Discovery of Disciplinary Core Vocabulary from Keyword Co-occurrence Networks: Postprint

Authors: YU Fengchang, Lu Wei

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Disciplinary foundational terms serve as crucial cornerstones of disciplinary knowledge, playing a significant role in understanding the composition of disciplinary knowledge systems, clarifying the intellectual context of disciplines, and advancing disciplinary education. However, they have long relied primarily on manual compilation, and efficient automatic mining of disciplinary foundational terms within a specific discipline has not yet been realized.

[Method/Process] This paper proposes a method for discovering relatively foundational terms within a discipline by utilizing keyword co-occurrence networks. The method automatically extracts foundational terms for a discipline from disciplinary keyword datasets by leveraging the characteristics that foundational terms exhibit relatively low word frequency while possessing relatively high centrality in the network.

[Results/Conclusion] The validity of the method is verified using keyword datasets from the computer science domain (full dataset) and two subtopics—user interfaces and information search and retrieval—from ACM paper collections spanning 1969 to 2012. Furthermore, the method is capable of discovering global foundational terms within datasets through relatively simple procedures.

Full Text

Discovery of Subject Basic Vocabulary from the Perspective of Keyword Co-occurrence Network

Yu Fengchang, Lu Wei
School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] Subject basic vocabulary serves as a crucial cornerstone of disciplinary knowledge, playing a vital role in understanding the composition of a discipline's knowledge system, clarifying its knowledge context, and promoting disciplinary education. However, its identification has long relied on manual summarization, and efficient automated mining of subject basic vocabulary within a specific discipline has not yet been achieved. [**Method/Process**] This paper proposes a method to discover foundational terms within a discipline using keyword co-occurrence networks. The method leverages the characteristics that basic vocabulary tends to have relatively low word frequency while maintaining relatively high centrality in the network, automatically extracting subject basic vocabulary from disciplinary keyword datasets. [**Result/Conclusion**] The validity of this method is verified using keyword datasets from ACM publications between 1969 and 2012, covering the entire computer science field (full dataset) and two sub-themes: *user interfaces* and *information search and retrieval*. The method can discover globally fundamental vocabulary in datasets through relatively simple steps.

1. Introduction

Basic vocabulary represents the carrier of foundational and important concepts and methods in a discipline, serving as an essential cornerstone for understanding a subject's knowledge. Research on discovering subject basic vocabulary is significant for comprehending the composition of a discipline's knowledge system, clarifying its knowledge context, and advancing disciplinary education.

Traditionally, the discovery of subject basic vocabulary has relied on manual summarization, primarily focusing on disciplines with relatively simple knowledge systems such as junior and senior high school curricula. Previous studies have shown that educators who systematically organized basic concepts in subjects like high school chemistry and politics, combined with appropriate pedagogical methods, effectively improved students' comprehension [**?**]. In the field of Traditional Chinese Medicine, a dictionary-style terminology database was established through the collaborative efforts of over 300 experts from more than 10 institutions over a decade [**?**], demonstrating that manual construction of disciplinary vocabulary databases entails enormous time and labor costs. In linguistics, research comparing the *Grade Division* of mainland Chinese with the *Basic Vocabulary Database* of Taiwan has revealed subtle differences in word usage across the strait [**?**]. These studies collectively confirm the academic value of vocabulary databases, particularly basic vocabulary databases, and suggest that automatic mining of disciplinary basic vocabulary would yield even greater significance.

Coarse-grained knowledge discovery in academic texts has long been studied, with co-occurrence networks and citation networks representing important research approaches. These methods model authors, papers, and journals as net-

work nodes, treating co-occurrence or citation relationships as edges to derive conclusions through network measurement. For instance, P. Chen et al. [?] constructed a network using citation relationships in physics literature and applied the PageRank algorithm to measure document centrality, identifying foundational and important papers widely recognized in the physics community. Y. H. Eom et al. [?] built networks from 24-language Wikipedia editions, applying PageRank, 2DRank, and CheiRank algorithms to identify 100 historically significant figures. S. Mukherjee et al. [?] utilized historical ODI cricket match data from 1877 to 2010 to construct directed, weighted competition networks for teams and captains, calculating node out-degree, PageRank values, and edge weights to determine the best teams and captains in history. Some scholars have modeled authors in digital libraries, incorporating PageRank values, author information, and paper abstracts to effectively recommend influential authors [?, ?]. C. Bigonha et al. [?] studied author influence ranking on Twitter, combining authors' positions in friend and retweet networks, tweet polarity, and text quality to achieve favorable results. Y. Ding et al. [?] integrated the PageRank algorithm with topic models for topic-based influence ranking of scholars in the information retrieval field. Y. L. Chen [?] pioneered the use of PageRank for journal ranking, incorporating citation analysis and expert opinions through particle swarm optimization to achieve good results. Z. Kozareva et al. [?] applied similar methods to word-level mining with satisfactory outcomes.

These studies demonstrate that co-occurrence networks possess strong capabilities for mining related content. Inspired by this research, we investigate the possibility of using keyword co-occurrence networks to discover basic vocabulary in the computer science domain from academic paper keywords.

2. Research Approach and Method

2.1 Research Approach Literature [?] indicates that mastering basic disciplinary concepts is crucial for acquiring other knowledge in the field. The underlying rationale is that other concepts and knowledge within a discipline are closely related to basic concepts, which play a central role in the disciplinary knowledge system—often forming a network structure emanating from basic concepts. In subsequent keyword co-occurrence networks, the basic vocabulary discovered by this method comprises words with relatively high centrality, representing a necessary condition for the method's effectiveness.

Academic papers document disciplinary development, and paper keywords distill key content, typically representing important concepts and methods in the paper. Therefore, the evolution of keywords in a discipline's academic papers reflects the discipline's development to some extent. For vocabulary-level knowledge discovery, keywords offer natural advantages, making them our research object for mining disciplinary basic vocabulary.

Literature [?, ?] notes that disciplinary themes evolve over time through five forms: emergence, disappearance, inheritance, division, and merging. Conse-

quently, basic vocabulary selection should be a dynamic process—cutting-edge technologies may become foundational as they mature. This is particularly relevant in computer science, where rapid technological iteration means basic vocabulary from a single time period cannot accurately represent the discipline's development trajectory. X. Jiang et al. [**?**] point out that observation time windows significantly impact ranking results, and although graph-based scholar ranking algorithms differ substantially from citation-count-based methods, their results remain highly correlated with citation counts. Therefore, decoupling these associations is necessary to obtain better results.

Based on these findings, we focus our basic vocabulary discovery on global basic vocabulary—identifying foundational terms from the perspective of entire computer science development. Considering that word frequency affects centrality, we use the difference between centrality ranking and frequency ranking as our metric to offset the influence of high frequency on centrality calculation.

Under the premise of global basic vocabulary, these terms must be independent of observation time windows. That is, if an effective method exists for discovering disciplinary basic vocabulary in datasets, basic vocabulary identified from longer observation windows should encompass that from shorter windows. For example, for a dataset spanning t_0 to t_m, let F_1 be the basic vocabulary set discovered from window t_0 to t_m, and F_2 from window t_0 to t_n ($t_0 < t_n < t_m$). Then $F_2 \subset F_1$ must hold. This time-window independence constitutes a necessary condition for the method's effectiveness.

According to the second necessary condition in our research approach, basic vocabulary comprises keywords with high PageRank values. We must extract keywords with high PageRank values from each time window. Based on the centrality calculation characteristics of keyword co-occurrence networks introduced in Section 2.2, if a node lacks sufficient centrality, only two reasons exist: (1) all connected nodes have low centrality, meaning the keyword rarely co-occurs with important keywords; or (2) while connected nodes have high centrality, they also have high degree, making the keyword a "universal" term with low precision for describing the discipline. Neither reason aligns with our definition of disciplinary basic vocabulary, meaning the negation of this proposition is false, thus making the proposition itself another necessary condition for discovering disciplinary basic vocabulary.

2.2 PageRank Algorithm This study uses computer literature keywords as network nodes and co-occurrence relationships between keywords in the same paper as edges. We employ the PageRank algorithm to calculate node centrality in the network. The PageRank algorithm is built upon the random surfer model of the internet [**?**], expressed as:

$$G_i = \lambda \frac{1}{N} + (1-\lambda) \sum_{j \in B(i)} \frac{G_j}{K_j} \quad (1)$$

A webpage's PageRank value consists of two parts. The right term represents contributions from all pages linking to page i, where the summation covers all neighboring nodes j of node i. K_j denotes the degree of page j, meaning page j contributes equally to all linked pages, with each receiving $1/K_j$ of j's PageRank value. The left term represents the contribution when jumping to page i from any random webpage, where N is the total number of webpages and λ is the damping coefficient. In equation (1), G_i approaches $1/N$ as λ increases, meaning an excessively large λ will cause all nodes' PageRank values to converge, reducing node distinguishability. Conversely, as λ decreases, G_i becomes more influenced by neighboring nodes—if neighboring nodes have high PageRank values, node i's PageRank value increases accordingly, facilitating differentiation of node importance. This study adopts the parameter setting $\lambda = 0.15$ from Page and Brin's original paper [**?**].

PageRank algorithm has three key characteristics: (1) When node i connects to nodes j and k with identical degrees, if node j's PageRank value exceeds node k's, node j contributes more to node i's PageRank value; (2) Among nodes with identical PageRank values connected to node i, those with fewer degrees contribute more to i's PageRank value; (3) When node i connects to numerous nodes, its PageRank value becomes larger.

3. Dataset

This study utilizes the full-text dataset from the Association for Computing Machinery (ACM). We selected 215,710 papers published between 1951-2012. After data cleaning, we retained 110,363 papers containing keywords published between 1969-2012. The dataset includes 364 sub-topic classifications. The two sub-themes with the most papers are *user interfaces* and *information search and retrieval*. We conduct experiments on the entire computer science discipline (full dataset), the *user interfaces* sub-theme, and the *information search and retrieval* sub-theme to verify our method's correctness.

[Figure 1: see original paper] shows the temporal distribution of papers in the dataset. Since the dataset was collected in mid-2012, paper counts (both total and keyword-containing) show an upward trend year by year except for 2012. [Figure 2: see original paper] displays the proportion of papers containing keywords each year (logarithmic scale on y-axis). Statistics begin in 1969 because earlier papers in the dataset contain no keywords. From 1990 onward, the proportion of keyword-containing papers increased annually, approaching 90% of total papers by 2011. Given that years with lower keyword inclusion rates also had relatively fewer total papers, our dataset adequately represents computer science research.

Keywords, as authors' summaries and refinements of their research achievements, reflect core paper content to some extent. This is why we select keywords as indicators of disciplinary basic knowledge. To verify keyword authenticity, we randomly sampled 30 papers by year and manually compared dataset keywords

with author-provided keywords in original texts—all matched perfectly.

4. Experiments

4.1 Experimental Design To verify whether our method effectively discovers basic vocabulary in computer science, experiments are divided into three groups targeting: (1) the entire computer science field (full dataset), (2) the *user interfaces* sub-theme, and (3) the *information search and retrieval* sub-theme. Below, we introduce the computer science field experiment as representative.

The experimental process follows these steps: (1) Segment the dataset by time windows; (2) Construct co-occurrence networks for keywords in each window; (3) Calculate PageRank values and TF values for each network, identifying keywords with large ranking differences; (4) Take the intersection of results from all time windows.

4.2 Observation Time Window Setting According to the first necessary condition in our research approach, the dataset must be divided into T time windows. This paper partitions papers into 5 overlapping time windows ($T = 5$) with approximately equal increments in paper counts. The dataset contains 110,363 keyword-containing papers. The five observation windows correspond to paper counts of 22,073; 44,146; 66,219; 88,292; and 110,363, with observation periods of 1969-2004, 1969-2007, 1969-2008, 1969-2010, and 1969-2012 respectively. We designed these windows to ensure each observation period starts from the dataset's temporal beginning, enabling study of disciplinary development from its origin.

4.3 Network Construction Based on the five observation windows, we used Python's NetworkX toolkit to construct five bidirectional, weighted co-word networks. Nodes represent keywords, edges represent two keywords appearing together in a paper, and each edge weight is set to 1. Specifically, for the 5th time window, the node set is V and the resulting graph is G.

4.4 Network Node Calculation For the five co-word networks, we calculated each node's centrality (PageRank value) using the PageRank algorithm and parameters described in Section 2.2. Simultaneously, we computed each keyword's frequency (TF) within its time window. We obtained two rankings for each window: PageRank ranking (GRank) and TF ranking (tfRank). [Figure 3: see original paper] shows the relationship between keyword frequency (TF) and average centrality (PageRank) in the 5th time window, where each point represents a keyword and the dashed line is an origin-passing fitted line. [Figure 4: see original paper] demonstrates that most keywords' frequencies and PageRank values follow a proportional relationship, consistent with literature [**?**] showing that higher frequency generally accompanies higher PageRank value—high tfRank keywords typically correspond to high GRank.

To identify keywords with large ranking differences in step (3) of Section 4.1, we select keywords in the top Top_g% of GRank and compute the difference between their GRank and tfRank. Negative results indicate keywords with low TF but high PageRank values. We sort these differences in ascending order and extract the top Top_t% as candidate basic vocabulary for each observation window.

Note that PageRank algorithm results are dimensionless values, while TF values count keyword occurrences with "times" as unit. Their different dimensions prevent direct subtraction, so we use rank differences to characterize disparity magnitude. Parameter Top_g determines the importance threshold of selected keywords within observation windows, while Top_t controls the ranking difference magnitude and final basic vocabulary quantity. Given that basic vocabulary in any discipline is limited, after multiple experiments we set $T = 5$, $Top_g = 3$, $Top_t = 33$ for computer science, and $T = 4$, $Top_g = 10$, $Top_t = 25$ for the *user interfaces* and *information search and retrieval* sub-themes.

4.5 Discovery of Subject Basic Vocabulary According to the second necessary condition, basic vocabulary comprises high-PageRank keywords. We extract keywords with high PageRank values from each time window. Based on the first necessary condition, global basic vocabulary must be time-window independent, so we take the intersection of results from all five windows as our final discovered basic vocabulary for computer science.

4.6 Result Validation No existing computer science basic vocabulary list is available for comparison. To verify correctness, we manually evaluated all three experimental results, submitting them to one associate professor or postdoctoral researcher from each respective domain. Before evaluation, we explained our definition of basic vocabulary as representing foundational and important concepts and methods in a discipline. Evaluators checked terms meeting this definition.

presents manual evaluation results. Accuracy is calculated as the ratio of checked basic vocabulary to total vocabulary provided. The three experiments yielded 232 basic vocabulary terms for computer science, 110 for *information search and retrieval*, and 153 for *user interfaces*, with accuracies of 91.81%, 84.55%, and 86.27% respectively. The higher accuracy in computer science likely reflects its broader scope and larger basic vocabulary pool relative to the limited number of candidate terms provided for evaluation.

compares frontier keywords (top 5 rows) with discovered basic vocabulary (bottom 5 rows). Frontier keywords exhibit both high PageRank values and high word frequency with small ranking differences, while basic vocabulary shows relatively low frequency but high PageRank values. Our method essentially identifies keywords with low frequency but high centrality.

[Figure 4: see original paper] illustrates the unique network topology of basic vocabulary. Using keywords from as roots, we construct trees where all con-

nected keywords are leaves. Each tree is a subgraph of G, with root node size set to 1 and leaf node sizes normalized by their PageRank values relative to the root. From a graph-theoretic perspective, frontier keywords have high GRank and high degree, but all leaf nodes have relatively low PageRank values. In contrast, disciplinary basic vocabulary has relatively low degree but all leaf nodes have relatively high PageRank values.

5. Analysis and Discussion

Our results include typical computer science basic vocabulary such as *data structure*, *network topology*, *microprocessors*, *time complexity*, *parallel algorithm*, *website*, and *program debugging*. For instance, data structure is fundamental for storing and organizing data in computers and represents a basic component of computer programming. Microprocessors (or central processing units) constitute the core hardware component executing circuit control and logical operations. Notably, frontier terms from the past decade—such as data mining, virtual reality, machine learning, and cloud computing—do not appear in our results.

We have designed a method for discovering disciplinary basic vocabulary and validated its effectiveness using ACM datasets and two sub-themes. The method is simple and efficient for identifying domain basic vocabulary. However, limitations exist: the method cannot rank basic vocabulary by fundamentality. Future research should improve measurement methods to further quantify the "basicness" of disciplinary vocabulary.

References

[?] Wen Haigang. Exploring the Teaching of Basic Chemical Concepts at the High School Level [?]. *Examination Weekly*, 2015(56): 131.

[?] Ma Xiaomin. Constructing a Knowledge Network of Junior High School Chemistry Basic Concepts—Application of Mind Maps in Teaching [?]. *Contemporary Teaching and Research*, 2015(2): 43-52.

[?] Yu Jing. Improving Performance Through Basic Concepts—On Concept Teaching in High School Politics [?]. *Contemporary Teaching and Research*, 2015(2): 116, 118.

[?] Jia Lirong, Li Haiyan, Yu Tong, et al. Analysis of the Basic Terminology Database of Traditional Chinese Medicine Language System [?]. *China Digital Medicine*, 2014, 9(2): 66-67.

[?] Qu Yinghua. Comparison and Reflection on the New Generation of Basic Chinese Vocabulary Lists Across the Taiwan Strait [?]. *Chinese Language Teaching and Research*, 2015(6): 317.

[?] Chen P, Xie H, Maslov S, et al. Finding scientific gems with Google's PageRank algorithm [?]. *Journal of Informetrics*, 2007, 1(1): 8-15.

[?] Eom YH, Aragón P, Laniado D, et al. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions [?]. *PLoS ONE*, 2015, 10(3): e0114825.

[?] Mukherjee S. Identifying the greatest team and captain—a complex network approach to cricket matches [?]. *Physica A: Statistical Mechanics and its Applications*, 2012, 391(23): 6066-6076.

[?] Mimno D, McCallum A. Mining a digital library for influential authors [?]//Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2007: 105-106.

[?] Lin L, Xu Z, Ding Y, et al. Finding topic-level experts in scholarly networks [?]. *Scientometrics*, 2013, 97(3): 797-819.

[?] Bigonha C, Cardoso TNC, Moro M, et al. Sentiment-based influence detection on Twitter [?]. *Journal of the Brazilian Computer Society*, 2012, 18(3): 169-183.

[?] Ding Y. Topic-based PageRank on author co-citation networks [?]. *Journal of the Association for Information Science and Technology*, 2011, 62(3): 449-466.

[?] Chen YL, Chen XH. An evolutionary PageRank approach for journal ranking with expert judgments [?]. *Journal of Information Science*, 2011, 37(3): 254-272.

[?] Kozareva Z, Hovy E. Insights from network structure for text mining [?]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2011: 1616-1625.

[?] Qu Jiabin, Ou Shiyan. Disciplinary Theme Evolution Analysis Based on Topic Filtering and Topic Association [?]. *Data Analysis and Knowledge Discovery*, 2018, 2(1): 64-75.

[?] Tang Guoyuan. Construction of Research Methods for Disciplinary Theme Evolution Based on Co-word Analysis [?]. *Library and Information Service*, 2017, 61(23): 100-107.

[?] Jiang X, Sun X, Zhuge H. Graph-based algorithms for ranking researchers: not all swans are white! [?]. *Scientometrics*, 2013, 96(3): 743-759.

[?] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web [?]. Stanford InfoLab, 1999.

Author Contributions:
Yu Fengchang: Data processing and analysis, paper writing;
Lu Wei: Theoretical guidance and research design.

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