УДК 004.853

МОДИФІКОВАНИЙ МЕТОД ПОШУКУ РЕЛЕВАНТНИХ ВЕБ-ДОКУМЕНТІВ З ВИКОРИСТАННЯМ ПОДВІЙНОГО ЗВАЖУВАННЯ

Чала Лариса, Харитонова Юлія

Харківський національний університет радіоелектроніки

Анотація

У доповіді запропоновано удосконалений метод пошуку релевантних веб-документів за допомогою модифікованого частотного критерію. Метод має на меті розширення вхідного запиту користувача синонімічними векторами з повторним зважуванням. Здійснено програмну реалізацію методу з використанням програмних засобів Ruby. В ході тестування розроблений метод показав суттєве поліпшення результатів за показниками точності з невеликою втратою у показниках повноти.

Abstract

The paper proposed an improved method of finding relevant Web documents using modified frequency criterion. The method is aimed at expanding the user's query input vectors synonymous with re-weighting. Done software implementation of the method using the software Ruby. During the testing of the developed method showed a significant improvement in results for performance accuracy with a small loss in terms of completeness.

Вступ

На сьогодні інформаційний пошук застосовується в різних прикладних галузях – від систем баз даних до веб-інформаційних пошукових систем. Мета такого пошуку полягає у знаходженні документів, що є релевантними запитам користувачів [1]. Є доцільним розглянути доцільність розробки та практичної реалізації методу інформаційного пошуку за модифікованим частотним критерієм, який окрім врахування релевантності слова враховував би також його семантичну вагу, покращуючи тим самим якість пошукового запиту. Це дасть змогу отримувати релевантні дані навіть у тому випадку, коли більшість слів запиту не містяться у корпусі, незважаючи на семантичну подібність між корпусом та запитом. У доповіді розглядається вирішення такої задачі.

Модифікований метод пошуку

Функцію частотного зважування термів *Idf* в інформаційних пошуових системах використовують, зазвичай, як складову функції *Tf-Idf*. Частота появи терма, що є зворотною частотою документа (*td*idf*-модель), використовується для обчислення ваги d_i для терма i в документі:

$$d_i = tf_i \cdot idf_i, \tag{1}$$

де tf_i є частотою появи терма i в документі, а idf_i є оберненою частотою появи терма i в усьому корпусі документів.

Модифікуємо формулу (1) для запитів для надання більшого рівня виразності термам у запитах.

Рівень подібності між запитом q і документом d згідно з моделлю векторного простору (VSM) визначається як косинус внутрішнього добутку між векторними представленнями документів:

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \sqrt{\sum_{i=1}^{n} d_i^2}}, \quad (2)$$

де q_i та d_i – відповідно вектори ваг запиту і документа.

Всі документи ранжуються відповідно до їх подібності введеному запиту. Відсутність спільних термінів у двох документах не обов'язково означає, що документи не є схожими семантично. Аналогічно, релевантні введеному запиту документи можуть не містити такі терміни. Семантично близькі поняття можуть бути виражені завдяким використанню різних слів у документах і запитів, що робить пряме порівняння за словами на основі VSM-моделі неефективним або взагалі неможливим. Запропонований у роботі метод надає можливість знаходження семантично подібних документів з використанням засобів WordNet та семантичних критеріїв подібності [2].

Запропонований метод передбачає реалізацію трьох етапів. На першому етапі здійснюється повторне зважування терма з урахуванням (2): вага q_i кожного терма i запиту коригується на основі його зв'язку з іншими семантично подібними термами j в межах одного вектора:

$$q_i = q_i + \sum_{sim(i,j)>t, i \neq j} sim(i,j),$$

де t – пороговий коефіцієнт, що задається користувачем.

Другим кроком роботи запропонованого алгоритму пошуку релевантних веб-документів є розширення терма. При цьому запит доповнюється синонімічними термами, а далі він доповнюється гіпонімами і гіпернімами, які є семантично подібними термам запиту. Кожний елемент користувацького запиту представляється деревовидною ієрархією WordNet.

На другому етапі здійснюється розширення сукупності термів за рахунок додаткових термів з подібністю, що перевищує порогове значення T. Кожному терму розширеної сукупності присвоюється вага згідно з наступним виразом:

$$q_i = \begin{cases} \sum_{sim(i,j)>T, i \neq j} sim(i,j) & \text{якщо } i \text{ – синонім} \\ q_i + \sum_{sim(i,j)>T, i \neq j} sim(i,j) & \text{якщо } i \text{ – гіпонім або гіпернім } q_i \end{cases}, \quad (3)$$

де n – кількість гіпонімів для кожного розкритого терма j, що входить до запиту. Можливий випадок, коли один терм представляє декілька термів, які вже існували в запиті на момент його виконання. Крім того, можлива поява протилежної ситуації, коли один і той же терм представляється більше декількома термами. Умова (3) передбачає прийняття до уваги ваги оригінальних термів запиту і те, що частка кожного терма у присвоєнні ваги термам запиту нормалізується кількістю його гіпонімів n.

Останнім кроком роботи алгоритму є визначення рівня подібності документів. Подібність між розширеним і повторно зваженим запитом q і документом d обчислюється як:

$$S(q,d) = \frac{\sum_i \sum_j q_i d_j s(i,j)}{\sum_i \sum_j q_i d_j} \quad (1)$$

де i і j – відповідно терми у запиті та документі.

Терми в запиті розширюються і повторно зважуються відповідно до попередніх кроків, в той час як терми документа d_j обчислюються як $tf_i \cdot idf_i$ - терми (вони не є ані розширеними, ані повторно зваженими). При цьому міра подібності нормується в діапазоні [0,1]. Розширення запиту пороговим значенням T вводить нові терми залежно від позиції термів у таксономії: більш конкретні терми у таксономії (нижчі в ієрархії) умови в залежності також від положення доданків в таксономії: Розширення і повторне зважування показує високу швидкість при обробці запитів (у більшості випадків запит містить лише декілька термів), але не для документів, які складаються з багатьох термів. Запропонований метод передбачає лише розширення запиту. Проте, функція подібності також враховує зв'язки між усіма семантично подібними термами в документі і в запиті (що не може бути забезпечено «чистою» VSM-моделлю).

Наведений алгоритм було реалізовано програмно з використанням мови програмування Ruby. Алгоритмічну процедуро реалізовано у вигляді консольної програми, що обробляє вхідні текстові дані і повертає текстовий документ, релевантний запиту користувача.

Оскільки обробляються вхідні веб-документи (гіпертекстові), то необхідна інформація з них отримується за допомогою XML-парсера. Одним з етапів обробки вхідного тексту є його нормалізація, тобто приведення всіх слів до нормальних форм (лем). Це виконується завдяки надсиланню HTTP запита POST зі словом в якості параметра на сервіс лематизатора, який повертає GET запит з нормальною словоформою. Для вхідного пошукового запиту користувача після зчитування проводиться процедура «розширення» (query expansion), завдяки чому запит поповнюється гіпернімами, зчитаними з бази WordNet.

Метод було програмно реалізовано та протестовано на корпусі наукових текстових документів. Слід зазначити, що запропонований у доповіді модифікований частотний критерій окрім релевантності слова враховує також його семантичну вагу, покращуючи тим самим якість виконання пошукового запиту. Це дає змогу отримувати релевантні дані навіть у тому випадку, коли більшість або всі слова запиту не містяться у корпусі, незважаючи на семантичну подібність між корпусом і запитом. В ході тестування розроблений метод пошуку релевантних веб-документів показав суттєве поліпшення результатів за показниками точності з невеликою втратою у показниках повноти.

Перспективною є подальша модернізація алгоритму для збільшення швидкості роботи і вдосконалення алгоритму для роботи з великими масивами даних.

Список використаних джерел:

1. Карпенко А.П. Меры важности концептов в семантической сети онтологической базы знаний [Электронный ресурс]/ А.П. Карпенко // Наука и образование: электронное научно- техническое издание, 2010, 7. (http://technomag.edu.ru/doc/151142.html).

2. Чала Л.Е. Оцінка семантичної близості текстових структур методом Bmatch [Текст]/ Л.Е. Чала, А.О. Зуб // Міжнародна науково-технічна конференція «Проблеми информатизації», тези доповідей 2-ї міжнар. наук.-техн. Конф., 12-13 квітня 2014р., Черкаси, Київ, Тольятті, Полтава, 2013. – С. 56.