Алгоритм подбора проектных команд ИТ-специалистов на основе данных проектных репозиториев

Н.Г. Ярушкина[1], А.С. Желепов[1]

[1]Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027

Аннотация. Из-за нехватки квалифицированных кадров в ИТ-сфере компании предоставляют своим сотрудникам возможность работать удаленно, чтобы искать новых специалистов на глобальном кадровом рынке. Компании, занимающиеся разработкой продуктовых решений, заинтересованы в поиске «сыгранных» проектных команд специалистов, которые в течение длительного времени уже успешно работали вместе. Поиск такой команды требует внесения корректив в HR-процессы: появляется необходимость анализировать деятельность, разработки не отдельных сотрудников, а команды разработчиков в целом. В данной статье описывается прототип системы, реализующей поиск и подбор проектных команд на основе данных открытых репозиториев исходного кода и связанных артефактов. Особое внимание в статье уделяется алгоритму выбора основных членов проектной команды из множества всех разработчиков, принимавших участие в проекте.

1. Введение

На фоне глобального роста сфера информационных технологий в России испытывает недостаток квалифицированных кадров [2]. Исследование сообщества компаний-разработчиков программного обеспечения «РУССОФТ» показало, что проблема особенно актуальна для региональных ИТ-компаний [3]. Из-за нехватки квалифицированных кадров компании меняют модель работы, предоставляя сотрудникам возможность трудиться удаленно, не находясь в офисе. Такой переход открывает возможность найма сотрудников компанией не только на локальном, но и на глобальном кадровом рынке.

Для продуктовых компаний наблюдается тенденция поиска не только отдельных специалистов, но уже «сыгранных» команд разработчиков [1, 4, 5]. Понятие «сыгранная» подразумевает, что группа специалистов имеет опыт совместной работы над проектами, соответственно процессы разработки уже выстроены, а взаимоотношения между членами команды налажены. Поиск «готовых» команд компаниями обуславливается особенностями современной продуктовой разработки: необходимость быстрой проверки гипотез и прототипирования, разработка MVP и т.д.

В данной статье приводится источник данных для организации глобального поиска проектных команд, базовые компоненты архитектуры разрабатываемой системы, алгоритм выбора основных членов проектной команды и примеры работы алгоритма.

2. Архитектура системы поиска проектных команд

Источником данных для тестирования прототипа системы поиска проектных команд и разрабатываемых алгоритмов был выбран сервис GitHub – «социальная сеть» разработчиков программного обеспечения, в которой зарегистрировано более 37 миллионов ИТ-специалистов и размещено более 100 миллионов проектных репозиториев. Взаимодействие разработанной системы и GitHub построено на основе открытого API сервиса.

На рисунке 1 представлена архитектура системы поиска проектных команд.

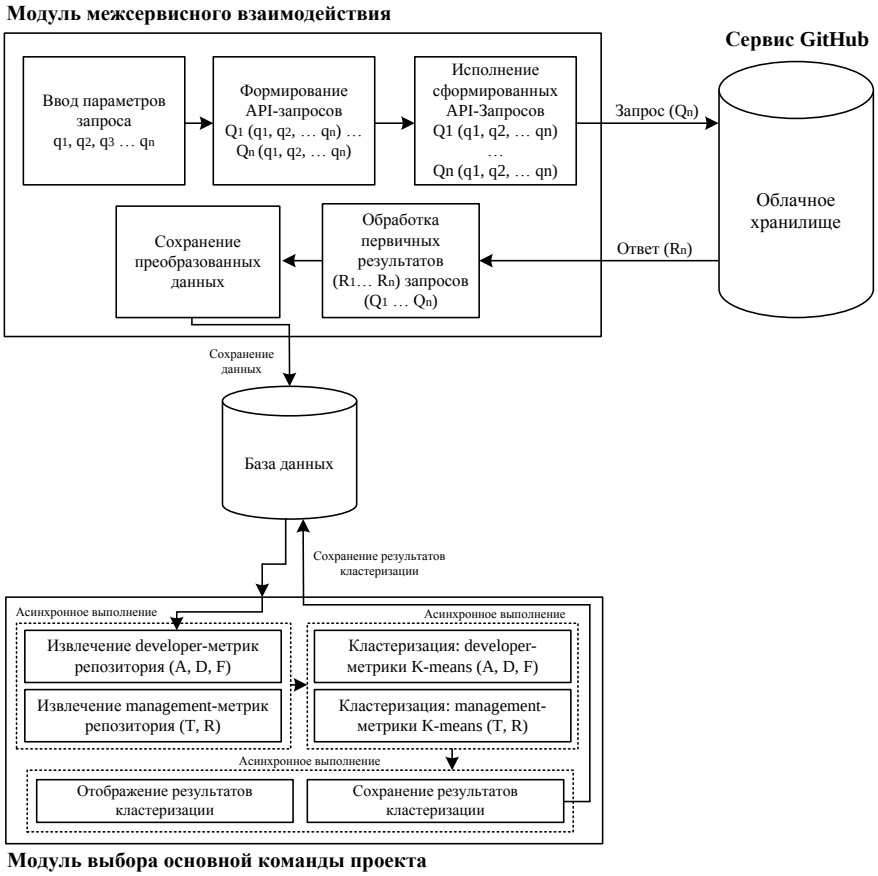

Рисунок 1. Архитектура системы поиска проектных команд.

Представленная система состоит из 2 модулей:

- модуль формирования запроса $Q\ (q_1, q_2, \dots q_n)$ обеспечивает обмен данными между сервисом GitHub и системой поиска проектных команд. $q_1, q_2 \dots q_n$ − параметры запроса: используемые технологии, число участников проекта и т.д.;
- модуль поиска основной команды проекта. GitHub-репозитории специфичны из-за того, что находятся в открытом доступе. Это дает возможность участия в проекте сообществу разработчиков, не входящих в основную команду проекта. Функция данного модуля – фильтрация и выборка членов основной команды.

3. Алгоритм выбора проектной команды

Алгоритм выбора предназначен для поиска основной группы разработчиков проектной команды на основе данных каждого из найденных проектных репозиториев. В основе работы алгоритма применяется метод кластеризации k-means [7], который позволяет задать искомое количество кластеров.

Основываясь на исследовании [6] предполагалось реализовать кластерный анализ с учетом трех кластеров: J (Junior), M (Middle), S (Senior), обозначающих соответствующие уровни квалификации специалистов.

Данные виды кластеров не могут быть выбраны при исследовании команды на основе количественных данных проектного репозитория. Например, senior-специалисты не всегда вносят в проект много изменений, а больше координируют действия команды, собирают стабильные версии программного обеспечения (релизы). Поэтому анализ на основе количественных метрик при таком выборе кластеров может дать серьезное отклонение.

С учетом особенностей были выбраны следующие кластеры:

- контрибьютор (К). Участник, вносящий относительно небольшой вклад в проект;
- участник (У). Специалист, который периодически принимает участие в улучшении проекта;
- предполагаемый член команды проекта (ПЧКП). Разработчик, активно развивающий проект;
- основной разработчик (ОР). Специалист, вносящий наибольший вклад в развитие проекта.

Процедура кластеризации проходит в два этапа: кластеризация на основе developer-метрик и management-метрик [8]. Разделение позволяет учесть то, что опытные специалисты могут быть больше менеджерами проекта, чем непосредственно исполнителями.

В качестве developer-характеристик были выбраны:

- Additions and Deletions (A, D). Метрика описывает количество строк кода, которые были добавлены и удалены каждым из разработчиков проекта;
- Changed Files (CF). Количество файлов, измененных разработчиком за время работы над проектом.

В качестве management-метрик были выбраны:

- Task Count (TC). Количество созданных задач, метрика характеризует умение правильно понимать и определять вектор развития проекта;
- Release Count (RC). Количество выпущенных релизов разработчиком. Характеристика описывает умение «сводить» все предложенные другими разработчиками изменения, быть ответственным за очередную версию продукта с точки зрения его корректной работы.

4. Применение разработанного алгоритма

Для тестирования алгоритма выбора основных участников команды были выбраны 10 проектных репозиториев, среди которых такие известные проекты, как:

- ClickHouse. Колоночная СУБД, разрабатываемая компанией Яндекс. Проект насчитывает в своей истории более 30 000 коммитов, 350 участников на GitHub, в том числе и основную команду проекта. Особенность: основная команда проекта не распределена.
- Yii 2 Framework. Популярный среди веб-программистов PHP-фреймворк. Количество коммитов: около 20 000, число участников: 950. Особенность: основная команда проекта распределена по всему миру.
- Albumentations и Catalyst. Современные фреймворки, которыми пользуются инженеры по машинному обучению. Проекты появились относительно недавно. Albumentations активно разрабатывается при участии компании X5 – основная команда не распределена, Catalyst – инициативный проект нескольких разработчиков.
- 6 проектов международной группы разработчиков Evil Martians: PostCss, BrowsersList, AutoPrefixer, NanoId, Gon, ImgProxy. Проекты широко применяются разработчиками программного обеспечения при создании веб-приложений.

Выбор репозиториев обусловлен тем, что их разработчики принимали участие в международной ИТ-конференции «Стачка», соорганизатором которой является соавтор данной статьи. Благодаря этому стало возможным практически проверить работу алгоритма на предмет совпадения участников проектной команды, выделенных алгоритмом, и реальных ее членов.

На основе developer- и management-метрик каждого из выбранных репозиториев был проведен кластерный анализ участников проектов, результаты которого представлены в таблице 1.

Таблица 1. Результаты кластерного анализа участников выбранных проектных репозиториев.

№	Наименование репозитория	Число участников	Распределение участников в результате анализа на основе developer-метрик	Распределение участников в результате анализа на основе management-метрик
1	Albumentations	46	К: 26 У: 16 ПЧКП: 2 ОР: 2	К: 1 У: 1 ПЧКП: 2 ОР: 0
2	PostCss	288	К: 208 У: 77 ПЧКП: 2 ОР: 1	К: 1 У: 0 ПЧКП: 0 ОР: 0
3	Yii2	983	К: 875 У: 101 ПЧКП: 4 ОР: 2	К: 1 У: 1 ПЧКП: 1 ОР: 1
4	ClickHouse	350	К: 311 У: 32 ПЧКП: 6 ОР: 1	К: 14 У: 8 ПЧКП: 1 ОР: 0
5	Catalyst	36	К: 27 У: 7 ПЧКП: 1 ОР: 1	К: 1 У: 1 ПЧКП: 0 ОР: 0
6	BrowsersList	100	К: 89 У: 8 ПЧКП: 2 ОР: 1	К: 1 У: 1 ПЧКП: 0 ОР: 0
7	AutoPrefixer	155	К: 140 У: 11 ПЧКП: 3 ОР: 1	К: 1 У: 1 ПЧКП: 0 ОР: 0
8	NanoId	55	К: 41 У: 13 ПЧКП: 1 ОР: 0	К: 1 У: 0 ПЧКП: 0 ОР: 0
9	Gon	59	К: 52 У: 4 ПЧКП: 2 ОР: 1	К: 1 У: 0 ПЧКП: 0 ОР: 0
10	Imgproxy	31	К: 22 У: 7 ПЧКП: 1 ОР: 1	К: 1 У: 0 ПЧКП: 0 ОР: 0

Вывод на основе результатов, представленных в таблице 1: распределение участников проекта по кластерам в результате анализа management-метрик значительно меньше. Это объясняется тем, что лишь небольшое количество специалистов отвечает за составление задач и подготовку стабильных версий программного обеспечения, в случае GitHub – это наиболее заинтересованные в развитии проекта разработчики или члены основной команды.

Практическая значимость работы алгоритма подтверждена участниками проектных команд ClickHouse, Yii 2, Catalyst, Albumentations и проектов Evil Martians. Результаты пересечения множеств участников (P1), выделенных алгоритмом, и множества действительных участников (P2), определенных экспертами, приведены в таблицах 2-3.

Разделение проектов по таблицам 2 и 3 неслучайно: таблица 2 содержит проекты, в которые большой вклад вносит несколько разработчиков в виду их сложности и широкого использования, таблица 3 – программные библиотеки, которые скорее упрощают процесс

разработки и не являются полноценными фреймворками или СУБД. В дальнейшем для учета социальной составляющей работы над проектом планируется добавить параметр: soft skills. Его численная метрика будет выражена в виде количества сообщений и связанных подсообщений обсуждений предлагаемых решений задач каждым из участников проекта.

Таблица 2. Сравнение множеств P1 и P2 участников проектов ClickHouse, Yii 2, Catalyst, Albumentations.

	ClickHouse	Yii 2	Catalyst	Albumentations
Множество участников P1 (учитываются только кластеры «ПЧКМ» и «ОР»)	BayoNet, Ivan Blinkov, Nikolai Kochetov, Vitaliy Zakaznikov, alexey-milovidov, chenxing-x, proller	samdark, NabiKAZ, SilverFire, 卄巡洋艦㊣, Carsten Brandt, Qiangxue	Tezikov Roman, Sergey Kolesnikov	Alexander Buslaev, Eugene Khvedchenya, Alex Parinov, Vladimir Iglovikov
Множество участников P2 (экспертная оценка)	alesapin, bayoNet, blinkov, 4ertus2, KochetovNicolai den-crane, proller, alexey-milovidov	Qiangxue, samdark, SilverFire, cebe	Tezikov Roman, Sergey Kolesnikov	Alexander Buslaev, Eugene Khvedchenya, Alex Parinov, Vladimir Iglovikov
Результат пересечения множеств $P_1 \cap P_2$	5 из 8	3 из 4	2 из 2	4 из 4

Таблица 3. Сравнение множеств P1 и P2 участников проектов PostCss, BrowsersList, AutoPrefixer, NanoId, Gon, Imgproxy.

	PostCss	BrowsersList	AutoPrefixer	NanoId	Gon	Imgproxy
Множество участников P1 (учитываются только кластеры «ПЧКМ» и «ОР»)	ai, ben-eb, jedmao	ai, AleshaOleg, akx, An-Tu	ai, bogdan0083, yepninja, Semigradsky	ai	gazay, torbjon, john-bai	DarthSim, koenpunt
Множество участников P2 (экспертная оценка)	ai	ai, akx	ai, Semigradsky	ai	gazay	DarthSim
Результат пересечения множеств $P_1 \cap P_2$	1 из 3	2 из 4	2 из 4	1 из 1	1 из 3	1 из 2

На основании экспертной оценки авторов проектов некоторые из действительных членов проектной команды не попали в кластеры «ПЧПК» и «ОР» в виду того, что вносят недостаточное количество изменений в проект. Это связано с рядом причин: участие в других проектах компании, выполнение функций поддержки продукта.

В ходе дальнейшего исследования алгоритм планируется улучшить, добавив дополнительный параметр для кластерного анализа: количество закрытых задач участником проекта.

5. Литература

[1] Катаев, А. Управление большой и распределенной командой (выступление на конференции Teamlead Conf 2018) [Электронный ресурс]. – Режим доступа: http://teamleadconf.ru/2018/abstracts/3205 (22.11.2019).

[2] Ярушкина, Н.Г. Исследование ИТ-кластера Ульяновской области / Н.Г. Ярушкина, Т.В. Афанасьева, О.В. Шиняева – Ульяновск: УлГТУ, 2013. – 137 с.

[3] Исследование РУССОФТ. Российской ИТ-индустрии предрекают острый кадровый голод [Электронный ресурс]. – Режим доступа: https://www.it-world.ru/it-news/it/140881.html (22.11.2019).
[4] Yarushkina, N. An Approach to Similar Software Projects Searching and Architecture Analysis Based on Artifcial Intelligence Methods / N. Yarushkina, G. Guskov, P. Dudarin, V. Stuchebnikov // Proceedings of the Third International Scientifc Conference Intelligent Information Technologies for Industry – Advances in Intelligent Systems and Computing. – 2018. – Vol. 1. – P. 341-352.
[5] Evil Martians [Электронный ресурс]. – Режим доступа: https://evilmartians.com/ (22.11.2019).
[6] Afanasyeva, T. Framework for Accessing Professional Growth of Software Developers / T. Afanasyeva, A. Zhelepov, I. Zagaichuk // 5th International Conference on Computer and Technology Applications, 2019.
[7] Li, Y. A Clustering Method Based on K-Means Algorithm / Y. Li, H. Wu // International Conference on Solid Devices and Materials Science, 2012.
[8] Kaur, K. Static and Dynamic Complexity Analysis of Software Metrics / K. Kaur, K. Minhas, N. Mehan, N. Kakkar // World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering. – 2009. – Vol. 3(8). – P. 1936-1938.

IT specialists search algorithm based on repositories

N.G. Yarushkina[1], A.S. Zhelepov[1]

[1]Ulyanovsk State Technical University, Severniy Venec street 32, Ulyanovsk, Russia, 432027

Abstract. Due to the lack of qualified personnel in the IT sector, companies provide their employees with the opportunity to work remotely. That opens an opportunity to seek new specialists in the global personnel market. Companies produced products are interested in finding already «played» project teams of specialists who have successfully worked together for a long time. The search for such a team requires making adjustments to HR processes: it becomes necessary to analyze the activities and development of not individual employees, but the development team as a whole. This article describes a prototype system that implements the search and selection of project teams based on data from open source code repositories and related artifacts. Particular attention is paid to the algorithm for selecting the main members of the project team from the set of all developers who took part in the project