УДК 336.66 DOI: 10.14451/1.240.144

Применение методов контент-ориентированной фильтрации и больших языковых моделей для улучшения системы мэтчинга резюме и вакансий

© 2024 **Ильин Матвей Игоревич**
Магистрант. Ведущий специалист по моделям предиктивной аналитики клиентских сервисов. Новосибирский национальный исследовательский государственный университет. ПАО «Группа Ренессанс Страхование».
E-mail: m.ilin@g.nsu.ru

© 2024 **Писаренко Михаил Максимович**
Студент. Новосибирский национальный исследовательский государственный университет.
E-mail: m.pisarenko@g.nsu.ru

Ключевые слова: рекомендательные системы, языковые модели, анализ резюме, рекрутинг, HR процессы, мэтчинг резюме, текстовая информация, большие данные.

В современном мире технологии быстро развиваются, а рынок труда становится все более динамичным. Эффективность процессов найма и отбора стала решающим элементом успеха компаний. Организации стремятся усовершенствовать свои HR-процессы, чтобы быстро и точно находить подходящих кандидатов. В этой ситуации значительную роль в сопоставлении резюме с вакансиями играют автоматизированные системы, что позволяет организациям значительно ускорить процесс поиска подходящих кандидатов. Традиционные методы подбора кандидатов, основанные на ручном анализе резюме являются неэффективными, так как они ограничены в масштабируемости и требуют высокой трудоемкости. Недостаточная эффективность традиционных методов подбора кандидатов стала особенно заметна в поседение годы. Это связано с тем, что все больше компаний переходят на удаленный формат работы, из-за чего в разы расширяется список потенциальных претендентов на вакансию. Рекомендательные системы фильтрации и мэтчинга, которые уже много лет применяются в различных сферах, таких как видеохостинги, аудио стриминговые сервисы, маркетплейсы, сложно реализовать в рамках рекрутинговой индустрии, из-за чего большинство российских агрегаторов вакансий используют обычный поиск по совпадающим словам. Такая особенностей связана с высокой степенью неоднородности представленной информации в рекрутинговой индустрии. Эффективно извлечь ключевую характеристику из записи резюме или вакансии, применяя только традиционные методики, практически невозможно. Особенно перспективным направлением является использование больших языковых моделей. Такие модели способны анализировать контекстуальные и семантические взаимосвязи даже в неструктурированной текстовой информации, что потенциально может позволить приводить

записи резюме или вакансий к общему, пригодному для автоматического анализа виду. Целью данной работы является исследование применимости совместного использования больших языковых моделей и техник контент-ориентированной фильтрации для создания системы мэтчинга резюме и вакансий в рекрутинговой индустрии. Предметом исследования являются модели фильтрации в рекомендательных системах и большие языковые модели. Объектом исследования — текстовые представления записей резюме и вакансий в рекрутинговой индустрии.

Рекомендательная система — это подтип системы фильтрации информации, который предоставляет предложения по элементам, наиболее релевантным для конкретного пользователя. Рекомендательные системы особенно полезны, когда пользователю необходимо выбрать элемент из потенциально огромного числа предложений, которые может предоставить сервис.

Любая рекомендательная система состоит из 4 основных компонентов, а именно:

– Сбор данных.

Рекомендательная система собирает данные о пользователях и контенте из различных источников. Эти данные могут быть явными (например, рейтинги, лайки, комментарии) и неявными (например, история просмотров, время на странице, клики). Система также может использовать демографические данные пользователей, такие как возраст, пол, местоположение, для создания более точных профилей.

– Обработка данных.

Так как собранная информация может хранится в любом из трех представлений, текстов, визуальном и аудио, ее необходимы перевести в пригодный для обработки формат. Как правило, для анализа используется модели машинного обучения и глубокого обучения, такие как векторные представления слов (Word Embeddings) для текстовой информации, свёрточные нейронные сети (CNN) для изображений и рекуррентные нейронные сети (RNN) для аудио.

– Рекомендательный алгоритм.

Существует три основных подхода к реализации рекомендательной фильтрации: контент-ориентированная фильтрация, коллаборативная фильтрация и гибридный метод. Так как первые два подхода не накладывают никаких ограничений на их совместное использование, как правило, применяется гибридный метод, который может использовать различные подходы комбинирования первых двух методик на разных этапах рекомендации. Чаще всего, в качестве гибридного подхода применяются взвешенные оценки, полученные после контент-ориентированной и коллаборативной фильтрации.

– Оценка степени релевантности.

На основе профилей пользователей и анализа контента система формирует список рекомендаций. Этот процесс может включать ранжирование рекомендаций по степени релевантности и вероятности интереса пользователя.

Такие рекомендательные системы широко и эффективно используются в различных отраслях, например:

- Электронная коммерция. Персонализированные рекомендации товаров и услуг позволяют увеличивать продажи и улучшать пользовательский опыт. Примеры включают системы рекомендаций на Amazon, которые предлагают товары на основе предыдущих покупок и просмотров пользователя.
- Стриминговые сервисы. Рекомендации фильмов, сериалов, музыки на основе предпочтений пользователя помогают увеличить время, проведенное на платформе, и улучшить удовлетворенность пользователя. Примером может служить Netflix, который использует алгоритмы для персонализации контента.

- Социальные сети. Рекомендации друзей, групп, контента на основе интересов и поведения пользователя способствуют увеличению вовлеченности и активности на платформе. VK, например, использует сложные алгоритмы для предложения друзей и групп, которые могут быть интересны пользователю.
- Образовательные платформы. Рекомендации курсов, лекций, учебных материалов в соответствии с учебными потребностями и прогрессом студента помогают улучшить результаты обучения и удовлетворенность образовательным процессом. Примером является Coursera, которая предлагает курсы на основе предыдущих занятий и интересов пользователя.

Коллаборативная фильтрация (CF) — один из методов, используемых в рекомендательных системах. Коллаборативная фильтрация имеет два смысла: узкий и более общий.

В современном контексте коллаборативная фильтрация представляет собой метод автоматического прогнозирования (фильтрации) интересов пользователя на основе анализа предпочтений и вкусов большого количества людей (коллаборации). Основное предположение подхода коллаборативной фильтрации заключается в том, что если два человека имеют одинаковое мнение по какому-либо вопросу, то, вероятнее всего, они будут иметь схожие мнения и по другим вопросам. Например, система рекомендаций на основе коллаборативной фильтрации для музыкальных произведений может предсказывать, какая музыка понравится пользователю, исходя из списка прослушиваний другого, похожего на него, пользователя. Эти предсказания специфичны для конкретного пользователя, но используют обобщенные данные от множества других пользователей. Это отличается от более простого подхода, где каждому элементу присваивается усредненная оценка, например, основанная на общем количестве голосов.

В более общем смысле, коллаборативная фильтрация — это процесс фильтрации информации или паттернов с использованием методов кооперации множества источников данных, агентов и точек зрения и т.д. Применение методов коллаборативной фильтрации обычно включает очень большие наборы данных. Обычно эти методы применяются к очень большим наборам данных и находят использование в различных сферах. Например, в экологическом мониторинге, где данные собираются с большого количества сенсоров на больших территориях, или в финансовой сфере, где интегрируются данные из множества финансовых источников. Метод коллаборативной фильтрации наиболее популярен в электронной коммерции, где на основе интересов групп пользователей генерируются рекомендации товаров и услуг.

Такой подход обладает как преимуществами, так и недостатками. Как преимущества можно выделить следующие пункты:

– Открытие новых предметов.

В отличие от контент-ориентированной фильтрации, коллаборативная фильтрация основывается на предпочтениях пользователей, а не на характеристиках контента, из-за чего способна делать неожиданные, но релевантные рекомендации, расширяя будущий ассортимент.

– Независимость от контентной информации.

Так как данный подход не требует информации о характеристиках контента, он особенно полезен в случаях, когда нельзя точно оценить характеристики информации или эффективно преобразовать информацию в пригодный для расчетов формат.

– Контролируемая масштабируемость.

Так как мы способны сами выбирать информацию о пользователе, которая будет доставаться и обрабатываться, можно контролировать объем вычислительных мощностей, необходимых для работы алгоритма.

Как недостатки данного подхода можно выделить:

– Проблема холодного старта.

CF-метод практически не имеет эффективности в случае недостатка достаточного количества пользователей, так как нельзя выделить схожие группы по интересам.

– Проблема популярности.

Часто возникает проблема, когда у большой части пользователей есть история взаимодействия с небольшой частью контента, из-за чего система не предлагает нишевый, но потенциально не менее релевантный контент, а также не способная рекомендовать появившийся недавно контент

– Проблема неординарных пользователей.

Существуют пользователи, интересы которых сильно отличаются от группы пользователей с схожими параметров, из-за чего они получают нерелевантные для них рекомендации.

– Правовые ограничения.

Ограничения на получение и обработку персональных данных сужают объем доступной для анализа информации и пользователе, тем самым снижая эффективность такого метода фильтрации.

В основном, для реализации коллаборативной фильтрации используется матричная векторизация или подходы, основанные на модели k-ближайших соседей. Матричная факторизация является одной из самых популярных техник в коллаборативной фильтрации. В основе этого метода лежит представление предпочтений пользователей и характеристик предметов в виде матриц.

Пусть

R – матрица оценок размером $m \times n$, где

m – количество пользователей,

n – количество предметов.

Элемент r_{ij} представляет оценку пользователя i предмета j.

Модель факторизации матриц.

Мы хотим представить матрицу R в виде произведения двух матриц меньшего ранга: матрицы пользователей U (размером $m \times k$) и матрицы предметов V (размером $k \times n$). Здесь k – количество скрытых факторов.

Формально, $R \approx U \cdot v$, Элемент r_{ij} аппроксимируется как $r_{ij} \approx u_i \cdot v_j$, где u_i – вектор скрытых факторов для пользователя i, а v_j – вектор скрытых факторов для предмета j.

Оптимизация.

Задача сводится к минимизации функции ошибки, которая измеряет разницу между реальными и предсказанными оценками:

$$\min_{U,V} \sum_{(i,j)\in\mathcal{K}} (r_{ij} - u_i \cdot v_j)^2 + \lambda(|U|^2 + |V|^2). \qquad (1)$$

Метод k-ближайших соседей основывается на поиске схожих пользователей, путем оценки степени их схожести [5].

Для каждого пользователя i ищется k наиболее похожих пользователей на основе меры сходства, такой как косинусное сходство:

$$\cos(u_i, u_j) = \frac{u_i \cdot u_j}{|u_i||u_j|}. \qquad (2)$$

Оценка для предмета j предсказывается как взвешенная сумма оценок k ближайших соседей:

$$\hat{r}_{ij} = \frac{\sum_{u\in\mathcal{N}_k(i)} \cos(u_i, u) \cdot r_{uj}}{\sum_{u\in\mathcal{N}_k(i)} |\cos(u_i, u)|}, \qquad (3)$$

где $\mathcal{N}_k(i)$ – множество k-ближайших соседей пользователя i.

Контент-ориентированный метод фильтрации (CBF) является одной из популярных техник, используемых в рекомендационных системах. В отличие от коллаборативной фильтрации, которая опирается на предпочтения других пользователей, контент-ориентированный метод основывается на характеристиках предметов и интересах конкретного пользователя. Контент-ориентированная фильтрация использует характеристики предметов для создания профилей

пользователей и предметов. Затем эти профили используются для предсказания, какие предметы могут заинтересовать пользователя на основе их сходства с ранее понравившимися предметами.

Плюсами такого подхода являются:

- Устойчивость к холодному старту. Так как в данной модели фильтрации отдельно рассматривается каждый пользователь, она хорошо работает в новых сервисах, без большой базы пользователей.
- Высокая степень персонализации. Каждый человек получает наиболее актуальные лично для него предложения.

Как недостатки можно выделить:

- Отсутствие новизны в рекомендациях. Данный подход способен рекомендовать только похожий на ранее просмотренный контент.
- Высокая скорость роста сложности алгоритма.

Характеристики каждого просмотренного пользователем предмета должны учитываться для дальнейшей фильтрации, из-за чего сложность алгоритма растет со временем, это особенно актуально для отраслей с неоднородным контентом/товарами, вроде маркетплейсов.

Для данной модели фильтрации плохо применим метод факторизации матриц, но так же, как и для коллаборативной фильтрации, используется метод k-ближайших соседей, основанный на мере косинусного сходства между интересами пользователя (историей его прошлых просмотров и оценок) и характеристикой нового предмета.

В отличие от других отраслей, вроде маркетплейсов, видеохостингов, аудио стриминговых сервисов, в рекрутинге каждый пользователь одновременно является контентом для другого пользователя. Если рассматривать компанию в роли пользователя сайта-агрегатора вакансий, то для него соискатели являются контентом. И, наоборот, одновременно с этим информация о вакансии и о резюме является практически однородной, что позволяет сразу рассчитывать косинусное сходство напрямую между ними, хоть такой метод не подходит под оригинальное определение коллаборативной или контент-ориентированной фильтрации, в дальнейших главах работы будет использоваться термин «контент-ориентированная фильтрация» в качестве обозначения используемого подхода, так как будет рассчитываться релевантность кандидата, исходя из описания его резюме. Потенциально данный факт может значительно повысить эффективность фильтрации, так как изначально полностью известны интересы пользователя и нет необходимости сбора достаточной истории его взаимодействия с контентом, в описании любой вакансии есть вся необходимая информация для поиска наиболее релевантного кандидата.

Библиографический список

1. A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods Based on Social Networks / R. Chen [et al.] // IEEE Access. — 2018. — Oct. — Vol. PP. — P. 1–1. — DOI: 10.1109/ACCESS.2018.2877208.
2. A survey on large language model based autonomous agents / L. Wang [et al.] // Frontiers of Computer Science. — 2024. — Mar. — Vol. 18, no. 6. — ISSN 2095-2236. — DOI: 10.1007/s11704-024-40231-1.
3. *Balakrishnan V., Ethel L.-Y.* Stemming and Lemmatization: A Comparison of Retrieval Performances // Lecture Notes on Software Engineering. — 2014. — Jan. — Vol. 2. — P. 262–267. — DOI: 10.7763/LNSE.2014.V2.134.
4. *Chen W.-Y.* Intelligent Tutor: Leveraging ChatGPT and Microsoft Copilot Studio to Deliver a Generative AI Student Support and Feedback System within Teams. — 2024. — DOI: 10.48550/ARXIV.2405.13024.
5. *Cui B.-B.* Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm // ITM Web of Conferences. — 2017. — Jan. — Vol. 12. — P. 04008. — DOI: 10.1051/itmconf/20171204008.
6. Explaining the black-box model: A survey of local interpretation methods for deep neural networks / Y. Liang [et al.] // Neurocomputing. — 2021. — Jan. — Vol. 419. — P. 168–182. — DOI: 10.1016/j.neucom.2020.08.011.

7. *Gohberg I.*, *Kaashoek M.*, *Spitkovsky I.* An Overview of Matrix Factorization Theory and Operator Applications // Operator Theory: Advances and Applications. — 2003. — Jan. — Vol. 141. — DOI: 10.1007/978-3-0348-8003-9_1.
8. *Gómez-Uribe C.*, *Hunt N.* The Netflix Recommender System // ACM Transactions on Management Information Systems. — 2015. — Dec. — Vol. 6. — P. 1–19. — DOI: 10.1145/2843948.
9. *Hassler M.*, *Fliedl G.* Text Preparation through Extended Tokenization // Data Mining VII: Data, Text and Web Mining and their Business Applications. — 2006. — June. — Vol. 37. — DOI: 10.2495/DATA060021.
10. *Knees P.*, *Neidhardt J.*, *Nalis I.* Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness //. — 12/2023. — P. 417–434. — ISBN 978-3-031-45303-8. — DOI: 10.1007/978-3-031-45304-5_27.
11. *Kumari A.*, *Shashi M.* Vectorization of Text Documents for Identifying Unifiable News Articles // International Journal of Advanced Computer Science and Applications. — 2019. — Aug. — Vol. 10. — P. 305. — DOI: 10.14569/IJACSA.2019.0100742.
12. Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities / H. Zhou [et al.] // IEEE Communications Surveys & Tutorials. — 2024. — Jan. — Vol. PP. — P. 1–1. — DOI: 10.1109/COMST.2024.3465447.
13. *Ricci F.*, *Rokach L.*, *Shapira B.* Recommender Systems: Techniques, Applications, and Challenges //. — 11/2021. — P. 1–35. — ISBN 978-1-0716-2196-7. — DOI: 10.1007/978-1-0716-2197-4_1.
14. *Thorat B. P.*, *Goudar R.*, *Barve S.* Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System // International Journal of Computer Applications. — 2015. — Jan. — Vol. 110. — P. 31–36. — DOI: 10.5120/19308-0760.
15. *Vanetik N.*, *Kogan G.* Job Vacancy Ranking with Sentence Embeddings, Keywords, and Named Entities // Information. — 2023. — Aug. — Vol. 14. — P. 468. — DOI: 10.3390/info14080468.