**УДК 519.71**

**К.т.н., доцент Білостоцький А. І., студент Стратієвський О.М.**

**Національний технічний університет України
«Київський політехнічний інститут»**

# МОДЕЛЬ ФІЛЬТРАЦІЇ ДОКУМЕНТІВ ЗА РЕЛЕВАНТНІСТЮ НА ОСНОВІ АВТОАНОТУВАННЯ

## Abstract

***Anatoliy I. Bilostotskiy, assoc. prof., PhD; Oleksandr Stratiyevs'kyy, student***
***Relevance documents filtering model for search in information warehouses***

*This work is due to the approach of relevance documents filtering model for full text search. A model for resolving the problem above, which uses automatic summarization of documents, is presented. The urgency and the adequacy of the proposed theoretical approach are proved.*

## Вступ

Традиційні моделі пошуку не показують належного результату в умовах феномену сучасного інформаційного буму. В той же час колекції документів містять в собі у явному чи прихованому вигляді набори додаткової інформації про тексти, що зберігаються, яку можна використати для організації пошуку. Зокрема для виявлення прихованих знань можна застосувати методи автоматичного аналізу текстів, названі Text Mining [1]. Робота присвячена розробці моделі фільтрації документів, що використовує подібні механізми для підвищення релевантності результатів процесу текстового пошуку. Такий підхід є актуальним, адже наявність динамічних автоматично побудованих структур, якими будуть описані документи в колекції, дозволить точніше співвідносити тексти з пошуковим запитом та більш повно задовольняти інформаційну потребу користувачів [1, 2].

## Постановка задачі

Метою роботи є розробка алгоритму фільтрації документів або модифікація чи комбінація вже існуючих з метою підвищення рівня релевантності процедури текстового пошуку за допомогою автоматичної побудови динамічного внутрішнього представлення текстових даних.

Об’єктом дослідження є процедура інформаційного пошуку, методи та технології представлення текстових даних у інформаційних системах.

Предметом дослідження у свою чергу є текстова інформація, алгоритми пошуку, методи оцінки відповідності текстових даних.

**Проблеми та особливості пошукових систем**

Існуючі інформаційно-пошукові механізми первісно проектувалися для забезпечення релевантності вибірки в поєднанні з вимогою повноти пошуку, але саме в цьому і полягає їхній головний недолік. Неконтрольований рівень пертинентності вибірки при цьому різко знижує ймовірність одержання користувачем потрібної інформації [2, 3].

Причини надлишковості результатів стандартного інформаційного пошуку можуть бути розділені на дві якісно різні категорії: дублювання інформації та інформаційна невідповідність [3]. Пошукові технології розроблюваної моделі розширюються за рахунок застосування додаткових семантичних засобів. Пропонується використати ідею попередньої обробки початкової сукупності документів, що має на меті сформувати деякий ефективний набір даних, що відображає її зміст і призначений для подальшого пошуку по ньому. Також перспективним з існуючих сьогодні напрямків організації пошуку є автоматичне групування результатів пошуку [1, 2]. Головна перевага автоматичного групування полягає в ієрархічній організації результатів пошуку, що дозволяє на першому етапі мати справу з обмеженим набором кластерів. Складність, однак, полягає в тому, що розбивка вибірки на групи здійснюється на підставі формальної близькості документів [3].

**Фільтрація з використанням анотування та групування даних**

Пропонується реалізація принципу попередньої обробки текстового матеріалу [3] за допомогою методик глибинного аналізу текстів, що передбачає автоматичне виділення найбільш значимої інформації і відсіювання "сміття" – фільтрацію. Це дозволить споживачеві працювати з обмеженими за обсягом наборами даних, і може істотно підвищити рівень пертинентності результатів пошуку. Головна ідея – релевантність документа визначається відносно деякого його образу, що називається анотацією [3].

Для побудови моделі фільтрування було використано ряд технологій Text mining. Латентно-семантичне індексування (LSI), що оперує поняттями статистично-незалежних концептів, які будуються на основі матриці значимих термінів, було обрано для побудови анотації тексту [4].

Проте для адаптації цього механізму до вищеозначеної задачі необхідно провести його певну модифікацію. Оскільки виявлення концептів проводитиметься в рамках одного документа, він розбивається на набір підтекстів, кількість та розмір яких залежить від розміру документа. Для оптимізації цього процесу проводиться попереднє виділення значимих термінів та відсіювання фрагментів з низькою частотою таких термінів.

Анотація, побудована таким чином, відповідає вимогам до вхідних даних, які висуваються процедурою кластеризації. Тому для уточнення тематики документа необхідно формалізувати алгоритм віднесення анотацій до того чи іншого кластеру. Так пропонується звести дану задачу до задачі кластеризації найбільш значимих термінів та термінів з заголовків текстів.

**Модель фільтрування документів за релевантністю**

На рисунку 1 представлено розроблювану модель у вигляді діаграми діяльності в нотації UML.

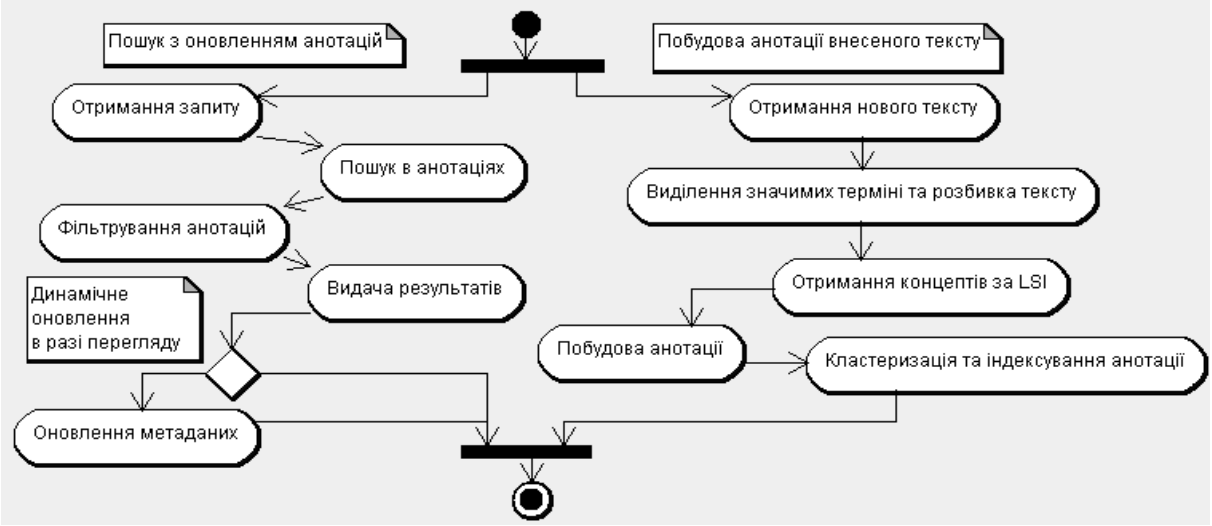


Рис.1. Модель підвищення релевантності за рахунок фільтрації

Пошук організовується на основі системи індексного опису кластеризованих анотацій внесених текстів з використанням схеми метаданих Дублінського базового комплекту елементів метаданих, що був затверджений як стандарт Z39.85 [2]. Передбачається динамічне оновлення комплекту термінами успішних запитів користувачів. Модифікація алгоритму латентно-семантичного індексування для побудови анотації показує, що підхід є цілком адекватним для виявлення тематики колекції термінів. Загалом можна сказати, що введення автоанотування позитивно вплинуло на процедуру пошуку в сенсі

запобігання видачі нерелевантних даних. Більшість експериментів показують поліпшення показників точності та ступеня фільтрації. Можна сказати, що за допомогою автоанотування досягається основна мета розробки – результати пошуку ранжуються за релевантністю. Привабливою темою для подальших досліджень є модифікація цього алгоритму з використанням байєсівських мереж.

## Висновки

У ході виконання роботи запропоновано модель фільтрування документів з заданим рівнем релевантності на основі автоанотування. Побудована математична модель разом з програмним забезпеченням, спроектованим на її основі, пройшли тестування на ЕОМ. Отримані результати показали високий ступінь адекватності отриманої моделі. Слід зауважити, що можливості результатів роботи LSI, за якими будувалася анотація, були використані не у повному обсязі, що дає перспективу для подальших досліджень.

Представлена модель орієнтована на практичну реалізацію і частково долає ряд технологічних обмежень, які містяться в обраному підході. Разом з тим, пропонована організація пошуку дозволить вирішити наступні важливі задачі:

1) автоматичне анотування документів;
2) автоматичне групування документів;
3) видача користувачу суто інформаційно-значимих документів.

Запропонована система організації пошуку дозволить істотно підвищити його привабливість із погляду "середньостатистичного" користувача.

## Література


1. *Ландэ Д. В*. Поиск знаний в Internet. Профессиональная работа: Пер с англ. – М.: Издательский дом “Вильямс”, 2005. – 272 с.
2. *Manning Christopher, Raghavan Prabhakar, Schütze Hinrich.* An Introduction to Information Retrieval . – Cambridge UP, 2009. – 569c.
3. *Григорьев А.Н., Ландэ Д.В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream//Труды Международного семинара "Диалог'2005". - 2005. - С. 109-111.
4. *Дерецький В.О.* Підхід до автоматичної побудови тематичної онтології документа для удосконалення інформаційного пошуку // Проблеми програмування. -2005-N3-С.76-82.