Enabling Roll-up and Drill-down Operations in News Exploration with Knowledge Graphs for Due Diligence and Risk Management

1st Sha Wang
Singapore Management University
sha.wang.2021@phdcs.smu.edu.sg

2nd Yuchen Li
Singapore Management University
yuchenli@smu.edu.sg

3rd Hanhua Xiao
Singapore Management University
hhxiao.2020@phdcs.smu.edu.sg

4th Zhifeng Bao
RMIT University
zhifeng.bao@rmit.edu.au

5th Lambert Deng
DBS Bank Limited
lambertdeng@dbs.com

6th Yanfei Dong
PayPal
dyanfei@paypal.com

***Abstract*—Efficient news exploration plays a critical role in real-world applications, particularly within the financial sector, where numerous control and risk assessment tasks rely on the analysis of public news reports. The current processes in this domain predominantly rely on manual efforts, often involving keyword-based searches and the compilation of extensive keyword lists. In this paper, we introduce NCEXPLORER, a framework designed with OLAP-like operations to enhance news exploration experience. NCEXPLORER offers users the ability to use *roll-up* operations for a broader content overview and *drill-down* operations for detailed insights. These operations are achieved through integration with external knowledge graphs (KGs), encompassing both fact-based and ontology-based structures. This integration significantly augments exploration capabilities, offering a more comprehensive and efficient approach to unveiling the underlying structures and nuances embedded within news content. Empirical studies through crowd-sourced evaluators on Amazon Mechanical Turk demonstrates NCEXPLORER's superiority over existing state-of-the-art news search methodologies across an array of topic domains, using real-world news datasets.**

I. INTRODUCTION

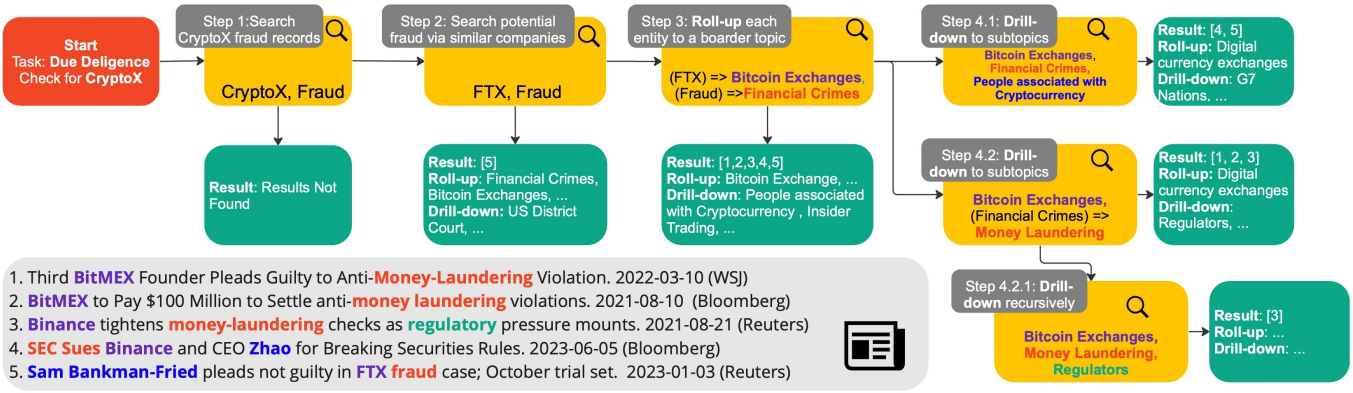
A constant element within every payment transaction is the presence of payment risk. For instance, global payment giant PayPal conducts Anti-Money Laundering (AML) and Counter-Terrorist Financing (CTF) risk assessments, as well as sanctions screening, for every new business clients [1]. Similarly, DBS, Southeast Asia's largest bank, shoulders additional responsibilities in the realm of Environmental, Social, and Governance (ESG) concerns [2]. DBS's ESG risk policy explicitly prohibits financial involvement in activities associated with illegal logging, forced or child labor, wildlife trading, and more. During risk assessments, analysts at these institutions rely on public news reports to determine whether the entities under scrutiny have ever been associated with suspicious activities. Due to the complex nature of these due diligence tasks, they are predominantly manual in nature. Compliance teams laboriously maintain extensive lists of financial crime terminology and sift through search results to distinguish genuine financial misconduct from unrelated or benign news. This creates a big operational overhead of running a business. According to a recent McKinsey survey[1], the expenditures related to compliance tasks in major banks have skyrocketed over the past decade, reaching an unsustainable crescendo. The situation becomes even more precarious for DNFBPs (Designated Non-Financial Businesses and Professions), such as precious metal dealers or real estate agents. While these entities may operate on a smaller scale, they are equally obligated to conduct rigorous checks for significant transactions, often with limited resources. The need for a more efficient approach to news exploration in this space has never been more evident.

In our research, we present an OLAP-inspired approach to news exploration. In the *roll-up* process, users input known terms, leading to the generation of broader topics. For example, "*FTX*" is expanded to "*Bitcoin Exchange*". Our system, NCEXPLORER, subsequently amplifies these topics by curating a list of relvant keywords for retrieval. The retrieved articles come with an array of related subtopics, granting users the capability to *drill-down* into specific news pieces and discover unanticipated topics.

To illustrate the enhanced productivity gained through our approach in the context of due diligence checks, consider a Know Your Customer (KYC) task involving a newly incorporated cryptocurrency exchange, "*CryptoX*", as depicted in Fig 1. Here, the exchange is seeking to open a business bank account in a jurisdiction that has recently passed regulations permitting digital currency companies. The KYC analyst initiates the process by querying "*CryptoX fraud*" but receives no results. Subsequently, with the understanding that "*CryptoX*" has a clean slate, the analyst shifts to focus on peer-related checks using queries such as "*FTX*" and "*Fraud*". This approach yields some results alongside *roll-up* options. Expand-

[1] https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-compliance-function-at-an-inflection-point

Fig. 1: NCEXPLORER ***roll-up*** and ***drill-down*** example



ing the search to industry-wide topics like "*Bitcoin Exchange*" and "*Financial Crime*" produces a more comprehensive set of results, each linked to entities relevant to the chosen topics (highlighted in color). Armed with this information, the expert delves deeper into understanding the prevalent fraud types in the crypto realm and regulatory implications. Throughout this journey, the analyst enjoys the leeway to alternate between roll-up and drill-down modes, mirroring the flexibility of navigating an OLAP cube. In comparison, traditional due diligence would demand the painstaking manual tweaking of keywords and in-depth examination of search outputs to discern interconnected entities and patterns.

Our proposed OLAP operations for news exploration are versatile, with potential in other novel applications. For instance, NCEXPLORER can detect media bias. When Elon Musk acquired Twitter, it stirred debates on wealthy individuals influencing media [3]. Using NCEXPLORER, users can expand from "*Elon Musk*" to unveil parallels like Jeff Bezos's acquisition of the Washington Post [4], Patrick Soon-Shiong's purchase of the Los Angeles Times [5], and Rupert Murdoch's buyout of The Wall Street Journal [6]. While articles on Musk leaned negative, others retained a neutral or positive bent. Such disparities underscore NCEXPLORER's prowess in discerning biases and news narratives.

This OLAP-like interaction is made possible through the integration of the ontology network of external Knowledge Graphs (KGs) [7], [8]. As illustrated in Fig 2, the KG encompasses not only millions of entities and relationships, forming a comprehensive fact network, but also contains extensive ontology information. We summarize our technical contributions as follows:

- We introduce a novel framework, NCEXPLORER, which is the first of its kind, supporting news exploration with semantic *roll-up* and *drill-down* operations by leveraging the ontology and fact networks of knowledge graphs (KGs).
- We develop effective ranking schemes to evaluate concept-article relevance, enabling smooth *roll-up* and *drill-down* operations. Additionally, we devise an efficient, unbiased sampling estimator for relevance score computation.
- We assessed NCEXPLORER using extensive news datasets, garnering feedback from Amazon Mechanical Turk's master-qualified participants. With over 3,900 evaluations, the findings affirm NCEXPLORER's superiority over leading news search techniques. Even after integrating GPT models into all compared baselines, the results remained consistent.
- We also release our implementations, datasets, and evaluation results at [2]. Our dataset contains 200k news articles, with 2.9 million entity and 3.7 million concept annotations linked to DBPedia [9]. Compared to existing news datasets, the inclusion of entity and concept annotations enables deeper analysis of news articles, leveraging the interconnectedness of the knowledge graphs.

II. RELATED WORK

Our work is broadly related to news search and recommendation, aiming to enhance user experience in these domains. Existing works in this area typically extend generic natural language understanding (NLU) models [10]–[12] to improve search and recommendation experiences for news articles. These works generally fall into two categories: (a) embedding-based approaches and (b) structure-based approaches.

Embedding-based approaches: These methods leverage embeddings to create document representations that capture the semantic meaning of news articles. Notable works in this category include DKN [13], KRED [14], and NewsGraph [15]. DKN integrates the embeddings of news titles, derived from a CNN approach [16], with the embeddings of news KG entities obtained through a knowledge graph embedding technique [17]. KRED extends this idea further by introducing a context embedding layer generated from a news entity's neighbors, as well as an attentive merging layer inspired by the Knowledge Graph Attention Network (KGAT). NewsGraph, on the other hand, builds a separate graph by removing irrelevant edges and introducing new edges that represent co-occurrences in the same news or visits from the same user. This alteration hydrates the static KG with dynamic interactions between entities in real life, making the latent vectors aware of current affairs.

[2] https://github.com/knowledge-fusion/ncexplorer

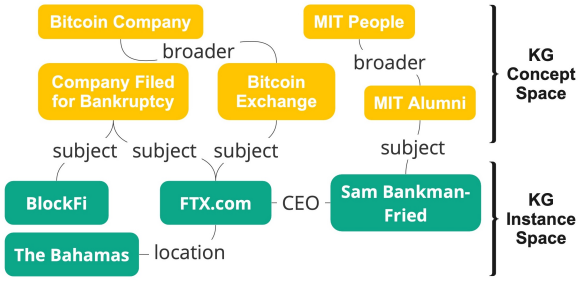


Fig. 2: KG Concept and Instance Spaces

Structure-based approaches: Our work is more related to the methods that focus on utilizing the KG structures to provide better explainability and semantic relevance between news articles. AnchorKG [18] and NewsLink [19] are two prominent works in this category. AnchorKG represents each news article as a compact subgraph containing essential news entities and their k-hop neighbors, generated using a reinforced learning framework. This interaction of anchor graphs provides explanations for semantic relevance between articles. NewsLink, a state-of-the-art news search approach, uses seed nodes identified from text fragments and a graph expansion algorithm to connect nodes in a single graph, adding hidden related nodes as auxiliary information for news semantic representation. To the best knowledge, our work is the first to facilitate systematic exploration with OLAP-like operations.

III. NCEXPLORER

NCEXPLORER leverages external KGs for news exploration. A KG is a multigraph $\mathcal{G} = (\mathcal{V}_C \bigcup \mathcal{V}_I, \mathcal{E}_C \bigcup \mathcal{E}_I, \Psi)$. $\mathcal{V}_C$ and $\mathcal{V}_I$ represent the concept entities and the instance entities respectively, represented by yellow and green nodes in Fig 2 respectively. Each concept edge in $\mathcal{E}_C$ links two concept entities, and an instance edge in $\mathcal{E}_I$ links two instance entities. Like [19], we add a reversed edge for each original edge so that $\mathcal{G}$ is bidirected. The associations between the instance space and the concept space are captured by the ontology relation Ψ. $\Psi(c)$ maps a concept entity $c \in \mathcal{V}_C$ to a set of instance entities while $\Psi^{-1}(v)$ maps an instance entity $v \in \mathcal{V}_I$ to a set of concept entities.

There are two major components in the architecture of NCEXPLORER: semantic *roll-up*, and *drill-down*:

Roll-up Operation. NCEXPLORER generates a list of entities from a news document d, users then replace one or more entities with KG concepts to form a concept pattern query Q. Given such a Q, a document d matches Q if for each concept $c \in Q$, there is an entity v in d that $v \in \Psi(c)$. In other words, we can find a set of entities in d that match Q. The relevance score of a matched document d to Q is the sum of the relevance scores among all concepts in Q, i.e.,

$$rel(Q, d) = \sum_{c \in Q} cdr(c, d). \tag{1}$$

NCEXPLORER returns all matched documents to the query Q, ranked by the relevance scores $rel(Q, \cdot)$. Furthermore, the entities that match the concepts in Q are annotated to explain why a news document is retrieved.

Drill-down Operation. Based on the matched news from the *roll-up* operation, NCEXPLORER suggests additional concepts as subtopics to a query Q. A suggested concept c', as a subtopic of Q, enables users to conduct a *drill-down* analysis. By selecting c', users narrow down the matched news to the augmented query $Q \cup c'$. Since a subtopic must appear as a concept in at least one matched document of Q, we can find all candidate subtopics by unionizing the concepts from all matched documents.

A. Concept Document Rank

We propose a novel ranking scheme for $cdr(c, d)$ by considering two key relevance dimensions: *ontology relevance* $cdr_o(c, d)$ and *context relevance* $cdr_c(c, d)$. A concept c is relevant to a document d if c is relevant in both dimensions:

$$cdr(c, d) = cdr_o(c, d) \cdot cdr_c(c, d) \tag{2}$$

Ontology Relevance. When a document d contains an entity v that matches a concept c under the ontology relation Ψ, i.e., $v \in \Psi(c)$, c is associated with d by ontology relevance. Since there could be more than one entity in d that matches c, we define the ontology relevance score function as follows:

$$cdr_o(c, d) = log \frac{|\mathcal{V}_I|}{|\Psi(c)|} \cdot \left(\max_{v \in ME(c,d)} tw(v, d) \right) \tag{3}$$

where $ME(c, d) = \{v |\ v \in d \text{ and } v \in \Psi(c)\}$. First, a concept that matches more entities in the ontology, i.e., lower specificity, should be less relevant to a document. Second, among all the matched entities, a *pivot* entity is selected as the one with the highest term weight $tw(v, d)$ in the document to match the concept. The term weight is used to capture the importance of v in d. If v plays a more significant role in d, then c is more relevant to d. We use the typical TF-IDF scheme for the term weight in our implementation and other schemes can be easily supported as well.

Context Relevance. An entity in a document could map to different concepts by ontology. Hence, ontology relevance can only differentiate the matched concepts of the same entity with the specificity score $\log(|\mathcal{V}_I|/|\Psi(c)|)$, which simply penalizes broad concepts. We thus propose to include the unmatched entities as contextual information to improve the relevance semantics. To measure the context relevance of c and d, we compute the relevance of c and the context entities $CE(c, d) = \{v |\ v \in d \text{ and } v \notin \Psi(c)\}$. Although the context entities do not match c, we take advantage of both the ontology relation and the instance space to link c with the context entities. We introduce a novel connectivity score $conn(c, d)$ to measure the KG connectivity between a concept c and a context entity set $CE(c, d)$ as follows:

$$conn(c, d) = \sum_{v \in CE(c,d)} \frac{\sum_{u \in \Psi(c)} \sum_{l=1}^{\tau} \beta^l \cdot |paths_{u,v}^{<l>}|}{|CE(c, d)|} \tag{4}$$

where $|paths_{u,v}^{<l>}|$ measures the number of l-hop simple paths connecting u and v in the instance space, and β is a damping factor that penalizes longer paths. Intuitively, the connectivity score is the average number of paths among all context entities connected to any KG instance entity that matches c, subject to a hop constraint of at most τ. Thus, better connectivity leads to a higher relevance score. Finally, we normalize the connectivity score to $[0, 1)$ as follows:

$$cdr_c(c, d) = 1 - \frac{1}{1 + conn(c, d)} \tag{5}$$

B. Connectivity Score Estimation

The connectivity score can be computationally expensive due to the massive number of path enumerations for each pair of entities and existing studies on s-t path enumeration can only produce *polynomial delay* algorithms [20], [21]. Inspired by the sampling approaches for online join aggregation queries [22]–[24], we devise a single random walk estimator (algorithm in appendix). To start a random walk, we first sample a source entity u from all entities that map to the concept c, and a target entity v from the corresponding context entities. A *non-repeating* random walk $r(u, v)$ from u tries to reach v samples via at most τ distinct nodes. Let u_i denote the ith node sampled by a random walk starting from the source u and $N(u_i)$ is the number of eligible neighbors of u_i to be sampled, we define the following sample estimator:

$$r(u, v) = \mathcal{I}(u, v) \cdot |\Psi(c)| \cdot \beta^{l-1} \cdot \prod_{i=1}^{l-1} N(u_i) \tag{6}$$

where $\mathcal{I}(u, v)$ is an indicator random walk: it returns 1 if u reaches v at the l-th sampled node for any $l \leq \tau$; otherwise, it returns 0.

The sampling process may suffer from slow convergence, especially when many of the sampled paths cannot reach the target context entity v. To enable faster convergence, we build a reachability index [25] on the KG instance space and only sample eligible neighbors that satisfy the hop constraint. The following theorem ensures our sample approach is unbiased.

Theorem 1: $r(\cdot, \cdot)$ is an unbiased estimator to $conn(c, d)$, i.e., $\mathbb{E}[r(\cdot, \cdot)] = conn(c, d)$ where the randomness is taken over the source u and target v as well as the random walk from u to v.

Proof Sketch. $conn(c, d)$ is a weighted sum of all the paths connecting an entity node $u \in \Psi(c)$ and a context entity $v \in CE(c, d)$ subject to a hop constraint of τ. For each path $s = \{u_1 = u, u_2, u_3, ..., u_l = v\}$ that connects u and v, the probability of s being sampled is $\mathbb{P}(s) = (|\Psi(c)| \cdot |CE(c, d)| \cdot \prod_i^{l-1} N(s_i))^{-1}$. Thus, we can use the Horvitz–Thompson estimator [26] to approximate the population sum without bias.

C. Drill-down Exploration Rank

NCExplorer can retrieve a significant amount of relevant documents for a concept query Q using the roll-up operation. To aid users in navigating the results, NCExplorer identifies related concepts in the retrieved documents and suggests them as subtopics for users to conduct *drill-down* analysis. Let $\mathcal{D}(Q)$ denote the set of retrieved documents for Q and a candidate subtopic is a concept c such that there is an instance entity $v \in \Psi(c)$ appearing in one of the documents $d \in \mathcal{D}(Q)$. We develop a score $sbr(\cdot, Q)$ over candidate subtopics and only suggest top-scored subtopics for Q. There are three key perspectives considered in $sbr(\cdot, Q)$ as below:

$$sbr(c, Q) = coverage(c, Q) \cdot specificity(c) \cdot diversity(c, Q) \tag{7}$$

- **Coverage.** The coverage score of a subtopic c is calculated as the sum of the relevance scores $cdr(c, d)$ for all documents d in $\mathcal{D}(Q)$. This ensures that only subtopics that are highly relevant to a large number of documents are suggested to the user for further exploration: $coverage(c, Q) = \sum_{d \in \mathcal{D}(Q)} cdr(c, d)$.
- **Specificity.** To avoid suggesting trivial subtopics like "*Person*" which may match a large number of entities, we prioritize concepts with higher specificity scores using the formula $specificity(c) = \log(|\mathcal{V}_I|/|\Psi(c)|)$.
- **Diversity.** The diversity score calculates the average number of distinct entities mapped to the concept c among all the documents relevant to $Q \cup \{c\}$: $diversity(c, Q) = \frac{|\bigcup_{d \in \mathcal{D}(Q)} ME(c,d)|}{|\mathcal{D}(Q \cup \{c\})|}$. This helps ensure fairness in the suggested concepts and prevents results from being biased toward concepts that match a small set of popular entities.

IV. Evaluation

Datasets. We use the June 2021 snapshot of DBPedia [9] as our backend KG. We crawl 200k articles from popular news portals: *Reuters* [27], *SeekingAlpha* [28] and *The New York Times* [29] to have a mixture of business and politics reports. Detailed statistics of released dataset are shown in table below.

News Source	Articles	Total Entities	Linked Entities
Seekingalpha	6823	97k	62k (63.9%)
NYT	3625	51k	35k (68.6%)
Reuters	171662	4539k	2336k (51%)

Compared Methods.

- Lucene implements a typical bag-of-words keyword match model. We use BM25 [30] for the term weighting scheme with the default library settings.
- Bert [10] is a popular neural text embedding model. We use a SBERT [31], a modification of the pre-trained Bert to map each news article to a vector of 768 dimensions.
- Newslink [19] is the state-of-the-art implicit news exploration method that expands a news document and a query by forming a common ancestor graph extracted from the KG. Each KG entity in the extracted graph is then treated as a matching keyword in the bag-of-words model.
- Newslink-Bert is a hybrid method that combines Newslink and Bert. It expands query entities into a subgraph using Newslink's algorithm and concatenates them to form a long text query.
- NCExplorer is our proposed approach. The parameters are set to $\tau = 2$ and $\beta = 0.5$ by default. The number of samples for connectivity score estimation is set to 50.

Implementation. NCEXPLORER and NEWSLINK are implemented in Python 3.9. BERT and NEWSLINK-BERT use Qdrant [32] as vector search engine. We use a server running on Ubuntu 20.04 with an AMD EPYC 7643 Processor @ 3.45GHz and 251G RAM.

A. Concept Document Relevance Study

As shown in Table I, we evaluate a total of six topics. Each topic is combined with either an entity group (can be a list of countries or companies) to form queries such as "*Elections in African countries*" or "*Lawsuits involving U.S. technology companies*". For each query, the top 5 news articles are retrieved from each method, resulting in 25 outcomes. These results are presented to each evaluator in a randomized order. To ensure a fair comparison, we hide NCEXPLORER's result explanations. The relevance level is rated for each concept in the query, with values ranging between 0 and 5. In total, we obtain 3,900 ratings from 78 evaluators. We use **NDCG@K** to evaluate the effectiveness of relevance ranking. We also feed each method's top-5 results to GPT to test whether performance be be further boosted. We ask GPT-3.5-turbo to evaluate the relevance between a topic and a news article with following prompt:

> < news article >. Is this article related to <topic>. please give a rating between 0.000 and 5.000, with 5.000 being most relevant. only give three decimal digit number.

Each method's top-K is sorted using GPT's ratinng and the new **NDCG@K** scores are displayed on right-hand side. The evaluation result for both scenario shows NCEXPLORER achieve the best or second-best performance in nearly all cases, except for the query "*Merger & Acquisition, U.S. biotechnology companies*". A detailed analysis reveals that evaluators show greater confidence in commonly known surface words like "*M&A*", "*acquisition*", "*buy*", and "*sell*", while expressing uncertainty about specialized terms such as "*takeover*". This finding highlights the effectiveness of NCEXPLORER's roll-up operation in helping analysts collect news articles featuring more domain-specific vocabularies. Embedding models also demonstrate remarkable performance in news exploration tasks, particularly when combined with KG information. However, there are two issues that arise when using pure embedding models. First, implicit matching may retrieve reports like daily trade price/volume[3], which, although crucial for real-time decision-making, have limited utility in exploration tasks. Second, the issue emerges when the concept entity isn't frequently mentioned alongside instance entities. For example, "*labor dispute*" are often reported in the news as strikes organized by various labor unions. The embedding of "*labor dispute*" alone cannot yield high-quality results. This issue can be partially mitigated with supplementary context provided by Newslink's subgraph KG embedding. For NEWSLINK, the performance is not stable due to the infrequent formation of densely connected single components by the subgraph embedding of multiple concept entities. Instead of identifying hidden nodes that connect existing query entities, the subgraph often results in a single concept entity accompanied by its N-hop neighbors. This dilutes the significance of concepts with smaller connected components, affecting the overall effectiveness of the news exploration process.

Table II shows the impact of GPT rerank. While performance is boosted for 84.4% of the measurement metrics, overall rank across different methods remain the same. One observation is that the impact is positive for all methods except LUCENE. Another interesting observation is the impact for NDCG@1 > NDCG@5 > NDCG@10, suggesting GPT can distinguish the subtle differences among top results.

B. Efficiency Study

Indexing Efficiency. NCEXPLORER processes each news document and constructs an index for query processing. To evaluate the indexing overhead, we select 100 articles from each news portal and report the average processing time in Fig. 3. LUCENE and BERT show sub-second execution time. NEWSLINK, NEWSLINK-BERT, and NCEXPLORER cost 2-3 seconds for each article. A breakdown analysis of the total cost showed that the top two overheads come from entity linking (91.8%) and relevance score calculation (7.1%). Entity linking is a common cost to all methods that require KG analysis. Improving the efficiency of entity linking is beyond the scope of this work. The relevance score calculation only takes 7.1% of the overall index time due to our efficient estimation approach via sampling. Furthermore, the overhead is manageable as the indexing cost incurs only once per document and the process can be easily accelerated with multi-threading or distributed approaches.

Retrieval Efficiency. Fig 4 displays the query efficiency of the compared methods when increasing the number of concepts while keeping corpus size fixed. The processing time for each data point is the average across 100 queries. LUCENE is a highly optimized system for text retrieval and takes the least amount of time. BERT used to take longer execution time due to the absence of index. Recent development on vector databases has greatly sped up embedding retrieval and result in LUCENE compatible speed. The performance for NCEXPLORER is similar to NEWSLINK where the duration correlates to number of KG entities in the query. NEWSLINK-BERT takes the sum of BERT and NEWSLINK. Overall, NCEXPLORER can answer a concept pattern query with reasonable overhead.

C. Context Relevance Score

Effectiveness and Parameter Study. To verify the effectiveness of the context relevance score $cdr_c(c, d)$ (Eqn. 4), we design a "negative sampling" approach. We randomly select 100 entries from the inverted index $\langle c, d\rangle$. For each entry $\langle c, d\rangle$, we sample a concept node c' from the KG to generate a "*negative*" concept. Fig. 5 demonstrates the effective differentiation between c and c' using the context relevance score, i.e., $cdc_c(c, d) > cdc_c(c', d)$, regardless of the hop constraint τ from 1 to 3. Notably, when τ is set to 1 and 2, the score differences are significant compared with $\tau = 3$,

[3]https://www.bloomberg.com/markets/stocks/futures

TABLE I: NCDG@K without/with GPT rerank. The best results are boldfaced and the second-best results are underlined.

	NDCG@1	NDCG@5	NDCG@10	NDCG@1	NDCG@5	NDCG@10
	wo/w GPT rerank	wo/w GPT rerank	wo/w GPT rerank	wo/w GPT rerank	wo/w GPT rerank	wo/w GPT rerank
Topic	**International Trade**			**Lawsuits**		
Lucene	0.688 / 0.572	0.557 / 0.532	0.737 / 0.720	0.574 / 0.571	0.593 / 0.597	0.763 / 0.766
BERT	0.856 / 0.856	0.882 / 0.882	0.951 / 0.951	**0.849** / 0.849	0.848 / 0.900	0.935 / 0.949
NewsLink	0.765 / 0.817	0.623 / 0.650	0.781 / 0.799	0.329 / 0.628	0.406 / 0.475	0.636 / 0.683
NewsLink-BERT	0.825 / 0.836	0.877 / 0.878	0.949 / 0.949	0.627 / 0.658	0.796 / 0.848	0.883 / 0.898
NCEXPLORER	**0.974 / 0.974**	**0.957 / 0.956**	**0.987 / 0.986**	0.844 / **1.000**	**0.919 / 0.948**	**0.959 / 0.979**
Topic	**Elections**			**Mergers & Acquisitions**		
Lucene	0.550 / 0.273	0.455 / 0.378	0.653 / 0.603	0.464 / 0.659	0.593 / 0.629	0.777 / 0.802
BERT	0.887 / 0.910	0.894 / 0.903	0.941 / 0.947	**0.728 / 1.000**	0.820 / **0.879**	0.915 / 0.955
NewsLink	0.554 / 0.554	0.450 / 0.466	0.649 / 0.660	0.305 / 0.493	0.445 / 0.471	0.675 / 0.693
NewsLink-BERT	**0.946** / 0.946	**0.972 / 0.972**	**0.990 / 0.990**	0.724 / 0.912	0.803 / 0.829	0.913 / 0.930
NCEXPLORER	0.924 / **0.947**	0.958 / 0.966	0.978 / 0.984	0.712 / 0.846	**0.843** / 0.870	**0.937 / 0.956**
Topic	**International Relations**			**Labor Dispute**		
Lucene	0.896 / 0.650	0.722 / 0.670	0.830 / 0.795	0.564 / 0.621	0.618 / 0.632	0.817 / 0.826
BERT	0.921 / 0.957	0.922 / 0.941	0.959 / 0.970	0.370 / 0.370	0.411 / 0.410	0.670 / 0.669
NewsLink	0.735 / 0.804	0.729 / 0.760	0.834 / 0.854	0.481 / 0.729	0.476 / 0.499	0.716 / 0.733
NewsLink-BERT	0.867 / 0.945	0.943 / 0.954	0.971 / 0.982	0.695 / 0.905	0.720 / 0.767	0.889 / 0.922
NCEXPLORER	**0.927 / 0.963**	**0.970 / 0.974**	**0.986 / 0.989**	**0.922 / 0.931**	**0.984 / 0.988**	**0.989 / 0.991**

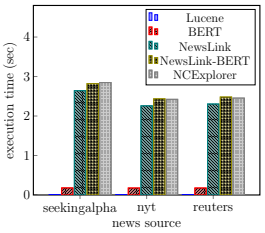


Fig. 3: Performance Study: indexing time

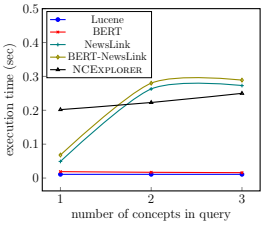


Fig. 4: Performance Study: retrieval time

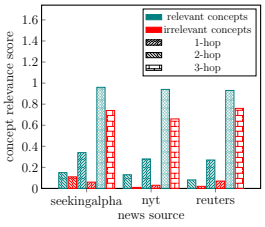


Fig. 5: Effectiveness study: context relevance

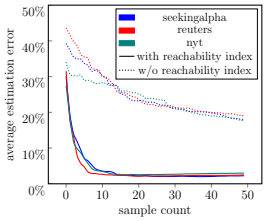


Fig. 6: Sampling error with increasing samples

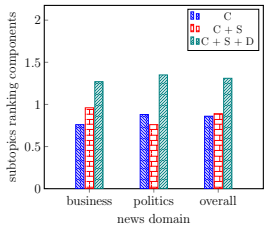


Fig. 7: Subtopics ablation study

TABLE II: Impact of GPT rerank

	NDCG@1	NDCG@5	NDCG@10	Overall
Lucene	-0.065	-0.016	-0.011	-0.031
BERT	0.055	0.023	0.022	0.030
NewsLink	0.143	0.032	0.022	0.066
NewsLink-BERT	0.055	0.018	0.010	0.028
NCEXPLORER	0.060	0.012	0.008	0.027

suggesting that a large τ may include irrelevant concepts due to a higher chance of linking concepts with documents. Our default τ is set to 2, as over half (55%) of the relevant scores are 0 when $\tau = 1$. In contrast, only 22.4% of the relevance scores are 0 for $\tau = 2$, striking a good balance between information linking and relevance differentiation.

RW Estimator Convergence. We evaluate the convergence rate of the proposed RW estimator on the connectivity score. Fig. 6 shows the average sampling error of $cdr_c(c, d)$ compared to the ground truth value. Solid and dotted lines represent RW *with* and *without* the guidance of k-hop reachability index respectively. With the k-hop index, our sampling approach can converge on all three datasets within 5% estimation error using 20 sampling iterations.

D. Ablation Study on Drill-down Operations

Given a search query, NCEXPLORER automatically suggests related concepts for *drill-down* operations. NCEXPLORER ranks the concepts by considering three key factors: *Coverage (C)*, *Specificity (S)* and *Diversity (D)*. To investigate the impact of each factor, we conduct user studies for the ablation analysis. We use the same queries from Sec. IV-A and select top-ranked concepts when only considering: (1) C; (2) C+S; (3) C+S+D. We build an interactive survey interface (listed in appendix) that allows the participants to click on different concepts and view associated results before assigning a distinct rating 1-3 for each subtopic. We recruited participants from the same crowd-source platform AMT [33] and obtain 518 survey results in total. The results are displayed in Fig. 7. We can observe that the specificity contributes slight improvement to the overall rating while diversity plays a more significant role in rating improvement.

V. CONCLUSION

NCEXPLORER is a useful tool designed to enhance the news article exploration experience by using OLAP-like operations to connect them with relevant concepts in a knowledge graph. Its ranking system considers both concept relevance and article context, aiding in a comprehensive understanding of the information. NCEXPLORER not only streamlines due diligence tasks but also improves a range of news analytics tasks. Its efficacy has been validated through crowd-sourced evaluations with a dataset of real-world news articles and a large knowledge graph. Its modular design ensures compatibility with various text-based systems, such as search engines and literature databases, facilitating information discovery and comprehension. Additionally, the release of 200,000 news articles, along with their KG annotations and concept relevance scores, allows for easy customization of the tool to meet specific research needs.

References

[1] Paypal: Anti-money laundering and know your customer. Accessed: 2023-10-01. [Online]. Available: https://publicpolicy.paypal-corp.com/issues/anti-money-laundering-know-your-customer

[2] (2021, Apr.) Dbs bank: Our approach to responsible financing. Accessed: 2023-10-01. [Online]. Available: https://www.dbs.com/iwov-resources/images/sustainability/responsible-banking/DBS%20Bank_Our%20Approach%20to%20Responsible%20Financing_Updated_26Apr2021.pdf

[3] G. Bensinger. (2022) Twitter under elon musk will be a scary place. Accessed: 2023-05-22. [Online]. Available: https://www.nytimes.com/2022/04/25/opinion/editorials/twitter-elon-musk.html

[4] E. Bell. (2013) Jeff bezos' shocking washington post buy was not a business deal — it was a cultural statement. Accessed: 2023-05-02. [Online]. Available: https://twitter.com/BusinessInsider/status/364551288845254656

[5] J. R. K. Meg James. (2018) Billionaire patrick soon-shiong reaches deal to buy l.a. times and san diego union-tribune. Accessed: 2023-05-02. [Online]. Available: https://www.latimes.com/business/hollywood/la-fi-ct-los-angeles-times-sold-20180207-story.html

[6] M. K. Sarah Ellison. (2007) Murdoch wins his bid for dow jones. Accessed: 2023-05-02. [Online]. Available: https://www.wsj.com/articles/SB118589043953483378

[7] T. P. Tanon, G. Weikum, and F. Suchanek, "Yago 4: A reason-able knowledge base," in *ESWC*, 2020, pp. 583–596.

[8] e. a. Lehmann, "Dbpedia, a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, 2015.

[9] Dbpedia snapshot 2021-06 release. Accessed: 2023-02-22. [Online]. Available: https://www.dbpedia.org/blog/snapshot-2021-06-release/

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[12] H.-S. Sheu and S. Li, "Context-aware graph embedding for session-based news recommendation," in *RecSys*, 2020, pp. 657–662.

[13] H. Wang, F. Zhang, X. Xie, and M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1835–1844.

[14] D. Liu, J. Lian, S. Wang, Y. Qiao, J.-H. Chen, G. Sun, and X. Xie, "Kred: Knowledge-aware document representation for news recommendations," in *Proceedings of the 14th ACM Conference on Recommender Systems*, 2020, pp. 200–209.

[15] D. Liu, T. Bai, J. Lian, X. Zhao, G. Sun, J.-R. Wen, and X. Xie, "News graph: An enhanced knowledge graph for news recommendation." in *KaRS@ CIKM*, 2019, pp. 1–7.

[16] J. Kim, A.-D. Nguyen, and S. Lee, "Deep cnn-based blind image quality predictor," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 1, pp. 11–24, 2018.

[17] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 2015, pp. 687–696.

[18] D. Liu, J. Lian, Z. Liu, X. Wang, G. Sun, and X. Xie, "Reinforced anchor knowledge graph generation for news recommendation reasoning," in *KDD*. ACM, 2021, pp. 1055–1065.

[19] Y. Yang, Y. Li, and A. K. Tung, "Newslink: Empowering intuitive news search with knowledge graphs," in *ICDE*, 2021, pp. 876–887.

[20] L. Qin, Y. Peng, Y. Zhang, X. Lin, W. Zhang, and J. Zhou, "Towards bridging theory and practice: hop-constrained st simple path enumeration," in *International Conference on Very Large Data Bases*. VLDB Endowment, 2019.

[21] S. Sun, Y. Chen, B. He, and B. Hooi, "Pathenum: Towards real-time hop-constrained st path enumeration," in *SIGMOD*, 2021, pp. 1758–1770.

[22] F. Li, B. Wu, K. Yi, and Z. Zhao, "Wander join: Online aggregation via random walks," in *SIGMOD*, 2016, pp. 615–629.

[23] Z. Zhao, R. Christensen, F. Li, X. Hu, and K. Yi, "Random sampling over joins revisited," in *SIGMOD*, 2018, pp. 1525–1539.

[24] Y. Park, S. Ko, S. S. Bhowmick, K. Kim, K. Hong, and W.-S. Han, "G-care: a framework for performance benchmarking of cardinality estimation techniques for subgraph matching," in *SIGMOD*, 2020, pp. 1099–1114.

[25] J. Cheng, Z. Shang, H. Cheng, H. Wang, and J. X. Yu, "Efficient processing of k-hop reachability queries," *The VLDB journal*, vol. 23, no. 2, pp. 227–252, 2014.

[26] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, p. 663–685, 1952.

[27] Reuters. Accessed: 2023-02-22. [Online]. Available: https://www.reuters.com

[28] Seeking alpha. Accessed: 2023-02-22. [Online]. Available: https://seekingalpha.com

[29] The new york times. Accessed: 2023-02-22. [Online]. Available: https://www.nytimes.com

[30] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[31] Sentence transformers all-mpnet-base-v2. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[32] "Vector search engine." [Online]. Available: https://qdrant.tech/

[33] I. Amazon Mechanical Turk. Amazon mechanical turk. [Online]. Available: https://www.mturk.com

APPENDIX

A. Connectivity Score Calculation

To calculate connectivity score, our sampling approach with the k-hop reachability index in Algorithm 1. For each random walk, we first sample the source u_1 and the target v. Subsequently, we iteratively sample the next node u_{l+1} by scanning the neighbors of u_l. With the help of the k-hop index, we first check if a neighbor n can reach v within the hop constraint of $\tau - l$ (Line 9) Then, we uniformly sample eligible neighbors by a Reservoir sampler, which only requires a single scan of the neighbors (Line 11) A random walk terminates when either the target v is sampled or the hop limit is reached.

Algorithm 1: Random walk estimator with k-hop index.

Input : KG $\mathcal{G}$, concept c, document d, sample size θ
Output: the estimated connectivity est

1 $est \leftarrow 0$;
2 **while** *there are less than* θ *random walks* **do**
3 $l \leftarrow 1$; $p \leftarrow 1.0$;
4 $u_l \leftarrow$ a random entity in $\Psi(c)$;
5 $v \leftarrow$ a random entity in $CE(c, d)$;
6 **while** $l \leq \tau$ **do**
7 $count \leftarrow 0$;
8 **foreach** $n \in N(u_l)$ **do**
9 **if** $hop(n, v) \leq \tau - l$ **then**
10 $count \leftarrow count + 1$;
11 $u_{l+1} \leftarrow n$ with probability $\frac{1}{count}$;
12 $p \leftarrow \frac{p}{count}$;
13 $l \leftarrow l + 1$;
14 **break** on $u_l = v$;
15 **if** $u_l = v$ **then**
16 $est \leftarrow est + \frac{\beta^{l-1}}{p}$;
17 **return** $\frac{est \cdot |\Psi(c)|}{\theta}$.

B. Survey Interface

To assess the effectiveness of our connectivity score and to understand the impact of different subtopics ranking, we have developed interactive interfaces, as shown in Figures 8 and 9, respectively. These interfaces allow users to visually explore how the connectivity score and subtopics ranking contribute to the overall performance of our system.

C. Extra Case Studies

In addition to these interactive studies, we have compiled a range of case studies spanning various topics in Table III. These case studies further exemplify the versatility and robustness of our approach in handling diverse content.

Furthermore, Figure 10 provides a glimpse into the sentiment analysis aspect of our system. It depicts the varied sentiments associated with the purchase of "*Mass media in the United States*" by "*American billionaires*", showcasing our system's ability to extract and analyze sentiment from news articles, which can be particularly useful for understanding public opinion and media portrayal of specific events or transactions.

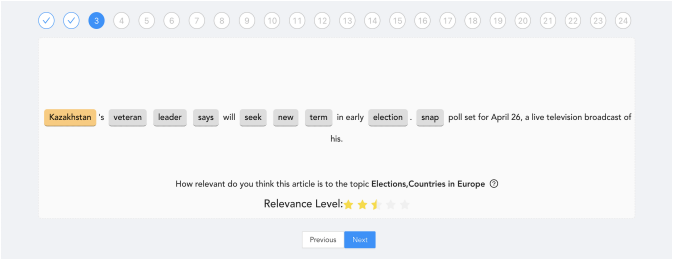


Fig. 8: Concept document relevance study survey interface. Each participant is asked to select two concepts and rate 25 documents.

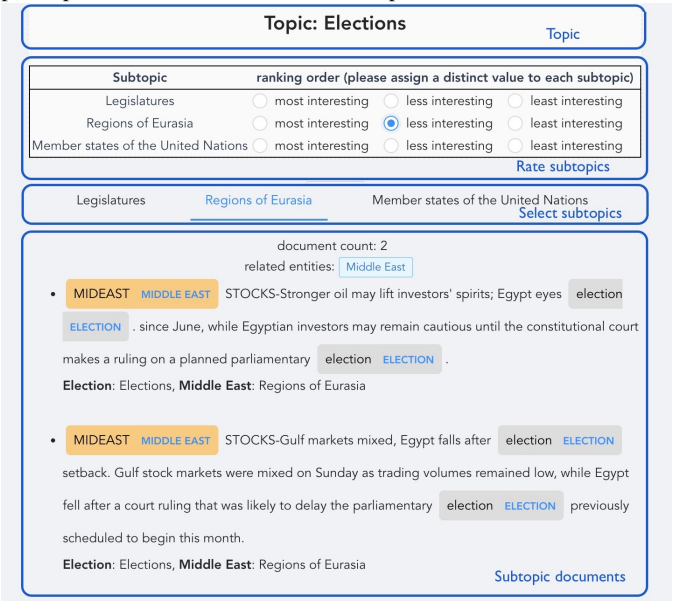


Fig. 9: Subtopic survey interface

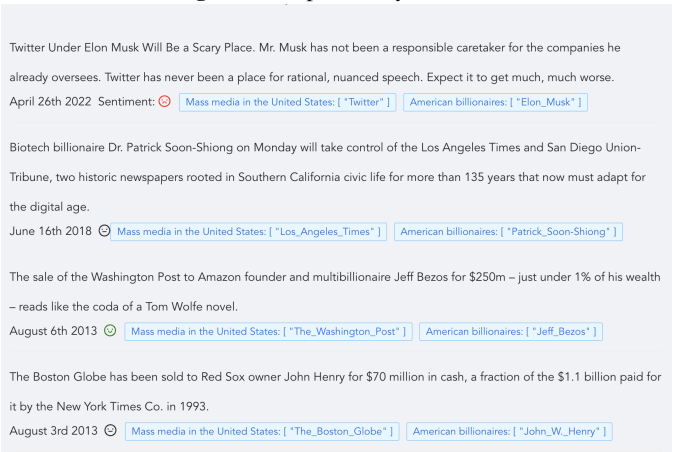


Fig. 10: Case study on Media Bias

TABLE III: Extra Case Studies

Query	Roll-up options	Drill-down options	Results
Asian Countries, Stock Market	Financial Markets	Forex Market, National Oil and Gas Companies	1) Indonesia Mining IPO Vaults Six Shareholders to Billionaire Status 2) China Tightens Oversight of Program Trading in Stock Market 3) TY Fashion, a Taiwanese-owned garment manufacturer in Cambodia, has received approval for listing on the Cambodia Securities Exchange
Biotechnology Companies of the U.S., M&A	Corporate Finance	Orphan Drug Companies, Multinational Food Companies	1) Unilever has expressed interest in acquiring GlaxoSmithKline's consumer healthcare business, a joint venture with Pfizer. 2) GSK leads race to buy Pfizer unit. 3) Pfizer to Purchase Cancer Drugmaker Seagen for $43 Billion
EU countries, Labor Disputes	social conflicts	Minimum Wage	1) Protesters in France demonstrate against President Emmanuel Macron's pension overhaul and the end of the longest transport strike in French history. 2) Greece's leftist government plans to raise the minimum wage and restore collective bargaining, as it seeks to renegotiate the terms of its international bailout 3) Amazon faces strike of logistical workers in Germany
Investment Banks, Commercial Crimes	Corporate Crimes	Hedge Funds, Online Brokerages, Investment Funds, Interest Rates	1) Broker Faces Insider Trading Probe Tied to Morgan Stanley Deals. 2) Ex-Morgan Stanley Adviser Gets Seven Years for Fleecing Clients in Ponzi Scheme. 3) Deutsche Bank Said to Probe Senior Russia Employee Over Bribes. 4) UBS has lost its appeal against a French tax evasion charge. 5) JPMorgan fined for wash trades in oil, gasoline