Knowledge Management System with NLP-Assisted Annotations: A Brief Survey and Outlook

Baihan Lin[1,*]

[1]Columbia University, New York, NY 10027, USA

Abstract

Knowledge management systems (KMS) are in high demand for industrial researchers, chemical or research enterprises, or evidence-based decision making. However, existing systems have limitations in categorizing and organizing paper insights or relationships. Traditional databases are usually disjoint with logging systems, which limit its utility in generating concise, collated overviews. In this work, we briefly survey existing approaches of this problem space and propose a unified framework that utilizes relational databases to log hierarchical information to facilitate the research and writing process, or generate useful knowledge from references or insights from connected concepts. Our framework of bidirectional knowledge management system (BKMS) enables novel functionalities encompassing improved hierarchical note-taking, AI-assisted brainstorming, and multi-directional relationships. Potential applications include managing inventories and changes for manufacture or research enterprises, or generating analytic reports with evidence-based decision making.

Keywords

knowledge management, insight annotation, relational databases, natural language processing, machine learning

1. Introduction

Knowledge management systems (KMS) are the driving engines of modern day information technologies (IT). These IT systems store data in parsed ways and retrieve knowledge insights to improve the information understanding, team collaboration and process alignment within organizations and groups. As an engineering entities in high demand for industrial researchers, chemical or research enterprises and evidence-based decision making, knowledge management systems are often used by organizations to affect innovation performance and generate accurate metrics on organizational capacity [1], but they can also be user-centric by centering the knowledge base around individual users or customers [2].

Take the application of reference management of academic researchers as an example. KMS are often used by researchers to keep track of papers or subsets of papers [3]. Usually, the research information of different papers or references has meta information that can be filtered and sorted. An example scenario would be: a scientist logs or inputs a particular paper into a system, with each entry containing many meta information about the papers. These meta information elements can be filtered or sorted (e.g., by year, journal, author, etc.). Each paper might contain multiple concepts or topics, and each topic might contain multiple paper. In some cases, we might want the system to be able to automatically assign topic to some papers based on text data mining. The user can filter the papers by topics. Within each paper, during the reading, the scientist might want to log an insight or note on certain paragraphs. Sometimes the notes can be about multiple papers, and their relationship can be in various types. These notes or insights also have topic tags, which can optionally be automatically curated. The system can also generate useful concepts or knowledge as well as their references to facilitate the research and writing process of the scientist.

We see from this example that the relationships between papers chosen in academic fields can have multiple, bidirectional relationships. Existing knowledge management systems for organizing research papers in scientific fields or organizing manufacture enterprises use directed acyclic graphs, Bayesian networks, and machine learning [3], which have limitations in categorizing and organizing these multi-faceted insights or relationships. This is because many traditional databases are usually disjoint with logging systems, which limit its utility in generating concise, collated overviews. In this work, we briefly survey existing approaches in the general field of these knowledge management systems, and propose a unified framework as a solution to these challenges. In our framework, we describe a knowledge management system that utilizes relational databases to log hierarchical information with connected concepts.

Back to the example problem of reference management, our KMS would utilize relational databases to log hierarchical information to facilitate the research and writing process, or to help generate useful knowledge from references or insights from connected concepts. This would enable novel functionalities encompassing improved hier-

CIKM 22: Workshop on Human-In-the-Loop Data Curation, October 21, 2022, Atlanta, GA

[*]Corresponding author.

baihan.lin@columbia.edu (B. Lin)

https://www.neuroinference.com/ (B. Lin)

0000-0002-7979-5509 (B. Lin)

archical notetaking, AI-assisted brainstorming, and multidirectional relationships. For instance, one can generate reports given keywords or topics collating hierarchical and intra-connected records. With these automatic annotations, the system can enable automatic curation of topic tags using text data mining. Other applications include managing inventories and changes for manufacture or research enterprises or generating analytic reports with evidence-based decision making.

Although we have seen successful system designs in commercial products such as Mendeley and recent community efforts such as Open Research Knowledge Graph (ORKG), we believe that our survey can still bring useful and new insights on the practical considerations on the intersections among machine learning, database management and human-system collaboration. In the following sections, we will first briefly survey the existing knowledge management systems approaches, and propose a unified bidirectional KMS (BKMS) framework that utilizes relational databases to log hierarchical information to facilitate the research and writing and generate helpful knowledge from references or insights from related concepts. We present a useful and novel system design for this bidirectional information management, formulate a few potential use-cases for this design, address the four-subset system of NLP-assisted annotations, and discuss future design considerations.

2. An Applied Perspective

2.1. Applications

There are different application domains for knowledge management systems with relational databases and insight annotation enabled by machine learning, including but not limited to reference manager for academic researchers, education and research tool, consulting firm report generator with evidence-based decision making, inventory management for manufacture or research enterprises, organizational tool for industries with high-volume data, and internal auditing tool for customized employee metrics.

2.2. User scenarios

Other than the reference management example in our introduction, we also include two additional applications. The first one is managing inventories and changes for manufacture, chemistry or research enterprises. The inventories or measurements of factories usually involves dependency and hierarchical interactions. A knowledge management system that uses a relational database instead of disjoint databases with separate logging systems can enable useful curation function to offer very useful and concise report regarding key events or phenomon (like the topics). These are important insights to keep the factories or warehouses in safety.

The second user scenario example is evidence-based decision making. In large business entities, critical decisions are usually made with a group of market researchers or consulting firms that come up with various analytic reports. A knowledge management system with AI-assisted insight annotation can provide a fast and evidence-based solution by generating a report (given the keyword or topic as input) which curates from hierarchical and interaconnected records. This hierarchical knowledge graph can serve as a useful primer in important decision making processes and guide the investigators to locate relevant resources.

2.3. Case studies

In this section, we outline three case studies that recent real-world knowledge management systems are likely adopt to become more interconnected and intelligent.

The concept of Internet of Things (IoT): The IoT advancements consist of a series of disruptive digital technologies, semantic languages, and virtual identities that can increases efficiency and effectiveness in daily life operations through interconnected communications among devices and systems [4]. Other than these organizational benefit, IoT stimulates the innovation process in various aspects, through fast iterations of knowledge flow and information gathering [5]. In [6], researchers employ structural equation modelling on a sample of 298 Italian firms from different sectors. Their study suggest that interconnected knowledge management systems facilitate the creation of a open and collaborative ecosystem by utilizing the internal and external flows of knowledge and increasing internal knowledge management capacity, which in turn increases innovation capacity.

Reference architecture: In the era of Industry 4.0 [7], smart warehouses are envisioned to host production that contains modular and efficient manufacturing systems and characterizes scenarios in which products control their own manufacturing process. As in our user scenario of warehouse inventory management, an optimal reference architecture would be the key to the warehouse knowledge management system. For instance, [8] describes a pipeline to perform a series of systematic analyses to identify the key concerns and processes and eventually arrive at potential architecture of smart warehouses. They conduct a case study at a large warehouse in the food industry and illustrates that an introduction of a reference architecture can be effective and practical.

Conversational recommendation systems: A conversational recommendation system (CRS) is a computer system that is able to have a conversation with a human user in order to make recommendations [9]. This is different from traditional recommendation systems, which

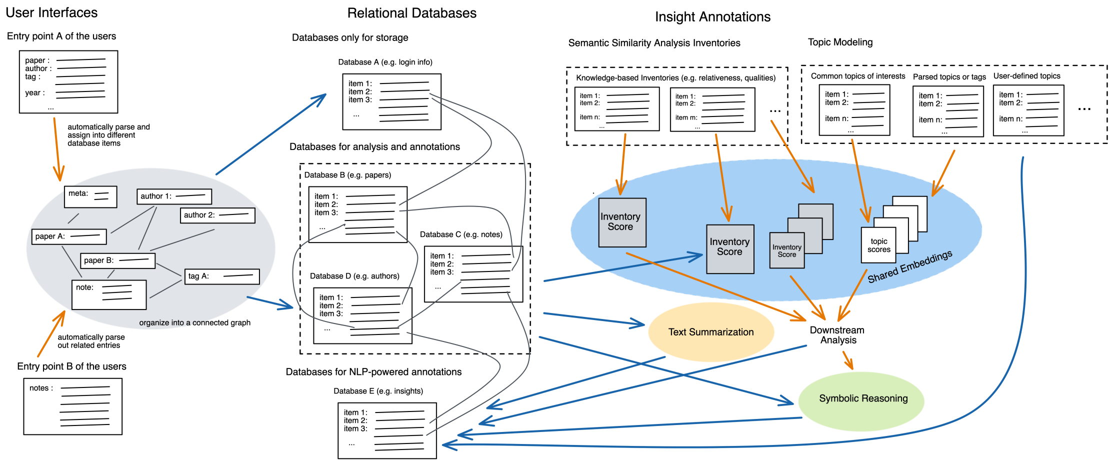
User Interfaces
Entry point A of the users
paper :
author :
tag :
year :
...
automatically parse and assign into different database items
meta:
author 1:
author 2:
paper A:
paper B:
tag A:
note:
organize into a connected graph
automatically parse out related entries
Entry point B of the users
notes :
Relational Databases
Databases only for storage
Database A (e.g. login info)
item 1:
item 2:
item 3:
...
Databases for analysis and annotations
Database B (e.g. papers)
Database C (e.g. notes)
Database D (e.g. authors)
Databases for NLP-powered annotations
Database E (e.g. insights)
Insight Annotations
Semantic Similarity Analysis Inventories
Knowledge-based Inventories (e.g. relativeness, qualities)
item n:
item m:
Topic Modeling
Common topics of interests
Parsed topics or tags
User-defined topics
Inventory Score
Inventory Score
Inventory Score
topic scores
Shared Embeddings
Text Summarization
Downstream Analysis
Symbolic Reasoning

Figure 1: A unified framework of a knowledge management system with relational databases and NLP-assisted annotation

do not interact with users. Often used in e-commerce, social media, and entertainment applications, CRS are becoming increasingly popular as they can provide a more personalized and interactive experience for users, but can pose additional challenges in managing different layers of knowledge at different states: the intent of the conversation, the entities matched by the intents, the long-term preferences of the users and similar users, their state-dependent preferences related to the current contexts, and the relationships between different entities, intents and users. One practical examples is recommending discussion topic to therapist during psychotherapy in real-time given automatically speech-transcribed dialogue records [10] and helpful visual analytics [11].

3. Bidirectional KMS Framework

Figure 1 outlines our framework of bidirectional knowledge management systems (BKMS) with relational databases and insight annotation powered by natural language processing (NLP). The user interface provides the entry points into our knowledge management systems. Different interfaces introduces different routes, but they all involve a parsing and extraction process to atomize the user inputs into nodes that connects in a small knowledge graph. This graph is then placed into a relational database where their links are preserved. The orange and blue arrows indicates intro- and inter-database data flows. The relational databases include three parts. Some databases in the relational databases are only used for storage. Some are used for analysis and annotations. And some databases are kept to store annotated insights or other downstream analytical artifacts, which provide an additional data flow direction.

4. NLP-Assisted Insight Annotation

As shown in the annotation component of Figure 1, there are several routes we can utilize natural language processing to generate and annotate insights within our databases. We will elaborate on how they play in knowledge management systems and survey modern machine learning methods in each of these routes below.

Semantic similarity: In principle, any sentence or paragraph embeddings can help us characterize our document and inventories of interest. For instance, the Doc2Vec embedding [12] is a popular unsupervised learning model that learns vector representations of sentences and text documents. It improves upon the traditional bag-of-words representation by utilizing a distributed memory that remembers what is missing from the current context. SentenceBERT [13] is another popular option which modifies a pre-trained BERT network by using siamese and triplet network structures to infer semantically meaningful sentence embeddings. With word or sentence embeddings, we can embed the document entries from our relational databases into vectors, and then compute the cosine similarity between the vector at certain turn and an inventory entry. With that, for each text, we obtain a N-dimension score for the said property. For instance, the inventory can be written guidelines that evaluate the usefulness of certain documents, say, a list of leadership principles that some companies use to evaluate a candidate's resume, work report or performance review form. And the relational database could be hosting an employee's self reported performance review form. The system can automatically compute a score based on each item of the guidelines and annotate these document entry accordingly. Other applications can be evaluating the patient-doctor alignment from an automatically transcribed psychotherapy sessions based on a clinical questionnaire inventory, as shown in [14, 15, 16].

Topic modeling: In natural language processing and machine learning, a topic model is a type of statistical graphical model that help uncover the abstract "topics" that appear in a collection of documents. The topic modeling technique is frequently used in text-mining pipeline to unravel the hidden semantic structures of a text body. This can be very handy in annotating the database entry. For instance, a user scenario could be in a clinical consumer-facing chatbot, where the dialogue between the client and agent is transcribed, and a topic modeling analysis is automatically performed and generate a list of discussed topics and their scores based on semantic similarity, as shown in [17]. Several state-of-the-art neural topic models include the Neural Variational Document Model (NVDM) [18] (an unsupervised text modeling approach based on variational auto-encoder), Gaussian softmax construction (GSM) [19] (a NVDM variant), the Wasserstein-based Topic Model (WTM) [20], the Embedded Topic Model (ETM) [21] among others.

Text summarization: When the scale of our databases increases, maintaining the interpretability of our knowledge management system becomes more and more challenging. This expanding availability of documents and entries inside the database cannot yield actionable insights without proper aggregation. The field of automatic text summarization deals with this problem by producing a concise and fluent summary while preserving key information content and overall meaning [22]. For instance, we can first group or cluster the database entries (such as paper abstracts, or reading notes as in our reference manager example) by their semantic similarity or inferred topics. And then, within each group, generate a condensed descriptions. A user case would be, automatically generating writing outlines or topics based on the available references and reading notes in a paper reference manager. In the active field of text summarization, extraction and abstraction are the two main approaches. The extractive summarization techniques generate summaries by choosing a subset of the sentences in the original text, by computing first an intermediate representation of the text, then a sentence score and finally a subset selection operation onto the original texts [23]. The abstraction approach uses latent semantic analysis, frequency-driven approaches [24] and topics modeling which we cover above.

Symbolic reasoning: While topic modeling offers interpretable subjects, and text summarization offers interpretable paragraphs, the logic and causal relationship between these insights can be arbitrary. The field of symbolic AI bridge this gap by introducing high-level and human-readable symbolic representations into these practical problems. They can potentially derive logic programming rules and semantic relationships that can be use as actionable knowledge graphs [25]. Recently, there have also been increasing interests in a modern approach called neuro-symbolic AI [26, 27], where the well-founded knowledge representation and reasoning from the symbolic perspective are integrated with deep learning from the statistical perspective. This offers both effective predictive power and necessary explainability for many real-world applications.

5. Practical Considerations

When designing a interconnected and intelligent knowledge management systems for a domain-specific application, here are some practical questions to be considered:

- Database consideration: What are the storage capacities of this technology?
- *User interface*: What visual and user interface is preferred by users?
- Organizational benefits: What specific organizational functionality would this system provide over current systems?
- Latency and responsiveness: What are the synchronization capacities of this technology across devices?
- Customization: Can users modify or customize this system to their own preferences?
- Security: Would this technology allow for secure encryption or storage of higher value data?
- Collaboration: Would this system allow for collaborative use by multiple stakeholders?
- Investigation: What kind of insights or investigations do we wish to gain from this system?
- I/O: Would this system allow import or export from other knowledge management systems?

Other than these practical questions to consider, a more thorough design process would involve market analysis (market size, emerging technologies, policies, challenges, new trends, and policies as in [28]), domain analysis (systematic activity for deriving, storing domain knowledge to support the engineering design process as in [29]), business process modeling (i.e. identifying the lead processes and subprocess of outgoing products [30]) and architecture design with viewpoints (stakeholder concerns, context diagram, decomposition view, uses view, and deployment view [31, 32]). Sometimes, case studies can also be useful to clarify the problem settings.

Since we are proposing the idea of introducing relational databases and various AI and symbolic techniques in knowledge management systems, there are additional future research challenges in relation to this proposition in terms of the human-system "collaboration" enabled by these systems. Methodologically, tne machine learning engine that powers many human-in-the-loop (HIL) solutions in data curation is reinforcement learning methods that have been demonstrated to effectively learn from human interactions with the speech- or text-based systems [33]. Operationally, from the human side, we need to encourage people to contribute their knowledge and expertise (e.g. crowdsourcing) by creating an effective user interface that allows people to easily log in, search for and find the information they need.From the system side, we need to ensure that knowledge is effectively captured and stored, consistently updated to keep the knowledge up to date and accuratem and manage different types of knowledge such that it is accessible to the right people. Finally, there are also ethical and societal considerations when we use machine learning and AI to encode knowledge related to human biometrics and well-beings, as reviewed in [34].

6. Conclusions

In summary, we describe the applied problem of a knowledge management systems that host information that contain multiple and bidirectional relationships in layers of meta data. We briefly survey the application domains, user scenarios and the existing approaches in the fields, and eventually propose a framework for a knowledge management system with relational database and NLP-assisted insight annotation. In our framework, a knowledge management system can comprise a user interface to provide input and present output relating to one or more documents or sensors. The system maintains a relational database storing information relating to the one or more documents, and a knowledge parsing unit, in communication to the user interface and the server, can determine at a first time instance the metadata information elements associated with the particular document entry. The databases can then be automatically annotated with NLP techniques such as semantic similarity analysis, topic modeling, text summarization and symbolic reasoning. A knowledge graph can then be learned from these language models to be used as interpretable insights for real-world downstream tasks.

References

[1] B. Lawson, D. Samson, Developing innovation capability in organisations: a dynamic capabilities approach, International journal of innovation management 5 (2001) 377–400.

[2] M. A. Kabir, J. Han, J. Yu, A. Colman, User-centric social context information management: an ontology-based approach and platform, Personal and Ubiquitous Computing 18 (2014) 1061–1083.

[3] Y. M. Yee, C. L. Tan, R. Thurasamy, Back to basics: building a knowledge management system, Strategic Direction (2019).

[4] V. Scuotto, A. Ferraris, S. Bresciani, Internet of things: applications and challenges in smart cities. a case study of ibm smart city projects., Business Process Management Journal (2016).

[5] Y. Malhotra, Knowledge management for e-business performance: advancing information strategy to “internet time”, Information Strategy: The Executive’s Journal 16 (2000) 5–16.

[6] G. Santoro, D. Vrontis, A. Thrassou, L. Dezi, The internet of things: Building a knowledge management system for open innovation and knowledge management capacity, Technological forecasting and social change 136 (2018) 347–354.

[7] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, M. Hoffmann, Industry 4.0, Business & information systems engineering 6 (2014) 239–242.

[8] M. van Geest, B. Tekinerdogan, C. Catal, Design of a reference architecture for developing smart warehouses in industry 4.0, Computers in industry 124 (2021) 103343.

[9] Y. Sun, Y. Zhang, Conversational recommender system, in: The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 235–244.

[10] B. Lin, G. Cecchi, D. Bouneffouf, Supervisorbot: Nlp-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning, arXiv preprint arXiv:2208.13077 (2022).

[11] B. Lin, Voice2alliance: automatic speaker diarization and quality assurance of conversational alignment, in: INTERSPEECH, 2022.

[12] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[13] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, Preprint arXiv:1908.10084 (2019).

[14] B. Lin, G. Cecchi, D. Bouneffouf, Deep annotation of therapeutic working alliance in psychotherapy, Preprint arXiv:2204.05522 (2022).

[15] B. Lin, Personality effect on psychotherapy outcome: A predictive natural language processing framework, arXiv preprint (2022).

[16] B. Lin, G. Cecchi, D. Bouneffouf, Working alliance transformer for psychotherapy dialogue classification, arXiv preprint arXiv:2210.15603 (2022).

[17] B. Lin, D. Bouneffouf, G. Cecchi, R. Tejwani, Neural topic modeling of psychotherapy sessions, Preprint arXiv:2204.10189 (2022).

[18] Y. Miao, L. Yu, P. Blunsom, Neural variational infer-

ence for text processing, in: International conference on machine learning, PMLR, 2016, pp. 1727–1736.

[19] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, in: International Conference on Machine Learning, PMLR, 2017, pp. 2410–2419.

[20] F. Nan, R. Ding, R. Nallapati, B. Xiang, Topic modeling with wasserstein autoencoders, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6345–6381.

[21] A. B. Dieng, F. J. Ruiz, D. M. Blei, Topic modeling in embedding spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453.

[22] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, Text summarization techniques: a brief survey, Preprint arXiv:1707.02268 (2017).

[23] A. Nenkova, K. McKeown, A survey of text summarization techniques, in: Mining text data, Springer, 2012, pp. 43–76.

[24] T. E. Dunning, Accurate methods for the statistics of surprise and coincidence, Computational linguistics 19 (1993) 61–74.

[25] M. Garnelo, M. Shanahan, Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, Current Opinion in Behavioral Sciences 29 (2019) 17–23.

[26] A. d. Garcez, L. C. Lamb, Neurosymbolic ai: the 3rd wave, Preprint arXiv:2012.05876 (2020).

[27] J. Zhang, B. Chen, L. Zhang, X. Ke, H. Ding, Neural, symbolic and neural-symbolic reasoning on knowledge graphs, AI Open (2021).

[28] G. Giudici, A. Milne, D. Vinogradov, Cryptocurrencies: market analysis and perspectives, Journal of Industrial and Business Economics 47 (2020) 1–18.

[29] Ö. Köksal, B. Tekinerdogan, Feature-driven domain analysis of session layer protocols of internet of things, in: 2017 IEEE International Congress on Internet of Things (ICIOT), IEEE, 2017, pp. 105–112.

[30] M. Weske, Business process modelling foundation, in: Business Process Management, Springer, 2019, pp. 71–122.

[31] P. Clements, D. Garlan, R. Little, R. Nord, J. Stafford, Documenting software architectures: views and beyond, in: 25th International Conference on Software Engineering, 2003. Proceedings., IEEE, 2003, pp. 740–741.

[32] E. Demirli, B. Tekinerdogan, Software language engineering of architectural viewpoints, in: European Conference on Software Architecture, Springer, 2011, pp. 336–343.

[33] B. Lin, Reinforcement learning and bandits for speech and language processing: Tutorial, review and outlook, arXiv preprint arXiv:2210.13623 (2022).

[34] B. Lin, Computational inference in cognitive science: Operational, societal and ethical considerations, arXiv preprint arXiv:2210.13526 (2022).