

Недоліком цього методу є низька пошукова здатність на мовах відмінних від російської та залежність результату запиту від реєстра його набору. Це не дозволяє отримувати релевантні відповіді на пошукові запити і в свою чергу, не дає можливості виходу системи на зарубіжний ринок. Даний недолік свідчить про відсутність універсальності пошукового алгоритму.

Найбільш близьким по сукупності ознак до запропонованого нами, є метод «Page Rank» пошукової системи «Google» в якому релевантність документа визначається на основі інформації про документи, які посилаються на нього, залежить від їх числа, а також релевантності цих посилань. Такий підхід дозволяє забезпечити видачу якісних результатів.

Недоліком цього методу є лінійне збільшення інформаційного «шуму» в залежності від збільшення кількості посилань. Це обумовлено тим, що документи можуть посилатися на ресурси з застарілою або видаленою інформацією, які не оновлюються власником. Це створює ризик отримати відповіді, які не будуть повністю задовольняти умовам запиту.

Актуальність дослідження. Актуальність поставленого дослідження обумовлена існуванням певних проблем у напрямку інформаційного пошуку, а саме:

- зріст інформаційного контенту, який супроводжується зростом інформаційного шуму та багаторазовим дублюванням інформації [8];
- низька структурованість даних;
- відсутність єдиної системи стандартизації та концепції розвитку Інтернет та відсутність центра управління розвитком глобальної мережі;
- постійна зміна технологій зберігання та представлення даних;
- зростання вимог з боку користувачів пошукових систем.

Формулювання мети статті. Метою даної статті є дослідження можливостей пошуку і ранжирування даних для надання найбільш якісного, релевантного результату. Необхідно розробити такі алгоритми пошуку, які могли б наслідувати можливості алгоритмів конкурентних систем, а також внести нові ідеї, тим самим зробивши пошукову систему унікальною. Основне вирішення даної проблеми полягає у впровадженні методів інтелектуальної обробки інформації.

Мета роботи пошукового механізму полягає в організації пошуку релевантної інформації в інформаційних мережах за довільним запитом.

Поряд з цим, метою системи є пошук таких документів колекції, які є найбільш релевантними по відношенню до довільних інформаційних потреб користувача. Документ називається релевантним, якщо, з точки зору користувача, він містить коштувану інформацію, яка задовольняє його інформаційні потреби.

Основна частина

Функціонування механізму пошукової системи можна поділити на два основні, незалежні один від одного завдання: індексація метаданих отриманих від пошукового агента і організація пошуку на підставі запиту користувача і індексованих в системі документів. Для отримання виграшу у швидкості пошуку, проводиться індексація ресурсів і на їх підставі виконується ранжирування інформації. Процес індексації складається з наступних етапів:

- збір документів, які підлягають індексації;
- представлення тексту у вигляді лексем;
- попередня обробка інформації (стеммінг, лематизація);
- індексація документа.

Індексація документа. У колекції кожен документ має власний ідентифікатор. У процесі побудови індексу даний ідентифікатор або присвоюється новим документом, або оновлюється, якщо документ вже існує. Вхідною інформацією для індексування є набір нормалізованих лексем для кожного документа, який розглядається як список пар «термін-документ».

Основним етапом в процесі індексування є сортування списку термінів, в результаті чого вони розташовуються в алфавітному порядку. Після цього зразки одного і того ж терміну групуються, а результат поділяється на таблицю атрибутів і таблицю словосполучень. Таким чином індексації документ стає придатним для проведення пошуку.

Алгоритми пошуку. Нами було розроблено та розглянуто три алгоритми пошуку які у поєднанні між собою підвищують його ефективність. Перший з них алгоритм зваженого зонного ранжирування.

Веб-документи супроводжуються метаданими, які кодуються у вигляді, доступному для розпізнавання сервером. Під метаданими представляється неоднорідна інформація про терміни документа, яка містить сукупність ознак таких як: присутність терміна в заголовку сторінки або мета-тегах, чи є термін посиланням на інший документ, жирним, курсивним, підкресленим тощо.

Такі ознаки називаються зонами і дозволяють збільшити або зменшити важливість терміна в порівнянні з іншими термінами оброблюваного документа. Така підтримка обробки запитів полягає в створенні таблиці атрибутів для індексованих документів яка вказує на приналежність терміна до тієї чи іншої зони. Даний підхід дозволяє забезпечити алгоритм зваженого зонного ранжирування.

Даний алгоритм привласнює для пари документа (d) та запита (q) значення релевантності на відрізьку [0..1], обчислюючи лінійну комбінацію зонних показників, до якої кожна зона документа вносить булеве значення. Нами було розглянуто колекцію документів, кожен з яких має l-зон. так що:

$$\sum_{i=1}^1 g_i = 1 \quad (1)$$

де g_i – коефіцієнти зон а також $g_i \in [0.1]$.

Зважаючи на цю умову зважену зонну релевантність розраховують по формулі:

$$\sum_{i=1}^1 g_i * s \quad (2)$$

де s – булева величина, що означає відповідність (або її відсутність) між запитом (q) і i -ою зоною.

Ваги $g_1 \dots g_i$ вказуються експертами або користувачем. Однак набагато частіше ваги визначаються на основі навчальних прикладів, оцінених заздалегідь. Цей метод відноситься до загального класу методів ранжирування в інформаційному пошуку під назвою «методи ранжирування на основі машинного навчання».

Наступним алгоритмом є ранжирування на основі частот входження термінів.

При використанні зваженого зонного ранжирування ранг документа залежить від наявності термінів запиту в зонах документа. Наступним кроком є визначення частот входження термінів у документах: документ, де термін запиту зустрічається частіше, слід вважати більш релевантним і привласнити йому більш високе значення релевантності. Це обґрунтовується тим, що терміни оброблюються інтерфейсом пошукової машини у вільному вигляді, без сполучних операторів.

Такий стиль, розглядає запит як безліч слів. Отже, для підрахунку показника документа достатньо підсумовувати показники його відповідності кожному з слів запиту. Для цього кожному терміну, виявленому в документі привласнюється вага, що залежить від кількості появ цього терміна в даному документі. Для оцінки відповідності між терміном запиту і документом використовується три схеми зважування.

Перша з них полягає в тому, що вага терміна дорівнює кількості входжень його до документу. Ця схема зважування називається частотою терміна і позначається як tf_{id} , де індекс (t) позначає термін, а індекс (d) - документ. В рамках цієї схеми точний порядок термінів у документі ігнорується, а основна увага приділяється кількості входжень кожного терміна до документу. Для коригування значимості термінів документа між собою застосовується модифікація обчислення частоти терміна названа сублінійним масштабуванням. Вона полягає у використанні логарифма до частоти термінів:

$$wf = \begin{cases} 1 + \log(tf), & tf > 0 \\ 0, & tf \leq 0 \end{cases} \quad (3)$$

де tf – частота терміну у документі, wf – сублінійна частота.

Представлена модифікація дозволяє зменшити важливість найбільш повторюваних термінів.

Така схема зважування має серйозний недолік: при ранжируванні документа за запитом всі терміни вважаються однаково важливими. Насправді, деякі терміни мають малу або нульову розрізняльну силу при визначенні релевантності. Для того щоб усунути зазначений недолік, вводиться механізм послаблення впливу «популярних» термінів. Для цього використовується друга схема зважування яка полягає у використанні документної частоти - кількості документів в колекції, що містять термін.

Це обумовлено тим, що, намагаючись знайти відмінності між документами з метою їх ранжирування за запитом, доцільніше використовувати статистичні показники самих документів (наприклад, кількість документів, що містять заданий термін), ніж статистичні показники колекції в цілому. Для коригування ваги терміна з використанням документної частоти необхідно визначити зворотню документну частоту за формулою:

$$idf = \log \frac{N}{df} \quad (4)$$

де N – загальна кількість документів в колекції, df – документна частота, idf – зворотня документна частота.

Найбільш ефективною схемою зважування є комбінація частоти терміна в документі і зворотної документної частоти. Дана схема привласнює кожному терміну в документі його коефіцієнт на підставі формули:

$$wf = idf = wf_{t,d} * idf_t \quad (5)$$

де wf – сублінійна частота для терміна (t) у документі (d), idf – зворотня документна частота, $wf-idf$ – коефіцієнт терміна у документі.

Вага $wf-idf$ має такі властивості:

- досягає максимального значення, якщо термін зустрічається багато раз в невеликій кількості документів (тим самим посилюючи їх відмінність від інших документів);

- зменшується, якщо термін зустрічається в документі лише кілька разів або зустрічається в багатьох документах (тим самим формуючи менш виражений сигнал про релевантність документа);

- досягає мінімального значення, якщо термін зустрічається практично у всіх документах [9].

Після визначення ваг всіх термінів запиту в усіх існуючих документах визначається міра перекриття: релевантність всіх документів дорівнює сумі входжень всіх термінів запиту в цей документ:

$$S(q, d) = \sum_{t \in d} (wf - idf) \quad (6)$$

де $wf-idf$ – коефіцієнт терміну у документі, $S(q,d)$ – сума коефіцієнтів термінів запиту (q) на документ (d).

Релевантним документом буде вважатися, той документ, в якому сума коефіцієнтів релевантності для всіх термінів запиту буде найбільшою.

