Research on Author Name Disambiguation Based on Semantic Fingerprint

Introduction

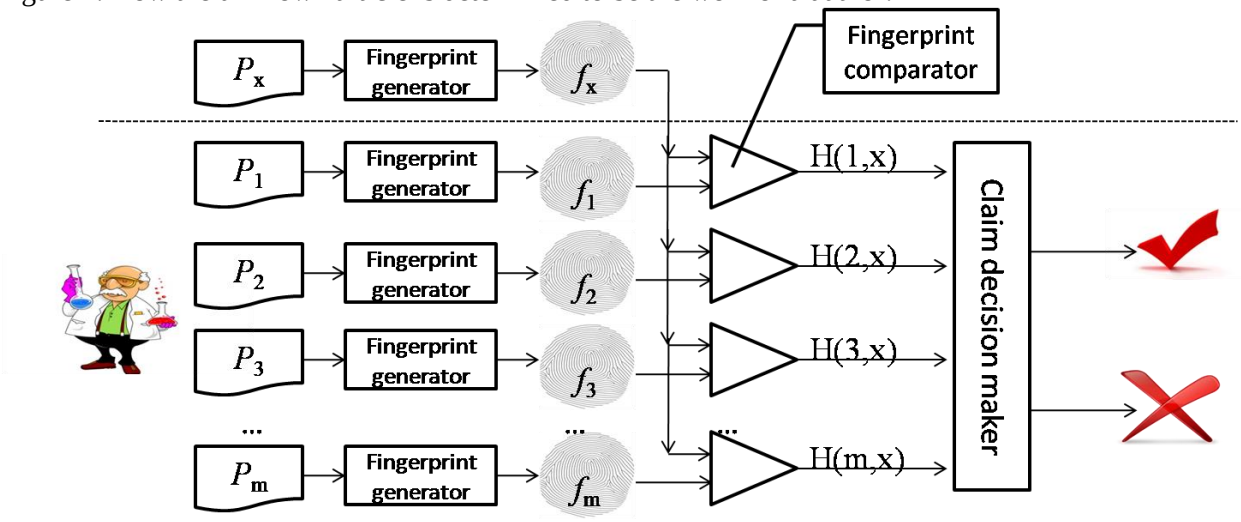
Author name ambiguity is a kind of uncertain phenomenon, where authors of scholarly documents often share names which makes it hard to distinguish each author's work [1, 4]. Due to the homonym, readers or analysts are often confused when they search literatures [2]. The author ambiguity problems have been an obstacle to efficient information retrieval in digital library age, causing incorrect identification between authors and their publications [5]. Author name disambiguation is a fundamental step in mapping knowledge domains and in other bibliometric and scientometrics analyses. It has great practical application value, which makes great influence on marketers who wish to direct their advertisements to specific individuals. And also it is crucial for establishing new resources such as co-author networks, citation networks and collaboration networks. In the personalized search, automatic question answering, multi-document summarization, hot figure tracking and discovery, and other fields have been widely applied. Our goal is to find all publications that belong to a given author and distinguish them from publications of other authors who share the same name, and also sort out the erroneous entities due to name disambiguation in a fast and efficient way. The existing methods are usually unable to meet the demand of practical application, especially under the condition of rapid growth of the scientific literature [3].

Methods

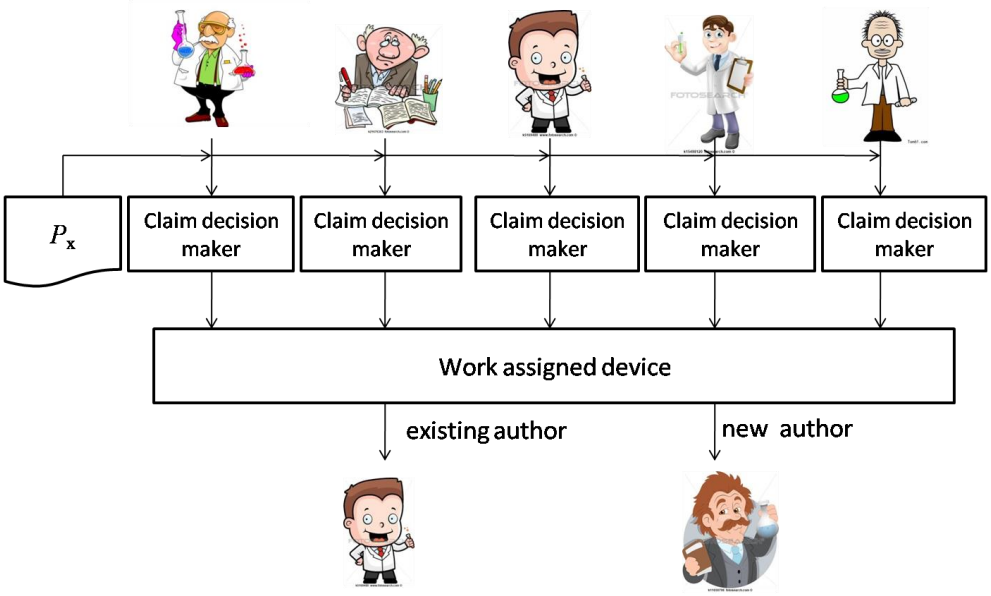
To deal with author name disambiguation, we design a mechanism to generate fingerprint for each article, and compare the fingerprint of a new article with fingerprints of articles having same-name author. Then we assign the new article to a certain author or a new one. If the new article is assigned to two or more authors, we use a arbiter to determine which author should have the article.

In this work, firstly, according to the characteristics of the scholarly documents, we extract the email addresses and the affiliations of authors from pre-processed documents. Then fingerprint generator is used to generate email fingerprints, affiliation fingerprints and text fingerprints, where text fingerprint is generated by semantic fingerprint algorithm (such as simhash) using the full text of a document. After that, the fingerprint comparator is used to compare fingerprint of an unknown article with fingerprints of same-name articles in database, where same-name authors in database have been disambiguated. Then claim decision maker is used to judge the unknown article belongs to which author in database or it is a new same-name author according to the result of the comparator. Fig.1 shows the process how the unknown article P_x is determined to be the work of a author or not.

Figure 1: How the unknown article is determined to be the work of a author.



If there are several same-name authors in database, we use a unit called work assigned device trying to assign the unknown article to one of them. Fig.2 shows the process of an unknown paper assignment.
Figure 2: The process of an unknown paper assignment.



When the unknown article is claimed, there may be three results as follows：(1) it is assigned to one author: this article belongs to this author; (2) it is assigned to two or more authors: we will use a unit called arbiter to deal with the problem. After arbitration, the article will be assigned to one of these authors, or submitted for manual handling; (3) it is not assigned: the article may belong to a new author.

Conclusions

In this paper, we propose name disambiguation method based on semantic fingerprint, the whole process does not involve the comparison of the original text, and the full text similarity is converted into the comparison of fixed length fingerprint. The method can dynamic build a fingerprint database and support incremental disambiguation instead of clustering all papers with same name authors in traditional methods.

References

[1] Anderson A. Ferreira, Marcos Andre Goncalves, and Alberto H. F. Laender. A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, 41(2):15–26, 2012.

[2] Hui Han, L. Giles, Hongyuan Zha, and C. Li. Two supervised learning approaches for name disambiguation in author citations. In *Acm/ieee-Cs Joint Conference on Digital Libraries*, pages 296–305, 2004.

[3] Neil R. Smalheiser and Vetle I. Torvik. Author name disambiguation. *Annual Review of Information Science & Technology*, 43(1):1–43, 2009.

[4] A Strotmann, D Zhao, and T Bubela. Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science & Technology*, 46(1):1–20, 2009.

[5] V. I. Torvik and N. R. Smalheiser. Author name disambiguation in medline. *Acm Transactions on Knowledge Discovery from Data*, 3(3):1–29, 2009.