DiversiNews: Enriching News Consumption with Relevant Yet Diverse News Articles Retrieval

Yiqun Sun
National University of Singapore
sunyq@comp.nus.edu.sg

Qiang Huang*
National University of Singapore
huangq@comp.nus.edu.sg

Yanhao Wang
East China Normal University
yhwang@dase.ecnu.edu.cn

Anthony K. H. Tung
National University of Singapore
atung@comp.nus.edu.sg

ABSTRACT

In the digital age, where echo chambers on social media and news platforms increasingly shape public opinion, there is a growing need for tools that present news consumers with a broad spectrum of perspectives. To this end, we introduce DiversiNews, a novel system designed to diversify news consumption by providing readers with articles that are not only relevant to their interests but also offer a variety of viewpoints. DiversiNews leverages state-of-the-art semantic text encoding techniques and implements advanced Diversity-aware k-Maximum Inner Product Search (DkMIPS) algorithms. Our demonstration highlights the potential of DiversiNews to broaden users' exposure to different viewpoints, thereby countering the polarizing effect of digital echo chambers. We showcase how DiversiNews can enrich the news reading experience, supporting the development of a more informed and balanced public discourse in digital news consumption applications.

PVLDB Reference Format:
Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony K. H. Tung. DiversiNews: Enriching News Consumption with Relevant Yet Diverse News Articles Retrieval. PVLDB, 17(12): 4277 - 4280, 2024.
doi:10.14778/3685800.3685854

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/dukesun99/DiversiNews/.

1 INTRODUCTION

In the social media era, digital news consumption has become predominant in daily news habits, with more people favoring social media over traditional news sources. Nevertheless, this shift has fostered the growth of *echo chambers* on these platforms: Users often join like-minded communities, sharing news that reflects their political beliefs, which narrows their exposure to diverse perspectives [5]. The echo chamber effect can create the illusion of consensus and promote misinformation and extremism.

Such problems are exacerbated by recommender systems, which reinforce users' existing biases to optimize engagement metrics like attention and click-through rates [6, 7]. Figure 1 shows how an echo chamber forms and is reinforced through the combined influence of Recommender systems, social media platforms, and user interactions. Initially, users encounter diverse opinions, but over time, implicit feedback mechanisms such as click history and viewing duration cause the systems to align with user preferences, reducing the diversity of content recommendations.

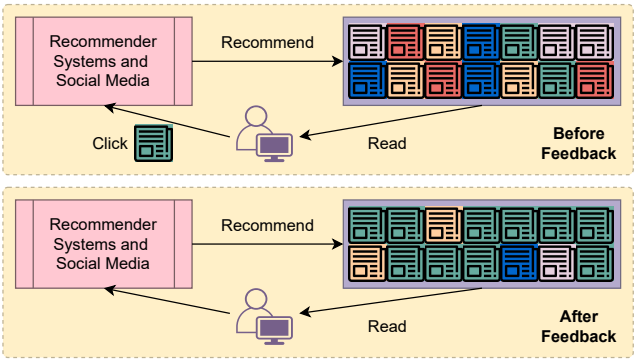


Figure 1: Illustration of how Recommender systems and social media can amplify the echo chamber effect.

To counteract the echo chamber effect, exposing users to diverse perspectives and opinions is crucial not only for individual enlightenment and protection against mis-, dis-, and mal-information but also for the promotion of meaningful, healthy societal discussions [1–3]. Platforms like AllSides[1] offer a model to present news from various political stances (*left, centrism, right*) and a system that works well for US politics, which are predominantly divided between the Democratic Party (*left and liberal*) and the Republican Party (*right and conservative*). However, this binary classification may not adequately represent the nuances of political discourse in other countries or non-political discussions. For example, Singapore's political opinions do not neatly align with a left-to-right, liberalism-to-conservatism scale. Given the complex nature of opinions and perspectives worldwide, relying solely on a left-to-right political spectrum fails to capture the subtle nuances in news perspectives across different discussions [12].

To enhance recommendation diversity, we recently investigated the problem of Diversity-aware k-Maximum Inner-Product Search (DkMIPS) [10], aiming to retrieve articles that are not only relevant to user interest but also exhibit a diverse range of perspectives.

[1] https://www.allsides.com/unbiased-balanced-news

*Corresponding author.

Proceedings of the VLDB Endowment, Vol. 17, No. 12 ISSN 2150-8097.
doi:10.14778/3685800.3685854

We introduce two novel algorithms for DkMIPS processing: (1) Greedy operates in k rounds and, in each round, adds an item that maximally increases the objective function to the result set; (2) DualGreedy operates in up to $2k$ rounds, maintains two sets of results greedily in turn, and returns the better one between them as the final result set. Furthermore, we integrate BC-Tree [9] into both algorithms to accelerate DkMIPS processing.

In this demonstration, we introduce DiversiNews, a novel news enrichment system developed around the DkMIPS problem to counteract the echo chamber effect. DiversiNews employs state-of-the-art pre-trained semantic text encoders [4, 8, 13, 14] that capture not only textual relevance but also discern latent political perspectives, which are manifested through features like writing style and word choice. By leveraging these encoders, DiversiNews converts news articles into embeddings (dense vectors) in an inner product space, where the inner product between any two article embeddings quantitatively measures their relevancy. When a user engages with an article, DiversiNews generates a corresponding query embedding using these encoders. Then, it employs DkMIPS approaches to recommend articles from the news corpus. The tunable balance between relevancy and diversity ensures that the recommended articles are pertinent and present a wide range of viewpoints. This feature of DiversiNews allows it to be seamlessly integrated as an enhancement tool into existing social media and news platforms, significantly elevating the news reading experience.

2 PROBLEM FORMULATION

Let $\mathcal{P}$ represent the corpus of encoded news articles, composed of n vectors in a d-dimensional inner product space $\mathbb{R}^d$. The vector $\boldsymbol{q}$ denotes the embedding of the article a user is currently reading. Our objective is to identify relevant yet diverse articles from $\mathcal{P}$, enhancing the user's reading experience. The inner product $\langle \boldsymbol{p}, \boldsymbol{q} \rangle = \sum_{i=1}^{d} p_i q_i$ between two vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d$ quantifies the relevancy of an article $\boldsymbol{p} \in \mathcal{P}$ to $\boldsymbol{q}$. The challenge of news enrichment lies in balancing the relevancy with the need for diversity in the articles presented. This task can be encapsulated in the DkMIPS problem [10]. DkMIPS goes beyond simply finding the top-k articles with the largest inner products with $\boldsymbol{q}$ [11]; it also integrates diversity criteria to ensure a wide range of perspectives. Formally,

Definition 2.1 (DkMIPS [10]). Given a set of n news articles $\mathcal{P}$ with each represented as a vector $\boldsymbol{p} \in \mathbb{R}^d$, a query vector $\boldsymbol{q} \in \mathbb{R}^d$, an integer $k > 1$, a balancing factor $\lambda \in [0, 1]$, and a scaling factor $\mu > 0$, find a set $\mathcal{S}^*$ of k article vectors such that

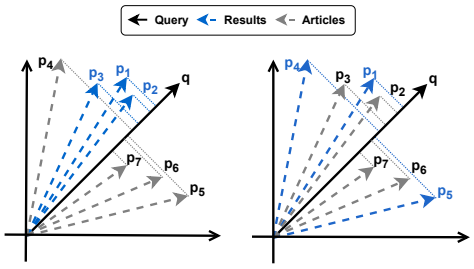
$$\mathcal{S}^* = \underset{\mathcal{S} \subset \mathcal{P}, |\mathcal{S}|=k}{\arg\max} \; f(\mathcal{S}), \tag{1}$$

where the objective function $f(\mathcal{S})$ is defined by Eq. 2 below:

$$f(\mathcal{S}) := \frac{\lambda}{k} \sum_{\boldsymbol{p} \in \mathcal{S}} \langle \boldsymbol{p}, \boldsymbol{q} \rangle - \frac{2\mu(1-\lambda)}{k(k-1)} \sum_{\boldsymbol{p}, \boldsymbol{p}' \in \mathcal{S} \wedge \boldsymbol{p} \neq \boldsymbol{p}'} \langle \boldsymbol{p}, \boldsymbol{p}' \rangle. \tag{2}$$

Definition 2.2 (k$MIPS). Setting $\lambda = 1$ in Eq. 2 yields the objective function of the kMIPS problem, which focuses solely on relevancy.

Figure 2 shows the difference between kMIPS and DkMIPS. For a query vector $\boldsymbol{q}$ and $k = 3$, kMIPS selects $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3\}$ for their largest inner products with $\boldsymbol{q}$, prioritizing relevancy but potentially lacking diversity. Conversely, DkMIPS chooses $\{\boldsymbol{p}_1, \boldsymbol{p}_4, \boldsymbol{p}_5\}$, which not only rank high based on inner products but also offer a wider range of perspectives. DkMIPS provides a result set that balances relevance (as denoted by the first term of Eq. 2) and diversity (as denoted by the second term of Eq. 2), ensuring a comprehensive coverage of viewpoints related to $\boldsymbol{q}$.



(a) kMIPS results. (b) DkMIPS results.

Figure 2: Illustration of retrieval results of kMIPS vs. DkMIPS for $k = 3$. Here, $\boldsymbol{p}_1$ has the largest inner product with $\boldsymbol{q}$, followed by $\boldsymbol{p}_2$, down to $\boldsymbol{p}_7$ with the smallest value.

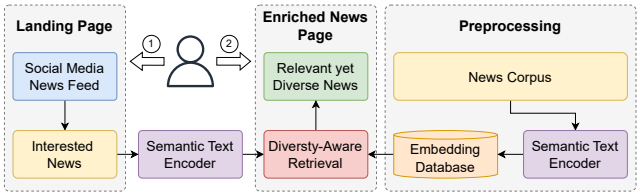


Figure 3: Overview of the DiversiNews system.

3 SYSTEM ARCHITECTURE

As described in Figure 3, the DiversiNews system comprises three core components as follows:

- **Preprocessing:** We construct the NewsSpectrum dataset[2] for news corpus, employ semantic text encoders to generate article embeddings, and store them in a vector database.
- **Landing Page:** Upon a user's arrival, (1) the journey begins by exploring a news article of interest in Social Media News Feed.
- **Enriched News Page:** After selecting a news article, (2) the user proceeds to the enriched news page. Our Diversity-Aware Retrieval mechanism (DkMIPS) uses it as a query to retrieve relevant yet diverse articles, broadening the user's experience.

Subsequently, we will explore each component in more detail.

3.1 Preprocessing

News Corpus. To effectively simulate and assess news reading enrichment scenarios, we developed a comprehensive news corpus, NewsSpectrum. Sourced from Pushshift's Reddit dump up to July 2022, it includes submissions with at least ten upvotes, from which URLs were extracted. We used the AllSides media list for ground-truth bias ratings, categorizing media outlets into five levels, i.e., *Left, Lean Left, Centrism, Lean Right, and Right*, quantified into a five-point rating $\{-2, -1, 0, 1, 2\}$. Each media outlet is associated with its bias rating and top-level domain information. To ensure balanced representation, we randomly selected 50,000 articles from each category, resulting in a diverse set of 250,000 articles.

[2]https://github.com/dukesun99/DiversiNews/tree/main/NewsSpectrum

Although NewsSpectrum mainly comprises new articles on US politics due to accessibility, the utility of our system extends beyond this realm. It is adaptable to various contexts, as political bias labels are not used for semantic encoder training or during retrieval.

Semantic Text Encoding. In DiversiNews, each text (article) is encoded into a vector representation, allowing the assessment of relevancy between texts through inner products. We utilize three advanced Transformer-based pre-trained text encoders without fine-tuning. Two of these, Sentence-BERT (all-MiniLM-L12-v2) [14] and AnglE (UAE-Large-V1) [13], are tailored for the Semantic Text Similarity (STS) task. Moreover, we incorporate LLAMA 2 (7B) [15], a general-purpose, decoder-only large language model, taking the last hidden state of the last token for article embeddings. Finally, we store all embeddings in the article embedding database.

3.2 Landing Page

The user interface simulates a typical online Social Media News Feed browsing experience. Users can scroll through and explore news articles in the feed on the landing page, selecting those of interest based on the headlines. Upon selecting an article, users are directed to the news enrichment page to read more about the article. Here, the selected article is encoded using the same semantic text encoder as in prepossessing and fed into the Diversity-Aware Retrieval component on the enriched news page.

3.3 Enriched News Page

Upon generating the query embedding for a selected article via the Semantic Text Encoder, DiversiNews activates its Diversity-Aware Retrieval component. This component employs DkMIPS methods, specifically BC-Greedy and BC-DualGreedy, which integrate the BC-Tree index [9] into Greedy and DualGreedy for efficiency, to curate a list of k relevant and diverse news articles [10]. Although they employ different greedy strategies, both aim to balance relevancy and diversity in the retrieved results. The parameter λ allows users to fine-tune this balance.

4 DEMONSTRATION SCENARIOS

This demonstration showcases the capability of DkMIPS to retrieve a range of politically diverse articles, enriching news reading experiences through quantitative and qualitative assessments. We quantitatively measure the average pairwise political bias rating in the result lists, demonstrating the superiority of DkMIPS over kMIPS in terms of diversity. Additionally, we provide concrete examples for qualitative evaluations.

4.1 Quantitative Results

Evaluation Measures. We first introduce two measures designed to assess relevancy and diversity, respectively.

- **Relevancy:** Given a result set $S = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k\}$ with k news articles and the query embedding $\boldsymbol{q}$, the relevancy is defined as the average inner product between each $\boldsymbol{p}_i \in S$ and $\boldsymbol{q}$, that is,

$$\text{Relevancy}(S, \boldsymbol{q}) = \frac{1}{k} \sum_{i=1}^{k} \langle \boldsymbol{p}_i, \boldsymbol{q} \rangle,$$

  where the inner product $\langle \boldsymbol{p}_i, \boldsymbol{q} \rangle$ measures the similarity between $\boldsymbol{p}_i$ and $\boldsymbol{q}$. A higher value means greater relevancy.

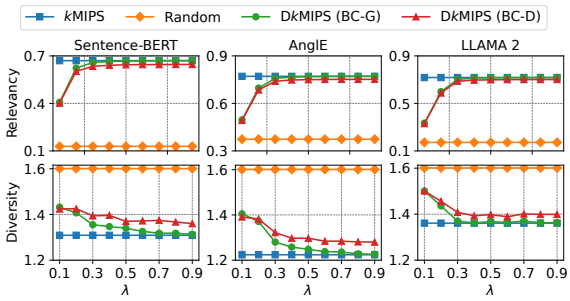


Figure 4: Quantitative results ($k = 10$).

- **Diversity:** As outlined in Section 3.1, each article $\boldsymbol{p}$ has a political bias rating $\delta(\boldsymbol{p}) \in \{-2, -1, 0, 1, 2\}$. Diversity is defined as the average pairwise media bias difference for articles in S, i.e.,

$$\text{Diversity}(S) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |\delta(\boldsymbol{p}_i) - \delta(\boldsymbol{p}_j)|.$$

  A higher value suggests greater diversity in the set S of articles in terms of political perspectives.

Benchmark Methods. In addition to BC-Greedy (BC-G) and BC-DualGreedy (BC-D) for DkMIPS, we also implement two baselines for comparative analysis, focusing on two key aspects:

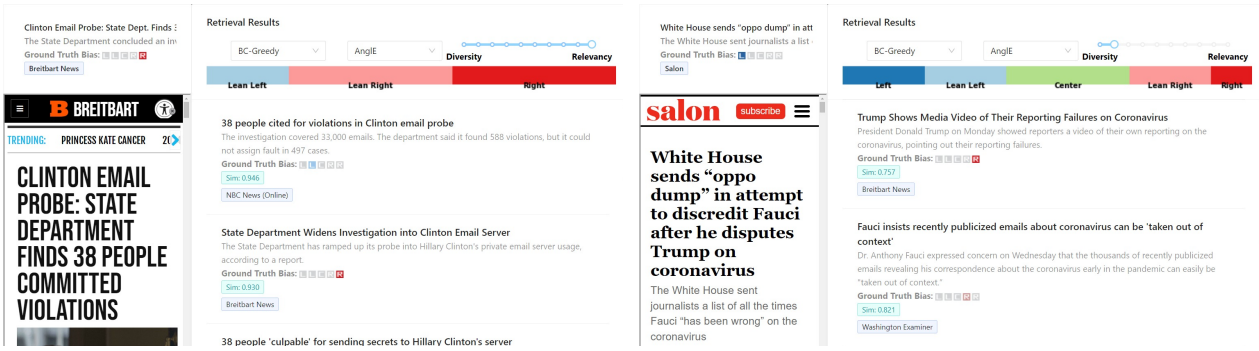
- **kMIPS:** The exact kMIPS method retrieves S solely based on their inner products to $\boldsymbol{q}$, establishing a benchmark for the *relevancy* measure as it represents the maximum of achievable relevancy.
- **Random:** The Random Oracle approach randomly selects a result set S from all articles, serving as a *diversity* benchmark since it does not consider the relevancy between S and $\boldsymbol{q}$, thus representing the maximum possible diversity.

Result Analysis. Figure 4 presents the relevancy and diversity results, comparing DkMIPS methods (BC-G and BC-D with varying λ values) against kMIPS and Random. The results show that DkMIPS methods boost diversity while maintaining similar relevancy levels, particularly at larger λ values. Adjusting λ values reveals a trade-off: as λ increases, relevancy grows while diversity diminishes. This demonstrates that DkMIPS methods can be customized to meet different user preferences for diversity and relevancy, enriching the reading experience. The following two scenarios offer a qualitative analysis of these outcomes, illustrated through case studies.

4.2 Scenario 1: Verifying News Authenticity

Figure 5 shows the split view of the enriched news page, with the selected article on the left and the DkMIPS-retrieved articles on the right. Users can interact with the interface through dropdown selectors for retrieval methods and encoding models and a draggable bar to adjust the balancing factor λ, allowing them to switch between retrieval algorithms and encoders and balance relevancy with diversity. Moreover, this page includes a media bias summary chart for the retrieved articles, offering users a sense of diversity in the retrieved articles. For each article in the result list, this page also displays the title, excerpt, AllSides media bias rating, similarity to the query article, and media outlet name.

In this scenario, we concentrate on verifying news authenticity. Referring to Figure 5(a), consider a user questioning the article "Clinton Email Probe: State Department Finds 38 People Committed



(a) Scenario 1: verifying news authenticity.

(b) Scenario 2: exploring diverse perspective.

Figure 5: Screenshot of the demonstration scenarios.

Violations" from Breitbart News, a right-wing media outlet. Using DiversiNews, users can increase the value of λ by dragging the selection bar to the right to discover highly relevant articles. The results, including one from left-leaning NBC News on the same event, indirectly corroborate the article's authenticity. This cross-validation from diverse media sources helps users assess the accuracy of the information, aiding in discerning its authenticity.

4.3 Scenario 2: Exploring Diverse Perspectives

Referring to Figure 5(b), this demonstration scenario illustrates how users, after reading an article titled "White House sends 'oppo dump' in attempt to discredit Fauci after he disputes Trump on coronavirus" from a left-leaning source such as Salon, can gain additional insights by being exposed to news articles from right-leaning sources offering differing viewpoints. This article criticizes the Trump administration's attempts to discredit Dr. Fauci as politically motivated and the prioritization of political image over scientific expertise and public health.

With DiversiNews, users can expand their views by adjusting the selection bar to access diverse viewpoints, including right-leaning media. The resulting list includes sources like the Washington Examiner, which offers a more neutral tone on Fauci's emails. This article balances criticism and defense, highlighting political responses and offering insights into Fauci's stance on navigating the complexities of a health crisis. DiversiNews enhances the reading experience by presenting articles from diverse political perspectives, allowing users to form a more extensive understanding of relevant events.

5 CONCLUSION

In this demonstration, we present DiversiNews, a system that significantly enriches the news reading experience by utilizing semantic text encoders and DkMIPS approaches to retrieve relevant yet diverse news articles. This system directly addresses the challenge of echo chambers in digital news consumption, highlighting the importance of diversity in news recommender systems. By offering the reader a broader spectrum of viewpoints, DiversiNews plays a pivotal role in promoting a more informed and balanced public discourse. This is especially crucial in an era where digital news is often restricted by algorithmic biases, leading to increased societal polarization and the proliferation of misinformation.

ACKNOWLEDGMENTS

This research is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001), the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, and the National Natural Science Foundation of China under grant No. 62202169.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore and the National Research Foundation, Singapore.

REFERENCES

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Soc. Netw. Anal. Min.* 13, 1, Article 30 (2023), 36 pages.

[2] Darrin Baines and Robert J R Elliott. 2020. *Defining misinformation, disinformation and malinformation: An urgent need for clarity during the COVID-19 infodemic.* Discussion Papers 20-06. Department of Economics, University of Birmingham.

[3] Elinor Carmi, Simeon J. Yates, Eleanor Lockley, and Alicja Pawluczuk. 2020. Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review* 9, 2 (2020), 22 pages.

[4] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-Tau Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In *NAACL*. 4207–4218.

[5] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. U.S.A.* 118, 9 (2021), e2023301118.

[6] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2022. The Effect of People Recommenders on Echo Chambers and Polarization. In *ICWSM*. 90–101.

[7] Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. 2022. Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways. In *WWW*. 2719–2728.

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*. 6894–6910.

[9] Qiang Huang and Anthony Tung. 2023. Lightweight-Yet-Efficient: Revitalizing Ball-Tree for Point-to-Hyperplane Nearest Neighbor Search. In *ICDE*. 436–449.

[10] Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony K. H. Tung. 2024. Diversity-Aware k-Maximum Inner Product Search Revisited. arXiv:2402.13858

[11] Qiang Huang, Yanhao Wang, and Anthony K. H. Tung. 2023. SAH: Shifting-Aware Asymmetric Hashing for Reverse k Maximum Inner Product Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 4312–4321.

[12] John Jost, Christopher Federico, and Jaime Napier. 2009. Political ideology: Its structure, functions, and elective affinities. *Annu. Rev. Psychol.* 60 (2009), 307–337.

[13] Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. arXiv:2309.12871

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*. 3982–3992.

[15] Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288