World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:10, 2017

Cleaning of Scientific References in Large Patent Databases Using Rule-Based Scoring and Clustering

Authors : Emiel Caron

Abstract : Patent databases contain patent related data, organized in a relational data model, and are used to produce various patent statistics. These databases store raw data about scientific references cited by patents. For example, Patstat holds references to tens of millions of scientific journal publications and conference proceedings. These references might be used to connect patent databases with bibliographic databases, e.g. to study to the relation between science, technology, and innovation in various domains. Problematic in such studies is the low data quality of the references, i.e. they are often ambiguous, unstructured, and incomplete. Moreover, a complete bibliographic reference is stored in only one attribute. Therefore, a computerized cleaning and disambiguation method for large patent databases is developed in this work. The method uses rule-based scoring and clustering. The rules are based on bibliographic metadata, retrieved from the raw data by regular expressions, and are transparent and adaptable. The rules in combination with string similarity measures are used to detect pairs of records that are potential duplicates. Due to the scoring, different rules can be combined, to join scientific references, i.e. the rules reinforce each other. The scores are based on expert knowledge and initial method evaluation. After the scoring, pairs of scientific references that are above a certain threshold, are clustered by means of single-linkage clustering algorithm to form connected components. The method is designed to disambiguate all the scientific references in the Patstat database. The performance evaluation of the clustering method, on a large golden set with highly cited papers, shows on average a 99% precision and a 95% recall. The method is therefore accurate but careful, i.e. it weighs precision over recall. Consequently, separate clusters of high precision are sometimes formed, when there is not enough evidence for connecting scientific references, e.g. in the case of missing year and journal information for a reference. The clusters produced by the method can be used to directly link the Patstat database with bibliographic databases as the Web of Science or Scopus.

Keywords : clustering, data cleaning, data disambiguation, data mining, patent analysis, scientometrics
Conference Title : ICCSIB 2017 : International Conference on Cybermetrics, Scientometrics, Informetrics and Bibliometrics
Conference Location : Barcelona, Spain
Conference Dates : October 30-31, 2017