Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(9), 3033-3044
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Ontology-Guided Semantic Retrieval in Virtual University Information Systems: An Empirical Benchmark Study

Mamadou Bakouan, Tiemoman Kone, Yao Konan

Département Informatique et Science du Numérique, Unité de Recherche et d'Expertise Numérique, Université Virtuelle de Côte d'Ivoire, Abidjan, Côte d'Ivoire
Email: mamadou.bakouan@uvci.edu.ci

How to cite this paper: Bakouan, M., Kone, T. and Konan, Y. (2026) Ontology-Guided Semantic Retrieval in Virtual University Information Systems: An Empirical Benchmark Study. *Open Journal of Applied Sciences*, **16**, 3033-3044.
https://doi.org/10.4236/ojapps.2026.169167

Received: August 16, 2026
Accepted: September 5, 2026
Published: September 8, 2026

Abstract

Semantic retrieval in virtual university information systems must address lexical variation and domain specific relationships. This study reports a controlled empirical evaluation of lexical, semantic, ontology-based, and combined retrieval configurations within a virtual-university benchmark. The benchmark contains 20 queries and 38 documents, with 70 judged query document pairs and 690 un judged pairs. Five configurations are evaluated using P5, R5, F1_5, MRR, and nDCG5, with query level analysis and multiplicity controlled statistical comparisons. The combined configurations achieve the highest recorded aggregate scores, including nDCG5 = 0.9894. The largest recorded descriptive difference relative to ontology only retrieval is ΔnDCG5 = 0.07308; the corresponding recorded Holmadjusted p-value is 0.05388 and therefore does not meet the predefined 0.05 significance threshold. Sensitivity analysis identifies a local performance plateau for the tested ontology-dominant interval $\alpha = 0$ and $\beta = 0.05$ - 0.30. The findings provide benchmark-specific descriptive and suggestive evidence for combining semantic and explicit domain knowledge signals, while indicating the need for larger benchmarks and independently reproducible query level statistical analyses.

Keywords

Semantic Retrieval, Ontology-Based Retrieval, Information Retrieval, Virtual University, Sensitivity Analysis

1. Introduction

Digital transformation in higher education has increased the volume and heterogeneity of institutional information that students, administrative staff, instructors,

and other users must access. In a virtual university information system (VUIS), relevant information may be distributed across academic regulations, programme descriptions, administrative procedures, library resources, institutional announcements, and other heterogeneous documents. Effective retrieval therefore requires mechanisms that can identify relevant information despite lexical variation and differences in document formulation.

Traditional information retrieval relies strongly on lexical representations and term-weighting schemes. TF-IDF provides a classical representation of term importance within documents and collections [1]. Such approaches remain useful for exact or near-exact matching, but vocabulary mismatch can prevent relevant documents from being retrieved when the query and document express related concepts using different terms [2].

Distributed representations provide a complementary mechanism by encoding words or texts in continuous vector spaces [3]. Contextual language models such as BERT [4] and sentence-level representations such as Sentence-BERT [5] further support semantic matching. Surveys of semantic first-stage retrieval and pre-trained-language-model-based dense retrieval show the evolution from classical term-based retrieval toward sparse, dense, and hybrid semantic approaches [2] [6].

Semantic representations do not necessarily provide an explicit representation of domain concepts and their relationships. Ontologies provide a complementary knowledge-representation mechanism for representing domain concepts and relations [7]. Recent retrieval studies have combined ontology-based representations with distributed semantic models to reduce vocabulary mismatch and improve domain-sensitive indexing and retrieval [8] [9].

Educational information access has also been addressed using ontology and semantic retrieval mechanisms. Examples include ontology-based educational-program counselling [10], ontology-based question answering for university admissions [11], semantic search in learning resources [12], ontology-driven search for remote-learning environments [13], and ontology-assisted search of higher-education regulations [14]. These studies demonstrate the relevance of explicit domain knowledge to educational information access, but their datasets, tasks, representations, and evaluation protocols differ substantially.

However, these studies do not provide a common experimental basis for quantifying the relative contribution of lexical, semantic, and ontology-based signals in a VUIS retrieval task. In particular, controlled query-level comparison and sensitivity analysis of hybrid weighting remain important for determining whether observed improvements are robust within a defined benchmark.

This study addresses this evaluation problem in a VUIS. Five retrieval configurations are compared under a common query-document benchmark using P5, R5, F1_5, MRR, and nDCG5. The evaluation combines descriptive performance analysis, five-fold query-level evaluation, sensitivity analysis, and multiplicity-controlled statistical comparisons. The standard information-retrieval evaluation ter-

minology follows [15], while nDCG is grounded in the cumulative-gain framework of [16].

The study addresses three research questions: RQ1, does hybrid retrieval improve effectiveness relative to the lexical baseline? RQ2, does combining the Semantic Proxy with ontology-based similarity improve effectiveness relative to ontology-only retrieval? RQ3, how sensitive is retrieval effectiveness to the semantic-ontology weighting within the tested parameter range?

The contribution is deliberately evidence bounded. The study evaluates lexical, Semantic Proxy, ontology-based, and combined retrieval configurations within a defined VUIS benchmark. It contributes a controlled comparison, query-level evaluation protocol, coefficient sensitivity analysis, and an explicit separation between descriptive performance and multiplicity-controlled statistical evidence. The component comparisons are not presented as a complete factorial ablation.

2. Related Work

Classical information retrieval relies on lexical representations and term weighting. TF-IDF remains a foundational baseline [1], but lexical matching is sensitive to vocabulary mismatch.

Distributed word representations [3], contextual language models [4], and sentence level representations [5] address aspects of semantic mismatch. Recent surveys consolidate semantic first stage retrieval and dense retrieval as major retrieval paradigms [2] [6].

Ontologies explicitly represent concepts and relationships within a domain [7]. Recent work combines ontology-based representations with distributed semantic models for document retrieval and semantic indexing [8] [9].

In educational settings, ontology-based systems have been investigated for counselling [10], university admissions [11], learning-resource search [12], remote learning [13], and higher-education regulations [14].

The research gap is therefore not the absence of semantic-ontology retrieval. Rather, it concerns controlled empirical comparison under a common benchmark, particularly the relative contribution of retrieval signals and the sensitivity of hybrid performance to weighting.

3. Methodology

3.1. Research Design

This study adopts a controlled empirical evaluation design. The same query document benchmark is used for all retrieval configurations. The general hybrid scoring function is defined as:

$$S_H(q,d) = \alpha S_{LEX}(q,d) + \beta S_{SP}(q,d) + \gamma S_{ONT}(q,d) \quad (1)$$

where $S_{LEX}(q,d)$ denotes lexical similarity, $S_{SP}(q,d)$ denotes the Semantic Proxy score, and $S_{ONT}(q,d)$ denotes ontology-based similarity. The coefficients α, β, and γ satisfy $\alpha + \beta + \gamma = 1$. The Semantic Proxy, denoted by $S_{SP}(q,d)$, is an ex-

perimentally defined semantic similarity signal used as an intermediate retrieval component in this study. It is distinct from a specific pretrained sentence encoder and is therefore not interpreted as an SBERT embedding. Its role is to provide a normalized semantic signal that can be combined with lexical similarity and ontology-based similarity under the constraint $\alpha+\beta+\gamma=1$. The present experimental documentation establishes the role and normalization of the Semantic Proxy but does not expose sufficient implementation-level information to reconstruct its complete preprocessing and scoring pipeline independently. We therefore avoid attributing the signal to a specific pretrained architecture or introducing undocumented preprocessing assumptions.

3.2. Benchmark and Ontology

The benchmark comprises 20 queries and 38 documents, yielding 760 possible query-document pairs. The experimental protocol records 70 judged query-document pairs as a draft-validated gold standard and 690 unjudged pairs. The same benchmark is retained across all retrieval configurations, with five query-level folds and four queries per fold.

The experimental ontology contains 52 concept nodes and 35 relations organized into 10 normalized relation types. Sixteen of the 20 benchmark queries have an exploitable ontology mapping, corresponding to 80% coverage. Four queries (Q07, Q09, Q10, and Q20) do not have an exploitable query-side ontology mapping. These queries were retained in the benchmark so that the evaluation population remained fixed at 20 queries across all compared configurations. For these queries, no exploitable query-side ontology mapping is available for computing the ontology-based similarity signal. Consequently, the evaluation population remains fixed at 20 queries across all compared configurations and should not be interpreted as evidence that ontology-based retrieval is equally applicable to all query types. The available artifacts do not provide a separate query-level attribution analysis sufficient to quantify the individual contribution of Q07, Q09, Q10, and Q20 to the aggregate configuration differences; this is therefore retained as a limitation of the present benchmark analysis.

The judged pairs constitute the available relevance evidence used for evaluation, whereas unjudged pairs are not converted into explicit negative relevance judgments. The complete benchmark inventory and experimental protocol are maintained in the associated experimental artifacts [17] [18]. The available documentation does not independently identify the judges or provide a complete adjudication history; this aspect is therefore retained as an evidence limitation rather than reconstructed retrospectively.

The available experimental documentation establishes the ontology size, relation structure, and mapping coverage, but does not provide sufficient implementation-level detail to reconstruct independently the complete concept-mapping and relation-weighting procedure. Accordingly, the present study treats $S_{ONT}(q,d)$ as the documented ontology-based similarity signal and does not introduce undocu-

mented assumptions concerning individual relation weights or path-computation rules. The main characteristics of the experimental benchmark and ontology are summarized in Table 1.

Table 1. Characteristics of the experimental benchmark and ontology.

Characteristic	Value
Queries	20
Documents	38
Possible query-document pairs	760
Gold-standard judged pairs	70
Unjudged pairs	690
Ontology concept nodes	52
Ontology relations	35
Normalized relation types	10
Queries with exploitable ontology mapping	16
Ontology coverage	80%
Cross-validation folds	5
Queries per fold	4

3.3. Retrieval Configurations

Five retrieval configurations are evaluated in this study, as summarized in Table 2. The lexical baseline is TF-IDF/Cosine, the semantic component is represented by the Semantic Proxy, and the ontology-only configuration isolates explicit domain knowledge.

Table 2. Retrieval configurations evaluated in the study.

Configuration	Description	α	β	γ
M1	TF-IDF/Cosine lexical baseline	1.00	0	0
M4	Semantic Proxy	0	1.00	0
M5	Ontology-only	0	0	1.00
MSO	CV-selected Semantic-Ontology	0	0.10	0.90
M6	Hybrid sensitivity configuration	0	0.05	0.95

The five-fold evaluation retains MSO = (0, 0.10, 0.90). M6 = (0, 0.05, 0.95) is an additional configuration used in pairwise analysis and sensitivity testing. The configuration values and aggregate performance values are directly recorded in the final performance artifact [18].

3.4. Evaluation Metrics

Retrieval effectiveness is evaluated at the query level using Precision at 5 (P5),

Recall at 5 (R5), F1 at 5 (F1_5), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain at 5 (nDCG5). P5 measures the proportion of judged relevant retrieved documents among the five retrieved positions. R5 measures the proportion of judged relevant documents retrieved within the top five positions relative to the total number of judged relevant documents for the query. F1_5 is computed for each query as the harmonic mean of P5 and R5. MRR is computed from the reciprocal rank of the first judged relevant result, and nDCG5 evaluates ranking quality within the top five positions. The reported aggregate values are macro-averages over the 20 benchmark queries rather than pooled counts.

An unjudged query-document pair is not converted into an explicit negative relevance judgment. An unjudged document may nevertheless occupy one of the five retrieved positions because the evaluation retains a fixed Top-5 cutoff. Thus, the absence of a relevance judgment is distinguished from an explicit non-relevance label, while the retrieved ranking remains defined over five positions. In particular, P5 retains the fixed denominator of five retrieved positions; this should not be interpreted as assigning an explicit negative relevance label to an unjudged document.

The available experimental artifacts do not expose the complete metric-generation implementation. Consequently, the exact implementation-level treatment of an unjudged item within each metric, particularly for ranking-based measures such as MRR and nDCG5, cannot be independently reconstructed from the manuscript and preserved artifacts alone. No undocumented post-hoc rule is therefore introduced. The reported values are retained as recorded experimental results, and this implementation-level limitation is explicitly acknowledged for reproducibility.

The metric definitions are given in Equations (2)-(6) [15] [16].

$$P_5 = \frac{\text{relevant retrieved documents in top 5}}{5} \tag{2}$$

$$R_5 = \frac{\text{relevant retrieved documents in top 5}}{\text{total judged relevant documents}} \tag{3}$$

$$F_{1,5} = 2P_5R_5 \big/ \left(P_5 + R_5\right) \tag{4}$$

$$MRR = \frac{1}{\text{rant of the first relevant result}} \tag{5}$$

$$\text{nDCG}_5 = \frac{\text{DCG at 5}}{\text{IDCG at 5}} \tag{6}$$

3.5. Statistical Analysis

The query is the paired unit of analysis. Pairwise comparisons are evaluated using the Wilcoxon signed-rank test for matched observations [17], followed by Holm's step-down procedure to control the family-wise error rate at a threshold of 0.05 [18]. The preserved experimental documentation does not explicitly expose the

complete hypothesis-family definition underlying the recorded Holm values. Therefore, the present manuscript does not infer an alternative correction family from the displayed p-values alone, and the adjusted values are treated as recorded experimental results rather than independently reconstructed quantities.

Statistical significance is claimed only when the Holm-adjusted p-value is below 0.05. Accordingly, the recorded adjusted value of 0.05388 for the M6-versus-M5 nDCG@5 comparison is not interpreted as statistically significant. Descriptive differences in retrieval scores are reported separately from inferential evidence, and an unadjusted p-value is not considered sufficient evidence of statistical significance.

3.6. Sensitivity and Evidence Controls

Sensitivity analysis varies the semantic and ontology coefficients while preserving $\alpha + \beta + \gamma = 1$. The analysis distinguishes an observed performance plateau from a globally optimal parameter setting. Unjudged query document pairs are not converted into negative relevance judgments, and the benchmark is not modified after observing results.

Metric definition audit: the manuscript distinguishes per query metric computation from subsequent macro-averaging. Because the available experimental artifacts do not expose the complete metric generation code, no additional assumption is made about the treatment of an unjudged item inside an individual top five list beyond the documented rule that unjudged pairs are not treated as nonrelevant.

4. Results

4.1. Aggregate Retrieval Performance

The Semantic Ontology and Hybrid configurations achieve the highest observed aggregate values across the five-evaluation metrics in the evaluated benchmark. The aggregate retrieval performance of the five evaluated configurations is reported in Table 3.

Table 3. Aggregate retrieval performance across the 20 benchmark queries.

Configuration	P5	R5	F1_5	MRR	nDCG5
M1 - TF-IDF/Cosine	0.460	0.9240	0.5479	1.000	0.9785
M4 - Semantic Proxy	0.460	0.9240	0.5479	1.000	0.9807
M5 - Ontology	0.460	0.9365	0.5511	0.950	0.9163
MSO - Semantic-Ontology	0.470	0.9490	0.5622	1.000	0.9894
M6 - Hybrid	0.470	0.9490	0.5622	1.000	0.9894

Metric aggregation note: the Final Performance artifact reports the 20-query means for all five configurations [18]. The underlying artifact records nDCG@5 = 0.9785183 (M1), 0.9807251 (M4), 0.9163307 (M5), and 0.9894133 for both MSO

and M6; the manuscript reports these values rounded to four decimals.

The per-query artifacts show that F1_5 is first computed per query and then averaged; therefore, the reported F1_5 is not the harmonic mean of the two aggregate P5 and R5 values. The F1 definition is given in Equation (4) [15].

Relative to M1, M6 increases the recorded macro-average P5 by 0.0100, R5 by 0.0250, F1_5 by 0.01429, and nDCG5 by approximately 0.01090, while MRR remains unchanged.

Relative to M5, the largest recorded descriptive difference is observed for macro-average nDCG5, with an increase of approximately 0.07308. The relevant nDCG definition is given in Equation (6) [16].

4.2. Statistical Comparisons

None of the tested pairwise differences reach the predefined 0.05 family-wise significance threshold after Holm correction. The statistical comparison results are reported in **Table 4** and are interpreted using the adjusted p-values rather than the raw p-values.

Table 4. Pairwise statistical comparisons using the Wilcoxon signed-rank test with Holm correction [18].

Comparison	Metric	Mean diff.	p_raw	p_Holm	95% CI	Significant?
M6 vs M1	P5	+0.01000	0.31731	0.95193	[0.00000, 0.03000]	No
M6 vs M1	R5	+0.02500	0.31731	0.95193	[0.00000, 0.07500]	No
M6 vs M1	F1_5	+0.01429	0.31731	0.95193	[0.00000, 0.04286]	No
M6 vs M1	MRR	0.00000	—	—	[0.00000, 0.00000]	No
M6 vs M1	nDCG5	+0.01090	0.27332	0.54664	[−0.00098, 0.03066]	No
M6 vs M5	P5	+0.01000	0.31731	0.95193	[0.00000, 0.03000]	No
M6 vs M5	R5	+0.01250	0.31731	0.95193	[0.00000, 0.03750]	No
M6 vs M5	F1_5	+0.01111	0.31731	0.95193	[0.00000, 0.03333]	No
M6 vs M5	MRR	+0.05000	0.15730	0.15730	[0.00000, 0.12500]	No
M6 vs M5	nDCG5	+0.07308	0.01796	0.05388	[0.02128, 0.13217]	No

The p-values reported in **Table 4** are the values recorded in the experimental results package [18]. In particular, the M6-versus-M5 comparison for nDCG@5 has a recorded raw p-value of 0.01796 and a recorded Holm-adjusted value of 0.05388. The unadjusted value would meet the nominal 0.05 threshold, whereas the recorded adjusted value does not. Accordingly, the manuscript does not claim statistical significance for this comparison. **Table 4** reports the recorded mean differences, raw p-values, Holm-adjusted p-values, and confidence intervals. The available results package is sufficient to verify the direction and magnitude of the aggregate differences, but the available artifacts do not support independent reconstruction of every query-level inferential quantity. The reported p-values are therefore retained as recorded experimental results, and no additional significance claim is introduced beyond the recorded Holm-adjusted values. The relevant

nDCG definition is given in Equation (6) [16].

4.3. Sensitivity Analysis

The sensitivity artifact evaluates a broader grid of coefficient triplets satisfying $\alpha + \beta + \gamma = 1$. For the tested configurations with $\alpha = 0$ and β from 0.05 to 0.30, all five aggregate metrics remain identical to the recorded MSO/M6 values. The broader grid shows that P5, R5, F1_5, and MRR remain unchanged across many tested coefficient combinations, whereas nDCG5 varies outside the $\alpha = 0$ plateau. For $\alpha = 0$ and $\beta = 0.05$ - 0.30, all five aggregate metrics remain identical to the recorded MSO/M6 values, with a maximum recorded nDCG5 of 0.9894. Accordingly, the sensitivity result is better characterized as a benchmark-specific performance plateau within the tested $\alpha = 0$ interval, combined with measurable nDCG sensitivity outside that interval, rather than as a general parameter insensitivity claim. The relevant nDCG definition is given in Equation (6) [16]. The sensitivity analysis results for the tested coefficient combinations are summarized in Table 5, showing an exact five-metric plateau for $\alpha = 0$ and $\beta = 0.05$ - 0.30.

Table 5. Sensitivity of retrieval performance to ontology-dominant semantic weighting.

α	β	γ	P5	R5	F1_5	MRR	nDCG5
0	0.05	0.95	0.470	0.9490	0.5622	1.000	0.9894
0	0.10	0.90	0.470	0.9490	0.5622	1.000	0.9894
0	0.15	0.85	0.470	0.9490	0.5622	1.000	0.9894
0	0.20	0.80	0.470	0.9490	0.5622	1.000	0.9894
0	0.25	0.75	0.470	0.9490	0.5622	1.000	0.9894
0	0.30	0.70	0.470	0.9490	0.5622	1.000	0.9894

Ablation interpretation: the evaluated configurations constitute a comparative component analysis, but not a complete factorial ablation study. The evidence therefore supports statements about the observed configurations only. It does not establish that removing or adding an individual component causes the observed aggregate differences independently of the other configuration changes.

The plateau should not be interpreted as evidence of a globally optimal weighting or universal insensitivity to the semantic coefficient. It is an empirical observation restricted to the tested benchmark and parameter interval.

5. Discussion

The descriptive results are consistent with the possibility that semantic and ontology-based signals provide complementary information in the evaluated VUIS benchmark. The ablation-oriented configuration set is useful because M1, M4, and M5 isolate individual retrieval signals, whereas MSO and M6 evaluate combinations. However, differences among these configurations should not be interpreted as causal component effects without a complete factorial ablation and in-

dependently reproducible query level analysis.

The strongest descriptive difference occurs for nDCG5 when M6 is compared with ontology only retrieval. This is consistent with the possibility that the hybrid configuration improves the ordering of relevant documents even when the change in the number of relevant documents retrieved at the cutoff is modest. The relevant nDCG definition is given in Equation (6) [16].

However, the inferential evidence is not conclusive after Holm correction. The uncorrected nDCG5 p-value of 0.01796 becomes 0.05388 after correction. The appropriate interpretation is not that the hybrid method is statistically superior, but that the benchmark provides a substantial descriptive difference, while the multiplicity-adjusted evidence remains inconclusive at the predefined 0.05 threshold. [18]. The relevant nDCG definition is given in Equation (6) [16].

The sensitivity analysis evaluates coefficient triplets satisfying $\alpha + \beta + \gamma = 1$. For the tested configurations with $\alpha = 0$ and β from 0.05 to 0.30, the recorded artifact shows identical values for all five-aggregate metrics: P5 = 0.470, R5 = 0.9490, F1_5 = 0.5622, MRR = 1.000, and nDCG5 = 0.9894. Across the broader tested grid, P5, R5, F1_5, and MRR remain unchanged while nDCG5 varies with the coefficient weighting. The maximum recorded nDCG5 is 0.9894133 for $\alpha = 0$ and $\beta = 0.05$ - 0.30. Accordingly, the sensitivity result is better characterized as a benchmark-specific performance plateau within the tested $\alpha = 0$ interval, combined with measurable nDCG sensitivity outside that interval, rather than as a general parameter-insensitivity claim [18].

Several limitations constrain external validity. The benchmark contains only 20 queries and 38 documents, and only 70 of the 760 possible query-document pairs have explicit relevance judgments. The ontology covers 16 of the 20 queries. The Semantic Proxy is treated as an experimentally defined semantic signal rather than a specific pretrained language model.

The results should therefore be interpreted as benchmark-specific empirical evidence. The aggregate performance values are directly recorded in the final performance artifact [18], while the protocol manifest confirms the benchmark and evaluation design [17]. Recorded inferential values are retained but are not presented as independently reproduced statistics. The study therefore supports descriptive and suggestive conclusions within the evaluated VUIS setting, not claims of universal superiority, causal component effects, or global parameter optimality.

6. Conclusion

The study evaluated lexical, semantic, ontology-based, and hybrid retrieval configurations in a virtual-university information environment. The final performance artifact records the highest aggregate values for the Semantic-Ontology and Hybrid configurations: P5 = 0.470, R5 = 0.9490, F1_5 = 0.5622, MRR = 1.000, and nDCG5 = 0.9894 [18]. The largest recorded descriptive improvement over ontology-only retrieval is observed for nDCG5 ($\Delta = 0.07308$), but the recorded Holm-adjusted p-value of 0.05388 does not meet the predefined 0.05 threshold.

Sensitivity analysis shows an exact five-metric plateau for $\alpha = 0$ and $\beta = 0.05$ - 0.30, while nDCG5 varies outside that tested plateau. The evidence therefore supports benchmark-specific descriptive and suggestive conclusions, while not supporting claims of universal superiority, causal component effects, or global parameter optimality. The relevant nDCG definition is given in Equation (6) [16].

Author Contributions

M.B., T.K., and Y.K. contributed to the conception, methodological development, experimental evaluation, and preparation of the manuscript. The precise allocation of individual contributions should be confirmed by all authors before submission. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[2] Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R. and Cheng, X. (2022) Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Transactions on Information Systems*, **40**, 1-42. https://doi.org/10.1145/3486250

[3] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. 1*st International Conference on Learning Representations* (*ICLR* 2013), *Workshop Track Proceedings*, Scottsdale, 2-4 May 2013, 1-12.

[4] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume* 1 (*Long and Short Papers*), Minneapolis, 2 June-7 June 2019, 4171-4186. https://doi.org/10.18653/v1/n19-1423

[5] Reimers, N. and Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings Using Siamese Bert-Networks. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, 3-7 November 2019, 3982-3992. https://doi.org/10.18653/v1/d19-1410

[6] Zhao, W.X., Liu, J., Ren, R. and Wen, J. (2024) Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Transactions on Information Systems*, **42**, 1-60. https://doi.org/10.1145/3637870

[7] Gruber, T.R. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**, 199-220. https://doi.org/10.1006/knac.1993.1008

[8] Sharma, A. and Kumar, S. (2023) Ontology-Based Semantic Retrieval of Documents Using Word2vec Model. *Data & Knowledge Engineering*, **144**, Article 102110. https://doi.org/10.1016/j.datak.2022.102110

[9] Sharma, A. and Kumar, S. (2023) Machine Learning and Ontology-Based Novel Se-

mantic Document Indexing for Information Retrieval. *Computers & Industrial Engineering*, **176**, Article 108940. https://doi.org/10.1016/j.cie.2022.108940

[10] Majid, M., Faisal Hayat, M., Zeeshan Khan, F., Ahmad, M., Jhanjhi, N., Arif Sobhan Bhuiyan, M., *et al.* (2021) Ontology-Based System for Educational Program Counseling. *Intelligent Automation & Soft Computing*, **29**, 373-386. https://doi.org/10.32604/iasc.2021.017840

[11] Thanh Sang Nguyen, T., Huu Trong Ho, D. and Tram Anh Nguyen, N. (2023) An Ontology-Based Question Answering System for University Admissions Advising. *Intelligent Automation & Soft Computing*, **36**, 601-616. https://doi.org/10.32604/iasc.2023.032080

[12] Tran, T.-D., Le, V.-T. and Nguyen, T.-N. (2020) An Approach for Semantic-Based Searching in Learning Resources. 2020 12*th International Conference on Knowledge and Systems Engineering* (*KSE*), **12**, 183-188.

[13] Cruz, X.M., Honrado, J.L., Coronel, A., Libatique, N.J. and Tangonan, G. (2023) Design and Development of an Ontology Driven Search Engine for a Mobile Cloud Asynchronous Remote Learning Platform. 2023 *IEEE Global Humanitarian Technology Conference* (*GHTC*), Radnor, 12-15 October 2023, 350-357. https://doi.org/10.1109/ghtc56179.2023.10354815

[14] Hidayah, N.W., Ali Ridho Barakbah, and Iwan Syarif, (2023) Semantic Information Search with Automatic Ontology Creation in Regulations National Standards for Higher Education in Indonesia. *The Indonesian Journal of Computer Science*, **12**, 1172-1185. https://doi.org/10.33022/ijcs.v12i3.3207

[15] Manning, C.D., Raghavan, P. and Schütze, H. (2008) Introduction to Information Retrieval. Cambridge University Press. https://doi.org/10.1017/cbo9780511809071

[16] Järvelin, K. and Kekäläinen, J. (2002) Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, **20**, 422-446. https://doi.org/10.1145/582415.582418

[17] Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**, 80-83. https://doi.org/10.2307/3001968

[18] Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.