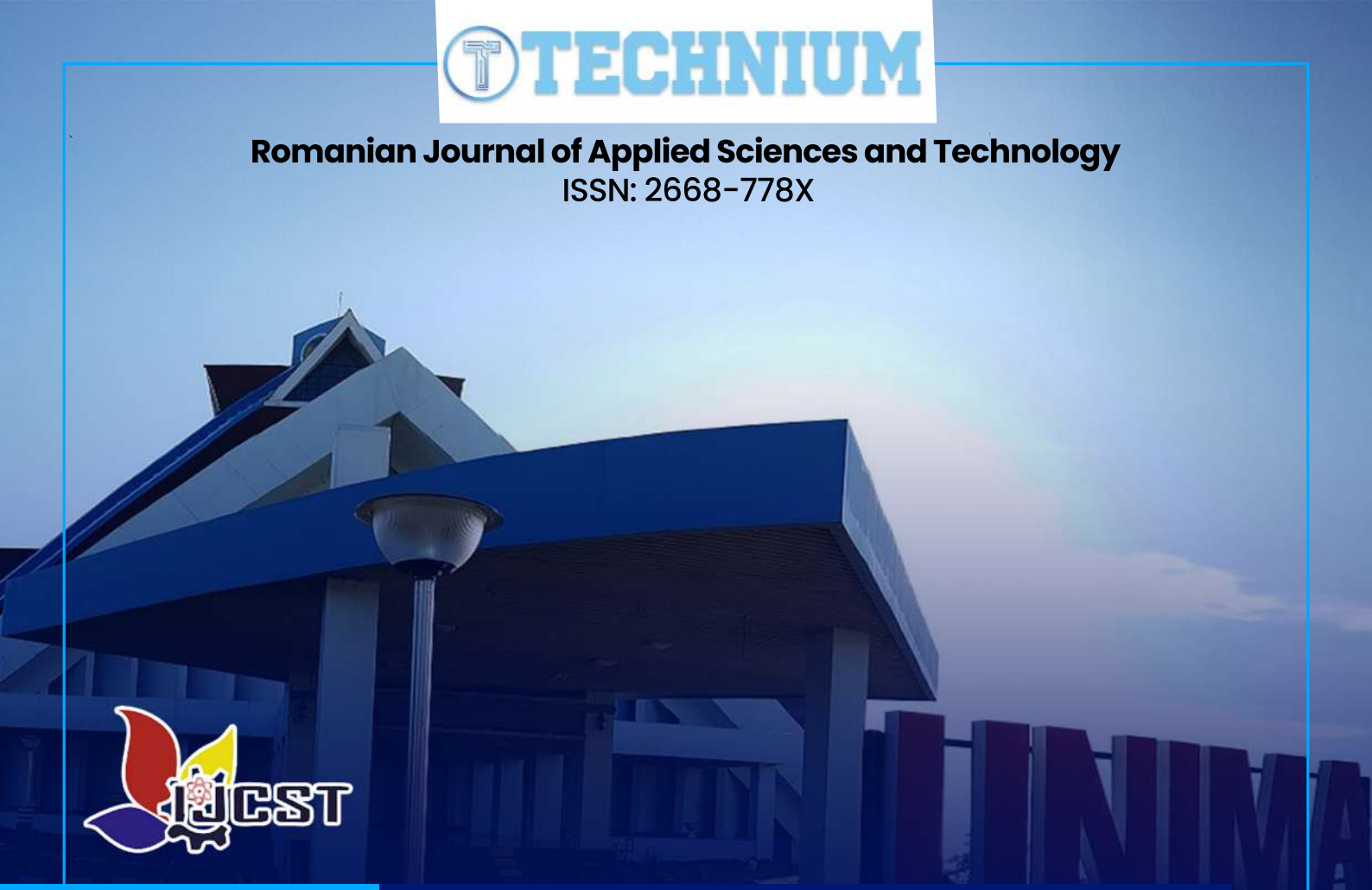
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 16 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 234-236, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Design and Implementation of Tourism News Information Retrieval System using Modified Cosine Similarity

Ika Oktavia Suzanti [1*], *Husni*[1], *Biru* Sultan Awaldhia [1], and *Mohammad* Syarief [1]

[1] Informatics Engineering Department University of Trunojoyo Madura, Bangkalan, Indonesia

Abstract. Technological developments cannot be denied a very meaningful impact on human life so that what was previously done traditionally is now completely digital, for example conventional news media which has been transformed into an online news portal so that it can still reach its readers. Online news portals provide a lot of up-to-date information, including tourism news, which is an industry that continues to grow and has the opportunity to create new jobs for the community. Currently, to search for tourism news, people only need to type a keyword (query) in the search engine which will then display the latest news about tourism. However, not all tourism news displayed matches what they are looking for, so readers have to re-check which takes a lot of time. Therefore, a tourism news retrieval system that can display the most relevant tourism news to the query is proposed. The Modified Cosine method shows good results in document clustering to bring the inter-cluster distance closer. This study uses the Modified Cosine method and TF-IDF weighting schema to determine the value of precision, recall, and f-measure in calculating the similarity of the query to tourism news documents. The system has been tested using 3 types of queries, with 5 different words each. The test results show that Modified Cosine method obtained best precision value in the test using 5 words in the query and the F-Measure value and the best recall on the 3 words test in query.

* Corresponding author: iosuzanti@trunojoyo.ac.id

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 234-236, October, 2023
ISSN: 2668-778X
www.techniumscience.com

1 Introduction

Tourism in Indonesia is a growing industry fast so that potential increase economy area and create field work [1]-[5]. Destination from tourist moment this no just for entertaining and relieving boredom but also for follow a current activity popular or often visited many people just for add collection photos on social media [6]. Machine search or search engines in this modern era often become the place first visited when look for information [7]-[12]. Search engines work with count similarity of query to document news existing tourism in the database and display the most relevant news [13]. SEBI (Search Engine Bahasa Indonesia) is one of the search engines capable of gather page web from the internet, preprocessing, building index, handle queries with perfect as well as categorize document [14].

There is a number of methods that can used in count similarity between queries and data in databases such as Cosine, Dice, Vector Support Machine (VSM), and others. In research conducted by Sahu [15] show that Cosine has less than optimal results in zoom out distance intra-cluster for group document so that researcher the using Modified Cosine for zoom out distance intra-cluster. Many studies have conducted for create a search engine as well look for similarity among two documents, as done by Wahyuni [16] for classify document thesis with apply Cosine Similarity algorithm and TF-IDF weighting show appropriateness system with percentage testing 88.3% when tested by expert archive. Study other done by Sahu for grouping document use improved K-means algorithm using Modified Cosine Distance Measure using Mahout with Hadoop shows that with Modified Cosine intra-cluster results obtained better compared Cosine method is 93% [15].

Putung, etc. [17] did a study for apply the cosine method in system meet return information on group document script and show the recall result is worth 1 for each test showing all relevant documents could found by the system [17]. Therefore, it is necessary to design and build an information retrieval system for tourism news to make it easier for users to find and obtain news about tourism that is most relevant to their information needs. The method proposed in this study is the Modified Cosine to calculate the similarity between the query and the collection of tourism news documents and the TF-IDF weighting scheme to normalize the importance of the term.

2 Methodology

Proposed architecture is shown in Figure 1. We have two type of users who can enter the query that will searching for and operator or automatic web crawler that collect many tourism news form Internet. News entered by operator will through Preprocessing, weighting and indexing. The last step is saving each term in an inverted index database. The entered query is also processed through Preprocessing stage (Case Folding, Tokenizing, Stopword removal/Filtering, Stemming). The query is weighted using TF-IDF. The System will takes each news from inverted index database and calculate the similarity to the query using Cosine method as well as Modified Cosine method. After calculation done the system will showing ranked tourism news which is most relevant to user query.

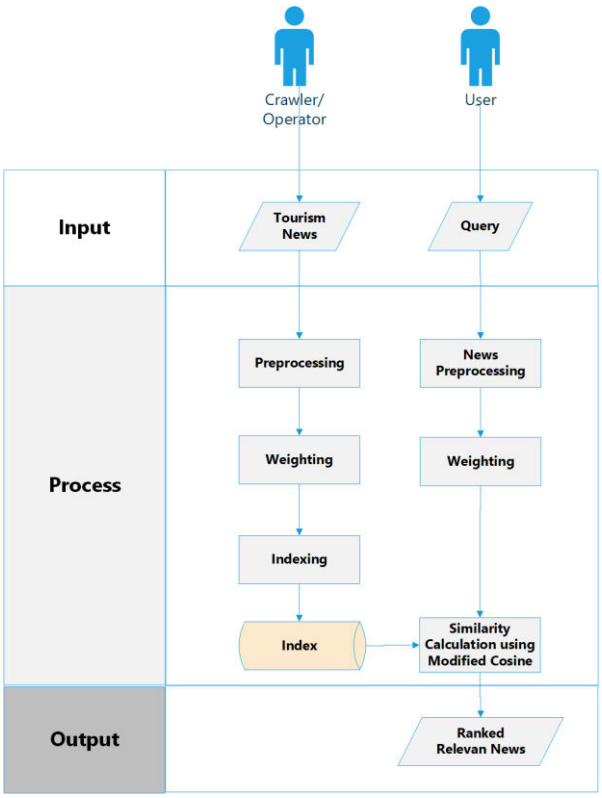


Fig. 1. Proposed architecture

3 Results and discussion

Tests carried out following this aim for knowing precision, recall, and f- measure values between trials search. Query used in this test as many as 15 sample queries. Each of the 5 queries consisting of 1 word, 3 words, and 5 words. Figures 2, 3 and 4 show chart testing that has been conducted with using 1 word, 3 words, and 5 words with 5 different words in each query.

Figures 2, 3 and 4 show results from every experiments that have been carried out and the results obtained show if best precision value found in test using 5 words in the query at 100% and the F-Measure value and the best recall in testing 3 words in the query with F-Measure value is 83.48% and recall is 84.84%.

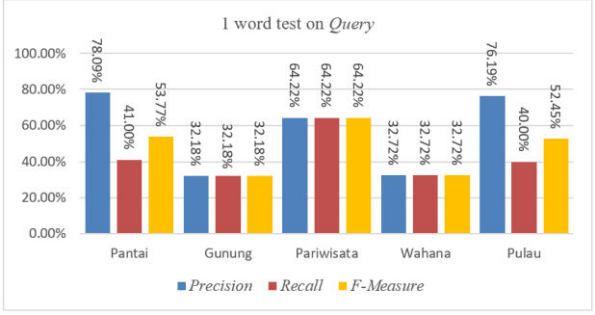


Fig. 2. Graphics 1 Word on Query

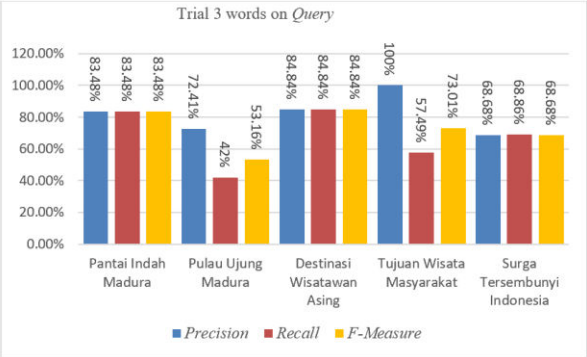


Fig. 3. Graphics 3 Words on Query

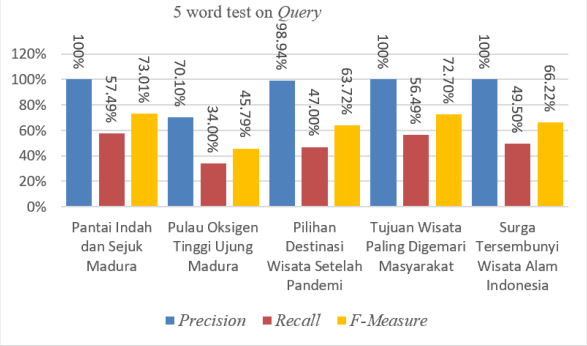


Fig. 4. Graphics 5 Words on Query

4 Conclusion

This tourism news information retrieval shows that Modified Cosine method has good result in similarity calculation between queries and tourism news where the best precision value found in test using 5 words in the query and the best recall on the 3 words in the query.

References

1. DY Reindrawati, NE Suriani, and S. Asmorowati, Local Community Participation in Development. 2019.
2. AK Jaelani, IGAKR Handayani, and L. Karjoko. "Development of tourism based on geographic indication towards to welfare state." International Journal of Advanced Science and Technology 29, no. 3s (2020): 1227-1234.
3. M Wood. Ecotourism: Principles, practices and policies for sustainability. UNEp, 2002.
4. A Lasso, and H Dahles. "Are tourism livelihoods sustainable? Tourism development and economic transformation on Komodo Island, Indonesia." Asia Pacific Journal of Tourism Research 23, no. 5 (2018): 473-485.
5. MP Hampton. "Enclaves and ethnic ties: The local impacts of Singaporean cross-border tourism in Malaysia and Indonesia." Singapore Journal of Tropical Geography 31, no. 2 (2010): 239-253.
6. T. Utomo, "Madura Tourism Based On Community Participation," J. Psychol., vol. 10, no. 1, pp. 72–84, 2019, doi: 0803973233.
7. A Halavais. Search engine society. John Wiley & Sons, 2017.
8. M Moran, and B Hunt. Search engine marketing, Inc.: Driving search traffic to your company's website. IBM Press, 2014.
9. Z Xiang and U Gretzel. "Role of social media in online travel information search." Tourism management 31, no. 2 (2010): 179-188.
10. BJ Jansen, and CM Eastman. "Limitations of advanced searching techniques on web search engines." Journal of Electronic Resources in Law Libraries 1, no. 1 (2006): 55-81.
11. J Zobel, and A Moffat. "Inverted files for text search engines." ACM computing surveys (CSUR) 38, no. 2 (2006): 6-es.
12. E Hargittai. "Second-level digital divide: Mapping differences in people's online skills." arXiv preprint cs/0109068 (2001).
13. D. Lewandowski, "Credibility in web search engines," Online Credibility. digits. Ethos Eval. Comput. comm., pp. 131–146, 2012, doi:10.4018/978-1-4666-2663-8.ch008.
14. Husni, IO Suzanti, YD Pramudita, SS Putro, and L. Heryawan, "Web Service for Indonesian Search Engines (SEBI)," J. Phys. conf. Ser., vol. 1569, no. 2, 2020, doi:10.1088/1742-6596/1569/2/022087.
15. L. Sahu and BR Mohan, "An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop," 9th Int. conf. eng. inf. syst. ICIIS 2014, 2015, doi:10.1109/ICIINFS.2014.7036661.
16. RT Wahyuni, D. Prastiyanto, and E. Supraptono, "Application of Cosine Similarity Algorithm and TF-IDF Weighting in Thesis Document Classification System," J. Tek. Electro, vol. 9, no. 1, pp. 18–23, 2017, doi:10.15294/jte.v9i1.10955.
17. KD Putung, ASM Lumenta, and A. Jacobus, "Implementation of Information Retrieval System in Thesis Document Collection," J. Tek. information. , vol. 8, no. 1, 2016, doi:10.35793/jti.8.1.2016.12227.