УДК 004.8

К.Ю. Тітов

РОЗРОБКА СЕМАНТИЧНОГО ФІЛЬТРУ НА ОСНОВІ ПЕРСОНАЛЬНИХ ВПОДОБАНЬ КОРИСТУВАЧА

Анотація. У даній статті були досліджені проблеми сучасних рекомендаційних систем, методи надання рекомендацій, а також отримання релевантної інформації. Також були розглянуті технології штучного інтелекту, за допомогою яких можливо вирішити дані проблеми. В статті викладений один із варіантів вирішення проблеми рекомендацій стосовно новинного контенту комбінуючи інструменти семантичного порталу та інтелектуального сервісу Personality insights.

Ключові слова: Semantic WEB, IBM, TRUST, пошук, новини, об'єктивність.

Вступ

Завдяки сучасним технологіям стрімко збільшуються об'єми новинної інформації, яку отримують користувачі. Але чим більше інформації надходить, тим важче її обробляти та аналізувати, а тим паче знайти саме ту, що цікавить користувача. Крім того деякі сучасні держави вдаються до нечесної гри, беручі на озброєння масмедіа, та розповсюджуючи за їх допомогою т.з. фейкові новини. Західні демократії вперше зіштовхнулися з таким явищем під час російської агресії на сході України та в Криму у 2014 року, коли величезна пропагандистська машина Кремля публікувала величезні об'єми неправдивої інформації стосовно подій у Криму, Україні та у країнах ЄС. Завдяки таким новинам можливо маніпулювати свідомістю громадян та впливати на демократичні процеси у західній спільноті, яка не була готова до боротьби с такою великою кількістю неправдивої або спотвореної новинної інформації.

Щоб знайти необхідну інформацію в мережі Інтернет, її треба проаналізувати, визначити пріоритети, відфільтрувати та відкинути зайве. Рекомендаційні системи вирішують цю проблему шляхом пошуку великого обсягу динамічної інформації, щоб забезпечити користувачам персоналізований контент та послуги. У даній роботі ми аналізуємо різні алгоритми формування рекомендацій та методи для

 ISSN 1562-9945

обчислення подібності вподобань користувачів, а також методи отримання інформації з цих систем. Ми запропонуємо новий підхід до надання рекомендацій користувачу та порівняємо його з вже існуючими методами та алгоритмами.

Еволюція рекомендаційних систем

За останні двадцять років було проведено багато досліджень на тему як автоматично надавати рекомендації користувачу. За цей час було запропоновано багато різних методів та значно збільшився інтерес до рекомендаційних систем. Рекомендаційні системи повинні співпрацювати із користувачем, отримуючи точні дані, щоб встановити його вподобання, зменшити засміченість у наборах даних і надати рекомендації стосовно контенту. Одним із перших кроків по впровадженню рекомендаційних систем було створення комп'ютерного бібліотекаря «Grundy»[1]. Це була досить примітивна система, яка групувала користувачів у т.з. стереотипи на основі невеликого інтерв'ю користувача та застосовуючи інформацію стосовно вподобань кожного зі стереотипів, яка в свою чергу була завантажена до системи і надавала рекомендації користувачам.

Однією з перших була Tapestry[2] – полу-автоматична, система колаборативної фільтрації. Вона дозволяла користувачу робити запити до інформаційного домену, та надавала рекомендації за досвідом попередніх запитів або дій користувачів. Це потребувало зусиль зі сторони користувача, але завдяки цьому стало можливим використовувати досвід попередніх користувачів. Незабаром були розроблені нові системи сумісної фільтрації, які вже автоматично визначали відповідні думки та узагальнювали їх за для надання рекомендацій. Нова система GroupLens[3] використовувала цей метод для ідентифікації статей, які б могли бути цікаві конкретному користувачу. Користувачам було необхідно надавати рейтинги до кожної статті, далі ці рейтинги система поєднувала з рейтингами інших користувачів й надавала персоналізовані рекомендації.

Наприкінці 90 х років почали з'являтися перші комерційні рекомендаційні системи. Можливо найбільш популярною рекомендаційною системою є Amazon – система на підставі історії покупок та переглядів товарів надає користувачу відповідні рекомендації. Після цього рекомендаційні системи почали інтегрувати в системи електронної комерції, адже завдяки цим сис-

темам збільшувалися об'єми продажу товарів. Рекомендаційні системи знову привернули значну увагу у 2006 році, коли Netflix запропонував заохочувати користувачів за для поліпшення своєї рекомендаційної системи. Метою цього конкурсу було створення алгоритму рекомендацій, який міг би перевершити свій внутрішній алгоритм CineMatch в автономних тестах на 10%. Це викликало значне збільшення активності як в академічних колах, так і серед звичайних користувачів.

Сьогодні існує декілька варіантів вирішення питання надання рекомендацій стосовно контенту. У цій статті ми розглянемо два з них – колаборативна фільтрація та рекомендації на основі контенту:

Рекомендації від подібних користувачів (collaborative filtering) – система визначає на скільки схожий користувач з іншими користувачами із бази даних.

Контентні рекомендації (content-based filtering) – система за оцінками інших користувачів передбачає яку саме оцінку поставив би користувач, враховуючи тих користувачів або продукти які більш схожі на даний.

На рис.1 зображені різні типи рекомендаційних систем.

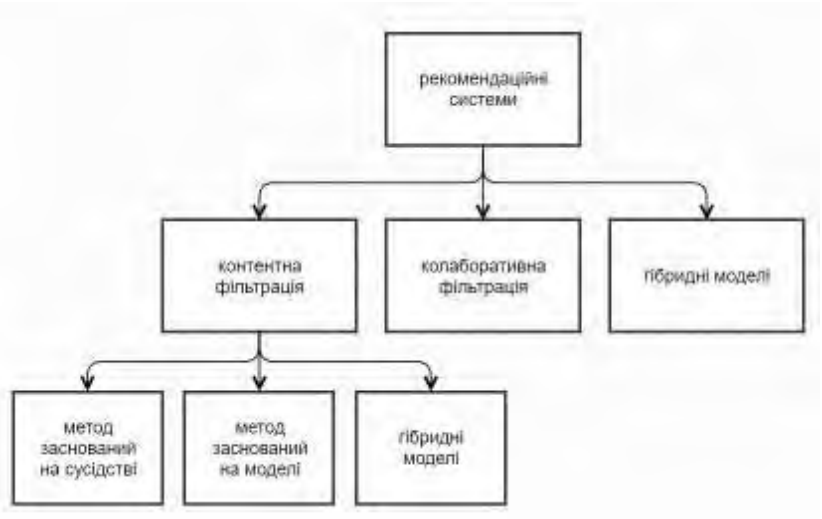

Рисунок 1

На початку 90х років за для вирішення питань з навантаженням в онлайнових просторах почали використовувати колаборативну фільтрацію. Основним припущенням цього методу є те, що думки інших користувачів можуть бути виділені та об'єднані таким чином, щоб забезпечити обґрунтоване прогнозування переваг активного користувача. Колаборативна фільтрація надає рекомендації, засновані

на моделі попередньої поведінки користувача. Ця модель може бути побудована виключно на основі поведінки певного користувача або з урахуванням поведінки інших користувачів зі схожими вподобаннями. У тих випадках коли колаборативна фільтрація бере до уваги поведінку інших користувачів вона використовує знання о цільових групах за для надання рекомендацій на підставі схожості поведінки і вподобань користувачів. Такі рекомендацій базуються на автоматичній співпраці багатьох користувачів, які демонструють схожі вподобання або поведінку[4].

Існує два основних метода колаборативної фільтрації:

Заснований на сусідстві – для користувача підбирається група користувачів з найбільш схожими до нього вподобаннями, комбінуючи оцінки підгрупи система надає рекомендації.

Заснований на моделі – цей підхід надає рекомендації, визначаючі параметри статичних моделей за для оцінювання користувачів. Моделі можуть бути побудовані за допомогою байесових мереж, кластерезації и т.п.

Ці методи можуть бути поєднані за для отримання найбільш релевантного результату.

Підхід заснований на моделі є більш комплексним та надає більш точні рекомендації. Цей підхід більш ефективно обробляє великі набори даних, на відміну від методу заснованого на сусідстві. Недоліком цього підходу є необхідність компромісу між точністю моделі та її розміром.

Не зважаючи на всі переваги, у колаборативної фільтрації є низка недоліків та проблем:

Розрідженість даних: більшість комерційних, рекомендаційних проектів мають велику кількість товарів або послуг, користувачі не завжди мають змогу виставити їм оцінку. Через це матриця виходить досить великою та розрідженою.

Масштабованість: зі збільшенням кількості користувачів стає важче її обчислювати, також збільшується час на обробку, а деякі системи повинні швидко реагувати на запити користувачів.

Шахрайство: кожна людина може ставити гарні оцінки своїм улюбленим товарам і навпаки низькі – конкурентам. Це призводить до того, що деякі недобросовісні гравці можуть шахрайським методом

піднімати рейтинги своїм товарам, а інколи занижувати рейтинги конкурентів.

Контента фільтрація формує рекомендацію на основі поведінки користувача. Цей метод може використовувати інформацію о переглядах або вподобаннях користувача накопичену за якийсь термін. Якщо користувач приділяє увагу до певних тематичних статей, видань, ресурсів, регулярно залишає коментарі під певними темами, виставляє рейтинги тощо, то контентна фільтрація використовує цю інформацію за для виявлення подібного контенту.

Контентна фільтрація має декілька переваг над колаборативною:

Незалежність користувачів: колаборативна фільтрація потребує рейтингу інших користувачів, щоб знайти подібність між користувачами, а потім надати рекомендації. Натомість метод, оснований на контенті, повинен лише аналізувати елементи та профіль користувача для рекомендації.

Прозорість: колоборативні методи надають рекомендації засновуючись на тому, що деякі користувачі мають вподобання схожі з вашими, а завдяки контентному методу користувач знає за якими характеристиками цей контент був запропонований.

Відсутній «швидкий старт»: на відміну від колаборативної фільтрації, контент може бути запропонований без попереднього збору інформації о вподобаннях користувачів.

З основних недоліків контентної фільтрації можливо виділити наступні:

Необхідність встановлювати зміст контенту.

Потреба у ручному або автоматичному індексуванні.

Необхідність встановлювати вподобання користувача, для цього необхідно витратити деякий час.

Інтуїтивна прозорливість – ефект при якому користувач випадково натрапляє на щось цікаве, у той час коли він шукав зовсім іншу за змістом інформацію [5].

На сьогодні існує декілька варіантів вирішення питання надання рекомендацій користувачу. Але колоборативна та контентна фільтрації, поодинці, мають низку недоліків, тому у великих компаніях створюють гібридні моделі, які поєднують в собі декілька класичних рекомендаційних методів методів. Тому було вирішено

розробити свій персональний фільтр який би міг не тільки фільтрувати контент відповідно вподобанням користувача, а ще й аналізувати і робити висновок стосовно інформації, що надходить. За для аналізу інформації, що надходить ми вирішили використовувати інтелектуальний сервіс від IBM Watson – Personality Insights.

2. Застосування сервісу Personality Insights від IBM Watson

IBM Watson це (когнітивна комп'ютерна система штучного інтелекту, яка здатна обробляти великі об'єми неструктурованих даних і відповідати на запитання, поставлені натуральною мовою, система здатна навчатися, розуміти та робити висновки.) проект, що представляє собою когнітивну систему, яка може навчатися, розуміти та робити висновки. Одним із інтелектуальних сервісів системи IBM Watson є Personality Insights, доступ до якого надається в хмарному середовищі. Сервіс Personality Insights використовує лінгвістичну аналітику, щоб визначити характеристики особистості, внутрішні потреби та цінності людей. Система аналізує публікації автора, ті які користувач вирішує зробити публічними, електронна пошта, текстові повідомлення, соціальні медіа, публікації на форумах, блогах тощо. Спочатку сервіс розмічає текст для створення представлення у n-мірному просторі. Сервіс використовує технологію word-embeding, з відкритим вихідним кодом, для того щоб отримати векторне представлення для усіх слів з тексту. Далі Personality Insights передає це представлення алгоритму машинного навчання, який описує профіль особистості із його характеристиками. Для навчання алгоритму сервіс використовує оцінки, отримані з опитувань, проведених серед тисяч користувачів, а також їх акаунтів соціальної мережі twitter.

Характеристики особистості, потреб та цінностей, використані у Personality Insights, допомагають компаніям краще зрозуміти своїх клієнтів та покращити рівень задоволеності клієнтів, передбачаючи потреби клієнтів та рекомендуючи майбутні дії. Це дозволяє підприємствам покращувати нові придбання, збереження та залучення клієнтів, а також зміцнювати їх відносини з існуючими клієнтами [6]. Цей сервіс може бути також використаний для аналізу публікації з висновками за трьома видами опису особистості, які зображені на Рисунок 2:

Рисунок 2

Характеристика особистості: сервіс може побудувати портрет особистості та визначити, як він взаємодіє із світом за п'ятьма основними характеристиками: відкритість, сумлінність, екстраверсія, агресивність і невротизм.

Потреби: сервіс може вивести певні аспекти продукту, який буде резонувати з індивідуумом за дванадцятьма потребами: хвилювання, гармонія, цікавість, ідеал, близькість, самовираження, свобода, любов, практичність, стабільність, виклики та структура.

Цінності: сервіс може визначити цінності, які описують мотиваційні чинники, що впливають на прийняття рішень людиною за п'ятьма вимірами: трансцендентність або допомога іншим, збереження традицій, гедонізм або насолода життям, самовдосконалення або досягнення успіху та відкритість для змін. Таким чином, сервіс виводить з соціальних медіа, де зберігається неоднорідна та не структурована інформація, психологічні портрети людей, що відображають їх особистості [7].

 ISSN 1562-9945

3. Розробка персонального новинного фільтру

Другою складовою нашого методу, яким ми пропонуємо вирішувати проблему надання рекомендацій, щодо новинного контенту, є модернізація веб-системи, яка здатна накопичувати новинний контент та його мета-опис, який характеризує автора статті. Такою веб-системою є онтологічний портал оцінки якості вищої освіти в Україні – портал, який був розроблений в міжнародному проекті Tempus «Національна система забезпечення якості і взаємної довіри в системі вищої освіти (TRUST)» як технічний засіб підтримки і гармонізації процесів з оцінки і забезпечення якості вищої освіти. Для побудови порталу використана технологія за стандартами Semantic Web. Портал розгортається автоматично, використовуючи знання про галузь освіти і науки та про особливості власного функціонування і інтерфейсу, що зберігаються в двох онтологіях доменній і сервісній. Портал надає можливість накопичувати факти, які описують ресурси предметної галузі, і аналітично обробляти їх, що забезпечує можливість прозорого та неупередженого контролю за якістю внутрішніх процесів, заснованих на оцінці ресурсів.

Подібні системи посилюють соціальний вплив на оцінку інформації та унеможливлюють контроль за нею лише зацікавленою стороною. Для цього подібні системи будуються за принципами соціальних мереж. Користувачі порталу є головними контролерами достовірності зареєстрованих фактів завдяки механізму соціальної верифікації. Для оцінювання ресурсів портал дозволяє створювати і застосовувати різні системи цінностей у вигляді гнучких багатовимірних показників якості, зважених за ступенем їх важливості для ранжування запиту оцінки. Таким чином, кожен користувач може оцінити якість ресурсів з різних точок зору, так званих «систем цінностей».

Інформація, що використовується для опису архітектури та функціональності порталу, представлена у онтологічному вигляді (в сервісній онтології). Гнучкість порталу забезпечується за рахунок розподілу його архітектури та функціоналу на дві онтології – доменну та сервісну. Доменна онтологія відповідає за поняття і властивості, які використовуються для опису предметної галузі. Таким чином, модифікуючи доменну онтологію, ми можемо повністю змінити предметну галузь порталу. Сервісна онтологія

використовується в якості незалежної допоміжної структури та для гнучкої взаємодії з доменною онтологією. Вносити будь які зміни до сервісної онтології не потрібно [8].

За для використання порталу у предметній області «новини» ми реконфігурували домену онтологію відповідно до нових вимог. Кореневим класом створеної онтології є «стаття», який має низку властивостей, зображених на мал.2, які характеризують особистість автора. Ці властивості є основними логічними розділами сервісу Personality Insights від IBM Watson, які включають в себе більш детальні характеристики – властивості, зображені на Мал.3, для опису індивідуальності автора статті.

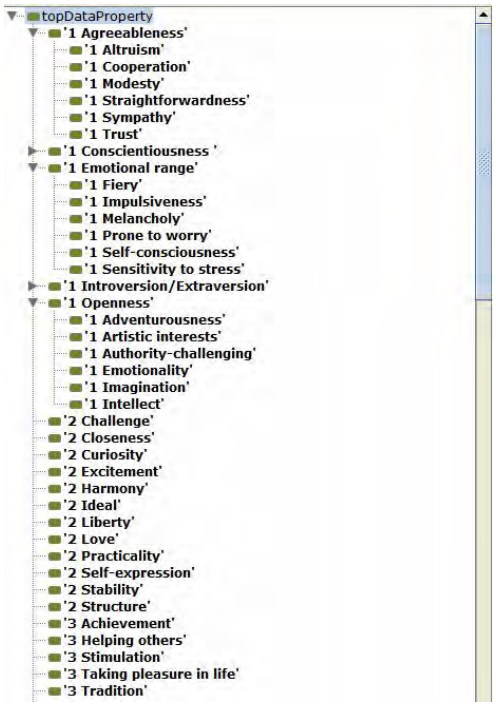

Рисунок 3

Таким чином, спираючись на властивості сервісу Personality Insights користувач, знаходячись в своєму особистому просторі, може створювати свої системи цінностей, встановлюючи вагові коефіцієнти відповідно до своїх вподобань. За для подальшої фільтрації публікацій ЗМІ ми повинні надати кожній статті рейтинг за допомогою зовнішніх експертів. У якості зовнішнього експерта ми використовуємо сервіс Personality Insights. Цей сервіс надає зручний API, що спрощує його розгортання та інтеграцію до будь-яких веб-систем. Сервіс аналізує та надає вагові коефіцієнти характеристикам статті, які зображені на Рисунок 4.

Рисунок 4

Далі користувач отримує відфільтрований, згідно із попередньо створеною системою цінностей, контент. Розробивши декілька таких систем цінностей користувач зможе отримувати контент який відповідає тільки його вподобанням або переконанням. Наприклад, з ранку користувачу потрібні новини практичність яких не була б нижчою за 90%, а ввечері він хоче отримувати емоційні новини але тільки приємні, в яких процент позитиву не менш ніж 85%. Завдяки цьому утворюються додаткові умови для проходження статей крізь фільтри користувачів, це повинно стимулювати авторів за для поліпшення якості написання новинного контенту. Таким чином, комбінуючи інструменти порталу та інтелектуального сервісу Personality Insights, ми можемо аналізувати і згодом фільтрувати новинний контент за нашими вподобаннями.

Висновки

Проаналізувавши існуючі методи надання рекомендацій ми встановили, що вони мають низку недоліків, які ускладнюють надання рекомендацій користувачам. Існуючі системи, в своїй більшості, використовуються в комерційних проектах та мають закритий код, що унеможливлює їх використання або модернізацію за для використання їх у інших галузях. Ми виявили, що рекомендаційні системи надають більш релевантні рекомендації, якщо вони працюють з однією предметною областю та засновані на гібридних моделях фільтрації. Тобто поєднання різних методів фільтрації в одній системі поліпшує якість наданих рекомендації.

Організація рекомендаційної системи у вигляді соціальної мережі, де користувачі зможуть викладати та ділитися своїми системами цінностей, повинно сприяти поліпшенню якості написання но-

винного контенту. Залучення широких мас до фільтрації новинного контенту ускладнить масове розповсюдження фейкових новин. У подальшому планується провести порівняльний аналіз фейкових та правдивих новин, задля встановлення спільних рис всіх фейкових новин або певні фундаментальні розбіжності. Це допоможе у майбутньому встановлювати правдивість новини.

ЛІТЕРАТУРА

1. E. Rich, “User modeling via stereotypes,” Cognitive Science, vol. 3, no. 4, pp. 329–354, October 1979.

2. D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” Communications of the ACM, vol. 35, no. 12, pp. 61–70, 1992.

3. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: an open architecture for collaborative filtering of netnews,” in ACM CSCW ’94, pp. 175–186, ACM, 1994.

4. G Adomavicius, A Tuzhilin Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, In: Knowledge and Data Engineering, IEEE Transactions on, Vol. 17, No. 6. (2005), pp. 734-749

5. Recommender Systems through Collaborative Filtering [Електронний ресурс]. – Режим доступу:
https://blog.dominodatalab.com/recommender-systems-collaborative-filtering/

6. Personality Insights, documentation [Електронний ресурс]. – Режим доступу:
https://console.bluemix.net/docs/services/personality-insights/getting-started.html#getting-started-tutorial

7. The science behind the service [Електронний ресурс]. – Режим доступу:
https://console.bluemix.net/docs/services/personality-insights/science.html#science

8. Terziyan V., Golovianko M., Shevchenko O., Semantic Portal as a Tool for Structural Reform of the Ukrainian Educational System, In: Information Technology for Development, Vol. 21, No. 3, 2015, Taylor & Francis, pp. 381-402.