International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2025, Vol. 9, No. 1, IJSKM860

Time-series Keyword Extraction Method and Its Application to Discovery Japanese Key Technology Transition Insights

Fan Cheng [*], Shota Tamaru [†],
Takafumi Nakanishi [*]

Abstract

This paper presents a time-series keyword extraction method and its application to the discovery of key technological transition insights in Japan. In general, it is one of the most important issues in understanding trends in science and technology. To grasp these trends, it will be possible to visualize trends in science and technology from time to time if a method is established to extract important keywords from time to time using the White Paper on Science, Technology, and Innovation published every year by the Japanese government as an example. In this method, words are extracted from text data from time to time, and a new Importance Transition Discovering Score (WITD-Score) is proposed as an index representing the likelihood of the occurrence of each word at that time, following the concept of F-Score. By extracting the transition of keywords from time to time from the change in the WITD-Score for each word, we can extract the transition of keywords from time to time. By implementing this method, we can discover important keywords from time-series text data and visualize the transition of keywords.

Keywords: time-series text data, time-series important word extraction, technology trend extraction, white paper on science, technology and innovation

1 Introduction

In recent years, science and technology have developed rapidly. To develop various strategies, it is important to grasp trends in science and technology correctly. The Japanese government publishes a White Paper on Science, Technology, and Innovation [1] every year.

To understand the trends in science and technology, it is necessary to read them correctly. If we can consolidate these white papers on science, technology, and innovation, and realize a

[*] Tokyo University of Technology, Japan
[†] Musashino University, Tokyo, Japan

method to extract trends in science and technology, it will be useful for formulating various future strategies.

Understanding the trends in science and technology is one of the most important issues. To understand these trends, it is possible to visualize trends in science and technology once a method has been established to extract key keywords by year from the White Paper on Science, Technology, and Innovation published annually by the government.

Text data arranged according to the order of time are called time series text data. We focused on extracting characteristic words for each unit of time from time-series text data. To extract characteristic words for each year, we need not only to calculate the frequency of occurrence, but also to define and calculate an index to calculate the certainty of occurrence for that year.

Although several methods such as TF-IDF, LDA, and SST have been widely used for text analysis, they are limited in their ability to capture the temporal transitions of individual keywords. TF-IDF can measure word importance within a static corpus but cannot reflect year-to-year changes. LDA and other topic modeling methods summarize co-occurring terms within latent topics, which makes it difficult to trace the evolution of specific words over time. SST, while effective for detecting change points in aggregated time-series signals, analyzes transitions at the topic level rather than at the word level. Therefore, there is a need for a method that directly evaluates the temporal dynamics of individual keywords—an issue addressed by the proposed WITD-Score.

This study introduces a time-series keyword extraction method and its application to the discovery of key technology transition insights in Japan.

This method extracts words from text data over time and proposes a new Importance Transition Discovering Score (WITD-Score) [2] [3], which follows the concept of F-Score as a measure of the occurrence possibility of each word at that time. By extracting the transition of keywords by time from the change in the WITD-Score by word, we can extract the transition of keywords by time. By realizing this method, we can discover important keywords by era from text data of time series and visualize the transition of keywords by time.

Furthermore, we applied this method to the actual White Paper on Science, Technology, and Innovation [1] from 2012 to 2022 as the text data of time series, and extracted the words with high WITD-Score in each year as the keywords that Japan focused on in that year.

The WITD-Score [2] [3] of important keywords as the trend of science and technology in Japan was represented by the year in the line graph.

By realizing this method, it is possible to consider the transition of important fields of science and technology in Japan and technology assuming various social events that may occur in the future.

This study makes the following contributions to the broader field of research:

- The keyword extraction method for time-series text data can be applied to extracting important keywords for each time series and observing past trends.

- The method extracts important words for each year using the new WITD-Score, which follows the F-Score concept used in the past to verify search accuracy.
- This method was applied to an actual White Paper on Science, Technology, and Innovation [1] from 2012 to 2022 to clarify specific keywords and trends in science and technology in Japan.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 presents the keyword extraction method for the time-series text data. Section 4 presents the results of the experiments and a discussion. Finally, Section 5 summarizes the proposed method and the evaluation results.

2 Related Works

Reference [4] describes a technique for extracting trends in time-series text data. We propose topic variation detection (TVD) for time-series text data.

In this method, a new WITD-Score based on the concept of F-Score is applied as an index for extracting important words at a specific time point in time-series text data, and the trend of time-series change is extracted from the change in WITD-Score at that time point.

The value of the time-series WITD-Score for each word extracted by this method can also be derived from the change point using SST, as in the method of Reference [4]. To put this into perspective, Reference [4] uses the value of the time-series topic share extracted by LDA, while the method in this study uses the value of the F-score.

Regarding the value of the topic share obtained from LDA, as in [4], it is difficult to observe a specific keyword when a topic consists of multiple word weights. On the other hand, this method calculates the WITD-Score for each word; therefore, if a word with a large WITD-Score can be extracted, important words can be identified and extracted for each word.

To clarify the novelty of the proposed WITD-Score compared with existing keyword extraction methods, we summarize the conceptual differences in Table 1. While TF-IDF evaluates term importance within static corpora and LDA estimates topic distributions across documents, both approaches have limited ability to capture the temporal transition of individual keywords. Similarly, SST focuses on detecting change points in aggregated topic time series, but not on word-level variations. In contrast, the proposed WITD-Score quantifies the temporal importance of each word directly by combining word-wise precision and frequency across years, enabling the visualization of how specific technological terms emerge and decline over time.

Table 1: Comparison of the proposed WITD-Score with conventional methods for time-series keyword extraction.

Method	Type	Temporal Resolution	Level of Analysis	Strength	Limitation
TF-IDF	Frequency-based	None (static)	Word-level	Simple and interpretable	Cannot track changes over time
LDA	Topic modeling	Weak (by topic)	Topic-level	Captures thematic structure	Blurs transitions of individual keywords
SST	Change-point detection	Strong (by time series)	Topic-level	Detects temporal shifts	Requires pre-defined topic signals
WITD-Score (proposed)	F-Score-based metric	Strong (by year)	Word-level	Captures explicit keyword transitions and trends	Requires word-level preprocessing

3 Proposed Method

A keyword transition extraction method for time-series text data is presented. A new WITD-Score [2] [3] is derived based on the concept of the F-Score, which extracts words at each time from text data and represents the occurrence possibility of each word at that time.

The transition of keywords at each time can be extracted by extracting the transition of keywords at each time from the change of WITD-Score at each word. By realizing this method, it is possible to discover important keywords in each period from time-series text data and to visualize the transition of keywords.

3.1 Overview

This section describes the proposed method. Figure 1 shows the system configuration diagram of the proposed method. The method consists of four functions: word extraction, word frequency count, WITD-score calculation, and visualization.

The method uses a white paper database on science, technology, and innovation. This database contains text data for each time series.

The time-series text data were read from the white paper database of science, technology, and innovation, and words were extracted from the time-series text data using the word extraction function. The word frequency count function calculates the occurrence frequency of words for each year, the WITD-Score calculation function calculates the WITD-Score for each word for each year, and the result is aggregated by the visualization function.

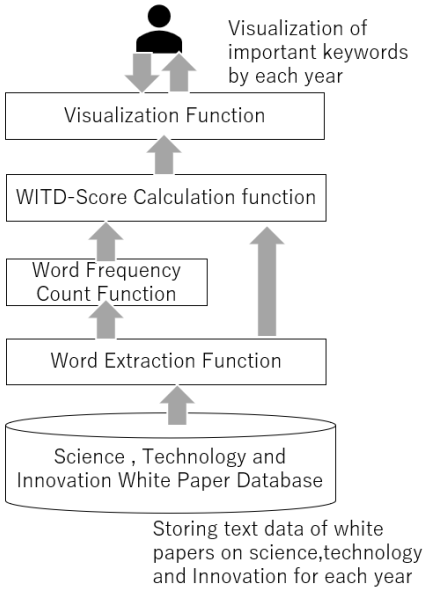

Figure 1: Overview of the proposed method. Our method consists of four functions: word extraction, word frequency count, WITD-score calculation, and visualization.

3.2 Word Extraction Function

This section describes the word extraction function. The word extraction function reads time-series text data from a database and extracts words.

For morphological analysis, we used MeCab (ver. 0.996) with the default IPA dictionary setting to tokenize and lemmatize Japanese text data. As part of text preprocessing, we removed stopwords using the open-source stopword library slothlib and further excluded numerical expressions such as years (e.g., “2012,” “2020”), non-independent functional words, and common suffixes that do not carry semantic content. All retained words were normalized to their base forms, and only nouns, verbs, and adjectives were used in subsequent analyses. This preprocessing ensures that the extracted words accurately represent meaningful linguistic units while minimizing noise from grammatical artifacts.

For reproducibility, the preprocessing script implementing these steps has been made available in our
public GitHub repository at:
https://github.com/[your-repo-name]/witdscore-preprocessing
(Access information will be provided upon publication if required by journal policy.)

To work with Japanese text data, the function consists of modules that separate sentences into words, convert words into their original form, and define and remove frequently occurring but less meaningful words as stop words.

Time-series text data are defined as text data linked to time at regular intervals. For example, a government white paper is typically published at the same time each year.

These data are text data at regular time intervals and can be considered a type of time-series text data. In addition, text posts for social networking services can be considered time-series text data.

The input to the word-extraction function is assumed to be time-series text data, as described

above. In particular, to realize the extraction of science and technology trends, this study assumes a series of data as input, including actual science and technology innovation white paper data from 2012 to 2022.

This function of this paper reads text data from 2012 to 2022 from the Science and Technology Innovation White Paper database and extracts nouns, verbs, and adjectives that represent features especially related to science and technology. The parts of speech to be extracted may vary depending on the type of time-series data to be extracted. That is, the function extracts word groups predicted to represent the characteristics of each period from each time series of text data.

3.3 Word Frequency Count Function

This section describes the word–frequency count function. The Word Frequency Count function counts the number of words extracted using the word extraction function described in Section 3.2.

Note that the word frequency was derived for each time series. This feature must pass the number of words per time series to the WITD-score calculator, as shown in Section 3.4. Because this feature applies to the Science, Technology, and Innovation White Paper for each year in this article, we derived the number of occurrences of each word for each year.

3.4 WITD-Score Calculation Function

This section describes the WITD-Score function. The WITD-Score calculation function derives the WITD-Score [2] [3] of each word in each time series from the occurrence frequency of each word in each time series extracted in section 3.3. Table 2 summarizes the methods used in this function.

Table 2: Overview of data that can be used for important word extraction methods (Confusion Matrix)

	Occurrence of word W_a	Other than the word W_a Occurrence of the word
Target year t_i	Number of times t_i a particular specific word W_a appears in the target year t_i —(1)	Number of occurrences of other than the word W_a in the target year t_i —(2)
Other than the target year t_i	Number of times a particular specific word W_a appears outside the target year t_i —(3)	Number of times the word other than W_a appears except in the target year t_i —(4)

Table 1 lists items that can be calculated by focusing on the number of occurrences of a particular word in a given year. The term frequency–inverse document frequency (TF-IDF) method was used to derive word importance. This method measures the importance of specific words in all sentences. In other words, the elements used were derived using only (1) and (2) in Table 1. However, because this study focuses on the relative word-occurrence relationship in each year, it is necessary to consider the importance of words in each year as well, and TF-IDF is insufficient for this purpose. When this table is viewed as a mixed matrix, the concept of the F-score is formed. In this study, we propose a new method for extracting important words, the Word Importance

Score (WI-Score), based on the F-Score, which, similar to the F-Score, calculates the fit and recall rates and takes the harmonic mean of the two values. This makes it possible to perform calculations using elements (1), (2), and (3) in Table 1, and to obtain vector data that takes into account the number of occurrences of specific words in all years, rather than only for sentences in one year.

Reference [3] shows how the F-Score value can be used to evaluate the certainty of a word occurrence for each category.Reference [3] introduces the WITD-Score as a new measure of the certainty of a word occurrence for each time series.

The specific WI-Score (Word Importance Score) is calculated as follows：

First, the precision function ($prec$) was calculated. The precision function calculates the ratio of the number of times a particular word appears to the number of times a particular word does not appear in a particular year. In other words, the probability of a particular word appearing is calculated from the entire word.

The word $w_i \in W$ and the time $t_j \in T$ obtained by the word extraction function are defined respectively. Note that time is a set and not a series. For example, the White Paper on Science, Technology, and Innovation has been published annually. Therefore, the time set T is $T = \{2012, 2013, \dots, 2022\}$.

The precision function $prec(w_i, t_j)$of the word w_i for each time t_j is defined as follows.

$$prec(w_i, t_j) = \frac{\#(w_i, t_j)}{\sum_{j=1}^{M} \#(w_i, t_j)}$$

In the formula, $\#(w_i, t_j)$ refers to the number of occurrences of the word w_i at time t_j.

Next, the frequency function ($freq$) is calculated. The frequency function calculates the ratio of the number of occurrences of a particular word in a particular year to that in other years. In other words, the probability of the occurrence of a particular word is calculated for the entire year.

The frequency of the word w_i at time t_j is defined as follows.

$$freq(w_i, t_j) = \frac{\#(w_i, t_j)}{\sum_{i=1}^{N} \#(w_i, t_j)}$$

Finally, the WITD-Score is obtained using the above two functions, $prec$ and $freq$. In other words, the probability of the occurrence of a specific word in a specific year is calculated from the whole year and the whole word by combining the two functions, Precision and Frequency, and taking the harmonic mean.

The formula for the WITD-score is as follows:

$$WITD - Score(w_i, t_j) = 2\frac{prec(w_i, t_j) freq(w_i, t_j)}{prec(w_i, t_j) + freq(w_i, t_j)}$$

We can derive the WITD-Score at each time t_j and each word w_i, and objectively calculate at which time each word became important. We consider the word w_i with the highest WITD-Score at each time $t_j$as having high importance at that time, and extract the top k words at each time t_j. Furthermore, we can see the evolution of word importance over time by looking at the change in the WITD-Score of each word w_i over time.

The novelty of the proposed method lies in its ability to capture the temporal transition of individual keywords from time-series text data, which has not been effectively achieved by conventional methods such as TF-IDF, LDA, or SST. Traditional approaches like TF-IDF evaluate word importance within a static corpus and cannot reflect how the relevance of a keyword changes over time. Topic modeling techniques such as LDA represent documents as mixtures of

topics, but they aggregate word importance within topics, making it difficult to trace the temporal evolution of specific terms. Similarly, time-series analysis methods such as SST focus on detecting change points in aggregated topic signals rather than identifying transitions at the word level.

In contrast, the proposed WITD-Score quantifies the importance of each word for every time unit by combining precision (representing word specificity within a year) and frequency (representing its relative prominence across years). This F-score-based formulation enables the direct visualization of how specific technological terms emerge, peak, and decline over time. Consequently, WITD-Score bridges the gap between static keyword extraction and topic-level time-series analysis, providing a fine-grained and interpretable measure of temporal keyword importance.

To further clarify the novelty and practical advantages of the proposed method, Table 1 presents a comparative summary of WITD-Score and conventional approaches such as TF-IDF, LDA, and SST. This table outlines their methodological characteristics, temporal resolution, levels of analysis, and respective strengths and weaknesses in extracting keywords from time-series text data. As shown, WITD-Score uniquely achieves fine-grained tracking of keyword-level temporal transitions, providing interpretability and analytical depth that conventional methods cannot capture.

3.5 Visualization Function

The visualization function is described in the following section. The visualization function aggregates and visualizes the WITD-Score of each word in each time-series output using the functions shown in Section 3.4. This function consists of a module that displays words with a high WITD-Score for each time series and a module that represents the change in the WITD-Score of each time series of words entered by the user as a line graph.
The former module displays keywords that describe the characteristics of each time-series.
Specifically, the module sorts the words in descending order of the F-score values of each time series derived in Section 3.4 above and outputs the top k words.
This module enables the user to know the key keywords of each time series.
The latter module takes a word of interest from the user as the input and displays the F-score value of each time series of the word as a line graph. Using this module, the user can see the change in the importance of the keyword of interest. In other words, the change in importance can be seen as a change in the trend.

4 Experiment

This section describes the experimental environment, experimental contents, considerations, and conclusions. The method proposed in Section 3 was implemented as an experimental system and the results were confirmed.

4.1 Experimental Environment

The experimental system was implemented in Google Collaboratory [7]. The White Paper on Science, Technology, and Innovation [1] from 2012 to 2022 was used as time-series text data. The word extraction function presented in Section 3.2 was applied to the Japanese morphological

analyzer MeCab [8]. The WITD-Score was derived using the library Scattertext [3] [9], which visualizes word features in each category. k=10 for the ability to display words with high WITD-Score, as described in Section 3.5.

4.2 Experiment 1 (Extraction of important words by year)

We derived the top 10 words with the highest WITD-Score for each year using this experimental system. These tables represent the keywords for each year mentioned in the White Paper on Science, Technology, and Innovation, from 2012 to 2022. Owing to space limitations, only the most characteristic years are presented here.

Table 3 shows the top ten words with the highest WITD scores in 2012. As a result, words such as "Earthquake Disaster" and "Fukushima Daiichi Nuclear Power Plant" were used. In fact, in 2011, the Tohoku Earthquake occurred in Japan and caused great damage. The White Paper on Science, Technology, and Innovation also appears to have drawn attention to the Great East Japan Earthquake.

These keywords reflect the strong focus on recovery and reconstruction following the Great East Japan Earthquake of 2011. The frequent appearance of terms such as “Fukushima Daiichi Nuclear Power Plant,” “Earthquake Disaster,” and “Tohoku Region” indicates that the Japanese government’s 2012 science and technology policy was heavily centered on disaster response and energy safety.

Table 3: Important word extraction result in 2012

	word	WITD-Score
1	Earthquake Disaster	1.000000
2	Fukushima Daiichi Nuclear Power Plant	0.999240
3	Accident	0.991385
4	After Earthquake Disaster	0.987909
5	Pacific Ocean Okinawa Earthquake	0.973909
6	Tohoku Earthquake	0.970153
7	Nuclear Power Plant	0.964601
8	Specialist	0.961831
9	Tohoku Region	0.960864
10	Earthquake	0.959101

Table 4 shows the top ten words with the highest WITD scores in 2016. As a result, words such as " Smart,” "Artificial Intelligence,” "IoT " are shown. 7th place "Learning" and 8th place "Deep", the word "Deep Learning" was separated during word extraction. 7th "Learning" is actually the English word learning, while "Learning" in 9th place is actually the Japanese word for learning. In 2016, artificial intelligence and big data also attracted attention in Japan.

The extracted keywords in 2016, including “Artificial Intelligence,” “Big Data,” and “IoT,” mark the rise of Japan’s national initiative toward Society 5.0. This year corresponds to the early stage of AI-driven industrial transformation, emphasizing data utilization and smart systems.

Table 4: Important word extraction result in 2016

	word	WITD-Score
1	Smart	1.000000
2	Computer	0.984364
3	Nobel Prize	0.983089
4	Artificial Intelligence	0.977337
5	Big Data	0.965898
6	Internet	0.964317
7	Learning	0.957074
8	Deep	0.956921
9	Learning(学習)	0.955530
10	IoT	0.949788

Table 5 shows the top ten words with the highest WITD scores in 2017. As a result, words such as " Open Innovation ", " venture ", and Start a Business were shown. The 1st place "Ohsumi" is Professor Yoshinori Ohsumi, who won the Nobel Prize in Physiology or Medicine in 2016. In 2017, the government focused on speeding up the use of research results by the private sector, such as startups and ventures.

The dominance of the term “SDGs” in 2018 indicates a transition toward sustainability-oriented policies and research agendas. The frequent occurrence of related institutional terms such as “National Research and Development Agency” highlights Japan’s efforts to align science and technology with global sustainable development goals.

Table 5: Important word extraction result in 2017

	word	WITD-Score
1	Ohsumi	1.000000
2	Autophagy	0.999383
3	Open Innovation	0.998125
4	Venture	0.994522
5	Start-up Company	0.994390
6	URA	0.985749
7	Start a Business	0.979343
8	Industry-University-Government Collaboration	0.978702
9	Collaborative Research	0.943287
10	Intellectual Property	0.932123

Table 6 shows the top 10 words with the highest WITD-Score in 2018. As a result, words such as "SDGs" etc. are shown. Since 2018, the importance of SDGs in science and technology has begun to draw attention.

Keywords such as “CO2,” “Future Society,” and “COVID-19” in 2020 reflect a dual focus on climate change mitigation and the global pandemic response. The simultaneous rise of “AI” and “5G” underscores continued attention to digital innovation amid the social challenges of the COVID-19 era.

Table 6: Important word extraction result in 2018

	word	WITD-Score
1	SDGs	1.000000
2	National Research and Development Agency	0.990052
3	Completion (of a course)	0.953716
4	Paper	0.948984
5	Ph.D	0.936635
6	Sector	0.930695
7	Society	0.913129
8	Innovation Conference	0.912509
9	Integrated Science and Technology	0.912509
10	Japan Agency for Medical Research and Development	0.909124

Table 7 shows the top ten words with the highest WITD scores in 2020. As a result, words such as "CO2," "Future Society," "Future Prediction," "Automatic Operation," "AI," "COVID-19," "5G," and "Year 2040" are shown. In 2019, the government emphasized basic research. In 2020, words related to environmental issues such as CO2, artificial intelligence, and future-oriented words for 2040 appeared. We can also see the keyword COVID-19, which was trending in 2019.

Keywords such as "CO2," "Future Society," and "COVID-19" in 2020 reflect a dual focus on climate change mitigation and the global pandemic response. The simultaneous rise of "AI" and "5G" underscores continued attention to digital innovation amid the social challenges of the COVID-19 era.

Table 7: Important word extraction result in 2020

	word	WITD-Score
1	CO2	1.000000
2	Future Society	0.987886
3	Future Prediction	0.986291
4	Automatic Operation	0.957643
5	AI	0.939378
6	COVID-19	0.938746
7	5G	0.927994
8	Industrial Revolution	0.923076
9	Novel Coronavirus	0.921011
10	Year 2040	0.920218

Table 8 shows the top ten words with the highest WITD scores in 2021. As a result, words such as "COVID-19", "AI", and "Novel Coronavirus" are shown. In 2021, the COVID-19 pandemic will 2020. However, there are also some future-oriented keywords, such as AI.

The 2021 results show a continued emphasis on pandemic-related issues such as "COVID-19," "Novel Coronavirus," and "Infection," reflecting Japan's sustained policy focus on public health and pandemic control. At the same time, the appearance of "AI," "Material," and "Virtual

Space" suggests an increasing interest in leveraging digital and advanced materials technologies to support social resilience and remote innovation environments. Furthermore, terms such as "Year 2050" and "Carbon Dioxide" indicate a renewed alignment of science and technology policies with long-term sustainability and decarbonization goals. Overall, these keywords illustrate a transitional phase in which Japan's science and technology policy shifted from immediate crisis management toward future-oriented recovery and sustainability.

Table 8: Important word extraction result in 2021

	word	WITD-Score
1	COVID-19	1.000000
2	AI	0.987332
3	Novel Coronavirus	0.979427
4	Material	0.979203
5	Society	0.971423
6	Social Science	0.970056
7	Infection	0.968031
8	Carbon Dioxide	0.962172
9	Year 2050	0.959999
10	Virtual Space	0.956875

Table 9 shows the top ten words with the highest WITD scores in 2022. As a result, words such as "Startup," "Moonshot," "DX," "Smart City" and "COVID-19" are shown. "DX" stands for Digital Transformation, and it became a hot topic in Japan. "Moonshot" is a moonshot plan launched by the Cabinet Office of Japan, which clearly laid out the science and technology to be realized by 2050. The keyword "Covid-19" is also at the top, indicating that the Japanese government is strongly promoting future orientation.

It is possible to find trends in science and technology that were noticed each year using important keywords derived from these experimental systems.

In 2022, keywords like "Startup," "Moonshot," and "DX (Digital Transformation)" represent Japan's strategic emphasis on next-generation innovation programs and digital ecosystems. The recurring presence of "COVID-19" indicates a continued policy concern with post-pandemic recovery and technological preparedness for future crises.

Table 9: Important word extraction result in 2022

	word	WITD-Score
1	Startup	1.000000
2	Moonshot	0.987125
3	Digital	0.986492
4	DX	0.978899
5	Research University	0.977945
6	Smart City	0.969847
7	Year 2050	0.961417
8	COVID-19	0.952568
9	Ecosystem	0.948897
10	Social Science	0.948825

4.3 Experiment 2 (Visualization of technological trend changing)

Figure 2 visualizes the trend change due to the change in WITD-Score using "Big Data," "IoT" and "AI" as examples. The result of the line graph is shown when the keywords "Big Data," "IoT" and "AI" start to attract attention and decline. IoT appeared in 2016 and converged in 2021 and 2022. Big data appeared in 2016, decreased in 2018, and then increased again in 2019. AI has been a hot topic since 2018, indicating that these keywords are not always hot.

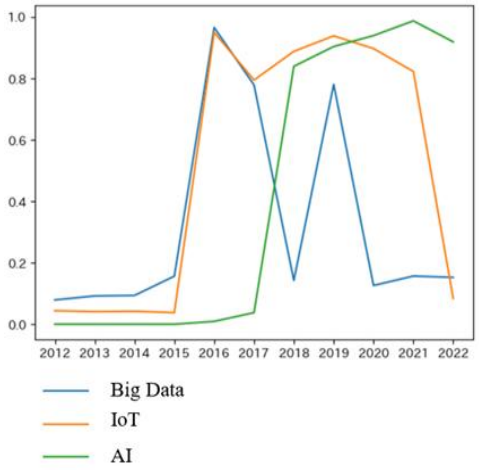

Figure 2: Experimental result in the case of words “Big Data", "IoT" and "AI".The x-axis represents the publication year (2012–2022), and the y-axis shows the corresponding WITD-Score values indicating the importance level of each keyword at that time.

5 Discussion

In Experiment 1, presented in Section 4.2, a list of words with a high WITD-score was presented by year. Consequently, it was possible to obtain a clear list of words that received annual attention in science and technology.

In Experiment 2 of Section 4.3, an example of visualizing trend changes due to changes in the WITD-Score is shown. It was possible to observe the period in which words related to science and technology were observed.

The results show that this method can visualize the tendency of time-series text data to change over time.

6 Conclusion

This paper describes a time-series keyword extraction method and its application to the discovery of key technological transition insights in Japan. Understanding trends in science and technology is one of the most important issues. To capture these trends, it is possible to visualize trends in science and technology if a method is established to extract key keywords by year from the White Paper on Science, Technology, and Innovation published by the government every year.

We proposed a new Importance Transition Discovering Score (WITD-Score), which follows the concept of F-Score as an indicator of the likelihood of occurrence of each word in time, and it was possible to calculate the likelihood of occurrence of a specific word in a specific year from the whole year and the whole word.

We implemented an experimental system and showed some results when it was applied as time-series text data in the White Paper on Science, Technology, and Innovation [1] from 2012 to 2022. By implementing this method, we were able to observe the technological trends in Japan.

Beyond the field of science and technology policy, the proposed WITD-Score method can be widely applied to other domains where understanding temporal dynamics in textual data is essential. For example, in economics, it can help identify transitions in policy priorities by analyzing annual government economic reports. In politics, it can reveal the evolution of key agenda topics across parliamentary records or election manifestos. Similarly, in healthcare, it can trace how medical and public health issues emerge and decline over time based on official documents or research abstracts. These examples demonstrate that the proposed method is not limited to a single domain but provides a general framework for visualizing and interpreting time-dependent keyword trends across a variety of contexts.

In the future, we will develop visualization applications for time-series text datasets. We will also apply it to other time-series text datasets, such as text data from networking services. We also apply it to various fields of study, such as political science and economics, where we need to show time-series trends.

By integrating these refinements in method comparison, data preprocessing transparency, and visualization clarity, the present study provides a more robust and reproducible foundation for

time-series keyword analysis. We expect that this approach will not only contribute to technology policy research in Japan but also serve as a practical analytical framework for identifying emerging trends and supporting evidence-based decision-making across diverse policy and research domains.

References

[1] Science, Technology, and Innovation white paper in Japan https://whitepaper-search.nistep.go.jp/white-paper/list

[2] A. A. Taha, A. Hanbury, “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,” BMC medical imaging, 15(1), pp.1-28, 2015.

[3] ScatterText package https://github.com/JasonKessler/scattertext#understanding-scaled-f-score

[4] S. Kato, T. Nakanishi, B. Ahsan, H. Shimauchi, Time-series topic analysis using singular spectrum transformation for detecting political business cycles, Journal of Cloud Computing: Advances, Systems and Applications, 10, 21, 2021.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, 3(Jan), pp.993-1022, 2003.

[6] T. Idé, K. Inoue, "Knowledge discovery from heterogeneous dynamic systems using change-point correlations," In Proceedings of the 2005 SIAM international conference on data mining, pp. 571-575, 2005.

[7] Google Colaboratory. https://colab.research.google.com/

[8] T. Kudo, Mecab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/, 2005.

[9] J. S. Kessler, "Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ," ACL System Demonstrations. 2017.