УДК 004.89

М.В. Токман, В.В. Сокол, Н.С. Лєсна

Харківський національний університет радіоелектроніки, Харків

МОДЕЛІ І МЕТОДИ ПОЛІПШЕННЯ РЕЛЕВАНТНОСТІ ПОШУКУ ТЕКСТОВИХ ДОКУМЕНТІВ

Проведено аналіз необхідності поліпшення релевантності пошуку текстових документів. Сформульовано задачу поліпшення пошуку. Розглянуто методи поліпшення пошуку текстових документів. Проаналізовано критерії відбору найбільш ефективних методів поліпшення релевантності пошуку текстових документів. Пояснено значення терміну якість пошуку стосовно пошуку текстових документів. Серед розглянутих методів виділено один, який найбільше відповідає вказаним критеріям. Описано переваги методу латентно-семантичного пошуку.

Ключові слова: *пошук текстової інформації, релевантність, критерії релевантності.*

Вступ

Постановка завдання та аналіз основних досліджень. У зв'язку з розвитком інформаційних систем і з обширністю текстових документів на комп'ютерах природно виникає необхідність в пошуку потрібної інформації. Серед інформаційних систем, що мають справу з текстовою інформацією, найбільш поширеними є системи текстового пошуку. Пошук інформації за допомогою комп'ютерів має вже майже піввікову історію. Перші автоматизовані інформаційні системи почали розроблятися ще в 50-х роках минулого століття, і головною їх функцією був саме пошук інформації. Тому їх назвали інформаційно-пошуковими системами. Залежно від характеру підтримуваних інформаційних ресурсів ці системи розділяють на дві категорії: фактографічні і документальні.

Фактографічні інформаційно-пошукові системи оперували фактами, представленими у вигляді сутностей реального світу і їх властивостей, які дозволяли знаходити сутності, що мають задані користувачем властивості, а також властивості заданих сутностей.

Документальні інформаційно-пошукові системи призначені для зберігання і пошуку документів, що утримують тексти на природних мовах. Такі інформаційно-пошукові системи є ранніми системами текстового пошуку.

Системи текстового пошуку, що розроблялися в цей період, називалися дескрипторними. У таких системах зміст кожного текстового документу і призначених для користувача пошукових запитів описується наборами слів або словосполучень, що називаються дескрипторами.

Однією з найбільш поширених сфер застосування дескрипторних систем був бібліографічний пошук. У таких системах зберігаються колекції бібліографічних описів документів, і система дозволяє знаходити публікації заданого автора, публікації, що випущені вказаним видавництвом і/або вийшли в деякому році і тому подібне. Багато бібліографічних дескрипторних інформаційно-пошукових систем використовуються до теперішнього часу.

Основна одиниця інформації в системах текстового пошуку називається документом. Документ – це одиниця інформації, представлена на якій-небудь природній мові. У ранніх інформаційно-пошукових системах документ розглядався як атомарна (неділима) одиниця. Для системи він виступав як "чорний ящик". У розвиненіших системах текстового пошуку зміст документу доступний системі для обробки і аналізу.

Повнотекстові системи текстового пошуку оперують електронними документами, тобто документами, що зберігаються в пам'яті комп'ютерів і доступними для автоматизованої обробки. Документи зберігаються в системі текстового пошуку для того, щоб задовольняти інформаційні потреби користувачів. Представлення інформаційних потреб користувача у формі, що сприймається програмним забезпеченням системи текстового пошуку, називається призначеним для користувача запитом (чи просто запитом). Необхідним компонентом змісту призначеного для користувача запиту є опис тих властивостей, які мають документи, що цікавлять користувача. Цей опис природно називати критерієм пошуку.

Слід підкреслити, що одиницею гранулярного пошуку, тобто найменшою одиницею інформації, яка може видаватися користувачеві в результаті обробки заданого їм запиту, у більшості систем текстового пошуку являється саме документ, а не яка-небудь його порція. Як правило, в результаті обробки призначеного для користувача запиту система видає безліч результуючих документів, що задовольняють заданому в запиті критерію.

Постановка завдання Було сформульовано наступну постановку завдання:

1. досліджувати моделі і методи пошуку текстових документів;

2. проаналізувати існуючі методи;
3. описати характеристики пошуку текстових документів;
4. порівняти методи;
5. вибрати метод, який забезпечить кращу якість пошуку текстових документів;
6. розробити програмне забезпечення на основі вибраного методу.

Виклад основного матеріалу

Різноманітність функціональних можливостей різних систем текстового пошуку пов'язана саме з відмінністю реалізованих в них моделей пошуку.

У багатьох системах використовуються прості моделі пошуку. До їх числа відносяться моделі, ґрунтовані на класифікаторах. У моделі, ґрунтованій на класифікаторові, документи представляються ідентифікаторами класів в ієрархічній структурі класифікатора, до яких відноситься цей документ. Представлення запиту в простому випадку також є ідентифікатором користувача класу заданого класифікатора, що цікавить. Критерієм релевантності документу є умова, що клас документу співпадає з яким-небудь класом в уявленні запиту або є його підкласом.

У складнішому випадку в моделях пошуку, ґрунтованих на класифікаторові, допускається вказівка в запиті декількох класів класифікатора. При цьому релевантними вважаються документи, що належать якому-небудь з вказаних в запиті класів або його підкласу. Така модель пошуку близька до булевої моделі, що розглядається далі.

Дещо складніший характер мають моделі контекстного пошуку. Застосування цих моделей стало можливим, коли досить високої продуктивності досягли процесори обчислювальних машин і виріс об'єм їх зовнішньої пам'яті прямого доступу. У моделях контекстного пошуку використовується представлення документу як сукупності всіляких слів, що зустрічаються в його тексті, і словосполучень, не рахуючи так званих стоп-слів. Стоп-слова – це службові слова, які зустрічаються практично у будь-якому документі. Пошук документів, що містять такі слова, привів би до видачі повної колекції документів у відповідь на запит. Тому такі слова не можуть використовуватися як індексуючі властивості документів.

У системах даного класу будується індекс за усіма словами, що зустрічаються в документах, і словосполученнями, окрім стоп-слів. При цьому для побудови індексу слова, виділені з тексту документу, наводяться спочатку до "канонічного виду" за допомогою підтримуваних в системі словників і засобів граматичного розбору. Призначений для користувача запит також піддається граматичному розбору, в процесі якого із запиту також виділяються слова, що зустрічаються в його тексті, і словосполучення. Документ вважається релевантним, якщо які-небудь слова або словосполучення із запиту трапляються з точністю до граматичних форм в тексті документу.

Іноді використовується жорсткіший критерій релевантності – входження в текст документу усіх названих в запиті слів і словосполучень і так далі.

У системах текстового пошуку широко використовуються булеві моделі пошуку. Користувач може формулювати запит у вигляді булевого вираження з використанням операторів І, АБО, НЕМАЄ. Терми булевого вираження можуть бути різними в різних варіаціях булевих моделей пошуку. Це може бути, наприклад, умова входження цього слова або словосполучення (з точністю до граматичних форм) в текст документу у булевому розширенні контекстної моделі пошуку. У булевому розширенні моделі пошуку по класифікаторах термами вираження можуть бути умови приналежності документу цьому класу класифікатора. У булевій моделі пошуку з використанням Дублінського ядра термом може бути рівність, що описує той факт, що деякий елемент метаданих для даного документу має задане в запиті значення.

Критерієм релевантності цього документу запиту у булевих моделях пошуку являється істинність булевого вираження, заданого в запиті.

Найбільш широке застосування в розвинених системах текстового пошуку мають векторні моделі пошуку. Використання таких моделей вимагає значно великих обчислювальних ресурсів в порівнянні з іншими моделями, проте вони забезпечують істотно більш високу якість пошуку.

У векторних моделях передбачається, що документи і запити представляються векторами. У простому випадку координати вектору відповідають термам тексту – словам або словосполученням, що належать словнику системи, який представляє загальномовну лексику або лексику предметної області. Кожному терму з такого словника зіставляється свій вимір у векторному просторі. Розмірність векторів, що представляють документи і призначені для користувача запити, в точності дорівнює кількості вимірів в цьому просторі.

Координаті вектору привласнюється одиничне значення у тому випадку, коли терм, що відповідає їй, зустрічається в цьому документі або, відповідно, в призначеному для користувача запиті. Інакше координаті вектору привласнюється нульове значення. Оскільки розмір словника може бути дуже великим, а документи або тексти запитів містять істотно меншу кількість термів, що містяться в ньому такі вектори виявляються дуже розрідженими. Тому треба використати яку-небудь техніку стислого їх представлення.

Для оцінки міри релевантності документу запиту (заходи їх близькості) у векторних моделях пошуку використовуються які-небудь векторні функції, аргументами яких виступають вектори, що представляють їх. Наприклад, можна використати в якості

такої міри косинус кута між вектором документу і вектором запиту. Важливо помітити, що, оскільки ненульові координати цих векторів відповідають тільки термам, що входять в текст документу і текст запиту, на значення функції – заходи в обох цих випадках – роблять вплив тільки терми, загальні для документу і запиту.

Для підвищення керованості векторних моделей пошуку часто ускладнюють ці моделі. Вводяться ваги термів, що характеризують їх значущість. Значення цих вагів використовуються як координати вектору документу, якщо його текст включає відповідні терми. Таким чином, входження різних термів в текст документу роблять різний вплив на значення функції близькості документу і запиту. Існують різні підходи до вибору вказаних вагів. Найчастіше для цієї мети використовують частоти входження терма в цей документ і частоти його входження в усі документи колекції в цілому. Зважуватися можуть також і терми запиту.

Відмінності між конкретними векторними моделями пошуку зводяться саме до різних способів призначення вагів термів і вибору заходів близькості. Векторні моделі дозволяють ранжирувати результуючу множину документів запиту.

Системи текстового пошуку в останні роки активно використовуються в найрізноманітніших областях діяльності. Тоді як спочатку вони розроблялися головним чином як інструмент для бібліотечної справи, нині вони знайшли застосування в різних організаціях для роботи з їх текстовими інформаційними ресурсами.

Нині проблематика текстового пошуку стала досить великою. Вона охоплює різні області теорії і розробки систем текстового пошуку, такі як:

– розвиток конкретних моделей пошуку;

– методологія проведення експериментів, тестування і оцінки систем;

– методи реалізації текстового пошуку;

– підходи до інтеграції технологій текстового пошуку і баз даних;

– пошук в середовищі Веб;

– методи стискування даних;

– оцінка ефективності обробки запитів;

– обробка природної мови;

– методи класифікації і кластеризації текстових документів;

– додатки інформаційного пошуку в електронних бібліотеках;

– глибинний аналіз текстів;

– технології індексування і пошуку мультимедійної інформації;

– інтерфейси "людина – комп'ютер" і так далі.

Розробники систем текстового пошуку приділяють велику увагу не лише вдосконаленню механізмів виконання їх базових функцій, але і розвитку ряду додаткових можливостей, що дозволяють істотним чином збільшити ефективність пошуку, підвищити керованість системи, забезпечити комфортніші умови для роботи користувача. Перерахуємо деякі такі можливості.

Підвищення точності пошуку. Деякі системи текстового пошуку дозволяють користувачеві надати ваги використовуваним в запиті термам з тим, щоб охарактеризувати їх значущість в запиті. Ця інформація використовується при обчисленні оцінок релевантності документів інформаційним потребам користувача, і тим самим істотно підвищується точність пошуку.

Документи, що зберігаються в системі, які відповідають призначеному для користувача запиту, називаються релевантними.

Релевантність документу не обов'язково повинна оцінюватися в термінах двозначної логіки("так – ні"). У деяких розвинених системах використовуються тонші оцінки, які обчислюються як значення спеціально підібраної числової функції (функції релевантності), що набуває значень в інтервалі [0 ÷ 1]. У таких випадках доречно говорити про міру релевантності документу, розуміючи її як значення цієї функції. Деякі системи текстового пошуку видають користувачеві безлічі документів, отриманих в результаті обробки запитів, упорядковуючи документи по убуванню міри релевантності. Таке впорядкування знайдених документів називають їх ранжируванням. Користувач ефективніше може аналізувати ранжирувану множину результуючих документів запиту. З великою вірогідністю документи, що найбільш цікавлять його, з числа знайдених знаходяться на початку списку документів, що виводиться.

В силу різних причин, пов'язаних, зокрема, з труднощами автоматизації розуміння природної мови, а також з неточністю відображення інформаційних потреб користувача в запиті, в результаті обробки призначеного для користувача запиту можуть бути знайдені документи, що не відповідають інформаційним потребам користувача. Таке явище називається інформаційним шумом.

Важливими характеристиками якості пошуку в системах текстового пошуку є повнота і точність пошуку. Повнота пошуку визначає відношення кількості релевантних документів, що видаються в результаті обробки призначених для користувача запитів, до кількості фактично наявних в системі релевантних документів. Для кількісної оцінки точності пошуку може служити доля релевантних документів у множині результуючих документів запиту.

Ранжирування результуючих документів запиту. В силу розглянутих раніше причин системи текстового пошуку не можуть гарантувати строгого задоволення інформаційних потреб користувача в результаті виконання пошукових запитів.

Кількість результуючих документів зазвичай буває значною. Тому дуже важливо упорядкувати документи, що видаються системою користувачеві, так, щоб на початку списку знаходилися документи, які, ймовірно, більшою мірою представляють інтерес для користувача. Операція такого роду називається ранжируванням документів. Розвинені системи текстового пошуку мають механізми, що забезпечують таку можливість. Залежно від моделі пошуку, що реалізовується ними, передбачається впорядкування безлічі документів, що видаються в результаті обробки призначеного для користувача запиту, за деякими оцінками міри їх релевантності запиту або вірогідності задоволення інформаційних потреб користувача.

Зворотний зв'язок релевантності – важлива функція, що дозволяє підвищити ефективність пошуку потрібних користувачеві документів. Річ у тому, що результати обробки запиту можуть не задовольняти користувача. У таких випадках у багатьох системах текстового пошуку користувачеві надається можливість уточнити запит. Для цього він може дати оцінку релевантності отриманих документів – вказати, які з них він вважає релевантними або не релевантними.

Оскільки кількість результуючих документів може бути досить великою, користувачеві пропонується оцінити хоч би декілька перших документів в ранжируваному списку, тобто тих документів, яким система призначила найвищі оцінки міри релевантності. Система може використати терми цих документів для формування нового, розширеного запиту, який, швидше за все, точніше виражатиме інформаційні потреби користувача.

Такий ітераційний процес обробки запиту і модифікації його за допомогою аналізу даних, отриманих на основі зворотного зв'язку користувача з системою, може повторюватися до тих пір, поки користувач не буде задоволений результатами пошуку. Зворотний зв'язок релевантності використовується в системах, ґрунтованих на різних моделях пошуку.

Автоматичне розширення призначених для користувача запитів. Мається на увазі розширення представлення запиту, спочатку запропонованого системі користувачем. Ця можливість також служить для підвищення ефективності пошуку.

Початкове представлення запиту може поповнюватися за рахунок:

– синонімів термів, що містяться в запиті, якщо система має в розпорядженні тезаурус, що підтримує відношення синонімії;

– термів, які знаходяться з термами запиту в деяких інших семантичних стосунках, визначених тезаурусом предметної області, наприклад представляють частину поняття, що відповідає деякому терму запиту, і тому подібне;

– термів результуючих документів, оцінених користувачем як релевантні або не релевантні, в системах, що забезпечують зворотний зв'язок релевантності;

– часто помилкових форм деяких термів запиту, що зустрічаються орфографічно, і так далі.

Автоматичне індексування документів. Дослідження, проведені ще на ранніх стадіях розвитку систем текстового пошуку, показали, що автоматичне індексування документів не поступається за якістю ручному індексуванню. Тому в сучасних розвинених системах використовується автоматичне індексування.

Мультимовний пошук. Деякі системи текстового пошуку дозволяють здійснювати пошук в колекціях, що містять документи, представлені на декількох природних мовах. Однією із складних проблем, які при цьому виникають, є ідентифікація мови, на якій представлений оброблюваний документ або його фрагменти.

Латентно-семантичний аналіз або індексування(LSA/LSI) – це теорія і метод для витягання "прихованих" контекстно-залежних значень термів і структури семантичних взаємозв'язків між ними шляхом статистичної обробки великих наборів текстових даних. Цей метод широко використовується в зоні пошуку і в завданнях класифікації інформації. Цей підхід дозволяє автоматично розпізнавати смислові відтінки слів залежно від контекстів їх використання. Цей підхід реалізує виявлення тематичної близькості термів, яка потім використовується для обчислення оцінок тематичної близькості документів. Метод LSA широко застосовується у факторному аналізі. Завданням факторного аналізу є виділення головних чинників з простору елементарних. У більшості випадків завдання знаходження головних чинників вирішується за допомогою методу алгебри головних компонент і сингулярного розкладання матриць. У разі інформаційного пошуку під чинниками розуміються деякі семантичні сутності, які частенько не мають певних назв, вибір яких – відкрите завдання.

Матричний латентно-семантичний аналіз. Метод кластерного аналізу LSA/LSI базується на сингулярному розкладанні матриць. Нехай масиву документів ставиться у відповідність матриця A, рядки якої відповідають документам, а стовпці – ваговим значенням термів (розмір словника термів – m).

Сингулярним розкладанням матриці рангу r розмірності $m \times n$ називається її розкладання вигляду

$$A = S \times U \times V ,$$

де U та V – ортогональні матриці розмірності $m \times r$ і $r \times n$, відповідно, а S – діагональна матриця, діагональні елементи якої ненегативні. Діагональні елементи матриці S називають сингулярними значен-

нями матриці A. Помітимо, що матриця S, на відміну від матриці A, квадратна. Приведене вище розбиття матриці має ту властивість, що якщо в матриці S залишити тільки k найбільших сингулярних значень (позначимо таку матрицю як), а в матрицях U та V – ті, що тільки відповідають цим значенням колонки то матриця A буде найкращою по Фробениусу апроксимацією початкової матриці матрицею з рангом, k, що не перевищує r.

Відповідно до методу LSA лише k найбільших сингулярних значень матриці A визначають k-мірний факторний простір, на який проектуються як документи (за допомогою матриці V), так і терміни(за допомогою матриці U). У отриманому факторному просторі документи і терміни групуються в масиви (кластери), що мають деякий загальний сенс, не заданий в явному виді, тобто латентний.

Вибір найкращого значення k для LSA – це проблема окремих досліджень. У ідеалі, k має бути досить велике для відображення усієї реально існуючої структури даних, але в той же час досить мало, щоб не враховувати випадкових залежностей.

У практиці інформаційного пошуку особливе значення відводиться матрицям U та V . Як вказувалося раніше, рядки матриці розглядаються як образи термів в k-мірному речовому просторі. Аналогічно, стовпці матриці розглядаються як образи документів в тому ж k-мірному просторі. Іншими словами, ці вектори задають шукане представлення термів і документів в k - мірному просторі прихованих чинників.

При інформаційному пошуку, в результаті того, що відкидаються найменш значимі сингулярні значення, формується простір ортогональних чинників, що грають роль узагальнених термів. В результаті відбувається "зближення" документів з близьких за змістом предметних областей, частково вирішується проблеми синонімії і омонімії термів.

Висновок

Враховуючи, що ідеальний результат пошуку повинен задовольняти вимогам єдності, повноти і несуперечності, отримуємо, що різні види пошуку визначають різні вимоги до функціональних можливостей системи в частині оцінювання результату. З точки зору використання комп'ютерної техніки, "інформаційний пошук" – сукупність логічних і технічних операцій, що мають кінцевою метою знаходження документів, відомостей про них, фактів, даних, релевантних запиту споживача. Найкращі результати дає латентно-семантичний метод, тому що:

– метод є найкращим для виявлення латентних залежностей усередині безлічі документів;

– метод може бути застосований як з навчанням, так і без навчання (наприклад, для кластеризації);

– використовуються значення матриці близькості, ґрунтованої на частотних характеристиках документів і лексичних одиниць;

– частково знімається полісемія і омонімія.

Список літератури

1. Економічна інформатика: Введення в економічний аналіз інформаційних систем: Підручник. – М.: Инфра-м, 2005. – 344 с.

2. Сычев А.В. Інформаційно-пошукові системи / А.В. Сычев. – К.: Наук. думка, 2002. – 228 с.

3. Маннинг К. Введення в інформаційний пошук / К. Маннинг, П. Рагхаван, Х. Шютце. – Вільямс, 2011. – 622 с.

Надійшла до редколегії 18.05.2012

Рецензент: д-р техн. наук, проф. Є.П. Путятін, Харківський національний університет радіоелектроніки, Харків.

МОДЕЛИ И МЕТОДЫ УЛУЧШЕНИЯ РЕЛЕВАНТНОСТИ ПОИСКА ТЕКСТОВЫХ ДОКУМЕНТОВ

М.В. Токман, В.В. Сокол, Н.С. Лесная

Проведен анализ необходимости улучшения релевантности поиска текстовых документов. Сформулирована задача улучшения поиска. Рассмотрены методы улучшения поиска текстовых документов. Проанализированы критерии отбора наиболее эффективных методов улучшения релевантности поиска текстовых документов. Объяснено значение термина качество поиска при поиске текстовых документов. Среди рассмотренных методов выделен метод, который наиболее соответствует указанным критериям. Описаны преимущества метода латентно-семантического поиска.

Ключевые слова: *поиск текстовой информации, релевантность, критерии релевантности.*

MODELS AND METHODS OF IMPROVEMENT OF RELEVANCE SEARCH OF TEXT DOCUMENTS

M.V. Tokman, V.V. Sokol, N.S. Lesna

The analysis of the need to improve the relevance of searching text documents was performed. The problem of improving search was described. Methods of improving the search text documents were analyzed. The analysis criteria most effective methods to improve the relevance of the search text documents was done. Explanation of the meaning of quality of search on the search of text documents was given. Among these methods selected the one that best matches the specified criteria. Description of the advantages of the method of latent-semantic indexing was given.

Keywords: *text information search, relevance, relevance criteria.*