УДК 004.9 **doi: 10.26906/SUNZ.2019.4.110**

Д. В. Шингалов, Є. В. Мелешко, В. В. Босько

Центральноукраїнський національний технічний університет, Кропивницький, Україна

ДОСЛІДЖЕННЯ ВІДКРИТИХ НАБОРІВ ДАНИХ ВЕБ-РЕСУРСІВ У КОНТЕКСТІ ЗАСТОСУВАННЯ ЇХ ДЛЯ ТЕСТУВАННЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ

Предметом вивчення у статті є процес тестування методів побудови рекомендаційних систем на основі відкритих наборів даних у мережі Інтернет. **Метою** є дослідження відкритих наборів даних веб-ресурсів у контексті застосування їх для тестування різних методів побудови рекомендаційних систем. **Завдання:** дослідити сучасні веб-платформи з відкритими наборами даних та можливість застосування їх даних для тестування якості роботи різних рекомендаційних систем. Отримані такі **результати:** Розглянуто найбільш популярні веб-платформи з відкритими наборами різнотипних мережевих даних. Здійснено порівняльний аналіз цих платформ з точки зору наявності вільного доступу до завантаження даних, їх функціональності та територіальної приналежності, формату даних та зручності для подальшого використання для машинного навчання, а також можливості застосування для тестування рекомендаційних систем. Також проведено оцінку актуальності даних, що зберігаються у репозиторіях з вільним доступом та наявності їх оновлення з часом. **Висновки.** Досліджено веб-платформи, що містять відкриті набори даних, які можна використати для тестування рекомендаційних систем. Основними перевагами більшості платформ є підтримка сучасних форматів даних та умовно вільний або вільний доступ. Серед недоліків розглянутих платформ слід зазначити недостатню структурованість деяких наборів даних, зокрема текстових, що значно обмежує їх застосування для тестування методів контентної фільтрації. Окрім того, одним з факторів, що обмежує використання відкритих наборів даних є їхня актуальність, тому що деякі набори, що зберігаються на платформах, є застарілими та не оновлюються. Усі розглянуті набори даних можуть бути застосовані для дослідницьких цілей та тестування роботи рекомендаційних систем.

Ключові слова: рекомендаційні системи, тестування, аналіз даних, відкриті набори даних, цифровий маркетинг.

Вступ

Рекомендаційні системи стають одним з найважливіших інструментів для маркетингу товарів та послуг в мережі Інтернет. Кількість інформації збільшується щодня, що призводить до перевантаження даними користувачів мережі. Визначення вподобань користувачів для формування ним індивідуальних списків рекомендацій стало одним з рішень даної проблеми

Рекомендаційні системи використовуються для фільтрування та сортування даних на основі вподобань користувачів.

Основною метою рекомендаційних систем є формування корисних пропозицій та рекомендацій з інформації, одержаної про користувачів веб-ресурсу.

Приклади сучасних рекомендаційних систем:

рекомендації книг на Amazon,

рекомендації фільмів на Netflix,

стрічка новин у Facebook, тощо.

Існує багато різних методів побудови рекомендаційних систем, їх можна розділити на три великі категорії [1, 2]:

контентна фільтрація,

колаборативна фільтрація,

гібридні методи.

Кожен з методів побудови рекомендаційних систем має свої переваги, недоліки та обмеження, тому в сучасних системах, як правило, використовуються складні гібриди різних методів контентної та колаборативної фільтрації [1], а також можуть додаватися різні специфічні методи [1, 3], такі як соціальна фільтрація, контекстна фільтрація, тощо.

Для розробки, навчання та тестування алгоритмів формування списку рекомендацій користувачам, як і для інших алгоритмів машинного навчання, необхідні набори даних, які відповідають певним вимогам, а також є доступними та змістовними.

На сьогоднішній день проблема полягає не в пошуку наборів даних, як було ще 10 років тому, а у підборі коректних та актуальних даних для досліджень. Хороші набори для здійснення наукових досліджень повинні бути цікавими та нетривіальними, відповідати вимогам дослідження: тематика, повнота тощо.

Метою даної роботи є дослідження існуючих відкритих наборів різнопланових даних для виявлення найбільш придатних з них для використання у рекомендаційних системах.

Поставлена мета реалізується шляхом вирішення наступних задач:

- дослідження характеру, тематики та типу даних у наборах;
- дослідження наявності таких функціональних критеріїв як теги, оцінки та коментарі у наборах даних;
- дослідження територіальної приналежності даних та обмежень доступу для завантаження безпосередньо з джерела.
- дослідження можливості застосування наборів даних з мережі безпосередньо до рекомендаційних систем.

Основний матеріал

У даній роботі проведено дослідження відкритих наборів даних, які можна використати для нау-

кових проектів, зокрема, для розробки та тестування рекомендаційних систем.

Тестування рекомендаційних систем – складний та неоднозначний процес, зокрема, й тому, що є різні погляди те, що розуміти під якістю роботи рекомендаційної системи.

Наприклад, користувач системи під якістю роботи буде розуміти здатність системи максимально точно спрогнозувати його вподобання та рекомендувати максимально релевантні його інтересам об'єкти.

В той же час власник системи буде вважати якісною таку рекомендаційну систему, яка буде збільшувати інтерес користувачів до сайту та кількість продажів товарів.

Взагалі існує два різні підходи до тестування рекомендаційних систем:

– offline-тестування системи на готових наборах даних;

– online-тестування системи за допомогою A/B тестування, коли на різних групах користувачів запускаються різні методи, та визначається, який метод дав кращий результат.

Відкриті набори даних, що містять дії користувачів веб-сайтів (наприклад, оцінки, перегляди, покупки, тощо) та, можливо, інформацію про користувачів (наприклад, демографічні дані), дозволяють здійснювати offline-тестування рекомендаційних систем.

Важливим фактором для розробки якісної рекомендаційної системи є можливість використання актуальних наборів даних для аналізу потреб користувачів, а також навчання рекомендаційної системи.

Нижче представлений огляд досліджених вебресурсів, що надають вільний доступ до великої кількості наборів відкритих мережевих даних.

Movielens. Проект науково-дослідної лабораторії кафедри комп'ютерних наук та інженерії в Університеті Міннесоти, що спеціалізується [4]:

в області рекомендаційних систем,

інтернет-спільнот,

мобільних технологій,

електронних бібліотек,

локальних і географічних інформаційних систем.

Містить набори даних, одержані в процесі роботи їх веб-сайту з рекомендаційною системою фільмів. Дані надаються усім користувачам на вільній основі.

Mlvis. Цей проект є першим, що об'єднав поняття сховища даних у реальному часі для візуального та інтерактивного аналізу даних, а також дослідницький аналіз в Інтернеті [5].

Статистичні методи в поєднанні з режимом реального часу для візуалізації даних дають можливість для дослідників легко знайти, вивчити, зрозуміти і відкрити для себе ключові моменти у великій кількості наборів даних.

Цей великий репозиторій даних є корисним для прийняття важливих наукових висновків, а також містить еталонні набори даних для різних додатків і областей, включає реляційні таблиці, просторові та часові ряди даних, а також нереляційні дані для машинного навчання.

Kaggle. Містить багато малих та середніх наборів даних з різних областей застосування: спорт, соціальні відносини, мобільні додатки, освіта, криптовалюти, тощо для різного роду аналізу даних [6].

Amazon. Репозиторій відкритих даних фірми Amazon існує, щоб допомагати людям відкривати і спільно використовувати набори даних, які доступні через ресурси Amazon Web Service [7]. Також сервіс надає можливості для хмарного аналізу даних.

Reddit. Надає дуже цікаві набори даних та пояснення до них. Ці набори даних налічують більш ніж терабайт корисних колекцій з можливістю вільного доступу, зокрема містить текстові дані для лінгвістичних досліджень [8].

Figshare. Це сховище, в якому користувачі можуть зробити всі свої результати досліджень доступними для суспільства.

Дозволяє користувачам завантажувати будь-який формат файлу, для перегляду в браузері, це будь-які дослідження від плакатів і презентацій до наборів даних і коду, може бути використаний іншими дослідниками [9].

Nodexlgraphgallery. Це велика колекція з мережевих карт і звітів, створених у NodeXL, де можна знайти [10]:

графи NodeXL,

архів наборів даних, завантажених користувачами спільноти NodeXL.

Мережа об'єднує групи дослідників, присвячені створенню відкритих інструментів для аналізу даних і хостингу відкритих даних.

Yelp. Веб-сайт для пошуку на місцевому ринку послуг, наприклад, ресторанів або перукарень, з можливістю додавати та переглядати рейтинги та огляди цих послуг.

Підтримує безкоштовний набір даних для використання в особистих, освітніх і наукових цілях [11].

Networkrepository. Найбільше мережеве сховище з тисячами наборів інтерактивних даних, призначених для візуалізації мережі та інтелектуального аналізу, містить декілька тисяч колекцій реальних мереж: від біологічних до соціальних [12].

UCI ML. Позиціонується як репозиторій машинного навчання та містить набори баз даних, які використовуються в машинному навчанні [13]. Спільнота призначена для емпіричного аналізу алгоритмів машинного навчання.

Github. Великий репозиторій для програмного забезпечення, де окрім програмного коду можна знайти різноманітні набори даних, включаючи дані, отримані з соціальних мереж, додатків тощо [14].

За допомогою інтерфейсу сервісу зручно переглядати оновлення потрібних баз даних.

Порівняльний аналіз розглянутих відкритих наборів даних наведений у табл. 1.

Таблиця 1 – **Порівняльний аналіз розглянутих відкритих наборів даних**

№	Джерело даних	Види даних	Наявність оцінок користувачів	Наявність тегів	Наявність коментарів	Тип ліцензії	Територіальна приналежність даних	Можливість застосування для тестування рекомендаційних систем	Формати даних
1	movielens.org	фільми	+	+	-	вільна	немає	+	csv
2	mlvis.com	різні	+	+	+	вільна	Європа, США	+	csv
3	kaggle.com	різні	+	+	+	вільна	США	+	csv, SQLite, JSON, BigQuery
4	aws.amazon.com	різні	+	+	-	вільна лише для наукових досліджень	Європа, США	+	csv, JSON
5	reddit.com	різні	+	+	+	вільна лише для наукових досліджень	СНД, Європа, США	+	csv
6	figshare.com	різні	+	+	-	вільна	Європа, США	+	csv, RDS
7	nodexlgraphgallery.org	різні	+	+	+	вільна	немає	+	graph
8	yelp.com	бізнес	+	+	+	вільна	немає	+	JSON
9	networkrepository.com	різні	+	+	+	вільна	немає	+	MTX
10	ics.uci.edu	різні	+	+	-	вільна	немає	+/-	csv, xls
11	github.com	різні	+	+	+	вільна	немає	+	csv, JSON, xls

Висновки

Досліджені відкриті набори даних містять колекції різних типів, розмірів та охоплюють різну за розмірами аудиторію. Також слід зазначити, що більшість з наборів даних мають вільний доступ до завантаження, інколи при умові, що користувач є зареєстрованим на платформі, але на деяких платформах зазначено, що дані можуть бути використані лише у дослідницьких цілях.

Перевагою переважної більшості програм є великий спектр форматів даних для завантаження, але найбільш популярним форматом для збереження даних є *.csv.

Крім того вагомим показником є територія, яку охоплюють представлені набори, подекуди вони значно обмежені лише зонами США та Західної Європи.

Зважаючи на всі переваги та недоліки розглянутих програм, можна зробити висновок, що найбільш придатними для застосування у розробці та тестуванні рекомендаційних систем наборами даних, є набори платформ Reddit, Networkrepository та Github. Дані репозиторії поєднують у собі набори придатні для статистичного та інтелектуального аналізу, мають широкі можливості візуалізації соціальних графів та підтримують більшість сучасних форматів файлів, що є дуже зручною перевагою як для дослідників так і для пересічних користувачів сервісу.

Крім того Github містить багато прикладів застосування конкретних наборів даних, а також інформацію про оновлення версій баз даних у часі, що надає багато можливостей для тестування та удосконалення рекомендаційних систем.

Також заслуговують на увагу такі платформи як Movielens, Mlvis та Kaggle, вони можуть бути досить корисними для тестування рекомендаційних систем, хоч і мають ряд обмежень.

Зокрема, першу платформу обмежено лише одним типом даних, тобто фільмами, останні дві – мають територіальне обмеження, дані зібрані переважно серед жителів США і подекуди не відповідають вітчизняним вимогам.

СПИСОК ЛІТЕРАТУРИ

1. “Recommender Systems Handbook” (2010) Editors Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, 1st edition., New York, NY, USA: Springer-Verlag New York, Inc., p. 842.

2. Segaran T. (2008) “Programming Collective Intelligence. Building Smart Web 2.0 Applications”, O'Reilly Media, p. 368.
3. Меньшикова Н.В., Портнов И.В., Николаев И.Е. (2016) “Обзор рекомендательных систем и возможностей учета контекста при формировании индивидуальных рекомендаций”, ACADEMY, №6, с. 20–22.
4. Harper, F.M. and Konstan J.A. (2016) “The MovieLens Datasets: History and Context”, ACM Transactions on Interactive Intelligent Systems (TiiS), available at: https://doi.org/10.1145/2827872
5. “Scientific data repository. Real-time visualization and exploration techniques interactive visual analytics”, available at: http://mlvis.com/
6. “Kaggle is the place to do data science projects”, available at: https://www.kaggle.com/datasets
7. “Registry of Open Data on AWS”, available at: https://registry.opendata.aws/
8. R/datasets”, available at: https://www.reddit.com/r/datasets
9. “Online open access repository Figshare”, available at: https://figshare.com/
10. NodeXL graph gallery”, available at: http://nodexlgraphgallery.org/Pages/Default.aspx
11. Yelp Open Dataset. An all-purpose dataset for learning”, available at: https://www.yelp.com/dataset
12. Network repository. A scientific network data repository with interactive visualization and mining tools” , available at: http://networkrepository.com/
13. UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems”, available at: https://archive.ics.uci.edu/ml/datasets.php
14. Github”, available at: https://github.com/search?q=dataset

Рецензент: д-р техн. наук, проф. С. Г. Семенов,
Національний технічний університет «Харківський політехнічний інститут», Харків
Received (Надійшла) 14.06.2019
Accepted for publication (Прийнята до друку) 31.07.2019

Исследование открытых наборов данных веб-ресурсов
в контексте применения их для тестирования рекомендательных систем

Д. В. Шингалов, Е. В. Мелешко, В. В. Босько

Предметом изучения в статье является процесс тестирования методов построения рекомендательных систем на основе открытых наборов данных в сети Интернет. **Целью** является исследование открытых наборов данных веб-ресурсов в контексте применения их для тестирования различных методов построения рекомендательных систем. **Задачи:** исследовать современные веб-платформы с открытыми наборами данных и возможность применения их данных для тестирования качества работы различных рекомендательных систем. Получены следующие **результаты.** Рассмотрены наиболее популярные веб-платформы с открытыми наборами разнотипных сетевых данных. Осуществлен сравнительный анализ этих платформ с точки зрения наличия свободного доступа к загрузке данных, их функциональности и территориальной принадлежности, формата данных и удобства для дальнейшего использования для машинного обучения, а также возможности применения для тестирования рекомендательных систем. Также проведена оценка актуальности данных, хранящихся в репозитариях со свободным доступом и наличия их обновления со временем. **Выводы.** Исследованы веб-платформы, содержащие открытые наборы данных, которые можно использовать для тестирования рекомендательных систем. Основными преимуществами большинства платформ является поддержка современных форматов данных и условно свободный или свободный доступ. Среди недостатков рассмотренных платформ следует отметить недостаточную структурированность некоторых наборов данных, в частности текстовых, что значительно ограничивает их применение для тестирования методов контентной фильтрации. Кроме того, одним из факторов, который ограничивает использование открытых наборов данных, является их актуальность, так как некоторые наборы, хранящиеся на платформах, устарели и не обновляются. Все рассмотренные наборы данных могут быть применены для исследовательских целей и тестирования работы рекомендательных систем.

Ключевые слова: рекомендательные системы, тестирование, анализ данных, открытые наборы данных, цифровой маркетинг.

Research of open data sets of web resources
in the context of their application for testing recommendation systems

D. Shynhalov, Ye. Meleshko, V. Bosko

The **subject matter** of the article is the process of testing methods of building recommender systems based on open data sets from the Internet. The **goal** is to research open data sets of web-resources in the context of using them to test various methods of building recommender systems. The **tasks** to be solved are: to explore modern web-platforms with open data sets and to research the possibility of using their data to test the work quality of various recommender systems. The following **results** were obtained: The most popular web-platforms with open sets of various network data were considered. The comparative analysis of these platforms was carried out in terms of the availability of free access to downloads of data, their functionality and geographical location, data format and convenience for future use for machine learning, as well as the possibility of using for testing recommender systems. Also, an assessment of the relevance of data stored in repositories with free access and their availability over time was made. **Conclusions**. Web-platforms containing open data sets that can be used to test recommender systems were explored. The main advantages of most platforms are the support of modern data formats and conditionally free or free access. Among the shortcomings of the considered platforms, it is worth noting the lack of structuredness of some data sets, in particular, text data, which significantly limits their use for testing content-based filtering methods. In addition, one of the factors that limit the use of open data sets is their relevance, since some of the sets stored on the platforms are outdated and not updated. All considered data sets can be applied for research purposes and for testing the work of recommender systems.

Keywords: recommendation systems, testing, data analysis, open data sets, digital marketing.