

Semantic User Modeling for Personalized News Recommendation and Information Filtering

Christephor H. Blerk

School of Computing, Clemson University, Clemson, SC, USA.

clclark@clemson.edu

Abstract

Personalized news recommendation and information filtering have become critical components of contemporary digital information infrastructures. Traditional recommender systems rely heavily on behavioral signals such as clicks, dwell time, and explicit ratings, but these signals alone often fail to capture the dynamic, context-sensitive, and semantically rich nature of user interests. This paper presents a system-level examination of semantic user modeling as a foundation for personalized news recommendation and information filtering. It analyzes the conceptual shift from purely behavioral profiles toward semantically structured representations that incorporate natural language understanding, knowledge organization, and temporal interest evolution. The discussion focuses on architectural dimensions, including candidate retrieval, semantic indexing, memory layers, and re-ranking pipelines, while also addressing infrastructure trade-offs related to latency, scalability, model refresh, and operational sustainability. The paper further examines structural tensions between personalization accuracy and information diversity, the risks of filter bubbles, and the governance challenges arising from opaque language models and user profiling. It argues that semantic user modeling should be treated not only as a computational problem but also as a socio-technical design challenge that requires fairness, accountability, robustness, and policy awareness. By synthesizing cross-domain perspectives from recommender systems, information retrieval, machine learning, and media governance, the paper identifies future directions for building personalized news systems that balance user relevance with broader democratic information access.

Keywords

semantic user modeling; personalized news recommendation; information filtering; large language models; algorithmic governance; fairness; socio-technical infrastructure.

1. Introduction

The contemporary digital news environment is characterized by continuous production, high volume, rapid topic evolution, and increasing heterogeneity of sources. In this environment, information overload has become a structural condition rather than a temporary inconvenience. Recommender systems have emerged as essential infrastructural mechanisms for selecting, ordering, and presenting news content to readers. Foundational work on recommender systems has established the core functions of user modeling, item representation, and relevance estimation across multiple domains [1]. Hybrid recommendation approaches have further demonstrated that combining collaborative signals with content-based representations can improve robustness in sparse and dynamic settings [2]. At the same time, the evolution of recommender systems has been shaped by the need to move beyond simple rating prediction toward context-aware, multi-criteria, and semantically grounded personalization [3]. In news recommendation, these needs are particularly acute

because news items have short life cycles, new items appear continuously, and user interests shift in response to real-world events.

A central limitation of conventional behavioral profiling is that it treats users largely as accumulations of past interactions. Clicks and reading time can indicate interest, but they do not fully expose why an item was relevant, which concepts within the item matched existing knowledge, or how the user interpreted the news event. Semantic user modeling attempts to address this gap by constructing meaning-oriented representations of users that align with the linguistic and conceptual content of news articles. Such models can support more explainable interest inference, better cold-start handling for unseen articles, and greater continuity of user understanding across different topics and time periods. However, building semantic user models at scale introduces significant architectural and operational complexity. It requires the integration of natural language processing components, dynamic knowledge structures, and long-term memory mechanisms into existing recommendation pipelines.

This paper examines semantic user modeling from a system-level perspective rather than as an isolated algorithmic technique. It considers structural trade-offs across architecture, infrastructure, deployment, sustainability, robustness, fairness, and governance. The objective is to provide an interdisciplinary synthesis that connects computational mechanisms with the broader socio-technical realities of personalized news distribution. By analyzing the conceptual foundations, architectural patterns, deployment constraints, and policy implications, the paper contributes a framework for understanding how semantic personalization can be designed responsibly in large-scale information filtering systems.

2. From Behavioral Signals to Semantic User Models

Traditional personalization systems frequently represent users through explicit ratings, purchase histories, or implicit behavioral traces such as clicks, skips, and reading duration. Matrix factorization techniques have been widely used to derive latent user and item vectors from interaction matrices, thereby enabling compact preference modeling in collaborative filtering settings [4]. Although these latent factors can capture statistical regularities in user behavior, they are not grounded in the meaning of the content being consumed. In news recommendation, this weakness is particularly consequential because the relationship between two news articles may depend on conceptual overlap, event context, or narrative framing that is not visible from interaction patterns alone. A user who reads articles about monetary policy, inflation, and central bank decisions may be semantically interested in macroeconomic governance even if the specific terms and article identifiers change rapidly.

Semantic representation methods offer a way to encode the meaning of words, sentences, and documents in continuous vector spaces. Early distributed word representations demonstrated that related words can be located close to one another in high-dimensional spaces, enabling similarity computation that reflects usage patterns in large corpora [5]. The introduction of attention mechanisms further improved the capacity of neural architectures to model long-range dependencies and contextual relevance in sequential data [6]. Pre-trained language models built on transformer architectures have since become foundational for many text understanding tasks because they capture rich contextual semantics and can be fine-tuned for downstream applications [7]. In the context of user modeling, these representations allow a shift from purely behavioral profiles to profiles that combine interaction history with linguistic and conceptual understanding.

Semantic user models can be constructed by aggregating representations of the articles a user has read, the entities mentioned in those articles, and the latent topics that connect them. Temporal dynamics are important because user interest is not static: a reader may be intensely focused on an election for several weeks and then shift attention to environmental policy. A semantically grounded user model can represent this drift by updating concept weights, decaying old interests, and introducing new semantic associations without requiring the user to provide explicit feedback. Furthermore, semantic modeling can support generalization across vocabulary. If a reader previously engaged with articles about renewable energy, a new article discussing photovoltaic technologies can be recognized as relevant even if the exact terms have not appeared in the user's interaction history. This capability is essential for news recommendation, where novelty is a core property of the item stream.

3. Architectural Dimensions of Semantic Personalization Systems

Large-scale news personalization systems typically decompose the recommendation process into multiple stages, including candidate retrieval, scoring, and re-ranking. This staged architecture balances the need for broad coverage with the computational cost of deep semantic inference. In early large-scale news personalization, scalable online collaborative filtering was used to cluster users and generate recommendations efficiently, demonstrating that practical deployments must prioritize latency and throughput alongside model accuracy [8]. Content-based recommendation systems have long provided a complementary path by comparing item features with explicit or inferred user interest profiles [9]. In semantic architectures, content-based filtering can be extended through embeddings that represent articles and user interests in a shared semantic space, allowing similarity-based retrieval across lexical variations.

Embedding-based news recommendation has been developed to handle millions of users and articles by learning distributed representations of users, articles, and contextual features [10]. Neural news recommendation models have further incorporated personalized attention mechanisms that weigh different parts of an article or different historical interactions according to the current candidate item [11]. These approaches illustrate a structural shift from precomputed item similarities toward learned interaction functions that can capture complex user-item relationships. However, this shift introduces new architectural demands. The system must maintain vector indexes for rapid nearest-neighbor search, manage the freshness of article embeddings, and ensure that user representations remain consistent across session boundaries and device contexts.

A semantic personalization architecture also requires a memory layer that persists user understanding beyond a single session. This memory layer may include short-term session context, mid-term topic interests, and long-term conceptual preferences. The separation of these temporal scales is important because different recommendation decisions require different granularities of user history. Short-term context may be decisive when recommending follow-up articles about a breaking event, whereas long-term preferences may be more relevant for discovering new topics that align with established intellectual interests. The architectural challenge is to coordinate these layers without allowing recent behavior to overwhelm stable semantic preferences or, conversely, allowing old preferences to prevent the discovery of new interests.

4. Infrastructure, Deployment, and Lifecycle Management

Deploying semantic user modeling in production environments requires more than accurate models. The surrounding infrastructure must support continuous ingestion of news content, low-latency semantic encoding, vector search, user profile update, and real-time scoring. These components must operate within strict resource budgets because news recommendation traffic can be highly variable and event-driven. Infrastructure decisions such as the choice of vector database, the granularity of caching, and the frequency of model refresh directly affect both operational cost and recommendation quality. Recent work on large language model based user memory systems has examined how persistent semantic representations can improve query matching and intent modeling in filtering pipelines, suggesting that memory-centric architectures may become increasingly important as language models are integrated into recommendation infrastructures [12]. The emphasis in such work is on maintaining coherent user state across interactions rather than treating each request as an isolated inference problem.

News recommender systems face specific lifecycle challenges because the item catalog changes continuously. Articles become stale quickly, entity graphs expand, and language use evolves. A semantic model trained on historical text may fail to capture emerging terminology or newly significant entities. Model refresh strategies must therefore include incremental updates, periodic retraining, and mechanisms for detecting distribution shift. Surveys of news recommendation have identified the dynamic nature of news content as a central obstacle and have called for architectures that can adapt quickly without sacrificing stability [13]. The lifecycle of a semantic user model is similarly dynamic. User interests evolve through events, seasonal changes, and gradual intellectual shifts, requiring continuous re-estimation of profile parameters.

Operational sustainability is another critical consideration. Large language models and deep semantic encoders can be expensive to train and serve. In a news filtering context, the cost of generating embeddings for every candidate article and every active user can become prohibitive. System designers must therefore evaluate trade-offs between model capacity, inference latency, and environmental impact. Caching, distillation, quantization, and selective computation may reduce resource use but can also introduce accuracy loss. These operational trade-offs are structural rather than incidental because they shape the feasible design space for semantic personalization. A system that provides rich semantic understanding but cannot scale economically will not be sustainable as an information infrastructure.

5. Robustness, Fairness, and Algorithmic Governance

The application of semantic user modeling to news recommendation raises significant normative questions because the selection and ordering of news content influence public knowledge and civic engagement. Research on the democratic role of news recommenders has emphasized that personalization should not undermine the diversity of information available to citizens [14]. The concept of the filter bubble has drawn attention to the possibility that personalized systems can isolate users within narrow informational environments [15]. In democratic societies, the health of public discourse depends partly on exposure to challenging perspectives, unexpected topics, and cross-cutting viewpoints [16]. Semantic user models that overfit to past preferences may inadvertently reduce this exposure even if they improve short-term engagement metrics.

Fairness concerns also arise from the way user models encode social categories, interests, and linguistic patterns. Recommendation systems can reproduce existing disparities in news access, amplify sensational content, or systematically underrepresent certain communities.

Research on the ethics of algorithms has mapped issues of opacity, bias, and accountability across algorithmic decision-making contexts [17]. Algorithmic accountability frameworks emphasize the need for transparency about how recommendations are generated and the ability to audit system behavior [18]. In semantic user modeling, auditability is complicated by the use of dense vector representations and deep language models whose internal reasoning is difficult to inspect. This opacity can obscure why a particular story was promoted or suppressed.

Large language models introduce additional robustness and bias concerns. Pre-trained models may encode societal stereotypes, historical biases, and problematic associations that can be transferred into user profiles or article representations. Critical analyses have shown that increasingly large language models are not automatically more trustworthy and can perpetuate harmful patterns when deployed in high-impact applications [19]. Fairness in machine learning is not reducible to a single metric; it requires careful specification of protected attributes, outcome definitions, and intervention points [20]. In news recommendation, fairness must be considered at the level of users, topics, sources, and affected communities. A semantically rich system can be used to diversify content intentionally, but only if diversity and fairness are treated as explicit design objectives rather than as post-hoc adjustments.

6. Sustainability and Operational Trade-offs

The long-term sustainability of semantic news personalization depends on the alignment of technical capabilities with organizational resources, environmental constraints, and social expectations. Large models may improve semantic understanding, but they also increase energy consumption and infrastructure complexity. System architects must decide when deep language understanding is necessary and when lighter semantic features are sufficient. For example, candidate retrieval may use efficient lexical or entity-based features, while a deeper semantic model may be applied only to a small set of final candidates. This staged use of semantic capacity can reduce cost while preserving most of the quality benefit. Similar trade-offs apply to user memory systems, where storing rich semantic states for every user can create substantial storage and update costs.

Robustness is closely linked to sustainability because systems that are brittle under distribution shift require frequent manual intervention and expensive retraining. A robust semantic user model should degrade gracefully when article encoders encounter new vocabulary, when user interests change abruptly, or when source metadata is incomplete. Redundant representations, fallback retrieval strategies, and confidence estimation can help maintain system performance under uncertainty. The integration of large language model based memory mechanisms may increase adaptability, but it also introduces new failure modes related to hallucination, state corruption, and unclear boundaries between user memory and model-generated inference [12]. Operational monitoring must therefore include not only traditional metrics such as click-through rate and latency but also semantic drift, profile coherence, and unexpected shifts in content exposure.

7. Policy Implications and Future Directions

The deployment of semantic user modeling in news recommendation intersects with emerging regulatory frameworks governing platform accountability, data protection, and artificial intelligence. Policymakers increasingly expect recommendation systems to provide meaningful transparency, user controls, and safeguards against harmful amplification. The democratic role of news recommenders depends on whether personalization systems expose

users to a sufficiently broad range of public discourse [14]. Algorithmic accountability mechanisms require that organizations can explain, audit, and justify the design choices embedded in their recommendation pipelines [18]. Semantic user models introduce particular challenges for these obligations because the profiles they construct are complex, dynamic, and difficult to communicate to end users. Future work should develop interpretable user model summaries that allow individuals to understand how their interests are represented and how those representations influence the news they see.

The governance of semantic personalization should also address the data used to train language models and the ways in which those models shape user understanding. Training data biases can be inherited by recommendation systems, even when the original task appears neutral [19]. Fairness considerations must move beyond individual-level parity to include structural effects such as source diversity, topic suppression, and the distribution of attention across communities [20]. This requires collaboration among computer scientists, media scholars, legal researchers, and civil society organizations. News personalization cannot be optimized solely for engagement or satisfaction because the informational environment has collective implications. A balanced governance approach would combine technical safeguards, external audits, and participatory mechanisms for users to contest or adjust their semantic profiles.

Future research directions include the development of semantically grounded diversification strategies that deliberately introduce curiosity-driven recommendations, the design of memory architectures that support user agency, and the evaluation of personalization systems using civic information health metrics rather than engagement alone. Another promising direction is the integration of cross-domain knowledge from information retrieval, digital journalism, and computational social science to better understand how semantic profiles evolve over long periods. These directions require long-term interdisciplinary research programs and cannot be addressed by isolated algorithmic improvements. The principal challenge is not merely to model meaning more accurately, but to govern the systems that act upon that meaning in ways that respect both individual autonomy and collective democratic interests.

8. Conclusion

Semantic user modeling offers a powerful foundation for personalized news recommendation and information filtering by enabling systems to represent user interests in terms of meaning rather than behavior alone. The shift from purely behavioral profiles to semantically structured user models can improve generalization, support cold-start handling, and enable more explainable interest inference. However, this shift also introduces substantial architectural, infrastructural, and governance challenges. Semantic models require coordinated pipelines for retrieval, scoring, memory management, and model refresh. They impose operational costs related to inference latency, storage, and energy consumption. Moreover, the use of language models and dense vector representations raises difficult questions about opacity, bias, filter bubbles, and the role of personalization in democratic societies.

A system-level perspective is therefore essential. Semantic user modeling should not be treated as an independent technique but as part of a larger socio-technical infrastructure that shapes how citizens encounter news and understand public affairs. The design of such systems must integrate technical performance with fairness, robustness, accountability, and sustainability. Future progress will depend on interdisciplinary collaboration and the

development of governance mechanisms that ensure semantic personalization serves both individual relevance and collective informational health.

References

1. Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Springer.
2. Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331-370.
3. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749.
4. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.
5. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30.
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186.
8. Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). Google news personalization: Scalable online collaborative filtering. *Proceedings of the 16th International Conference on World Wide Web*, 271-280.
9. Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325-341). Springer.
10. Okura, S., Tagami, Y., Ono, S., & Tajima, A. (2017). Embedding-based news recommendation for millions of users. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1933-1942.
11. Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). NPA: Neural news recommendation with personalized attention. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2576-2584.
12. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
13. Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems: Survey and roads ahead. *Information Processing & Management*, 54(6), 1203-1227.
14. Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), 993-1012.
15. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.
16. Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

17. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21.
18. Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.
19. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
20. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.