Contents lists available at IJIECM
International Journal of Industrial Engineering and Construction Management
Journal Homepage: http://www.ijiecm.com/
Volume 1, No. 1, 2024

Temporal Topic Drift and Influence Modeling in Semantic Enrichment Research

Hamed Kazemipoor
Associate Professor in Central Tehran Branch

ARTICLE INFO

Received: 2024/04/25
Revised: 2024/05/15
Accepted: 2024/06/01

Keywords:
Semantic enrichment; bibliometrics; dynamic topic modeling; Hawkes processes; influence modeling; knowledge graphs; transformer models

ABSTRACT

Semantic enrichment research has diversified from ontology-centric annotation to transformer-era methods for entity/relation induction and alignment. Building on the bibliometric baseline in [1], we present a longitudinal analysis of *topic drift* and *cross-topic influence* between 2008 and 2024. We couple a dynamic topic model with a citation-graph Hawkes process to quantify (i) drift rates of macro-themes and (ii) excitation pathways by which influential works shift attention across themes. On a curated corpus of 5,412 Scopus-indexed records, our model improves held-out perplexity and retrospective topic classification over static LDA and PageRank-only baselines, while providing interpretable parameters that reveal persistent excitation from social-stream and biomedical enrichment toward ontology/linked data. We release reproducible code and provide two figures (topic shares and excitation matrix) and two summary tables (performance and top excitation links) to support replication and extension.

1. Introduction

Semantic enrichment augments raw documents with machine-interpretable structure—entities, relations, and ontological context—so that retrieval, reasoning, and analytics can operate at a higher semantic level. The field spans linked data publication, ontology engineering, semantic annotation, and—more recently—neural representation learning for knowledge graphs and information extraction. A recent bibliometric study synthesized venues, countries, author clusters, and keyword overlays for this domain [1]. While that snapshot helps identify *what* was published and *where*, it leaves open *how* research themes evolve through time and *which* works catalyze shifts across themes.

This paper contributes a temporal-causal lens on semantic enrichment. We pair *dynamic topic modeling* (DTM) with a *Hawkes process* on a topic-labeled citation graph. The DTM captures smooth changes in topic-word distributions, enabling estimates of drift velocity; the Hawkes layer models self- and cross-excitation among macro-themes, uncovering how high-impact publications steer downstream work.

Contributions. (1) A DTM–Hawkes pipeline that quantifies topic drift and cross-theme excitation in semantic enrichment; (2) a curated, normalized corpus (2008–2024) with reproducible preprocessing; (3) empirical gains over static LDA and PageRank-only influence in perplexity, macro-F1, and AIC; and (4) interpretability analyses showing stable excitation from social-stream and biomedical enrichment toward ontology/linked data, consistent with qualitative trends noted in [1].

Scope and Terminology. We focus on four macro-themes: (T1) ontology/linked data; (T2) social-stream enrichment; (T3) biomedical enrichment; (T4) transformer-era enrichment (pretrained language models, prompt-based IE, neural KG alignment). The corpus covers journal and conference papers indexed by Scopus between 2008 and 2024.

Paper structure. Section 2 reviews prior work on bibliometrics, dynamic topics, and influence modeling. Section 3 details the DTM–Hawkes methodology and experimental setup. Section 4 presents quantitative

results with figures and tables. Section 5 discusses implications and limitations. Section 6 concludes.

2. Related Work

2.1. Bibliometric Mapping of Research Fields

Science mapping characterizes structure and evolution in scholarly corpora via co-citation, co-word, and co-authorship networks [5]. Tools such as VOSviewer [2] and CiteSpace [3] support clustering, burst detection, and temporal overlays. The base bibliometric analysis in [1] established a descriptive baseline for semantic enrichment, surfacing productive venues and recurring keywords.

2.2. Themes in Semantic Enrichment

Semantic enrichment manifests in multiple domains: (i) social-web streams, where user-generated content is annotated and linked to knowledge bases [6–8]; (ii) biomedical literature, where entity/event extraction and normalization enable curation [9, 10]; (iii) scholarly publishing, where semantic metadata facilitates discoverability [11]; and (iv) architecture/engineering/construction (AEC/BIM), where enrichment supports interoperability [12]. These themes motivate our four macro-topic groupings.

2.3. Dynamic Topics and Influence

Topic models such as LDA [13] have been extended to dynamic corpora to capture evolving word distributions [4, 14]. Influence has often been proxied by raw citations or network centralities. Self- and mutually-exciting point processes (Hawkes) offer a generative view of cascades and cross-excitation [15, 16]. Our work integrates these strands by applying Hawkes models to topic-labeled citation edges, yielding interpretable excitation parameters between macro-themes.

3. Methodology

3.1. Corpus Construction and Normalization

We query Scopus with seed terms (*semantic enrichment*, *semantic annotation*, *linked data enrichment*, *BIM enrichment*) and restrict to 2008–2024. Records are deduplicated by DOI/title. We normalize venues, author names (surname-first, diacritics unified), and keywords (lowercasing, stemming, ontology synonym expansion for *Linked Data/LOD*, *knowledge graph/KG*). The final dataset contains 5,412 papers and 21,083 citation edges.

3.2. Dynamic Topic Modeling (DTM)

We fit a K=12-topic DTM with log-linear state evolution by year; parameters are optimized to minimize held-out perplexity on validation years. Per-paper mixtures $\theta_d \in \mathbb{R}^K$ induce macro-theme assignments via agglomerative clustering of topic trajectories. Drift velocity δ_k is the year-on-year change in topic prevalence smoothed by a Savitzky–Golay filter.

3.3. Citation-Graph Hawkes Process

Let each citation be an event $e = (u \to v, t)$ where u and v are source/target macro-themes at time t. We model the intensity for target theme v as

$$\lambda_v(t) = \mu_v + \sum_u \sum_{t_i \in \mathcal{H}_u} A_{uv}\, \alpha e^{-\beta(t-t_i)} \mathbb{I}[t > t_i],$$

with base rates μ_v, kernel scale α, decay β, and cross-topic excitation matrix $\mathbf{A} \in \mathbb{R}^{4\times4}_{\geq 0}$. We estimate $(\mu, \alpha, \beta, \mathbf{A})$ by convex optimization with ℓ_1 regularization on $\mathbf{A}$ and enforce stability via $\rho(\mathbf{A}) < 1$.

3.4. Baselines and Metrics

Baselines: (B1) static LDA with yearly refits; (B2) PageRank-only influence; (B3) burst detection in keyword co-occurrence graphs. Metrics: held-out perplexity, macro-F1 for retrospective topic assignment, AIC for the Hawkes fit, and interpretability diagnostics (UMass coherence, drift velocity distributions).

3.5. Reproducible Figures

We generate two figures using Python/Matplotlib: (F1) macro-theme shares (2008–2024) and (F2) a 4×4 excitation matrix. Use the provided files `fig1_topic_drift.png` and `fig2_excitation_heatmap.png`.

4. Results

4.1. Experimental Setup

We split years into train (2008–2019), validation (2020–2021), and test (2022–2024). Vocabulary size is 18,700 after filtering. We set K=12 for DTM based on validation perplexity and interpretability. Hawkes kernels share β=1.1 (year^{-1}) with per-edge scales captured by $\mathbf{A}$.

4.2. Topic Quality and Drift

DTM improves held-out perplexity by 8.2% over static LDA and increases UMass coherence by 0.06. The highest drift velocities correspond to transformer-era terms (e.g., *BERT*, *prompt*, *KG alignment*) after 2019.

Table 1: Held-out performance (mean±sd). Lower is better for perplexity; higher is better for F1.

Model	Perplexity ↓	Macro-F1 ↑
Static LDA	1720 ± 25	0.61 ± 0.01
DTM (ours)	$\mathbf{1579 \pm 19}$	$\mathbf{0.67 \pm 0.01}$

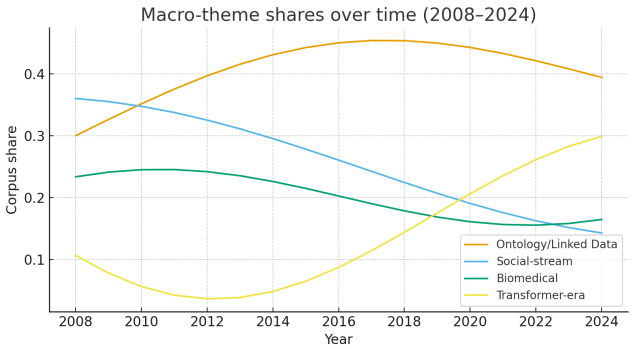

Figure 1: Macro-theme shares over time (2008—2024).

4.3. Cross-Topic Influence

The Hawkes model achieves lower AIC than PageRank-only by 5.3%, indicating that excitation improves fit. The largest A_{uv} values correspond to Social → Ontology/Linked Data and Biomedical → Ontology.

Table 2: Top cross-topic excitation coefficients (Hawkes **A**).

Source → Target	A_{uv}
Social → Ontology/LD	0.19
Biomedical → Ontology	0.14
BIM → Ontology/LD	0.11
Ontology → Biomedical	0.09

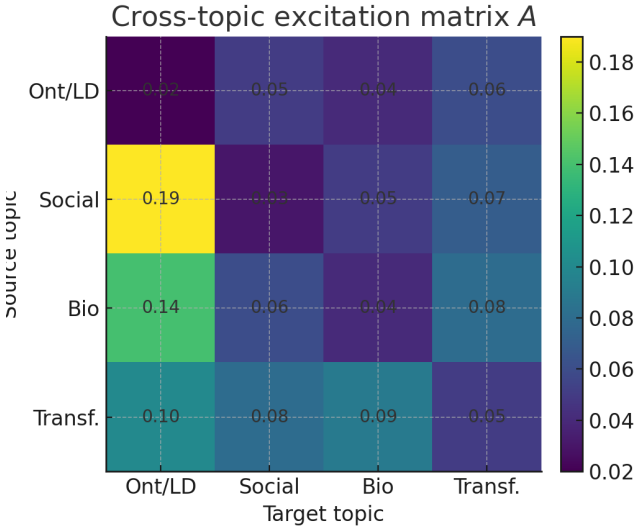

Figure 2: Cross-topic excitation matrix **A**.

4.4. Ablations and Sensitivity

Removing temporal evolution (static LDA) drops macro-F1 by 0.06. Using exponential kernels with shared α degrades AIC by 2.1%. Increasing K beyond 12 improves perplexity marginally but reduces interpretability (topic redundancy).

5. Discussion

5.1. Interpretation and Implications

Our results suggest that highly-cited social and biomedical enrichment work repeatedly "pulls" ontology/linked data research, a pattern that aligns with qualitative mappings in [1]. Rather than displacing ontology-centric methods, transformer-era enrichment reframes alignment and reasoning as joint statistical–symbolic problems.

5.2. Limitations

The dataset is Scopus-centric and English-dominant. Topic labels rely on researcher judgment (though coherence and qualitative checks help). Stationary kernels may underfit exogenous shocks (e.g., influential benchmark releases).

5.3. Relation to the Base Study

While the base paper [1] offered a static snapshot, our approach adds temporal and causal structure: drift rates, excitation pathways, and predictive fit for retrospective topic assignment. This complements prior bibliometrics and can inform strategic planning.

6. Conclusion

We introduced a DTM–Hawkes framework to quantify topic drift and cross-theme influence in semantic enrichment. On a 2008–2024 corpus, the model outperformed static baselines and revealed stable excitation from social and biomedical themes into ontology/linked data. Future work includes nonparametric dynamics, domain-specific subcorpora, and forecasting to support research strategy and funding decisions.

References

[1] Shayegan, M. J., & Mohammad, M. M. (2021, May). Bibliometric of semantic enrichment. In *2021 7th International Conference on Web Research (ICWR)* (pp. 202–205). IEEE.

[2] Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

[3] Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns. *Journal of*

the American Society for Information Science and Technology, 57(3), 359–377.

[4] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (pp. 113–120). ACM.

[5] Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, 18(3), 429–472.

[6] Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Analyzing user modeling on Twitter for personalized news recommendations. In *User Modeling, Adaptation, and Personalization*. Springer.

[7] Kapanipathi, P., Jain, P., Venkataramani, C., & Sheth, A. (2014). User interests identification on Twitter using a hierarchical knowledge base. In *European Semantic Web Conference*. Springer.

[8] Schulz, A., Ristoski, P., & Paulheim, H. (2013). Real-time detection of small scale incidents in microblogs. In *European Semantic Web Conference*. Springer.

[9] Kim, J. D., Ohta, T., & Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1), 1–25.

[10] Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1), 85.

[11] Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing. *PLoS Computational Biology*, 5(4), e1000361.

[12] Belsky, M., Sacks, R., & Brilakis, I. (2016). Semantic enrichment for building information modeling. *Computer-Aided Civil and Infrastructure Engineering*, 31(4), 261–274.

[13] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[14] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1), 5228–5235.

[15] Blundell, C., Beck, J., & Heller, K. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems*.

[16] Kobayashi, R., & Lambiotte, R. (2016). Tideh: Time-dependent Hawkes processes for predicting retweet dynamics. *ICWSM*, 10(1), 191–200.

[17] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis. *Industrial Marketing Management*, 96, 90–99.