УДК 004.652.4+004.827

Н.Б.Шаховська, Р.Ю.Нога, О.Б.Вовк
Національний університет «Львівська політехніка», Львів, Україна

ІНТЕЛЕКТУАЛЬНА СИСТЕМА АНОТУВАННЯ НОВИН ДЛЯ ОЦІНЮВАННЯ ДОСТОВІРНОСТІ ЇХ ДЖЕРЕЛ

У статті описано елементи системи анотування новин для подальшої оцінки достовірності їх джерел. Розглянуто процес функціонування та алгоритми, що можуть використовуватися при реалізації такої системи.

Вступ

Сучасне інформаційне суспільство дає можливість без особливих зусиль (в основному) отримувати інформацію про актуальні події у світі. Численні ЗМІ надють їх своїй аудититорії через друковану пресу, радіо, телебачення та Інтернет. Останній канал розповсюдження є на сьогодні найперспективнішим (а для багатьох уже став основним), що не дивно, зважаючи на такі його переваги, як от поширення, гнучкість, зручність.

Тим не менше, існує ряд проблем, з якими зустрічається аудиторія при отриманні массової їнформації. До них належать:

- відбір актуальних (за тематикою і часом запиту) для користувача сюжетів з величезного об'єму доступних,
- усунення повторів сюжетів, які виникають при наданні відомостей про ту ж подію різними провайдерами масової інформації,
- фільтрування з отриманих новин тих, які в тій чи іншій мірі не відповідають дійсності (містять грубі неточності, спотворення фактів чи відверту брехню).

Остання проблема набирає особливої ваги в сфері сучасних інформаційних технологій, зокрема зважаючи на «unsolid» представлення інформації в Інтернеті. Вона пов'язана з попередньою, оскільки анотування передує і переплітається з існуючими стратегіями оцінки достовірності джерел.

Тому стаття присвячена побудові алгоритмів анотування та оцінювання достовірності новин, а також архітектури системи.

Аналіз останніх досліджень, публікацій та наявних рішень

Математичний апарат для оцінювання достовірності подій ґрунтується на баєсівських мережах. Проте, зважаючи на дослідження в [4], для більше ніж 7 параметрів необхідно використовувати додаткові евристики, що є надзвичайно складним завданням у цій предметній області, оскільки існує велика залежність отриманих евристик від тестового набору даних.

Проблема анотування новин на даний момент має ряд практичних вирішень. Більшість з них, щоправда, базуються на ручній праці фахівці, що

© Н.Б.Шаховська, Р.Ю.Нога, О.Б.Вовк

має свої недоліки з точки зору об'єктивності, швидкості, обсягу роботи та уніфікованості. Існують і сервіси, що використовують автоматичний підхід, наприклад Google News і Яндес.Новости. Останній, зокрема, навіть пропонує україномовну версію.

Проте, у широкому застосування немає сервісів, які б дозволяли проводити оцінку достовірності джерел чи окремих подій, хоча існують напрацювання, що підводять до цього

Основний матеріал

Алгоритм агрегування новин

Процедура агрегування потоку новин може включати такі етапи (на основі процедури Яндес.Новости):

1) Завантаження новин через PCC чи інші канали;
2) Сегментація повідомлень (виділення заголовка, опису, основного тексту, картинок, відео і т. п.);
3) Виділененя сюжетів – віднесення новини до тієї чи іншої події (кластеризація повідомлень на основе аналізу їх текстів);
4) Анотування сюжетів (подання в скороченому вигляді основного змісту сюжетів);
5) Виділення в межах сюжету повідомлень, що його підтверджують або спростовують.

Завантаження та сегментація (за структурованого представлення повідомлень, наприклад, XML/RSS) представляє собою нескладну технічну задачу.

Виділененя сюжетів можна здійнити за допомогою того чи іншого алгоритму кластеризації, наприклад, K-means.

Аннотування – створення короткої версії деякого тексту чи множини текстів. Створення аннотаціїлюдиною – поширена задача. Виділють задачу анотування одного та декількох документів. Перший випадок – це коротке представлення основного змісту даного документу. Другий - виявлення різних документів на задану тему з врахуванням часового фактора, тобто деякі документи можу втратити свою актуальність. Задача анотування новинних сюжетів є якраз задачею багатодокументного анотування з врахуванням часового фактору.

Алгоритм ймовірності появи новин

Імовірнісна модель появи пари «тема-слово»:

$$p(d,w) = \sum_{s \in S} p(s)p(w \mid s)p(d \mid s) = \sum_{s \in S} p(s)p(w \mid s)p(s \mid d) = \sum_{s \in S} p(w)p(s \mid w)p(d \mid s)$$

де S – множина тем; $p(s)$ – невідомий апріорний розподіл тем у всій колекції; $p(d)$ – апріорний розподіл на множині документів, емпірична оцінка $p(d) = n_d / n$, де $n = \sum_d n_d$ – сумарна довжина всіх документів; $p(w)$ –

апріорний розподіл на множині слів, емпірична оцінка $p(w) = n_w / n$, де n_w – число входжень слова *w* у всі документи; шукані ймовірності розподілу *p* (*w*|*s*), *p* (*s*|*d*) виражаються через *p*(*s*|*w*), *p*(*d*|*s*) за формулою Байєса:

$$p(w \mid s) = \frac{p(s \mid w)\, p(w)}{\sum_{w'} p(s \mid w')\, p(w')}; \; p(s \mid d) = \frac{p(d \mid s)\, p(s)}{\sum_{s'} p(d \mid s')\, p(s')}.$$

Для ідентифікації параметрів тематичної моделі (теми новини) за колекцією документів (джерел даних) застосовується принцип максимуму правдоподібності, який призводить до задачі мінімізації

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log p(d, w) \to \min_{\Phi, \Theta} \text{ за обмежень}$$

$$\sum_{w} p(w \mid s) = 1, \sum_{s} p(s \mid d) = 1, \sum_{s} p(s) = 1$$

де n_{dw} – число входжень слова *w* в документ *d*.

Для вирішення даної оптимізаційної задачі застосовано Expectation-maximization алгоритм (ЕМ-алгоритм).

Оцінювання достовірності новин

Наступним кроком є оцінювання достовірності отриманих новин. Один з її варіантів може базуватися на моделі, запропонованій в [1]. Модель передбачає надання користувачем новин А двох оцінок провайдеру В:

$t_{AB} \in (0; 1)$ – оцінка довіри до джерела, та

$u_{AB} \in (0; 1]$ – невпевненість в оцінці довіра.

Початковими значеннями приймаються $t_{AB} = 0.5$ та $u_{AB} = 1$, тобто 50%-ва довіра до джерела з максимальною можливою невпевненістю.

Для уточнення значень оцінок провайдера В користувач А вибирає з наданих йому новин (у кількості n_B) ті, про які він може сказати з деякою мірою впевненості, чи правдиві вони (їх кількість позначимо mB). Нехай серед них є r_B правдивих і s_B неправдивих, $r_B + s_B = m_B$. Тоді частота неправдивих новин у В складе $r_B / (r_B + s_B)$. Далі у [1] визначають t_{AB} як очікуване значення бета-розподілу з параметрами $\alpha = r_B^* + 1$, $\beta = s_B^* + 1$, де r_B^* та s_B^* представляють усереднення попередньо зібраних значень r_B та s_B, здійснене з урахуванням так званого фактору старіння [2], а u_{AB} – як нормалізоване значення дисперсії цього ж розподілу.

Запропонуємо інший спосіб і використаємо модель довіри, в якій є пара незалежних коефіцієнтів підтвердження (k1) та спростування (k2) деякої події ($k_1, k_2 \in [0; 1]$) за інформацією різних провайдерів. Початкові значення приймаються як $k_1 = k_2 = 0$. Далі для і-го провайдерів, що описують подію, визначається коефіцієнт підтверження чи спростування події w_i, і коефіцієнти перераховуються: якщо подія підтверджується, то $k_1 = k_1 + W_i(1-k_1)$, інакше $k_2 = k_2 + W_i(1-k_2)$. Результуючий коефіцієнт довіри k розраховується різниця коефіцієнтів k_1, k_2: $k = k_1 - k_2$, $k \in [-1; 1]$. Перевага цього методу у відсутності

необхідності безпосередньої оцінки користувачем правдивості частини новин.

Розроблення архітектури системи

На рис. 1 зображена діаграма діяльності основного бізнес-процесу поведінки користувача.

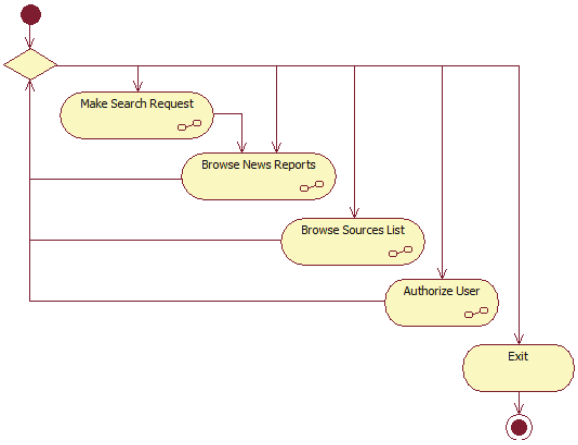

Рис 1. Діаграма діяльності основного бізнес-процесу поведінки користувача

Пояснимо можливі дій користувача системи:

Make Search Request – відправлення запиту – користувач може задати обмеження, які теми новин його у даний момент цікавлять.

Browse News Reports – перегляд новин – користувач може переглядати запропоновані системою теми або подані як відповідь на його запит, та вказувати свій ступінь довіри до тих чи інших новин.

Browse Sources List – перегляд списку джерел – користувач може переглядати список джерел новин системи і вказувати свій ступінь довіри до них.

Authorize User – авторизація – користувач може зареєструватися, залогуватися або працювати анонімно.

Діаграма перегляду новин зображена на рис. 2. Поянимо її дії.

View Source List – безпосередньо перегляд списку тем новин.

View Source Details – перегляд детальної інформації по конкретній темі, а саме всіх повідомлень, що належать до неї. Ця дія відкриває доступ до наступних.

Discuss Source – прочитати наявні джерела або залишити свої коментарі щодо теми.

Share Source – поділитися посиланням на тему через соціальні мережі або електронним листом.

Go to Source Webpage – перейти на сторінку джерела новин, що містить тему.

Authorize User – авторизувати користувача.

Rate Materials – авторизовані користувачі можуть ставити оцінки довіри до окремих матеріалів.

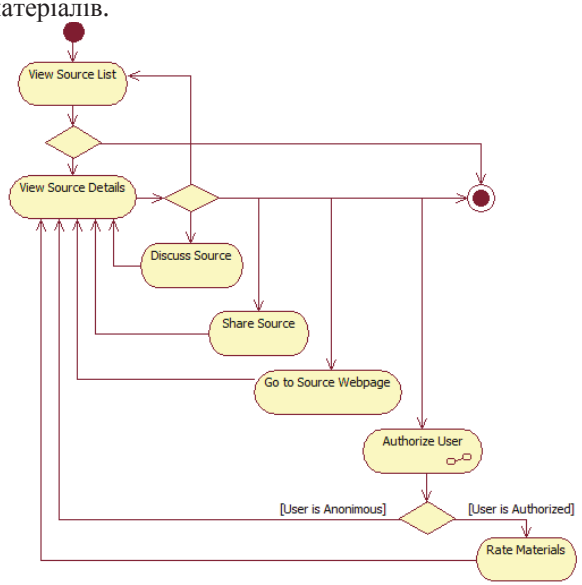

Рис 2. Діаграма діяльності перегляду новин

Висновки

Розроблено архітектуру системи анотування інформаційних повідомлень від різних постачальників, а також оцінювання надійності новин та довіри до окремих ЗМІ. Це дасть змогу користувачам краще і швидше орієнтуватися в доступних ЗМІ та новинах, які їх цікавлять.

1. *Shahraev A*, "Avtomatycheskoe annotyrovanye novostnoho potoka." [Automatic annotation of news flow], Natalia Ostapuk's page on slideshare.net, 29 Nov 2011. [Online]. Available: SlideShare.net, http://www.slideshare.net/NataliaOstapuk/ss-10380447 [Accessed: 1 Oct 2013]

2. *A. Korshunov, A. Homzyn*, "Tematycheskoe modelyrovanye tekstov na estestvennom yazyke" [Topical modeling of natural language texts], Trudy ISP RAN [Proceedings of ISP RAS], vol. 23, 2012. Available: Open Access Library "KyberLenynka", http://cyberleninka.ru/article/n/tematicheskoe-modelirovanie-tekstov-na-estestvennom-yazyke [Accessed: 1 Oct 2013]

3. *E. Staab, V. Fusenig and T. Engel*, "Towards Trust-Based Acquisition of Unverifiable Information", in Proc. 12th International Workshop, CIA 2008, Prague, Czech Repb.: Springer, vol. 5180, pp. 41-54

4. *Бідюк, П.І.* Байєсівські мережі в технологіях інтелектуального аналізу даних / П.І.Бідюк, О.М. Терентьєв, М.М. Коновалюк // Искусственный интеллект. – 2010. – N2. – С.104-113.

Поступила 3.03.2014р.