УДК 004.67

Д.В. Ланде
Інститут проблем реєстрації інформації НАН України

ПІДХОДИ ДО АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ТЕРМІНОЛОГІЧНИХ ОСНОВ ОНТОЛОГІЙ

На даний час актуальним залишається питання визначення важливих структурних елементів тексту, що виявляються *інформаційно-значущими*, визначають інформаційну структуру. Використання таких елементів дозволяє формувати тезауруси, пошукові образи документів, онтології.

Ключові слова для пошуку в тексті, опорні слова для автоматичного екстрагування значущих фрагментів текстів або формування автоматичних рефератів, вибираються з урахуванням такої властивості слів, як «*дискримінантна сила*», для визначення яких існує три класичних метода – *TFIDF*, *дисперсійний*, *ентропійний*. Також застосовуються мережеві методи, що, зокрема, базуються на побудові *графів горизонтальної видимості* і *мереж природних ієрархій термінів*.

Показник *TFIDF* [1] для кожного терміну з текстового корпусу у канонічному виді дорівнює добутку частоти слова у фрагменті тексту (Term Frequency) на логарифм від величини, зворотної кількості окремих фрагментів тексту, в яких це слово зустрілось (Inverse Document Frequency). Для кожного слова i, що входить до текстового корпусу, який складається з N документів, підраховується кількість документів $df(i)$, в яких міститься це слово, а також загальна частота входження даного слова i у текстовий корпус – $n(i)$. Після цього розраховується середнє значення TFIDF для кожного слова:

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{df(i)}\right).$$

Алгоритм *TFIDF* дає не зовсім коректні результати на текстових масивах із документів з різною за довжиною. Тому

запропоновано його модифікацію – *Okapi BM25*, де на відміну від звичайного *TFIDF* береться до уваги довжина документа:

$$TF \cdot IDF\left(w_i\right) = IDF\left(w_i\right) \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

де $f\left(w_i, D\right)$ – частота слова w_i у документі D; $\left|D\right|$ – довжина документа D (число слів); *avgdl* – середня довжина документа в колекції; k_1 і b – параметри, що вибираються експертно. $IDF\left(w_i\right)$ обчислюється за формулою:

$$IDF(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5},$$

де N – загальна кількість документів у масиві, $n(w_i)$ – кількість документів, що містять термін w_i.

Дисперсійна оцінка важливості термінів обчислюється для окремих термінів наступним чином. Позначимо середній інтервал між появами слова A у тексті через $<\Delta A>$, а середній квадрат значень цих інтервалів через $<\Delta A^2>$. Дисперсійна оцінка слова σ_A розраховується як

$$\sigma_A = \frac{\sqrt{\left\langle \Delta A^2 \right\rangle - \left\langle \Delta A \right\rangle^2}}{\left\langle \Delta A \right\rangle}.$$

Ідея дисперсійної оцінки близька до TFIDF, однак більш коректно застосовується до цілих текстів, а не до масивів з великої кількості документів, як TFIDF.

У теорії інформації, якщо ймовірність події a дорівнює $p(a)$, то кількість інформації пропорційна $-\log_2 p(a)$. Якщо розглядати терміни як події, можемо знайти кількість інформації, що відповідає терміну w, тобто оцінку ваги терміну (*Pointwise Mutual Information, PMI*) [2]:

$$m(w) = -\log_2 \frac{p(w \mid d)}{p(w)} = -\log_2 \frac{n_{w,d}}{n_w},$$

де $n_{w,d}$ – кількість входжень терміну w до тексту документа, n_w – загальна кількість входжень.

Ряди з цифрових значень, що відповідають словам із текстів, тексти як послідовності термінів можуть перетворюватися у *графи горизонтальної видимості* (*Horizontal Vizibility Graph*) [3], вузлам яких відповідають не тільки цифрові значення, але й слова, що несуть певне змістовне значення. Мережа слів за цим алгоритмом будується у три етапи. *На першому етапі* на осі абсцис за номерами входжень у текст відмічається ряд вузлів, кожний з яких відповідає слову. По осі ординат відкладаються вагові оцінки (наприклад, значення *TFIDF*). *На другому етапі* будується традиційний граф горизонтальної видимості. При цьому між вузлами визначається зв'язок, якщо вони знаходяться у «прямій видимості», тобто їх неможливо поєднати горизонтальною лінією, що не перетинає ніяку вертикальну лінію. На *третьому етапі* мережа компактифікується, тобто всі вузли, що відповідають тим самим словам поєднуються у один вузол. Всі зв'язки також поєднуються. Після побудови відповідної мережі слів новою оцінкою слів вважається ступень вузлів, що відповідають цим словам.

Для уточнення вагових значень термінів широко застосовуються мережеві підходи, зокрема автором запропоновано розгляд *мережі природних ієрархій термінів* [4]. У цьому випадку вага термінів (окремих слів або словосполучень) визначається як вагові коефіцієнти значень вузлів цієї мережі. При побудові мережі природних ієрархій термінів вибирається відповідна кількість найбільш вагомих за наведеними вище алгоритмами (зокрема, *HVG*) уніграм (одиничних слів), біграм і триграм. З відібраних термінів будується мережа природних ієрархій термінів, у якої як вузли розглядаються самі терміни, а зв'язки відповідають входженням одних термінів у інші. Якщо одиничне слово входить до біграми або триграми, або біграма входить до тригами, утворюється зв'язок, який позначається стрілкою. Множина вузлів, яким

відповідають терміни, і зв'язків і утворює трирівневу мережу природної ієрархії термінів.

Для вирішення задачі визначення ваги термінів використовується алгоритм ранжирування *HITS* (Hyperlink Induced Topic Search), запропонований Дж. Клейнбергом [5], що забезпечує вибір з мережі кращих «авторів» (вузлів, на які введуть посилання) і «посередників» (вузлів, від яких йдуть посилання). Для кожного вузла мережі v_j рекурсивно обчислюється його вага як автора $a(v_j)$ і посередника $h(v_j)$ за формулами:

$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i).$$

Найцікавішими з семантичного погляду у цій мережі виявляються вузли з найбільшими значеннями авторства і посередництва.

Література

1. *Salton G., McGill M.J.* Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.
2. *Church K.W., Hanks P.* Word association norms, mutual information, and lexicography // Comput. Linguist, 1990. – № 16 (1). – P. 22–29.
3. *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Physical Review E, 2009. – P. 046103-1 – 046103-11.
4. *Ландэ Д.В., Снарский А.А.* Подход к созданию терминологических онтологий // Онтология проектирования, 2014. – № 2(12). – С. 83-91.
5. *Kleinberg J.* Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.