From knowledge discovery to knowledge creation: How can literature-based discovery accelerate progress in science?

N.R. Smalheiser, University of Illinois at Chicago, United States

G. Hahn-Powell, University of Arizona, United States

D. Hristovski, University of Ljubljana, Slovenia

Y. Sebastian, Charles Darwin University, Australia

Introduction

This essay gives an overview and describes prospects for generating new scientific knowledge from disparate datasets, as viewed by four active practitioners from around the globe (Illinois, Arizona, Slovenia and Australia). Although artificial intelligence (AI) and machine learning (ML) are central techniques employed in the field, the key concepts in this essay are undiscovered public knowledge (UPK) and literature-based discovery (LBD). These comprise a variety of situations, including some not yet tackled via ML.

UPK was originally coined by Swanson (1986) and expanded by Davies (1989). It suggests that scientific findings, hypotheses and assertions may exist within the published literature without anyone being aware of them. They may be undiscovered because no one alive has read the articles (e.g. they were published in obscure journals or lack Internet indexing). In other cases, different snippets of evidence or assertions may be scattered among multiple documents and need to be pieced together. For example, one article may raise a hypothesis that is tested in another, without any one individual being aware the two are related. As another example, multiple types of evidence may exist across different studies that address the same issue but are not integrated readily with each other (e.g. epidemiologic study vs. case reports); this is in contrast to meta-analyses, which attempt to collate comparable studies.

LBD generally refers to the fascinating possibility that one can create entirely new, plausible and scientifically non-trivial hypotheses by combining findings or assertions across multiple documents. If one article asserts that “A affects B” and another that “B affects C”, then “A affects C” is a natural hypothesis. The potential number of such transitive relations in the literature is astronomical. Thus, the LBD problem is to filter or identify which assertions of the type “A affects C” are novel, scientifically plausible, non-trivial and sufficiently interesting that a scientist would find them worthy of study. LBD differs from AI data mining efforts such as Knowledge Discovery from Databases that use statistics and interestingness[1] measures to

identify explicitly stated findings or significant associative trends in the data. In contrast, LBD attempts to identify *unknown* knowledge that is implicitly rather than explicitly stated.

Advances in AI, ML and computational linguistics are key to improving such systems. For example, better extraction of entities and relations, and better natural language inference and causality models, will improve the precision of "A affects B" and "B affects C" assertions. This, in turn, will greatly help assess whether a potential link can be explained in terms of known mechanisms. Advances in machine reading (teaching computers to read and comprehend natural language text), and especially "deep learning" neural network architectures applied to text, show great potential in identifying assertions in scientific articles and implicit relationships.

What LBD tools are available?

The first computer-assisted tools for carrying out LBD analyses were the following:

Arrowsmith 1-node and 2-node search tools (http://arrowsmith.psych.uic.edu) (Swanson and Smalheiser, 1997; Torvik and Smalheiser 2007). In the 2-node search, users define two sets of biomedical articles (hereby termed literature sets A and C) by carrying out two searches within the PubMed search engine. The Arrowsmith software then identifies title words from both literature sets to identify one or more connecting terms/phrases ($B_{i=1, 2, 3, ...}$) in common. These phrases are then ranked according to their predicted relevance for linking A and C in a meaningful manner. For each connecting term B_i, the system displays the instances of B_i in the A literature next to instances of B_i in the C literature, making it easy to see if there is an interesting A - B_i - C relationship. In the 1-node search, the user defines a single literature A that studies a given problem (e.g. Alzheimer's disease). The system then identifies disparate literatures C_i ranked by how many intermediate terms or concepts they share with A.

BITOLA (https://ibmi.mf.uni-lj.si/en/node/253) is based on co-occurrences of medical subject headings integrated with genetic background knowledge. This makes it especially useful for identifying candidate genes (Hristovski, 2005).

SemBT (http://sembt.mf.uni-lj.si) uses semantic relations extracted from the biomedical literature combined with microarray results (microarrays are used in laboratory settings to detect simultaneous expressions of thousands of genes). LBD in this case can be used both for microarray results interpretation (Hristovski, 2009) and for drug repurposing.

Mine the Gap! (accessible in https://h2020-minethegap.eu/) is a variant approach in which the user specifies a given set of literature, whereupon the software identifies "gaps" within that field. These gaps are pairs of topics that separately are studied frequently within the field, yet have never been discussed in the same article in that field.

Influence Search provides direct and indirect search over a graph of influence relations mined from English and Portuguese scholarly documents indexed by PubMed and SciELO. Each edge is weighted by how frequently its corresponding relation is discussed across documents. It is also weighted as a measure of the certainty of the relationship based on the degree of hedging in its description (Hahn-Powell, Valenzuela-Escárcega and Surdeanu, 2017; Barbosa et al., 2019).

Finally, **Lion-LBD** (https://lbd.lionproject.net) looks for relationships among instances of diseases, genes, mutations, chemicals, cancer hallmarks and species mentioned within biomedical articles rather than among documents or sets of documents.

All of these systems are implemented as free, public biomedical web tools. As well, proprietary systems include IBM Watson for Drug Discovery and Biovista's Biolab Experiment Assistant.

New and emerging models of LBD

To date, most research on LBD has come from practitioners in computer science, information science and bioinformatics. It has largely dealt with methodological questions that employ the ABiC model. For example, should Bi terms be extracted from title, abstract, specific document sections or the full text of literatures A and C? Should Bi terms represent text or ontological concepts? How can LBD be modelled on knowledge graphs, for example, by predicting which unlinked nodes are likely to become connected in the future?

Emerging approaches extend the ABC model in various ways. For example, instead of A – Bi – C, one may wish to create longer paths or chains of assertions (A – B1 -B2 – B3 – C) bridging any two literatures or concepts (Hossain et al., 2012). Alternatively, instead of connecting textual artefacts (documents or concepts), one may envision connecting investigators to identify potential collaborators (or potential reviewers). “Dr Smith” and “Dr Jones”, for example, may not know each other or attend the same meetings. However, they may be implicitly linked if they published on similar topics or even co-authored with some of the same scientists. If they share certain common interests or attributes, they might be expected to collaborate fruitfully, perhaps synergistically, on a particular hypothesis or scientific problem.

Even more interesting is when the collaborators come from complementary domains. Recently, a semantics-based methodology for cross-domain collaboration recommendations has been proposed (Hristovski et al., 2015) and later implemented with a graph database (Hristovski et al., 2016). This methodology proposes not only pairs of potential collaborators but also an explanation for why such a collaboration makes sense. Another approach along these lines is to define research communities, looking for links across disparate fields of research (Hahn-Powell, 2018). This would identify knowledge gaps and key ideas that can bridge disciplines and foster the kind of collaboration that accelerates scientific progress.

Early LBD studies focused on identifying novel links that represent potential new hypotheses. However, it is increasingly clear that the real goal is not novelty per se but rather finding hypotheses that domain scientists will find interesting, non-trivial and worthy of further study. Assertions that represent small increments from current knowledge may be very likely to be true. For example, if dexamethasone helps patients with COVID-19, then similar steroids may help COVID-19 as well. Yet these assertions are the least surprising and, because they are obvious, perhaps the least interesting from the standpoint of investigators in the field. Thus, there is an apparent trade-off: the more divergent a predicted hypothesis is from current knowledge, the more surprising it is but (all things being equal) the least likely it is to be true. On the other hand, previously published findings that were neglected or apparently refuted using the methods available at the time may actually represent the raw material for new hypotheses and even new paradigms. This is especially true if one looks at them in the light of more recent findings and methods (Swanson, 2011; Smalheiser, 2013; Smalheiser and Gomes, 2014; Peng, Bonifield and Smalheiser, 2017).

The above suggests a need for “interestingness” measures that can automatically score and rank hypotheses in terms of their surprisingness and potential impact on science. These would help guide users to focus on those that have “bang for the buck”. While no such dataset exists, judgements on aspects of “interestingness” could be collected through user interaction with an information retrieval system. These judgements could then be used to train a personalised recommendation system that learns to combine features derived from knowledge graph structures with user profiles and behaviours (Zhao, Wu and Liu, 2016; Guo et al., 2020). Such a system could be continuously improved through a virtuous cycle of human-machine collaboration.

Future LBD systems may also need to consider radically new approaches in synthesising knowledge that assess multiple weak findings across disparate sources. For example, multiple medical case reports sometimes publish quite similar findings, albeit in different contexts (Smalheiser, Shao and Yu, 2015). Under this scenario, one cannot undertake a conventional meta-analysis. Yet, intuitively, the presence of multiple independent reports should point to a real signal among the “noise” of individual cases. In materials

sciences, scientific documents commonly report a limited number of material samples being synthesised and characterised from non-comparable experiments. Again, conventional meta-analysis is not appropriate. Instead, new ways of combining information across disparate contexts are needed (Tshitoyan et al., 2019; Szczypiński et al., 2021).

LBD can be fruitfully integrated with other AI methods, such as neural networks, to provide explanation capabilities. In Zhang et al. (2021), several methods for knowledge graph completion (link prediction) using neural networks were used for drug repurposing for COVID-19. The medical doctor responsible for the evaluation may not always find it easy to directly interpret the rationales behind the proposed drug repurposing. In these cases, LBD's 2-node searches (such as those provided by the Arrowsmith system) can be used to provide explanations for paths in the knowledge graph between the drugs and COVID-19.

How can informatics scientists best collaborate with bench scientists, especially in biology and medicine?

Many biomedical hypotheses emerging from LBD analyses have been published. The earliest examples suggested that magnesium supplementation could prevent or treat migraine headaches, and that fish oil could treat Raynaud's disease. Indeed, the entire field of drug repurposing owes its underlying strategy to LBD. For example, one may rank drugs according to whether they elicit changes in gene expression that occur in directions opposite to those that occur in a given disease. LBD or bioinformatics practitioners themselves carried out most of the published analyses. Roughly 25 specific hypotheses have appeared among the Swanson-Smalheiser group, the Hristovski-Rindflesch group, the Wren-Garner group and a few others. Some have been experimentally tested and confirmed. As well, several independent biomedical investigators have employed Arrowsmith software to generate and assess hypotheses related to their laboratory studies (Kell, 2009; Manev and Manev, 2010).

The problems that LBD tools are solving (generating potentially novel hypotheses) are inherently more difficult and specialised than searching the research literature (as done by PubMed and Google Scholar). This may partially explain their limited use by the biomedical community to date. As well, LBD tools may need to become more user-friendly, fast responding and interactive. Perhaps they could display only the few best hypotheses generated by these systems that need to be investigated and evaluated. The tools should also be able to explain why the proposed novel hypotheses are attractive. In other words, explainability is also essential for wider adoption. Finally, LBD tools need to be publicly accessible as web-based tools (not merely code archived on Github) that operate over a continuously updated document collection.

There are also social and organisational obstacles to wider LBD adoption by the biomedical community. Most LBD research is published and presented at venues attended by biomedical informaticians, who are not the real end-users. Conversely, the biomedical curriculum does not train students to search systematically for new hypotheses. Moreover, in general, many investigators have little expertise or formal training with computer programming, data provenance issues and so forth. Investigators design and conduct their experiments (or should do) in collaboration with statisticians. In the same way, LBD analyses should be undertaken in dialogue or partnership between biomedical end-users and informatics consultants in response to specific research questions. For example, what molecular pathways are most promising to study in Alzheimer's disease?

Extending LBD analyses beyond text

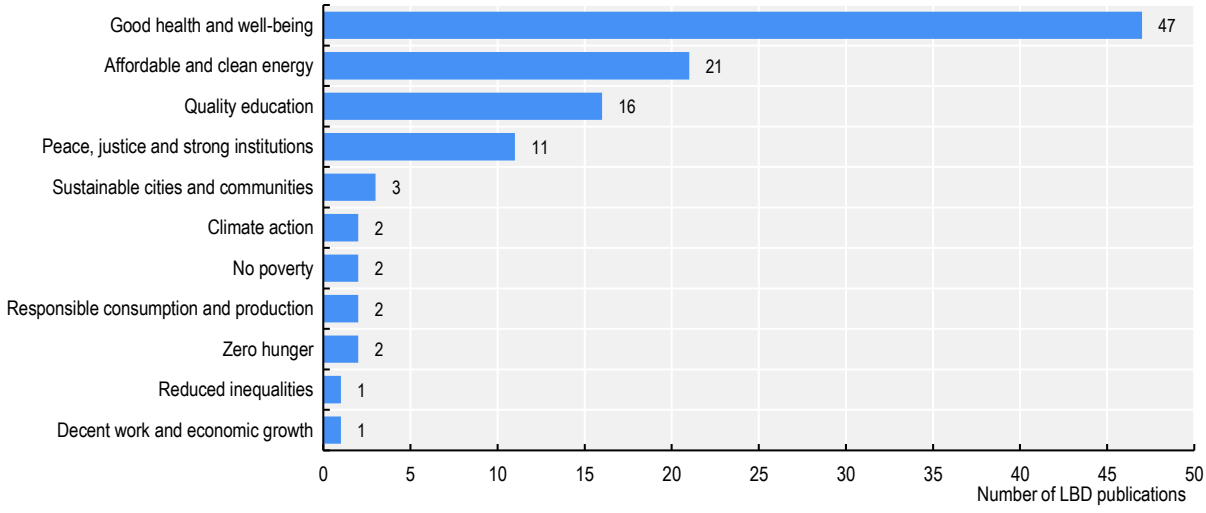
Besides linking assertions or findings in articles and other documents, the next-generation LBD systems are likely to use information in non-natural language forms. These could include numerical tables, charts

and figures, programming codes, microarrays, next-generation sequencing results, phenotypes, clinical data, etc. This is in parallel with the increasing awareness that different scientific fields communicate differently, implying diverse emphases on various formats of information (National Academies of Sciences, Engineering and Medicine, 2017). In fact, progress is being made in this direction that makes non-textual information more amenable to text mining (Pyarelal et al., 2020; Suadaa et al., 2021).

Prospects for LBD accelerating scientific progress outside biomedicine

An increasing number of LBD applications are being reported outside of biomedicine. In materials sciences, a group at the Lawrence Berkeley National Laboratory (Tshitoyan et al., 2019) recently demonstrated an LBD-style computer algorithm for discovering new materials. The algorithm uses static word embeddings to discover latent associations between an existing material (e.g. a crystal structure) and its previously unexplored thermoelectric applications. Word embeddings are vector representations of words built from millions of materials science publications. These vector representations can capture complex relationships among materials concepts without requiring explicit chemical knowledge to be specified *a priori* (e.g. periodic table). Using this method, computers can be used to automatically recommend new or existing materials for novel applications *long before* their discoveries. This saves money and time given that conventional materials engineering approaches typically rely on slow and arduous experimentations to discover or repurpose new materials (Szczypiński et al., 2021). As an example, researchers at MIT recently demonstrated the discovery time of new materials can be dramatically reduced from 50 years by conventional analytical methods to merely 5 weeks with the help of artificial neural networks (Janet et al., 2020).

Figure 1. The distribution of literature-based discovery research publications according to their alignment with selected UN Sustainable Development Goals (1989-2021)



Note: Bars indicate the number of publications containing the keyword "literature-based discovery" published between 1989 and 2021.
Source: www.dimensions.ai (accessed on 9 October 2021).

Figure 1 illustrates the far-reaching potentials of LBD in terms of the UN Sustainable Development Goals (SDGs). LBD researchers have previously attempted analyses on 10 of 17 goals. However, Figure 1 also points to a problem: less than 6% of all LBD publications (108 of 1 928) can be mapped to at least one SDG. Limitations of bibliographic indexing aside, this may suggest that the practicality of new LBD methods

and algorithms needs to be better contextualised within real-world problems (Mejia and Kajikawa, 2021). Doing so could help increase the uptake of LBD by the scientific and non-scientific community at large.

Future developments of AI-driven knowledge creation tools must be accompanied by the increasing availability of open research data. Platforms such as Figshare (https://figshare.com), Dryad (https://datadryad.org/stash) and Zenodo (https://zenodo.org) provide open access to research data as figures, datasets, images or videos. Cloud-based bibliography management solutions (Mendeley, Zotero) and academic social networking sites (ResearchGate, Academia.edu) could also open exciting possibilities for more author and community-centric LBDs. Finally, catalysts can also be found in public data initiatives such as The Australian Research Data Commons (https://researchdata.edu.au), the US Government's Open Data (https://www.data.gov) and the EU ORD Pilot (https://data.europa.eu).

Conclusion

UPK and LBD are simple, intuitive concepts that have profound implications for the philosophy and practice of science. Investigators now realise that publications are not simply archives of prior studies. They can also be a fertile raw material for making new and testable hypotheses that represent potential discoveries. LBD techniques work hand in hand with AI methods in machine learning, ontologies, knowledge graphs and computational linguistics, which are themselves making rapid progress. Thus, LBD analyses should continue to expand in biomedicine, the physical and social sciences, and even the humanities.

The greatest challenge is to integrate LBD analyses into real-life scientific workflows. There is no "killer app" akin to Google Scholar used by the general scientific community on a daily basis. Instead, tools are more specialised and require some training, not unlike the training required to use statistics packages or computer programming environments. Perhaps the best way forward is not to require bench and clinical investigators to become LBD experts themselves but rather to create partnerships and collaborations with informatics consultants fluent with LBD tools. One might also envision holding workshops and conferences that address specific problems (e.g. climate change) and carry out brainstorming in conjunction with domain experts assisted by LBD analyses. Maybe, in the not-so-distant future, AI software agents could serve the role of an intermediary between LBD tools and their intended users.

References

Barbosa G.C.G. et al. (2019), "Enabling search and collaborative assembly of causal interactions extracted from multilingual and multi-domain free text", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations),* Minneapolis, https://doi.org/10.18653/v1/N19-4003.

Davies, R. (1989), "The creation of new knowledge by information retrieval and classification", *Journal of Documentation*, Vol. 45/4, pp. 273-301, https://doi.org/10.1108/eb026846.

Guo, Q. et al. (2020), "A survey on knowledge graph-based recommender systems", *arXiv,* abs/2003.00911, https://doi.org/10.48550/arXiv.2003.00911.

Hahn-Powell, G. (2018), "Machine reading for scientific discovery", PhD dissertation, University of Arizona, Tucson, https://repository.arizona.edu/handle/10150/630562.

Hahn-Powell, G., M.A. Valenzuela-Escárcega and M. Surdeanu (2017), "Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph", in *Proceedings of ACL 2017, System Demonstrations*, Association for Computational Linguistics, Vancouver, https://doi.org/10.18653/v1/P17-4018.

Hristovski, D. et al. (2016), "Implementing semantics-based cross-domain collaboration recommendation in biomedicine with a graph database", in *Proceedings of the Eighth International Conference on Advances in Databases, Knowledge, and Data Applications*, Lisbon, https://www.iaria.org/conferences2016/DBKDA16.html.

Hristovski, D. et al. (2015), "Semantics-based cross-domain collaboration recommendation in the life sciences: Preliminary results", in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, Calgary, https://doi.org/10.1145/2808797.2809300.

Hristovski, D. et al. (2013), "Using literature-based discovery to identify novel therapeutic approaches", *Cardiovascular & Hematological Agents in Medicinal Chemistry,* Vol. 11/1, pp. 14-24, https://doi.org/10.2174/1871525711311010005.

Hristovski, D. et al. (2009), "Semantic relations for interpreting DNA microarray data" in *AMIA Annual Symposium Proceedings,* Vol. 255/9, American Medical Informatics Association, Rockville.

Hristovski, D. et al. (2005), "Using literature-based discovery to identify disease candidate genes", *International Journal of Medical Informatics*, Vol. 74/2-4, pp. 289-298, https://doi.org/10.1016/j.ijmedinf.2004.04.024.

Janet, J.P. et al. (2020), "Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization", *ACS Central Science,* Vol. 6/4, pp. 513-524, https://doi.org/10.1021/acscentsci.0c00026.

Kell D.B. (2009), "Iron behaving badly: Inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases", *BMC Medical Genomics* Vol. 2/2, https://doi.org/10.1186/1755-8794-2-2.

Manev, H. and R. Manev (2010), "Benefits of neuropsychiatric phenomics: Example of the 5-lipoxygenase-leptin-Alzheimer connection", *Cardiovascular Psychiatry and Neurology* 2010:838164, https://doi.org/10.1155/2010/838164.

Mejia, C. and Y. Kajikawa (2021), "Exploration of shared themes between food security and Internet of Things research through literature-based discovery", *Frontiers in Research Metrics and Analytics,* Vol. 6/25, https://doi.org/10.3389/frma.2021.652285.

National Academies of Sciences, Engineering, and Medicine (2017), *Communicating Science Effectively: A Research Agenda,* National Academies Press, Washington, DC, https://doi.org/10.17226/23674.

Peng, Y., G. Bonifield and N.R. Smalheiser (2017), "Gaps within the biomedical literature: Initial characterization and assessment of strategies for discovery", 22 May*, Frontiers in Research Metrics and Analytics,* https://doi.org/10.3389/frma.2017.00003.

Pyarelal, A. et al. (2020), "Automates: Automated model assembly from text, equations, and software", *arXiv,* arXiv:2001.07295, https://arxiv.org/abs/2001.07295v1.

Sebastian, Y., E.G. Siew and S.O. Orimaye (2017), "Emerging approaches in literature-based discovery: Techniques and performance review", *The Knowledge Engineering Review*, Vol. 32, p. e12, https://doi.org/10.1017/S0269888917000042.

Smalheiser, N.R. (2017), "Rediscovering Don Swanson: The past, present and future of literature-based discovery", *Journal of Data and Information Science*, Vol. 2/4, pp. 43-64, https://doi.org/10.1515/jdis-2017-0019.

Smalheiser, N.R. (2013), "How many scientists does it take to change a paradigm? New ideas to explain scientific observations are everywhere – we just need to learn how to see them", *EMBO Reports*, Vol. 14/10, pp. 861-865, https://doi.org/10.1038/embor.2013.125.

Smalheiser, N.R (2012), "Literature-based discovery: Beyond the ABCs", *Journal of the American Society for Information Science and Technology*, Vol. 63/2, pp. 218-224, https://doi.org/10.1002/asi.21599.

Smalheiser N.R. and O.L. Gomes (2014), “Mammalian Argonaute-DNA binding?”, *Biology Direct*, Vol. 10/27, https://doi.org/10.1186/s13062-014-0027-4.

Smalheiser, N.R., W. Shao and P.S. Yu (2015), “Nuggets: Findings shared in multiple clinical case reports”, *Journal of the Medical Library Association*, Vol. 103/4, pp. 171-176, https://doi.org/10.3163/1536-5050.103.4.002.

Suadaa, L.H. et al. (2021), "Towards table-to-text generation with numerical reasoning", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, on line, https://doi.org/10.18653/v1/2021.acl-long.115.

Swanson, D.R. (2011), “Literature-based resurrection of neglected medical discoveries”, *Journal of Biomedical Discovery and Collaboration*, Vol. 6, pp. 34-47, https://doi.org/10.5210/disco.v6i0.3515.

Swanson, D.R (1986), “Undiscovered public knowledge”, *The Library Quarterly: Information, Community, Policy,* Vol. 56/2, p. 103118, https://doi.org/10.1086/601720.

Swanson, D.R. and N.R. Smalheiser (1997), “An interactive system for finding complementary literatures: A stimulus to scientific discovery”, *Artificial Intelligence*, Vol. 91/2, pp. 183-203, https://doi.org/10.1016/S0004-3702(97)00008-8.

Szczypińsk, F.T. et al. (2021), “Can we predict materials that can be synthesised?”, *Chemical Science*, Vol. 12/3, pp. 830-840, https://doi.org/10.1039/D0SC04321D.

Torvik, V.I. and N.R. Smalheiser (2007), “A quantitative model for linking two disparate sets of articles in MEDLINE”, *Bioinformatics*, 1 July, Vol. 23/13, pp.1658-1665, https://doi.org/10.1093/bioinformatics/btm161.

Tshitoyan, V. et al. (2019), “Unsupervised word embeddings capture latent knowledge from materials science literature”, *Nature,* Vol. 571/7763, pp. 95-98, https://doi.org/10.1038/s41586-019-1335-8.

Zhang, R. et al. (2021), “Drug repurposing for COVID-19 via knowledge graph completion”, *Journal of Biomedical Informatics*, Vol. 115, p. 103696, https://doi.org/10.1016/j.jbi.2021.103696.

Zhao, W., R. Wu and H. Liu (2016), “Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target”, *Information Processing and Management,* Vol. 52/5, pp. 976-988, https://dl.acm.org/doi/abs/10.1016/j.ipm.2016.04.004.

Note

[1] “Interestingness” in data mining is a broad concept encompassing such ideas as reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability.