

АНАЛІЗ МЕТОДІВ РОЗШИРЕННЯ ЗАПИТІВ ІНФОРМАЦІЙНОГО ПОШУКУ ЯК ЗАСОБІВ ЗБІЛЬШЕННЯ ПОШУКОВИХ МОЖЛИВОСТЕЙ

Артюхова Юлія Ігорівна, студентка 4 курсу
Науковий керівник – Мочалова Н.С., старший
викладач Національний університет
кораблебудування ім. адмірала Макарова Україна, м.
Миколаїв

Анотація: У статті подано основні методи розширення інформаційних запитів та проведено аналіз цих методів на доцільність їх використання, щоб підвищити ступінь релевантності видачі результатів пошуку, при розробці інформаційно-пошукових систем.

Ключові слова: інформаційний-пошук, пошуковий запит, розширений запит.

Основне завдання інформаційного пошуку – це пошук документів, які задовольнятимуть інформаційну потребу користувача. Проте користувач не завжди може точно і в повному обсязі сформулювати свій запит. Такі неповні запити не дозволяють пошуковій системі задовольнити потребу користувача в інформації. Рішенням цієї проблеми може бути модифікація, тобто розширення інформаційних запитів, які зможуть правильно і, що більш важливо, повністю виразити інформаційну потребу користувача.

Метою статті є аналіз методів розширення інформаційних запитів, що дозволить визначити принципи їх роботи, основні характеристики та доцільність їх використання.

Традиційно для розширення запитів інформаційного пошуку використовувалися семантичні словники – тезауруси. «Інформаційно-пошуковий тезаурус (ІПТ) – це контрольований словник термінів створений на природній мові, який явно вказує відносини між термінами і призначений для інформаційного пошуку» [1].

Основною одиницею тезаурусів є терміни, які поділяються на дескриптори (авторизовані терміни) і недескриптори (аскриптори). «Дескриптор – це ключове слово, вибране з групи умовно еквівалентних ключових слів і представляє цю групу при індексуванні та пошуку інформації. У ролі дескрипторів виступають терміни, що позначають окремі поняття деякої предметної області та задовольняють принципи загальноживності, поширеності, стислості та термінологічної точності» [2]. Під аскрипторами розуміють лексичні одиниці інформаційно-пошукового тезауруса, які не

можуть використовуватись для індексування в пошукових образах документів і підлягають заміні на відповідний дескриптор. Отже, кожне ключове слово не є дескриптором, але за допомогою смислових відношень у тезаурусі воно пов'язується із дескриптором свого класу [3].

Типи зв'язків між дескрипторами частіше за все не відображаються в детальному переліку відносин тезауруса, а записуються за допомогою невеликого набору відносин, які зазвичай поділяються на: ієрархічні і асоціативні.

«ПРИКЛАДНА ЛІНГВІСТИКА 2019: ПРОБЛЕМИ ТА РІШЕННЯ»

XV ВСЕУКРАЇНСЬКА НАУКОВО-МЕТОДИЧНА КОНФЕРЕНЦІЯ СТУДЕНТІВ І МОЛОДИХ НАУКОВЦІВ

Ієрархічні відносини мають властивості транзитивності та антисиметричності, які можуть бути використані при надмірному індексуванні в інтересах підвищення ефективності інформаційного пошуку. Основними ієрархічними відношеннями є родо-видове відношення (нижче-вище) та відношення частина-ціле. Родо-видовий зв'язок встановлюється між двома дескрипторами, якщо обсяг поняття нижчого дескриптора входить в обсяг поняття вищого. Основним призначенням встановлення асоціативних відносин між дескрипторами інформаційно-пошукового тезауруса є те, що встановлення такого зв'язку може вказати додаткові дескриптори, корисні при індексуванні або при пошуку [4].

Тезаурусами також називають лінгвістичні ресурси типу WordNet і EuroWordNet, що описують відносини між лексичними значеннями природної мови як ієрархічну систему груп синонімів – синсетів.

Синсети в WordNet пов'язані між собою різними семантичними відносинами:

- ☐ гіперонімічні (відношення вище-нижче): (сніданок → прийом їжі);
- ☐ гіпонімічні (відношення вище-нижче): (факультет → професор);
- ☐ меронімічні (відношення частина-ціле): (стіл → ніжка);
- ☐ антонімічні: (лідер → послідовник);
- ☐ синонімічні: (солдат → воїн) [4].

Тезаурусний метод є досить простим в реалізації, а також ключовим в розширенні запитів, адже отриманий документ може бути релевантним, навіть якщо він не містить слів, які точно збігаються з запитом, але концептуально однакові. За своєю суттю тезаурус є штучною мовою опису, що будується на основі природної мови, а тому цей метод дозволить швидко знайти необхідну інформацію з мінімальними зусиллями з боку користувачів.

Наступним методом оптимізації пошукової системи є метод вивчення журналів пошукових запитів. Ці журнали запитів надають цінні вказівки для розуміння типів документів, які користувачі мають намір отримати, формулюючи запит з набором конкретних термінів. Цей метод використовують для аналізу уподобань користувача та додавання відповідних термінів до запиту.

Ханг Цуй розробив систему, яка витягує додаткові терміни на основі поведінки користувача та зберігає їх у вигляді журналів запитів. Ці журнали

зберігають список всіх документів, які були відвідані за конкретним запитом. Обчислення ймовірності перегляду документа, коли в ньому присутнє певне слово запиту, дозволяє визначити релевантність цього документа.

З журналів запитів можна витягти багато сеансів запитів. Кожен сеанс містить один запит і набір документів, на які користувач натиснув. Центральна ідея цього методу полягає в тому, що якщо набір документів часто обирається для тих самих запитів, то терміни в цих документах тісно пов'язані з умовами запитів. Коли подається новий запит, спочатку витягуються терміни запиту (після видалення стоп-слів). Тоді для кожного терміну запиту всі корельовані терміни документа вибираються на основі умовної ймовірності. Об'єднуючи ймовірності всіх термінів запиту, можна обчислити значення згуртованості терміну документа для нового запиту [5]. Проте такий метод не здобув популярності, оскільки він негативно впливав на ефективність пошуку при добре сформованому оригінальному запиті.

Існують також локальні методи розширення пошукових запитів. Такі методи змінюють запит з урахуванням документів, знайдених за вихідним запитом. До локальних відносяться такі методи:

- ☐Зворотній зв'язок за релевантністю;
- ☐Зворотній зв'язок за псевдорелевантністю;
- ☐ (Глобальний) неявний зворотний зв'язок за релевантністю [6].

Ідея зворотного зв'язку за релевантністю (Relevance Feedback) полягає в залученні користувача до процесу пошуку, щоб покращити кінцевий список результатів. Зокрема, користувач повідомляє систему про релевантність документів у первинному наборі результатів. Процедура починається з того, що користувач обирає найбільш релевантні документи у наборі відповідей, отриманих з початкового запиту. Далі система виділяє репрезентативні терміни (або вирази) обраних документів та включає їх до нового формулювання запиту. Система визначає краще представлення потрібної інформації на основі відгуку користувача та відображає переглянутий набір результатів пошуку [6].

Метод RF може передбачати одну або декілька ітерацій. В основі цього процесу лежить ідея, згідно з якою користувач не в змозі сформулювати точний запит, не знаючи добре змісту колекції, але може оцінити документи. Тому доцільно виконати кілька таких ітерацій, щоб уточнити запит. В рамках цього сценарію зворотний зв'язок за релевантністю також може бути ефективним у відстеженні змін інформаційних потреб користувача. Перегляд деяких документів може допомогти користувачеві уточнити свої уявлення про інформацію, яку він шукає. Розширення запиту з використанням псевдорелевантності (Pseudo-Relevance Feedback) є корисною та популярною методикою переконання переконання запиту.

«ПРИКЛАДНА ЛІНГВІСТИКА 2019: ПРОБЛЕМИ ТА РІШЕННЯ»

XV ВСЕУКРАЇНСЬКА НАУКОВО-МЕТОДИЧНА КОНФЕРЕНЦІЯ СТУДЕНТІВ І МОЛОДИХ НАУКОВЦІВ

В рамках цього методу спочатку виконується пошук і знаходиться вихідна сукупність документів, в якій перші N документів передбачаються релевантними, а потім ці документи використовуються для переформулювання запиту з використанням додаткових термінів для поліпшення результатів. Потім цей розширений запит буде оновлено для отримання остаточного набору документів, які будуть доступні користувачеві. Актуальність термінів розширення залежить від початково отриманих документів [7].

М.К. Чінакотла у своїй роботі «Multilingual Pseudo-Relevance Feedback: Performance Study of Assisting Languages» пропонує використовувати допоміжну мову для поліпшення результатів пошуку пошукової системи. Запит перекладають на допоміжну мову і двічі виконують зворотний зв'язок за псевдорелевантністю, один раз для мови запитів та ще один для допоміжної мови. Потім вони об'єднують розширенні терміни, отримані з обох екземплярів зворотного зв'язку, використовуючи переклад та отримують документи для розширеного запиту [8].

Арджун Атрея пропонує дуже перспективний підхід, який враховує структуру документів, при якому система надає пріоритет термінам розширення. Ідея полягає в тому, що термін, який знаходиться у заголовку документа, є більш важливим для цього документа, ніж термін, який знаходиться в тілі. Заголовок більш компактно представляє весь документ, і тому малоймовірно, що назва буде мати слово, яке не пов'язане з документом. Такий тип розширення може використовуватись, наприклад, при виборі термінів розширення для загального пошуку з таких підмножин, як Вікіпедія, де документи зазвичай мають мінімально фіксований набір полів, а саме: заголовок, тіло, інфобокс та категорії. Отримані таким чином результати є дуже перспективними і підвищують точність інформаційного пошуку [9].

Псевдорелевантність зворотного зв'язку відображає важливі терміни тільки на основі їх спільної появи в документі. Але для коректності результатів недостатньо лише взаємного поєднання. Потрібно розглянути семантичні та лексичні властивості слова. Крім того, оскільки зворотний зв'язок не залежить від користувача, існує ймовірність відхилення від теми. Це відхилення відбувається в результаті додавання термінів, які не мають ніякого зв'язку з темою запиту.

Висновки.

Розглянувши основні методи розширення інформаційних запитів можна зробити висновок, що для створення пошукової системи по сайту найбільш доцільно використовувати тазаурусний метод та метод зворотного зв'язку за псевдорелевантністю. Методи зворотного зв'язку за псевдорелевантністю не залежать від будь-якого зовнішнього лексичного ресурсу і, таким чином, невідоме слово не перешкоджає ефективності пошуку інформації. Такий метод витягує лексично і семантично пов'язані терміни для запиту і, таким чином, можна отримати документи, які контекстуально відносяться до запита, але не містять його термінів. Втім, через те, що користувач не може впливати на результати видачі, існує ймовірність відхилення від початкової теми запиту.

Тазаурусний метод є ключовим в розширенні запитів, адже отриманий документ може бути релевантним, навіть якщо він не містить слів, які точно збігаються з запитом, але концептуально однакові. За своєю суттю тезаурус є штучною мовою опису, що будується на основі природної мови, а тому цей метод дозволить швидко знайти необхідну інформацію з мінімальними зусиллями з боку користувачів. Процес індексування тезаурусу базується на лінгвістичних, граматичних знаннях, а також знаннях предметної області, що підвищує результативність пошуку.

Результати, отримані при проведенні цього дослідження, можуть бути використані при подальшому аналізуванні та вдосконаленні методів розширення запитів інформаційного пошуку, а також розвитку такого розділу комп'ютерної лінгвістики та інформатики як штучний інтелект. Проведений аналіз наявних методів розширення інформаційних запитів показав актуальність подальшого дослідження та розробки універсального метода, який буде максимально ефективно задовольняти пошукові потреби користувача.

СПИСОК ЛІТЕРАТУРИ

1. Федотов, А.М.(2013). Современные проблемы информатики и вычислительной техники. Взято з http://www.nsc.ru/win/elbib/data/show_page.dhtml?77+204.
2. Мельник, Р. А. (2018). Програмування веб-застосувань (фронт-енд та бек-енд).(с.44-45). Львів: Видавництво Львівської політехніки.
3. Телецька, А. О., Загнітко, А. П., & Надутенко, М. В.(2018). Класифікація інформаційно-пошукових мов. Взято з [jvestnik-sss.donnu.edu.ua/article/download/5477/5504](http://vestnik-sss.donnu.edu.ua/article/download/5477/5504)
4. Соловьев, В.Д., Добров, Б.В., Иванов, В.В., & Лукашевич, Н.В.(2006).(с.90) Онтологии и тезаурусы (Учебное пособие) Казань, Москва.
5. H, Cui, J.-R. Wen, J.-Y. Nie, & W.-Y. Ma. (2002). Probabilistic Query Expansion Using Query Logs. Взято з https://www.researchgate.net/publication/2920597_Probabilistic_Query_Expansion_Using_Query_Logs_Ha

«ПРИКЛАДНА ЛІНГВІСТИКА 2019: ПРОБЛЕМИ ТА РІШЕННЯ»

XV ВСЕУКРАЇНСЬКА НАУКОВО-МЕТОДИЧНА КОНФЕРЕНЦІЯ СТУДЕНТІВ І МОЛОДИХ НАУКОВЦІВ

6. Manning C. D., Raghavan P. & Schütze H. (2008) Introduction to Information Retrieval (c.177-178). Cambridge: Cambridge University Press
7. Vaidyanathan R., Das S., & Srivastava N. (2014) Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval. *International Journal of Computer Applications*, 105(8), 1-6
8. Chinnakotla, M. K., Raman, K., & Bhattacharyya, P. (2010) Multilingual pseudo-relevance feedback: performance study of assisting languages. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1346–1356. Sweden: Association for Computational Linguistics.
9. Atreya, A., Kakde, Y., Bhattacharyya, P., & Ramakrishnan. G. (2013) Structure cognizant pseudo relevance feedback. *In Proceedings of IJCNLP 2013*, 982–986.

Artiukhova Yu. I. Mochalova N. S., Analysis of methods of query expansion as means of increasing search capabilities

Abstract: *The article describes the main methods of query expansion and analyzes these methods on the expediency of their use in order to increase the relevance of the search results in the development of information retrieval systems.*

Keywords: *informational search, query, advanced query.*

Артюхова Ю. И., Мочалова Н.С., Анализ методов расширения запросов информационного поиска как средств увеличения поисковых возможностей

Аннотация: *В статье представлены основные методы расширения информационных запросов и проведен анализ этих методов на целесообразность их использования, чтобы повысить степень релевантности выдачи результатов поиска, при разработке информационно-поисковых систем.*

Ключевые слова: *информационный-поиск, поисковой запрос, расширенный запрос.*