Post-print Identification of Characteristic Words for Breakthrough Research Based on Abstract and Citation Text Corpora

Authors: Yang Xuemei, Wang Xue, Du Jian, Tang Xiaoli

Date: 2023-04-01T16:15:55+00:00

Abstract

[Purpose/Significance] Based on the perspectives of authors' descriptive evaluations of their own research and subsequent researchers' commentative citations, this study extracts characteristic terms of breakthrough research using abstract and citation corpora, thereby understanding the characteristics of these corpora to aid in identifying breakthrough research.

[Methods/Process] Key literature selected as "Breakthrough of the Year" by Science and "key publications" of Nobel Prize laureates were chosen as breakthrough research corpus data, integrating paper abstracts and citation corpora for characteristic term extraction. In the extraction process, the Stanford CoreNLP tool was first utilized to perform tokenization and term frequency statistics on the corpora, and characteristic term primitives were extracted in conjunction with expert opinions. These characteristic term terms were then used as seed words, and semantic expansion was conducted using semantic relationships from medical texts. Finally, recall and precision rates were employed to comparatively evaluate the retrieval and identification effectiveness of characteristic terms for abstracts and citations before and after expansion.

[Results/Conclusion] The breakthrough research corpus yielded 8 characteristic term primitives from abstract corpora and 8 from citation corpora. In characteristic term retrieval and identification, expanded characteristic terms from both abstracts and citations achieved the highest recall rate, citation characteristic terms achieved the highest precision rate, and expanded citation characteristic terms demonstrated a balanced effectiveness in both recall and precision rates.

Full Text

Identifying Feature Words of Breakthrough Research Based on Abstract and Citation Text Corpus

Yang Xuemei[1], **Wang Xue**[1], **Du Jian**[2], **Tang Xiaoli**[1] [1]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005 [2]National Institute of Health Data Science, Peking University, Beijing 100191

Abstract

[**Purpose/Significance**] From the perspectives of authors' descriptive evaluation of their own research and subsequent researchers' critical citations, this study extracts feature words of breakthrough research using abstract and citation corpora to understand the characteristics of breakthrough research texts and facilitate the identification of such research. [**Method/Process**] Key documents selected by *Science* as "Breakthrough of the Year" and "key publications" of Nobel Prize winners were used as breakthrough research corpora, integrating paper abstracts and citation texts for feature word extraction. First, the Stanford CoreNLP tool was employed for tokenization and word frequency statistics, and feature tokens were extracted in combination with expert opinions. These feature words were then used as seed words for semantic expansion based on semantic relationships in medical texts. Finally, retrieval and identification effectiveness before and after feature word expansion were compared through recall and precision rates. [**Result/Conclusion**] Eight feature tokens were selected from abstract corpora and eight from citation corpora. In retrieval and identification, expanded feature words from both abstracts and citations achieved the highest recall rate, while citation feature words achieved the highest precision rate, with expanded citation feature words demonstrating the best overall performance.

Keywords: breakthrough research; feature words; abstract text; citing sentence
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.11.014

"Innovation-driven development" has become a strategic initiative for China to accelerate the transformation of its economic development model. Breakthrough technological innovation that can bring about dual changes in industrial technology architecture and components and disrupt markets is a crucial move in this strategy. Early discovery of breakthrough research can directly drive breakthrough technological innovation. If breakthrough research can be identified early in its development, it can promote the deployment of breakthrough innovation research and accelerate the process of breakthrough technological innovation. Early identification of breakthrough research is of great significance for building China into a strong power in science and technology innovation.

This paper explores the semantic features of abstracts and citation corpora of breakthrough research from the perspective of feature word discovery, using extracted feature words to help identify breakthrough research papers in large-scale literature datasets and provide theoretical support for breakthrough research deployment strategies. In the field of scientometrics, research methods can be divided into two categories: one is identifying breakthrough research based on specific indicators from bibliometrics. E. Garfield initially identified some scientific discoveries using simple citation counts by observing the impact of scientific papers over time [1]. I.V. Ponomarev et al. [2] combined quantitative methods based on publication citation dynamics to early detect and identify candidate breakthrough papers. Y.H. Huang et al. [3] discovered that the emergence of breakthrough research would cause ruptures in citation chains based on article citation paths, thus proposing a "rupture score" to identify breakthrough research in biomedicine and other fields. In addition to citation-related indicators, some studies also considered author collaboration, delayed recognition index, and other metrics. A joint study by Clarivate Analytics and the U.S. National Cancer Institute [4] incorporated publication time, co-author networks, and other field characteristics into a random forest model, combined with expert selection, to identify candidate breakthrough papers as early as possible after publication. Du Jian et al. [5] proposed using the delayed recognition index and patent citation indicators to identify transformative research.

The other category of methods identifies breakthrough research using linguistic features of evaluative statements. Many scholars have adopted citation content analysis or citing sentence analysis methods because citation corpora contain valuable information. D.R. Radev et al. used citation semantic mining to summarize the "contribution points" of cited papers from citing sentences [6]. H. Small selected citing sentences containing clue words "discover*" and their corresponding references, using semantic features of citing sentences for machine learning experiments to automatically determine whether references were "scientific discoveries" [7]. These studies have identified major discoveries from the perspective of citation data. However, since authors have complex motivations when citing references, research based on citation counts and paths has the limitation that simple citation metrics and relationships cannot provide sufficient information for accurate identification. Although identification based on citation statement features highlights subsequent researchers' evaluations of relevant research, each citation cannot comprehensively summarize the author's research results. Therefore, identification based on statement features should also incorporate authors' own evaluations of their research. Additionally, H. Small's use of clue words to limit candidate breakthrough research literature provides a new approach, but single clue words may exclude some potential breakthrough research literature.

1. Related Research

1.1 Breakthrough Research Identification Currently, research on identifying breakthrough research mainly focuses on two approaches. The first approach uses bibliometric indicators. E. Garfield initially identified some scientific discoveries using simple citation counts [1]. I.V. Ponomarev et al. [2] combined quantitative methods based on citation dynamics for early detection. Y.H. Huang et al. [3] proposed using "rupture scores" based on citation chain disruptions. Some studies also integrated author collaboration networks and delayed recognition indices [4]. Du Jian et al. [5] used delayed recognition and patent citation indicators.

The second approach utilizes evaluative language features in citation texts. Many scholars analyze citation content to identify breakthrough research [6]. H. Small used "discover*" clue words in citing sentences for machine learning classification [7]. While these studies provide valuable insights, they face limitations: citation metrics alone are insufficient, single citations don't fully represent research contributions, and reliance on single clue words may exclude potential breakthroughs.

1.2 Text Feature Extraction Text feature extraction is a fundamental step in natural language processing, widely applied in text classification, indexing, and retrieval. The main idea is to construct an evaluation function to calculate feature weights, rank them, and select the top n features as the final subset [8].

Current methods fall into two categories: statistical and semantic approaches [9]. Statistical methods like TF-IDF [10] treat features as independent. Gu Jun et al. [11] used ICTCLAS for tokenization and improved TF-IDF for hotspot term extraction in patents. While simple and intuitive, statistical methods ignore semantic relationships, leading to incomplete feature extraction.

Semantic methods establish relationships between features through context analysis. The Word2vec model [12] extracts semantic representations based on context. Chen C. [13] used Word2vec to identify uncertainty cue words. While effective for defined categories, semantic methods struggle with undefined categories and inter-class evaluation.

2. Feature Word Extraction Method

Building on previous research, this study integrates authors' self-evaluation and peer evaluation to extract feature words from known breakthrough research abstracts and citations using statistical and semantic methods. The framework consists of four steps: data source and preprocessing, feature token extraction based on word frequency statistics, semantic expansion, and effectiveness evaluation [Figure 1: see original paper].

2.1 Data Source and Preprocessing Defining breakthrough research is essential for corpus selection. While no consensus definition exists, I.V. Pono-

marev et al. [14] found breakthrough papers receive many citations and provide new research directions. *Science*'s news editor R. Coontz stated that breakthroughs either solve long-standing problems or open doors for new research [15]. This study defines breakthrough research as major discoveries in incremental studies or disruptive transformations that provide new research directions.

We selected *Science*'s "Breakthrough of the Year" references and Nobel Prize winners' "Key Publications" in biomedicine as corpora [16]. Abstracts were retrieved from PubMed using PMIDs. Citation texts were obtained from the Colil platform [17], a Japanese life science database center developed from PMC-OAS that batch-retrieves citation texts using PMIDs [18].

Retrieved corpora required cleaning to remove invalid characters, references, URLs, and other non-standardized text that would affect statistical analysis and coding.

2.2 Feature Token Extraction Based on Word Frequency Statistics Feature selection focused on words common across multiple documents, making TF-IDF unsuitable. We used traditional word frequency statistics with Stanford CoreNLP [19] for tokenization and frequency counting. The process involves: tokenization → lemmatization → POS tagging → frequency statistics, filtering punctuation and cardinal numbers (CD) to reduce noise. [Figure 2: see original paper] illustrates this process with the example sentence: "The sulfur atom is supplied by a separate cluster in the enzyme."

2.3 Semantic Expansion of Feature Tokens After obtaining feature tokens via NLP, we expanded them using semantic relationships from all medical literature in PMC. Word2vec offers CBOW and Skip-Gram models, with Skip-Gram being more suitable for large corpora [20]. We trained Skip-Gram on PMC Open Access (PMCOA) texts using 1-5 gram sliding windows [21], building token vectors to create an N-gram library containing all significant information for distributional similarity modeling. Random indexing [22] summed index vectors of words in context windows to obtain vector spaces. After neural network training, cosine distances between words determined semantic similarity, with closer distances indicating stronger semantic relationships. [Figure 3: see original paper] shows the PMCOA Word2vec model construction flowchart.

2.4 Evaluation Method Recall and precision [23] are key information retrieval metrics for evaluating feature extraction effectiveness. Precision measures the noise ratio (relevant retrieved documents vs. total retrieved). Recall measures success in retrieving relevant documents from the collection (relevant retrieved vs. total relevant). In our evaluation: Precision = TP/(TP+FP), Recall = TP/(TP+FN), with definitions shown in .

Positive examples were papers in Faculty of 1000 (F1000) database rated as "New-Finding" (showing novel data/models as breakthroughs). Negative exam-

ples were papers rated as “Negative/Null Result” (showing no valuable results). This created a clear distinction for evaluation.

3. Feature Word Extraction for Breakthrough Research

3.1 Data Acquisition and Preprocessing We included Nobel Prize winners’ Key Publications (1981-2018) and *Science* Breakthrough of the Year papers (1996-2018) in biomedicine. Retrieval from *Science* and Nobel websites yielded 556 breakthrough papers and 103 Nobel Key Publications (648 unique papers after deduplication). Abstract corpora contained 467 records from PubMed. Citation corpora contained 135,526 records from Colil (131,767 after cleaning).

3.2 Feature Token Screening and Extraction Frequency statistics extracted 7,058 words (54,394 total frequency) from abstracts and 70,995 words (3,184,578 total frequency) from citations. Both corpora were dominated by nouns (NN), adjectives (JJ), and verbs (VB), though proportions differed. Citation corpora showed more stable POS distributions due to larger size [Figure 4: see original paper].

From the top 500 words, we filtered medical terms from MeSH and breakthrough-irrelevant words. Three information science experts independently reviewed candidate tokens in context. Tokens identified as breakthrough indicators by $ $2 experts were confirmed as feature tokens. Final selections: 8 abstract tokens (new, novel, potential, key, change, evidence, basis, base) and 8 citation tokens (change, first, potential, new, novel, since, discovery, discover) .

3.3 Semantic Expansion Analysis and Visualization Feature tokens were input into Word2vec (abstracts: -n50 for “new, novel, potential, key, change, evidence, basis, base”; citations: -n50 for “change, first, potential, new, novel, since, discovery, discover”). shows top 10 semantically related words for “key” and “discovery,” serving as expansion candidates.

Expert review and contextual analysis yielded 30 expanded abstract features and 36 expanded citation features. Co-occurrence analysis revealed frequent joint appearances. In abstracts [Figure 5: see original paper], 27 nodes showed high-frequency words like “based,” “changes,” and “findings,” with “new” appearing most frequently and co-occurring often with “evidence.” In citations [Figure 6: see original paper], 35 nodes formed a tighter network, with “first” and “since” showing strong co-occurrence and frequent “since…first…” patterns.

3.4 Effectiveness Analysis of Feature Word Extraction We evaluated effectiveness using F1000 database papers: 183 abstracts and 1,895 citations rated $ $5 times as “New-Finding” (positive), and 125 abstracts and 1,840 “Negative/Null Result” papers (negative). Reverse retrieval using abstract features, abstract expanded features, citation features, citation expanded features, and combined features calculated recall and precision rates [Figure 7: see original paper].

Results showed abstract and citation expanded features achieved the highest recall (94.54%). Citation features achieved the highest precision (70.77%). Citation expanded features showed the best comprehensive performance. Researchers can select different feature types based on recall vs. precision needs .

4. Conclusion and Outlook

This study used *Science* Breakthrough of the Year papers and Nobel Prize winners' Key Publications in biomedicine to extract feature words by integrating abstracts and citations through statistical and semantic methods. Both corpora showed similar POS distributions overall. Stanford CoreNLP extracted 8 abstract tokens and 8 citation tokens, expanded to 30 and 36 features respectively using Word2vec. Co-occurrence analysis confirmed frequent joint appearances, especially the "since…first…" pattern in citations.

Evaluation on F1000 data showed recall rates above 90%, but precision remains insufficient for standalone identification. These feature words serve as a first step to narrow candidate literature ranges. Future work should combine machine learning with full semantic information from citations and abstracts to accurately identify breakthrough research from candidate pools.

References

[1] GARFIELD E. The 1976 articles most cited in 1976 and 1977. 1. Essays of an information scientist, 1979, 13(4): 81-99.

[2] PONOMAREV I V, WILLIAMS D E, HACKETT C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs[J]. Technological forecasting and social change, 2014, 81(1): 49-55.

[3] HUANG Y H, HSU C N, LERMAN K. Identifying transformative scientific research[C]//2013 IEEE international conference on data mining (ICDM). Melbourne: IEEE, 2013: 291-300.

[4] WOLCOTT H N, FOUCH M J, HSUE R, et al. Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers[J]. Scientometrics, 2016, 107(2): 807-817.

[5] Du Jian, Sun Yinan, Zhang Yang, et al. Scientometric characteristics of transformative research and early identification methods[J]. Chinese Science Fund, 2019, 33(1): 90-100.

[6] RADEV D R, AMJAD D. Rediscovering ACL discoveries through the Lens of ACL anthology network citing sentences[C]//Proceedings of ACL 2012 special session on the 50th anniversary of ACL. Stroudsburg: Association for Computational Linguistics, 2012: 1-12.

[7] SMALL H, TSENG H, PATEK M. Discovering discoveries: identifying

biomedical discoveries using citation texts[J]. Journal of informetrics, 2017, 11(1): 46-62.

[8] SIOLAS G. Support vector machines based on a semantic kernel for text categorization[C]//Proceedings of the international joint conference on neural networks. Como: IEEE Computer Society, 2000: 205-209.

[9] Liu Lizhen, Song Hantao. Feature selection in text classification[J]. Computer engineering, 2004, 30(4): 14-15.

[10] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1987, 24(5): 513-523.

[11] Gu Jun, Yan Ming. Research on new technology term identification based on Chinese patents[J]. Information science, 2013, 31(2): 144-149.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2020-02-23]. https://arxiv.xilesou.top/pdf/1301.3781.pdf.

[13] CHEN C, SONG M, HEO G E. A scalable and adaptive method for finding semantically equivalent cue words of uncertainty[J]. Journal of informetrics, 2018, 12(1): 158-180.

[14] PONOMAREV I V, WILLIAMS D E, HACKETT C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs[J]. Technological forecasting and social change, 2014, 81: 49-55.

[15] Science newsletters[EB/OL]. [2020-02-23]. http://www.sciencemagchina.cn/highlights141219.aspx.

[16] Breakthrough of the year[EB/OL]. [2020-02-23]. http://en.wikipedia.org/wiki/Breakthrough_{{{of}}}{{{th

[17] Colil[EB/OL]. [2020-02-23]. http://colil.dbcls.jp/browse/papers/.

[18] FUJIWARA T, YAMAMOTO Y. Colil: a database and search service for citation texts in the life sciences domain[J]. Journal of biomedical semantics, 2015, 6(1): 38.

[19] DING Y, ROUSSEAU R, WOLFRAM D. Text mining with the Stanford CoreNLP[J]. Replicable science of science studies, 2014(10): 215-234.

[20] Liu Xin, She Xiandong, Tang Yongwang, et al. Short text clustering algorithm based on feature word vectors[J]. Journal of data acquisition and processing, 2017, 32(5): 1052-1060.

[21] PYYSALO S, GINTER F, MOEN H, et al. Distributional semantics resources for biomedical text processing[J]. Proceedings of languages in biology and medicine, 2013.

[22] KANERVA P, KRISTOFERSON J, HOLST A. Random indexing of texts for latent semantic analysis[J]. Proceedings of the annual meeting of the Cognitive Science Society, 2000, 22(22): 1036-1036.

[23] CLEVERDON C. The cranfield tests on index language devices[J]. Aslib proceedings, 1967, 19(6): 173-194.

Author Contributions: Yang Xuemei: Method optimization and implementation, manuscript writing Wang Xue: Data acquisition and preprocessing Du Jian: Research design Tang Xiaoli: Research design optimization and manuscript review

Identifying Feature Words of Breakthrough Research Based on Abstract and Citation Text Corpus

Yang Xuemei[1], Wang Xue[1], Du Jian[2], Tang Xiaoli[1] [1]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005 [2]National Institute of Health Data Science, Peking University, Beijing 100191

Abstract: [Purpose/significance] Based on authors' descriptive evaluation of their research and critical citations from subsequent researchers, this study extracts feature words from abstract and citation corpora to understand breakthrough research characteristics and aid identification. [Method/process] Key documents selected as *Science* "Breakthrough of the Year" and Nobel Prize winners' "key publications" served as breakthrough research corpora, integrating abstracts and citations for feature extraction. Stanford CoreNLP performed tokenization and frequency statistics, with feature tokens extracted using expert opinions. These served as seed words for semantic expansion using medical text semantic relationships. Retrieval effectiveness before and after expansion was compared via recall and precision rates. [Result/conclusion] Eight feature tokens were selected from abstract corpora and eight from citation corpora. Expanded features achieved highest recall for both, while citation features achieved highest precision (70.77%), with expanded citation features showing best comprehensive performance.

Keywords: breakthrough research; feature words; abstract text; citing sentence

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