11. Ferguson C. Diglossia. Arabic Natural Language Processing: Challenges and Solutions / C. Ferguson. – 1959.
12. Ferguson C. Epilogue: Diglossia revisited / C. Ferguson // Contemporary Arabic Linguistics in Honor of El-Said Badawi / The American University in Cairo. – Cairo, 1996.
13. McCarthy J. A prosodic theory of nonconcatenative morphology / J. McCarthy. – Linguistic Inquiry. – 1981. – Vol. 12, No. 3. – P. 373–418.
14. Soudi A. Arabic Computational Morphology: Knowledge-Based and Empirical Methods / A. Soudi, A. van den Bosch, G. Neumann. – Springer, 2007. – 308 p. – (Text, Speech, and Language Technology).
15. Versteegh K. The Arabic Language / K. Versteegh. – NY : Columbia University Press, 1997.

Глибовець А. М., Мусбаг Заід Енвігі Мохаммед

ОСОБЛИВОСТІ АВТОМАТИЧНОЇ ОБРОБКИ АРАБСЬКОЇ МОВИ

Складність арабської мови ставить перед методами обробки природної мови великі виклики і вимагає докладних досліджень. Ця стаття є першим кроком до розуміння проблем та спробою дати поштовх до пошуку їх вирішення в автоматичній обробці арабської мови.

Ключові слова: арабська мова, обробка природної мови, NLP.

Матеріал надійшов 02.09.2015

УДК 681.3

Олецький О. В.

ПРО ПІДХІД ДО АВТОМАТИЧНОГО ФОРМУВАННЯ РЕКОМЕНДАЦІЙ ДЛЯ ВІДВІДУВАЧІВ ВЕБ-ПОРТАЛУ НА ОСНОВІ ТЕОРІЇ НЕЧІТКИХ МНОЖИН

Розглянуто задачу автоматичного формування рекомендацій для відвідувачів тематичного порталу щодо того, які сторінки видаються найбільш перспективними для подальшого перегляду. При цьому взято до уваги, що рекомендовані матеріали не повинні бути ні надто схожими на поточну сторінку, ні надто віддаленими від неї. Розглянуто функцію залежності між мірами релевантності та відстанями між документами, для опису якої використовується апарат теорії нечітких множин. Запропоновано методику розрахунку мір релевантності на основі відповідного нечіткого правила, наведено конкретний приклад такого розрахунку.

Ключові слова: тематичний портал, рекомендаційна система, міри релевантності, нечітке правило.

Вступ

У роботах [1–5] розвивається напрям, пов’язаний з автоматичним формуванням рекомендацій щодо добору найбільш релевантних навчальних матеріалів на тематичному порталі. Мова йде про ситуацію, в якій основною метою відвідувача порталу є отримання якомога більш повної та всебічної інформації з певного питання, інакше кажучи, максимізація рівня своїх знань з певної

теми. Особливий інтерес становить ситуація, коли для повного задоволення інформаційної потреби одного документа недостатньо і потрібно комбінувати інформацію, яка міститься в різних документах.

Тематичний портал містить значну кількість матеріалів з різних тем; для цих матеріалів характерна достатньо висока якість та зв'язність. З іншого боку, побудова рекомендаційної системи ускладнюється у зв'язку з такими рисами інформаційного наповнення:

– значна частина інформації дублюється, але матеріали можуть бути розраховані на різні категорії відвідувачів, зокрема, на різний рівень підготовки;

– множина матеріалів у репозиторії не є незмінною, вони можуть додаватися не тільки власниками та модераторами порталу, але й іншими відвідувачами.

Інформаційне наповнення має бути тісно пов'язаним з онтологією предметної області, яка має стати ключовою компонентою рекомендаційної системи.

Суттєвою проблемою при цьому є те, що заздалегідь зазвичай невідомо, які саме матеріали слід рекомендувати конкретному відвідувачеві з огляду на його поточний рівень знань, індивідуальні особливості тощо. Один з можливих підходів до вирішення цієї проблеми полягає в навчанні рекомендаційної системи на основі методу проб і помилок, зокрема, на основі марковських процесів прийняття рішень [6]. Як певну альтернативу можна розглядати вироблення рекомендацій на основі аналізу мір близькості між документами та вузлами онтології (хоча зазначені підходи не є взаємовиключними та можуть комбінуватися між собою). Природно говорити про деяку функцію залежності мір релевантності від мір близькості.

Але слід звернути увагу на важливу обставину. Якщо відвідувач перебуває на сторінці, пов'язаній з певним навчальним матеріалом, яку наступну сторінку йому запропонувати як найбільш перспективну? Очевидно, що не найбільш віддалену, але й не найменш віддалену – найменш віддалена сторінка буде надто схожою на цю і існує значний ризик того, що розміщена на ній інформацію буде вже знайома відвідувачеві. Інакше кажучи, функція залежності міри релевантності від міри близькості має спочатку зростати, досягати максимуму в районі середніх значень мір близькості і потім знову спадати.

Об'єктивна складність отримання точних кількісних оцінок зумовлює необхідність формулювати закономірності на якісному, лінгвістичному рівні. Ця обставина дає підстави говорити про доцільність залучення нечітких множин для опису нечітко сформульованих понять і про нечіткий характер шуканої функції залежності.

Основний зміст роботи

Описані міркування з приводу того, що документи, перехід до яких з певного вузла є найбільш перспективним, не повинні бути ні надто схожими на цей вузол, ні надто віддаленими від нього, можна сформулювати як нечітке правило: «якщо відстань СЕРЕДНЯ, релевантність ВИСОКА».

Якщо відстані задаються як нечіткі множини, міри релевантності, або перспективності, мають бути розраховані за методом центру тяжіння композиції максимум-мінімум [7; 8], а саме:

– задати нечітку множину A, яка характеризує поняття «СЕРЕДНЯ ВІДСТАНЬ»;

– задати нечітку множину B, яка характеризує поняття «ВИСОКА РЕЛЕВАНТНІСТЬ»;

– задати нечітку множину A', яка характеризує відстань до конкретного вузла;

– обчислити нечітку множину B' за формулою

$$B' = A' \bullet (A \Rightarrow B),$$

де знак $\Rightarrow$ позначає відношення нечіткої імплікації, знак $\bullet$ – композицію максимум-мінімум;

– знайти центр тяжіння множини B', отриману величину можна прийняти як шукану міру релевантності.

Можна розглядати різні відношення нечіткої імплікації. Зокрема, в [8] наводяться такі відношення:

– min-імплікація:

$$\mu^{\min}_{A\to B}(u,v) = \min(\mu_A(u), \mu_B(v));$$

– нечітке розширення класичної імплікації:

$$\mu^{Kls}_{A\to B}(u,v) = \max(\mu_B(v), 1-\mu_A(u));$$

– нечітка імплікація Лукасевича:

$$\mu^{Luc}_{A\to B} = \min(1, 1-\mu_A(u)+\mu_B(v)).$$

Але часто буває так, що відстані задаються конкретними числовими значеннями. Тоді для обчислення нечітких мір релевантності можна скористатися прийомом, який у [8] охарактеризований як метод простої підстановки нечіткого значення. А саме: шукана міра релевантності v отримується з рівняння

$$\mu_B(v) = \alpha,$$

де α – ступінь належності заданої відстані до нечіткої множини A.

Нечіткі міри релевантності слід розрахувати для всіх документів, які видаються потенційно корисними, і далі ранжувати ці документи за отриманими мірами.

При цьому можна розглядати відстані як між документами, так і між вузлами онтології. Ці міри можуть вводитися по-різному. Найпростіший підхід полягає в побудові матриці «документ-термін» *M*, рядки якої відповідають документам, а стовпці – термінам предметної області [9]. Елемент m_{ij} означає кількість входжень *j*-го терміна до *i*-го документа (або ж його інверсна частота). Тоді відстань між документами визначається як геометрична відстань між відповідними рядками (а відстань між термінами як вузлами онтології – як відстань між відповідними стовпчиками). Більш складні міри близькості можна отримувати на основі деяких формалізованих моделей інформаційного наповнення порталу, які залучають до розгляду зв'язки між документами та вузлами онтології предметної області (одна з таких формалізацій розглядалася в [3]); на основі врахування не тільки термінів, які трапляються в документі безпосередньо, але й пов'язаних з ними понять тощо.

Проілюструємо описану методику на простому прикладі. Нехай маємо множину документів $D=\{D_1,...D_{10}\}$ та множину термінів $T=\{T_1,...,T_8\}$. Нехай відвідувач у певний момент переглядає документ D_1. Потрібно розрахувати міри релевантності інших документів на основі описаного вище нечіткого правила. Відстані та міри релевантності будемо вважати нормованими, тобто їхнє максимальне значення дорівнює 1.

Передовсім потрібно розрахувати відстані від D_1 до всіх інших документів. Для скорочення викладення опустимо цей тривіальний етап. Вважаємо, що відстані вже підраховані та упорядковані за зростанням:

$\rho(D_1,D_2) = 0;$
$\rho(D_1,D_3) = 0,05;$
$\rho(D_1,D_4) = 0,1;$
$\rho(D_1,D_5) = 0,2;$
$\rho(D_1,D_6) = 0,25;$
$\rho(D_1,D_7) = 0,4;$
$\rho(D_1,D_8) = 0,6;$
$\rho(D_1,D_9) = 0,8;$
$\rho(D_1,D_{10}) = 1.$

Відповідно, якщо рекомендаційна система керується міркуваннями про те, що найбільш релевантними є найменш віддалені документи, вона має розташувати посилання на них у зазначеному порядку – від D_2 до D_{10}.

Задамо функцію належності для поняття «ВИСОКА РЕЛЕВАНТНІСТЬ» (доцільно, щоб ця функція була монотонно неспадною, але вона може бути як лінійною, так і нелінійною; ми взяли нелінійну функцію належності):

$\mu_B(0) = 0;\ldots;\ \mu_B(0,5) = 0;\ \mu_B(0,55) = 0,05;\ \mu_B(0,65) = 0,2;\ \mu_B(0,75) = 0,4;\ \mu_B(0,8) = 0,5;\ \mu_B(0,85) = 0,7;\ \mu_B(0,9) = 0,8;\ \mu_B(0,95) = 0,9;\ \mu_B(1) = 1.$

Задамо функцію належності для поняття «СЕРЕДНЯ ВІДСТАНЬ» (як зазначалося раніше, вона має спочатку зростати, а потім спадати):

$\mu_A(0) = 0;\ \mu_A(0,05) = 0,3;\ \mu_A(0,1) = 0,5;\ \mu_A(0,15) = 0,75;\ \mu_A(0,2) = 1;\ \mu_A(0,25) = 0,9;\ \mu_A(0,3) = 0,7;\ \mu_A(0,5) = 0,6;\ \mu_A(0,6) = 0,3;\ \mu_A(0,9) = 0,1;\ \mu_A(1) = 0.$

Оскільки в нашому випадку відстані задані числовими значеннями, ми можемо скористатися методом простої підстановки. Його застосування дає такі міри перспективності для документів (якщо в таблиці не було відповідного значення, ми брали найближче):

$v(D_2) = 0;\ v(D_3) = 0,7;\ v(D_4) = 0,8;\ v(D_5) = 1;\ v(D_6) = 0,95;\ v(D_7) = 0,82;\ v(D_8) = 0,7;\ v(D_9) = 0,6;\ v(D_{10}) = 0,5.$

Таким чином, документи $D_2,...D_{10}$ мають бути проранжовані за релевантністю до документа D_1 в такому порядку:

$D_5, D_6, D_7, D_4, D_3, D_8, D_9, D_2, D_{10}.$

Тому, якщо відвідувач перебуває на сторінці, пов'язаній з документом D_1, рекомендаційна система має розмістити посилання на інші документи саме в такому порядку. Крім того, якщо для аналізу мір важливості сторінок розглядати марковський процес, аналогічний тому, який використовується в алгоритмі Page Rank [9], можливо, зі змінними перехідними ймовірностями між сторінками, то ці перехідні ймовірності можна пов'язувати з отриманими мірами перспективності.

Висновки

У роботі розглянуто можливість вибору документів, перехід до яких у рамках тематичного порталу видається найбільш перспективним з огляду на те, що перехід не має здійснюватися

ні до надто схожих документів, ні до надто несхожих. Розглядається деяка функція залежності міри релевантності документів від мір близькості між ними. Для опису цієї залежності запропоновано використовувати нечітке правило: «якщо відстань СЕРЕДНЯ, релевантність ВИСОКА». На основі цього запропоновано методику розрахунку мір релевантності, наведено конкретний приклад такого розрахунку.

Але очевидно, що корисність описаної методики критичним чином залежить насамперед від того, наскільки адекватно задаються функції належності та які міри близькості використовуються. Відповідні величини мають підбиратися на основі статистичних даних, які можуть бути отримані шляхом аналізу реальної поведінки відвідувачів порталу та, можливо, ефективності рекомендацій, які надавалися рекомендаційною системою. Перспективним видається і можливість підбору параметрів функцій належності на основі генетичних алгоритмів [10]. При цьому як функцію, що має мінімізуватися, можна розглядати розбіжність між оцінками релевантності, які були обчислені рекомендаційною системою, і тими, які дала б людина-експерт [11].

Усе це має стати предметом подальших досліджень.

Список літератури

1. Олецький О. В. Організація онтологічно-орієнтованих засобів автоматизованого добору інформаційних ресурсів на тематичному порталі / О. В. Олецький // Наукові записки НаУКМА. – 2009. – Т. 99 : Комп'ютерні науки. – С. 66–69.
2. Олецький О. В. До проблеми моделювання потоку відвідувань на онтологічно-орієнтованому тематичному порталі / О. В. Олецький // Моделювання та інформаційні технології : зб. наук. пр. – К., 2010. – Т. 2 : Спецвипуск. – С. 321–326.
3. Олецький О. В. Побудова формалізованого опису графа «онтологія-документ» як моделі інформаційного наповнення тематичного порталу / О. В. Олецький // Наукові записки НаУКМА. – 2012. – Т. 138 : Комп'ютерні науки. – С. 57–60.
4. Олецький О. В. Про застосування марковських процесів прийняття рішень для автоматизованого добору навчальних матеріалів у системах blended learning / О. В. Олецький // Наукові записки НаУКМА. – 2013. – Т. 151 : Комп'ютерні науки. – С. 115–118.
5. Олецький О. В. Про оптимізацію структури веб-порталу на основі марковських процесів прийняття рішень / О. В. Олецький // Вісник КНУ імені Тараса Шевченка. Серія фізико-математичні науки. – 2013. – Спецвипуск. – С. 134–137.
6. Рассел С. Искусственный интеллект: современный подход / С. Рассел, П. Норвиг. – М. : Вильямс, 2006. – 1408 с.
7. Прикладные нечеткие системы / К. Асаи, Д. Ватада, С. Иваи [и др.] ; под ред. Т. Тэрано, К. Асаи, М. Сугэно. – М. : Мир, 1993. – 368 с.
8. Глибовець М. М. Штучний інтелект : підручник для студентів вищих навчальних закладів, що навчаються за спеціальностями «Комп'ютерні науки» та «Прикладна математика» / М. М. Глибовець, О. В. Олецький. – К. : Вид. дім «КМ академія», 2002. – 366 с.
9. Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – М. : Вильямс, 2011. – 528 с.
10. Глибовець М. М. Еволюційні алгоритми / М. М. Глибовець, Н. М. Гулаєва. – К. : НаУКМА, 2013. – 828 с.
11. Олецький О. В. Принципи застосування генетичних алгоритмів до задачі онтологічного інформаційного пошуку / О. В. Олецький // Наукові записки НаУКМА. – 2010. – Т. 112 : Комп'ютерні науки. – С. 49–54.

O. Oletsky

AN APPROACH TO FORMING RECOMMENDATIONS AT THE WEB-PORTAL ON THE BASE OF FUZZY SETS

The problem of forming recommendations for users of a web-portal about the most relevant pages is regarded. It should be taken to account that such pages should be neither too close to the current page nor too far from it. The function of dependency of relevancy measures from distances between documents is regarded; fuzzy sets are involved to describing this function. A method for calculating relevancy measures on the base of the appropriate fuzzy rule is suggested. A numerical example of such evaluating is provided.

Keywords: Web-portal, recommending system, relevancy measures, fuzzy rule.

Матеріал надійшов 15.09.2015