УДК 004.048
В.В. Семенюк
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ЗАСТОСУВАННЯ ФАКТОГРАФІЧНОГО ПІДХОДУ ДЛЯ ПОШУКУ ПОВ'ЯЗАНИХ ТА АКТУАЛЬНИХ ДАНИХ В СИСТЕМІ КОНСОЛІДАЦІЇ СОЦІОКОМУНІКАЦІЙНИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ВЕЛИКИХ ДАНИХ

Semeniuk

APPLICATION THE FACTUAL METHOD TO FIND RELATED AND ACTUAL DATA IN THE SYSTEM FOR CONSOLIDATION OF SOCIO-COMMUNICATION INFORMATION RESOURCES WITH USING TECHNOLOGIES NATURAL LANGUAGE PROCESSING AND BIG DATA

Ефективний пошук одиниць контенту на основі введеного чи складеного запиту в множині інформаційних ресурсів системи, з забезпеченням отримання релевантних результатів та високого показника пертінентності, є однією з ключових функціональних можливостей, які передбачаються діаграмою прецедентів [1].

Застосування одного з сучасних методів пошуку дозволяє видавати релевантні результати, базуючись на відповідності вмісту документу запиту та гіперпосиланнях на нього [2], але не має можливості оцінювати та враховувати в пошуковій видачі достовірність і актуальність інформації на основі збігів фактів з іншими джерелами.

Потреба в оцінці актуальності та достовірності даних чи пошуку схожого контенту на основі заданих фактів може виникати, наприклад, під час археологічних досліджень, а саме під час пошуку нової інформації та при перевірці гіпотез.

Для вирішення цього пропонується підхід до фактографічного пошуку контенту з впровадженням ймовірнісних оцінок схожості, достовірності та актуальності виділених фактів, які впливатимуть на ранжування документів в пошуковій видачі та зберігатимуться в відповідній розподіленій базі даних.

Пропонується два способи пошуку – пошук схожих даних, де кожній одиниці буде даватись імовірнісна оцінка схожості (та достовірності) до факту із запиту та пошук актуальних даних, де на основі фактів, результати будуть видані в хронологічній послідовності з оцінкою актуальності. Сторінка результатів пошуку міститиме список знайдених співпадінь в контенті ресурсів, відображаючи навпроти кожного ймовірнісні оцінки, отримані шляхом обробки даних розробленими моделями.

Перевірка фактів здійснюватиметься на основі інформації всіх ресурсів системи з допомогою методів обробки природної мови та великих даних, використовуючи також дані відкритих джерел.

Разом з цим, при формуванні результатів пошуку, враховуватимуться додаткові несемантичні фактори (внутрішня оцінка авторитетності ресурсу, на якому розміщений документ, наявність мультимедійного вмісту та коментарів, внутрішньосистемний рейтинг документу, розповсюдження в соціальні мережі та інші), пріоритетність яких визначатиметься ціллю запиту та спільними переважаючими ознаками тематичної множини ресурсів.

Також, оскільки факти мають властивість втрачати актуальність чи спростовуватись, потрібно буде регулярно оновлювати їх сховище та здійснювати переоцінку на основі нової інформації. Оновлення вимагатимуть і оцінки авторитетності ресурсів (наприклад, у випадку відкликання свідоцтва про реєстрацію

конкретного ЗМІ), які є ключовим додатковим фактором при оцінюванні фактів.

Передбачається врахування в моделях обробки даних синонімічних, морфологічних особливостей та помилок написання слів при кластеризації і створення методів оцінки при відсутності компетентних джерел перевірки, способів уникнення впливу малозначних фактів та перешкоджаючих факторів при аномаліях даних.

Реалізація підходу передбачає розглядання та опрацювання наступних питань:

1. Визначення способів отримання інформації;
2. Формування вимог до сховищ даних;
3. Моделювання способів виділення фактів, розробка методів обробки, кластеризації, оцінки та перевірки актуальності фактів;
4. Ситуація відсутності компетентних джерел перевірки;
5. Виділення і зменшення пріоритезації малозначних фактів;
6. Способи зберігання оброблених даних;
7. Додаткові фактори ранжування фактів;
8. Отримання та ранжування результатів пошуку;
9. Розробка методів оновлення ймовірнісних оцінок;
10. Додаткові способи використання бази фактів.

Ситуація відсутності компетентних джерел перевірки факту може виникати тоді, коли ресурс, на якому він розміщується, є єдиним джерелом. В цьому випадку, застосування раніше запропонованого підходу, що базується на наявності декількох джерел, факти яких співставляються, буде неможливим через їх відсутність.

Для часткового вирішення цієї проблеми використовуватимуться додаткові несемантичні фактори ранжування, але вони не даватимуть об'єктивну ймовірнісну оцінку, оскільки для текстового аналізу буде використано тільки поточне джерело. Тому для оцінки достовірності використовуватимуться оцінки ситуативно схожих наявних фактів з поточним і здійснюватиметься прогностичне моделювання ситуацій.

Отримана таким чином оцінка достовірності та актуальності буде наближеною і частково дозволить збільшити об'єктивність оцінки за відсутності інших даних та може бути скорегована при появі більш точних джерел перевірки.

Оскільки достовірність виділених фактів є ймовірнісною оцінкою, що базується на основі знімку контенту ресурсів на час аналізу, який може бути зміненим та доповнюватиметься, то, як згадувалось вище, деякі з фактів можуть ставати неактуальними чи спростовуватись. Тому обов'язковим є регулярне оновлення оцінок та додаткових факторів ранжування контенту на основі оновленої інформації.

З технічної сторони, підхід вимагає не тільки ефективних алгоритмів і моделей обробки та валідації даних, а і великих обсягів обчислювальних та фізичних ресурсів, тому для розгортання системи доцільним буде використання хмарної інфраструктури.

Додатково, накопичена база фактів дозволить пришвидшити пошук інформації про знайдені археологічні об'єкти на основі широкого спектру даних про вже досліджені схожі одиниці. Також, використовуючи запропонований підхід, базу фактів, географічні дані та методи їх обробки з врахуванням історичних періодів, можна буде прогнозувати приблизні місця розташування майбутніх знахідок.

Література

1. Пасічник В. В., Кунанець Н. Е., Дуда О. М., Липак Г. І., О Мацюк. В., Семенюк В. В. Актори та діаграми прецедентів системи консолідації соціокомунікаційних інформаційних ресурсів "Розумних міст". Науковий вісник НЛТУ України. 2017. Вип. 27(10). С. 129–136.

2. Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. Proc. VLDB Endow., 8(9):938–949.