МОДЕЛІ ТЕХНОЛОГІЙ ІНТЕЛЕКТУАЛЬНОГО ПОШУКУ В ЕЛЕКТРОННИХ БІБЛІОТЕКАХ

Т.Р. Стисло, Я.А. Пилип, О.Р. Стисло

Україна, м. Івано-Франківськ, Івано-Франківський національний технічний університет нафти і газу

76019, м. Івано-Франківськ, вул. Карпатська 15

Існуючі системи видобування інформації даних та знань використовують методи індексації при роботі з документами. В даному контексті під індексом розуміють ключове слово, або групу ключових слів, що мають певну семантичну інтерпретацію. З програмної точки зору індекс – це певне слово, що зустрічається в тексті документу, що є частиною певної колекції. Тому побудова методу видобування базуючись на індексації є достатньо простим процесом з точки зору реалізації, проте існує ряд проблем: 1. Реалізація даної процедури виходить з припущення про те, що семантика інформаційного запиту користувача може бути вирішена на основі множини індексних входжень (семантичних термів). Проте як показує досвід такий підхід є суттєвим спрощенням проблеми семантичного аналізу, оскільки у більшості випадків семантична інтерпретація запиту природною мовою не може бути замінена множиною ключових слів, тобто прийняття такого спрощення на практиці означає втрату певної частини семантичних коннотацій; 2. Семантичні втрати також неминучі при співставленні шаблону ключових слів в документі з відповідним шаблоном запитів, що базуватиметься на неточно описаному просторі індексованих ключових слів (термів); 3. У більшості користувачів відсутній достатній досвід ефективного формулювання інформаційних запитів на основі ключових слів.

Найбільш класичними моделями видобування інформації та документів вважаються булева, векторна та імовірнісні моделі [1,2]. У булевих моделях документи і запити представляються як множини індексованих термів. У векторних моделях, документи і запити представляють як векторні, тобто засобами алгебраїчної інтерпретації. В імовірнісних моделях структура документів і запитів базується на методах і засобах теорії імовірності.

В традиційних системах видобування інформації колекції документів розглядаються як статичні, а запити користувача, як динамічні. Така модель використовується у переважній більшості застосувань, хоча в деяких задачах необхідна зворотна функціональність з статичними запитами і відповідно динамічною циркуляцією документів. В даному контексті також доцільно розглядати задачу фільтрації з введенням профілю користувача, як моделі на основі його преференцій. Наявність такого профілю дозволить виконувати сортування вхідних документів шляхом визначення для кожного документа відповідної цільової групи користувачів які можуть бути у ньому зацікавлені. В умовах університету, де автоматизація рішень електронної бібліотеки базується на системі АБІС «УФД/Бібліотека», дана властивість є особливо актуальною при сортуванні нових інформаційних входжень, оскільки виконання ефективної фільтрації дозволятиме визначати цільові аудиторії потенційних користувачів бібліотечних сервісів. Крім того, система може виконувати ранжування фільтрованих документів і виводити такі результати користувачу, що дозволить економити час користувача виходячи з припущення, що ранжовані документи вверху списку будуть найбільш релевантними стосовно задачі інформаційного пошуку. Таким чином, застосування ранжувань дозволяє визначити множину потенційних релевантних документів, при чому релевантність може оцінюватися деяким пороговим значенням. Найбільш застосовною для ранжування документів є саме векторна модель. Проте ключовим питанням у даному контексті залишається механізм формування ефективних профілів користувача, що дозволить відображати його преференції. В існуючих підходах до побудови профілю користувача слід виділити такі рішення: 1) опис профілю на основі множини ключових слів є простим рішенням, проте користувач не завжди може чітко вказати усі ключові слова, що задають його преференцій; 2) динамічні методи побудови профілю користувача

використовують лише деякі початкові набір ключових слів для ініціалізації профілю. Наступне розширення профілю користувача відбувається шляхом відслідковування його вибору щодо релевантних та не релевантних документів. Такий метод динамічної підтримки профілю користувача є ефективним при відносній стабільності його інтересів.

Таким чином, формулювання ефективних методів ранжування документів доцільно виконувати на основі виділення деякої базової моделі, що визначатиме процес видобування інформації у цілому. Очевидно, що у якості складових такої моделі слід розглядати: 1) множину складених логічних представлень для колекцій документів; 2) множину складених логічних представлень, що описує інформаційні потреби користувача (у першому наближенні такі представлення слід розглядати як запити); 3) ведення деякої загальної структури для моделювання представлень документів та запитів, та відношень між ними; 4) ведення функцій ранжування для виконання зв'язування представлень документів та представлення запитів, що дозволить виконати впорядкування множини документів стосовно послідовності запитів. Основними аспектами реалізації, таким чином є представлення документів і представлення інформаційних потреб користувачів. Використовуючи такі представлення ми можемо утворювати формальні структури для їх моделювання. Крім того, такі формальні структури повинні також забезпечувати засоби для побудови функцій ранжування. Наприклад, для класичної булевої моделі така формальна структура утворюється з множини документів і відповідно основними інструментами є стандартні операції, що вводяться для множин. Для векторно-алгебраїчної моделі така структура утворюється з векторного простору заданої розмірності і відповідних операцій лінійної алгебри над векторами. Відповідно для імовірнісної моделі така структура утворюється з множин стандартних імовірнісних операцій та теореми Байєса як центрального елемента.

Запити, що задаються як булеві вирази можуть бути інтерпретовані у термінах визначення семантики, тому враховуючи простоту і чітку формалізованість булевої моделі, вона ефективно застосовується в багатьох бібліотечних системах, зокрема у системі АБІС «УФД/Бібліотека». Проте, незважаючи на успішність застосування дана модель має також ряд недоліків: 1) стратегія видобування базується на бінарному критерії рішення, тобто кожен документ розглядається як релевантний або як не релевантний, що суттєво знижує ефективність процесу видобування, тому булева модель у більшості випадків розглядається як модель видобування саме даних; 2) незважаючи на семантичну чіткість та точність булевих виразів завдання переведення інформаційного запиту користувача у відповідний булевий вираз не є простим та очевидним, тому ряд користувачів не можуть без попередньої підготовки чітко будувати свої запити у формі булевих виразів. На рівні представлення булева модель розглядає індексні терми як такі, що наявні або відсутні у документі, тому вагові значення індексних термів розглядаються як бінарні, а запити утворюються з індексних термів, що зв'язуються трьома сполучниками (і, або, заперечення). Таким чином, запит можна розглядати як звичайний булевий вираз, що представляється як дизюнкція кон'юктивних векторів. Таким чином булева модель дозволяє визначати відповідно релевантність або нерелевантність кожного документа. Основним недолік є те, що бінарна структура формальної основи дозволяє відповідно отримувати або надто багато результатів або недостатню кількість результатів, крім того в рамках булевої моделі неможлива реалізація відбору документів по частковому співпадінню.

Векторна модель дозволяє враховувати обмеження щодо бінарних ваг булевої моделі і відповідно пропонує користувачам структуру у якій можливо реалізувати часткове співпадання. Дане завдання вирішується шляхом присвоєння не бінарних вагових значень до індексних термів в запитах та документах. Такий підхід дозволяє обчислювати ступінь схожості між запитом користувача і кожним документом, що зберігається у базі даних. Відповідно видобуті документи сортуються в спадному порядку по степеню подібності, що дозволяє векторній моделі розглядати документи з частковим співпаданням по термах запиту. Практичне застосування показує, що результатуюча множина для векторної моделі є більш точною щодо відповідності семантичному значенню запиту порівняно з відповідною

результатуючою множиною для булевої моделі, оскільки відповідно до ступеня подібності документ може видобуватися навіть якщо його відповідність запитам є частковою або принаймі більшою за встановлене граничне значення.

Таким чином ввведена в векторній моделі ступінь схожості по внутрішній кластеризації може бути квантифікована шляхом вимірювання частоти входження виділеного терма всередині документа і відповідно дозволятиме визначити наскільки даний терм описує контент документа. У якості основних переваг векторної моделі можна виділити покращення ефективності видобування за рахунок нової схеми порових коефіцієнтів термів, можливість видобування документів, що наближено (частково) задовольняють умови запиту за рахунок наявної стратегії часткового співпадання, можливість ранжування документів по ступеню схожості. Проте, основним недоліком є те, що дана схема функціонує на основі припущення про те, що індекси термів є взаємонезалежні. Крім того ранжування множини в результаті можна покращити тільки шляхом розширення запиту та підвищення релевантності зворотного зв'язку.

Основна особливість імовірнісних моделей полягає у розгляді проблеми видобування даних інформації та документів в рамках імовірнісних представлень. А саме, для заданого користувачем запиту очевидно існує множина документів, що містить тільки релевантні входження. Така множина документів називається ідеальною множиною відповідей. Маючи опис такої ідеальної множини можна гарантовано видобути документи, що у ній містяться і таким чином процес побудови запиту можна розглядати як процес специфікації властивостей ідеальної множини відповідей. Проте проблема полягає у тому, що не завжди такі властивості є відомими. Відомо тільки, що існують індексні терми, семантика яких повинна бути використана для характеристики даних. Оскільки це не можливо зробити під час виконання запиту, то потрібно попередньо згенерувати імовірнісний опис ідеальної множини відповідей, що буде використана для побдови певної множини документів. Після чого можна ініціалізувати взаємодію з користувачем з метою покращення імовірнісного опису ідеальної множини можливих відповідей. Для цього користувач перевіряє список видобутих документів і вказує які з них є релевантними. Система використовує дану інформації для уточнення опису ідеальної множини відповідей. Багаторазове повторення даної процедури дозволить отримати достатньо точний опис. Таким чином для заданого запиту і документа у колекції імовірнісна модель намагається оцінити імовірність того, що користувач оцінить певний документ як релевантний. В даній моделі імовірність релевантності залежатиме від запиту і способу представлення документів. У якості наступного кроку виконується припущення про те, що існує підмножина документів які користувач розглядатиме як множину відповідей для запиту. Побудова такої множини відповідей дозволить максимізувати загальну імовірність релевантності. Проте, складність полягає у обчисленні імовірності релевантності. Таким чином, у якості основної переваги імовірнісної моделі є ранжування результатів по спаданню ступеня їх релевантності. В якості недоліків слід визначити необхідність розподілу множини документів на релевантні та не релевантні, не врахування частоти входжень індексних термів у документі, початкове припущення щодо незалежності індексних термів.

Таким чином, проведений аналіз показує, що серед класичних моделей видобування інформації даних та документів булева модель є найслабшою, оскільки не дозволяє використовувати метод часткових співпадань. Найкращі результати щодо практичного впровадження в загальному випадку дає імовірнісна модель Проте для добре структурованих колекцій документів ефективними є також алгебраїчні моделі.

Список використаної літератури

1. Grossman D.A., Frieder O. – Information Retrieval: Algorithms and Heuristics.– Kluwer Academic Publishers.– 1998.–300p.

2. Gudivana V.,Raghavan V., Grosky W., Kasanogottu R.– Information retrieval on the World Wide Web.– IEEE Internet Computinge.– Oct-Nov:58-67.– 1997.–P.100-150.