Інформаційні технології

УДК 004.3

Халін Олег Ігорович

студент

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

ДОСЛІДЖЕННЯ КОЕФІЦІЄНТУ РЕЛЕВАНТНОСТІ НОВИННИХ СТАТТЕЙ З МЕТОЮ ПРЕДСТАВЛЕННЯ НЕОБХІДНОЇ КОРИСТУВАЧАМ ІНФОРМАЦІЇ

Анотація. *У роботі представлено аналіз та визначення способу обчислення коефіцієнту релевантності новинних статей з метою надання користувачам необхідної актуальної інформації. Наведено постановку задачі, аналіз подібних публікацій, сформульовано проблему та описано шляхи та засоби її вирішення.*

Ключові слова: *рекомендаційні системи, новинні статті, релевантність новин, аналіз поведінки користувачів, рекомендації.*

Вступ. Кожен з нас зараз проживає в умовах дезінформованості, люди не знають де шукати інформацію. Більшість інформації подається в суміші з розважальним контентом і з якоюсь метою, чи то реклама, чи то завоювання лояльності аудиторії.

Щоб процес пізнання новин та визначення інформації щодо всього, що виникає у світі проходив зрозуміло і результативно – людям подається релевантна інформація основуючись на індексах, що описані нижче. Також для надання користувачам повідомлення про небезпеку чи необхідної термінової інформації, використовується геолокація користувача.

В 21 сторіччі інформація і слово являють собою найпотужнішу зброю для маніпулювання людьми, можливості нав'язати комусь певну думку, відвернути від значущої події або «розправитися» з конкурентом. Фейкову інформацію умільці маскують під правду так креативно, що довірливі читачі сприймають як правду.

Щоб отримувати інформацію у всьому світі людям потрібне єдиний додаток, куди можна заходити, щоб отримувати лише ділові новинні статті, а не розважальний контент в переважності.

Щоб отримувати необхідну інформацію в конкретний момент, користувачі мають отримувати саме релевантні новинні статті.

Тож серед переваг використання рекомендаційних систем можна виділити такі можливості:

Чітке розділення новинного і розважального контенту

На даний момент людям подається строгий контент у перемішку з розважальним контентом. Люди попросту не знають точно, що із запропонованої їм інформації є корисним і діловим, а що просто дезінформація або насмішка.

Покращення залученості користувачів

Для читача та людини, що хоче знати актуальні і правдиві новини є потреба у виборі контенту. Загалом, можна сказати, що основною метою введення рекомендаційних систем є збільшення коефіцієнта конверсії, тобто кількості користувачів, які отримують рекомендації і можуть з легкістю знайти необхідну інформацію і бути впевненим, що ця інформаціє є цілком корисною і правдивою.

Підвищення лояльності користувачів

Користувач буде лояльним до додатку, який під час відвідування впізнає старого користувача, розуміє й знає його вподобання, теми які близькі користувачу. Тож цінною особливістю рекомендаційних систем є те, що вони враховують взаємодію та попередні оцінки, прочитані статті,

направленість діяльності користувача, це дозволяє рекомендаційним системам надавати релевантні рекомендації.

Актуальність. Побудова системи релевантності новин є актуальною задачею сьогодення. При її проектуванні необхідно брати до уваги велику кількість факторів, у тому числі й поведінку користувача, та його суб'єктивні точки зору на деякі події у світі. Дані є слабоструктурованими, що призводить до надання недостовірної інформації і не точних результатів.

Проведено ряд експериментів для визначення суті проблеми, одним з яких є опитування в одній із областей України чи можуть люди самостійно відрізнити правду від фейку. 65% із 95 опитаних людей впевненні, що можуть [1]. Однак лише одиниці змогли дати відповідь і аргументувати як саме вони відрізняють правду від фейку.

В європейському дослідженні «Trust in media» [2] встановлено, що 88% населення країн Європейського Союзу висловлюють недовіру до соціальних мереж, 68% – Інтернету, 53% – пресі, 50% – телебаченню. Найбільшою довірою користується радіо – 59%. Статистика якомога краще показує всю серйозність ситуації.

Розгляд даної тематики має за мету - вирішення проблем багатьох людей. По-перше показувати людям ті новини, котрі їх інтересують або мають велику значимість у світі. Основуючись на рейтингу кожного користувача і його минулим діям, а також по рейтингу самого поста формулювати релевантність і значимість новини.

Мета дослідження. Метою роботи є дослідження методів побудови системи релевантності новин. Це підвищить усвідомленості населення України щодо новинних статей шляхом наданням безкоштовного продукту, який надає змогу читати актуальні релевантні новини та дає змогу бути обізнаним у випадку коли небезпека досить близько до користувача.

Підвищення ефективності підбору рекомендацій за допомогою виведених формул підрахунку релевантності новин беручи до уваги велику кількість факторів, у тому числі поведінку користувача.

Завдання дослідження. Завданням дослідження є розробка методу, що дозволить прогнозувати релевантні рекомендації новин для користувачів додатку на основі коефіцієнтів, що найбільше впливають на процент релевантності новинних статей для користувачів.

Об'єктом дослідження роботи є процеси формування рекомендацій новинних статтей на основі поведінки користувачів додатку.

Предметом дослідження є аналіз підходів до методу аналізу релевантності новини з метою вирахування оптимального коефіцієнта релевантності.

Аналіз існуючих способів обчислення релевантності

Релевантна інформація важлива для прийняття рішень тому, що вона містить дані, які слід використовувати для розрахунків при підготовці інформації для майбутніх дій. Нерелевантна інформація - це несуттєві, надлишкові дані про витрати та доходи. Використання нерелевантної інформації може призвести до таких наслідків:

- прийняття помилкового рішення в результаті викривлення інформації, яка описує проблемну ситуацію, щодо якої слід прийняти рішення;
- зниження оперативності та підвищення трудомісткості процесу прийняття рішення, тобто відсутнє викривлення інформації, хоча користувач отримує зайву інформацію, яка збільшує час прийняття рішення.

Релевантний підхід - зосередження уваги лише на релевантній інформації в процесі прийняття рішення, що при значному обсязі інформації дозволяє полегшити та прискорити процес прийняття найкращого рішення.

Існує багато алгоритмічних формул для вирішення питання релевантності. Найкращими прикладами реалізації даних алгоритмів є пошукові система та соціальні мережі. Вони, як ніхто інший гарно проробили алгоритми і дають користувачам найбільш необхідні посилання, що допомагають людям.

Однак дані системі не викривають своїх алгоритмів, оскільки тоді люди змогли би їх використовувати у корисних цілях.

Є багато догадок і перевірених часом стратегій, однак дані системі весь час розвивають свої алгоритми задля досягнення ідеалу.

Спосіб вирахування релевантності новинної статті для надання необхідної користувачу інформації

Кожна людина має суб'єктивні враження щодо новин і їх значимості. І виходячи з того який саме читач і з якою ціллю він шукає новини ми маємо вираховувати йому релевантність контенту по різному.

Обчислення загального коефіцієнту релевантності новинних статей буде складатися з обчислення декількох допоміжних коефіцієнтів, таких як:

- Рейтинг статті;
- Індекс приналежність до певної категорії;
- Індекс приналежність ключових слів до категорії (адаптована категорія);
- Індекс релевантності в залежності від тегів;
- Індекс релевантності в залежності від кількості згадки статті;
- Індекс релевантності в залежності від геолокації користувача.

Для обчислення кожного з індексів використовується своя формула. Це розбиття було виконано з метою досягнення якомога чіткіших результатів.

Рейтинг статті

1 – максимальний рейтинг, 0- мінімальний

$Ri = \frac{n60*.8+n20*.3+n5*.2}{Np}$, де:

- n60 – кількість людей що читали статтю більше хвилини;
- n20 – кількість людей що читали статтю більше 20 секунд;
- n5 – кількість людей що читали статтю більше 5 секунд;
- Np – загальна кількість переглядів статей;
- Ri – рейтинг статті.

Дана формула враховує показники перегляду статей і кількість часу, яку люди приділяють вивченню статі. Таким чином ми обрали тестові коефіцієнти, на які множимо кількість людей що перебували на сторінці на відповідну кількість часу. Так як дані коефіцієнти являються тестовими, дана формула потребує вдосконалення з часом і проведення A/B тестування.

Індекс приналежності до певної категорії

Якщо новини належить до даної або подібної категорії (від 0 до 1)

$Cati = \frac{\sum (n*k)}{Nc}$, де:

- n – кількість статей в даній категорії;
- k – коефіцієнт близості даної категорії (якщо категорії співпадають - k=1);
- Nc – загальна кількість категорій.

Даний індекс використовується задля виявлення приналежності або близості статі до категорії. Таким чином, завдяки вподобанням користувача, ми можемо виявити його улюблені категорії та виділити подібні категорії (наприклад категорії автомобілі та мототехніка є досить близькими категоріями по суті). Опираючись на дані результати - вивести індекс для подальший формули релевантності.

Індекс приналежності до адаптованої категорії

Якщо новини належить до даної або подібної категорії по ключам в статті (від 0 до 1)

$Cat'i = \frac{\Sigma \quad (\frac{Nk}{Na} * n * k)}{Nc}$, де:

- Nk – кількість ключових слів що належать до даної категорії;
- Na – загальна кількість ключових слів;
- n – кількість статей в даній категорії (категорії за ключовими словами);
- k – коефіцієнт близості даної категорії (якщо категорії співпадають - k=1);
- Nc – загальна кількість категорій.

Даний індекс враховує ключові слова для виявлення тематики статті і розбиття її на умовні категорії за сенсом контенту.

Індекс релевантності в залежності від тегів

Якщо новини мають суміжні теги (від 0 до 1)

$Tagi = \frac{tags}{tagp}$, де:

- tags – кількість суміжних тегів у статті до тегів людини;
- tagp – загальна кількість улюблених тегів людини.

Індекс залежить від тегів. Подібно поділу за категоріями, теги використовуються як допоміжні частини статті (ключові слова). Вони мають досить великий вплив, так як теги більш точкові і можуть показувати більше детально тематику статті.

Індекс релевантності в залежності від кількості згадок статті

Якщо новини були замічені у інших виданнях статті (від 0 до 1)

$Ki = \frac{ks}{Np}$, де:

- ks – кількість згадок у інших виданнях;

- Np – загальна кількість переглядів новини.

Даний коефіцієнт більшість розробників та сео-спеціалістів називають "вагою посилань". Тож чим більше згадок у інших виданнях, репостів, відправок іншим людям у соціальних мережах - тим більш вагома новина.

Індекс релевантності в залежності від геолокації користувача

Якщо коефіцієнт більше 0 – використовуємо, якщо ні – він дорівнює 0 (від 0 до 1)

$Gi = 1 - Gkm > 0 \; ? \; 1 - Gkm : 0$, де:

- Gkm – відстань від користувача до події в 100 тис. км.

Таким чином ми намагаємося по локації події виявити на скільки критична для користувача новина. Ми прирівнюємо коефіцієнт до 0 для тих новин. які відбулися або пов'язані з тими місцями що знаходяться далі ніж за 100 тис. км. від локації юзера. Інакше мі вираховуємо даний коефіцієнт за наведеною формулою.

Загальний індекс релевантності

$$Fi = \frac{Ri + Cati + Cat\imath i + Tagi + Ki + Gi}{6}.$$

Дана формула використовує описані вище індекси та формує фінальний показник релевантності. Вона не є ідеальною і являє собою скоріш фундамент для розвитку. Необхідно тестувати та оцінювати реальний результат. Проводити роботи над вдосконаленням і брати до уваги набагато більшу кількість факторів.

Висновки. У даній роботі був розглянутий спосіб обчислення коефіцієнту релевантності статті, який базується на декількох параметрах та факторах поведінки користувачів задля надання користувачам необхідної інформації.

Перевагами описаного вище алгоритму є досить висока точність результатів за рахунок використання лінійної математичної моделі.

Недоліком можна вважати те, що врахована не максимально можлива кількість параметрів для обчислення індексів.

Також можна покращувати систему і додавати більш складні індекси, наприклад індекс рейтингу публіциста, людей, що оцінили статтю та аналіз самого контенту.

Література

1. Опитування «Чи можуть люди відрізнити фейк від правдивої інформації?» [Електронний ресурс]: [Веб-сайт]. URL: https://www.facebook.com/DonbassLiveMedia/posts/743514775984119
2. Research «Trust in media». [Електронний ресурс]: [Веб-сайт]. – Електронні дані. URL: https://www.ebu.ch/publications/research/login_only/report/trust-in-media;
3. Наукова стаття «The Challenges of Explicit and Implicit Communication: A Relevance-Theoretic Approach», Jodłowiec M., 2015.
4. Наукова стаття «Research on relevance in information science: A historical perspective», Saracevic T., 2012.