2022 IEEE International Conference on Web Services (ICWS) | 978-1-6654-8143-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICWS55610.2022.00069

Knowledge Base 4.0: Using Crowdsourcing Services for Mimicking the Knowledge of Domain Experts

Amin Beheshti
School of Computing, Macquarie University, Sydney, Australia
amin.beheshti@mq.edu.au

***Abstract*—Intelligence is the ability to learn from experience. Knowledge workers in knowledge-intensive processes develop invaluable domain-specific expertise and knowledge over time. Accordingly, it is vital for organizations to capture this knowledge (which is hidden in the biological Neural Network of subject-matter experts) and enable novices/inexperienced knowledge workers to benefit from that in choosing the best next steps. This position paper proposes linking weak supervision and crowdsourcing techniques to incorporate knowledge in a continual fashion based on estimating uncertainty or errors from the existing knowledge and learning models. Applications may include handling cold start and concept drift situations. We discuss the design of an intelligent Knowledge Base (KB), namely KB 4.0, for mimicking the knowledge of domain experts in knowledge-intensive processes, and using this knowledge to facilitate auto labelling of the data to be used in learning algorithms. We present a motivating scenario in police investigation processes, and argue how inexperienced investigators can benefit from such a domain-specific KB in law enforcement.**

***Index Terms*—Knowledge Base, Crowdsourcing, weak supervision, Knowledge-Intensive Business Processes**

I. Scope and Motivation

In the age of Big Data, Artificial Intelligence (AI) development needs to focus more on *data* and the *context* around that. Therefore, this work proposes linking weak supervision and crowdsourcing techniques to incorporate knowledge in a continual/online fashion based on estimating uncertainty or errors from the existing knowledge and learning models.

Knowledge Base. A Knowledge Base (KB) is traditionally defined as a human/machine-readable library of information about concepts such as a person, product, organization, service, and topic [1]. Such concepts can be organized into a taxonomy, instances for each concept, and relationships among the concepts [2]. A KB can focus on general knowledge (e.g., 'wikidata.org/', 'dbpedia.org/', and 'yago-knowledge.org/' that provides general knowledge about people, cities, countries, movies, and organizations) or can be domain-specific (e.g., METASPACE [3] which is a community-populated KB of spatial metabolomes in health and disease). Many KBs are interlinked to form the backbone of the Web of Linked Data [4] with the goal of evolving the Web into a global data space. Such a data space will enable data-centric and knowledge-intensive applications (e.g., personalization and recommendation, entity linking, deep Q&A, and semantic search) to be more intelligent. Google Knowledge Graph [5] is an example of this.

A KB can be manually created. WordNet, i.e., a large lexical database, is an example of a manually created KB. Recent advances in building KBs focused on automating the building of large KB's [6]. These approaches mainly use Information Extraction (IE) techniques and harnessing private/public knowledge sources to automatically identify instances for each concept (e.g., a named entity such as Barack Obama extracted from a textual data and assigned it to the concept Person) as well as the relationships among them (e.g., identify the relationship between Barack Obama, an instance of the concept Person, and the United States, an instance of the concept Country). Other related work put one step forward to building KBs automatically and automating the curation of large KBs. For example, in our previous work DataSynapse [7], [8], we offered a curation pipeline to extract, enrich, and annotate information items related to instances of concepts in KBs to facilitate understanding of the relationships among the instances of concepts. DataSynapse focuses on building a domain-specific KB to offer a machine-readable library of information about concepts related to the government budget. The approach extends the state-of-the-art in Weak Supervision [9], [10] by breathing domain-specific knowledge into learning pipelines.

Weak Supervision. Machine Learning (ML) encompasses a broad range of algorithms that can improve automatically through experience [11]. Recent progress in ML has been driven by the availability of Big Data, i.e., the large amount of data generated on open, private, social, and Internet of Things (IoT) data islands [12]. However, the main challenge for learning algorithms is the Poor Quality of Data. Data labelling is the process of understanding the context around raw data and using labels to breathe meanings into the data, so that the ML model can learn from the contextualized labelled data. The main goal here is to provide ground truth data for ML models. Labelling the data is challenging, and the current state of the art in Machine Learning relies on massive sets of hand-labelled training data. Hand-labelling is quite expensive and requires domain experts to breathe their knowledge into the data (in the form of labels) by spending time and moving through each information item in the data set. Transfer learning [13] tries to address this challenge by using models that are already

978-1-6654-8143-4/22/$31.00 ©2022 IEEE
DOI 10.1109/ICWS55610.2022.00069

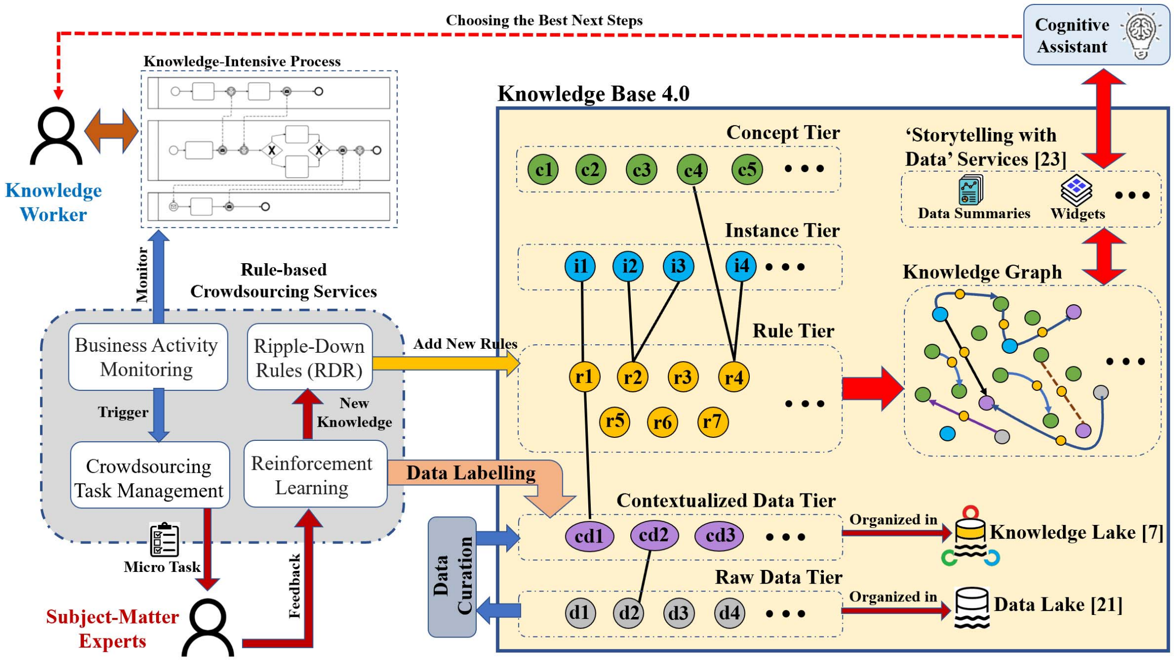

Fig. 1. Proposed framework for Knowledge Base 4.0.

trained and applying them to similar tasks. However, this approach cannot be generalized to all different tasks. To fill this gap, weak supervision approaches [9], [14], [15] recently focused on leveraging higher-level and/or noisier input from subject-matter experts, i.e., domain experts who have extensive knowledge and experience in a specific domain. An example could be domain experts with 20+ years of experience in the field, such as a police investigator who is an expert in criminal investigations, a teacher who is an expert in identifying creative students, or a psychologist who is an expert in identifying patterns of mental health disorder.

Motivation: Knowledge-intensive Processes. Over the last decade, many Business Processes (BPs) across and beyond the enterprise boundaries have been implemented. The ongoing explosion in the availability of Big Data and advances in Data Science, made data-centric and knowledge-intensive processes more prevalent [16]. Knowledge-intensive processes contain a set of coordinated tasks and activities, controlled by knowledge workers to achieve a business objective or goal. Subject-matter experts in these processes gain invaluable experience dealing with such ad-hoc processes over time. In this context, intelligence can be defined as learning from experience and using domain experts' knowledge to facilitate decision-making. For example, a police investigation is a data-centric and knowledge-intensive process. Police investigators (i.e., subject-matter experts in the investigation process) develop invaluable knowledge and experience over time to ascertain the hypothesis, methods, motives, interview witnesses, and more. We argue that understanding and organizing this knowledge is challenging, and leveraging rule-based crowdsourcing services and combining them with reinforcement techniques will enable us to mimic the knowledge of domain experts while they are involved in investigation processes.

II. Building Knowledge Base 4.0

The fourth industrial revolution, Industry 4.0, is fast transforming how businesses operate. Artificial Intelligence, and Machine Learning at its core, have brought about a new type of data-centric and knowledge-intensive processes. In this context, it will be vital for organizations to capture this intelligence and enable novices/inexperienced knowledge workers to benefit from this in choosing the best next steps. To address this challenge, we propose using rule-based crowdsourcing services to mimic the knowledge of domain experts in data-centric and knowledge-intensive processes and use this knowledge to build an intelligent KB, namely Knowledge Base 4.0. The goal is to facilitate auto labelling of the data to be used in learning processes. Figure 1 illustrates the proposed framework for building Knowledge Base 4.0. In the following, we present the main component of the proposed framework.

The *Rule-based Crowdsourcing Services* component contains the *Business Activity Monitoring* agent [17] which monitors business activities and can identify situations where a knowledge worker or a learning model has difficulties in choosing the best next steps. Examples may include handling cold start and concept drift situations. This, in turn, will trigger the *Crowdsourcing Task Manager* service to generate a microtask and share that with subject-matter experts. The feedback provided by the experts will then feed into the *Reinforcement Learning* service, which may leverage state-of-the-art techniques, e.g., CrowdRL [18], to learn from subject-matter experts' feedback and use this knowledge for integrating task selection, task assignment and truth inference [19] together. This knowledge will be used to automatically label the data, and at the same time, will be fed into the *Ripple-Down Rules (RDR) [20]* component, which is an approach to building knowledge-based systems incrementally, while the KB is in routine use.

In particular, RDR is responsible for adding new rules to the *Rule Tier* in the Knowledge Base, where a rule r is accountable for identifying relationships among concepts, instances of the concepts, and information items (that are stored in *Raw Data Tier* and *Contextualized Data Tier*). The *Contextualized Data Tier* consists of: (i) Curated Data: curation process is responsible for transforming the raw data (stored in the Data Lake [21]) into contextualized data and knowledge (stored in the Knowledge Lake [7]. At this stage, our previous work [22] can be used to automate the curation process; and (ii) Labelled Data: recall that the *Reinforcement Learning service* component automates the labelling process based on the knowledge/feedback provided by the domain experts. The ’Data Labelling‘ arrow in Figure 1, connecting the Reinforcement Learning component into the Contextualized Data Tier, will offer ML models to learn from the contextualized labelled data. This in turn will facilitate discovering deep insights that are trapped in the relationships among entities in the Knowledge Graph.

Rules may have different types. For example, in Figure 1: (i) rule $r4$ identified the relationship between concept $c4$ and instance $i4$; (ii) rule $r2$ identified the relationship between two instances $i2$ and $i3$; and (iii) rule $r1$ identified the relationship between instance $i1$ and information item $cd1$. Rules can be more complicated (similar to our previous work [22]) and identify the complex relationship among several concepts, instances, information items, and/or patterns in the data. The rule tier, as illustrated in Figure 1, will facilitate the construction of the Knowledge Graph, which generates a graphical representation of the relationships among the data points in KB 4.0. The ‘Storytelling with Data’ [23] layer will: (i) provide services to generate data products [24] and data summaries; and (ii) provide a services layer to provide secure access to a cognitive assistant, similar to our previous work iCOP [25], to assist knowledge workers in choosing the best next steps.

III. Summary and Discussion

Injecting domain knowledge into learning algorithms is not a new concept. However, the main novelty in asking this question now is that the advancement in Artificial Intelligence (AI) and Data Science is transforming business processes fundamentally by assisting knowledge workers in communicating analysis findings, supporting evidence, and making decisions. Accordingly, the core of the idea for the concept of Knowledge Base 4.0 is to assist organizations in identifying novel applications of AI and Data Science, from process automation to cognitive assistants.

References

[1] D. B. Lenat, “Cyc: A large-scale investment in knowledge infrastructure,” *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[2] X. Chai, O. Deshpande, N. Garera, A. Gattani, W. Lam, D. S. Lamba, L. Liu, M. Tiwari, M. Tourn, Z. Vacheri *et al.*, “Social media analytics: The kosmix story.” *IEEE Data Eng. Bull.*, vol. 36, no. 3, pp. 4–12, 2013.

[3] T. Alexandrov *et al.*, “Metaspace: A community-populated knowledge base of spatial metabolomes in health and disease,” *BioRxiv*, 2019.

[4] T. Heath and C. Bizer, “Linked data: Evolving the web into a global data space,” *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[5] A. Singhal, “Introducing the knowledge graph: things, not strings,” *Official google blog*, vol. 5, p. 16, 2012.

[6] F. M. Suchanek and G. Weikum, “Knowledge bases in the age of big data analytics,” *PVLDB*, vol. 7, no. 13, pp. 1713–1714, 2014.

[7] A. Beheshti, B. Benatallah, R. Nouri, and A. Tabebordbar, “CoreKG: a knowledge lake service,” *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 1942–1945, 2018.

[8] A. Beheshti, B. Benatallah, A. Tabebordbar, H. R. Motahari-Nezhad, M. C. Barukh, and R. Nouri, “Datasynapse: A social data curation foundry,” *Distributed Parallel Databases*, vol. 37, no. 3, 2019.

[9] T. S. A. L. Blog, “Weak supervision: A new programming paradigm for machine learning,” http://ai.stanford.edu/blog/weak-supervision/, May 2022.

[10] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, “Snorkel: Rapid training data creation with weak supervision,” in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11, no. 3, 2017, p. 269.

[11] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[12] A. Beheshti, B. Benatallah, Q. Z. Sheng, and F. Schiliro, “Intelligent knowledge lakes: The age of artificial intelligence and big data,” in *WISE*, vol. 1155. Springer, 2019, pp. 24–34.

[13] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[14] C. Shin, W. Li, H. Vishwakarma, N. Roberts, and F. Sala, “Universalizing weak supervision,” *arXiv preprint arXiv:2112.03865*, 2021.

[15] P. Lison, J. Barnes, and A. Hubin, “skweak: Weak supervision made easy for nlp,” *arXiv preprint arXiv:2104.09683*, 2021.

[16] S. Beheshti, B. Benatallah, S. Sakr, D. Grigori, H. R. Motahari-Nezhad, M. C. Barukh, A. Gater, and S. H. Ryu, *Process Analytics - Concepts and Techniques for Querying and Analyzing Process Data*. Springer, 2016.

[17] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, “Real-time business activity monitoring and analysis of process performance on big-data domains,” *Telematics and Informatics*, vol. 33, no. 3, pp. 793–807, 2016.

[18] K. Li, G. Li, Y. Wang, Y. Huang, Z. Liu, and Z. Wu, “Crowdrl: An end-to-end reinforcement learning framework for data labelling,” in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 289–300.

[19] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, “Truth inference in crowdsourcing: Is the problem solved?” *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[20] P. Compton, L. Peters, G. Edwards, and T. G. Lavers, “Experience with ripple-down rules,” in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2005.

[21] A. Beheshti, B. Benatallah, R. Nouri, V. M. Chhieng, H. Xiong, and X. Zhao, “CoreDB: a data lake service,” in *CIKM*. ACM, 2017, pp. 2451–2454.

[22] B. Benatallah, M. Barukh, A. Beheshti, and S. Zamani, “Method and system for data curation,” 2019, uS Patent WO2019173860A1.

[23] A. Beheshti, A. Tabebordbar, and B. Benatallah, “iStory: Intelligent storytelling with social data,” in *Companion of The 2020 Web Conference*. ACM / IW3C2, 2020, pp. 253–256.

[24] J. Yang, Y. Tang, and A. Beheshti, “Design methodology for service-based data product sharing and trading,” in *Next-Gen Digital Services*, ser. Lecture Notes in Computer Science, vol. 12521. Springer, 2021, pp. 221–235.

[25] F. Schiliro, A. Beheshti, S. Ghodratnama, F. Amouzgar, B. Benatallah, J. Yang, Q. Z. Sheng, F. Casati, and H. R. Motahari-Nezhad, “iCOP: Iot-enabled policing processes,” in *ICSOC*, vol. 11434. Springer, 2018, pp. 447–452.