An Automatic Method for Extracting Innovative Ideas Based on the Scopus® Database

Lielei Chen*, Hui Fang**

Nanjing University, School of Electronic Science and Engineering, Nanjing 210023, China,
*<542197681@qq.com>, **<fanghui@nju.edu.cn> (corresponding author)

Lielei Chen received a bachelor's degree in electronic and information engineering from Hohai University in Nanjing, China, in 2016. She is now a graduate student at the School of Electronic Science and Engineering, Nanjing University. Her research interests include natural language processing and information science.

Hui Fang received a bachelor's degree in radio engineering (in 1990) and a master's degree in signal processing (in 1993) from Southeast University in Nanjing, China, and the PhD in electroanalytical chemistry from Nanjing University in Nanjing, China, in 1998. He is now an associate professor at the School of Electronic Science and Engineering, Nanjing University and is affiliated with the State Key Laboratory of Analytical Chemistry for Life Science. His research interests include information processing, data mining, artificial intelligence, instruments and instrumentation, and bibliometrics.

Chen, Lielei and Hui Fang. 2019. "An Automatic Method for Extracting Innovative Ideas Based on the Scopus® Database." *Knowledge Organization* 46(3): 171-186. 74 references. DOI:10.5771/0943-7444-2019-3-171.

Abstract: The novelty of knowledge claims in a research paper can be considered an evaluation criterion for papers to supplement citations. To provide a foundation for research evaluation from the perspective of innovativeness, we propose an automatic approach for extracting innovative ideas from the abstracts of technology and engineering papers. The approach extracts N-grams as candidates based on part-of-speech tagging and determines whether they are novel by checking the Scopus® database to determine whether they had ever been presented previously. Moreover, we discussed the distributions of innovative ideas in different abstract structures. To improve the performance by excluding noisy N-grams, a list of stop-words and a list of research description characteristics were developed. We selected abstracts of articles published from 2011 to 2017 with the topic of semantic analysis as the experimental texts. Excluding noisy N-grams, considering the distribution of innovative ideas in abstracts, and suitably combining N-grams can effectively improve the performance of automatic innovative idea extraction. Unlike co-word and co-citation analysis, innovative-idea extraction aims to identify the differences in a paper from all previously published papers.

Received: 26 October 2018; Revised: 20 February 2019; Accepted: 1 March 2019

Keywords: innovative ideas, research, corpus, Scopus®

1.0 Introduction

Research evaluation is important for employing researchers, making grant decisions, and determining researcher promotions (Kosten 2016). Currently, one widely recognized method is citation analysis. Articles with high numbers of citations reflect their contribution to a certain extent. However, the use of citation-based indicators to evaluate research is the subject of much debate (Wu 2015). Those with low citations may also be valuable despite currently having little impact (Garfield 1972). Garfield (1979) questioned the rationality of assessing the quality of publications based on only the number of citations. A research evaluation based solely on citations is not objective (Fiala et al. 2017).

Innovation is considered to be the soul of science; it promotes scientific research (Xu 2001), and the pursuit of innovation is closely related to social and scientific development. Innovation embodies the creation, evolution, exchange, and application of new ideas for the advancement of society (Rogers 1993). Therefore, innovativeness can reflect the contribution of individual scientific publications. To assess the innovativeness of a research paper, we should first extract its innovative ideas. In the academic stratification system, peer review plays a central role in the evaluation of academic work (Cole et al. 1974), and novelty is a major and frequently used criterion (Guetzkow et al. 2004). However, automatically evaluating the innovativeness of research remains difficult.

Innovation is considered one evaluation criterion of scientific papers. Methods for identifying the original and innovative works of research efficiently and accurately have been researched in recent years (Wieringa et al. 2006).

Existing innovation-idea identification methods for individual papers are based on context features extracted or learned from manually judged documents. However, these methods can extract the ideas of a paper whose authors believe the ideas are innovative, but in reality, have been proposed previously. To avoid this situation, we present an innovative-idea extraction method that checks a widely used document database to determine whether the ideas extracted from a paper are innovative.

In this study, we performed an analysis based on a series of aspects concerning innovative ideas, and as a result, we propose an automatic method for innovative-idea extraction that checks the innovativeness of the ideas extracted. We combined N-gram extraction and web search techniques to extract innovative ideas from papers without requiring any domain corpus assembled by experts. We selected abstracts of articles published from 2011 to 2017 with the topic of "semantic analysis" as our experimental texts and considered ideas that had not been proposed previous to an article's publication year as innovative. Through experiments, we investigated factors that could improve the performance of this method and analysed the reasons causing defects in the method, thereby suggesting how the method can be further improved in future work. The proposed method provides a foundation technique for future evaluation of innovation in research papers. This work is an application of knowledge organization research.

2.0 Related work

2.1 Types of innovative ideas

To support paper evaluation criteria in the requirements engineering field, Wieringa and others (2006) classified research papers into six classes: "evaluation research," "proposal of solution," "validation research," "philosophical research," "opinion papers," and "personal experience papers." Among these classes, "proposal of solution" and "philosophical research" papers generally contain novel and original technologies or concepts. Frame (2008) divided technology innovations into the following three types: "derivative" (an extension of existing technology), "platform" (a new application of existing technology), and "breakthrough" (an entirely new technology). Mullins, Snizek, and Oehler (1988) proposed an analysis of innovation evaluation based on the structure of scientific papers (i.e., introduction, methods, and results). Based on that research, Dirk (1996) described a research work as a combination of established (E) or new (N) elements of "theory-methods-results" and suggested eight types to evaluate the novelty of original work, ranging from E-E-E (established theory—established methods—established results) to N-N-N (new theory—new methods—new results). This typological assessment was recommended for performing peer reviews of innovation (Dirk 1999).

2.2 Structure of abstracts

Mullins and others (1988) proposed an analysis of innovation evaluation based on the structure of scientific papers, including the introduction, methods, and results.

Abstract writing guidelines have been studied to improve the quality and consistency of abstracts. Milas-Bracović and Zajec (1989) suggested using the IMRAD format for an abstract, that is, introduction (I), methods (M), results (R), and discussion (D). Endres-Niggemeyer (1998) defined five moves—background (B), purpose (P), methodology (M), result (R), and discussion (D)—that constitute the abstract of research articles. By investigating abstracts from a variety of journals, researchers revealed several most-frequent abstract elements. Hartley and Betts (2009) showed that most paper abstracts in the social sciences included the goals, methods, results, and conclusions. Jamar, Šauperl and Bawden (2014) demonstrated that the most common combination of structural elements in the abstracts of technical sciences papers is moves B-M-R. Cross and Oppenheim (2006) found that moves M and R were present in all experimental abstracts.

By using this five-move framework, Kanoksilapatham (2013) provided a linguistic characterization of information presented in abstracts. The study indicated that move B functions by preparing the topic focus for readers and by highlighting the importance of topics using words or phrases such as "challenging," "increasingly important," and "improve" or by introducing the current development of the topic with present tense verbs such as "are," "is," "can," and "exhibit." Move P usually follows move B and is explicitly stated. The phrase "this study" is commonly found in this move, and the present tense, in active or passive voice, is preferred. Move M is typically expressed using research activity verbs such as "were conducted," "was tested," "estimated," and "included" when the subject of the research is an experiment or algorithm. To express move R, the verbs "show" and "find" in either present or past tense are usually used. In move D, phrases such as "were attributed," "is estimated," and "should consider" are used to discuss the implications, significance, interpretations, and explanations of the results and findings.

2.3 Terminology extraction and the Stanford Parser

Terminology extraction methods mainly include linguistic (Chen and Cshen 1994; Justeson and Katz 1995), statistical (Church and Hanks 1990; Smadja 1993), and hybrid approaches (Bounhas and Slimani 2009; Maynard and Ananjadou 1999; Oliver and Vàzquez 2015). Linguistic ap-

proaches utilize the linguistic features of sentences such as parts of speech and structure to identify terms. Statistical approaches consider statistical indicators such as term frequency, mutual information variants, co-occurrence, TFIDF (term frequency–inverse document frequency), and other methods to measure the association of terms. Recent terminology extraction methods combine terminology extraction methods with statistical approaches into a hybrid method to achieve better performance. Hybrid measures first utilize linguistic analysis to extract all candidates and then apply statistical analysis for further selection.

Software including Word Segmenter, EnglishTokenizer, and Parser, developed by the Stanford Natural Language Processing Group, has been widely used in natural language processing (NLP) research (https://nlp.stanford.edu/software/). Here, we used the Stanford Part-Of-Speech Tagger (POS Tagger) (Toutanova et al. 2003) as a part-of-speech annotation tool to analyse sentences. POS Tagger assigns parts of speech to each word using tokens such as noun (NN), verb (VB), and adjective (JJ). The tag set in Table 1 shows thirty-six POS tags used by Stanford POS Tagger (Taylor et al. 2003).

2.4 Innovative idea extraction

Innovation is an unusual recombination of prior knowledge (Nelson and Winter 1982; Basalla 1988; Weitzman 1996; Fleming 2001). The identification of innovative ideas in the scientific literature can be classified into those at sentence-level and at phrase-level based on the extraction unit (Wen, Xu, Lai and Wen 2005; Leng et al. 2013). There are two primary categories of extraction technology: feature-based methods (Wen, Wen, Xu and Pan 2005) and machine learning methods (Freitag 1998).

Feature-based methods use the linguistic features of the sentences in which the innovative ideas are located to extract the candidates. Dahl (2008) constructed a list of linguistic features that potentially indicate new research contributions to identify knowledge claims automatically. Wen, Wen, Xu and Pan (2005) established the information relationship between innovation and knowledge claims; thus, a feature-based knowledge spectrum was aggregated by related sentences to extract innovative ideas. However, one limitation of this method is that the feature-based rules are constructed manually by linguistics experts. In addition, the selected features and the rule formulations cannot cover all the linguistic phenomena of the target text.

Machine learning methods primarily create rules by learning a pre-annotated corpus. Experts in fields initially annotate the corpus using certain specifications; subsequently, a system trained on that corpus handles new texts automatically (Leng et al. 2013). Soderland (1997) proposed a knowledge extraction system that used covering algorithms and assembled a set of text analysis rules.

Tag	Part of speech	Tag	Part of speech
CC	Coordinating conj.	PP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition	SYM	Symbol
JJ	Adjective	TO	Infinitival *to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past sense
MD	Modal	VBG	Verb, gerund/present pple
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBZ	Verb, 3rd ps. sg. present
NNP	Proper noun, singular	VBP	Verb, 3rd ps. sg. present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Table 1. Part of the speech tag set used in the Stanford POS Tagger.

Huang and others (2012) transformed the problem of innovation extraction into a classification problem. Classification features such as word frequency, sentence length, and verb characteristics were selected to train the classifier. A machine learning method is faster than a manual method but requires sufficient training data. In addition, these supervised learning methods require manual training set annotation, and the system's performance is affected by the marked corpus.

These methods extract potentially innovative ideas from paper abstracts but do not validate whether the extracted ideas are truly innovative. For example, the authors of a paper might deem their work to be innovative even though such work had been previously proposed, because the authors had not read the related previous paper. Thus, they describe the work as innovative using the linguistic features usually used to introduce innovative ideas. No matter how well the existing innovation extraction methods perform, they will identify such work as innovative when it is not. To ensure that the ideas extracted are innovative, we propose an automatic approach that combines N-gram extraction and web search techniques to extract and identify the innovative ideas of scientific papers. We used the Scopus® database as the official corpus to identify the extracted ideas. Our main contribution is to provide the foundation for a scientific paper content analysis and evaluation system.

3.0 Innovative ideas in research papers

The innovative ideas in research papers considered here are the ideas that have not appeared in previously published papers. Authors of research papers need to demonstrate that their works are rational. Therefore, they prefer to use terminology that is known. Otherwise, it is difficult for researchers to use academic papers to communicate. Authors also need to express the differences between their work and existing studies, as is the case for the examples given in the last paragraph of this section.

Here, innovative ideas are extracted according to their novelty, and they range from very small new ideas to major innovations. This pilot work aims to establish a convenient and reliable method to extract innovative ideas from research papers, and thus, the developed method provides innovative ideas for future works to further grade them.

The present work extracts the innovative ideas of a paper at the word level. Some were expressed by a single phrase, such as probabilistic latent semantic analysis (PLSA) in the paper by Hofmann (1999). Some of the other innovative ideas were expressed as a combination of two or more phrases. These phrases may be technologies presented before the paper was published or may be applications of existing technologies, indicating that many research studies were built upon previous endeavours. For example, Li and others (2016) first combined the Biterm topic model and K-means clustering algorithm when they sought to discover topics from blogs. In another example, Recchia and Louwerse (2016) applied cognitive science approaches to Indus script to estimate the provenance of artefacts with unknown origins (the geographic origin of artefacts from the Indus Valley Civilization), an application of this technique that had not been previously proposed.

4.0 Data and methodology

Our innovative-idea extraction method was limited to the abstracts of each scientific paper; we did not analyse the full texts for the following reasons: 1) an abstract can represent the important content of the paper (Salager-Meyer 1990), and a well-written abstract can be considered key to understanding the original argument (Swales 1990). Therefore, the abstract can be employed as a summary of the main work of the whole research; 2) because abstracts are much shorter in length than the full text, judging innovative ideas from the abstract corpus is an efficient approach; 3) an English-language abstract can help overcome language barriers (Cross and Oppenheim 2006; Small et al. 2014). Many articles written in other languages also provide an English-language abstract containing the central themes to widen readers' access to research; and, 4) access to the full texts of papers is often restricted for some journals; however, the abstracts of papers are always freely obtainable if the institution subscribes to a document database such as Scopus® that indexes the journals in which the papers appeared. Therefore, using abstracts broadened the scope of our investigation.

In this pilot work, we limited our investigation to technology and engineering papers, because the abstracts of theoretical research papers often include analysis that interferes with automatically extracting innovative ideas; thus, automatically extracting innovative ideas from abstracts of technology and engineering papers is both more probable and simpler. Automatically extracting innovative ideas from theoretical research papers will be attempted in future studies. Specifically, we used semantic analysis papers to test our method, because it is not difficult for us to understand the content of papers in this area.

Papers on semantic analysis (excluding theoretical research papers) were used to exemplify the presented automatic innovative-idea extraction method. We downloaded 1,663 abstracts from Scopus®, limited to those whose title, abstract, or keywords contained "semantic analysis," whose publication year was from 2011 to 2017, and whose document type was article or article in press. Our research objective addresses engineering papers. Therefore, we identified 1,014 articles that do have certain engineering innovation

and excluded reviews, questionnaire analyses, comparisons of existing methods, and specific technology evaluations.

Ideas within a paper include what the authors intended to do and how they did it. As most papers published currently report positive results, ideas within a paper comprise not only what its authors intended but also what they achieved to do, including the purpose, technology, application, etc. The knowledge organization theory (Smiraglia and van den Heuvel 2013) shows that works are made up of ideas and that ideas are made up of concepts, which can be expressed by words. The expressions and applications of concepts have been extensively researched. Interested readers can refer to reviews (e.g., Dahlberg 2006; Hjørland 2017; Kleineberg 2017; Arboit 2018; Mazzocchi 2018) and the references therein. Therefore, our work automatically extracted innovative ideas at the word-level, that is, we extracted the N-grams that reflect the main content of the article and then judged whether the work is innovative. An N-gram is a contiguous sequence of N items; here, an N-gram is defined as a noun phrase, because the main concepts of sentences are carried primarily by noun phrases (Kamp 2008). N refers to the number of words the noun phrases contain; it is variable and determined by the extraction results of the Stanford POS Tagger.

We evaluated the performance of the automatic innovative-idea extraction method by comparing its results with manual judgements. The two authors of this work read the abstracts, provided their judgements on the research innovations in the corresponding papers, and retrieved the ideas using Scopus® to determine whether the ideas had emerged previously. We eliminated any disagreements by discussion to construct a final standard for assessing the automatic innovative-idea extraction method developed below. This tedious work was time consuming and limited the number of abstracts that could be used in the experiment.

We define $\varphi(a)$ as the set of innovative ideas extracted by the automatic approach and $\varphi(b)$ as the standard set based on the manual judgements used for comparison. Here, $\varphi(a)$ contains three subsets: $\varphi(a_1)$ is the subset of innovative ideas that are completely included in the standard, i.e., $\varphi(a_1) = \varphi(a) \cap \varphi(b)$; $\varphi(a_2)$ contains synonyms or different expressions of the elements in $\varphi(b)$, and $\varphi(a_3)$ consists of noise candidates. The metrics used to evaluate the performance of the method are recall and precision. Recall is the proportion of the manually judged innovative ideas extracted by the automatic method, which can be notated as Recall = $|\varphi(a) \cap \varphi(b)| / \varphi(b)$. Precision is the proportion of the automatically extracted innovative ideas that match the artificially judged standard or their synonyms—in other words, Precision = $|\varphi(a_1) \cup \varphi(a_2)| / \varphi(a)$.

Figure 1 shows that the automatic innovative-idea extraction method consists of the following four steps:

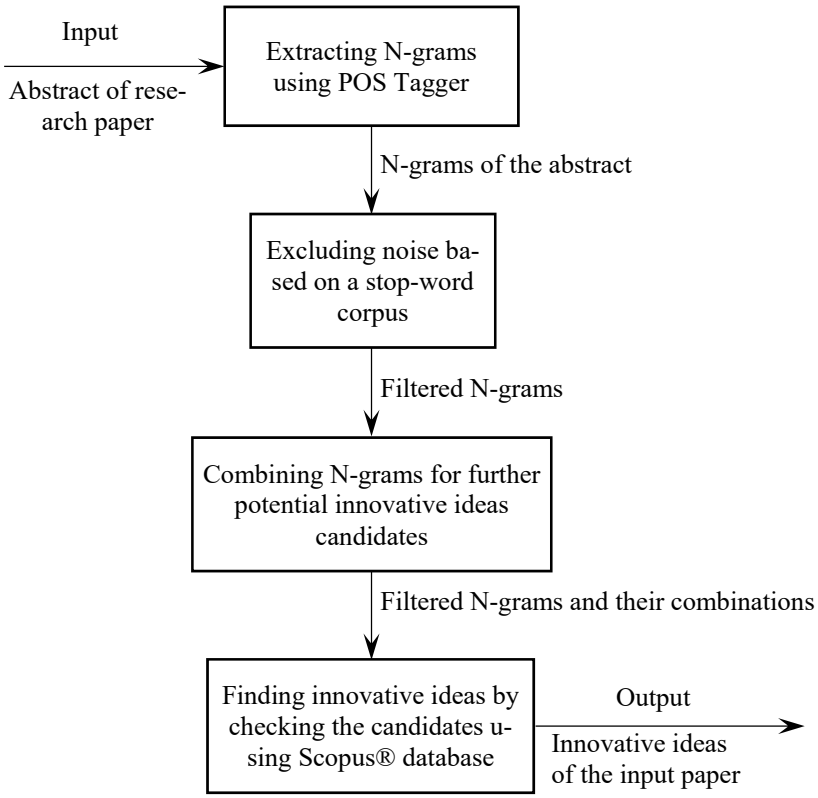

Figure 1. Flowchart of the automatic innovative-idea extracting method.

1) Extracting N-grams from each abstract using POS Tagger, provided by the Stanford Natural Language Processing Group;
2) Excluding noise based on a stop-word corpus, extended from the work of Liu and others (2015), and excluding N-grams not included in move P or move M;
3) Combining two N-grams if necessary;
4) Checking the extracted N-gram candidates using the Scopus® database to determine whether they represent innovative ideas.

More details of these four steps are explained below.

4.1 N-gram extraction

In the first step, we used the Stanford POS Tagger for POS tagging and extracted all the N-grams, excluding symbols, markers, numbers, special characters and tokens, which were tagged for example as personal pronouns (PRP) and determiners (DT). For the reasons explained above, the N-grams here were confined to noun phrases; thus, the extracted N-grams should be expressed and tagged as the following forms:

– A sequence of nouns (e.g., "collection," "text classification," "information retrieval technology")
– Noun-grams following one or several adjectives (e.g., "conceptual representation," "salient semantic analysis")
– "Noun-grams" with "adjective-grams" and the conjunction "and." This type of N-gram should be divided into two N-grams, because the conjunctions can be noise when retrieved from the database. There are two dividing situations:
  a) N-grams expressed as "adjective-gram(s) noun-gram(s)1 and noun-gram(s)2" should be divided into "adjective-gram(s) noun-gram(s)1" and "adjective-gram(s) noun-gram(s)2." For example, the N-gram "characteristic extraction and detection" should be divided into "characteristic extraction" and "characteristic detection."
  b) N-grams expressed as "adjective-gram(s)1 and adjective-gram(s)2 noun-gram(s)" should be divided into "adjective-gram(s)1 noun-gram(s)" and "adjective-gram(s)2 noun-gram(s)." For example, the N-gram "geo-tagged and time-tagged data" should be divided into "geo-tagged data" and "time-tagged data."

4.2 Noise exclusion

The last step obtained an N-gram set from each abstract. However, this set includes some descriptive adjectives and words that do not carry pertinent information and that are used only for writing and thus are not ideas. For example, the N-grams "ubiquitous network text," "proposed machine learning algorithm," and "well-known retrieval algorithm" contain adjectives such as "ubiquitous," "proposed" and "well-known" that should be removed when they are the first words of N-grams, because they do not express an innovative idea. For the same reason, certain nouns such as "sample," "method," and "approach" should be removed when they are the last term of an N-gram. In addition, certain writing words with low information that can be used in papers in many domains should also be removed. Examples are "algorithm," "framework," and "importance," which contain low or no specific concept information when used alone for writing purposes.

Liu and others (2015) assembled a set of noun-phrase filtering terms for the same purpose. Here, we extended their set and assembled two sets of stop-words—descriptive adjectives and terms used for writing purposes. Table A1 shows the descriptive adjective stop-words. When one such adjective is the first term of an N-gram, we remove it and retain the remaining words in the N-gram for subsequent steps. Table A2 shows two types of writing phrase stop-words.

In addition, some concepts are used in abstracts for enumeration following the phrase "such as." When a sentence has an "A such as B, C and D" structure, the concepts B, C and D are generally attached to A. The focus of this sentence is concept A. If B, C and D are important, there should be other sentences describing them. Therefore, we can ignore B, C and D if they are not described elsewhere. Texts also contain concepts used for comparison following the characterization of, for example, "other," "different from," "unlike," and "in contrast to." The concepts listed after these characterizations are not the main idea of the abstract and thus should also be removed. For example, in the sentence, "This paper introduces the construction of the Semantic Lexicon of Dermatology by using the theory and technology of Natural Language Processing (NLP) which can provide the database, such as automatic semantic analysis, word sense disambiguation, for NLP" (Zhou et al. 2016), the concepts following "such as" consist of an enumeration of NLP technology, which is not the focus of the sentence. In the sentence, "Unlike some traditional forecasting model based on several movie-related features, this paper comprehensively utilizes the real-time social media, microblog, to realize a more accurate weekly box office forecasting model" (Chen et al. 2016), the concepts following "Unlike" are existing technology used for comparison purposes and should be removed.

In addition, in the experiment, we found that the ideas are mainly distributed in the move P and move M portions of an abstract. The phrases listed in Table 2 are the characterizations that mark the sentence as the beginning of a research description and appear after move B, while the

are designed; are developed; are presented; are proposed; are shown;
design/methodology/approach;
is designed; is developed; is presented; is proposed; is shown;
materials and methods; methods/methods;
our;
the article; the article here; the paper; the present; the study; this article; this context; this contribution; this letter; this paper; this present; this publication; this research; this study; this work;
was designed; was developed; was presented; was proposed; was shown;
we; were designed; were developed; were presented; were proposed; were shown

Table 2. Characterizations in sentences that indicate the beginning of a research description.

analysis revealed that; as result; are demonstrated;
comparative experiments; comparison experiments; conclusion; conclusions; contrast experiment;
evaluation experiments; evaluation show; experimental data shows; experimental results; experimental study; experiments demonstrate; experiments on; experiments reveal; experiments show; experiments were performed;
final conclusion; findings -; findings; findings indicate; for evaluation; for evaluation;
in experiment; in experiments; in sum; is evaluated; is demonstrated;
our experiment; our result;
perform experiments; practical experiment; promising result;
result; result achieved; result indicates that; result proves; results:; results are compared to; results demonstrate; results provide; results show; results show that; experimentation showed that that; results suggest;
shows comparable performance; simulated experiment;
the experiment; to illustrate; test showed that;
was tested; we demonstrate; was evaluated; we evaluate; we perform; when compared to

Table 3. Characterizations in sentences that indicate the end of a research description.

phrases listed in Table 3 are the characterizations that mark the sentence as the beginning of move R or the end of the research description. The proposed method excludes N-grams that are not included in move P and move M.

4.3 N-gram combination

Considering that some innovative ideas are expressed as a combination of N-grams, we applied a rule-based approach to combine certain N-grams from the filtered N-gram set. Appertaining means that purpose and meaning occur together in one sentence (Thorleuchter 2008); thus, such innovative ideas might be represented by a combination of certain concepts that occur together in the same sentence. Therefore, we combine two filtered N-grams that are not new methods or concepts but are adjacent in a sentence that contains "of," "to," "with," "by," "for," or a characterization word such as "based," "utilize," "apply," "combine," "and," or "conducted" (Liu et al. 2015) and use the combination as a new N-gram in the following step.

We created this combination rule for two reasons. First, the association between elements in a sentence is stronger than associations in different sentences (Thorleuchter 2008), and within a sentence, the association is stronger between adjacent elements than between non-adjacent ones. Second, combination rules should be highly efficient, because there is a usage limitation per week for one Scopus® API key (see the next sub-section). Suppose one sentence contains M N-grams that must be combined. There would be $M \times (M - 1)/2$ combinations between any two N-grams but only $M - 1$ combinations based on our rule (combining only adjacent N-grams). The number of combined N-grams of the former is $M/2$ times the latter.

4.4 Innovation judgement of N-grams

From the aforementioned steps, we now have a set of N-grams with filtered individual and combined concepts that must be classified into innovative ideas or existing concepts. The criterion used to judge these ideas as innovative ideas is that the idea should not have been previously proposed. We used all the abstracts collected by the Scopus® database as a corpus and retrieved the N-gram candidates from the Scopus® database automatically using the Scopus® application programming interface (API) to determine whether they were innovative ideas.

The combination of terms in Section 4.3 was realized in this step. For example, to check whether the combination of the Biterm topic model and K-means clustering algorithm in the study by Li and others (2016) is an innovative idea, we conducted a search of the two terms with the "and" operation in Scopus®.

Scopus® APIs allows researchers to integrate content and data from the Scopus® database into their own websites and applications. Curated abstracts and citation data of all scholarly journals indexed by Scopus®, Elsevier's abstract and citation database, can be retrieved using Scopus® APIs (https://dev.elsevier.com/sc_apis.html/). There is an API key for each API, and there is a usage quota enabled for each API key per week (https://dev.elsevier.com/api_key_settings.html/). The quota for our abstract retrieval here is 10,000 per week (i.e., we can send up to 10,000 retrieval requests every week to Scopus® using our API key).

When using the Scopus® APIs to automatically retrieve extracted N-grams from the Scopus® database to judge whether the N-grams are innovative, we limited the retrieval scope to abstracts with publication dates before the publication year of the paper inspected. For a candidate idea in the abstract of a paper inspected (P_{ins}), if no abstract of a paper published before the year of the publication of P_{ins} mentioned the candidate idea, the idea was classified as innovative. Here, we checked whether an idea is innovative using the Scopus® platform rather than Web of Science, because we have not found any API for the latter platform.

4.5 An example

Here, we exemplify the proposed method with the following paper: "Dai, W, You, Y, Wang, W, Sun, Y, Li, T. (2011) Search engine system based on ontology of technological resources. *Journal of Software, 6*(9): 1729-1736." Its abstract is as follows:

> Internet has become a huge and updating information warehouse, and provides a new source for us to build a well technological resources sharing system to support our research work and development activities. However, the technological resources on Internet is usually diverse, professional and complex. They are difficult to be retrieved precisely and completely by traditional search engines. This paper proposed a new search engine system based on ontology of technological resources. In that system, a database with ontology knowledge warehouse was designed to store all related conceptions and the relationships of technological domains. By semantic analysis of users' queries and a heuristic search, the expected technological resources can be retrieved more precisely and completely to satisfy their intentions.

The N-grams extracted from its abstract are as follows: Internet, information warehouse, new source, technological resources sharing system, research work, development activities, technological resources, traditional search engines, paper, new search engine system, ontology, technological resources, system, database, ontology knowledge warehouse, related conceptions, relationships, technological domains, semantic analysis, queries, heuristic search, technological resources, intentions.

After excluding noise and the N-grams not in move P or M, we obtained the following N-gram candidates: search engine system, ontology, technological resources, ontology knowledge warehouse, technological domains, semantic analysis, queries, heuristic search, technological resources, intentions.

Using the combination strategy mentioned in Section 4.3, we added the following combined N-grams to the N-gram candidates: "search engine system and ontology," "ontology and technological resources," "semantic analysis and queries," "queries and heuristic search," "heuristic search and technological resources."

By retrieving the N-gram candidates from the Scopus® database automatically using the Scopus® API, we found that the following N-gram candidates had not appeared in the publication year of the example paper: "ontology knowledge warehouse," "heuristic search and technological resources."

Obviously, the combination N-gram candidate "heuristic search and technological resources" is not a specific innovative idea and is an error in the results. Building an ontology knowledge warehouse as a database that includes all related conceptions and relationships of the technological domain as the query conditions is an innovative idea (although it is a small new idea) in this example paper for precise and complete retrieval.

5.0 Results

From the 1,014 abstracts used in the experiment, 4,399 N-grams were finally extracted and classified automatically as innovative or non-innovative ideas by our method. In addition, 2,295 manually judged innovative ideas were used as the standard for evaluating the method. Among the 4,399 extracted innovative ideas, 2,272 matched the manually judged innovative ideas; thus, the precision was 51.6%. Among the 2,295 innovative ideas judged manually, 1,991 were extracted by the automatic method; thus, the recall was 86.8%.

5.1 Effects of noise exclusion

When stop-words and the terms for enumeration and comparison were not excluded, more resulting N-grams appeared combined with these stop-words and terms as noise, but they were classified as innovative ideas automatically, which reduced the precision to 35.8%.

5.2 Mistakes of data and POS Tagger

Spelling mistakes in the original abstracts and the errors by the POS Tagger made during the N-gram extraction process using the Stanford NLP tool also reduced the performance of our method. Across all 1,014 of the experimental abstracts, seventy-three text mistakes were found, and examples of them appear in Table 4. Additionally, the POS Tagger made forty-five errors, and examples of these errors are listed in Table 5. When these mistakes and the subsequent noise combinations are included, the precision decreases to 49.8%. These errors prevented the extraction of three combined innovation ideas, which reduced the recall to 86.6%.

5.3 Location distribution of innovative ideas in abstracts

As shown in Table 6, when the text to be processed contained move B but does not contain move R, 5,780 N-grams were classified as innovative ideas by the automatic extraction method, reducing the precision to 39.4%. When the text to be processed contained move R but not move B, 5,681 N-grams were classified as innovative ideas automatically, reducing the precision to 40.1%. Only four and five abstracts mentioned innovative ideas in moves B and R, respectively. Recall increases slightly when the method considers move B or R, as shown in Table 6. The results show that limiting the text to be processed to that occurring between move B and move R excludes much interference and improves the efficiency and accuracy of our work.

Table 6 shows the compared results of our experiments as discussed above. Because of the rule of combining two adjacent N-grams in the same sentence, when stop-words are not excluded, they sometimes prevent the combination of the two surrounding technology concepts; thus, these instances miss the chance to be judged as innovative ideas. Without removing stop-words, recall decreased to 80.3%. In addition, without combining two adjacent N-grams,

Errors	Correct
timesequential images	time sequential images
bag-of- word model	bag-of-word model
shot segmentationsin videos	shot segmentations in videos
spectralanalysis	spectral analysis
weightestimation algorithms	weight estimation algorithms
machine learning techniques In this study	machine learning techniques. In this study
models word sense disambiguationand	models word sense disambiguation and
event relation ship	event relationship
automated semantic analyses.We	automated semantic analyses. We
manyanonymity algorithms	many anonymity algorithms
concept-basedand	concept-based and

Table 4. Examples of textual errors in the original abstracts.

No	Sentence	Term	Wrong POS	Correct POS
1	help the government offer more effective assistance	offer	NN	VB
2	guide the lexicographer through his/her task	his/her	NN	PRP
3	performance benefits from a syntactic-based definition	benefits	NNS	VB
4	link identification numbers with a semantic enrichment process	link	NN	VB
5	The PAM first extract the dominant color compositions	first extract	JJ NN	RB VB
6	The BIOMedical Search Engine Framework	search	VB	NN
7	(PLSA) method is developed to leverage attribute information	leverage	NN	VB
8	develop a lexical database of Punjabi verbs leveraged in the form of a dictionary of verbs	leveraged	NN	VBN

Table 5. Examples of POS Tagger errors.

Process	Precision	Recall
Without removing stop-words	35.8%	80.3%
Including textual and parser tool errors	49.8%	86.6%
Containing text to be processed before move B	39.4%	86.9%
Containing text to be processed after move R	40.1%	87.0%
Without combining adjacent N-grams	56.7%	37.1%
Final result with 4 improvement steps	51.6%	86.8%

Table 6. Precision and recall of the proposed method under different conditions.

precision increased to 56.7% because of the reduction in noisy combinations, but recall decreased to 37.1%.

6.0 Discussion

This paper introduced an automatic approach to extract innovative ideas from the abstracts of technology and engineering research papers. The results show the feasibility and effectiveness of the method; however, its performance can still be improved.

One challenge is that the performance of the innovative-idea extraction method depends to some extent upon the quality of the abstracts. One type of quality in abstracts involves clarity of presentation (Timmer et al. 2003). Abstracts that clearly, concisely, and unambiguously present the main points of the research are ideal targets for our work. In reality, most of the abstracts used in this study proved to be sufficiently good to achieve the novelty extraction purpose, but there were some unsatisfactory examples for which the extraction failed. In addition, some unstructured abstracts lack obvious characteristics to identify the portion that describes the central work of the research. For example, some abstracts do not contain the features often used in the first sentence of the result or comparison descriptions of the research in abstracts, such as "experimental result" (see Table 3), which resulted in noise candidates and reduced the precision.

Synonymy caused by different authors' writing styles is also a significant challenge in our work. Synonymy means that meanings can be expressed in several different forms (Miller et al. 1990), which leads to a problem in automatically judging innovative ideas; an N-gram in an abstract may be an alternative expression of an existing concept. However, the retrieved result for that N-gram from Scopus® indicates that it had not been previously proposed before the paper's publication; thus, the N-gram is mistakenly identified as innovative. For example, "latent context features" (Ren and Wang 2016) has the same meaning as "context-based latent features;" however, the former expression could not be retrieved from Scopus® before the paper's publication year; thus, the method misjudged this candidate as an innovative idea. Synonyms led to the greatest reduction in the method's precision. Therefore, we plan to introduce Wordnet (Miller et al. 1990) into future work to reduce the negative influence of synonyms.

The unique experimental tools, data, or platforms used in some research also form noise that reduced the precision. For example, Ben Aouicha and others (2016) exploited seventeen datasets for semantic similarity purposes and semantic relatedness evaluation. These datasets, including *RG65*, *MC30*, *AG203*, etc., had not been used in other research based on Scopus® retrieval, and, therefore, they were automatically classified as innovations by our method. Although using these seventeen datasets can be considered novel research to some extent, they are not innovative ideas by themselves, and including them increases the number of noise candidates and reduces precision.

One shortcoming of the method is the rule of combining only two adjacent N-grams in one sentence to express the potential innovative ideas. This limitation might miss innovative ideas that combine three or more technologies, that are described in several steps in different sentences, or that are represented as two non-adjacent concepts. For example, Renu and Mocko (2016) aimed to enable retrieval and knowledge sharing of text-based assembly process plans, and one innovative idea of their research lies in combining the four text-mining algorithms "word overlap," "Jaccard score," "term frequency-inverse document frequency," and "latent semantic analysis;" however, this innovation cannot be extracted when only two N-grams are allowed to be combined. Another example is Yuan and others (2016), who introduced an approach to analyse and model relationships among image sequences and key postures. They described their work in four steps. Our automatic method does not extract innovative ideas reflected in a combination of several steps proposed in different sentences. To address this problem, in future research, we will attempt to both combine N-grams efficiently from one sentence and concepts emerging in different sentences.

Another shortcoming of the combination of N-grams is that due to the search quota limitation of the Scopus® API, we did not recheck the retrieved results from Scopus®. Our method retrieves a combined concept using a

strategy that relates its two terms using the operator "AND;" this approach returns all the abstracts that include the two N-grams. However, the two terms can appear in different sentences in some result abstracts; thus, the association of those two terms might be weak in those abstracts. The Scopus® retrieval rules allow researchers to use a location qualifier operator to limit the distance between two search terms in the abstract to a specific value. This capability is beneficial for limiting the two terms to one sentence. However, with this limitation, the retrieval results might miss similar work describing the combined concepts in different sentences. In the future, to improve the accuracy of the results, we plan to recheck the retrieval results by inspecting the association of the concepts included in combination candidates in the returned abstracts.

There is another reason for rechecking the returned abstracts in the future version of the method. There are two types of rules for retrieving N-grams in Scopus®: exact match and approximate match. Exact match uses braces ({}) around the phrase to be retrieved, and the results must contain the exact phrase that occurs between the braces. Approximate match uses quotes ("") around the phrase to be retrieved, and in this case, the results contain the adjacent words of the phrase but might also contain punctuation between them. In addition, when an approximate match uses the singular form of a word in the strategy, the results may include its singular, plural, and possessive forms for most words. Thus, we use approximate match in our method, because doing so can reduce omissions caused by different word forms. However, because the approximate match method ignores punctuation, when we retrieve the N-grams "Word1 Word2 Word3," for example, "Natural Language Processing," punctuation might occur between the three continuous words in the returned documents, for example, "Natural Language, Processing," which does not meet our expectations. We randomly inspected 1,027 retrieved phrases with more than two words in returned abstracts and found that eighteen results contained punctuation between the retrieved continuous terms, corresponding to an error rate of 1.75%. Therefore, we plan to recheck the N-grams in returned abstracts to determine whether punctuation exists in the continuous terms, which will help to ensure the consistency of strategies and returns. The rechecking work is time consuming but could increase the recall.

7.0 Implications

In contrast to co-word and co-citation analysis, which are used to investigate the relationships between papers, innovative-idea extraction reflects the differences in a paper from all previously published papers. Co-word and co-citation analysis are two major clustering methods used to delimitate science subfields (Olmeda-Gómez et al. 2017) for exploring the structure of scientific literature (Small and Griffith 1974). One function is to detect the research front (Zitt and Bassecoulard 1994). The results show the differences between different classes of research papers and the similarities between papers in a same class. Our method characterizes the novelty of a paper by comparing it with all previous papers, even though the difference may be slight. In short, clustering methods show differences among papers at the class level, while our method shows differences among papers at the article level.

Research evaluations and scientific research policies that affect researchers' careers influence researchers' behaviours. For example, the policy that university funding should be based on only the number of publications, which was implemented in Australia in 1995, mostly led to greater publishing activity in low-quality journals (Butler 2003). Currently, research evaluation is mainly based on the number of research publications and citations. This notion encourages researchers to study hot topics, as papers on hot topics are more likely to be accepted for publication and to receive more citations. Such research belongs to normal science. Another kind of research is scientific "revolution," that is, the creation of a paradigm shift (Kuhn 1962). Scientific "revolution" is caused by breakthrough and is excellent science (Spier and Poland 2013). These studies can be recognized by identifying their differences from previous studies at the article level or by identifying the new knowledge that they provide to human society. Additionally, if research evaluation considers innovativeness, it will encourage researchers to pursue breakthroughs.

Innovation involves a paradox: innovation is important for science development, but ideas with a higher level of originality have a higher risk of rejection by audiences (Staw 1995; Cooper 2007; Mueller et al. 2012), even by academic journals (Starbuck 2003). Readers prefer normal-science work or innovative works with fewer new elements for two reasons. One reason is that existing research work has provided partial recognition for the contribution; the other reason is that the professional knowledge of audiences is occasionally not consistent with that in innovative works (Trapido 2015). Our method can help readers, including journal editors and reviewers, to know and consider carefully the innovative ideas of research papers.

This work is an extension of knowledge organization research. Mazzocchi (2018) proposed that there are two main items that characterize knowledge organization. One is the knowledge organization process, and the other is the knowledge organization system. This work is based on the knowledge organization theory and uses an abstract database of academic papers. Therefore, it is related to another item that characterizes knowledge organization: knowledge organization application.

8.0 Limitation

In the experiments of this work, we used a self-judged standard to assess the performance of the proposed method. Although we established this standard carefully and prudently, it is inevitable that bias is included in this standard, thus leading to some errors in the performance evaluation. However, comparisons of the performance of the method with those of different measures show that the method can be improved in appropriate ways. The discussion section also provides potential methods for further improvement. Together, these data and proposed improvements fulfil the aim of this paper, which is automatically extracting innovative ideas from papers and has the possibility to be fully achieved and applied in practice, though the method needs further refinement.

9.0 Conclusions

We performed an experimental investigation into the innovative ideas that exist in semantic analysis papers. We proposed an automatic approach to extract the knowledge claims in abstracts and judge whether they are innovative by comparing them with retrievals from the Scopus® database. This approach does not require manually assembling a domain corpus. A list of stop-words and the characteristics of research descriptions were developed to exclude noise. By considering the distribution of text describing innovation in abstracts and excluding stop-words, the performance of our method was improved. We believe that with further improvement, our research will be helpful to the development of a research evaluation system.

References

Arboit, Aline Ellis. 2018. "Knowledge Organization: From Term to Concept, From Concept to Domain." *Knowledge Organization* 45: 125–36.

Basalla, George. 1988. *The Evolution of Technology*. Cambridge: Cambridge University Press.

Ben Aouicha, Mohamed, Mohamed A. Hadj Taieb and Abdelmajid Ben Hamadou. 2016. "LWCR: Multi-Layered Wikipedia Representation for Computing Word Relatedness." *Neurocomputing* 216: 816–43.

Bounhas, Ibrahim and Yahya Slimani. 2009. "A Hybrid Approach for Arabic Multi-Word Term Extraction." In *2009 International Conference on Natural Language Processing and Knowledge Engineering*. [Piscataway, N.J.]: IEEE, 429–36. doi:10.1109/NLPKE.2009.5313728

Butler, Linda. 2003. "Explaining Australia's Increased Share of ISI Publications the Effects of a Funding Formula Based on Publication Counts." *Research Policy* 32: 143–55. doi:10.1016/S0048-7333(02)00007-0

Chen, Kuang-hua and Hsin-Hsi Cshen. 1994. "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation." In *32nd Annual Meeting of the Association for Computational Linguistics: Proceedings Of The Conference: 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA.* [Morristown, N.J.]: Association for Computational Linguistics, 234–41.

Chen, Runyu, Wei Xu and Xinghan Zhang. 2016. "Dynamic Box Office Forecasting Based on Microblog Data." *Filomat* 30: 4111–24.

Church, Kenneth W. and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16: 22–9.

Cole, Jonathan R., Stephen Cole, and Donald D. Beaver. 1974. "Social Stratification in Science." *American Journal of Physics* 42: 923–4.

Cooper, Robert G. 2007. "Managing Technology Development Projects." *IEEE Engineering Management Review* 35: 67–76.

Cross, Cate and Charles Oppenheim. 2006. "A Genre Analysis of Scientific Abstracts." *Journal of Documentation 624*: 428–46.

Dahl, Trine. 2008. "Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics." *Journal of Pragmatics* 40*:* 1184–201.

Dahlberg, Ingetraut. 2006. "Knowledge Organization: A New Science?" *Knowledge Organization* 33: 11–9.

Dirk, Lynn. 1996. "From Laboratory to Scientific Literature: The Life and Death of Biomedical Research Results." *Science Communication* 18: 3–28.

Dirk, Lynn. 1999. "A Measure of Originality: The Elements of Science." *Social Studies of Science 29:* 765–76.

Endres-Niggemeyer, Brigitte. 1998. *Summarizing Information: Including CD-Rom "SimSum", simulation of summarizing, for Macintosh and Windows*. Berlin: Springer.

Fiala, Jaroslav, Jiří J.Mareš, and Jaroslav Šesták. 2017. "Reflections on How to Evaluate the Professional Value of Scientific Papers and Their Corresponding Citations." *Scientometrics* 112: 697–709.

Fleming, Lee. 2001. "Recombinant Uncertainty in Technological Search." *Management Science* 47: 117–32.

Frame, J. Davidson. 2008. Review of *Reinventing Project Management: The Diamond Approach to Successful Growth and Innovation, by Aaron J. Shenhar and Dov Dvir. Project Management Journal* 39: 96. doi:10.1002/pmj.20027

Freitag, Dayne. 1998. "Multistrategy Learning for Information Extraction." In *Machine Learning: Proceedings of the Fifteenth International Conference (ICML '98)*, ed. Jude Shavlik. San Francisco: Morgan Kaufmann, 161–9.

Garfield, Eugene. 1972. "Citation Analysis as a Tool in Journal Evaluation." *Science* 178: 471–9.

Garfield, Eugene. 1979. *Citation Indexing*. New York: Wiley.

Guetzkow, Joshua, Michele Lamont, and Gregoire Mallard. 2004. "What is Originality in the Humanities and the Social Sciences?" *American Sociological Review* 69: 190–212.

Hartley, James and Lucy Betts. 2009. "Common Weaknesses in Traditional Abstracts in the Social Sciences." *Journal of the Association for Information Science and Technology* 60: 2010–8.

Hofmann Thomas. 1999, "Probabilistic Latent Semantic Analysis." In *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference (1999), July 30-August 1, 1999, Royal Institute of Technology (KTH), Stockholm, Sweden*, ed. Kathryn B. Laskey and Henri Prade. San Francisco, CA: Morgan Kaufmann, 289–96.

Hjørland, Birger. 2017. "Classification." *Knowledge Organization* 44: 97–128.

Huang, Ke-Chun, Charles C. Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. 2012. "Classification of PICO Elements by Text Features Systematically Extracted from PubMed Abstracts." In *Proceedings: 2011 IEEE International Conference on Granular Computing; Kaohsiung, Taiwan, Nov. 8-10, 2011*, ed Tzung-Pei Hong. Piscataway, NJ: IEEE, 279–83.

Jamar, Nina, Alenka Šauperl and David Bawden. 2014. "The Components of Abstracts: The Logical Structure of Abstracts in the Areas of Materials Science and Technology and of Library and Information Science." *New Library World* 115: 15–33.

Justeson, John S. and Slava M. Katz. 1995. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text." *Natural Language Engineering* 1: 9–27.

Kamp, Hans. 2008. "A Theory of Truth and Semantic Representation." In *Formal Semantics: The Essential Reading,* ed. Paul Portner and Barbara H. Parte. Oxford: Blackwell, 189–222.

Kanoksilapatham, Budsaba. 2013. "Generic Characterisation of Civil Engineering Research Article Abstracts." *3L Southeast Asian Journal of English Language Studies* 19: 1–10.

Kleineberg, Michael. 2017. "Integrative Levels." *Knowledge Organization* 44: 349–79.

Kosten, Joost. 2016. "A Classification of the Use of Research Indicators." *Scientometrics* 108: 457–64.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Leng, Fuhua, Rujiang Bai, and Qingsong Zhu. 2013. "A Hybrid Semantic Information Extraction Method for Scientific Research Papers." *Library and Information Service* 57: 112–9.

Li, Weijiang, Yanming Feng, Dongjun Li, and Zhengtao Yu. 2016. "Micro-Blog Topic Detection Method Based on BTM Topic Model and K-Means Clustering Algorithm." *Automatic Control and Computer Sciences* 50: 271–7.

Liu, Haixia, James Goulding and Tim Brailsford. 2015. "Towards Computation of Novel Ideas from Corpora of Scientific Text." In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II*, ed. Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, João Gama, Alípio Jorge, and Carlos Soares. Lecture Notes in Computer Science 9285. Cham: Springer, 541–56. doi:10.1007/978-3-319-23525-7_33

Mazzocchi, Fulvio. 2018. "Knowledge Organization System (KOS): An Introductory Critical Account." *Knowledge Organization* 45: 54–78.

Maynard, Diana and Sophia Ananiadou. 1999. "Identifying Contextual Information for Multi-Word Term Extraction." In *TKE '99: Terminology and Knowledge Engineering; Proceedings, Fifth International Congress on Terminology and Knowledge Engineering, 23-27 August 1999, Innsbruck, Austria*, ed. Peter Sandrini. Vienna: TermNet, 212–21.

Milas-Bracović, Milica and Jasenka Zajec. 1989. "Author Abstracts of Research Articles Published in Scholarly Journals in Croatia (Yugoslavia): An Evaluation." *Libri* 39: 303–18.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. 1990. "Introduction to WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3: 235–44.

Mueller, Jennifer S., Shimul Melwani and Jack A. Goncalo. 2012. "The Bias Against Creativity: Why People Desire but Reject Creative Ideas." *Psychological Science* 23: 13–7.

Mullins, N., W. Snizek and K. Oehler. 1988. "The Structural Analysis of a Scientific Paper." In *Handbook of Quantitative Studies of Science & Technology*, ed A. F. J. van Raan. Amsterdam: North-Holland, 81–105.

Nelson, Richard R. and Sidney G. Winter. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

Oliver, Antoni and Mercè Vàzquez. 2015. "TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction." In *International Conference Recent Advances in Natural Language Processing Hissar, Bulgaria 7–9 September, 2015: Proceedings*, ed. Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov. Shoumen, Bulgaria: Incoma, 473–9.

Olmeda-Gómez, Carlos, Maria-Antonia Ovalle-Perandones, and Antonio Perianes-Rodríguez. 2017. "Co-word Analysis and Thematic Landscapes in Spanish Information Science Literature, 1985–2014." *Scientometrics* 113: 195–217.

Recchia, Gabriel L. and Max M. Louwerse. 2016. "Archaeology Through Computational Linguistics: Inscription Statistics Predict Excavation Sites of Indus Valley Artifacts." *Cognitive Science* 40: 2065–80.

Ren, Kai and Shi-Wen Wang. 2016. "Improved Convolutional Neural Network for Biomedical Word Sense Disambiguation with Enhanced Context Feature Modelling." *Journal of Digital Information Management* 14: 342–50.

Renu, Rahul S. and Gregory Mocko. 2016. "Computing Similarity of Text-Based Assembly Processes for Knowledge Retrieval and Reuse." *Journal of Manufacturing Systems* 39: 101–10.

Rogers, Debra M. Amidon. 1993. "Knowledge Innovation System: The Common Language." *Journal of Technology Studies* 19: 2–8.

Salager-Meyer, Françoise. 1990. "Discoursal Flaws in Medical English Abstracts: A Genre Analysis per Research- and Text-Type." *Text* 10: 365–84.

Smiraglia, Richard P. and Charles van den Heuvel. 2013. "Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction." *Journal of Documentation* 69: 360–83.

Small, Henry, Kevin W. Boyack, and Richard Klavans. 2014. "Identifying Emerging Topics in Science and Technology." *Research Policy* 43: 1450–67.

Small, Henry and Belver C. Griffith. 1974. "The Structure of Scientific Literature. I: Identifying and Graphing Specialties." *Science Studies* 4: 17–40.

Smadja, Frank. 1993. "Retrieving Collocations from Text: Xtract." *Computational Linguistics* 19: 143–77.

Soderland, Stephen G. 1997. "Learning Text Analysis Rules for Domain-Specific Natural Language Processing." Ph.D. diss, University of Massachusetts.

Spier, Ray E. and Gregory A. Poland. 2013. "What is Excellent Science and How Does it Relate to What We Publish in Vaccine?" *Vaccine* 31: 5147–8.

Starbuck, William H. 2003. "Turning Lemons into Lemonade: Where is the Value in Peer Reviews?" *Journal of Management Inquiry* 12: 344–51.

Staw, Barry M. 1995. "Why No One Really Wants Creativity." In *Creative Action in Organizations*, ed. Ford Cameron and Gioia Dennis. Thousand Oaks, CA: Sage, 161–6.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taylor, Ann, Mitchell Marcus and Beatrice Santorini. 2003. "The Penn Treebank: An Overview." In *Treebanks*, ed. Anne Abeillé. Text, Speech and Language Technology 20. Dordrecht: Springer, 5–22. doi:10.1007/978-94-010-0201-1_1

Thorleuchter, Dirk. 2008. "Finding New Technological Ideas and Inventions with Text Mining and Technique Philosophy." In *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, ed. Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker. Berlin: Springer, 413–20.

Timmer, Antje, Lloyd R. Sutherland and Robert J. Hilsden. 2003. "Development and Evaluation of a Quality Score for Abstracts." *BMC Medical Research Methodology* 3: 1–7. doi:10.1186/1471-2288-3-2

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Proceedings of The Conference and Associated Workshops*. East Stroudsburg, PA: Association for Computational Linguistics, 252–9.

Trapido, Denis. 2015. "How Novelty in Knowledge Earns Recognition: The Role of Consistent Identities." *Research Policy* 44: 1488–500.

Weitzman, Martin L. 1996. "Hybridizing Growth Theory." *American Economic Review*, *86*: 207–12.

Wen, Youkui, Guohua Xu, Bonian Lai, and Hao Wen. 2005. *Knowledge Element Mining*. Xi'an: Xi'an Electronic Science & Technology University Press.

Wen, Youkui, Hao Wen, Duanyi Xu, and Longfa Pan. 2005. "Knowledge Element Mining in Knowledge Management." *Journal of the China Society for Scientific and Technical Information* 24: 663–8.

Wieringa, Roel, Neil Maiden, Nancy Mead, and Colette Rolland. 2006. "Requirements Engineering Paper Classification and Evaluation Criteria: a Proposal and a Discussion." *Requirements Engineering* 11: 102–7.

Wu, Zhiqian. 2015. "Average Evaluation Intensity: A Quality-Oriented Indicator for the Evaluation of Research Performance." *Library & Information Science Research* 37: 51–60.

Xu, YanZhang. 2001. "Innovation: The Soul of Science." *Science & Technology Review* 19: 8–11.

Yuan, Hejin, Chunhong Duo and Weihua Niu. 2016. "A Human Behavior Recognition Method Based on Latent Semantic Analysis." *Journal of Information Hiding and Multimedia Signal Processing* 7: 489–98.

Zhou, Yang, Nan Xiang, Ruixiang Wang, Yao Liu, Xingliang Qi and Zhenguo Wang. 2016. "Construction of Semantic Lexicon of Dermatology." *ICIC Express Letters* 10: 1725–30.

Zitt, Michel and Elise Bassecoulard. 1994. "Development of a Method for Detection and Trend Analysis of Research Fronts Built by Lexical or Co-citation Analysis." *Scientometrics* 30: 333–51.

Appendix

Table A1 shows the descriptive adjective stop-words. When one such adjective is the first term of an N-gram, we remove it and retain the remaining words in the N-gram for subsequent steps. Table A2 shows the writing phrase stop-words, of which there are two types. The first type contains the words that should be removed when they are extracted as a single N-gram. The other type occurs when a word such as "sample," "method," or "approach" is the last term of an N-gram; in such cases, we remove the last word and retain the remaining words in the N-gram.

When the following words are the first terms of N-grams, remove the first word and retain the rest.
above; acceptable; accurate; additional; advanced; aforementioned; apparent; appropriate; arbitrary; available;
bad; basic; best; bewildering; brief;
careful; certain; challenging; chaotic; chief; chosen; collected; common; comprehensive; considerable; corresponding; creative; credible; critical; crucial; current;
detailed; different; distinct; distinctive; developed; diverse; difficult;
easy; effective; efficient; elaborate; emphasis; entire; essential; established; eventual; excellent; exhaustive; existing; existent; extracted;
felicitous; few; final; first; following; follow-up
general; great; given; good;
helpfulness; high quality; high-performance; high-quality; huge;
important; improved; improper; incomplete; independent; inappropriate; incremental; insightful; insufficient; interesting; inventive;
judicious;
known;
large; large-scale; large-granular; latest; longstanding;
main; major; many; mass; mass-use; massive; meaningful; methodological; modern; more; most;
namely; necessary; new; next; not; novel; numerous;
obtained; off-the-shelf; old; only; overall; own;
particular; personal; plausible; popular; possible; potential; powerful; practical; precise; pre-existing; previous; primary; prior; progress; promising; proposed;
recorded; related; reasonable; recent; relevant; reliable; respective; rich; robust;
same; satisfying; second; several; sharing; significant; simple; small-scale; so-called; so-called; special; specific; state-of-the-art; strong; subsequent; successful; such; sufficient; suitable; superior
then; total; tough; traditional; trend-breaking; turn; typical;
ubiquitous; understandable; unique; unknown; unnecessary; unreliable; useful; usual;
valid; valuable; various; vast;
well-defined; well-established; well-known; whole; wrong

Table A1. Stop-words of descriptive adjectives used for writing purposes.

The following words should be removed when they are extracted as a single word.
ability; absence; accomplishment; accuracy; achievement; activity; adaptation; addition; adequacy; advance; advantage; advantageous function; agenda; algorithm; amount; analogy; analysis; answer; application; approach; approximation; architecture; area; article; aspect; associate; assumption; attempt; attribute;
background; basis; bulk;
capacity; case; case study; category; cause; challenge; characteristic; class; code; coefficient; collocation; combination; comment; community; companionship; comparison; competence; competency; competitive advantage; complete procedure; completeness; completion; complex; complicated problem; computing precision; concept; conception; conclusion; condition; connotation; consolidated statement; construct; construction; content; contribution; convenience; core; core aim; core attribute; core feature; core idea; correlation; course; creativity;
data; database; dataset; definition; description; design; detail; determination; development; difference; diffusion; dimension; disclosure; discovery; discussion; dissertation; diversity; domain; drawback;
effectiveness; efficacy; efficiency; effort; enhancement module; enrichment; entity; essay; estimation; ethics; evaluation; event; eventual; everyday; evidence; exemplary tasks; example; experiment; explanation; explication; exploration; expression; extensibility;
facet; facility; feature; feature guarantee; field; figure; flexibility; focus; form; formalism; formation; formed indicator; former; frame; framework; function; further; further improvement; further validation;
gap; generation; goal; graph; group;
heavy; heuristic; hiding; high correlation; hypotheses;

idea; identification; image; impact; impetus; implementation; implication; importance; impossibility; improvement; inclusion; incompleteness; inconsideration; inconsistency; increase; individual; information; interconnection; innovative; input; insight; item; inter; interest; integration; interpretable way; investigation; issue;
key factor; kind; knowledge;
label; lack; latter; level; limitation; list; literature;
manifest validity; mean; meaning; measurement; mechanism; medline; mistake; mix; merit; method; methodology; model; modelling; model parameter; module; multifold; multiplicity;
need; node; notion; novelty; number;
object; objective; ones; operation; opinion; order; original source; outlook; output;
pace; paper; paradigm; parallel application; parameter; part; participant; people; performance; performance characteristic; period; phenomena; phrase; piece; platform; point; popular approach; portion; possibility; practicality; practice; precision value; precondition; prework; present; presence; present paper; principle; probe; problem; problem situation; procedure; process; processing; product attribute; project; promising approach; proposal; proposition; purpose;
quality; quantity; question;
range; raw data; realization; record; redundancy; reference; reflection; reformation; relation; relationship; relative improvement; relative value; reliability; remodelling; representation; requirement; research; research theme; research trend; resolution; respect; responsiveness; restriction; restructure; result; rigorous; role; rule;
sample; satisfaction; scale; scientist; scope; score; sentence; set; shortcoming; signature; significance; similarity; simulation; single experiment; size; solution; specificity subtopic; standard measure; standing; step; stock; strategy; structure; study; subject; subject matter; success; superiority; supplement; support; survey; sustainable good performance; system;
tally; target; task; technique; technologist; technology; technological problem; text; theme; theoretical principles; thesis; time; tool; toolkit; topic; training example; training set; transformation; truth; type; typicality;
underlying mechanism; understanding; unnecessity; usability; usage; use; usefulness; user;
validity; value; variability; variance; variety; vein; version; void; volume;
way; weakness; weight; whole research; wide scope; word; work
When the following words are the last word of an N-gram, remove the last word and retain the rest.
approach; case; comparison; complexity; cost; efficiency; example; experiment; insight; measure; mechanism; method; methodology; one; problem research; researcher; sample; score; stage; task; technique; theory; phenomenon; principle; property; quality

Table A2. Stop-words of terms used for writing purposes.