УДК 002.513.5

О.О. Шумейко, Г.Я. Шевченко

РАНЖУВАННЯ ДОКУМЕНТІВ ЗА ІНФОРМАЦІЙНИМ ЗАПИТОМ

Анотація. Запропоновано метод побудови текстового ранжирування наявного корпусу документів у відповідності з інформаційним внеском в них складових інформаційного запиту.

Ключові слова: ранжування документів, інформаційний пошук, ентропія.

Вступ

Під терміном ранжирування, як правило, розуміють процес вибірки пошуковою машиною документів з бази даних і впорядкування їх за ступенем відповідності з інформаційним (пошуковим) запитом [1-3]. Існує поняття текстового ранжирування і ранжирування за гіпертекстовими посиланнями [2, 3]. У першому випадку пошуковою машиною враховуються такі чинники як, наприклад, щільність ключових фраз в текстах статей (документів), оформлення заголовків [3, 4]. При ранжируванні за гіпертекстовими посиланнями беруться до уваги посилання на сайт з інших ресурсів.

В основі більшості алгоритмів ранжирування документів лежить ідея TF*IDF (див. [4]), яка використовує частоту використання ключових слів в поточному документі та в інших документах корпусу. Окрім того, серед використаних параметрів-

- виділення ключових слів тегами і їх відстань до початку документа;
- довжина документа;
- число пар слів, які йдуть підряд в запиті і в такому ж вигляді зустрічаються в тексті;
- число ключових слів із запиту які взагалі зустрічаються в тексті;
- чи зустрічається весь запит в тексті та ін.

Існуючі методи ранжирування документів досить механістичні і часто спираються на формальний вигляд документів, які підлягають ранжируванню. У разі, коли документ представлений малою кількіс-

 ISSN 1562-9945

тю ключових слів і відсутня можливість повнотекстового пошуку, використовуються модифікації TF*IDF, PCA та інше.(див. [2-3])

У даній роботі запропоновано метод побудови текстового ранжирування наявного корпусу документів у відповідності з інформаційним внеском в них складових інформаційного запиту.

Використання інформаційної ентропії для ранжування документів

Нехай наявний корпус документів, кожен з яких визначений частотним словником словоформ, які входять до нього $D_k = \left\{ w_i^k : n_i^k \right\} (k = 1, \ldots, N)$, а через $S = \{s_i\}$ позначимо інформаційний запит.

Потрібно провести ранжирування корпусу документів $\{D_k\}\big|_{k=1}^{N}$ у відповідності із інформаційним запитом S. В основі запропонованого методу ранжирування лежить ідея використання зміни значення ентропії при об'єднанні документів. Зазначимо, що такого роду конструкції використовуються при розв'язку оптимізаційних задач теорії інформації, наприклад, при побудові дерев рішень C4.5 та ін. (див. [5]).

У подальшому нам будуть потрібні наступні поняття:

В якості міри невизначеності випадкового об'єкту (системи) A зі скінченною множиною можливих станів $A_1, A_2, \ldots, A_n$ та відповідною імовірністю $p_1, p_2, \ldots, p_n$, Клод Шеннон запропонував використовувати функціонал названий ентропією.

$$H(A) = H(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log p_k,$$

Логарифми беруться при довільній основі, але у випадку, якщо за одиницю вимірювання ступеня невизначеності прийняти невизначеність, що міститься в досліді з двома рівно імовірними результатами (наприклад, наявний елемент в деякій множині або відсутній), то слід брати основу яка дорівнює двом. Зазначимо, що при заданому n величина ентропії максимальна і дорівнює $\log n$ лише у випадку, коли всі p_i рівні між собою, тобто

$$p_1 = p_2 = \ldots = p_n = \frac{1}{n}.$$

Таким чином, $H(D_k) = -\sum_{i=1}^{N_k} \frac{n_i^k}{N_k} \log_2 \frac{n_i^k}{N_k}$, де N_k – загальне число словоформ у документі D_k, а $n_i^k = num(w_i^k)$ – число входжень словоформи w_i^k у даний поточний документ ($num(s)$ - число входжень слова s).

Для двійкового випадку, у разі, коли серед n станів системи A наявні m, які мають деяку властивість V, то ентропія по відношенню до властивості V буде дорівнювати

$$H(A,V) = -\frac{m}{n}\log_2\frac{m}{n} - \frac{n-m}{n}\log_2\frac{n-m}{n}.$$

Якщо використовувати якийсь атрибут Q, який має q значень, то необхідно визначити приріст інформації, що вимірює очікуваний рівень ентропії (різницю між інформацією з A та інформації, що необхідна для визначення елементу з A після того, як значення атрибуту Q було визначене, тобто, приріст інформації завдяки атрибуту Q):

$$G(A,Q) = H(A) - \sum_{j=1}^{q} \frac{|A_j|}{|A|} H(A_j, V).$$

де A_j - множина станів A, для яких атрибут Q приймає i-те значення, а |X|- число елементів множини X.

Для нашого випадку, величина ентропії документу D_k відносно слова s_i з інформаційного запиту S буде дорівнювати

$$H(D_k, s_i) = -\frac{num(s_i)}{N_k}\log_2\frac{num(s_i)}{N_k} - \frac{N_k - num(s_i)}{N_k}\log_2\frac{N_k - num(s_i)}{N_k}.$$

Величина приросту ентропії буде дорівнювати

$$H(D_k, S) = H(D_k) - \sum\left\{\frac{num(s_i)}{N_k} H(D_k, s_i) \middle| s_i \in S\right\}.$$

Чим більше буде значення приросту ентропії, тим більше наш документ буде відрізнятися від інформаційного запиту.

З іншої сторони, значення ентропії залежить від кількості станів системи (у нашому випадку від кількості словоформ, які описують документ), тому для ранжирування нам треба визначення не абсолютного значення зміни значення ентропії, а відносного, тобто

$$\bar{H}(D_k, S) = \frac{H(D_k) - \sum\left\{\frac{num(s_i)}{N_k} H(D_k, s_i) \middle| s_i \in S\right\}}{H(D_k)},$$

яке дозволяє оцінити зменшення рівня ентропії документу, якщо відома інформація щодо ключових слів (складових інформаційного пошуку). Значення $\bar{H}(D_k, S) = 1$ вказує на той факт, що даний документ D_k ніякого відношення до даного інформаційного пошуку не має, тобто інформація про $s_i \in S$ не змінює загальній обсяг інформації що до D_k. І чим менше значення $\bar{H}(D_k, S)$, тим менша ступінь невизначеності D_k відносно S.

Розглянемо приклад.

№	Документ	Опис документу
1	Васильев Ф.П. "Методы оптимизации"	оптимизация, функция, минимизация, дифференциальные уравнения, численные методы
2	Корнейчук Н.П., Лигун А.А., Доронин В.Г. "Аппроксимация с ограничениями"	аппроксимация, неравенство, приближения, сплайн
3	Лоран П. - Ж. "Аппроксимация и оптимизация"	сплайн, аппроксимация, интерполяция, экстраполяция, оптимизация
4	Самарский А.А., Гулин А.В. "Численные методы математической физики"	аппроксимация, разности, дифференциальные уравнения
5	Лебедев П.Д., Ушаков А.В. "Аппроксимация множеств на плоскости оптимальными наборами кругов "	сеть, круг, аппроксимация, кривая, многоугольник
6	Бляшке В. "Круг и шар"	круг, шар, минимизация, симметрия
7	Леонтьев В. "Экономические эссе"	круг, интерес, экономика, политика
8	Смит Р.С, Эренберг Р.Дж. Современная экономика труда.	труд, политика, экономика

Додаємо до опису документу словоформи з назви документу і отримаємо для кожного документу частотний словник. Після обчислення значення ентропії отримаємо наступні дані

№	Частотний словник	Ентропія $\left(H\left(D_k\right)\right)$
1	оптимизация-2, функция-1, минимизация-1, дифференциальные уравнения-1, численные методы-1	2,251629
2	аппроксимация-2, неравенство-1, приближения-1, сплайн-1, ограничения-1	2,251629
3	сплайн-1, аппроксимация-2, интерполяция-1, экстраполяция-1, оптимизация-2	2,235926
4	аппроксимация-1, разности-1, дифференциальные уравнения-1, численные методы-1, математическая физика-1	2,321928
5	сеть-1, круг-2, аппроксимация-2, кривая-1, многоугольник-1, плоскость-1	2,500000
6	круг-2, шар-2, минимизация-1, симметрия-1	1,918296
7	круг-1, интерес-1, экономика-2, политика-1	1,921928
8	труд-2, политика-1, экономика-2, современность-1	1,918296

В якості прикладу інформаційного запиту візьмемо текстовий рядок «аппроксимация круговыми сплайнами», який після перетворення в словоформи буде мати вигляд «аппроксимация, круг, сплайн».

Наступним кроком знайдемо кількість інформації, що необхідна для визначення елементу з поточного документу, якщо відоме слово (словоформа) з інформаційного пошуку.

№	аппроксимация	круг	сплайн
1	0,000000	0,000000	0,000000
2	0,918296	0,000000	0,650022
3	0,863121	0,000000	0,591673
4	0,721928	0,000000	0,000000
5	0,811278	0,811278	0,000000
6	0,000000	0,918296	0,000000
7	0,000000	0,721928	0,000000
8	0,000000	0,000000	0,000000

 ISSN 1562-9945

Значення 0 показує, що дане слово в описі документу відсутнє, тому ніяким чином не впливає на співвідношення поточного документу з даним словом з інформаційного пошуку.

Далі знайдемо загальний обсяг інформації, необхідної для визначення елементу з поточного документу, по всій множині складових інформаційного пошуку, потім, обчислимо абсолютне значення зміни ентропії за умовою наявності інформації що до складових інформаційного пошуку, і, нарешті, значення відносної зміни рівня ентропії і за отриманими значеннями маємо рейтинг документів відносно даного інформаційного пошуку

№	$\sum\left\{\frac{num(s_i)}{N_k}H(D_k,s_i)\middle\vert s_i\in S\right\}$	$H(D_k,S)$	$\bar{H}(D_k,S)$	Рейтинг
1	0,000000	2,25163	1,000000	7
2	0,414436	1,83719	0,815940	1
3	0,331131	1,9048	0,851905	4
4	0,144386	2,17754	0,937816	6
5	0,405639	2,09436	0,837744	2
6	0,306099	1,6122	0,840432	3
7	0,144386	1,77754	0,924875	5
8	0,000000	1,9183	1,000000	8

Зазначимо, що так як даний алгоритм не несе семантичної складової, то, на жаль ніяким чином не враховується загальна спрямованість документу, тому має сенс зроби деяке узагальнення алгоритму, формуючи новий запит на основі отриманої інформації.

Так як у результаті ми маємо значення вкладу слів інформаційного пошуку в кожен документ D_k, то, можна вважати, що і інші слова даного документу пов'язані з словоформами інформаційного пошуку. Таким чином, вага кожного слова документа (відносно словоформ інформаційного пошуку) може бути розрахована наступним чином

$$\bar{W}(D_k,S)=\frac{\sum\left\{\frac{num(s_i)}{N_k}H(D_k,s_i)\middle\vert s_i\in S\right\}}{H(D_k)},$$

окрім оригінальних словоформ інформаційного пошуку, вага яких дорівнює 1, тобто

$$\bar{W}(D_k, s_i) = \begin{cases} \bar{W}(D_k, S), & s_i \notin S, \\ 1, & s_i \in S. \end{cases}$$

№	1	2	3	4	5	6	7	8
$\bar{W}(D_k, S)$	0	0,18406	0,1481	0,06218	0,16226	0,15957	0,07513	0

Загальне вагове значення для кожного слова з нового інформаційного запиту обчислимо наступним чином

$$\bar{W}(s_i) = \begin{cases} \dfrac{\sum\left\{\bar{W}(D_k, S) \middle| s_i \in D_k\right\}}{\sum\left\{1 \middle| s_i \in D_k\right\}}, & s_i \notin S, \\ 1, & s_i \in S. \end{cases}$$

Сформуємо новий інформаційний пошук, який буде складатися з усіх слів, які є в документах корпусу, окрім тих, що мають нульовий рейтинг і знайдемо відносний приріст ентропії з урахуванням ваги

$$\bar{H}(D_k, S) = \frac{H(D_k) - \sum\left\{\bar{W}(s_i)\dfrac{num(s_i)}{N_k} H(D_k, s_i) \middle| s_i \in S\right\}}{H(D_k)},$$

і вже за даною величиною визначимо рейтинг документів вже з урахуванням всієї інформації, яка входить в корпус і пов'язана з інформаційним пошуком.

№	$\sum\left\{\bar{W}(s_i)\dfrac{num(s_i)}{N_k} H(D_k, s_i) \middle| s_i \in S\right\}$	$\bar{H}(D_k, S)$	Рейтинг
1	0,0693569	0,9692	7
2	0,4742572	0,78937	2
3	0,3926891	0,82437	4
4	0,1802972	0,92235	6
5	0,5372771	0,78509	1
6	0,3895175	0,79695	3
7	0,1952600	0,8984	5
8	0,0311366	0,98377	8

Таким чином, рейтинг документів змінився з урахуванням словоформ розширеного інформаційного пошуку.

Можна, взагалі, зняти пріоритет словоформ інформаційного запиту і прирівняти права запиту з правами документу. В такому разі ми вважаємо, що на першому кроці мі вказали на пріоритети пошуку, а на другому дозволяємо підправити пошук з урахуванням всіх документів, тобто

$$\bar{W}(s_i) = \frac{\sum\left\{\bar{W}(D_k, S)\middle| s_i \in D_k\right\}}{\sum\left\{1\middle| s_i \in D_k\right\}}$$

для всіх словоформ. В такому випадку маємо

№	$\sum\left\{\overline{W}(s_i)\frac{num(s_i)}{N_k}H(D_k, s_i)\middle| s_i \in S\right\}$	$\overline{H}(D_k, S)$	Рейтинг
1	0,069357	0,9692	7
2	0,202380	0,91012	3
3	0,175197	0,92164	4
4	0,080671	0,96526	6
5	0,265499	0,8938	1
6	0,190553	0,90067	2
7	0,101409	0,94724	5
8	0,031137	0,98377	8

Висновок

Отримані результати дозволяють ефективно організувати текстове ранжирування невеликих за кількістю множин документів у відповідності з інформаційним (пошуковим) запитом.

ЛІТЕРАТУРА

1. Шумейко А.А. Интеллектуальный анализ данных / А.А.Шумейко, С.Л.Сотник .— Днепропетровск: Белая, 2012 .— 212 с . – Режим доступу: http://pzs.dstu.dp.ua/ Data/dm.pdf
2. Маннинг К. Введение в информационный поиск / К.Маннинг, П.Рагхаван, Х.Шютце .— М: Вильямс, 2014 .— 528 с.
3. Шокин Ю.И. Проблемы поиска информации / Ю.И.Шокин, А.М.Федотов, В.Б.Барахнин. Новосибирск: Наука, 2010 .— 220 с . – Режим доступу: http://elib.sbras.ru:8080/jspui/ bitstream/SBRAS/7176/ 1/search.pdf
4. Гулин А. Алгоритм текстового ранжирования Яндекса на РОМИП-2006 // А.Гулин, М.Маслов, И.Сегалович . – Режим доступу: http://cache-kiev10.cdn.yandex.net/ download.yandex.ru/ company/03_yandex.pdf
5. Сегаран Т. Программируем коллективный разум / Т.Сегаран .— СПб: Символ-Плюс, 2008 .— 368 с.
6. Михалев А. Компьютерные методы интеллектуальной обраотки данных: учебное пособие / А.Михалев, Е.А.Винокурова, С.Л.Сотник .— Днепропетровск: НМетАУ, ИК "Системные технологии", 2014 .— 209 с. Режим доступу: http://sotnyk.com/Articles/cmidpbook_20140818.pdf
7. Дюк В.А. Data Mining. Учебный курс / В.А.Дюк, А.П.Самойленко .— Сп.б.: Питер, 2002 .— 368 с.