УДК 004.853

Л.Е. Чала, Ю.Ю. Харитонова, А.О. Зуб

Харківський національний університет радіоелектроніки, Харків

МЕТОД ПОШУКУ РЕЛЕВАНТНИХ ВЕБ-ДОКУМЕНТІВ З ВИКОРИСТАННЯМ МОДИФІКОВАНОГО ЧАСТОТНОГО КРИТЕРІЮ

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.

Ключові слова: *інформаційний пошук, релевантність, зважування термів, WordNet.*

Вступ

Інформаційний пошук був і є однією з найбільш розвинених галузей сучасної науки про штучний інтелект. Це зумовлено в першу чергу фантастичними темпами збільшення кількості інформації, яка зберігається в межах мережі Інтернет. Автоматичні системи інформаційного пошуку використовують для зменшення так званого «інформаційного перевантаження». Чимало користувачів використовують системи інформаційного пошуку для полегшення доступу до веб-документів (книжок, журналів, наукових статей та повідомлень). Результат пошуку – перелік знайдених документів, для кожного з яких, як правило, видається його назва, URL, розмір, дата створення і фрагмент тексту, що дозволяє судити про зміст сторінки (форма видачі результатів варіюється в різних системах і часто може вибиратися користувачем). При розширеному пошуку формується складний запит з використанням кількох ключових слів або виразів, зв'язаних логічними операторами. Ефективність виконання пошуку визначається тим, чи знайдені необхідні користувачу документи, і наскільки багато не релевантних запиту документів було йому запропоновано. Якість виконання пошуку залежить від засобів подання запиту, засобів подання знань про інформаційні ресурси, засобів їхнього зіставлення та обсягу інформаційних ресурсів, доступних пошуковій системі. 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]. Актуальною та важливою науково-технічною задачею є розробка методів та відповідних програмних засобів для підвищення ефективності пошуку та аналізу даних у мережі Інтернет на основі використання нових критеріїв оцінювання їх релевантності. Зокрема, є доцільним розглянути можливість 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. В статті розглядається вирішення такої задачі.

Постановка задачі

У сучасних умовах, щоб знайти релевантну інформацію, яка відповідає критеріям користувача, необхідно витратити чимало часу на обробку різноманітних джерел за тематикою, яка його цікавить. Нажаль, суттєво не покращують ситуацію і пошукові системи, які іноді видають за одним запитом тисячі релевантних результатів, що також не сприяє підвищенню ефективності пошуку (під ефективністю пошуку зазвичай розуміють знаходження необхідної та достовірної інформації за мінімально можливий час). З іншого боку, останнім часом активно розвиваються спеціалізовані системи видобування даних із Web-джерел. Їх основна особливість полягає в структурованості оброблюваної інформації за ключовими атрибутами, що дозволяє у загальному випадку підвищити ефективність пошуку релевантних даних за запитами користувачів. Але більшість існуючих методів такої структуризації має ряд суттєвих обмежень та припущень, пов’язаних зі складністю алгоритмів аналізу та обробки необхідних даних за допомогою спеціалізованих пошукових систем.

Задачею цієї статті є розробка модифікованого методу пошуку релевантних веб-документів з використанням подвійного зважування. Важливість слова в документі, який є частиною масиву або корпусу наукових документів, пропонується визначати за допомогою розширеної оцінки TF-IDF. Семантично близькі поняття можуть бути виражені завдяким використанню різних слів у документах і запитів,

що робить пряме порівняння за словами на основі стандартної VSM-моделі неефективним або взагалі неможливим. Запропонований у роботі метод надає можливість знаходження семантично подібних документів з використанням засобів WordNet та семантичних критеріїв подібності. Це дозволить проводити пошук більш релевантних документів у корпусі згідно з пошуковим запитом.

Моделі інформаційного пошуку

Основна мета задачі інформаційного пошуку – допомогти користувачу знайти інформацію, яка йому необхідна. Процес інформаційного пошуку в загальному вигляді включає в себе послідовність операцій, які направлені на збір, обробку і надання необхідної інформації зацікавленим особам.

В процесі становлення інформаційного пошуку було сформовано моделі, які з часом стали класичними, а саме: булева модель, ймовірностна модель, векторна модель, дескрипторна модель та моделі, базовані на класифікаторах.

В булевій моделі документ подається за допомогою набору термінів, які зберігаються в індексі. Кожен термін представлений як булева змінна. Ефективність пошуку невисока і неможливо ранжування документів за релевантністю.

В основі ймовірностних моделей лежить принцип ймовірністного ранжування (Probabilistic Ranking Principle, PRP). Згідно з цим принципом найбільша ефективність пошуку досягається тоді, коли результуючі документи ранжуються за зменшенням ймовірності їх релевантності запиту користувача [2].

Векторні моделі (Vector Space Model), на відміну від булевих, дозволяють ранжувати результуючу множину документів запиту. Документи (та запити до них) представляють собою набір векторів в n-мірному просторі [3]. Простір містить n базисних нормалізованих векторів, де n – загальна кількість різних термів в усіх документах. Значення компонентів вектору визначає вага терму (терміну). Показник відповідності (релевантності) визначається як оцінка кореляції між векторами.

Дескрипторна модель є найпростішою моделлю пошуку. В ній документ задається в вигляді набору асоційованих з ним зовнішніх атрибутів. У простих системах дескрипторного пошуку подання документу описується сукупністю слів (дескрипторів) або фраз лексики предметної області (ПрО), які характеризують зміст документу. Моделі, базовані на класифікаторах, також належать до найпростіших моделей пошуку. Документ у цій моделі, як і у дескриптивних системах, характеризується сукупністю асоційованих з ним атрибутів.

Ефективність пошуку в інформаційно-пошукових системах аналізується і регулюється перш за все за рівнем релевантності й пертинентності в частині вдосконалення організації запитів користувачів, пошуку за параметрами, за рахунок кластеризації, пошуку за подобою, ранжуванням відгуків, використання «сюжетних підходів» та використання семантичних методів.

Одним з найголовніших етапів запропонованого алгоритму пошуку є пошук семантично схожих термів за допомогою WordNet.

WordNet – семантичний словник для англійської та російської мов. У ньому слова мови розбито на групи синонімів – синсети (від англ. synset, synonym set), та надається коротке загальне визначення, та семантичні стосунки між цими словами. Мета подвійна: по-перше, це створення комбінації словника і тезауруса, більш інтуїтивно придатних для використання, а по-друге, підтримка автоматичного аналізу текстів та розробок в галузі штучного інтелекту. WordNet містить близько 100000 термів, організованих в таксономічні ієрархії. Іменники, дієслова, прикметники і прислівники, згруповані в синсети. Синсети також організовані в сенси (тобто відповідні різним значенням одного терма або концепта). Синсети (або концепти) пов'язані з іншими синсетами, які знаходяться вище або нижче в ієрархії, різними типами відносин.

Таким чином, перспективним є застосування можливостей засобів WordNet для вирішення поставленої задачі.

Модифікований метод пошуку

Для застосування методу, що пропонується, необхідно попередньо здійснити векторизацію документів та запитів на підставі синтаксичного аналізу, визначити найбільш вживані терми, а потім сформувати морфологічний словник. Терм будемо визначати як не стоп-слово після процедури лематизації, що проводиться шляхом морфологічному аналізу кожного слова для пошуку нормальних словоформ у морфологічному словнику. Морфологічний словник (лексикон) має містити всі словоформи однієї мови (в нашому випадку англійської або російської).

Маючи морфологічний словник, можна приступати до морфологічному розбору тексту. Кожній словоформі приписується певний набір граммем (морфологічна інтерпретація слів), які є значеннями морфологічних категорій (рід, число, відмінок і т.д.). Крім цього, в словнику кожної морфологічної інтерпретації мають бути задані нормальні форми слів (леми). Таким чином, для кожної словоформи S словник видає набір пар <M, L>, де M – морфологічна інтерпретація S, а L – лема словоформи S. Якщо для словоформи знаходиться більше однієї пари <M, L>, то виникає проблема дозволу морфологічної омонімії, оскільки в заданому вхідному

тексті, як правило, тільки одна морфологічна інтерпретація є вірною.

Для методу, що розглядається, структуру словника пропонується представити у вигляді реляційної схеми, яка складається з таблиць Lemmata, FlexiaModels, AccentModels та Ancodes.

Таблиця Lemmata містить перелік всіх лем даного словника, для кожної леми дано її властивості: псевдооснова слова, тобто спільна для всіх словоформ даного слова строчка (поле BaseStr); посилання на набір закінчень (поле FlexiaModelId); посилання на набір наголосів (поле AccentModelId); посилання на набір приставок (поле PrefixSetId); посилання на призначену для користувача сесію, при якій була внесено остання зміна цього запису (поле SessionId); посилання на загальні граммеми даної леми (поле Ancode) (може бути порожнім). Загальні граммеми є семантизованими граммемами, які повинні бути приписані всім словоформам відповідної леми (наприклад, граммема «прізвище»). Набір приставок леми – це ті префікси, з якими лема утворює повне слово мови. До набору префіксів може входити порожній префікс, що означає можливість безпрефіксного використання леми. Таблиця FlexiaModels містить заданий перелік можливих закінчень всіх лем. Унікальним ключем тут є поля FlexiaModelId і FormNo. Поле FormNo містить порядковий номер закінчення в даному наборі закінчень, відповідно, FormNo не перевищує максимальна к-ть словоформ в одне парадигмі. Поле PrefixStr містить префікс даної словоформи (можливо, порожній). Поле FlexiaStr містить закінчення даної словоформи (можливо, порожнє). Поле Ancode містить морфологічну інтерпретацію даної словоформи. Припустимо, що ми маємо запис Q з таблиці Lemmata, а P – один з її можливих префіксів, взятих по полю Q.PrefixSetId. Для того, щоб отримати і-ю словоформу даної леми, треба знайти в таблиці FlexiaModels запис R, таку, що Q.FlexiaModelId = R.FlexiaModelId і R.FormNo = i, тоді і-я словоформа буде дорівнює сумі (P + R.PrefixStr + Q.BaseStr + R.FlexiaStr).

Таблиця AccentModels містить перелік можливих номерів ударних голосних для словоформ. Унікальним ключем є поля AccentModelId і FormNo. Поле FormNo виконує таку ж роль, що і в таблиці FlexiaModels. Поле AccentedCharNo містить номер ударної гласною з кінця слова. Для кожної словоформи в словнику має бути зазначено наголос, якщо наголосу немає, тоді використовується спеціальна константа (255). Таблиця Ancodes містить всі можливі морфологічні інтерпретації. Ключем є поле Ancode. Поле PartOfSpeech містить частина мови (С, Г, П, ...), а поле Grammems набір граммем.

Така схема побудови словника має деякі принципово важливі переваги. Слід відзначити, що словник зберігає інформацію щодо можливих закінчень, можливих приставок, які можуть приєднуватися або до окремих словоформ, або до всіх словоформ даної парадигми. Крім того, словник зберігає інформацію про наголоси.

Функцію частотного зважування термів Idf в інформаційних пошукових 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 і в документі:

$$d_i = tf_i \cdot idf_i , \qquad (1)$$

де tf_i є частотою появи терма і в документі, а idf_i є оберненою частотою появи терма і 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:

$$S(d,q) = \frac{\Sigma_i q_i d_i}{\sqrt{\Sigma_i q_i^2}\sqrt{\Sigma_i d_i^2}}, \qquad (2)$$

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 завдяки 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 [4].

Модифікований 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:

$$q_i = q_i + \sum_{S(i,j)\ge t}^{i\neq j} q_j S(i,j),$$

де t – пороговий коефіцієнт, що задається користувачем.

На другому етапі здійснюється розширення сукупності термів за рахунок додаткових термів з подібністю, що перевищує порогове значення T.

По-перше, запит доповнюється синонімічними термами (береться найбільш поширений «сенс»).

Згодом запит доповнюється гіпонімами і гіпернімами, які є семантично подібними термам запиту.

Рис. 1 ілюструє цей процес: кожен елемент користувацького запиту представляє представлений своєю деревовидною ієрархією WordNet.

Потім визначаються околиці терма і всі терми з подібністю більше порогової T (у цій роботі T = 0.9), також мають бути включені в вектор запиту.

Це розширення може включати такі терми, які знаходяться на один рівень вище або нижче в ієрархії.

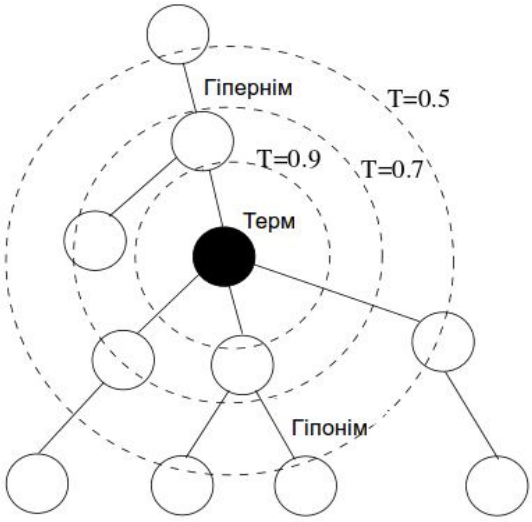

Рис. 1. Розкриття терму за допомогою WordNet

Кожному терму розширеної сукупності присвоюється вага згідно з таким виразом:

$$q_i = \begin{cases} \sum\limits_{S(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j S(i,j), \\ \qquad\qquad \text{якщо i – новий терм;} \\ q_i + \sum\limits_{S(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j S(i,j), \\ \qquad\qquad \text{якщо терм i – ваги } q_i, \end{cases}$$

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. Наведена умова 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:

$$S(q,d) = \frac{\sum\limits_i \sum\limits_j q_i d_j S(i,j)}{\sum\limits_i \sum\limits_j q_i d_j}, \qquad (3)$$

де i та 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).

Реалізація та тестування методу

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. Для оцінки якісних характеристик розробленого методу було здійснено тестування процедур пошуку релевантної інформації в мережі Інтернет для різних типів запитів.

Результати тестування методу підтвердили його працездатність (кількість незадовільного визначення релевантості знайдених документів запитам користувачів в середньому не перевищує 2%). 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 [5].

Висновки

В ході тестування розроблений метод пошуку релевантних веб-документів показав суттєве поліпшення результатів за показниками точності з невеликою втратою у показниках повноти.

Запропонована модифікація методу пошуку з використанням подвійного зважування дозволяє враховувати зв'язки між усіма семантично подібними термами в документі і в запиті, що не може бути забезпечено стандартною VSM-моделлю.

Перспективною є подальша модернізація алгоритму на основі застосування паралельних обчислювань для збільшення швидкості роботи і вдосконалення алгоритму для роботи з великими масивами даних.

Список літератури

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. – Режим доступа: (http://technomag.edu.ru/doc/151142.html).

2. Jabberwacky [Электронный ресурс]. – Режим доступа: www. http://www.liveenglish.ru/about/jabberwacky.html. – 26.05.2014 г. – Загл. с экрана.

3. Домашний робот: от идеи к продукту [Электронный ресурс]. – Режим доступа: www/ URL: http://habrahabr.ru/company/cubicrobotics/blog/222655 – 23.05.2014 г. – Загл. с экрана.

4. Чала Л.Е. Оцінка семантичної близькості 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.

5. Чала Л.Е. Модифікований метод пошуку релевантних веб-документів з використанням подвійного зважування [Текст] / Л.Е. Чала, Ю.Ю. Харитонова // Internet education science. IES-2014. Proceedings of the ninth international scientific-practical conference. Ukraine, Vinnytsia, VNTU, 2014. – С. 18-20.

Надійшла до редколегії 16.12.2014

Рецензент: д-р техн. наук, проф. С.Г. Удовенко, Харківський національний університет радіоелектроніки, Харків.

МЕТОД ПОИСКА РЕЛЕВАНТНИХ ВЕБ-ДОКУМЕНТОВ С ПРИМЕНЕНИЕМ МОДИФИЦИРОВАННОГО ЧАСТОТНОГО КРИТЕРИЯ

Л.Э. Чалая, Ю.Ю. Харитонова, А.О. Зуб

В статье предложен усовершенствованный метод поиска релевантных веб-документов с помощью модифицированного частотного критерия. Метод подразумевает расширение входного запроса пользователя синонимичными векторами с повторным взвешиванием. Осуществлена программная реализация метода с использованием программных средств Ruby. В ходе тестирования разработанный метод показал существенное улучшение результатов по показателям точности с небольшой потерей в показателях полноты.

Ключевые слова: *информационный поиск, релевантность, взвешивание термов, WordNet.*

THE METHOD OF FINDING RELEVANT WEB-DOCUMENTS USING THE MODIFIED FREQUENCY CRITERION

L.E. Chala, Yu.Yu. Kharytonova, A.A. Zub

The article 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.

Keywords: *information retrieval, relevance, weighing terms, WordNet.*