TCI: Towards a Rigorous Citation Index
Shu-hsien TSENG¹, Tzu-heng CHIU², Chia-ning CHIANG³*

Abstract
Building a citation database like the Taiwan Citations Index is very different from conventional bibliographic databases. When building citation indexes, citation entries are found duplicated and variant by nature at the time of data entry. Previous research showed that the earlier years of ISI citation databases still have many unresolved errors and variations. Although Scopus claims better consistency than WoS, researchers also found incomplete editing of overlapping sources for the article references. It is unclear whether they utilize tools to automatically recognize citations or process duplicates. However, it is essential to find a rigorous method to digest the large amount duplicate into a day-to-day process, as well as an effective way to apply the standard record in place of variant citations for the betterment of citation quality.

For accuracy purpose, the TCI team adopted the idea of checking references against the original sources to develop the Search-Apply-Merge module. Consequently, in the TCI system, source records may be used as the benchmark record for citations. This method not only de-duplicates but also improves the accuracy of resulting citation entries. This paper describes the design concept and functions to digest duplicates and obtain better quality on-the-fly for data entry.

1. The Problem
- Citation entries are found duplicated and variant by nature
  - High error rates in major fields & numerical data
  - Inaccurate, sloppy referencing, external characteristics of journals, referencing conventions, language problems, author identification problems, unfamiliarity with foreign author names, data capturing procedures (Mooi, 2005)
- Inaccuracy hinders reliability and credibility
  - The reliability of citation-based analyses
  - The retrieval of all of the work by an author for getting credit for his/her work (Oermann et al, 2001)
- Error correction solutions
  - Manually corrections and/or tools (WoS, KCI)
  - Autonomous (CiteSeer)
  - Dictionaries (CSSCI)

2. Design Approach
Aim: collapsing duplicates while also obtain better quality on-the-fly
- Questions to be answered:
  1. How to get correct citation counts?
  2. How to standardize the duplicated citations with errors and variants?
  3. What would be feasible for de-duplicates and benchmarking? Computer algorithm or manual?
  4. How to solve duplicates entered on-the-fly?
  5. How to handle the collapses happening in parallel?
  6. How should the Benchmark Record be designed and utilized?

3. Search Duplicates and Collapse Module
1. A working area to support on-the-fly collapsing
2. The identification of a standard record
3. An Apply-Standard function to point the variants to the standard record
4. The variant records make use of the data in standard record
5. Using overnight batch process for merging all collapses

4. Results: Search-Apply & Collapse
   - 2013.2 TSSCI & THCI converted into the TCI
   - 2013.5 TCI: Towards a Rigorous Citation Index
   - 2013.6 Batch comparison and grouping (approx. 10,000 groups)
   - "Times Cited" generated

Collapsing Duplicates -- Quality Control

---

1. Director General, National Central Library
2. Director, Knowledge Services Division, National Central Library
3. TCI Project Coordinator, Knowledge Services Division, National Central Library

*Email*: cnchiang@ncl.edu.tw