PatentsView-Evaluation：推进发明人姓名消歧研究的评估数据集与工具

Olivier Binette[1,2], Sarvo Madhavan[1], Jack Butler[1],Beth Anne Card[1],Emily Melluso[1],和 Christina Jones[1]

[1]杜克大学[2]美国教育研究院

arXiv:2301.03591v1 [cs.DL] 9 Jan 2023

摘要——**我们介绍 PatentsView-Evaluation，这是一个 Python 软件包，可帮助研究人员评估发明人姓名消歧系统（如 PatentsView.org ）的性能。该软件包包含基准数据集和评估工具，旨在通过提供高质量的评估数据并提升评估标准，推进发明人姓名消歧研究。**

索引词——**数字图书馆、发明人姓名消歧、PatentsView、统计评估、开源软件**

I. 引言

发明人姓名消歧是指在专利数据集中识别不同发明人的任务（Li 等人，2014年；Toole 等人，2021年）。这需要利用上下文信息来区分姓名相同的不同发明人，并解决姓名变体问题。由于美国专利中没有发明人的唯一标识符，因此消歧需要借助统计算法来提供近似解。该任务与数字图书馆中的作者姓名消歧密切相关（Ferreira 等人，2012年；Smalheiser 等人，2009年；Subramanian 等人，2021年），也是实体消歧的一个特例（Binette andSteorts，2022年；Christen，2012年；Christophides 等人，2021年）。

遗憾的是，该领域的进展受到误导性评估方法和缺乏代表性基准数据集的阻碍（Wang 等人，2022年）。在基准数据集上直接计算性能指标（即精确率和 F分数），会导致估计产生偏差，并且在许多情况下颠倒竞争算法的排名（Binette 等人，2022年）。这是因为实体消歧的性能并非简单地随规模变化：尽管对小型基准数据集进行消歧很容易，但出错的可能性会随着数据集规模的增大而呈平方级增长。此外，一些基准数据集已经过时，或未向公众开放。

为应对这些挑战，我们发布了 PatentsView-Evaluation，这是一个 Python 软件包，可通过 github.com/patentsView/patentsView-Evaluation/ 获取，并可使用以下命令从 PyPI 安装（PyPI 作者，2022年）：

```
pip install pv-evaluation
```

这是一个开源 Python 软件包，包含一套基准数据集和评估工具，可用于代表性性能评估。该软件包包含以下数据集：美国专利商标局（USPTO）2015年消歧竞赛中使用的数据集、此前未向公众开放的 Azoulay 学术生命科学数据集，以及一个新数据集。该新数据集基于 Binette 等人（2022年）中的 PatentsView 所开发的内容扩展而来，专门用于评估。为便于开展性能评估，该软件包还包含具有代表性的精确率和召回率估计量，以及一套汇总统计和可视化工具。

本文其余部分的结构如下。第 II 节概述该软件包的各个模块，包括可用数据和性能估计量。第 III 节总结我们的贡献，并概述我们对未来研究的展望。

II. 软件包概述

PatentsView-Evaluation 基于 ER-evaluation Python 软件包（Binette，2022年）构建，该软件包提供核心的实体消歧评估功能。它包含两个主要子模块。基准测试模块提供数据、汇总统计和可视化。模板模块提供模板化报告，可使用 Quarto 发布系统（quarto.org）编译为 HTML。

A. 基准数据集

发明人消歧是将**发明人提及**关联到唯一的发明人标识符。这里，发明人提及由专利号和署名顺序编号组合而成，从而构成**提及项 ID**。例如，提及项 ID“US11379060-0”指的是美国专利号 11379060 中列出的第一位发明人。

我们的基准数据集是以发明人提及为索引、值对应唯一发明人标识符的 pandas Series（Wes McKinney，2010）。请注意，尽管基准数据集旨在提供一组发明人消歧的真实值，但由于发明人消歧本身具有不确定性且难度较高，数据集仍可能包含错误。

我们提供的发明人基准测试如下所列。可通过包中名为 load_*_inventors_benchmark() 的函数获取这些数据。

1) 文件“patents 2005 12”中的**学术生命科学（ALS）**数据集由 _ _

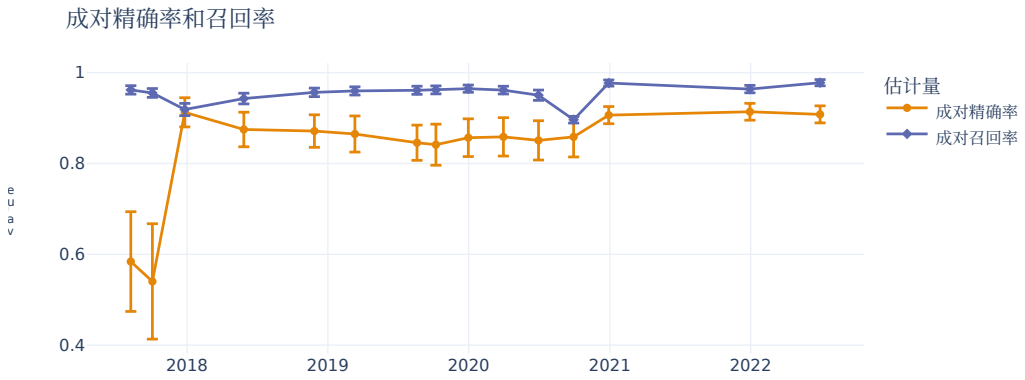

图 1. PatentsView 消歧历史中成对精确率和召回率的估计值。

Pierre Azoulay（个人通信）慷慨提供，并允许发布相应的发明人提及聚类结果。Azoulay 等（2007，2011）和 Ventura 等（2015）都曾提及该数据集及其变体。我们根据专利号和发明人提及名称，为每条记录关联了提及 ID。

2）Trajtenberg 和 Shiff（2008）的**以色列发明人基准数据集**。3）Li 等（2014）的**Li 的 2011 年发明人基准数据集**。4）来自 PatentsView 的 2015 年消歧竞赛（PatentsView，2015）的**工程师和科学家发明人基准数据集**。5）**PatentsView 的 2021 年发明人基准数据集**来自 Monath 等（2021），其中包含一组特别含混的发明人提及。6）**Binette 2022年发明人基准**，该数据集扩展自 Binette 等人（2022），涵盖 1976 年至 2021 年 12 月 31 日授予的美国专利。这是一个发明人随机样本，抽样概率与每位发明人的专利数量成正比。

B. 性能估计量

如前所述，直接计算基准数据集上的精确率和召回率会产生误导性数值。因此，PatentsView-Evaluation 借鉴 Binette 等人（2022）的方法来估计具有代表性的性能。给定一组针对 1976 年至 2021 年 12 月 31 日期间授予的美国专利的发明人消歧结果，函数 inventor_estimates_trend_plot() 会绘制每个消歧结果的精确率和召回率估计值，并提供不确定性量化（± 一个标准差）。默认情况下，这些估计值基于 Binette 2022年发明人基准。通过将其他基准数据集作为附加参数传入，可以获得与之对应的估计值。图 1 展示了使用默认参数生成的结果图。

C. 汇总统计与可视化

除了性能指标估计量之外，PatentsView-Evaluation 还基于 ER-Evaluation 软件包提供了一套汇总统计可视化工具。这些工具可以监测匹配率、姓名变体率、姓名同名率以及簇大小分布熵等指标。关于这些指标定义的更多信息，请参阅 Binette（2022）。函数 inventor_summary_trend_plot() 为可视化 PatentsView 的发明人消歧历史提供了一个入口。图 2 展示了其输出。请注意，在 2021 年前后，同名率从约 0.2 变为接近 0.4，随后又下降至接近 0.05。这些是发明人消歧中的重大变化，但匹配率并未反映出来。

D. 模板化 *HTML* 报告

PatentsView-Evaluation 的最后一个组件是模板化报告，可使用 Quarto 编译为 HTML。它可以通过汇总统计、评估指标和错误可视化，对一组发明人消歧结果进行比较。其入口是函数 render_inventor_disambiguation_report()，该函数接收一组消歧文件作为参数。

III. 讨论

本文介绍了 PatentsView-Evaluation，这是一个提供评估数据和工具、以推进发明人姓名消歧工作的 Python 软件包。我们概述了该软件包，并举例展示了它的一些功能。

PatentsView 改进发明人姓名消歧的愿景，建立在其经验和现有系统的成功基础之上。我们旨在通过分别创新其三个主要组件，提高 PatentsView 系统的可维护性、模块化程度和性能：(1) 特征工程组件，用于定义针对给定专利属性的成对比较指标；(2) 相似度建模组件，用于估算成对匹配概率；(3) 聚类组件，用于解析具有传递性的发明人聚类。对于 (1)，我们旨在开发更多特征

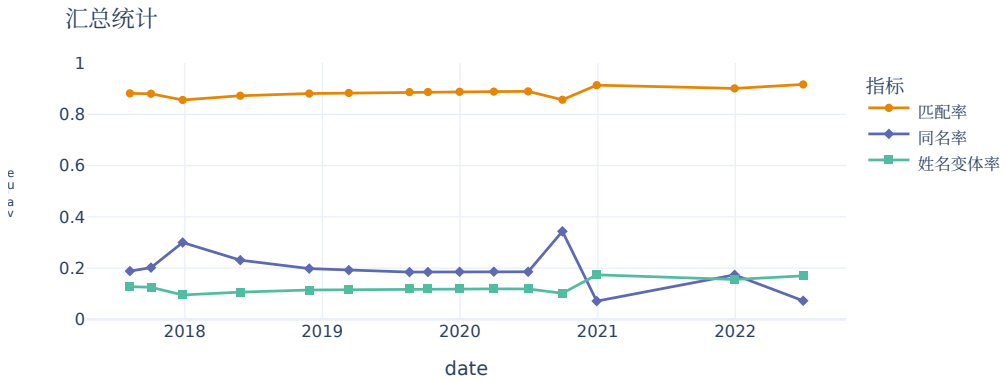

图2. PatentsView 消歧历史中汇总统计数据的演变。

，借助现代文本分析和自然语言处理方法来实现。对于 (2)，我们旨在开发灵活的半监督方法，以应对特征之间的依赖关系以及基准数据集中的偏差。最后，对于 (3)，我们旨在更好地调整聚类算法，以优化关键性能指标。借助 PatentsView-Evaluation 软件包中提供的严谨性能评估工具，现在可以对新的方法学进展进行严格测试。

参考文献

Azoulay, P., W. Ding 和 T. Stuart（2007）。教师专利申请行为的决定因素：人口统计特征还是机遇？《经济行为与组织杂志》*63*(4), 599–623。Azoulay, P., J. S. Graff Zivin 和 G. Manso（2011）。激励与创造力：来自学术生命科学领域的证据。《兰德经济学期刊》*42*(3), 527–554。Binette, O.（2022）。ER-Evaluation：实体消歧系统的端到端评估框架。可在 GitHub 上获取：https://github.com/OlivierBinette/ER-Evaluation。Binette, O. 和 R. C. Steorts（2022）。实体消歧面面观。《科学进展》*8*(12), eabi8021。Binette, O., S. A. York, E. Hickerson, Y. Baek, S. Madhavan 和 C. Jones（2022）。评估实体消歧算法的性能：从 patentsview.org 得到的经验教训。*arXiv* 电子预印本。arXiv:2210.01230。Christen, P.（2012）。《数据匹配：记录链接、实体消歧和重复项检测的概念与技术。数据中心系统与应用。柏林—海德堡：施普林格出版社。Christophides, V., V. Efthymiou, T. Palpanas, G. Papadakis 和 K. Stefanidis（2021）。大数据端到端实体消歧概述。《*ACM*计算综述》*53*(6), 1–2。Ferreira, A. A., M. A. Gonc¸alves 和 A. H. Laender（2012）。作者姓名自动消歧方法简述。《*ACM SIGMOD*记录》*41*(2), 15–26。Li, G. C., R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu 和 F. Lee（2014）。美国专利发明人数据库中的合著网络消歧及（1975–2010）。《研究政策》*43*(6), 941–955。Monath, N., C. Jones 和 S. Madhavan（2021）。PatentsView：发明人、受让人和地点的消歧。技术报告，美国教育研究院，弗吉尼亚州阿灵顿。PatentsView（2015）。发明人消歧研讨会总结。在线获取：https://patentsview.org/events/workshop-2015。PyPI 作者（2022）。Python 软件包索引 - PyPI。Smalheiser, N. R., V. I. Torvik 等（2009）。作者姓名消歧。《信息科学与技术年度评论》*43*(1), 1。Subramanian, S., D. King, D. Downey 和 S. Feldman（2021）。S2and：作者姓名消歧的基准测试与评估系统。载于*2021*年 *ACM/IEEE*数字图书馆联合会议（*JCDL*），第 170–179 页。Toole, A., C. Jones 和 S. Madhavan（2021）。PatentsView：推动科学技术政策发展的开放数据平台。技术报告，USPTO 经济学工作论文。Trajtenberg, M. 和 G. Shiff（2008）。以色列专利发明人的识别与流动。技术报告，Pinhas Sapir。Ventura, S. L., R. Nugent 和 E. R. Fuchs（2015）。关注非明星：以往消歧方法中的一些偏差来源，以及一种利用带标签记录的新型公共工具。《研究政策》*44*(9), 1672–1701。Wang, T., H. Lin, C. Fu, X. Han, L. Sun, F. Xiong, H. Chen, M. Lu 和 X. Zhu（2022）。弥合实体匹配现实与理想之间的差距：重新审视与基准重构。*arXiv* 电子预印本。arXiv:2205.05889。Wes McKinney（2010）。Python 中的统计计算数据结构。载于 St´efan van der Walt 和 Jarrod Millman（编），第九届 *Python*科学会议论文集，第 56–61 页。