Content-Based Book Search Using TF-IDF and Cosine Similarity

Hoàng Quốc Nam
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student625@ictu.edu.vn

***Abstract*—This undergraduate thesis investigates book search using TF-IDF and cosine similarity. Readers need a way to retrieve relevant books from catalogue text and compare alternative results. The student collects and preprocesses a book dataset, vectorizes titles with TF-IDF, ranks query matches by cosine similarity, and adds a score based on rating and rating count in a demonstration search function. The manuscript shows the search code and examples, reporting that 6810 book records were processed in its setup. It presents ranked results and a demonstration interface. The thesis notes that the interface and more advanced models remain unfinished. It does not report a standardized retrieval benchmark, user study, or verified improvement over a baseline. The contribution is a documented student study or prototype whose stated scope can guide further testing.**

***Index Terms*—book search, TF-IDF, cosine similarity, information retrieval**

I. Introduction

Readers need a way to retrieve relevant books from catalogue text and compare alternative results. This thesis investigates book search using TF-IDF and cosine similarity. It begins from a practical engineering need and explains the technologies and workflow chosen for the case. The objective is to document a student implementation or comparative investigation, together with the evidence available in the manuscript. Broader operational or economic effects require separate measurements and are not assumed from the existence of a prototype [1]. Earlier ICTU student work on related designs is included here for context [2], [3], [4], [5]. Related published studies provide context for the subject and methods [6], [7], [8], [9], [10]. These contextual works were selected for this English summary; they are not asserted to appear in the original thesis bibliography.

II. Methods

The student collects and preprocesses a book dataset, vectorizes titles with TF-IDF, ranks query matches by cosine similarity, and adds a score based on rating and rating count in a demonstration search function. The source is summarized through its stated design, implementation material, examples, and conclusion. Features shown in the manuscript are distinguished from proposals for later development. This review does not add a new experiment, fill missing measurements with estimates, or treat a planned capability as an observed result. The source thesis remains the provenance for the reported work [1].

Supervisor: Đoàn Thị Bích Ngọc, Thai Nguyen University of Information and Communication Technology (dtbngoc@ictu.edu.vn).

III. Results

The manuscript shows the search code and examples, reporting that 6810 book records were processed in its setup. It presents ranked results and a demonstration interface. The report's screens, configurations, examples, or measurements constitute the evidence for this summary. A capability described only as a design intention is not counted as a validated operational outcome. No additional performance, adoption, financial, or safety figure has been introduced here. The result is interpreted within the scope of the student's documented setup [1].

IV. Discussion

The thesis notes that the interface and more advanced models remain unfinished. It does not report a standardized retrieval benchmark, user study, or verified improvement over a baseline. Further work could document the dataset and evaluation queries, measure precision and response time, and test the interface with readers. These limitations affect how the evidence should be read. The current manuscript can support understanding of the design and reported demonstration, while independent tests would be needed for stronger claims. Any later deployment should document its environment and separate observed outcomes from intended benefits [1].

V. Conclusion

The thesis documents book search using TF-IDF and cosine similarity. The manuscript shows the search code and examples, reporting that 6810 book records were processed in its setup. It presents ranked results and a demonstration interface. The thesis notes that the interface and more advanced models remain unfinished. It does not report a standardized retrieval benchmark, user study, or verified improvement over a baseline. Further work could document the dataset and evaluation queries, measure precision and response time, and test the interface with readers. Accordingly, the present contribution is a source-grounded account of the student work and its tested or illustrated scope. Claims about broader effectiveness should be assessed in later work using explicit evaluation conditions [1].

References

[1] Hoàng Quốc Nam, ”Content-Based Book Search Using TF-IDF and Cosine Similarity,” internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2024.

[2] Lưu Xuân Chiến, ”Construction of a Bookstore Website for Minh Sang Bookstore,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2023.

[3] Nguyễn Minh Toàn, ”Construction of an Online Bookstore Website for HEBooks,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2023.

[4] Nguyễn Văn Điệp, ”Construction of an Online Bookstore Website for Thai Nguyen Book and Equipment Company,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2009.

[5] Lê Minh Thuận, ”Construction of an Online Bookstore Website Using ASP.NET Technology,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2008.

[6] Y Nurdiansyah et al., ”New book classification based on Dewey Decimal Classification (DDC) law using tf-idf and cosine similarity method,” Journal of Physics: Conference Series, 2019. doi:10.1088/1742-6596/1211/1/012044.

[7] Mizanul Ridho Aohana et al., ”Tourism Destination Article Search Features using TF-IDF and Cosine similarity,” DIELEKTRIKA, 2023. doi:10.29303/dielektrika.v10i2.338.

[8] Gottapu Mohana Rao et al., ”A Novel Approach to Identifying the Cosine Similarity using TF-IDF,” IJARCCE, 2017. doi:10.17148/ijarcce.2017.6640.

[9] Rashika S, ”Personalized Book Recommendation System using TF-IDF and KNN Hybrid,” International Journal for Research in Applied Science and Engineering Technology, 2022. doi:10.22214/ijraset.2022.45736.

[10] S. Tata and J. M. Patel, ”Estimating the Selectivity of tf-idf Based Cosine Similarity Predicates,” ACM SIGMOD Record, vol. 36, no. 2, pp. 7-12, 2007. DOI: 10.1145/1328854.1328855.