Technical Framework and Research Advances in Knowledge Discovery from Scholarly Figures (Postprint)

Authors: Ding Pei

Date: 2023-04-01T16:02:59+00:00

Abstract

[Purpose/Significance] Against the backdrop of deep integration of scientific and technological resources, academic chart knowledge discovery provides a novel knowledge discovery approach beyond text-based knowledge discovery, constitutes a crucial component in refining literature knowledge discovery, can enhance researchers' efficiency in scientific discovery and knowledge creation, and promotes the upgrading of knowledge services in digital libraries. [Method/Process] This paper systematically reviews the evolutionary trajectory of academic chart knowledge discovery, elaborately demonstrates the contents of its technical framework, and substantiates that the technology for academic chart knowledge discovery is gradually maturing. By examining application services of academic chart knowledge discovery, it is argued that academic chart knowledge discovery possesses broad application prospects across multiple dimensions of technological innovation. [Results/Conclusion] Looking forward to the future of academic chart knowledge discovery, we need to: prioritize academic chart knowledge discovery and integrate it within the literature knowledge discovery system; refine the semantic knowledge organization system for academic charts and construct specialized semantic knowledge bases for academic charts; and develop novel applications for academic chart knowledge discovery.

Full Text

Preamble

Volume 65, Issue 23, December 2021

The Technical Framework and Research Progress of Knowledge Discovery in Academic Figures and Tables

Ding Pei, Shenzhen University Library, Shenzhen 518060

Abstract:
[**Purpose/Significance**] Against the backdrop of deep integration of scientific and technological resources, knowledge discovery in academic figures and tables provides a new approach to knowledge discovery beyond textual knowledge discovery. It represents a crucial component in perfecting document-based knowledge discovery, enhancing researchers' efficiency in scientific discovery and knowledge creation, and promoting the upgrading of knowledge services in digital libraries. [**Method/Process**] This paper traces the evolutionary trajectory of knowledge discovery in academic figures and tables, elaborates on its technical framework, and demonstrates the gradual maturation of related technologies. Combined with application services, it demonstrates that knowledge discovery in academic figures and tables has broad application prospects across multiple facets of scientific and technological innovation. [**Result/Conclusion**] Looking ahead, we need to: prioritize knowledge discovery in academic figures and tables and integrate it into the literature knowledge discovery system; improve the semantic knowledge organization system for academic figures and tables and construct specialized semantic knowledge bases; and develop novel knowledge discovery applications for academic figures and tables.

Keywords: academic figures and tables; knowledge discovery; knowledge organization; information extraction
Classification Number: G254
DOI: 10.13266/j.issn.0252-3116.2021.23.015

In the context of deep integration of scientific and technological information resources, a new data-intensive scientific discovery paradigm has emerged as an innovation ecosystem. Breakthroughs in artificial intelligence and deep learning technologies have brought new transformations and requirements to knowledge discovery services that support this ecosystem. In the digital library domain, knowledge discovery services centered on literature-based knowledge discovery are increasingly characterized by multi-source heterogeneous objects, fine-grained content organization, cross-type semantic associations, machine understandability, and machine-driven knowledge discovery. Traditional knowledge discovery centered on academic texts faces challenges from heterogeneous carriers and new service demands. Academic figures and tables are digital objects used in scientific literature for content description, argument support, and data comparison. N. Siegel's analysis of arXiv and PubMed revealed that only 20% of PDF papers in arXiv lacked relevant figures and tables, while only 10% of XML files in PubMed were without them [1]. In the biomedical domain, nearly every journal article contains academic figures and tables, which represent evidence-based content more effectively than any other information type [2]. Research has shown that academic figures and tables provide more information than text alone and can effectively improve users' efficiency in discovering literature [3]. P. Lee found that more influential papers tend to contain more academic figures and tables [4]. Academic figures and tables support research

reuse, explain important research content, and serve as critical carriers of scientific knowledge at the intersection of scientific literature and data resources.

However, due to factors such as diverse formats and complex information extraction, machine understanding of academic figures and tables remains at a weak semantic level, preventing their effective integration into existing literature knowledge discovery systems. Future academic knowledge service systems require fine-grained knowledge organization, semantic-based knowledge association, knowledge discovery for all resource types, and cognitive computing that effectively supports intelligent Q&A and precise intent representation. As typical heterogeneous academic objects, research on knowledge discovery in academic figures and tables is both necessary and urgent for improving the literature knowledge discovery system, promoting deep integration of scientific resources, advancing knowledge discovery for non-textual data, and innovating digital library knowledge services.

This study used "image/table," "information extraction," and "scientific literature/paper" as core search terms, expanding to related concepts such as "image recognition/table recognition," "image annotation/table annotation," "knowledge discovery," "named entity recognition," and "figure/table relationship extraction." Searches were conducted in Web of Science, Scopus, and CNKI databases with a cutoff date of August 2021. Based on abstract reading, 85 highly relevant papers were identified, and an additional 135 relevant papers were added through reference expansion, forming the foundation of this research. This paper reviews the evolution of knowledge discovery in academic figures and tables, surveys the research branches and progress of various technical points using the technical framework and process as a skeleton, and finally prospects future research directions.

2. Evolution of Knowledge Discovery in Academic Figures and Tables

Knowledge discovery in academic figures and tables has evolved from object discovery to knowledge discovery. Object discovery refers to the process of extracting, organizing, retrieving, and discovering academic figures and tables from scientific literature. This evolution has progressed through three stages: simple object discovery, literature-associated object discovery, and multi-dimensional object discovery. In the first stage, scholars focused on extracting individual figures or tables from scientific literature and organizing basic metadata for keyword-based discovery. In the second stage, researchers incorporated contextual content as an important information source for discovery, establishing associations between figures/tables and their source documents and attempting to integrate them into scientific literature discovery systems. During this period, research on academic image classification flourished, with image classification organization becoming a new feature. In the third stage, major digital resource providers (e.g., PubMed, CNKI) participated in object discovery, exploring more discovery methods such as using image features for figure-to-figure

discovery and employing natural language processing and machine learning algorithms to automatically extract figures/tables, their textual content, and document metadata for large-scale discovery, while attempting to introduce semantic knowledge organization systems (e.g., thesauri) for semantic expansion. Table 1 summarizes the research and practices across these stages.

Object discovery has partially met researchers' needs for finding non-textual resources, but it only reveals explicit information without identifying hidden knowledge. Moreover, object discovery separates figures/tables from text, hindering knowledge exchange and integration. In recent years, rapid advances in machine vision, text mining, and semantic organization technologies have matured, pushing academic figure/table discovery from mere object discovery toward discovering hidden knowledge within them.

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data [15]. Knowledge discovery in academic figures and tables is the process of automatically constructing and discovering new knowledge patterns from massive amounts of figures and tables in large volumes of literature. This is not a manual process of deduction, induction, or reasoning, but a machine learning process. Academic figures and tables possess dual-modal characteristics of textual and visual information representation, requiring statistical machine learning algorithms, robust database support, linguistic features for text processing and pattern training, and machine vision to mine visual features for hidden knowledge patterns.

Compared to object discovery, knowledge discovery in academic figures and tables achieves three breakthroughs: First, it no longer separates figures/tables from text but eliminates modality barriers through digital knowledge representation patterns, enabling cross-modal discovery at the knowledge level. Computers truly understand figures/tables as integral components of scientific literature. Second, knowledge discovery faces massive data processing, making natural language processing, automatic image classification, automatic text classification, automatic semantic annotation, and information extraction critical supporting technologies. Semantic knowledge organization serves as the primary organization method, assisting multi-source heterogeneous systems in retrieval and fine-grained content discovery. Third, pattern discovery is the core focus. Based on domain knowledge organization systems (e.g., ontologies) and manually annotated corpora, knowledge discovery integrates visual object recognition, term extraction, semantic annotation, and relationship extraction to automatically extract and model complex knowledge.

3. Technical Framework for Knowledge Discovery in Academic Figures and Tables

Knowledge discovery is a process-oriented endeavor. The technical framework for text knowledge discovery comprises four components: free text preprocess-

ing, text representation and encoding, text classification or clustering, and information/knowledge extraction. Similarly, knowledge discovery in academic figures and tables consists of several key technical nodes forming a framework. Based on the basic process of knowledge discovery and the characteristics of academic figures and tables, four key technical nodes are identified: recognition and acquisition of figures/tables and their text, information representation and modeling, classification of figures/tables and their text, and information extraction. Figure 1 [Figure 1: see original paper] illustrates the flow relationships among these technical components.

3.1 Recognition and Acquisition of Academic Figures/Tables and Their Text

3.1.1 Recognition and Acquisition of Academic Figures and Tables

Knowledge discovery must first identify, locate, and extract academic figures and tables from scientific documents and establish connections with surrounding text. Standardized markup formats (HTML/XML) and PDF are the two mainstream document formats, requiring different technologies for figure/table recognition. Figure 2 [Figure 2: see original paper] shows the technical differences across formats.

In the early days of HTML, researchers built table DOM tree models based on ASCII files, optical character recognition, or special tab characters to identify academic tables in HTML documents [16]. In the XML era, figure/table data is stored separately from XML documents, enabling direct acquisition by establishing associations through tags and paths. When tables exist directly in XML, structural analysis is required, using wrapper learning methods with specific tags to acquire and recombine table content [17].

PDF document image recognition has been studied extensively [18]. Graphic data in PDFs is typically embedded as raster formats (PNG, JPEG) or vector formats (SVG, EPS). Two approaches exist for identifying and separating images: (1) Image-based figure-to-figure recognition, which converts the entire PDF to an image and uses bitmap segmentation [19], region classification [20], or connected component methods [21]; and (2) Formatted tag identification, which converts PDFs to structured XML/HTML and extracts images based on tags. Tools like Apache PDFBox [22], PDFMiner [23], Xpdf [24], and Poppler [25] perform such conversions but struggle with vector graphics, extracting individual components (e.g., a bar in a histogram) rather than entire images. To address this, researchers have proposed using regular expressions (heuristics) to identify figure captions and employing clustering algorithms based on caption positions [26] or classification algorithms to exclude irrelevant vector images [27-28]. P.Y. Li et al. separated text from graphic content in PDFs, used connected component analysis to detect images, and restored caption relationships based on layout information [29].

PDF table recognition follows three technical routes: (1) Converting PDF to

XML/TXT and extracting tables based on tags and text features [30]; (2) For tables stored as images, using image recognition techniques with grayscale transformation, smoothing, edge detection, binarization, and skew correction [31]; and (3) Parsing PDF table features (text grids, borders) directly using algorithms to restore table structure [32]. Related tools include Tabula [33], TEXUS [34], and TAO [35].

3.1.2 Recognition and Acquisition of Text in Academic Figures and Tables (1) Internal Text Extraction. Internal text refers to legends, annotations, and text within figures. J. Sas [36] and F. Böschen [37] summarized general methods for text extraction from academic images, including binarization, feature vector calculation, connected component labeling, OCR recognition, and special character filtering.

To improve unstable accuracy, researchers use domain-specific methods. In map images, color quantization algorithms with morphological operators and OCR detect and separate text [38]. Vertical and horizontal projection histograms recursively classify regions as text or non-text [39]. Geometric, regional, example, and contour features with SVM classify text in biomedical publication images [40]. Deep learning and OCR extract molecular entities and interactions from pathway diagrams [41].

Table text extraction is relatively mature: (1) Convert tables to images and use layout, lines, text position, word spacing, and font size features with Bayesian classification or tree traversal [42]; (2) Use rule-based heuristics or templates to identify axis labels and values, extracting and reconstructing relationships [43].

(2) Contextual Text Acquisition. Y. Hong found that researchers lose 30% of information when understanding figures without contextual references [44]. Contextual acquisition must balance coverage and accuracy. Key research focuses on extracting titles, captions, and in-text references.

Title and caption extraction uses rule-based or layout-based methods. Rule-based methods employ specific fields (e.g., `<caption>`, `<table-note>`) or regular expressions based on naming conventions [45], requiring filters to reduce noise (e.g., selecting phrases ending in punctuation, bold/italic text, or font-changed phrases) [46]. Layout-based methods use spatial relationships, employing image recognition to extract captions positioned below figures or above tables [28-29].

Matching titles to figures/tables is crucial. XML documents typically provide reference IDs for direct matching. PDF documents require algorithms considering 1-to-1, N-to-N, and N-to-M relationships between figures/tables and captions based on layout [29].

Contextual mention extraction uses two approaches: (1) Identifying explicit references using keywords like “fig” or “table” [47]; (2) Using figure/table titles or explicit references as anchors to find semantically similar sentences/paragraphs based on topic relevance [48-49].

Overall, recognition and acquisition tasks have developed different technical routes for various document types. While figure/table recognition achieves good results, contextual mention acquisition remains challenging, requiring balance between coverage and precision.

3.2 Information Representation and Modeling of Academic Figures and Tables

Information representation transforms natural language descriptions and visual information into computer-processable digital knowledge patterns. Three representation types exist: textual representation, visual feature representation, and annotated text representation (Figure 3 [Figure 3: see original paper]).

3.2.1 Textual Representation In text knowledge discovery, discrete word representations form the basis, with Bag-of-Words being the most common model, extended to vector space, probabilistic [50], and inference network models [51]. TF-IDF is a standard weighting method, while distributed word embedding is a popular neural network-based approach [52]. These methods extend to figure/table titles, captions, and contexts.

3.2.2 Visual Feature Representation Visual feature representation describes image content for machine understanding, forming the basis of Content-Based Image Retrieval (CBIR). The process involves region selection, feature representation, and feature clustering.

Early region selection used fixed partitioning, which was simple but disrupted visual content. Image segmentation is the most studied approach, including supervised [53], weakly supervised [54], and unsupervised [55] algorithms. While effective in specific domains, automatic segmentation remains suboptimal in general domains. Salient point selection optimizes region selection by identifying distinctive points [56].

After region selection, visual features (color, texture, shape, spatial relationships) are extracted and represented using descriptors, forming Bag of Visual Words (BVW). Common local feature methods include SIFT [57], SURF, and HOG [58]. High-dimensional feature vectors require dimensionality reduction via principal component analysis [59], singular value decomposition [60], or locality-sensitive hashing [61].

Recent deep learning research explores visual-semantic embedding learning [62], consensus-aware visual-semantic embedding [63], and graph attention [64] to mine latent semantic structures between images and text, enabling image-to-text or text-to-image retrieval. These methods unify visual and textual representations but struggle to balance global and local features, limiting application to cross-modal tasks like image captioning and visual Q&A.

3.2.3 Annotated Text Representation Pure visual features cannot bridge the semantic gap between machine and human understanding. Image annotation establishes mappings between low-level visual features and high-level semantic concepts. Academic image annotation uses manual or automatic methods to represent visual features as semantic text, enabling machine understanding [65]. Five mainstream automatic annotation methods exist: generative models, nearest neighbor models, discriminative models, label completion, and deep learning [66]. Deep learning-based annotation using CNNs, RNNs, LSTM, and autoencoders is a current hotspot [67], though manual annotation remains dominant in academic images, with tools like QuickAnnotator [68] and DicomAnnotator [69] supporting semi-automatic or crowdsourced annotation.

The dual-modal nature of academic figures and tables creates representation fragmentation. While annotated text representation attempts to bridge this gap, the lack of initial annotation knowledge bases and incomplete core semantic representation models hinder large-scale application. Unifying visual and textual representation in a shared space remains a promising direction requiring attention to global-local feature balance and visual-semantic reasoning.

3.3 Classification of Academic Figures/Tables and Their Text

Classification is fundamental for retrieval and other applications. Text classification uses predefined frameworks or rules based on logical (decision trees), probabilistic (naive Bayes), or geometric (SVM) models [70].

Academic figure/table text classification has two subtasks: (1) Context classification (e.g., introduction, methods, results, discussion) for summarization; and (2) Internal text classification. Some text in figures (legends, axis labels) has clear meanings and can be classified. J. Poco et al. built a pipeline for academic image text analysis, detecting text, performing OCR, merging words, and classifying text into entity types [71]. S. Kim classified tables in scientific papers into background, system/method, experiment, comment, and comparison categories [72].

Academic image classification has been extensively studied. Research combines low-level image features and text features using SVM [73], CNN [74], and diverse density algorithms [75] to automatically classify images like bar charts, pie charts, line graphs, and ray diagrams. Composite figure recognition and sub-figure classification are current hotspots.

Composite figure recognition uses three methods: (1) Text features (e.g., “A.”, “b:”, “(c)” labels) identified via regular expressions [76] or SVM [77]; (2) Visual features (e.g., blank spaces between sub-figures) using boundary detection [78-79], connected component detection [80], or intensity statistics [81]; and (3) Hybrid features. Sub-figure classification is a multi-label task: (1) Segment composite figures and apply single-figure classification [82]; or (2) Create multi-label models using caption text and visual features [83].

Table classification is less studied, focusing on form and function, such as Tabex identifying web tables as vertical lists, horizontal lists, calendars, or forms [84].

Both figure/table classification and text classification improve information extraction. Current text classification in figures remains functional; future work should explore semantic depth by combining image types to investigate semantic associations between image types and text (e.g., flowchart text representing steps, hierarchical relationships in tree diagrams). Due to domain-specific image types and numerous composite figures, comprehensive image type coverage remains challenging.

3.4 Information Extraction from Academic Figures and Tables

Information extraction is the most critical step, extracting structured information from unstructured data to obtain initial knowledge patterns. Named entity recognition and relationship extraction are core processes. Academic figure/table information extraction includes: (1) Text extraction from documents (well-documented elsewhere); and (2) Figure/table-specific extraction, including entity recognition/annotation and relationship extraction (Figure 4 [Figure 4: see original paper]).

3.4.1 Entity Recognition and Annotation in Academic Images This involves non-text object recognition and text-based named entity recognition.

Non-text object recognition uses image segmentation and machine vision to identify research objects (genes, proteins) in photos, medical images, and micrographs, establishing boundaries and categories. Examples include SLIF for biological microscopy images [85], the Human Brain Project for brain imaging regions [86], and EMAP for mouse embryo annotation [87]. Agricultural researchers use CNNs to identify crop diseases and pests with good results on small datasets [88-89].

Text-based named entity recognition identifies text objects and performs NER based on image content. T. Kuhn et al. identified gene/protein entities in gel diagrams with ~65.3% accuracy [90].

3.4.2 Relationship Extraction from Academic Figures and Tables (1) Table Relationship Extraction. Researchers extract table text and use ontologies or semantic mappings to identify relationships. Z.Q. Zhang proposed TableMiner, an incremental, mutual-recursive, weakly supervised method for semantic annotation of one-dimensional tables [91]. H.P. Cao et al. used ontologies to transform observation data tables into understandable events [92]. C.S. Bhagavatula et al. built TabEL, an entity linking system that determines column types and relationships by co-occurrence analysis [93].

(2) Image Relationship Extraction. Building on extracted text, objects, and values, relationships can be extracted via rules or classification. A. Kembhavi et al. introduced Diagram Parse Graphs (DPG) to identify visual elements

in illustrations (e.g., food chains, atmospheric cycles) and establish semantic relationships [94]. P. Lee et al. proposed PhyloParser, a hybrid algorithm for extracting phylogenies from dendrograms [95]. Y. He studied bar chart detection, segmentation, and information extraction in scientific literature, using CNNs to extract soybean gene-phenotype relationships [96].

Information extraction is comprehensive, requiring representation and classification as foundations and deep semantic integration. Current research achieves partial semantic extraction in specific figure/table types using domain dictionaries or custom relationships. Building a complete semantic knowledge organization system integrated with domain knowledge would enable more precise extraction.

4. Application Services for Knowledge Discovery in Academic Figures and Tables

Application services are the ultimate goal. Current applications focus on three areas: retrieval, automatic summarization, and visual question answering.

4.1 Retrieval of Academic Figures and Tables

Retrieval is the most widespread application, involving figure/table recognition, classification, and annotation. Platforms like CSA Illustrata extract tables and figures, building independent indexes through "deep indexing" with manually annotated metadata for keyword-based retrieval. NIH's Open-i platform integrates research images from PMC, MedPix, and other sources, offering keyword, MeSH term, and image-based retrieval.

Advancing retrieval features: (1) Increased use of machine learning for automatic classification; (2) Semantic annotation for ontology-based term recommendation; (3) Automatic summarization via text classification and similarity computation.

4.2 Automatic Summarization of Academic Figures and Tables

Summaries help researchers quickly grasp figure/table meanings without reading full papers and support modular knowledge services. Summarization uses context extraction, text classification, and information extraction. Two types exist: extractive (selecting existing sentences) and abstractive (generating novel sentences). Current approaches are mostly extractive.

Methods are categorized as supervised or unsupervised. Supervised methods train classifiers (naive Bayes, SVM) to select sentences similar to figure/table titles [49]. S. Agarwal et al. developed FigSum, extracting structured summaries classified as introduction, methods, results, and discussion [97]. Unsupervised methods use multi-objective optimization (MOO) without training, as in MOOFigSum [98] and FigSum++ [99]. J. Chen et al. used unsupervised hierarchical multi-modal RNNs for text+image news summarization [100].

4.3 Visual Question Answering

Visual Question Answering (VQA) integrates computer vision and natural language processing. Users input images and natural language questions, receiving natural language answers. This involves object recognition, annotation, and other knowledge discovery technologies.

Current VQA focuses on natural images, with methods including feature fusion, entity attention, multi-step reasoning, knowledge injection, and relationship modeling [101]. In academic images, researchers develop domain-specific VQA and datasets. A. Kembhavi et al. used DPG attention models to extract illustration elements and text, building a VQA system [95]. K. Kafle introduced DVQA for bar chart data retrieval and reasoning [102]. Microsoft created FigureQA with 180K charts and 200K+ Q&A pairs [103]. Similar datasets include LEAF-QA [104].

Overall, academic image VQA has broad prospects but requires substantial technical development.

5. Research Outlook

Based on the technical framework and applications, knowledge discovery in academic figures and tables has made progress in recognition, representation, classification, and extraction, with emerging services in retrieval, summarization, and Q&A. To maximize its role in future academic knowledge service systems, we propose the following strategies:

5.1 Prioritize Knowledge Discovery in Academic Figures and Tables and Integrate It into Literature Knowledge Discovery

Textual knowledge discovery has long dominated literature-based knowledge discovery. In the data-intensive scientific discovery ecosystem, researchers increasingly value figures and tables, necessitating their integration into existing systems.

Integration can: (1) Expand retrieval objects beyond text, providing richer multi-dimensional figure/table displays and extending to underlying scientific data; (2) Enable evidence-based precise discovery, advancing services toward multi-modal knowledge; (3) Deepen computer understanding through extraction and annotation, laying foundations for deep relationship mining.

Specifically: (1) Build semantic representation models for figure/table-centric knowledge units using ontology learning, integration, and alignment; (2) Develop specialized semantic knowledge bases; (3) Leverage deep learning to breakthrough key technologies (unified semantic representation, automatic classification, semantic annotation, intelligent recommendation, knowledge extraction, automatic summarization) for multi-modal knowledge discovery.

5.2 Improve the Semantic Knowledge Organization System and Build Specialized Semantic Knowledge Bases

Semantic knowledge bases integrate knowledge discovery and organization, providing semantic data support for NER, similarity computation, and extraction. While mature in text domains, academic figure/table semantic knowledge bases remain nascent.

Building figure/table-centric knowledge unit semantic representation models is imperative. Current organization relies on traditional metadata, with ontologies and knowledge graphs emerging. Existing discovery rarely leverages knowledge organization systems, limiting large-scale effectiveness. Ontological semantic models can serve as frameworks for semantic categories and associations, playing crucial roles in normalization and disambiguation throughout the retrieval-to-Q&A pipeline.

We must construct semantic representation models and attribute systems applicable to different types, domains, and problems, using various knowledge organization methods to build application ontologies, domain ontologies, and knowledge graphs. Semantic annotation technologies should be applied to build foundational corpora and knowledge bases.

5.3 Develop Novel Knowledge Discovery Applications for Academic Figures and Tables

Knowledge services manifest value through applications. Retrieval is the foundation and priority. Providers like PMC, ProQuest, and CNKI use retrieval as an entry point. Retrieval should combine semantic organization, multi-factor ranking, and intelligent recommendation to create semantic intelligent discovery engines.

Automatic summarization is crucial for rapid literature comprehension and modular knowledge representation. Future work should combine extractive and abstractive methods, integrating image features for multi-modal summarization.

Q&A and reasoning services expand academic knowledge applications. For example, historical weather statistics can predict meteorology and correlate with crop characteristics. Additionally, figure/table misuse is a concern in research integrity. Fine-grained semantic annotation and visual similarity computation can build plagiarism detection systems to prevent improper reuse and data fabrication.

References

[1] SIEGEL N, LOURIE N, POWER R, et al. Extracting scientific figures with distantly supervised neural networks[C]//Proceedings of the 18th ACM-IEEE on joint conference on digital libraries. Texas: ACM, 2018: 223-232.

[2] YU H, LEE M. Accessing bioscience images from abstract sentences[J]. Bioinformatics, 2006, 22(14): 547-556.

[3] STELMASZEWSKA H, BLANDFORD A. From physical to digital: a case study of computer scientists' behaviour in physical libraries[J]. International journal on digital libraries, 2004, 4(2): 82-92.

[4] LEE P, WEST J D, HOWE B, et al. Vizometrics: analyzing visual information in scientific literature[J]. IEEE transactions on big data, 2018, 4(1): 117-129.

[5] PYREDDY P, CROFT W B. TINTIN: A system for retrieval of text tables[C]//Proceedings of the second ACM international conference on digital libraries. Philadelphia: ACM, 1997: 193-200.

[6] LIU F, JENSEN T, NYGAARD V, et al. FigureSearch: a figure legend indexing and classification system[J]. Bioinformatics, 2004, 20(16): 2880-2882.

[7] TENOPIR C, SANDUSKY R, CASADO M. The value of CSA deep indexing for researchers (executive summary)[J]. School of information sciences publications and other works, 2006(1): 1-8.

[8] LIU Y, BAI K, MITRA P, et al. TableSeer: automatic table metadata extraction and searching in digital libraries[C]//Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2007: 91-100.

[9] XU S H, JAMES M C, MICHAEL K. Yale image finder (YIF)[J]. Bioinformatics, 2008, 17(24): 1968-1970.

[10] HONG Y, LIU F, RAMESH B P. Automatic figure ranking and user interfacing for intelligent figure search[J]. Plos one, 2010, 5(10): e12983.

[11] NCBI. PMC[EB/OL].[2020-08-31]. https://www.ncbi.nlm.nih.gov/pmc/.

[12] CNKI. CNKI image retrieval[EB/OL].[2020-08-31]. http://image.cnki.net/Default.aspx.

[13] SIEGEL N, HORVITZ Z, LEVIN R, et al. FigureSeer: parsing result-figures in research papers[C]//European conference on computer vision. Amsterdam: Springer International Publishing, 2016: 664-680.

[14] National Library of Medicine. Open-i[EB/OL].[2020-08-31]. https://openi.nlm.nih.gov/.

[15] FAYYAD U M, PIATETSKY-SHAPIRO G, SMYTH P. From data mining to knowledge discovery in databases[J]. AI magazine, 1996, 17(3): 37-54.

[16] TANG H. Research and implementation of a table data extraction method for PDF files[D]. Beijing: Beijing University of Posts and Telecommunications, 2015.

[17] LIU Y. Research on table information extraction based on web structure[D]. Hefei: Hefei University of Technology, 2012.

[18] CHAO H, FAN J. Layout and content extraction for PDF documents[C]//Document analysis systems 2004. Florence: Springer, 2004: 213-224.

[19] CHOUDHURY S R, GILES C L. An architecture for information extraction from figures in digital libraries[C]//International conference on document analysis & recognition. IEEE Computer Society. Washington, DC: IEEE, 2013: 887-891.

[20] CHHATKULI A, FONCUBIERTA-RODRÍGUEZ A, MARKONIS D, et al. Separating compound figures in journal articles to allow for subfigure classification[C]//Medical imaging 2013: Advanced PACS-based imaging informatics and therapeutic applications. Orlando: SPIE Medical Imaging, 2013: 86740J.

[21] LI P, JIANG X, KAMBHAMETTU C, et al. Compound image segmentation of published biomedical figures[C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. Newark: ACM, 2016: 143-152.

[22] Apache Software Foundation. Apache PDFBox[EB/OL].[2021-05-02]. https://pdfbox.apache.org/.

[23] YUSUKE S. PDFMiner[EB/OL].[2021-05-02]. https://github.com/euske/pdfminer.

[24] Glyph & Cog. Xpdf[EB/OL].[2021-05-02]. http://www.xpdfreader.com/.

[25] Kristian Høgsberg. Poppler[EB/OL].[2021-05-02]. http://poppler.freedesktop.org/.

[26] LUIS D L, JINGYI Y, CECILIA N, et al. An automatic system for extracting figures and captions in biomedical PDF documents[C]//2011 IEEE international conference on bioinformatics and biomedicine. Atlanta: IEEE, 2011: 578-581.

[27] PRACZYK P A, NOGUERAS-ISO J, MELES S. Automatic extraction of figures from scientific publications in high-energy physics[J]. Information technology and libraries, 2013, 32(4): 25-52.

[28] CLARK C, DIVVALA S. PDFFigures 2.0: mining figures from research papers[C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. Newark: ACM, 2016: 143-152.

[29] LI P Y, JIANG X Y, SHATKAY H, et al. Figure and caption extraction from biomedical documents[J]. Bioinformatics, 2019, 35(21): 4381-4388.

[30] YILDIZ B, KAISER K, MIKSCH S. Pdf2table: a method to extract table information from PDF files[C]//Proceedings of the 2nd Indian international conference on artificial intelligence. Pune: Springer, 2005: 1-13.

[31] LI H, LIU J, MING D, et al. A table recognition method based on statistical feature point grid distribution[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2002, 30(9): 60-63.

[32] ZHANG B. Research on table recognition technology based on PDF text stream[D]. Beijing: North China University of Technology, 2010.

[33] MANUELA A, MIKE T, JEREMY B M. Tabula[EB/OL].[2021-05-02]. https://tabula.technology/.

[34] RASTAN R, PAIK H Y, SHEPHERD J. TEXUS: A unified framework for extracting and understanding tables in PDF documents[J]. Information processing & management, 2019, 55(3): 895-918.

[35] PEREZ-ARRIAGA M O, ESTRADA T, ABAD-MOTA S. TAO: system for table detection and extraction from PDF documents[C]//Proceedings of the 29th international Florida artificial intelligence research society conference. Florida: AAAI, 2016: 591-596.

[36] SAS J, ZOLNIEREK A. Three-stage method of text region extraction from diagram raster images[J]. Advances in intelligent systems and computing, 2013, 226: 527-538.

[37] FALK BÖSCHEN, ANSGAR SCHERP. A comparison of approaches for automated text extraction from scholarly figures[C]//International conference on multimedia modeling. Reykjavik: Springer, 2017: 15-27.

[38] CHIANG Y Y, KNOBLOCK C A. Recognizing text in raster maps[J]. Geoinformatica, 2015(19): 1-27.

[39] XU S H, MICHAEL K. A new pivoting and iterative text detection algorithm for biomedical images[M]. Elsevier Science, 2010.

[40] DE S, STANLEY R J, CHENG B, et al. Automated text detection and recognition in annotated biomedical publication images[J]. International journal of healthcare information systems and informatics, 2014, 9(2): 34-63.

[41] HE F, WANG D, INNOKENTEVA Y, et al. Extracting molecular entities and their interactions from pathway figures based on deep learning[C]//2019 IEEE international conference on bioinformatics and biomedicine (BIBM). San Diego: IEEE, 2020: 1191-1193.

[42] NAGY G. Learning the characteristics of critical cells from web tables[C]//International conference on pattern recognition. Tsukuba: IEEE, 2012: 1554-1557.

[43] SETH S C, NAGY G. Segmenting tables via indexing of value cells by table headers[C]//International conference on document analysis & recognition. Washington, DC: IEEE, 2013: 887-891.

[44] HONG Y, AGARWAL S, JOHNSTON M. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension[J]. Journal of biomedical discovery & collaboration, 2009, 4(1): 1-10.

[45] CHOUDHURY S R, MITRA P, KIRK A, et al. Figure metadata extraction from digital documents[C]//International conference on document analysis & recognition. Washington, DC: IEEE, 2013: 887-891.

[46] LOPEZ L D, YU J, ARIGHI C N, et al. An automatic system for extracting figures and captions in biomedical PDF documents[C]//IEEE international conference on bioinformatics & biomedicine. Atlanta: IEEE, 2012: 578-581.

[47] BALAJI P R, SETHI R J, HONG Y, et al. Figure-associated text summarization and evaluation[J]. Plos One, 2015, 10(2): e0115671.

[48] YU H. Towards answering biological questions with experimental evidence: automatically identifying text that summarizes image content in full-text articles[C]//Annual symposium proceedings/AMIA symposium. Washington, DC: AMIA, 2006: 834-838.

[49] BHATIA S, MITRA P. Summarizing figures, tables and algorithms in scientific publications to augment search results[J]. ACM transactions on information systems, 2010, 30(1): 1-24.

[50] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to information retrieval[M]. Beijing: Posts & Telecom Press, 2010.

[51] TURTLE H R, CROFT W B. Inference networks for document retrieval[C]//13th international conference on research and development in information retrieval. Brussels: ACM, 1990: 1-24.

[52] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer science, 2013, arXiv:1301.3781.

[53] SHUAIZ, CHENG M M, WARRELL J, et al. Dense semantic image segmentation with objects and attributes[C]//2014 IEEE conference on computer vision and pattern recognition (CVPR). Columbus: IEEE, 2014: 3214-3221.

[54] VEZHNEVETS A, FERRARI V, BUHMANN J M. Weakly supervised structured output learning for semantic segmentation[C]//2012 IEEE conference on computer vision and pattern recognition. Providence: IEEE, 2012: 845-852.

[55] HUI Z, FRITTS J E, GOLDMAN S A. Image segmentation evaluation: a survey of unsupervised methods[J]. Computer vision & image understanding, 2008, 110(2): 260-280.

[56] PEDERSEN K S, LOOG M, DORST P. Salient point and scale detection by minimum likelihood[C]//Proceedings of machine learning research. Bletchley Park: PMLR, 2007: 59-72.

[57] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[58] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE computer society conference on computer vision & pattern recognition. San Diego: IEEE, 2005: 886-893.

[59] NG R T, SEDIGHIAN A. Evaluating multi-dimensional indexing structures for images transformed by principal component analysis[C]//Proceedings

volume 2670, storage and retrieval for still image and video databases IV. San Jose: SPIE, 1996: 50-61.

[60] PHAM T T, MAILLOT N E, LIM J H, et al. Latent semantic fusion model for image retrieval and annotation[C]//Proceedings of the sixteenth ACM conference on information and knowledge management. Lisbon: ACM, 2007: 439-444.

[61] INDYK P. Approximate nearest neighbors: towards removing the curse of dimensionality[C]//Proceedings of the 30th ACM symposium on theory of computing. Dallas: ACM, 1998: 604-613.

[62] YANG Z. Research on visual semantic embedding based on deep learning and word embedding[D]. Chongqing: Southwest University, 2019.

[63] WANG H, ZHANG Y, JI Z, et al. Consensus-aware visual-semantic embedding for image-text matching[J]. IEEE transactions on circuits and systems for video technology, 2020(99): 1-1.

[64] WEN K, GU X, CHENG Q. Learning dual semantic relations with graph attention for image-text matching[C]//2020 European conference on computer vision. Glasgow: Qrxiv, 2020: 18-34.

[65] CHEN T, SHAN R, LI H. Research on semantic annotation of image resources in digital humanities[J]. Journal of Library and Information Science in Agriculture, 2020, 32(9): 6-14.

[66] BHAGAT P K, CHOUDHARY P. Image annotation: then and now[J]. Image and vision computing, 2018(80): 1-23.

[67] ADNAN M M, RAHIM M, REHMAN A, et al. Automatic image annotation based on deep learning models: a systematic review and future challenges[J]. IEEE access, 2021(9): 50253-50264.

[68] MIAO R, TOTH R, ZHOU Y, et al. Quickannotator: an open-source digital pathology based rapid image annotation tool[J]. The journal of pathology, 2021, 7(6): 542-547.

[69] DONG Q, LUO G, HAYNOR D, et al. DicomAnnotator: a configurable open-source software program for efficient DICOM image annotation[J]. Journal of digital imaging, 2020, 33(6): 1514-1526.

[70] SUN T, DING P, HUANG Y, et al. Review of text mining technology applications in agricultural knowledge services[J]. Journal of Library and Information Science in Agriculture, 2021, 33(1): 4-16.

[71] POCO J, HEER J. Reverse-engineering visualizations: recovering visual encodings from chart images[J]. Computer graphics forum, 2017, 36(3): 353-363.

[72] KIM S, LIU Y. Functional-based table category identification in digital libraries[C]//2011 international conference on document analysis and recognition. IEEE, 2011: 1364-1368.

[73] SAVVA M, KONG N, CHHAJTA A, et al. ReVision: automated classification, analysis and redesign of chart images[C]//User interface software and technology. New York: ACM, 2011: 393-402.

[74] NKWENTSHA X, HOUNKANRIN A, NICOLLS F. Automatic classification of medical X-ray images with convolutional neural networks[C]//2020 international SAUPEC/RobMech/PRASA conference. Cape Town: Springer, 2020: 1-4.

[75] HUANG W, ZONG S, TAN C L, et al. Chart image classification using multiple-instance learning[C]//Workshop on applications of computer vision. Texas: ACM, 2007: 27-27.

[76] PELKA O, FRIEDRICH C M. FHDO biomedical computer science group at medical classification task of ImageCLEF 2015[C]//Working notes of CLEF 2015 conference. Toulouse: CEUR-WS, 2015.

[77] LI P, SORENSEN S, KOLAGUNDA A, et al. UDEL CIS working notes in ImageCLEF 2016[C]//Working notes of CLEF 2016 conference. Portugal: CEUR-WS, 2016: 334-346.

[78] CHHATKULI A, FONCUBIERTA-RODRÍGUEZ A, MARKONIS D, et al. Separating compound figures in journal articles to allow for subfigure classification[C]//Proceedings of SPIE medical imaging: Advanced PACS-based imaging informatics and therapeutic applications. Orlando: SPIE, 2013: 86740.

[79] YUAN X, ANG D. A novel figure panel classification and extraction method for document image understanding[J]. International journal of data mining and bioinformatics, 2014, 9(1): 22-36.

[80] LI P, JIANG X, KAMBHAMETTU C, et al. Segmenting compound biomedical figures into their constituent panels[C]//International conference of the cross-language evaluation forum for European languages. Dublin: Springer, 2017: 199-210.

[81] TASCHWER M, MARQUES O. Compound figure separation combining edge and band separator detection[C]//International conference on multimedia modeling. Miami: Springer, 2016: 162-173.

[82] SANTOSH K C, AAFAGUE A, ANTANI S, et al. Line segment-based stitched multi-panel figure separation for effective biomedical CBIR[J]. International journal of pattern recognition and artificial intelligence, 2017, 31(6): 1757003.

[83] YU Y. Research on key technologies of image pattern recognition for medical literature[D]. Dalian: Dalian University of Technology, 2018.

[84] CRESTAN E, PANTEL P. Web-scale table census and classification[C]//Proceedings of the fourth ACM international conference on web search and data mining. Hong Kong: ACM, 2011: 545-554.

[85] MURPHY R F, VELLISTE M, YAO J, et al. Searching online journals for fluorescence microscope images depicting protein subcellular location patterns[C]//IEEE international symposium on bioinformatics & bioengineering. Bethesda: IEEE, 2001: 119-128.

[86] GERTZ M, SATTLER K U, GORIN F, et al. Annotating scientific images: a concept-based approach[C]//Proceedings 14th international conference on scientific and statistical database management. Los Alamitos: IEEE, 2002: 59-68.

[87] EMAGE. Data Annotation Methods[EB/OL].[2020-11-02]. http://www.emouseatlas.org/emage/about/dat

[88] TOO E C, YU J L, NJUKI S, et al. A comparative study of fine-tuning deep learning models for plant disease identification[J]. Computers and electronics in agriculture, 2018, 161(1): 272-279.

[89] BARBEDO J A. Plant disease identification from individual lesions and spots using deep learning[J]. Biosystems engineering, 2019, 180(1): 96-107.

[90] KUHN T, NAGY M, LUONG T B, et al. Mining images in biomedical publications: Detection and analysis of gel diagrams[J]. J biosemantics, 2014, 5(1): 1-9.

[91] ZHANG Z. Towards efficient and effective semantic table interpretation[C]//International semantic web conference. New York: Springer-Verlag, 2014: 487-502.

[92] CAO H, BOWERS S, SCHILDBAUER M P. Approaches for semantically annotating and discovering scientific observational data[C]//Database and expert systems applications. Berlin: Springer, 2011: 526-541.

[93] MARTIN M, NUFFELEN B V, ABRUZZINI S, et al. The digital agenda scoreboard: a statistical anatomy of Europe's way into the information age[EB/OL].[2021-05-02]. http://www.semantic-web-journal.net/sites/default/files/swj283.pdf.

[94] KEMBHAVI A, SALVATO M, KOLVE E, et al. A diagram is worth a dozen images[C]//Computer vision-ECCV 2016. Amsterdam: Springer, 2016: 235-251.

[95] LEE P, YANG T S, WEST J, et al. PhyloParser: a hybrid algorithm for extracting phylogenies from dendrograms[C]//14th IAPR international conference on document analysis and recognition (ICDAR). Kyoto: IEEE, 2017: 1087-1094.

[96] HE Y. Research and application of bar chart information extraction in PubMed Central literature[D]. Wuhan: Wuhan University of Technology, 2018.

[97] AGARWAL S, YU H. FigSum: automatically generating structured text summaries for figures in biomedical literature[C]//American medical informatics association annual symposium. San Francisco: PMC, 2009: 6-10.

[98] SAINI N, SAHA S, POTNURU V, et al. Figure summarization: a multiobjective optimization-based approach[J]. Intelligent systems, 2019, 34(6): 43-52.

[99] SAINI N, SAHA S, BHATTACHARYYA P, et al. Textual entailment-based figure summarization for biomedical articles[J]. ACM transactions on multimedia computing communications and applications, 2020, 16(1s): 1-24.

[100] CHEN J, ZHU G H. Extractive summarization of documents with images based on multi-modal RNN[J]. Future generation computer systems, 2019, 99(1): 186-196.

[101] WU C. Research on visual question answering based on relationship modeling[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.

[102] KAFLE K, PRICE B, COHEN S, et al. DVQA: understanding data visualizations via question answering[C]//2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 5648-5656.

[103] KAHOU S E, MICHALSKI V, ATKINSON A, et al. FigureQA: an annotated figure dataset for visual reasoning[J]. Computer science, 2018, arXiv:1710.07300.

[104] CHAUDHURY R, SHEKHAR S, GUPTA U, et al. LEAF-QA: locate, encode & attend for figure question answering[C]//2020 IEEE winter conference on applications of computer vision (WACV). Snowmass Village: IEEE, 2020: 3512-3521.

The Technical Framework and Research Progress of Knowledge Discovery in Academic Figures and Tables

Ding Pei
Shenzhen University Library, Shenzhen 518060

Abstract:
[**Purpose/Significance**] Against the backdrop of deep integration of scientific and technological resources, knowledge discovery in academic figures and tables provides a new approach to knowledge discovery beyond textual knowledge discovery. It represents a crucial component in perfecting document-based knowledge discovery, enhancing researchers' efficiency in scientific discovery and knowledge creation, and promoting the upgrading of knowledge services in digital libraries. [**Method/Process**] This paper traces the evolutionary trajectory of knowledge discovery in academic figures and tables, elaborates on its technical framework, and demonstrates the gradual maturation of related technologies. Combined with application services, it demonstrates that knowledge discovery in academic figures and tables has broad application prospects across multiple facets of scientific and technological innovation. [**Result/Conclusion**] Looking ahead, we need to: prioritize knowledge discovery in academic figures and tables and integrate it into the literature knowledge discovery system; improve the semantic knowledge organization system for academic figures and tables and

construct specialized semantic knowledge bases; and develop novel knowledge discovery applications for academic figures and tables.

Keywords: academic figures and tables; knowledge discovery; knowledge organization; information extraction

Note: Figure translations are in progress. See original paper for figures.