Postprint: Research on R&D Team Identification Based on Patent Inventor Name Disambiguation

Authors: Zhang Jing, Zhang Zhiqiang, Zhao Yajuan

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] Talent is the core of technology research and development. R&D teams are the key focus of technological development in various fields and an important manifestation of an institution's R&D strength.

[Method / Process] Taking patent documents from the Derwent Innovation Index (DII) as the analysis object, clarifying the rules for inventor name disambiguation, using inventor co-occurrence clustering to identify major R&D teams, and then conducting an empirical analysis of R&D team identification after name disambiguation using patents related to digital light processing in 3D printing.

[Results / Conclusion] It is demonstrated that inventor name disambiguation in patents facilitates accurate analysis of inventor patent counts.

Full Text

Identification of R&D Teams Based on Patent Inventor Name Disambiguation

Zhang Jing[1],[2],[3], Zhang Zhiqiang[4], Zhao Yajuan[1]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]University of Chinese Academy of Sciences, Beijing 100049
[3]Archives of Chinese Academy of Sciences, Beijing 100190
[4]Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu 610041

Abstract

[Purpose/Significance] Talent constitutes the core of technological research and development. R&D teams represent both the primary focus of technological advancement across various fields and a critical indicator of an institution's

research capabilities. [**Method/Process**] This study employs Derwent Innovation Index (DII) patent documents as its analytical corpus, establishes explicit rules for inventor name disambiguation, and identifies key R&D teams through inventor co-occurrence clustering. An empirical analysis is then conducted on patents related to Digital Light Processing (DLP) technology in 3D printing to validate the proposed disambiguation approach and team identification methodology. [**Result/Conclusion**] The findings demonstrate that patent inventor name disambiguation significantly enhances the accuracy of patent quantity analysis for individual inventors, thereby improving the precision of R&D team identification.

Keywords: patents; inventors; R&D team identification; name disambiguation
Classification Numbers: G353.1; G306

---

1. Research Progress on Name Disambiguation

The essence of technological R&D lies in human talent. In today's information-rich environment, the exponential growth of data makes collaborative research teams indispensable for technological advancement. When formulating specific policies such as talent recruitment, decision-makers must focus not only on principal investigators but also on key personnel who play central roles within R&D teams. Identifying R&D teams constitutes a crucial component of patent analysis, as it facilitates the recognition of core team members and the discovery of non-lead critical talent, thereby providing enhanced support for policy formulation and key researcher identification. However, researcher names exhibit strong ambiguity, manifesting as both polysemy (multiple individuals sharing the same name) and synonymy (the same individual represented by different name formats). Consequently, name disambiguation becomes the foremost challenge in R&D team identification research, with accuracy serving as its core objective.

Name disambiguation research treats cross-document name coreference as a fundamental problem. A. Bagga et al. [1] initiated explorations into cross-document name disambiguation as a coreference resolution task in 1998. The Web People Search (WePS) evaluation workshops in 2007, 2009, and 2010 specifically addressed web-based name disambiguation. In China, the CIPS-SIGHAN-2012 conference [2] stimulated growing research on Chinese name recognition and disambiguation. Numerous studies have focused on extracting entity features from web resources and performing clustering for name disambiguation, employing methods such as social network analysis, threshold determination principles, and probabilistic approaches. For instance, G. Mann et al. [3] in 2003 constructed feature vectors by customizing templates to extract personal biographical information from web pages. M. B. Fleischman et al. [4] in 2004 extracted name features, webpage features, overlap features, and semantic features, applying a maximum entropy model to calculate the probability that two names refer to the same entity. B. Malin [5] in 2005 proposed a social network-based ap-

proach to name disambiguation. K. Balog et al. [6] in 2007 utilized trained language models to compute the probability that a name in a webpage referred to a specific entity, then determined thresholds for disambiguation. Y. Chen et al. [7] in 2007 extracted noun phrase-based and named entity-based features, subsequently applying hierarchical agglomerative clustering. S. Ono et al. [8] in 2008 performed document clustering using hybrid features including named entity coreference, keywords, and topical information. L. Romano et al. [9] in 2009 introduced the XMedia system employing a quality threshold clustering algorithm. Zhang Shunrui et al. [10] in 2010 applied hierarchical clustering algorithms for Chinese name disambiguation, while Chen Chen et al. [11] in 2011 investigated Chinese name disambiguation by examining the impact of different social network edge weights and graph partitioning criteria.

As name disambiguation research has matured, studies targeting specific data sources and employing multi-method, stepwise approaches have increased to improve accuracy. In 2012, Yang Xinxin et al. [12] expanded feature documents using four types of search engine query rules and applied a two-layer clustering algorithm [13] for name disambiguation. In 2013, Li Guangyi et al. [14] set weights according to feature types and performed multiple clustering iterations. In 2014, S. Christian et al. [15] constructed social network graphs using citation relationships among database documents for source-specific name disambiguation. In 2015, Yang Yilin et al. [16] integrated three different clustering algorithms using contextual, entity, and social relationship features to enhance disambiguation accuracy. D. H. Han et al. [17] employed extreme learning machines to propose clustering algorithms for individual names and name sets. M. Song et al. [18] constructed specialized training sets for the PubMed database and proposed new publication feature sets to improve accuracy.

Overall, current research primarily focuses on web resources or academic authors, with methodological emphasis on extracting more name-related features through algorithmic improvements or employing multiple/multi-layer clustering for comparative judgment. These approaches inherently contain certain disambiguation errors, which are directly determined by algorithms without clear identification of potentially erroneous name instances, thus creating a “black box” problem where analysts cannot ascertain the scope of potential errors.

Research specifically addressing patent document characteristics for inventor name disambiguation remains scarce. Patent inventor name formats vary across databases, typically involving both Chinese and foreign name disambiguation. Furthermore, patent inventor name disambiguation for policy support must prioritize accuracy while improving efficiency. Therefore, patent-based name disambiguation requires specific exploration grounded in the structural characteristics of inventor names within patent databases to enhance accuracy and reduce error uncertainty associated with “black box” problems.

2. Patent Inventor Name Disambiguation

Derwent Innovation Index (DII) represents intellectually processed patent data offering batch accessibility, natural language search capabilities, and unified reclassification of patents from diverse sources, making it a commonly used resource for patent analysis. This study focuses on this database, supplemented by Thomson Innovation (TI) patent database features, to develop patent inventor name disambiguation rules based on inventor affiliation and country information.

2.1 Name Disambiguation Process This study primarily compares name similarity based on inventor name structural features, then utilizes available inventor characteristic information from patent documents to make disambiguation judgments, as illustrated in [Figure 1: see original paper].

2.2 Patent Inventor Name Structural Features and Their Impact The structural characteristics of inventor names from different countries affect name ambiguity differently. Analysis of actual data reveals two primary name structure categories: Western-style names and Chinese-style names, as shown in . These structural differences determine that Western-style names exhibit higher probabilities of different representations referring to the same person, while Chinese-style names show higher probabilities of identical representations referring to different individuals.

2.3 Patent Inventor Feature Information In DII and TI databases, available patent inventor feature information includes name abbreviations, full names, addresses (including country information), patent accession numbers, affiliations, and collaborators, as detailed in . The completeness of this information varies across databases: (1) TI generally contains more complete name information than DII; (2) However, TI's full name field sometimes contains the same values as the abbreviation field, indicating incompleteness; (3) Country information in addresses is more complete than city-level information; (4) Patent accession numbers and collaborator information are relatively complete.

2.4 Name Disambiguation Rules Name disambiguation first requires identifying potentially ambiguous name representations through similarity assessment, with specific criteria presented in . Notably, this assessment disregards punctuation such as periods and hyphens in name representations. Based on the patent inventor name structural features () and available feature information (), and validated through actual data, the following disambiguation rules can be constructed by priority for Western-style and Chinese-style names.

2.4.1 Western-Style Name Disambiguation Rules Given the structural characteristics of Western-style names, the disambiguation focus lies in consolidating multiple name representations of the same individual into a single

representation. Therefore, disambiguation proceeds from name abbreviations as the entry point, which both excludes representations belonging to different individuals and maximizes the inclusion of representations for further judgment. Specific rules are described in .

2.4.2 Chinese-Style Name Disambiguation Rules For Chinese-style names, the disambiguation focus is distinguishing different individuals who share identical name representations. Again starting from name abbreviations to differentiate non-identical individuals, specific rules are described in .

It is important to note that the disambiguation process yields probabilistic rather than deterministic conclusions. The rules require presenting specific entries for extended verification supplemented by manual judgment to reach final conclusions. Following disambiguation, primary R&D team identification standards can be established based on co-owned patent quantities or proportions, enabling team identification through inventor co-occurrence clustering.

3. Empirical Study on DLP R&D Team Identification Based on Name Disambiguation

This study employs patents related to Digital Light Processing (DLP) technology in 3D printing as a case study to empirically validate the proposed name disambiguation and team identification methodology.

3.1 Comparative Statistics Before and After Name Disambiguation Following retrieval and expert review, 274 patents (810 patent families) related to DLP technology were obtained from the DII database. After deduplication by Derwent accession number and inventor name representation, the original DII data involved 640 inventor name representations, while TI data involved 652 representations. Applying the disambiguation rules described in Section 2.4 and counting name abbreviations from TI data, 120 inventor name representations were found to have multiple formats for the same individual, while 90 representations corresponded to cases of multiple individuals sharing the same name. The final determination identified 602 unique inventors involved in DLP technology development.

A comparison of primary inventors (those with more than 3 patents) before and after disambiguation reveals changes in patent counts. For instance, HULL CHARLES W's patent count increased from 5 to 6, and KRITCHMAN Eliahu M.'s count increased from 4 to 5 (highlighted in), demonstrating that name disambiguation yields more accurate rankings and statistics for primary inventors.

3.2 R&D Team Identification After Name Disambiguation Building upon the disambiguated data, Bibexcel was used to generate an inventor co-occurrence matrix and node data for visualization. Pajek software then produced the inventor clustering network shown in [Figure 2: see original paper].

Among the 602 inventors in the DLP technology field, 63 participated in the clustering analysis. Based on the data, this study defines an R&D team as comprising at least 3 inventors.

The network clearly reveals 7 R&D teams from 6 institutions in the DLP technology domain. details these teams. Notably, the two teams from HUNTSMAN have no connecting personnel in the DLP field and are thus clearly distinguished. The team from 3D SYSTEMS INC comprises 11 members, which can be further divided into two sub-teams (labeled A and B in), with HULL CHARLES W and PARTANEN JOUNI P serving as the connecting links between them, appearing as a large cluster in [Figure 2: see original paper].

3.3 Empirical Study Summary Since the name disambiguation rules proposed in this study are tailored to the specific data structures of particular databases, they lack universal applicability; therefore, no performance evaluation of the rules themselves was conducted. However, the comparison in Section 3.1 demonstrates that name disambiguation improves the accuracy of patent count statistics for primary inventors, reducing identification errors in R&D teams caused by uncertainty regarding name coreference and facilitating more precise team identification.

4. Conclusion

The accuracy of name disambiguation directly impacts the reliability of patent analysis results, which in turn influences competitor identification and talent policy decisions based on such analyses. Consequently, name disambiguation represents a critical issue that must be addressed as patent analysis continues to deepen.

This study argues that name disambiguation in patent R&D team identification should prioritize accuracy. The proposed rules adopt feature vector similarity determination concepts but differ from other methods in two key aspects: First, they are derived from specific patent database structural characteristics, making them more targeted; second, they supplement probabilistic rules that cannot yield deterministic conclusions with manual judgment, thereby avoiding the "black box" problem of uncertain errors inherent in direct algorithmic determination.

Empirical validation demonstrates that the proposed disambiguation rules enhance the accuracy of inventor patent count analysis. However, it should be noted that these rules are based on specific patent document data and thus have limited general applicability, though they offer greater accuracy for the targeted data. Future research should further explore and refine name disambiguation methods, expand the applicable scope of these rules, and thereby improve R&D team identification.

References

[1] BAGGA A, BALDWIN B. Entity-based cross-document conferencing using the vector space model[C]//COLING'98: Proceedings of the 17th international conference on computational linguistics. New York: ACM Press, 1998: 79-85.

[2] China Academic Conference Online. The 2nd CIPS-SIGHAN International Conference on Chinese Language Processing[EB/OL]. [2014-04-10]. http://www.meeting.edu.cn/meeting/meetingAction-29689!detail.action.

[3] MANN G, YAROWSKY D. Unsupervised personal name disambiguation[C]//CONLL' 03: Proceedings of the 7th conference on natural language learning at HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003: 33-40.

[4] FLEISCHMAN M B, HOVY E. Multi-document person name resolution[EB/OL]. [2014-03-14]. http://acl.ldc.upenn.edu/W/W04/W04-0701.pdf.

[5] MALIN B. Unsupervised name disambiguation via social network similarity[C]//Proceedings of 2005 SIAM international conference on data mining. Newport Beach: Siam Workshop on Link Analysis, 2005: 93-102.

[6] BALOG K, AZZOPARDI L, RIJKE M D. UVA: Language modeling techniques for web people search[C]//Proceedings of the 4th international workshop on semantic evaluations. Prague: International Workshop on Semantic Evaluations, 2007: 468-471.

[7] CHEN Y, MARTIN J. Towards robust unsupervised personal name disambiguation[EB/OL]. [2014-03-14]. http://www.dsi.unifi.it/~jain/…/pdf?origin=publication_{detail}.

[8] ONO S, SATO I, YOSHIDA M, et al. Person name disambiguation in web pages using social network, compound words and latent topics[C]//Proceedings of the 12th pacific-asiaconference on advances in knowledge discovery and data mining. Berlin: Pacific-asiaconference on advances in knowledge discovery and data mining, 2008: 260-271.

[9] ROMANO L, BUZA K, GIULIANO C. XMedia: Web people search by clustering with machinelylearned similaritymeasures[EB/OL]. [2014-03-14]. https://www.researchgate.net/publication/228569058_{{{XMedia}}}{{Web}}{{{People}}}{{Search}}{{{by}

[10] ZHANG Shunrui, YOU Hongliang. Chinese personal name disambiguation based on hierarchical clustering algorithm[J]. New Technology of Library and Information Service, 2010(11): 64-68.

[11] CHEN Chen, WANG Houfeng. Cross-document name disambiguation based on social network[J]. Journal of Chinese Information Processing, 2011(5): 76-82.

[12] YANG Xinxin, LI Peifeng, ZHU Qiaoming. Person name disambiguation based on query expansion[J]. Computer Applications, 2012, 32(9): 2488-2490, 2507.

[13] YANG Xinxin, LI Peifeng, ZHU Qiaoming. Person name disambiguation based on dependency features of web text[J]. Computer Engineering, 2012(19): 133-136.

[14] LI Guangyi, WANG Houfeng. Chinese named entity recognition and disambiguation based on multi-step clustering[J]. Journal of Chinese Information Processing, 2013, 27(5): 29-34.

[15] CHRISTIAN S, AMIN M, ALEXANDER M P, et al. Exploiting citation networks for large-scale author name disambiguation[J]. EPJ data science, 2014, 3(11): 1-12.

[16] YANG Yilin, ZHOU Jie, LI Bicheng. Person name disambiguation algorithm based on clustering ensemble[J/OL]. Application Research of Computers, 2015: 33. [2016-05-30]. http://www.cnki.net/kcms/detail/51.1196.TP.20151028.1121.120.html.

[17] HAN D H, LIU S Q, HU Y C, et al. ELM-based name disambiguation in bibliography[EB/OL]. [2016-04-13]. http://link.springer.com/article/10.1007%2Fs11280-013-0226-4.

[18] SONG M, KIM E H, KIM H J. Exploring author name disambiguation on PubMed-scale[J]. Journal of informetrics, 2015(4): 924-941.

Author Contributions

Zhang Jing: Conceived the research idea, prepared data, and wrote the manuscript; **Zhang Zhiqiang:** Refined the research framework and provided revision suggestions; **Zhao Yajuan:** Refined the research framework and provided revision suggestions.

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