An Inductive Method for "Term Clumping": A Case Study on Dye-Sensitized Solar Cells

Yi Zhang [1], Alan L. Porter [2], Zhengyin Hu[3]

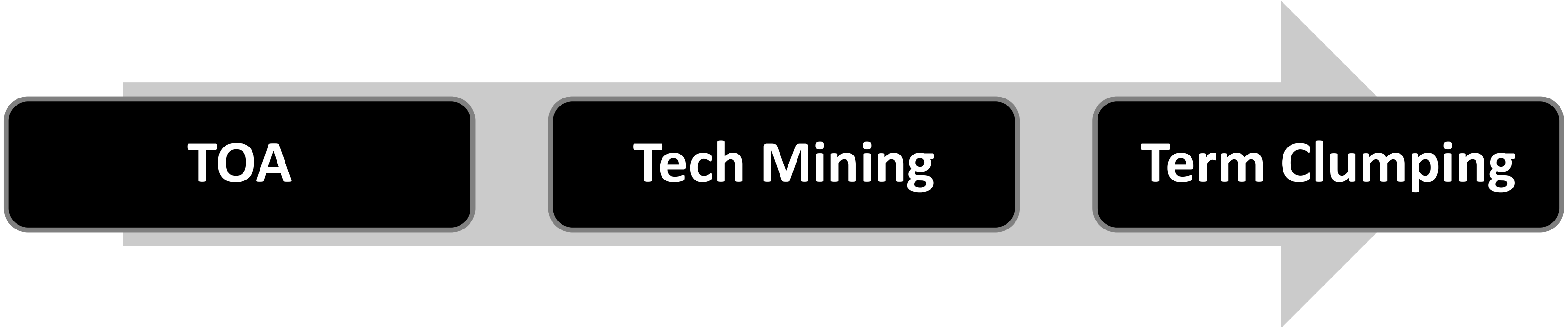
[1] Beijing Institute of Technology

[2] Georgia Institute of Technology and Search Technology, Inc.

[3] Chinese Academy of Sciences

Introduction: Term Clumping

- Technology Opportunities Analysis (TOA) and Tech Ming
  - Approaches for retrieving usable information on the prospects of particular technological innovations from Science Technology and Innovation (ST&I) resources.
  - Focus on processing huge search results from ST&I databases. Such searches provide terms that can indicate significant topics during the emergence of a technology.
  - Aim to explore the methods of cleaning and consolidating the rich sets of topical phrases in order to generate “better topical phrases” for further analyses.

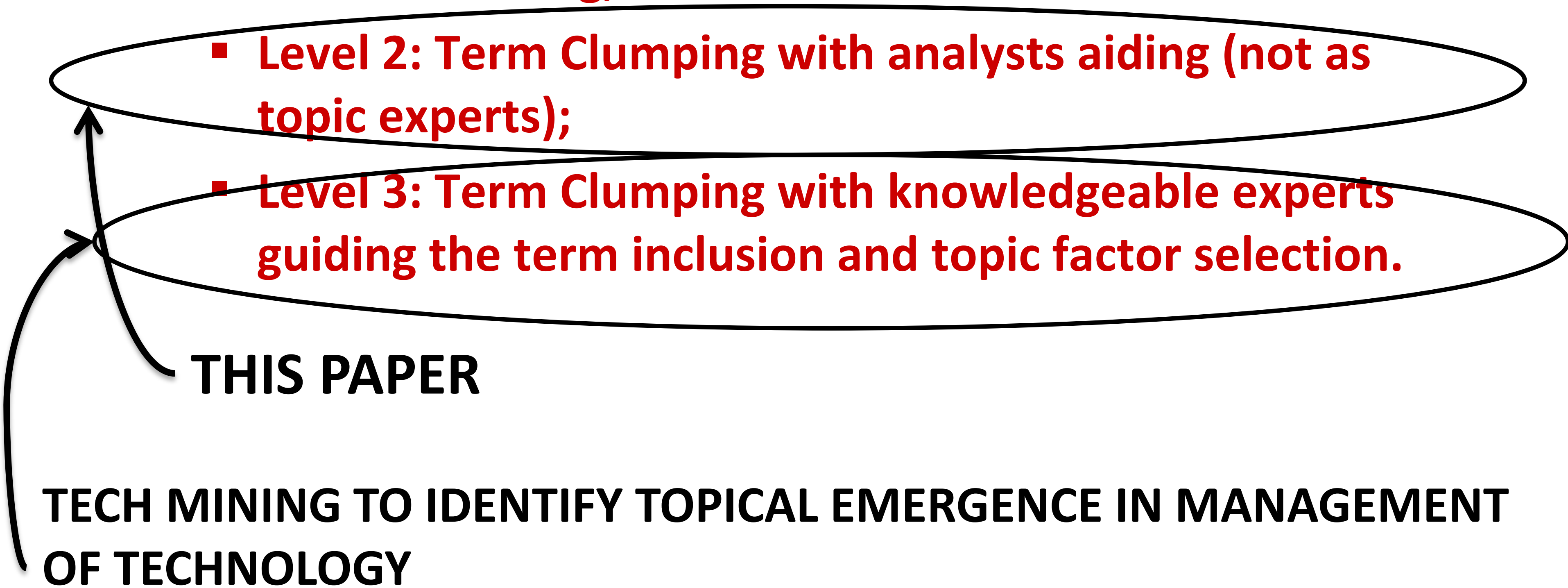


Introduction: Term Clumping

- Term clumping
  - The steps to clean and consolidate rich sets of topical phrases and terms, which pertain to a technology under study, in a collection of documents.
- Definitions
  - Experts: Professional researchers who are broadly knowledgeable across the specific domain;
  - Analysts: Professional researchers in data retrieval and analysis who have analytical skills in handling text;
  - Technicians: Software operators who follow the analysts' guidance, operate the software, and are able to program for specific scripts or functions as needed.

Introduction: Term Clumping

- 3-Level Human Intervention Model
  - Level 1: Automated Term Clumping with almost no human cleaning;



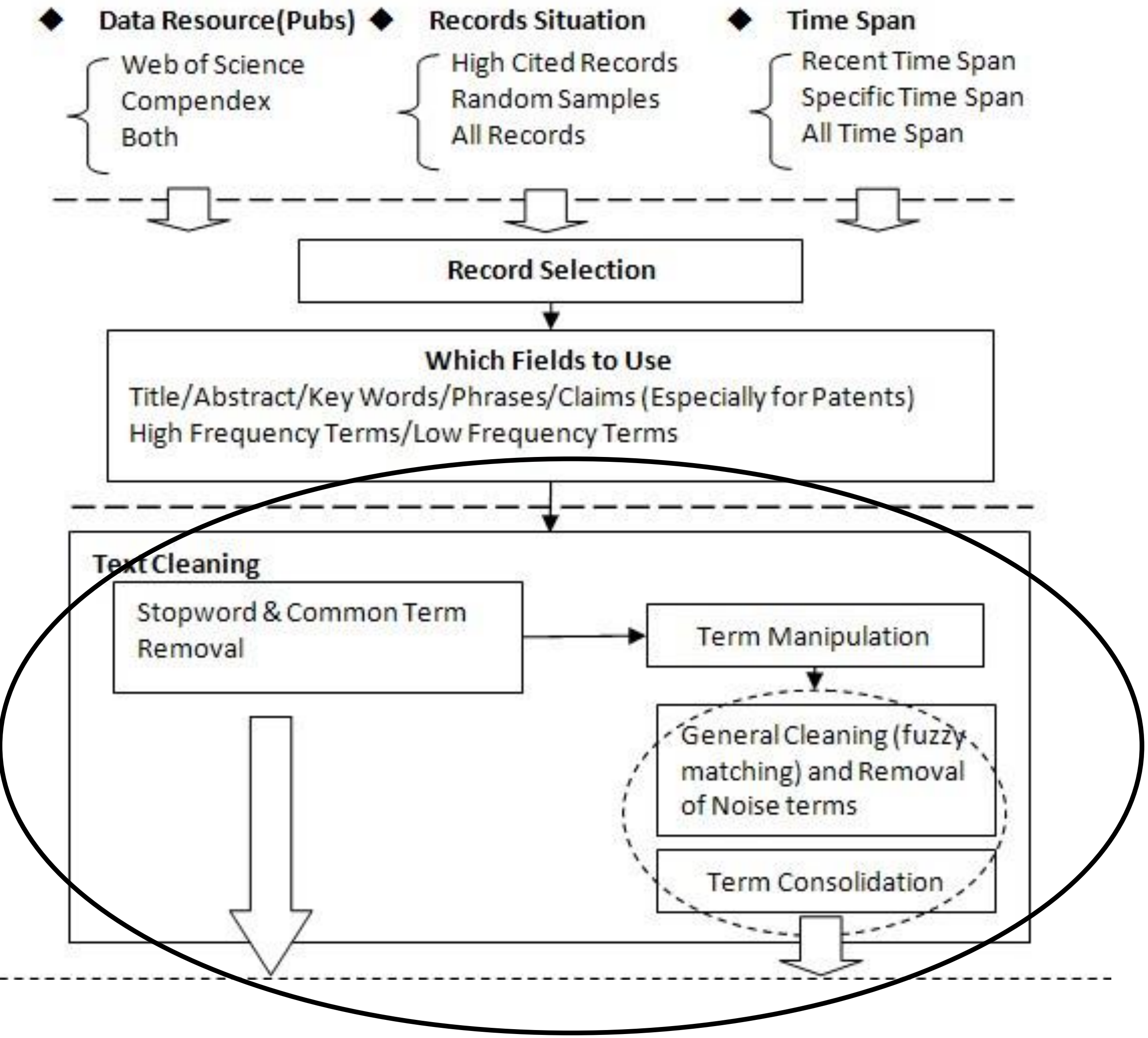
  - Level 2: Term Clumping with analysts aiding (not as topic experts);
  - Level 3: Term Clumping with knowledgeable experts guiding the term inclusion and topic factor selection.

THIS PAPER

TECH MINING TO IDENTIFY TOPICAL EMERGENCE IN MANAGEMENT OF TECHNOLOGY

Alan L. Porter, Yi Zhang, Nils C. Newman

Methodology: Framework for Term Clumping



◆ Data Resource(Pubs)

Web of Science
Compendex
Both

◆ Records Situation

High Cited Records
Random Samples
All Records

◆ Time Span

Recent Time Span
Specific Time Span
All Time Span

Record Selection

Which Fields to Use
Title/Abstract/Key Words/Phrases/Claims (Especially for Patents)
High Frequency Terms/Low Frequency Terms

Text Cleaning

Stopword & Common Term Removal

Term Manipulation

General Cleaning (fuzzy matching) and Removal of Noise terms

Term Consolidation

- **Software:** ***VantagePoint***

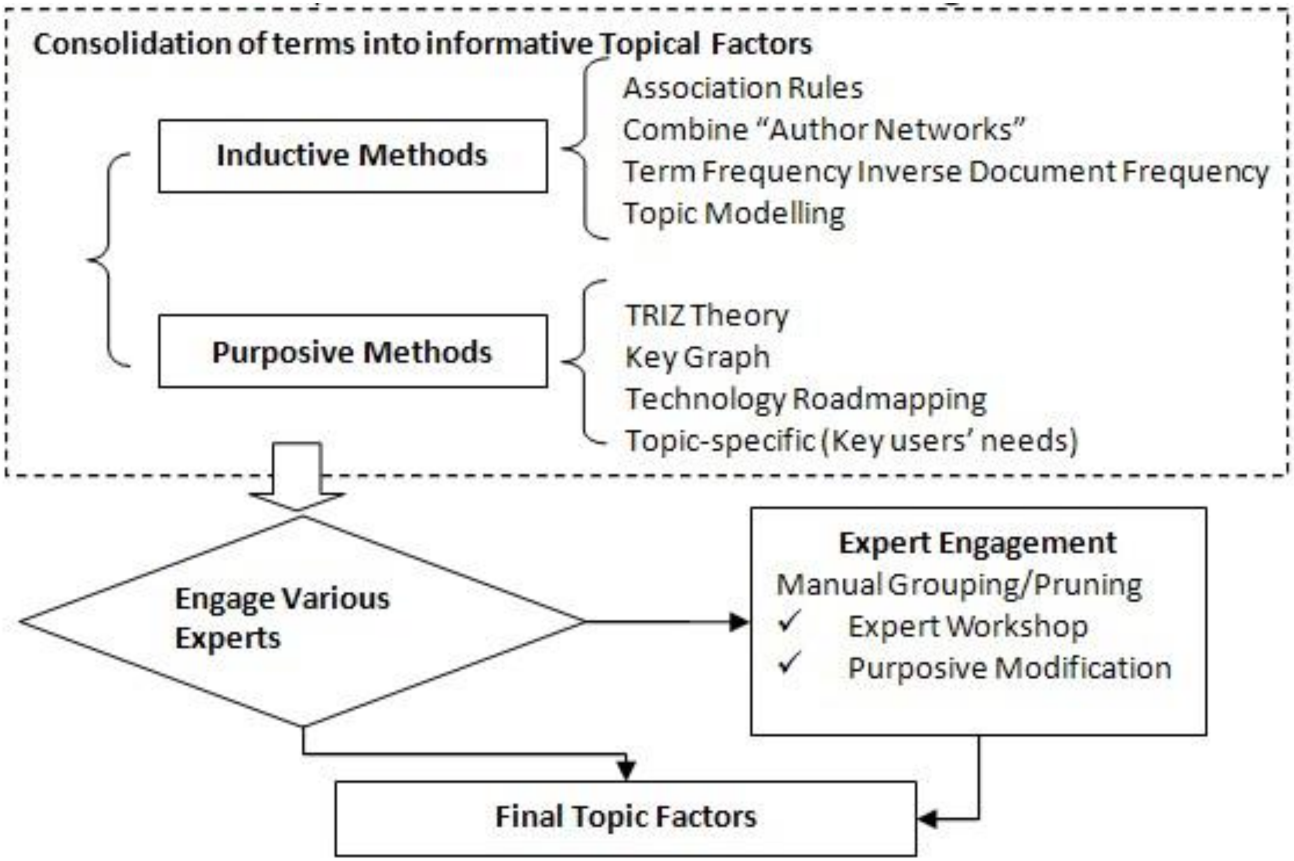
□ ***Thesaurus***

□ ***Macros***

- **Human aiding:** ***Analysts***

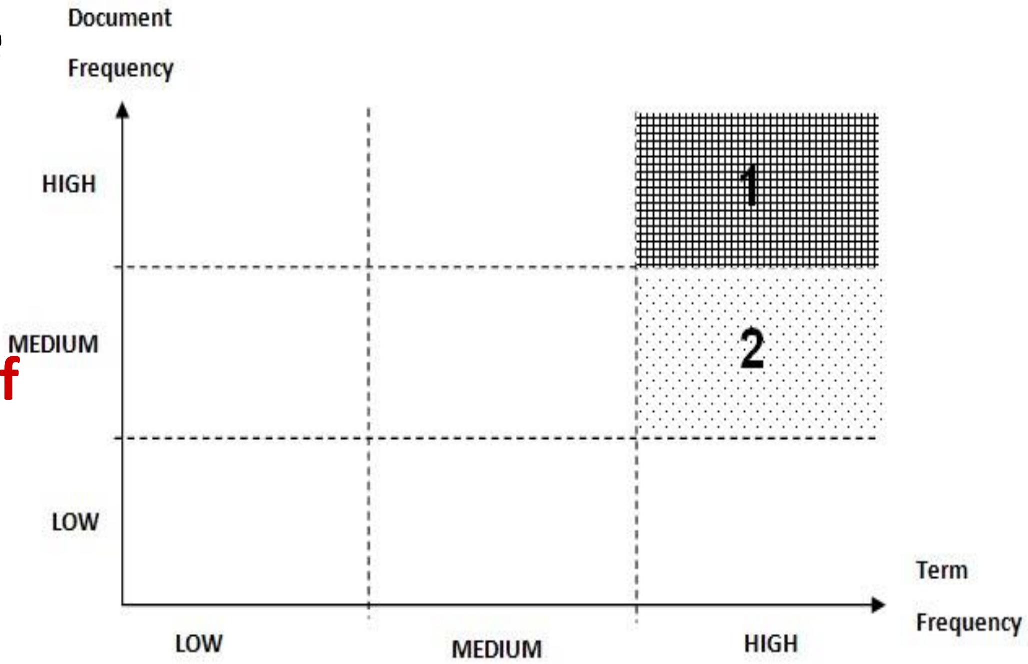
Methodology: Framework for Term Clumping

- <u>Inductive Method</u>: Emphasizes where we work to consolidate terms into topical factors, and works from the dataset without a priori criteria to target particular terms.
- <u>Purposive Method</u>: Comes to the given text compilation with pre-conceived key terms.



Methodology: Framework for Term Clumping

- **Combine Author Network**
  - Consolidates authors and their main co-authors before the association analysis, which helps us to find core authors easily.
  - We transfer the idea to deal with terms.
- **Term Frequency Inverse Document Frequency**
  - Evaluates not only the frequency of the term, but also the frequency of the records where the term appears.



Case Study: Dye Sensitized Solar Cells

- Record Selection
  - From 2001 to 2010;
  - Web of Science (4104 Records) + EI Compendex (3730 Records) Database = 5784 Records;
- Field Selection
  - Title + Abstract + Keywords

with Nature Language Processing = 90980 Noun Phrases

Target Phrases for Term Clumping

Case Study: Dye Sensitized Solar Cells

- Text Cleaning
  - Stopword Removal: 89355 Terms (apply Thesauri)
  - Further Removal: 82701 Terms (Extra Thesauri)
  - General Cleaning: 65379 Terms (Fuzzy Matching)
  - Pruning
    - Remove Single Terms (frequency < 2): 23311 (*Critical*)
    - Analysts review the term list, remove HTML codes, organization titles, etc: 20178

Case Study: Dye Sensitized Solar Cells

- Inductive Methods
  - Combine Author Network ("CAN") Analysis : 8181
    - Consolidates Terms with Similar Meaning
    - E.g. Almost 2000 "TiO2" Terms are consolidated into "TiO2," "TiO2 film," "TiO2 electrode," and "nanotube TiO2";
  - Term Frequency Inverse Document Frequency Analysis
    - Take Terms above the threshold "10.0": 2,367 high TFIDF terms ;

Case Study: Dye Sensitized Solar Cells

- Inductive Methods
  - Compare the 2,367 high TFIDF terms and 2,367 high Frequency terms in CAN list
    - The 3rd highest term in the TFIDF list is “ZnO”, which is the 16th highest term in the high-frequency CAN list;
    - Several terms that appear 14 times and belong to the high or medium frequency terms (Top 1000 Period), such as “Molecular calculations” and “Free Organic-Dyes”, are nearly in the Top 3000 Period of TFIDF value.
    - Oppositely, several terms that only appear 2 or 3 times but have high TFIDF values, such as “dye-sensitized monolithic solar cells,” “ZnO photoanode,” and ”ZnO nano array.” Of course, these terms relate closely to DSSCs.

Conclusions

- **Define “term clumping” as the steps to clean and consolidate rich sets of topical phrases in a collection of documents pertaining to a technology under study.**
- **Present a framework for term clumping, employing a number of established and some relatively novel bibliometric and text-mining techniques.**
- **Results demonstrate the term clumping process and show promise for semi-automation to get usable term clusters to perform Technology Opportunities Analysis and other Future-oriented Technology Analyses.**

Future Work

- Inductive Methods
  - TFIDF analysis with different parameters;
  - Further Comparison with TFIDF and CAN results;
  - Compare the Term Clumping Steps on different topics (e.g., more or less technical; physical vs. bio sciences)
- Purposive Methods
  - TRIZ Theory: "Problem + Action = Solutions" Pattern;
  - Technology Roadmapping: Visualized Approaches for Topical Analysis

Q&A

Thank You!