае запрос и документ из коллекции документов можно сравнивать между собой. Расширение запроса – это обычная практика, используемая в ИПС для повышения полноты поиска [7, 8]. Такие поисковые системы как Google или Yandex расширяли запрос с помощью статистического анализа предыдущих запросов самого пользователя и других пользователей. Сейчас крупные поисковые системы используют для уточнения запроса большие языковые модели.

Вместо сложных статистических моделей и LLM для улучшения характеристик поиска среди документов некоторой предметной области, например, программирования

предлагается использовать тезаурус, построенный по этой области. Тезаурус предлагается строить автоматически на основе относительно небольшого датасета. Структурная схема ИПС с использованием тезауруса представлена на рисунке 1.

Отношения в тезаурусе в области программирования

Отношения являются важной частью информационно-поискового тезауруса как онтологического ресурса.

Во-первых, именно отношения позволяют расширить запрос.

Во-вторых, отношения играют важную роль при разрешении многозначности. Для предметной области «Программирование», в которой многозначность терминов – частое явление, такая функция очень важна и при расширении запроса, и при анализе выдачи документов.

Для реализации любой из этих функций необходим своеобразный логический вывод: встретив понятие в запросе необходимо выполнять многошаговые проходы по отношениям.

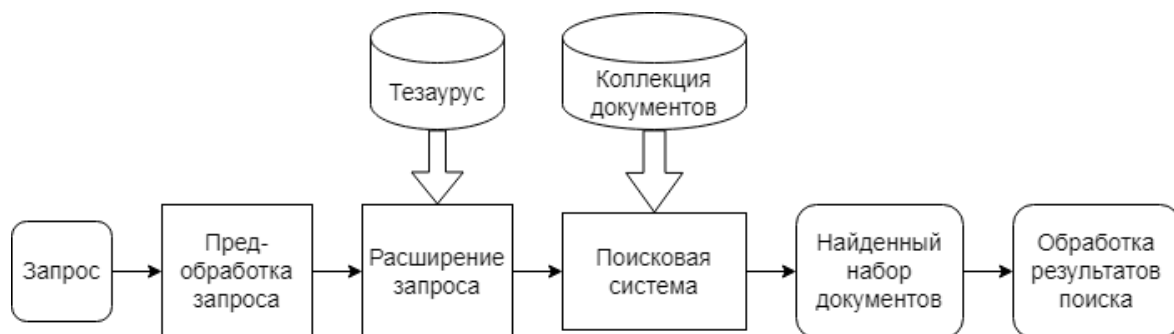


Рисунок 1 – Структура информационно-поисковой системы с использованием тезауруса

В информационно-поисковом тезаурусе по предметной области «Программирование», как и в большинстве информационно-поисковых тезаурусов [9, 10], используются отношения «выше-ниже», «часть-целое» и отношение ассоциации. При этом с отношением ассоциации могут возникать сложности из-за его субъективности.

И хотя в литературе неоднократно возникают предложения по замене информационно поисковых тезаурусов, а более формализованные онтологии с более подробной системой отношений, но в [3] указано, что использование сложных онтологических формализмов нецелесообразно. Поэтому в информационно-поисковом тезаурусе для предметной области «Программирование» используется обычный набор отношений для такого рода ресурсов.

Принципы построения отношений

Отношение «выше-ниже» – это родовидовое отношение типа «класс-подкласс», обладает свойствами наследования и транзитивности. Относится к таксономическим отношениям.

Отношение «выше-ниже» обладают свойствами онтологических отношений «класс-подкласс», такими как [3]:

каждый пример видового понятия в любой момент своего существования должен быть примером родового понятия;

видовое понятие должно относиться к тому же семантическому классу, что и родовое понятие;

видовое понятие должно наследовать основные свойства родового понятия.

Таковыми же свойствами обладают отношения между ролевым понятием и понятием класса, если экземпляры только этого класса могут выступать в данной роли (МЕТОД КЛАССА – ФУНКЦИЯ).

Еще один тип отношений, обладающий такими свойствами, – это отношение между фазой какой-либо сущности и собственно этой сущностью (НАЧАЛЬНОЕ ЗНАЧЕНИЕ ПАРАМЕТРА – ПАРАМЕТР, ПОЛУЧЕННЫЙ В РЕЗУЛЬТАТЕ РАСЧЕТА).

Таким образом, отношение «выше-ниже» обладает свойствами несимметричности и транзитивности:

выше (X,Y) ∧ выше (Y,Z) → выше (X,Z)
ниже (X,Y) ∧ ниже (Y,Z) → ниже (X,Z)
выше (X,Y) → ниже (Y,X)

Но ещё больше сложностей возникает при построении тезауруса смешение типов и ролей в одной иерархии. Отношения тип-тип (анонимная функция – функция) и тип-роль (анонимная функция – замыкание) могут пройти все тесты установления родовидовых отношений. Но анонимная функция в любом случае останется функцией, и при этом она может быть замыканием, а может не быть им.

Отношение «часть-целое» играет очень важную роль в информационно-поисковых тезаурусах. Это связано с определением релевантности документа запросу пользователя. Можно предположить, что, если документ посвящен описанию части какого-то объекта, то он релевантен запросу, содержащему термин, определяющий целое. При этом сам термин, обозначающий термин, в документе может отсутствовать. Если, например, в документе описываются правила определения методов класса, то он с большой долей вероятности относится к понятию «класс в ООП» и «Объектно-ориентированному программированию». Примеры отношений «часть-целое» из тезауруса в области программирования приведены в таблице 1.

Таблица 1 - Примеры отношений «часть-целое» из тезауруса в области программирования

Часть	Целое
Метод	Класс
Оператор	Программа
Локальная переменная	Функция
Счётчик цикла	Цикл
Элемент массива	Массив
Объект ядра ОС	Ядро ОС
Стек потока	Поток
Контроллер	Паттерн MVC
Владелец мьютекса	Мьютекс

В лингвистике для определения отношения «часть-целое» используются лингвистические тесты – заданные предложения, в которые подставляются анализируемые сущности. Например, X – это часть Y, которое должно нормально звучать для некоторых X и Y. X (часть) называется меронимом, а Y (целое) – холонимом [4]. Это один из подходов, который используется при определении отношения «часть-целое» для тезауруса программирования. Задаются шаблонные фразы, по которой определяется, что X часть Y. Например, «X входит в Y», «X является частью Y» и т. п.

Второй подход является онтологическим. Важным принципом определения отношения «часть-целое» является следующее утверждение: уничтожение (изменение) предполагаемой части оказывает влияние на предполагаемое целое. Например, если из КОМПЬЮТЕРНОЙ ПРОГРАММЫ удалить все ОПЕРАТОРЫ, то КОМПЬЮТЕРНАЯ ПРОГРАММА прекратит своё существование. Наличие или отсутствие у МЬЮБЕКСА ВЛАДЕЛЬЦА МЬЮТЕКСА меняет состояние МЬЮТЕКСА (свободное/занятое).

Важным свойством отношения «часть-целое» является транзитивность. Благодаря этому свойству можно строить цепочки логического вывода. Например (читается слева направо):

МЕТОД – КЛАСС – ОБЪЕКТНО-ОРИЕНТИРОВАННОЕ ПРОГРАММИРОВАНИЕ
АТРИБУТ ЗАЩИТЫ – ОБЪЕКТ ЯДРА ОС – ЯДРО ОС

Отношение ассоциации – еще тип отношений, который присутствует в информационно-поисковом тезаурусе в области программирования. Отношение ассоциации, как и отношение «часть-целое», связано с релевантностью документа запросу. Считается, что если два понятия C1 и C2 связаны отношением ассоциации, то тексты, содержащие понятие C1, релевантны запросу, содержащему понятие C2, и наоборот. Например, такими понятиями являются ПОТОК и НИТЬ.

Вывод

Задача информационного поиска является сложной и актуальной задачей. Одним из подходов к решению этой задачи – является использование информационно-поискового тезауруса в конкретной области, в частности, в области программирования.

В работе рассмотрены особенности и сложности формирования тезауруса в области программирования. Определены отношения, встречающиеся в таком тезаурусе – «ниже-выше», «часть-целое» и ассоциации. Описано, какие сложности возникают при определении этих отношений.

Литература

1. Маннинг, К.Д. Введение в информационный поиск / К.Д. Маннинг, П. Рагхаван, Х. Шютце. – М.: ООО «И.Д. Вильмс», 2011. – 528 с.
2. Large Language Models for Information Retrieval: A Survey / Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, etc, - URL <https://paperswithcode.com/paper/large-language-models-for-information-retrieval> (дата обращения 01.03.2024).
3. Fine Tuning Large Language Model (LLM). – URL <https://www.geeksforgeeks.org/fine-tuning-large-language-model-llm/> (дата обращения 01.03.2024).
4. Филимонов, В. Ю. Большие языковые модели и их роль в современных научных открытиях / В. Ю. Филимонов // Философские проблемы информационных технологий и киберпространства. – 2024. – №1(25). – С. 42-57.
5. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Издательство Московского университета, 2011. – 512 с.
6. Коломойцева, И.А. Применение тезауруса для расширения запроса информационно-поисковой системы на примере предметной области «Программирование» / И.А. Коломойцева // Материалы XIII Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование» (ИУСМКМ-2022). – Донецк: ДОННТУ, 2022. – С. 155-159.
7. Атаева, О. М. Формирование расширенных поисковых запросов на основе тезауруса предметной области в онтологии знаний семантической библиотеки / О. М. Атаева, В. А. Серебряков, Н. П. Тучкова // Электронные библиотеки. – 2020. – Т. 23, № 3. – С. 271-291.
8. Рогаленко, Н. А. Расширение поисковых запросов с использованием семантической сети с учетом контекста поиска / Н. А. Рогаленко // XII Конгресс молодых ученых : сборник научных трудов, Санкт-Петербург, 03–06 апреля 2023 года. – Санкт-Петербург: Национальный исследовательский университет ИТМО, 2023. – С. 373-377.
9. Гончаров, М. В. Применение тезаурусов при обработке поисковых запросов: от локального использования - к связанным данным / М. В. Гончаров, К. А. Колосов, Е. Ф. Бычкова // Научные и технические библиотеки. – 2022. – № 12. – С. 85-103.
10. Анализ тематических запросов удалённых пользователей Белорусской сельскохозяйственной библиотеки (с использованием дескрипторов тезауруса AGROVOC) / Р. А. Муравицкая, Н. С. Шакура, Е. В. Аксютю, В. В. Слемнева // Библиотечно-информационный дискурс. – 2023. – Т. 3, № 1. – С. 45-51.

Коломойцева И.А., Бердюкова С.С. Особенности определения отношений в тезаурусе в области программирования. В статье выполнен анализ существующих модели информационного поиска. Рассмотрены особенности формирования тезауруса в области программирования. Обоснована необходимость использования такого тезауруса в информационном поиске. Описаны структура информационно-поисковой системы с использованием тезауруса и особенности отношений между понятиями в тезаурусе («ниже-выше», «часть-целое», ассоциации), позволяющие его отнести к формальным онтологиям.

Ключевые слова: обработка естественного языка, информационный поиск, тезаурус, онтология, отношения, программирование.

Kolomoitseva I.A., Berdiukova S.S. Specifics of defining relationships in the thesaurus in the sphere of programming. The article analyzes the existing models of information retrieval. The features of the thesaurus formation in the field of programming are considered. The necessity of using such a thesaurus in information search is substantiated. The structure of an information retrieval system using a thesaurus and the features of the relationship between concepts in the thesaurus ("below-above", "part-whole", associations) are described, allowing it to be attributed to formal ontologies.

Keywords: natural language processing, information retrieval, thesaurus, ontology, relationships, programming.

Статья поступила в редакцию 24.06.2024
Рекомендована к публикации профессором Федяевым О. И.