Laure Normand

THIS ARTICLE IS MADE FOR YOU

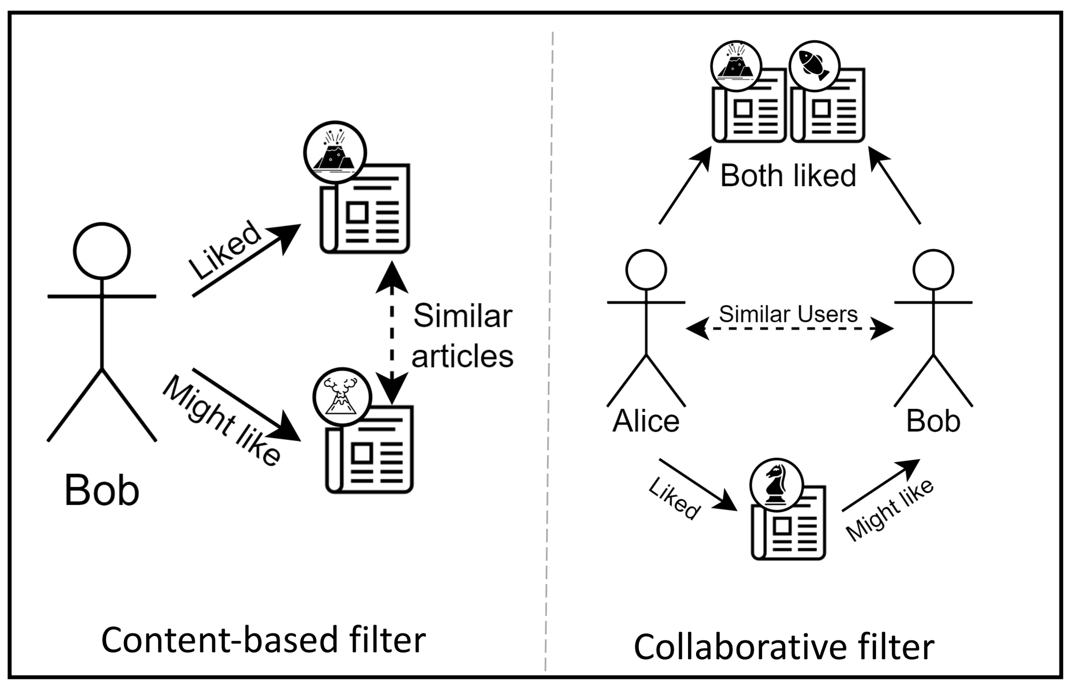
Or how to optimize newspaper article recommendations

Arthur Gallois — Nathan Chalumeau — Ali Ramlaoui — Dimitri Martin

News has become easily and massively accessible since the 1990s, thanks to the Internet. Collaborative recommendation filters, used by social media platform, rely on mass interactions [1]. These filters face challenges such as data volume, computational resources, and filter bubbles [2]. This paper develops new content-based recommendation filters using article writing style.

Recommendation Methods

Content-based: rely on topic similarities between articles
Collaborative: uses user preferences similarities to suggest new articles

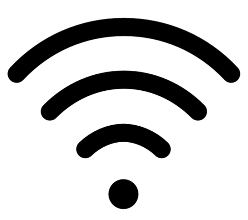


Collaborative vs Content-based filters

Other useable data:
Geographic, recently visited shops (google maps), internet search...

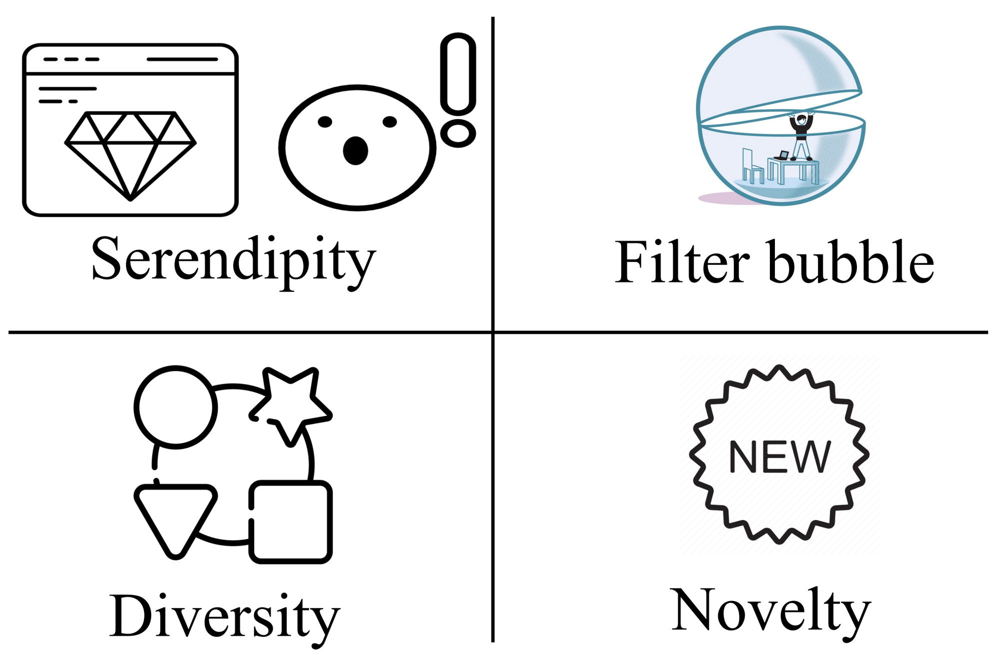
Evaluation Methods

Offline: fixed dataset that includes user feedback without real-time update



Online: real-time interaction with users to assess models.

Qualitative method tries and focuses on new criteria:

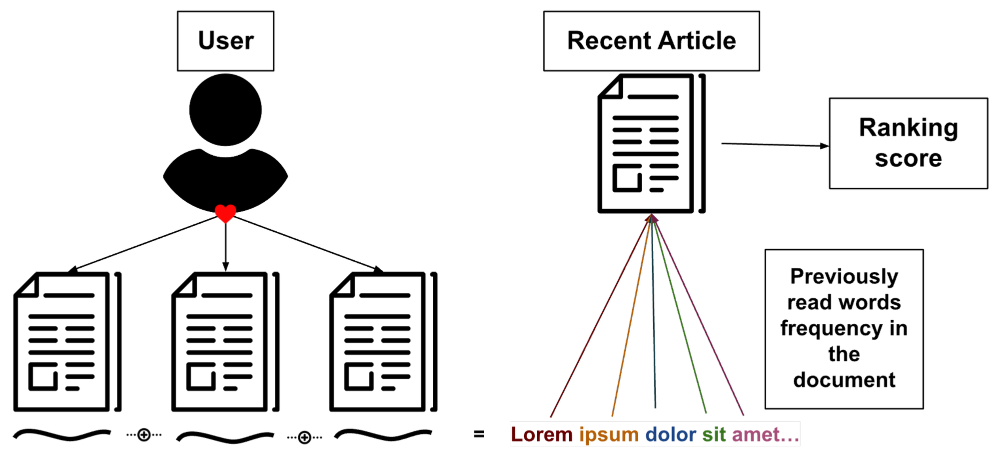


N.B. : **Serendipity** = New + Surprising

Filter bubble: Not a criterion but situation the algorithm should avoid

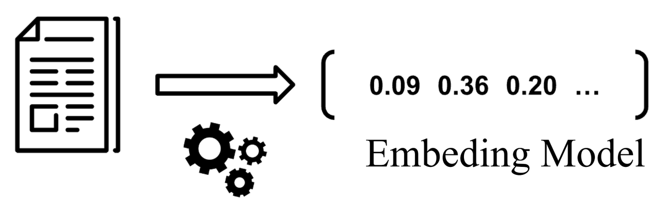
Content-Based Algorithms

BM25: Widely used document similarity scoring algorithm based on **word frequency**



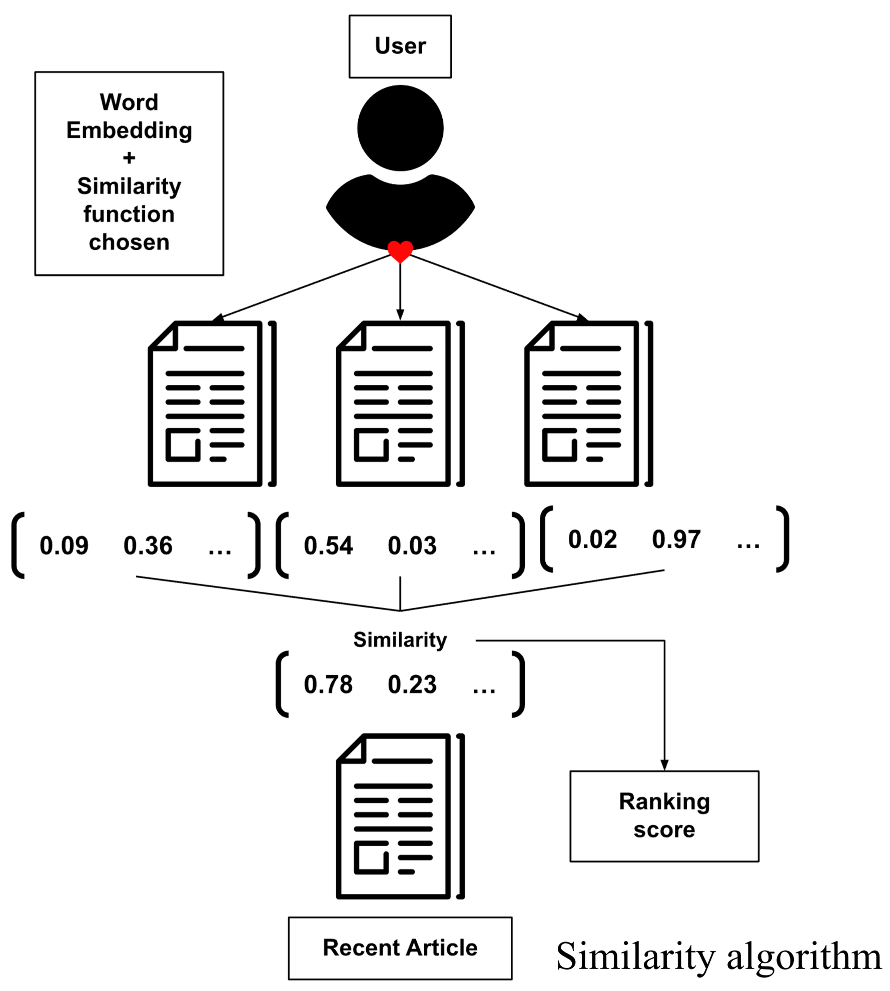
BM25 Ranking Algorithm

Word embedding techniques compute **ranking scores** using a **similarity function**.



Embeding Model

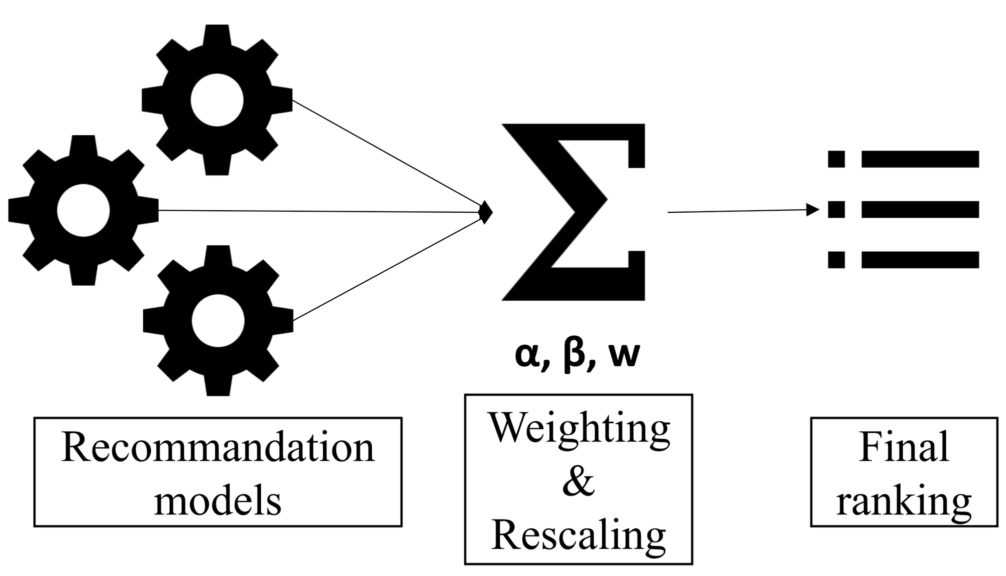
Pertinent word **embedding models** range from **semantic** models to **style representation** models.



Similarity algorithm

Combining Rankings

Reswhy (***Rescaled*** and ***Weigthed Hybridization***): Models combined to extract information from all.

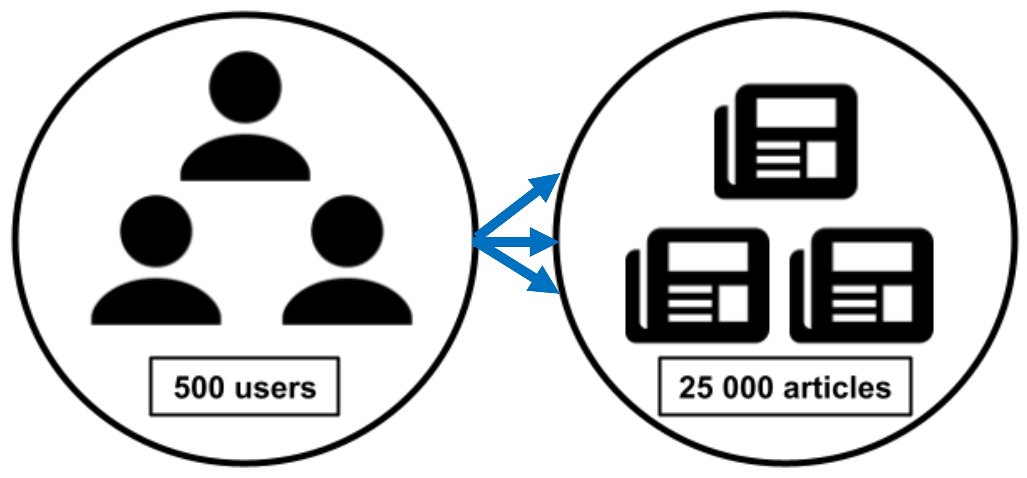


Reswhy is interesting on a combination of semantic model (***BM25***) and style-based model (***DBert-ft***).

CentraleSupélec | université PARIS-SACLAY

Training Dataset

The dataset we use is made of ***Twitter*** users and articles they shared.

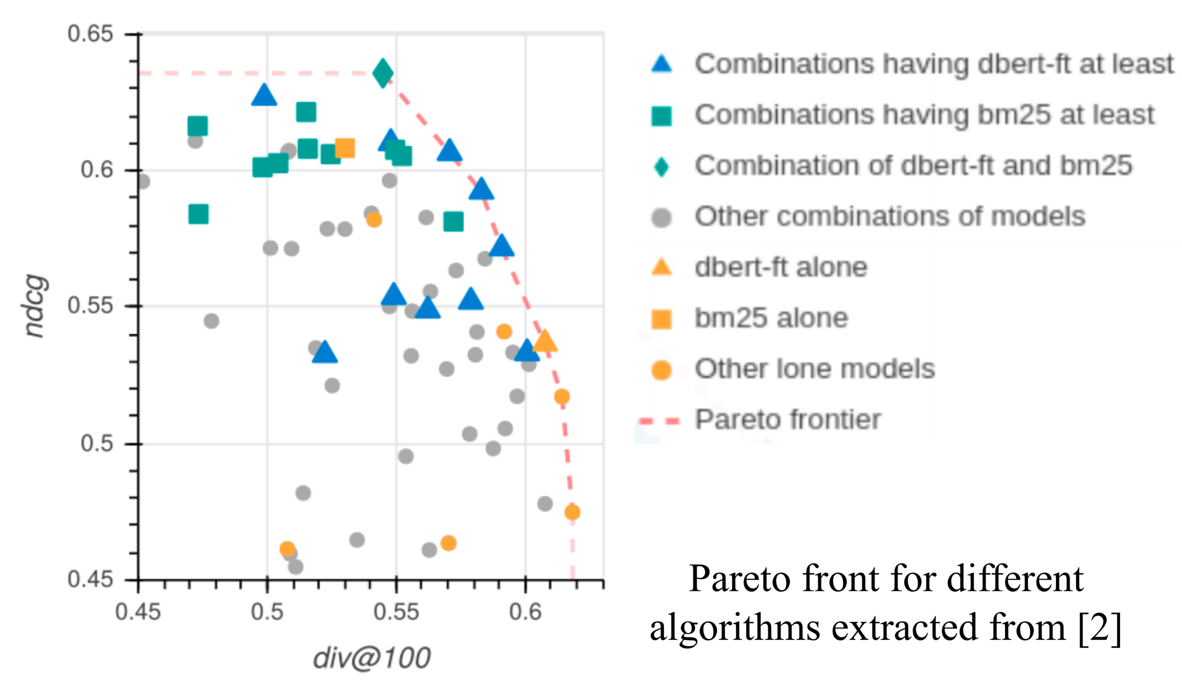


Results

	nDCG	MRR
BM25	0.45	0.27
DBert-ft	0.53	0.46
Reswhy	0.55	0.45

Results obtained to validate [2]

nDCG: Precision score considering item positions with decreasing weights
MRR: Score indicating how close the first relevant items in rankings are to the first position



Pareto front for different algorithms extracted from [2]

Pareto fronts are used to identify the best algorithms that favor multiple criteria simultaneously.

Style-based models give complementary information to **semantic** models and are interesting to explore with multi-modal data (video, images, text...)

References

[1] Claypool, M. (1999) « Combining Content-Based and Collaborative Filters in an Online Newspaper. SIGIR'99 Workshop on Recommender Systems : Algorithms and Evaluation »

[2] Hay, J. (2021) « Apprentissage de la représentation du style écrit, application à la recommandation d'articles d'actualité »

[3] Burke, R. (2007) « Hybrid Web Recommender Systems ».