JASIST 数据集介绍

The Introduction of the JASIST dataset.

This document is preceded by the **Chinese version** and followed by the **English version**.

信息科学与技术协会会刊（Journal of the Association for Information Science and Technology，JASIST）是国际领先的信息科学领域同行评议学术期刊，其课题涵盖了信息发现、存储、表示、检索和分析等领域的技术与应用，是图书情报领域的核心期刊之一，研究它的学术文献全文在科学计量研究中具有重要意义。结合自然语言处理和文本挖掘技术对文献内部知识单元进行定量分析和评价是当前的热点研究范式，我们爬取了 2010 年至 2020 年 6 月发表到 JASIST 期刊（https://www.asist.org/publications/jasist/）上的论文全文数据，从句子和实体两个层面对其进行了数据加工，制作了知识图谱和实体数据集。

本数据集包含四部分：

1 知识图谱: 基于关联数据的“实体-关系”知识图谱。

实体（含句子）类型分为 16 类：

研究问题、研究方法、研究结果、研究展望、引文作者、引文时间、引文的研究问题、引文的研究方法、引文的研究结果、时间、数据资源与平台（含标准数据集、自建数据集、数据源）、工具、模型、数学公式、表、图。

其中研究问题、研究方法、研究结果、研究展望、数据资源与平台、工具、模型的概念参见第 2 和第 3 部分。其余实体概念如下：

（1）引文作者指被引用文献的作者姓名。

（2）引文时间指的是被引用文献的发表时间。

（3）引文的研究问题指被引句中应引的研究问题部分。

（4）引文的研究方法指引文中的研究方法、工具、手段、技术和解决方案。

（5）引文的研究结果指在引文句中的实验结果和结论部分。

（6）时间指描述时间的词汇。

（7）数学公式指文中被整个标注出来的公式；出现 ` formula xxx `并被标记为` formula xxx `。

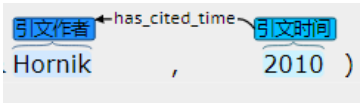
（8）表指的是文中被标注出来的表格和相应的解释。

（9）图指的是文中被标注出来的图和相应的解释。

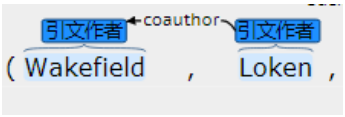
关系类型分为 10 类：

solve、supports、produces、has_cited_time、coauthor、field_similar_as、results、uses、leads to、select_time。其概念如下：

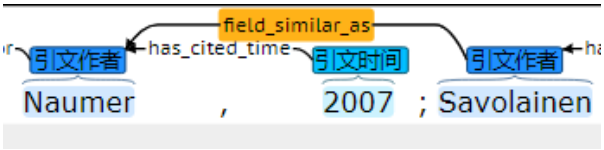
（1） has_cited_time：引文作者与引文发表时间之间的关系。表示为:引文时间_ Time - has_ cited_ Time -引文作者_作者(被引时间)。



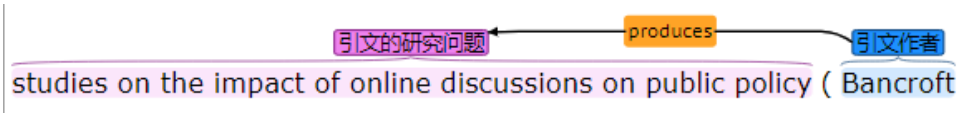
（2） coauthor：协作是指同一或不同研究领域的研究人员为解决具体问题或项目而进行的协作。表示为:引文作者-合著者-引文作者(合著)。



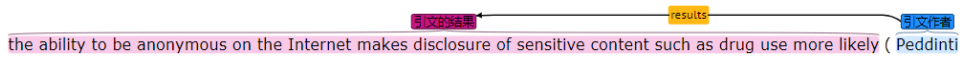
（3） field_similar_as：相似研究领域的研究人员之间的关系。表示为:引文的作者-领域_ similar_ as -引文的作者(相似研究领域)。



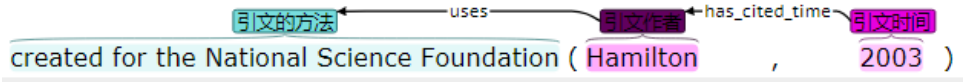
（4） produces：引文的作者与提出的研究问题之间的关系。表示为:引文作者-produces-引文研究问题_研究问题(生成)。



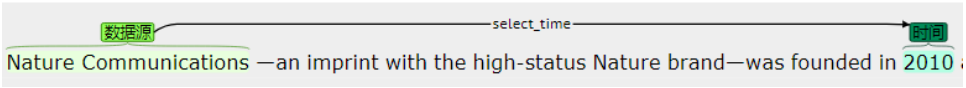
（5） results：引用作者与获得的研究结果之间的关系。表示为:被引作者- results -被引研究结果_研究结果。



（6） uses：引用作者与所使用的研究方法之间的关系。这表现为:作者的引文- uses -引用研究方法_研究方法(使用)。



（7） select_time：有两种类型的关系:数据源与其选择的时间之间的关系，以及软件工具与其选择的时间之间的关系。表示为:数据资源和平台-Select_ Time - Time (selection)和软件工具 tools - Select_ Time - Time (selected)



（8） solve：研究中的解决方案与其解决的研究问题之间的关系。

（9） supports：研究方法、工具等与通过使用这些方法和工具得到的研究问题之间的关系。

（10） leads to：研究发现或理论基础与基于这些发现或基础的进一步研究发现之间的关系。

2 研究句子数据集：包含 2010-2020.6 月的全文数据、数据精度一般

句子类型分为 4 类：研究问题、研究方法、研究结果、研究展望（含不足、局限等）。其实体概念如下：

(1)研究问题指文章想要解决的问题。此处的研究问题指在学术全文本中指明文献研究问题的句子。

(2)研究方法指关于解决应用领域问题的方法、工具、手段、技术和方案。此处的研究方法指在学术全文本中指明文献研究方法的句子。

(3)研究结果指文章实验得出的结果和结论。此处的研究结果指在学术全文本中指明文献研究结果的句子。

(4)研究展望指文章某些方面的研究还不够深入，还存在问题有待进一步解决等。此处的研究展望指在学术全文本中指明文献研究展望的句子。

3 知识实体数据集：包含部分摘要数据、人工加工校对过。

实体类型分为 6 类：研究对象、研究方法/技术、工具、模型、评价指标、数据资源与平台

（1）研究对象指研究所探讨的问题，研究方法的作用目标等。一般是一组名词性结构短语。

（2）研究方法/技术是指研究所采用的方法技术的名称，涉及具体算法模型、理论模型的不纳入此项实体类型。

（3）工具，即现有的可直接使用的各类软件工具包，包括统计分析软件、信息计量、科学计量与可视化软件工具、自然语言处理工具、编程语言等。

标注范例 1：

（4）模型包括理论模型、算法模型等。

（5）评价指标是指一项研究或实验对于研究结果的合适评价标准、指标，包括统计学指标、计量学指标和模型性能评估指标等。

（6）数据资源与平台，包括研究采用的通用数据集或公开数据集名和研究人员所采集数据的来源两类，即“数据源”和“数据集”。

4 摘要数据集（句子+实体）：包含全部摘要数据，为模型的预标注数据集，数据精度一般。

摘要数据集由第 3 部分的知识实体数据集进行训练后预测得到，实体类型同知识实体数据集，包含研究对象、研究方法/技术、工具、模型、评价指标、数据资源与平台六类实体。

Journal of the Association for Information Science and Technology (JASIST) is a leading international peer-reviewed academic journal in the field of information science, whose topics cover technologies and applications in the areas of information discovery, storage, representation, retrieval, and analysis. It is one of the core journals in the field of library and intelligence, and the study of its full-text scholarly literature is of great significance in scientometric research. Combining natural language processing and text mining techniques for quantitative analysis and evaluation of knowledge units within the literature is a current hot research paradigm, we crawled the full-text data of papers published to the JASIST journal (https://www.asist.org/publications/jasist/) from 2010 to June 2020, and analyzed them from sentence and The data were processed at both sentence and entity levels to produce a knowledge graph and entity dataset.

This dataset contains four parts:

1、 Knowledge Graph："Entity-relationship" knowledge graph based on Linked Data.

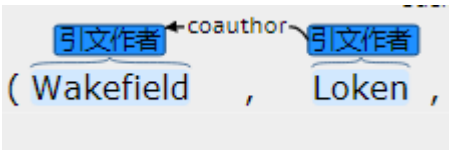
The types of entities (including sentences) are divided into 16 categories:

Research Questions, Research Methods, Research Results, Research Prospects, Citation Author , Citation Time, Citation Research Questions, Citation Research Methods, Citation Research Results, Time, Data Resources and Platforms (including standard datasets, self-built datasets, data sources), Tools, Model, Mathematical formulas, Table, and Figures.

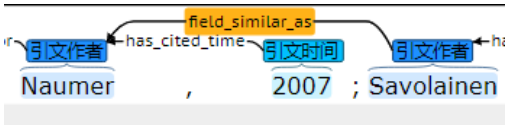
Among them, the concepts of Research Questions, Research Methods, Research Results, Research Prospects, Data resources and Platforms, Tools, and Model are referred to sections 2 and 3. The remaining entity concepts are as follows:
(1) Citation author refers to the name of the author of the cited document.
(2) Citation Time refers to the time of publication of the citation.
(3) Citation Research Questions refers to the part of the research question that should be cited in the quoted sentence.
(4) Citation Research Methods refers to the research methods, tools, instruments, techniques, and solutions in the cited text.
(5) Citation Research Results refer to the experimental results and the conclusion section in the cited sentence.
(6) Time refers to the words describing time.
(7) Mathematical formulas, with the entire formula marked out; The 'formula xxx' appears and is marked with 'formula xxx'.
(8) Tables refer to mark out the table and corresponding explanations.
(9) Figure refer to t mark out the diagram and corresponding explanations.

Relationship types are divided into 10 categories:
solve, supports, provides, has_cited_time, co-author, field_similar_as, results, uses, leads to, select_time. the concepts are as follows:
(1)has_cited_time: The relationship between the citation author and the time when the citation was published. Represented as: Citation time_ Time—has_ cited_ Time - Author of the citation Citation_ Author (cited time)

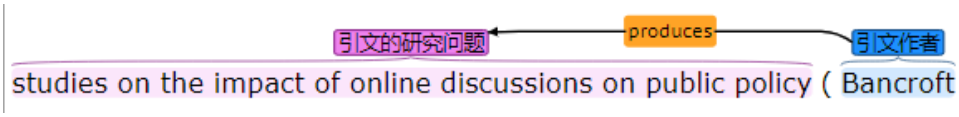
(2)coauthor: Collaborative collaboration refers to the collaboration between researchers in the same or different research fields to solve specific problems or projects. Represented as: author of the citation - co author - author of the citation (co writing)



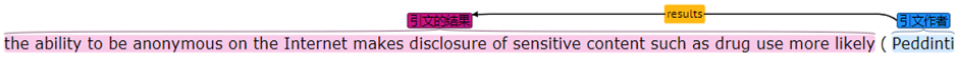
(3)field_similar_as:The relationship between researchers in similar research fields. Represented as: the author of the citation - field_ similar_ As - Author of the citation (similar research field)



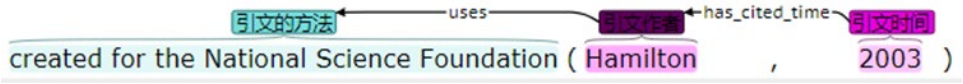
(4)produces:The relationship between the author of the citation and the research questions raised. Represented as: the author of the citation - products - citation research question_ Research Question (generated)



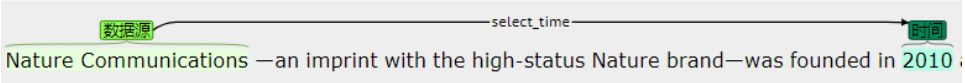
(5)results:The relationship between the author of the citation and the research results obtained. Represented as: author of the citation - results - citation research results_ ResearchResult



(6)uses:The relationship between the citation author and the research methods used. This is manifested as: the author of the citation - uses - citation research method_ Research Method (used)



(7)select_time:There are two types of relationships: the relationship between the data source and its selected time, and the relationship between the software tool and its selected time. Represented as: Data Resources and Platforms Select_ Time - Time (selection) and software tools Tools - select_ Time - Time (selected)



(8)solve:The relationship between the solution in the study and the research problem it solves.
(9)supports:The relationship between research methods, tools, etc., and the research questions obtained through the use of these methods and tools.
(10)Leads to： The relationship between research findings or theoretical foundations and further research findings based on these findings or foundations.

2、 Research sentence dataset: Contains full-text data from 2010 to June 2020.

The types of sentences are divided into 4 categories: Research Questions, Research Methods, Research Results, and Research prospects (with deficiencies, limitations, etc.). Their concepts are as follows:

(1) Research problem refers to the problem that the article aims to solve. The research question here refers to the sentence indicating the literature research question in the academic full-text.

(2) Research methods refer to methods, tools, means, technologies, and solutions for solving problems in the application field. The research method here refers to the sentence indicating the literature research method in the academic full-text.
(3) The research results refer to the results and conclusions drawn from the experiment in the article. The research results here refer to the sentences that indicate the literature research results in the academic full-text.
(4) Research outlook refers to the lack of in-depth research on certain aspects of the article, and there are still issues that need to be further addressed. The research outlook here refers to the sentence indicating the literature research outlook in the academic full-text.

3、 Knowledge entity dataset: Contains partial abstract data, manually processed and proofread.

The types of entities are divided into 6 categories: Research subjects, Research methods/techniques, Tools, Model, evaluation indicators, Data resources and Platforms.
(1) The research subjects refers to the problems explored by the research institute, the purpose of the research method, etc. Generally, it is a group of noun structured phrases.
(2) Research methods/technologies refer to the names of the methods and technologies used in the research, and those involving specific algorithm models and theoretical models are not included in this entity type.
(3) Tools in general, namely, existing software toolkits that can be used directly, including statistical analysis software, information metrology, scientific metrology and visualization software tools, natural language processing tools, programming languages, etc.
(4) Various algorithms and models, including theoretical models, algorithm models, etc.
(5)Evaluation indicators refer to the appropriate evaluation criteria and indicators for the research results of a study or experiment, including statistical indicators, metrology indicators and model performance evaluation indicators.
(6) The data resources and platforms here include two categories: the common dataset or public dataset names used in the research, and the sources of data collected by the researchers, namely "data sources" and "datasets".

4、Abstract dataset (sentences + entities): All abstract data dataset.

The Abstract dataset is predicted from the knowledge entity dataset in part 3 after training, and the entity type is the same as the knowledge entity dataset, containing six types of entities: Research subjects, Research methods/techniques, Tools, Model, evaluation indicators, Data resources and Platforms.

5 附件（Appendix）：

5.1 句子类型及实体标注示例 （ Examples of annotations of different types of sentences and entities. ）

（1）研究问题（Research Questions）

Example1:10.1002/asi.22617-2-24 Now , we come to the main point of our article , namely , the relation between the h-index h and the impact factor IF (average number of citations per publication) in the shifted Lotka model .

Example2:10.1002/asi.22618-1-8 The purpose of this article is to report the development and the evaluation of an automated semantic annotation system (called CharaParser) that is designed to overcome the scalability problem in manual annotation and to promote adaptability of the system across various taxon groups .

Example3:10.1002/asi.22622-1-12 This article focuses on these ill-defined information contexts <cit> , where it aims at presenting a method and tool for information slaves to become emancipated from masters ' attempts to dictate a truth .

（2）研究方法（ Research Methods ）

Examples1:10.1002/asi.22638-2-9 Another approach is to focus on a small subset of websites to build a linking subgraph based on the small websites subset as vertices , and specifying the edges between these websites

Examples2:110.1002/asi.22620-2-37 The visibility of the articles was analysed through a normalization of the citation per paper .

Examples3:110.1002/asi.22620-2-6 Topic searches are a common technique used to analyse growth and trends in the scientific literature of specific subject areas .

（3）研究结果（ Research Results ）

Examples1:110.1002/asi.22617-4-2 In this article , we presented , based on a shifted power model , an implicit relation between the h-index and the total number of sources , and between the h-index and the impact factor (average number of items per source) .

Examples2:110.1002/asi.22617-4-3 In this way , we extended earlier work to the case that the impact factor can have a value lower than one .

Examples3:110.1002/asi.22617-4-4 Although the relation between the h-index and the impact factor looks roughly linear , their relation (in the shifted power model) is not linear at all .

（4）研究展望（ Research Prospects ）

Examples1:10.1002/asi.22617-4-5 We hope that the shifted model may help in explaining observed phenomena .

Examples2:110.1002/asi.22617-4-6 In this context , we refer to our own work where , indeed , inclusion of a zero class would probably have improved the presentation .

Examples3:110.1002/asi.22617-4-7 We hope to be able to do this in the near future .

（5）引文作者（ Citation Author ）
Examples1：10.1002/asi.22617-1-8 The fact that many journals have an impact factor lower than one illustrates this ; see also Hodge and Lacasse, to which we will return later .

Examples2：10.1002/asi.22617-1-15 The two-parameter has been studied in depth by Burrell and was also used by Glänzel.

Examples3：10.1002/asi.22617-1-16 In Egghe and Rousseau (in press) , we proved the following results for the shifted Lotka function , extending similar results for the classical case .

（6）引文时间（ Citation Time ）
Examples1：10.1002/asi.22620-1-21 By 1994 the confusion within the field had already led to a level of scepticism from scientists and policy makers who had , in the past , been keen supporters of the field .

Examples2：10.1002/asi.22654-2-8 In 2003 {ref[#asi 22654-bib-0024]} , Jansen and Spin .

Examples3：10.1002#asi.21186 During automatic indexing , frequently occurring word forms having no real purpose in describing document contents are removed for two main reasons (Manning , Raghavan , & Schütze , 2008)

（7）引文的研究问题（ Citation Research Questions）
Examples1：10.1002/asi.22650-4-36 Tibbo and Meho use Web search engines to look for finding aids based on the premise that users will easily use Web search engines to look for information .

Examples2：10.1002/asi.22650-4-37 Daniels and Yakel observe the search behaviors of users in online finding systems and what makes users ' searches successful.

Examples3：10.1002/asi.22650-4-35 Yakel studied the usability of electronic finding aids on institutional Web pages .

（8）引文的研究方法（ Citation Research Methods ）
Examples1：10.1002/asi.22618-3-4 Multiple noncontent cues such as fonts and layout information may be used to assist extraction .

Examples2：10.1002/asi.22618-3-5 Probst , Ghani , Krema , Fano , & Liu , however , used plain text descriptions as the description source .

Examples3：10.1002/asi.22618-6-48 Probst and colleagues took a classification approach to `` consistently " separate product attributes from their values .

（9）引文的研究结果（ Citation Research Results ）
Examples1：10.1002/asi.22617-1-16 In Egghe and Rousseau(in press) , we proved the following results for the , extending similar results for the classical case.

Examples2：10.1002/asi.22618-3-22 Taylor 's work shows that the special purpose syntactic parsers perform well on morphological descriptions .

Examples3：10.1002/asi.22618-6-21 The performance scores of CharaParser as shown in Table 3 exceed any previously reported character-level annotations.

（10）时间（Time）
Examples1：10.1002/asi.22620-2-7 To identify an appropriate search string, we downloaded all publications from top level bibliometric journals in the period 2001 -- 2010.

Examples2：10.1002/asi.22620-2-23 Using WoS and the search string outlined above, publications were retrieved during the period 1991 -- 2010.

Examples3：10.1002/asi.22620-2-39 The number of citations per article received by publications in each year from 1991 onwards is normalized using the average number of citations per article (CPA) received in the same year in journals included in the JCR-LIS.

（11）数据资源与平台（ Data Resources and Platforms ）
a.期刊、数据库和网站 （ journals, databases, and websites）
Examples1：10.1002/asi.22620-2-42 The adapted version is relatively easy to calculate using the online version of the Science Citation Index (SCI) .

Examples2：10.1002/asi.22620-2-23 Using WoS and the search string outlined above , publications were retrieved during the period 1991 -- 2010 .

Examples3：10.1002/asi.22620-1-33 Bibliometric journals form part of the Institute for Scientific Information Journal Citation Reports (ISI JCR) category library science and information science .

Examples4：10.1002/asi.22640-5-4 We collected data on referral keywords , visitor traffic , and advertising revenue data on BuenaMusica.com from June 1 , 2010 through October 31 , 2010.
b.公开可用的精炼数据集（ publicly available, refined, and recognized datasets ）
Examples1：10.1002/asi.22623-5-7 Three standard news corpora, Reuters 21587, TDT1 and TDT2, are used.

Examples2：10.1002/asi.22609-2-63 As suggested in previous studies, we used names and affiliations in combination to identify the right person in the DBLP.

Examples3：10.1002/asi.22650-6-80 The three most frequently used digital archival collections are the Trans-Atlantic Slave Trade Database: Voyages (here after Voyages); the Library of Congress digital collections; and the Digital National Security Archive (here after DNSA).
c.作者自己构建的数据集（ the dataset constructed by the author himself ）
Examples1：10.1002/asi.22624-2-5 We collected the data from September 2009 to October 2011 , and all author reviews posted in this period were obtained .

Examples2：10.1002/asi.22626-3-53 A subset of the data gathered from Aug City was used for this article .

Examples3：10.1002/asi.22647-2-4 Starting with this database , we extracted the hard sciences publications authored by Italian universities during the period 2004 -- 2008 , amounting to a total of 167,179.

Examples4：10.1002/asi.22609-2-73 We call this dataset EB_IS_2009 for `` editorial boards in IS listed in JCR edition 2009 .”

（12）工具（Tools）
Examples1：10.1002/asi.22618-6-43 This set of results seems to suggest that CharaParser may do away with the HR3 and HR4 adjustments , making the program more efficient .

Examples2：10.1002/asi.22618-6-44 In other words , POS-tagged domain terms alone may be sufficient in adapting Stanford Parser for the biosystematics domain .

Examples3：10.1002/asi.22618-1-43 SDD is used by key generation software such as Lucid to generate organism identification keys .

（13）模型（ Model ）
Examples1：10.1002/asi.22618-3-8 The classification algorithm used was semisupervised co-expectation maximization (EM) with supervised Naïve Bayesian classifier .

Examples2：10.1002/asi.22621-1-10 Methods relying on supervised machine learning train classifiers (support vector machines , random forests , etc.) on a hand-labeled training set containing pairs of articles where similarly named authors are identified as being the same or different persons .

Examples3：10.1002/asi.22621-1-14 Unsupervised algorithms , by contrast , require no hand-labeled training data , instead defining a metric of similarity between pairs of articles and applying an unsupervised clustering algorithm such as k-means or spectral clustering .

（14）数学公式（ Mathematical formulas ）

Examples1：10.1002/asi.22617-3-13 One may also refer to Equation (3) : if C and α are large , T is not necessarily large .

Examples2：10.1002/asi.22617-3-14 A similar remark applies to Equation (4) .

Examples3：10.1002/asi.22617-2-28 Equation (15) yields an implicit relation h (<model>IF</model>) , or equivalently IF (h) , with C as a parameter .

（15）表（ Table ）
Examples1：10.1002/asi.22617-3-15 Yet as pointed out by this referee , T is at times extremely high for this data set (see Table 1) , especially if we assume an α - value near 2 .

Examles2：10.1002/asi.22618-5-41 Table 2 shows the descriptive statistics of FNA.v19 and TIP.H test sets .

Examples3：10.1002/asi.22618-5-44 Table 3 shows the precision and recall scores on structure , character , and relation elements , as well as sentence-wise averages .

（16）图（Figure）
Examples1：10.1002/asi.22617-2-34 Figure 4 gives an example for C = 1,000 .

Examples2：10.1002/asi.22617-2-39 Figure 5 , showing h (IF) for C = 200 , reveals only the concave part .

Examples3：10.1002/asi.22617-2-16 Figure 2 shows small values of T (T = 10 ... 200) , while Figure 3 shows somewhat larger values for T (T = 100 ... 2000) .

5.2 知识图谱示例（ knowledge graph ）
Example: