МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНБАСЬКА ДЕРЖАВНА МАШИНОБУДІВНА АКАДЕМІЯ
КАФЕДРА КОМП'ЮТЕРНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАРПОВ О. С.

ДОСЛІДЖЕННЯ МЕТОДІВ, МОДЕЛЕЙ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ОЦІНКИ РЕЛЕВАНТНОСТІ НАУКОВИХ ПУБЛІКАЦІЙ

RESEARCH OF THE METHODS, MODELS AND INFORMATION TECHNOLOGIES FOR ASSESSING THE RELEVANCY OF SCIENTIFIC PUBLICATIONS

Спеціальність 122 – Комп'ютерні науки

АВТОРЕФЕРАТ
на здобуття кваліфікації
магістра з комп'ютерних наук

Краматорськ – 2021

Дипломна робота виконана на кафедрі комп’ютерних інформаційних технологій Донбаської державної машинобудівної академії.

Науковий керівник: професор, д.т.н., завідувач кафедри КІТ Тарасов О. Ф.

Захист дипломної роботи відбудеться «28» травня 2021 року об 11-00 годині у Донбаській державній машинобудівній академії за адресою: 84313, Донецька обл., м. Краматорськ, бул. Машинобудівників, 39, ауд. 2218, кафедра «Комп’ютерні інформаційні технології».

Summary

The purpose of the master's thesis is to increase the productivity and quality of search and analysis of scientific articles based on the use of relevance assessment methods. The object of research is the process of assessing the relevance of the scientific literature. The subject of the study is the comparative effectiveness of the use of relevance assessment methods. When performing the work, the existing methods of text analysis are analyzed. Mathematical models of three methods for assessing the relevance of the text are determined. A method of comparative research of methods for assessing relevance using expert evaluation is formed. Developed program for the implementation of selected methods. A study has shown that for tasks that require a more accurate search result, you can use the combined method of assessing relevance, and for a more complete search - the method of assessing relevance based on the weights of word pairs.

Keywords: relevance, evaluation, search, keywords, scientific publication, tf-idf method, method based on the weights of word pairs, combined method.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми.

Наукова діяльності в кінцевому підсумку це отримання нових знань і створення нових технологій, що сприяють поліпшенню якості життя. Для того щоб отримані результати могли бути використані, необхідно донести інформацію про них до реальних «споживачів», а також до «виробників» нових наукових результатів. Для оцінки рівня досліджуваності питання і передбачуваного попиту на результат роботи можна використовувати ряд методів, таких як аналіз літератури та наукових статей або аналіз відносної кількості пошукових запитів в різних пошукових системах. Але з кожним роком кількість інформації збільшується. Для обґрунтування пошуку та орієнтування в такому обсязі інформації створюють різні пошукові засоби. Всі ці системи мають певні переваги, в числі яких простота та зручність використання, що дозволяє непідготовленому користувачеві відразу приступити до пошуку інформації, сортування результатів пошуку від найбільш релевантних до менш релевантних, відображення заголовка сторінки і невеликого екстракту (зазвичай 2-3 рядки) поряд з посиланням на сайт, що дозволяє скласти перше враження про вміст сайту або виданого результату.

Найновітніші методи оцінки релевантної інформації використовуються в сучасних пошукових системах Інтернету. Але ці методи є скритими, бо вони мають комерційну цінність. На сьогоднішній час є дуже багато літератури, в якій описується використання тих чи інших відкритих методів при оцінці тексту або сайту. Але майже немає порівняння результатів використання цих методів при аналізі наукової літератури.

Тому пошук інформації із набору публікацій за допомогою методів оцінки релевантності дасть змогу виконувати швидко аналіз і фільтрацію наукової літератури, яка з кожним днем збільшується. А дослідження і порівняння цих методів дасть змогу зрозуміти, які з них допоможуть швидко і якісно отримати потрібну інформацію.

Зв'язок роботи з науковими програмами, планами, темами.

У відповідності до Наказу ректора №07-3 від «22» січня 2021 року, автором самостійно було виконано дослідження методів, моделей та інформаційних технологій оцінки релевантності наукових публікацій, що виконується на кафедрі комп'ютерних інформаційних технологій Донбаської державної машинобудівної академії.

Мета і завдання дослідження.

Метою даної робити є збільшення продуктивності і якості пошуку та аналізу наукових статей на основі використання методів оцінки релевантності.

В відповідності з метою роботи виділені наступні завдання:

– аналіз роботи інформаційно-пошукових систем та методів оцінки релевантності;

– розробка математичних моделей методів оцінки релевантності та методики проведення дослідження даних методів;

– вибір засобів розробки програмного забезпечення та розробка технічного завдання на створення програмного комплексу оцінки релевантності наукових публікацій;

– розробка програмно-методичного комплексу;

– проведення дослідження методів оцінки релевантності наукових публікацій;

– аналіз результатів дослідження методів оцінки релевантності.

Об'єкт дослідження.

Процес оцінювання релевантності наукової літератури.

Предмет дослідження.

Порівняльна ефективність використання методів оцінки релевантності.

Методи дослідження.

У роботі використовувалися методи оцінки релевантності, такі як метод TF-IDF, метод релевантності вагів по парам слів, та комбінований метод на основі методів TF-IDF та вагів по парам слів. Для їх аналізу використовувалися наступні методи експериментальних досліджень: експертний аналіз для оцінки релевантності

наукових публікацій та порівняльний аналіз експертних оцінок з результатами роботи методів оцінки релевантності. Порівняльний аналіз результатів роботи відбувався за критеріями ефективності інформаційного пошуку.

Наукова новизна.

Розроблена математична модель приведення до спільного діапазону значень релевантності при комбінуванні методів.

Проведено дослідження якості роботи методів оцінки релевантності за критеріями ефективності інформаційного пошуку при аналізі наукових публікацій та статей.

Практичне значення отриманих результатів.

На підставі отриманих результатів дослідження методів оцінки релевантності був розроблений програмно-методичний комплекс оцінки релевантності наукових публікацій. Даний комплекс дозволяє користувачеві отримати оцінку релевантності наукових публікацій до обраних ключових за різними методами, а також визначити які саме публікації є релевантними. Це допомагає збільшити продуктивність і якість пошуку потрібної літератури. Структура створеного ПМК дозволяє розширювати фунціонал шляхом додання нових методів оцінки релевантності для подальшого порівняльного аналізу даних методів.

Публікації.

Опубліковано статтю у збірнику «Студентський Вісник ДДМА 2021», тези доповідей у матеріалах V Всеукраїнській науково-технічній конференції «Сучасні інформаційні технології, засоби автоматизації та електропривод».

Структура та обсяг роботи.

Дипломна робота складається із вступу, шести розділів, висновків, переліку використаної літератури із 41 найменування, 29 рисунків, 70 таблиць та 7 додатків. Загальний обсяг дипломної роботи складає 155 сторінок, включаючи 132 сторінок основної частини та 23 сторінок додатків.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** визначено проблеми пошуку інформації, обґрунтовано актуальність роботи, сформульована мета та завдання дослідження методів оцінки релевантності, а також об’єкт та предмет даного дослідження.

У **першому розділі** розглянуті інформаційно-пошукові системи, їх призначення та класифікації, виділення поняття релевантності та принцип використання цього поняття. Виділено дві групи систем: системи, які мають доступ до бази документів і системи, що використовують мережу Інтернет, а також проаналізовані програмні продукти та їх властивості у цих видах систем. Також виділено три типи аналізу тексту: кластерний аналіз, лінгвістичний аналіз та статичний аналіз і проаналізовані методи оцінки релевантності цих груп.

У **другому розділі** обрані теоретичний та експериментальний методи дослідження, а саме експертний аналіз як теоретичний метод для експертної оцінки набору публікацій на відповідність до теми та порівняльний аналіз як експериментальний метод для порівняння результатів алгоритмів та експертної оцінки. Розроблені математичні моделі обраних методів оцінки релевантності: методу TF-IDF, методу пошуку релевантності на основі вагів по парам слів, комбінованого методу, на основі TF-IDF та вагів по парам слів. Розглянуті критерії ефективності інформаційного пошуку та побудований план дослідження ефективності цих методів.

У **третьому розділі** проаналізовано механізм оцінки релевантності тексту. На основі цього була розроблена логічна та фізична модель ПМК оцінювання релевантності наукових публікацій «Article Analyzer». Реалізація ПМК відбувалася за допомогою шаблонів проектування «Абстрактна фабрика» та «Адаптер». Також продемонстрований інтерфейс користувача.

У **четвертому розділі** проведений експертний та порівняльний аналіз, який показав, що найбільш точним є комбінований метод пошуку релевантності, а найбільш повним на результат пошуку є метод релевантності на основі вагів пар слів.

Усі три обрані методи повернули більшу частину публікацій, які за експертною думкою є релевантними.

У **п’ятому розділі** при економічному порівнянні зі схожими програмними продуктами було виявлено, що розроблений ПМК має перевагу в ряді оціночних факторів, а саме надійність, супровідність, зручність застосування, коректність. З’ясовано, що заявлена робота має високий ранг новизни і готовності до використання, бо присутній високий ранг опрацьованості теми дослідження.

У **шостому розділі** проведено аналіз небезпечних і шкідливих виробничих факторів. На основі цього розроблені заходи щодо забезпечення безпечних умов праці та підвищення стійкості роботи в умовах надзвичайної ситуації.

ЗАГАЛЬНІ ВИСНОВКИ

Аналіз процесу пошуку релевантної інформації дав змогу зрозуміти, якими основними принципами керуються пошукові системи при аналізі та видачі потрібної інформації. Аналіз існуючих інформаційних систем допоміг зрозуміти, на які типи вони поділяються та які переваги і недоліки мають такі системи. Аналіз існуючих методів пошуку інформації допоміг визначити, де саме і як вони застосовуються, а також які саме методи будуть застосовані для дослідження та комбінування. Із цього аналізу для дослідження було виділено три методи оцінки релевантності:

- метод TF-IDF;
- метод пошуку релевантності на основі вагів по парам слів;
- комбінований метод, на основі TF-IDF та вагів по парам слів.

Результати роботи методів порівнювалися з експертним аналізом публікацій. Це дало змогу оцінити, наскільки точно і якісно працюють дані методи. Також були розроблені математичні моделі методів оцінки релевантності, за допомогою який стало можливим програмно реалізувати дані методи для дослідження. Розроблений план дослідження допоміг визначити послідовність дій і критерії, за якими були проаналізовані методи оцінки релевантності. Розробка технічного завдання на створення ПМК оцінки релевантності наукових публікацій дала змогу зрозуміти, що

саме та як повинне бути реалізовано в програмному продукті, за допомогою якого будуть аналізуватися методи.

Розробка логічної моделі дослідження методів оцінки релевантності наукових публікацій показала основні можливості системи оцінки релевантності, а саме вибір ключових слів та критеріїв релевантності для методів оцінки та можливість аналізу як однієї публікації, так і всього набору. Відображення процесу та послідовності дій для оцінки релевантності публікацій різними методами дало змогу зрозуміти принцип роботи цих обраних методів оцінки релевантності.

Використання шаблонів проектування «Абстрактна фабрика» та «Адаптер» при розробці покращило написання коду, а саме це дозволило зробити код створення об'єктів універсальним і встановити зручний зв'язок між паралельними ієрархіями класів, стало можливим багаторазове використання коду. Розробка інтерфейсу дозволила зручно взаємодіяти з програмою та виводити результати роботи методів у зрозумілому для людини форматі.

При виконанні дослідження були обрані ключові слова та 20 наукових публікацій. Експертний аналіз показав, що релевантними були 11 із 20 публікацій. За допомогою розробленого програмного продукту було проаналізовано цей набір публікацій на релевантність до обраних ключових слів. Порівняльний експеримент цих результатів показав, що найбільш точним став комбінований метод оцінки релевантності з оцінкою 0,86 при зниженні пошукового шуму до 0,14, а найбільш повним став метод оцінки релевантності на основі вагів пар слів з повнотою пошуку 0,82 із втратою інформації лише 0,18.

Із цього можна визначити, що для задач, які потребують більш точного результату пошуку можна використовувати комбінований метод оцінки релевантності, а для більш повного пошуку - метод оцінки релевантності на основі вагів пар слів.

Реалізація методів дозволила збільшити продуктивність та зменшити час для аналізу літератури та наукових статей в потрібній предметній області, а структура ПМК дозволяє розширювати фунціонал для подальшого покращення аналізу публікації.

ПЕРЕЛІК НАУКОВИХ ПРАЦЬ

Карпов О. С., Тарасов О. Ф., Автоматизація оцінки релевантності наукових публікацій в процесі інформаційного пошуку: матеріали V Всеукраїнській науково-технічній конференції «Сучасні інформаційні технології, засоби автоматизації та електропривод». – [Укр. мова.]

2. Карпов О. С., Тарасов О. Ф., Автоматизація оцінки релевантності наукових публікацій в процесі інформаційного пошуку. [Прийнята до друку]: Студентський вісник ДДМА : тематичний збірник наукових праць. – Краматорськ : ДДМА, 2021.

Анотація

Метою магістерської роботи збільшення продуктивності і якості пошуку та аналізу наукових статей на основі використання методів оцінки релевантності. Об’єктом дослідження є процес оцінювання релевантності наукової літератури. Предметом дослідження є порівняльна ефективність використання методів оцінки релевантності. При виконанні роботи проаналізовано існуючі методи аналізу тексту. Визначені математичні моделі трьох методів оцінки релевантності тексту Сформована методика порівняльного дослідження методів оцінки релевантності з використанням експертного оцінювання. Розроблений ПМК для реалізації обраних методів. Проведено дослідження, яке показало, що для задач, які потребують більш точного результату пошуку можна використовувати комбінований метод оцінки релевантності, а для більш повного пошуку - метод оцінки релевантності на основі вагів пар слів.

Ключові слова: релевантність, оцінка, пошук, ключові слова, наукова публікація, метод tf-idf, метод вагів по парам слів, комбінований метод.