# Knowledge mining: A cross-disciplinary survey

August 14 2023

| Metric name | Metric expression |
|---|---|
| All_confidence | $\min\left(p\left(A\middle\vert B\right),\ p\left(B\middle\vert A\right)\right)$ |
| Max_confidence | $\max\left(p\left(A\middle\vert B\right), p\left(B\middle\vert A\right)\right)$ |
| Kulczynski | $\frac{1}{2}\left(p\left(A\middle\vert B\right)\ +\ p\left(B\middle\vert A\right)\right)$ |
| Cosine | $p\left(A,B\right)/\sqrt{p(A)p(B)}$ |
| Jaccard | $p\left(A,B\right)/\left(p\left(A\right)+p\left(B\right)-p\left(A,B\right)\right)$ |

Some of the most basic metrics to evaluate the interestingness of association rules. Credit: *Machine Intelligence Research* (2022). DOI: 10.1007/s11633-022-1323-6

Knowledge mining is a widely active research area across disciplines such as natural language processing (NLP), data mining (DM), and machine learning (ML). The overall objective of extracting knowledge from data source is to create a structured representation that allows researchers to better understand such data and operate upon it to build applications.

Each mentioned discipline has come up with an ample body of research, proposing different methods that can be applied to different data types. A significant number of surveys have been carried out to summarize research works in each discipline. However, no survey has presented a cross-disciplinary review where traits from different fields were exposed to further stimulate research ideas and to try to build bridges among these fields. In this work published on Machine Intelligence Research, the researchers present such a survey.

Automatic extraction of knowledge from diverse sources of data is a challenging task across different fields. For example, in natural language processing (NLP), research on the extraction of structured knowledge bases from natural language text has received much attention due to its applications.

In data mining (DM), a wide area of research has focused on mining rules from structured databases that can help people discover novel associations between items or features and make decisions in diverse contexts such as business or education.

Furthermore, in the field of machine learning (ML), plenty of effort has been advocated towards extracting knowledge, mainly in the form of logic rules, from both machine learning system's predictions and parameters in order to build an interpretable representation that helps to explain the system's decisions (the so-called interpretability problem); a scenario highly sought in medicine, for example.

Extracting or mining knowledge from data (be it unstructured, structured, or behavioral data) is an open problem that has been tackled across different research fields. This wide scenario has not only led to different definitions and ways to represent the construct of knowledge (and consequently, to define the task of knowledge mining), but it has also resulted in diverse research perspectives, which seem to use

different methodologies to extract knowledge and different metrics to evaluate the consistency of the knowledge extracted.

On the other hand, in the NLP field, a knowledge base is usually represented as a tensor structure where each entry usually corresponds to a probabilistic assignment of the belief of a fact.

Finally, in the field of machine learning, the problem of knowledge mining has been motivated by the problem of trying to understand and validate ML systems which due to their complexity are not easy to be inspected manually. Similarly, the choice of the representation of knowledge has been constrained to be understandable by humans, where a widely common and accepted representation in this area are logic rules.

From this brief overview of knowledge mining across fields, it can be observed that the diversity of objectives and constructs and the wide scenario researchers claimed at the beginning, which leads them to the questions: How is knowledge mining characterized across research fields? What are their proposed approaches and shared traits? And how can researchers consolidate them?

Researchers note that while there are already several in-depth surveys in the literature of each field showing the methods and algorithms to extract knowledge, it is supposed that there is no survey that jointly traverses these research areas to answer the above questions.

Furthermore, the importance of mining knowledge has permeated different fields and has also impacted the industry. Therefore, researchers believe that a cross-disciplinary literature review, in a landscape-oriented approach, that encompasses all these varying degrees of freedom underlying the problem of mining knowledge from data is on the call.

In this paper, rather than surveying a plethora of methods and previous works across these three research areas, researchers intend to overview the nuances, and attached idiosyncrasies, of the approaches taken to extract knowledge from a target data source.

Hence, this paper advocates for an additive overview of the problem of extracting knowledge across the fields of natural language processing, data mining and machine learning to show their key objectives, methods, and evaluations, and how some previous works have made links among these areas for the task of knowledge mining.

The final aim of this paper is to stimulate and provoke new ideas and research agendas among researchers from the different disciplines so that new bridges among the areas surveyed can emerge to further advance in the task of knowledge mining. Following this approach, researchers avoid providing a single definition of knowledge and knowledge mining, and rather present how these constructs have been embraced across fields. Thus, researchers depart from a common starting point across fields. They fix the choice of knowledge representation to that of logic, or logic-like formulas, which is a representation highly used across these fields.

Based on this knowledge representation, in Sections 2–4, researchers walk through the different goals and key approaches of each field, in a problem-oriented perspective, to gain a refined insight into how knowledge mining is embodied and what traits they find in these research areas. Section 2 is about knowledge extraction from natural language text which include six parts.

Firstly, researchers provide preliminaries of state-of-the-art methods and models in NLP. Secondly, they introduce the most common learning approaches for information extraction, namely supervised learning (classification and sequence labeling), distant supervised learning, and

unsupervised learning.

Then, they provide an account of the two IE problems that have received much attention in the NLP community, namely named entity recognition in part 3 and relation extraction in part 4, as well as the methods to evaluate how well an NLP system performs at any of these tasks in part 5. Finally, researchers review some current challenges in NLP related to the problem of IE in the last part.

Section 3 is about knowledge mining from transactional databases. It consists of four parts: Part 1 surveys some of the main approaches to the problem of frequent itemset generation. Part 2 refers to association rule mining. Part 3 shows methods for pruning and evaluating candidate rules and part 4 is about current challenges.

Section 4 is about knowledge extraction from machine learning systems. In this section, researchers present different approaches to extracting the knowledge learned by complex ML systems, also known as black-box systems, due to their un-interpretability.

Similar to previous sections, researchers mainly target works in the literature where the knowledge extracted is in the form of logic rules (this is one of the most popular types of knowledge representation in the interpretability literature). Most of the black-box systems they review in this section are neural networks due to their wide acceptance and use in ML and related fields.

Finally, in Section 5, researchers firstly identify five dimensions that they believe characterize the knowledge extraction work across fields, namely objectives, methods, research orientation, data, and evaluations. In what follows, they provide a comparison of the knowledge mining problem for the NLP, DM, and ML fields across these five traits.

Finally, they provide what they believe to be a long-term research direction for knowledge mining. Researchers believe this paper will contribute to creating future research directions for the task of knowledge mining that encompass the three, so far unlinked, research areas of NLP, DM, and ML.

**More information:** Yong Rui et al, Knowledge Mining: A Cross-disciplinary Survey, *Machine Intelligence Research* (2022). DOI: 10.1007/s11633-022-1323-6

Provided by Beijing Zhongke Journal Publising Co.