Синтаксические конструкты поисковых запросов Sourcegraph

Багаль Илья Геннадьевич

УО «Брестский государственный университет имени А.С. Пушкина»

Sourcegraph – специализированная поисковая система, для индексации исходных текстов, находящихся в открытом доступе. Sourcegraph предоставляет удобный интерфейс для поиска необходимого кода. В основе работы Sourcegraph лежат алгоритмы поиска по строго заданным последовательностям. Чем более размыто сформулирован текст, тем менее релевантную выдачу получает пользователь. Для уточнения в интернете используется язык запросов. Каждая ПС разработала свои правила [1].

В основе принципа поисковой работы Sourcegraph лежит семантическое индексирование кода. Поисковые запросы могут состоять только из слов, можно увидеть результаты, в которых эти слова отображаются по порядку, во всех файлах во всех хранилищах. Многие запросы также будут использовать ключевые слова. Ключевые слова помогают фильтровать результаты поиска, определять тип поиска и многое другое.

Пример ключевых слов для поиска кода:

Keywords (all searches)

The following keywords can be used on all searches:

1. *regexp-pattern* – plain words are actually interpreted as regular expressions (using the standard RE2 syntax). Multiple words are joined with \s* to construct the combined pattern – (open\|close)file
2. *“any string”* – surround a string in double quotes to find exact matches (including whitespace and punctuation). Use the \" and \\ escapes if needed – "system error 123"
3. *repogroup:group-name* – only include results from the named group of repositories (defined by the server admin). Same as using a repo: keyword that matches all of the group’s repositories. Use *repo:* unless you know that the group exists – *repogroup:backend* [2].

Sourcegraph поддерживает 19 языков и может работать с GitHub, Bitbucket и Phabricator. Сервис Sourcegraph имеет возможность подключения внешних серверных обработчиков, поддерживающих протокол LSP (Language Server Protocol). Функция подключения внешних обработчиков используется для разбора семантики языка, статического анализа, а также последующего заключения. В персонал серверной части входят:

Сервисы для обеспечения работы фронтэнда (web-интерфейса);

Прокси (в видах интеграции с GitHub);

Git-сервер для зеркалирования репозиториев (в своём хосте);

Индексатор для построения поискового индекса «получи и распишись» на основании содержимого репозитория с учётом семантики на разных языках; [3].

В синтаксических конструкторах сервиса Sourcegraph существуют поисковые запросы, задачи которых могут быть похожими или совпадать с задачами запросов Google и Yandex:

1. Исключение нежелательного в конечном результате: -lang: и «-»
2. Для поиска точных совпадений: "system error 123"
3. Поиск на выбор из вариантов: «\|” (using the standard RE2 syntax) и «|” [2].

Особенности сервиса Sourcegraph заключаются в том, что выполняется полнотекстовый поиск по исходному коду и поддерживаются как регулярные выражения, так и точные запросы. Sourcegraph выполняет поиск во всех репозиториях автоматически. Синтаксис поисковых запросов Sourcegraph позволяет выполнять сложные запросы. Такие запросы, как поиск по любой ветви или фиксации, сужение поиска по языку программирования или шаблону файла и др.

Поскольку все поисковые методы должны быть определены вместе в одном модуле, они объединяются в «кейсы». Такой кейс в Sourcegraph – это анализ группирования альтернативных модулей в целом поисковом разделе. При необходимости обработки целого кейса, структуры данных как единого целого, запросы Sourcegraph могут подразделяться на три типа:

1. Кейс типа (включает объявления по умолчанию);
2. Тип данных (включает конструкторы и функции выбора);
3. Пример класса для определенного типа данных.

В сервисе Sourcegraph присутствуют нюансы:

1) Слишком большое количество репозиториев.

Steps to reproduce:
1. *Go to sourcegraph.com.*
2. *Enter "pubsub".*
Expected behavior:
Search results come back with some relevant repos that mention pubsub.
Actual behavior:
It returns "Too many matching repositories", and four links to suggested refinements that all yield further empty search results [4].

2) Неправильный синтаксис текущего запроса.

Нахождение repos, у которых есть файл "*baz*", который содержит "*foo*", а также содержит файл "*bar*".

Пример кода: *file:baz foo*

Этот запрос не может быть выражен в синтаксисе текущего запроса, который был разработан в основном для поиска совпадений в тексте/регулярном выражении в коде [5].

Стоит отметить, что, несмотря на незначительные нюансы, в 2018 году продукт Sourcegraph был выложен на Github по открытой лицензии Apache. Сообщество оценило высокую скорость работы продукта и отметило, что релиз может повлечь за собой важные изменения в индустрии. SourceGraph представляет другой подход к анализу программного обеспечения. Это полезно для программистов, интересующихся тем, как взаимодействуют объекты в их кодовой базе. Охват SourceGraph может выходить за рамки любого конкретного редактора или интерфейса. В то же время формат отчета для передачи результатов анализа пользователю может быть дополнен интерактивным интерфейсом. Подобное дополнение используется для упрощения манипулирования и понимания визуализаций граф. Набор анализов, которые SourceGraph применяет к графам вызовов, в настоящее время довольно ограничен. Существует несколько алгоритмов анализа, которые планируется внедрить в ближайшем будущем, а также рассматриваются альтернативные формы визуализации графов вызовов, чтобы их было легче понять [6]. Общая концепция может быть распространена на различные языки. Возможность визуализации графов вызовов можно добавить в самом SourceGraph благодаря таким проектам, как библиотека Benedikt Huber's Language.C [7], которая предоставляет функциональность, аналогичную haskell-src-exts.

Список использованных источников

1. Miljenovic I. L : the SourceGraph Program / – School of Mathematics and Physics. The University of Queensland, Queensland, Australia. – Mode of access: https://ivanmiljenovic.files.wordpress.com/2010/03/sourcegraph_pepm10_reprint.pdf

2. Sourcegraph [Electronic resource]. – Mode of access: https://sourcegraph.com/help/user/search/queries. – Date of access: 07.02.2019

3. OpenNET [Electronic resource]. – Mode of access: https://www.opennet.ru/opennews/art.shtml?num=49382. – Date of access: 10.02.2019

4. Jason Poovey : A Performance Evaluation of Open Source Graph Databases / – Georgia Institute of Technology. – Mode of access: http://www.stingergraph.com/data/uploads/papers/ppaa2014.pdf

5. GitHub [Electronic resource]. – Mode of access: https://github.com/sourcegraph/sourcegraph/issues/127. – Date of access: 08.02.2019

6. Gansner E. R and North S. C. An Open Graph Visualization System and Its Applications. – Mode of access: http://graphviz.org.

7. Ediger D, McColl R, Riedy J, and D. A. Bader. High performance data structure for streaming graphs: High Performance Extreme Computing Conference (HPEC), – Sept. 2012. – Mode of access: https://www.cc.gatech.edu/~bader/papers/STINGER.html