ADVANCED ENERGY MATERIALS
www.advenergymat.de

Construction and Application of Materials Knowledge Graph Based on Author Disambiguation: Revisiting the Evolution of $LiFePO_4$

Zhiwei Nie, Yuanji Liu, Luyi Yang, Shunning Li, and Feng Pan**

Due to the recent innovations in computer technology, the emerging field of materials informatics has now become a catalyst for a revolution of the research paradigm in materials science. Knowledge graphs, which provide support for knowledge management, are able to collectively capture the scientific knowledge from the vast collection of research articles and accomplish the automatic recognition of the relationships between entities. In this work, a materials knowledge graph, named MatKG, is constructed, which establishes a unique correspondence between subjects and objects in the materials science area. An emphasis is placed on the disambiguation of authors, addressed by a deduplication model based on machine learning and matching dependencies algorithms. Specifically, MatKG is applied to perform tracking on research trends in the study of $LiFePO_4$ and to automatically chronicle the milestones achieved so far. It is believed that MatKG can serve as a versatile research platform for amalgamating and refining the scientific knowledge of materials in a variety of subfields and intersectional domains.

Data mining of published works has emerged from the field of computer science and recently found its place in the research of materials science.[1] In contrast to the structural information that has been stored in several well-established materials databases,[2] the scientific knowledge scattered in publications in the form of text is still, up to now, gathered and analyzed manually by individual researchers, which is generally time-consuming and far from complete. Computer-automated analysis of materials science knowledge in the literature can help elucidate the collective association between different scientific articles, and would therefore greatly advance our macroscopic and comprehensive understanding of the evolution of scientific knowledge. "Knowledge graph", first proposed by Google in 2012 for search engines, is critical for the construction of such materials knowledge network.[3] However, the application of knowledge graph in material science is still in its infancy. As a structured semantic knowledge base for describing concepts and their physical relationships in a symbolic form, knowledge graph is expected to enable researchers to retrieve the interrelated data, such as the chemical names and their extracted features, in an automatic fashion.[4] More importantly, by integrating the historical track record of the deductive logic in past publications, knowledge graph may even facilitate future discovery of advanced materials that are hidden in the literature from different domains.[1b] Hence, we are optimistic that the application of knowledge graph techniques to materials science can open up new opportunities for fast extracting and handling of large-scale information for the materials research community.

In a knowledge graph, the constituent units are the "entity-relationship-entity" triads, in which the entity and its related attribute-value pairs are connected by relationships to form a network of knowledge structures. Accordingly, a knowledge graph could only be constructed on the premise that clear identification of both the subject (e.g., author) and the object (e.g., material) is achieved. **Figure 1** illustrates what a materials knowledge graph consists of. We would like to note that, due to the enormity and complexity of published literature, it is a laborious task to disambiguate the authors by manual methods. Especially, the information about the authors could change during their carrier. In the field of materials science, previous studies on the data mining of scientific literature mainly emphasized on the retrieval of information regarding the object,[1] while research in disambiguation of the subject is scarce. By harnessing the recent technological advancements in artificial intelligence and database technologies, we are in a better position to develop novel disambiguation method for the construction of a consolidated knowledge graph for materials science.

In this work, machine learning (ML) and matching dependencies (MD)[5] are combined to construct a deduplication model for the disambiguation of authors in scientific articles. Based on the literature database and taking $LiFePO_4$ as an example, a knowledge graph, which we call MatKG, is constructed and employed to shape the evolution of scientific knowledge on materials. The unique correspondence between subjects and objects is established for the first time in the field of materials science, thus permitting automatic tracking of research trends

Z. Nie, Y. Liu, Dr. L. Yang, Dr. S. Li, Prof. F. Pan
School of Advanced Materials
Peking University
Shenzhen Graduate School
Shenzhen 518055, P.R. China
E-mail: lisn@pku.edu.cn; panfeng@pkusz.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aenm.202003580.

DOI: 10.1002/aenm.202003580

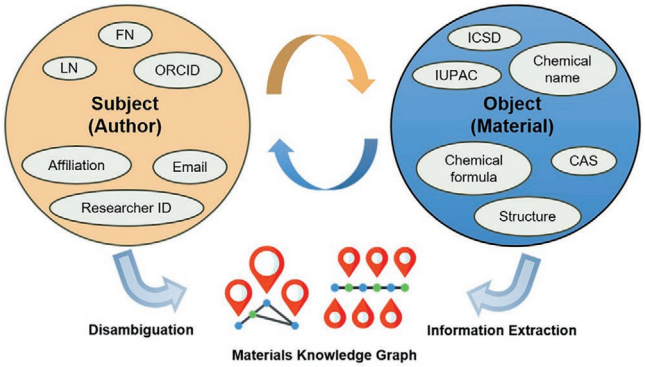

Figure 1. Structure of a materials knowledge graph.

and milestones. The proposed information-to-knowledge research route can not only expedite the development of known materials for different applications but also reform the path and paradigm of materials research which will accelerate the discovery of novel materials.

The construction of MatKG starts from data collection of research articles and author information. More than 2.9 million articles, as well as their author information, are collected in the field of materials science through the application programming interfaces of Elsevier's Scopus, Science Direct, and Web of Science. Over 1.05 million records of author information are gathered, including the author's first name (FN), last name (LN), Open Researcher and Contributor ID (ORCID), email, and affiliation. We also extract the semantic information from the titles, abstracts, and keywords of the articles for author deduplication.

It is worth noting that most of the author information is incomplete, which prevents us from reliably differentiating one author from the other. This leads to duplicates or ambiguity of the subject in the constructed knowledge graph. Here, we take the name "Jun Li" as an example. In **Table 1**, t_1, t_2, and t_3 are distinguished as three different authors in the database, due to the difference in either FN or affiliation. The acronym "J." cannot be automatically correlated with "Jun", but in fact, t_1 and t_2 represent the same author. Even more striking is the case where all the FNs, LNs, and affiliation names are different, as exemplified by t_4 and t_5 in Table 1, which can confuse a person without context information. By tracing the biography of the author known as "Barner-Kowollik", we can find that t_4 and t_5 correspond to the same person. Since the discrepancies in author information will result in redundancy and uncertainty of the database, the identification and fusion of database records that belong to the same author would be a prerequisite for constructing a materials knowledge graph. This data cleaning task can be done by a deduplication model incorporating the information in the titles, abstracts, and keywords of the corresponding articles.

Table 1. Examples of author records in the database.

	FN	LN	ORCID	Affiliation
t_1	J.	Li	Null	Shanghai Jiao Tong University
t_2	Jun	Li	Null	School of Materials Science and Engineering, Shanghai Jiao Tong University
t_3	Jun	Li	Null	Sichuan University
t_4	Christopher	Barner-Kowollik	0000-0002-6745-0570	Karlsruhe Institute of Technology
t_5	C.	Kowollik	Null	University of Göttingen

The flowchart of the construction of MatKG is presented in **Figure 2**. Four modules are designed in this framework: ML-based pre-training (module I), MD-based collective blocking (module II), ML-based classification (module III), and breadth-first search (module IV). Modules I, II, and III correspond to the deduplication process, in which the integration of ML and MD methods can enable fast and highly-accurate detection of the duplicate records.

In module I, a widely-used ML approach, Naïve Bayes Classifier, is employed for pre-training the original records. The algorithm is based on Bayes' theorem with the assumption that all attributes are independent given the value of the class variable. It is more suitable than other common ML classification algorithms when dealing with text information (Figure S1, Supporting Information).[6] In this stage, the authors are classified into several major areas according to the information in the abstract. Only the first 60 words instead of the full text of the abstract are taken into account, which can result in better classification performance (Figure S2, Supporting Information).[7]

In module II, an MD-based collective blocking is performed to split the database records into blocks. MD plays the role of a declarative logical rule for evaluating the similarity between two records, not only by their own information (e.g., FN, LN) but also by the semantic relational information contained in the titles and keywords of the articles. It means that if two article records are placed in the same block, the corresponding records of authors with similar names should be stored in the same block as well. Then, we only need to compare the records in the same block during the following deduplication, while any two records in different blocks are regarded as non-duplicates.

In module III, Naïve Bayes Classifier is utilized again for duplication detection in each block, with the information in titles, abstracts, and keywords taken into consideration. The probability of two author records referring to the same person is computed, and a threshold of 90% is set, above which the pair of records are merged into a single one. This module produces a duplicate-free database consisting of around 0.63 million records of author information.

The efficiency of a knowledge graph depends on the speed of resource discovery, which can be tackled by a breadth-first search algorithm (module IV). CTANE,[8] a level-wise algorithm based on the well-known functional dependencies mining method TANE,[9] can effectively accomplish the search task. In order to reduce the time complexity and cope with the increasing amount of data, a pruning algorithm is carried out during the CTANE. The candidate set, that is, the search space, is pruned, which yields a large leap in the search speed. The improved algorithm is named Pruned-CTANE. After the search procedure, a knowledge graph is constructed, with the authors and the research fields constituting the logical starting points for the inquiry into the literature.

The advantages of pre-training and pruning strategies in search speed and information retrieval quality are demonstrated in **Figure 3**. To test the search speed, four different

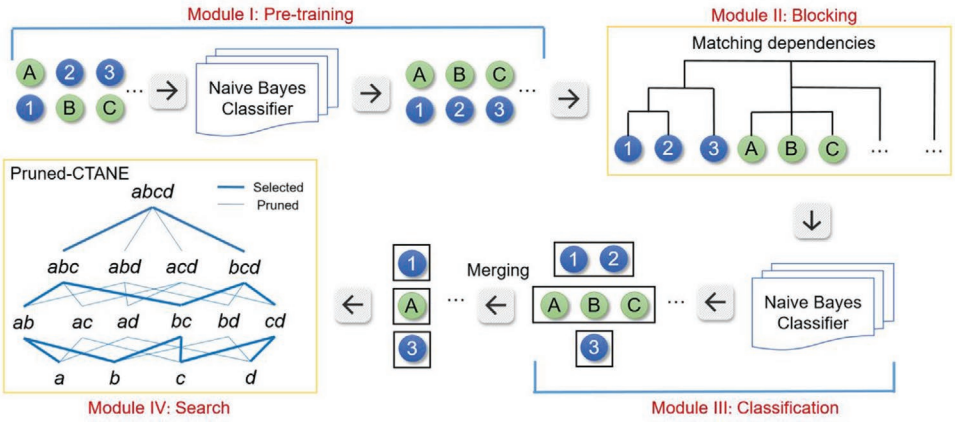

Figure 2. The flowchart of the construction of MatKG.

frequency situations are examined (Figure 3a). Here, the frequency is in inverse proportion to the number of candidate sets in each layer of the search process. The smaller the frequency, the larger the number of candidate sets in each layer will be. It is shown that for all frequencies, the pruning strategy can shorten the search time by more than two-thirds when taking the traditional CTANE algorithm as a reference. On the other hand, the quality of information retrieval can be evaluated by metrics including precision and recall, which are shown in Figure 3b. Cora dataset[10] is used for test, and the calculation details are given in Supporting Information. Higher precision means that more candidate duplicate record-pairs are true, while higher recall means that more true candidate record-pairs have been found. We can see that a single ML or MD model with both precision and recall above 50% is practically beyond attainment. Apparently, the method using either ML or MD alone could hardly meet the requirements for author disambiguation. The ERBlox project,[11] a recently proposed technique combining ML and MD for entity resolution, performs quite well but at a higher cost of computational time and resource than the disambiguation model in this work. It is worth emphasizing that the performance of MatKG, with a precision of 89% and a recall of 93%, is even better than that of ERBlox. Therefore, the proposed pre-training and pruning strategies can not only dramatically outperform the traditional methods in accelerating the search process, but also accurately capture the subject–object relationship that guarantees the successful construction of materials knowledge graph in the present work.

We take $LiFePO_4$ as a case example to illustrate how the MatKG is employed for mapping the evolution and development of $LiFePO_4$ in the field of lithium-ion batteries (LIBs). In 1997, John B. Goodenough pointed out in his article that the olivine phosphate, $LiFePO_4$, can be used as a cathode material for rechargeable LIBs.[12] Since then, numerous researchers have contributed to the study of $LiFePO_4$ due to its stable cycle performance, low cost, resource abundance, and environmental

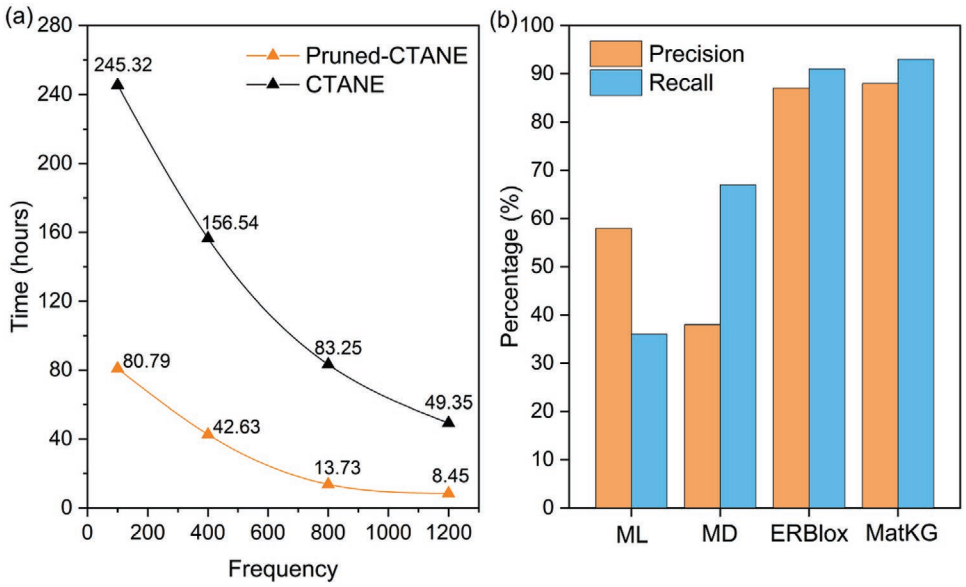

Figure 3. The evaluation of search speed and information retrieval quality for MatKG. a) Comparison between Pruned-CTANE and CTANE algorithms on search time. b) Comparison of precision and recall among single ML model, single MD model, ERBlox, and MatKG using Cora dataset.

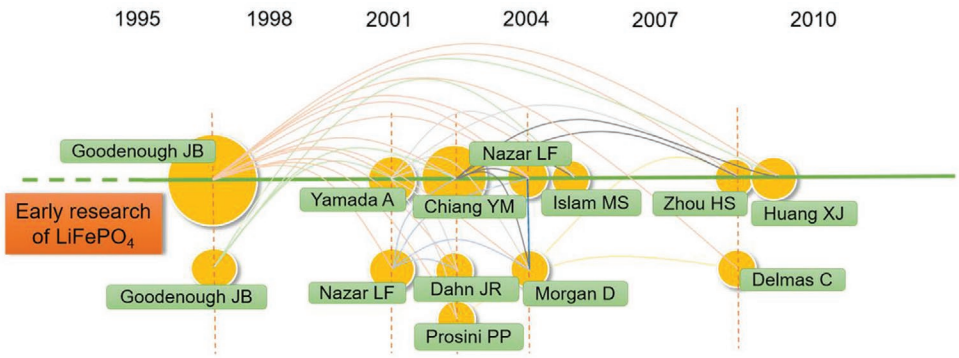

Figure 4. The milestone map of $LiFePO_4$ development for LIBs.

benignity. Building on the efforts made in the past decades, $LiFePO_4$ cathodes have been successfully commercialized and become one of the most reliable cathodes regarding safety issues. Using MatKG, we can unambiguously list the authors belonging to this field and associate the related articles to each author in an automatic framework.

First, the word "$LiFePO_4$" and other names such as "LFP" and "lithium iron phosphate" are selected as the starting points for inquiry. We notice that most of the early works on $LiFePO_4$ focused on the crystal structure[13] and physicochemical properties, including the magnetism,[14] magnetoelectric effect,[15] vibrational interactions,[16] and electron density distribution.[17] By further appending the search term "LIBs" as well as its equivalents into the inquiry, the search results would be more relevant to $LiFePO_4$ as a cathode material for LIBs. Then, with the established knowledge graph, the authors are accurately distinguished according to the interlinked articles. With the aid of the citation information in the article database, a milestone map for $LiFePO_4$ research and development is obtained, as depicted in **Figure 4**.

It is noteworthy that the corresponding authors of the research articles are generally the main contributor, and therefore they are chosen as the representative of the authors for each article. In Figure 4, a timeline is drawn, along which the size of the circles denotes the impact of the corresponding articles on the works of other authors. Larger circle would suggest a greater contribution of the author to the research trend in this field. For clarity and ease of presentation, only articles with over 600 citations are shown. We find that the first two works come from A. K. Padhi in 1997,[12,18] but after a careful inspection, we realize that these works are actually from John B. Goodenough mentioned above. Therefore, they are isolated cases where the corresponding author is not selected as the representative. Moreover, it is noticed that the author LN "Goodenough" appears in other research fields, and there are several affiliations associated with this LN, including University of Texas at Austin, University of Reading, University of Oxford, Massachusetts Institute of Technology, etc. MatKG can efficiently distinguish the authors with the same LN and merge all the duplicate information, as shown in **Table 2**. In Table 2, s_1–s_5 represent five author records in the database, and "Group A/B/C" represents the classification results. The identification of authors tabulated in Table 1 is also solved. These results allow us to employ this knowledge graph for the automatic examination of correlations between scientific advancements.

Table 2. Identification of "John B. Goodenough" from the author records with an LN of "Goodenough".

	Author name	Affiliation	Group
s_1	John B. Goodenough	University of Texas at Austin	A
s_2	J. B. Goodenough	Massachusetts Institute of Technology	A
s_3	John Goodenough	University of Oxford	A
s_4	Daniel A. Goodenough	Harvard Medical School	B
s_5	Peter W. Goodenough	University of Reading	C

In the MatKG graph, the accessed information of research discoveries in $LiFePO_4$ is summarized as follows: The research in discharge capacity, electrode preparation, and Li intercalation mechanism of $LiFePO_4$ emerged from 1997.[12,18] In 2001, a specific capacity approaching the theoretical value was achieved at room temperature, and $LiFePO_4$/C composite was found to exhibit very good rate capability, excellent stability, and high theoretical capacity.[19] In 2002, the influence of carbon content in $LiFePO_4$/C composite electrode was evaluated,[20] while other studies focused on Li intercalation mechanism of $LiFePO_4$ and the electronic conductivity.[21] In 2004, researchers focused more on improving the electronic conductivity.[22] Later on, atomic-scale investigation and nanostructure design were popular.[23] It can be seen that the scientific knowledge in a subfield that was only known by specialized experts in the past, can be programmatically codified via materials informatics, thus facilitating rapid identification of research trends. In this regard, we envision MatKG to be a promising platform for amalgamating and refining the scientific knowledge in various materials science domains.

In summary, we have shown that by leveraging the knowledge from data mining and artificial intelligence, it is possible to construct a materials knowledge graph as an infrastructure for easy access to research findings in materials science. We introduce MatKG as a research platform for systematically codifying the scientific knowledge documented in research publications. A million-scale database containing research articles and author information in the field of materials science is implemented. Author disambiguation, as one of the major obstacles in the construction of knowledge graph, has been addressed through a combination of ML and MD models. The utility of MatKG is exemplified by the automatic generation of a milestone map for the investigation of $LiFePO_4$, which can enable rapid and unbiased tracking of the current trends in the academia for

researchers in different fields. We anticipate that a predictive system based on MatKG can be obtained by incorporating more information on materials properties, and this functionality will open up a new paradigm of materials discovery and design.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was financially supported by National Key R&D Program of China (2016YFB0700600), the Chemistry and Chemical Engineering Guangdong Laboratory (Grant No.1922018), and the Shenzhen Science and Technology Research Grant (No. JCYJ20200109140416788).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Research data are not shared.

Keywords

lithium iron phosphate, machine learning, matching dependencies, materials knowledge graphs

Received: November 15, 2020
Revised: February 12, 2021
Published online: March 11, 2021

[1] a) P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73; b) V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Q. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95; c) S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, L. Cronin, *Science* **2020**, *370*, 101; d) D. Mrdjenovich, M. K. Horton, J. H. Montoya, C. M. Legaspi, S. Dwaraknath, V. Tshitoyan, A. Jain, K. A. Persson, *Matter* **2020**, *2*, 464; e) L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, A. Jain, *J. Chem. Inf. Model.* **2019**, *59*, 3692.

[2] a) A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2011**, *50*, 2295; b) S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, *58*, 218; c) J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*, 1501; d) J. Jie, M. Weng, S. Li, D. Chen, S. Li, W. Xiao, J. Zheng, F. Pan, L. Wang, *Sci. China: Technol. Sci.* **2019**, *62*, 1423; e) M. Weng, Z. Wang, G. Qian, Y. Ye, Z. Chen, X. Chen, S. Zheng, F. Pan, *Sci. China: Chem.* **2019**, *62*, 982.

[3] S. A, Official Google Blog: Introducing the Knowledge Graph: things, not strings, https://blog.google/products/search/introducing-knowledge-graph-things-not/, (accessed: May 2012).

[4] a) Y. Shen, K. Q. Yuan, J. C. Dai, B. Z. Tang, M. Yang, K. Lei, *J. Med. Syst.* **2019**, *43*, 9; b) H. Paulheim, *Semant. Web* **2017**, *8*, 489.

[5] a) S. Song, L. Chen, presented at Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, November, **2009**; b) L. Bertossi, S. Kolahi, L. V. S. Lakshmanan, *Theor. Comput. Syst.* **2013**, *52*, 441; c) J. Jie, Z. Hu, G. Qian, M. Weng, S. Li, S. Li, M. Hu, D. Chen, W. Xiao, J. Zheng, L.-W. Wang, F. Pan, *Sci. Bull.* **2019**, *64*, 612.

[6] a) C. C. Aggarwal, C. Zhai, *A Survey of Text Classification Algorithms*, Springer US, Boston, MA **2012**; b) T.-S. Lim, W.-Y. Loh, Y.-S. Shih, *Machine Learning* **2000**, *40*, 203; c) M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, *J. Mach. Learn. Res.* **2014**, *15*, 3133; d) T. Bayes, *Philos. Trans. R. Soc. London* **1763**, *53*, 370; e) I. Rish, presented at *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, Seattle, WA, August **2001**; f) S. R. Gunn, *ISIS Tech. Rep.* **1998**, *14*, 5; g) T. Joachims, presented at *Proc. of the 24th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New Orleans, LA **2001**; h) T. Joachims, *Learning to Classify Text Using Support Vector Machines*, Springer Science & Business Media, Berlin, Germany **2002**; i) S. R. Safavian, D. Landgrebe, *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660; j) P. H. Swain, H. Hauska, *IEEE Transactions on Geoscience Electronics* **1977**, *15*, 142.

[7] a) K.-M. Schneider, presented at *Int. Conf. on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, February **2005**; b) S.-B. Kim, K.-S. Han, H.-C. Rim, S. H. Myaeng, *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1457; c) J. Chen, H. Huang, S. Tian, Y. Qu, *Expert Syst. Appl.* **2009**, *36*, 5432.

[8] W. Fan, F. Geerts, J. Li, M. Xiong, *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 683.

[9] Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen, *Comput. J.* **1999**, *42*, 100.

[10] A. K. McCallum, K. Nigam, J. Rennie, K. Seymore, *Inf. Retr. J.* **2000**, *3*, 127.

[11] Z. Bahmani, L. Bertossi, N. Vasiloglou, *Int. J. Approx. Reasoning* **2017**, *83*, 118.

[12] A. K. Padhi, K. S. Nanjundaswamy, J. B. Goodenough, *J. Electrochem. Soc.* **1997**, *144*, 1188.

[13] a) O. V. Yakubovich, M. A. Simonov, N. V. Belov, *Sov. Phys. – Doklady* **1977**, *22*, 347; b) E. A. Genkina, Y. K. Kabalov, B. A. Maksimov, O. K. Mel'nikov, *Sov. Phys. – Crystallogr.* **1984**, *29*, 27.

[14] a) R. P. Santoro, R. E. Newnham, *Acta Crystallogr.* **1967**, *22*, 344; b) J. G. Creer, G. J. Troup, *Phys. Lett. A* **1970**, *32*, 439.

[15] M. Mercier, P. Bauer, B. Fouilleux, *C. R. Hebd. Seances Acad. Sci., Ser. B* **1968**, *267*, 1345.

[16] M. T. Paques-Ledent, P. Tarte, *Spectrochim. Acta, Part A* **1974**, *30*, 673.

[17] V. A. Streltsov, E. Belokoneva, V. Tsirelson, N. Hansen, *Acta Crystallogr. B* **1993**, *49*, 147.

[18] A. K. Padhi, K. S. Nanjundaswamy, C. Masquelier, S. Okada, J. B. Goodenough, *J. Electrochem. Soc.* **1997**, *144*, 1609.

[19] a) H. Huang, S. C. Yin, L. F. Nazar, *Electrochem. Solid-State Lett.* **2001**, *4*, A170; b) A. S. Andersson, J. O. Thomas, *J. Power Sources* **2001**, *97–98*, 498; c) A. Yamada, S. C. Chung, K. Hinokuma, *J. Electrochem. Soc.* **2001**, *148*, A224.

[20] Z. Chen, J. R. Dahn, *J. Electrochem. Soc.* **2002**, *149*, A1184.

[21] a) S.-Y. Chung, J. T. Bloking, Y.-M. Chiang, *Nat. Mater.* **2002**, *1*, 123; b) P. P. Prosini, M. Lisi, D. Zane, M. Pasquali, *Solid State Ionics* **2002**, *148*, 45.

[22] a) P. S. Herle, B. Ellis, N. Coombs, L. F. Nazar, *Nat. Mater.* **2004**, *3*, 147; b) D. Morgan, A. Van der Ven, G. Ceder, *Electrochem. Solid-State Lett.* **2004**, *7*, A30.

[23] a) M. S. Islam, D. J. Driscoll, C. A. J. Fisher, P. R. Slater, *Chem. Mater.* **2005**, *17*, 5085; b) Y. Wang, Y. Wang, E. Hosono, K. Wang, H. Zhou, *Angew. Chem., Int. Ed.* **2008**, *47*, 7461; c) C. Delmas, M. Maccario, L. Croguennec, F. Le Cras, F. Weill, *Nat. Mater.* **2008**, *7*, 665.