УДК 004.412:519.237.5

FUZZY SEARCH МЕТОДИ ДЛЯ ФІЛЬТРАЦІЇ СПИСКІВ

Бойченюк А.В.[1]; Молодець Б.В., PhD.[2]
[1,2]Національний технічний університет "Дніпровська політехніка"
[1,2]Україна, Дніпро
[1]antonboichenuk@gmail.com; [2]bogdan.molodets@gmail.com

Анотація. *В роботі розглянуто особливі моменти застосування методів нечіткого пошуку Fuzzy Search для фільтрації списків у сучасних інформаційних системах. Визначено основні алгоритми нечіткого пошуку, їхні переваги, недоліки та області застосування. Проаналізовано ефективність використання Fuzzi Search у задачах обробки текстових запитів, пошуку даних у базах та інтерфейсах користувачів.*
Ключові слова: *фільтрація списків; нечіткий пошук; Levenshtein; Jaro-Winkler; n-grams; Soundex.*

Вступна частина. Сучасні інформаційні системи, орієнтовані на збір, обробку та відображення великих обсягів даних не можуть існувати без інструментів пошуку та фільтрації. При цьому перед розробниками інформаційних систем ставиться задача розробити зручний інструмент, який має враховувати наступний фактор, як орфографічні помилки, введення скорочень та інші неточності, які допускаються програмами-скраперами і живими користувачами при пошуку та зборі даних з інформаційних систем.

Для вирішення цієї задачі використовуються методи нечіткого пошуку Fuzzy Search, які ґрунтуються на визначенні ступеня подібності між текстовими рядками. На відміну від точного пошуку, який використовує однозначне співпадіння, алгоритм нечіткого пошуку порівнює введений запит із кожним елементом бази даних, обчислюючи відстань або коефіцієнт схожості. Чим менше значення відстані або вищий коефіцієнт, тим більш релевантним вважається результат. Це особливо корисно при роботі з не строго структурованими даними, покращуючи зручність користування та забезпечуючи інтуїтивну взаємодію користувача з системою.

Метою дослідження. Метою даної роботи є дослідження ключових методів нечіткого пошуку Fuzzy Search, що використовуються для фільтрації списків, проведення порівняльного аналізу їхніх характеристик, визначення переваги та недоліків методів нечіткого пошуку у системах, які працюють з великими масивами текстової інформації.

Основна частина. Методи нечіткого пошуку базуються на різних алгоритмах, які дозволяють оцінити ступінь подібності між текстовими рядками. Вони широко застосовуються для обробки неточних запитів, фільтрації списків і підвищення зручності пошуку[[1]]. До таких алгоритмів відносяться:

- Відстань Левенштейна (Levenstein Distance). Визначає мінімальну кількість змін – вставок, видалень або замін символів – необхідних для перетворення одного рядка в інший [2] Наприклад, щоб змінити слово kitten на sitting, потрібно виконати три операції. Чим менше значення цієї відстані, тим більш схожими вважаються рядки. Алгоритм реалізовано в таких бібліотеках, як FuzzyWuzzy, RapidFuzz та Fuse.js.

- Метод Jaro. Оцінює схожість рядків, враховуючи кількість спільних символів і їх порядок. Модифікація методу, Jaro-Winkler, надає додаткову вагу збігам на початку рядка, що особливо корисно для пошуку імен та назв. Наприклад, слова Martha і Marhta мають високу оцінку подібності, незважаючи на перестановку символів[1].

- Метод n-грам (n-grams). Цей підхід передбачає розбиття рядка на підрядки з n символів. Наприклад, для n=2 слово text перетворюється на набір: {“te”, “ex”, “xt”}. Далі обчислюється частка спільних фрагментів між рядками. Метод простий у реалізації та добре підходить для швидкої фільтрації великих обсягів даних у реальному часі[3].

- Soundex. Фонетичний алгоритм, який кодує слова за їх звучанням. Завдяки цьому можна знаходити слова, що звучать подібно, але мають різне написання, наприклад Smith і Smyth. Метод активно використовується в базах даних для пошуку імен, прізвищ та географічних назв.

Сучасні системи застосовують також векторизацію тексту. Вони включають наступні підходи: TF-IDF, Word2Vec, BERT. Відбувається порівняння не лише формально, а й семантичної схожості між рядками. Ці методи дозволять отримати контекст і значення слів, що робить їх незамінними в інформаційних пошукових системах, чат-ботах та інструментах обробки природної мови[4].

У рамках експериментального дослідження було реалізовано двійковий алгоритм пошуку для вирішення задачі нечіткого пошуку в текстових даних. Для програмної реалізації використовувалась бібліотека Fuzzy JS, яка надає інструменти для обробки нечітких запитів і порівняння текстових рядків на основі різних метрик схожості.

У межах експерименту було сформовано вхідний масив даних обсягом приблизно 91 106 записів, отриманих із датасету movies_metadata.csv база TMDB[**Ошибка! Источник ссылки не найден.**]. Кожен запис містив:

- назву фільму;
- короткий опис;
- список жанрів.

Пошук здійснювався за допомогою двійкового алгоритму після попереднього сортування даних. Для перевірки стійкості алгоритму до орфографічних помилок були сформовані тестові запити з навмисними спотвореннями назв жанрів наприклад, acton, romnce, comdy, war, crime.

Для аналізу ефективності було побудовано матрицю невідповідностей, що представлена у таблиці 1. Вона ілюструє розбіжності правильність роботи алгоритму, де істинно позитивні – алгоритм знайшов правильно знайшов жанр, хибно позитивні – алгоритм визначив інший жанр як очікуваний при пошуку, хибно хибні – алгоритм не виявив жанр як той що необхідно знайти, істинно хибні – алгоритм вірно відфільтрував нерелевантні записи.

Таблиця 1

Табличне представлене матриця невідповідностей для всіх проведених експериментів

Тип стану	Істинно позитивні (True positive)	Хибно позитивні (False Positive)	Хибно хибні (False Negative)	Істинно хибні (True Negative)
Кількість випадків	1412	14559	30731	408828

Також сформовано таблицю оцінки, що містить середній час пошуку, відсоток коректних збігів і рівень похибки. Це таблиця 2 оцінок експерименту з нечітким пошуком фільмів за жанрами, яка містить ключові метрики Precision, Recall, F1-score, час виконання запитів та приклади результатів. Вона дозволяє порівняти ефективність пошуку для кожного жанру та оцінити загальну якість алгоритму при наявності орфографічних помилок.

Таблиця 2

Оцінки якості та швидкості пошуку

Запит	Очікуваний жанр	Precision	Recall	F1-score	Час пошуку
acton	action	0.098	0.033	0.049	1.255s
romnce	romance	0.126	0.024	0.040	1.109s
wa	war	0.034	0.244	0.060	0.599s
comdy	comedy	0.333	0.036	0.065	0.901s
crim	crime	0.154	0.056	0.082	0.787s

Середня точність становить 14.9%, що свідчить про необхідність налаштування порогових значень – знизити нижній поріг схожості, бо відносна похибка в коротких запитах більша. Середній час обробки одного запиту – близько 1 секунди, що свідчить про прийнятну ефективність методу для великих текстових баз даних.

Висновки. Методи нечіткого пошуку Fuzzy Search є важливим інструментом сучасних інформаційних систем, що забезпечує гнучкість і адаптивність пошукових механізмів. Серед переваг є можливість обробки неповних або фрагментарних запитів, а також легка інтеграція. Але ці алгоритми потребують індивідуального налаштування порогових значень схожості, особливо для коротких запитів, що приводять до більш неоднозначних відповідей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Zhang, K., Shao, T., Sun, Y., Xu, X., Zhang, X., Zhou, X., Ding, K., & Huang, S. (2025). Interpretable research of fuzzy methods: A literature survey. *Information Fusion, 103524.* https://doi.org/10.1016/j.inffus.2025.103524.

[2] Berger, B., Waterman, M. S., & Yu, Y. W. (2020). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Transactions on Information Theory, 1.* https://doi.org/10.1109/tit.2020.2996543

[3] Martinez-Hinarejos, C., Benedi, J., & Tamarit, V. (2014). Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1*. https://doi.org/10.1109/taslp.2014.2377595

[4] Wang, J., & Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information, 11(9), 421.* https://doi.org/10.3390/info11090421

[5] Kaggle. *GroupLens Research. The Movies Dataset.* Retrieved October 16, 2025, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Boichenuik A.V., Molodets B.V.

FUZZY SEARCH METHODS FOR FILTERING LISTS

Abstract. *The paper considers the special aspects of the application of Fuzzy Search methods for filtering lists in modern information systems. The main fuzzy search algorithms, their advantages, disadvantages and areas of application are determined. The effectiveness of Fuzzy Search in processing text queries, searching for data in databases and user interfaces is analyzed.*

Keywords: *list filtering; fuzzy search; Levenshtein; Jaro-Winkler; n-grams; Soundex.*