Identifying Potential Sleeping Beauties Based on Dynamic Time Warping Algorithm

Zewen Hu [1, *], Yu Chen[1], Jingjing Cui[1]

[1] School of Management Science and Engineering, Nanjing University of Information Science and Technology, 219 Ningliu Road, Pukou District, Nanjing 210044, Jiangsu Province, China

Abstract

Sleeping beauty is recognized as delayed highly-cited or high-impact literature. Precise and efficient identification of potential sleeping beauties from massive literature can maximize their value in science and technology development. Therefore, in this study, a new time series similarity method, named dynamic time warping (DTW) algorithm, is designed to precisely and efficiently identify sleeping beauties from massive literature. First, top 1% of the highly cited papers (5425 articles) between 1990 and 2010 in the field of artificial intelligence were identified based on data collected from the Web of Science database. Then, the DTW algorithm was designed and implemented to identify potential sleeping beauties based on the citation curve of a classic sleeping beauty. Among the findings: (1) The DTW algorithm can quickly and effectively identify potential sleeping beauties with help from the citation trajectory of benchmark sleeping beauties, thereby matching the high recognition accuracy of curve fitting and high recognition efficiency of the objective indicator method. (2) The DTW method displayed strong robustness, automatically and accurately identifying different kinds of highly influential publications including sleeping beauties, Nobel Prize papers, highly-cited papers, and hot papers, based on publication citation trajectories.

Keywords

Sleeping beauty, Citation curve, Identification and prediction, Dynamic time warping algorithm, Quadratic function fitting

1 Introduction

In the science world, scientific and technical manuscripts have a finite lifetime after their publication, Although some publications are widely cited and reach a citation peak during an initial short citation window after publication, the rate and number of citations dwindle to an extremely low level after a certain period [1]. Considering the enormous number of publications from different disciplines, most manuscripts ate either uncited or citation level is low after a long citation window following publication [2, 3, 4]. Undeniably, there have also been a few popular or high-quality publications that steadily accumulate citations during a long citation lifetime and finally become highly cited papers [5]. Except for the common low-citation, non-citation, and high-citation phenomena, some scholars also found another kind of sleeping beauty phenomenon that merges low- and high-citation patterns. A sleeping beauty (SB) in science refers to a publication, the importance and relevance of which have not been recognized, whereby the manuscript does not receive much attention during the initial citation window following its publication, and unexpectedly starts being frequently cited followed by a sudden spike of popularity [1, 6]. Depending on the rate and number of citations since publication, scientific and technical manuscripts can be classified into three: publications with low-citation or non-citation, highly-cited or hot publications, and sleeping beauties which refer to publications from the 1960s receiving delayed recognition. Barber discovered the phenomenon of delayed recognition upon noticing that some major scientific discoveries in the scientific community were not widely used and cited when they were first published [7]. However, many years after their publication, their scientific value and significance began to attract the attention of researchers, whereupon they were widely utilized and cited. In 1980, Garfield introduced the theory of “delayed recognition,” which refers to manuscripts lacking attention at the time of publication, suddenly being highly cited after a certain period of time [8]. Some classical sleeping beauties include the Einstein, Podolsky, and Rosen “paradox” paper which was used as the primary sleeping beauty in this study.

Identifying "sleeping beauties" from a massive number of papers and recommending them to the scientific world would enable their full recognition in terms of scientific and technological value, thereby driving the development of science and technology [9]. Considering the delayed recognition phenomenon of "sleeping beauties," the rapid and efficient identification of SBs through various methods and models has become more significant in ensuring their full utilization potential. Most academic databases or platforms such as Web of Science and Scopus, have implemented the recommendation function of highly cited or hot papers. However, a recommendation function has not been designed for sleeping beauties or other outstanding publications. Therefore, highly efficient methods or algorithms for

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023), June 26 - 30, 2023, Santa Fe, New Mexico, USA and online
*Corresponding author.
EMALL:huzewen915@163.com (Z. Hu); 202212240023@nuist.edu.cn(Y. Chen); 20201224004@nuist.edu.cn(J.Cui)

identifying and recommending "sleeping beauties" would have significant application value.

Since Van Raan proposed the concept of sleeping beauty [6], a series of quantitative studies on the identification and application of sleeping beauties were implemented and published [10, 11, 12, 13]. These quantitative studies can be categorized into two types; sleeping beauty identification based on curve fitting methods and those based on indicators.

(1) Identifying sleeping beauties through curve fitting. The curve fitting method entails the selection of the appropriate curve type, use of a mathematical expression or model to fit the annual citation frequency of each document, and classification of the manuscript type based on the document's citation curve. Curve fitting provides the advantages of simple operation, intuitive results, and easy analysis. However, when the number of documents is too large, the fitting efficiency is extremely low.

(2) Identifying sleeping beauties by indicator-based methods. Including subjective indicators and objective indicators. The subjective indicator method can also be referred to as artificial parameter setting. Van Raan [6] first proposed a definition for sleeping beauty and three subjective indicators to identify SBs. The three indicators used to characterize sleeping beauties were: a sleeping period of greater than or equal to five years, a sleeping period of less than or equal to two years, and a state of awakening of greater than 20 years. However, up to now, there has been no unified standard on the setting of values of subjective indicators including the length and depth of sleeping and awakening intensity, which is variable and greatly influenced by interference factors and subjective perception of scholars.

The objective indicator method is different from artificial parameter setting. In 2015, Ke et al. [1] designed a no-parameter indicator "Beauty Coefficient (B)" based on the citation frequency of the literature to identify sleeping beauties and quantitatively analyze the distribution of the number of SBs in different disciplines. A sleeping beauty can be quickly identified through the Beauty Coefficient, which does not reflect citation after the citation peak. Ye and Bornmann improved the Beauty Coefficient by introducing a dynamic citation angle β to quantitatively identify sleeping beauties [14]. Li improved the Gini index in the field of economics to identify sleeping beauties [15]. But, the objective indicator method has the problem of ignoring the specific citation curve of sleeping beauties and may be influenced by parameters such as the length and depth of sleeping and length of citation period.

In this study, we designed and implemented the dynamic time warping (DTW) algorithm, which is a much more robust distance measure for time series, allowing similar shapes to match even if they are out of phase along the temporal axis. Because of this flexibility, DTW is widely used in science, medicine, industry, and finance [16, 17, 18, 19]. In bioinformatics and chemical engineering, DTW algorithms have been successfully applied to RNA expression data, synchronization, and monitoring of batch processes in polymerization [20, 21]. The DTW algorithm is different from the traditional sleeping beauty identification method in that it measures the distance between time series curves [22], to identify documents with similar citation trajectories in the same or different citation periods. This method not only considers the citation curve of a document's entire lifetime, but also measures a specific DTW-value, and combines the advantages of the curve fitting and objective indicator methods, thereby displaying high robustness. This method can identify potential sleeping beauty citation curves exhibiting slow citation rate with flat to fast growth, as well as those with large annual citation frequency fluctuations. "benchmarking sleeping beauty" refers to a standardized sleeping beauty satisfying the three aforementioned indicators proposed by van Raan [6].

The innovations of this study are primarily reflected in the following aspects. First, the DTW algorithm was applied to identify potential sleeping beauties based on any given benchmarking sleeping beauty citation curve. Second, the accuracy of the DTW algorithm was improved by combining it with the three indicators proposed by Van Raan [6] to identify sleeping beauties with a standardized sleeping period and not some highly cited papers. Finally, the identification results of DTW method were compared with those of the quadratic function fitting methods.

2. Materials and Methods

2.1 Data

The research uses papers and citation data from the "Computer Science, Artificial Intelligence" categories in Web of Science as research samples to evaluate the performance and application value of the DTW algorithm. The data from the Web of Science were restricted to three document types: articles, proceeding papers, and reviews. We set the publication period from 1990 to 2010 and the citation period from1990 to 2019, making the shortest and longest citation periods 10 and 30 years, respectively. A total of 524,599 papers with corresponding citation and bibliometric data were collected. By selecting different lengths of citation periods, the effect of the length of citation period on the identification of sleeping beauties can be verified using the DTW algorithm, to further suggest robust DTW algorithms.

Although sleeping beauty combines low and high citation, SBs belong to highly cited publications to some extent as there is a higher percentage of sleeping beauties among highly cited publications. The top 1% is a commonly used criterion for identifying highly cited publications, as defined by Essential Science Indicators from Clarivate Analytics. Therefore, the top 1% highly cited articles were selected as samples to identify potential sleeping beauties. Table 1 shows the number of highly cited as well as total number of papers in the field of artificial intelligence during the indicated 21 years.

2.2. Methodology

Dynamic time warping (DTW) is a dynamic programming method that combines time warping with distance measurement. The basic idea is to find the smallest alignment matching path to minimize the distance between two sequences. The DTW algorithm can calculate the distance between sequences of different lengths and thus is not sensitive to time series offset. Therefore, the DTW algorithm can quickly identify potential sleeping beauty documents that conform to the "benchmarking sleeping beauty" citation curves from massive documents.

Given two time series of citations for two papers: Q and C, with series length n and m, respectively, where

$$Q=q_1, q_2, ..., q_i, ..., q_n \tag{1}$$

$$C=c_1, c_2, ..., c_j, ..., c_m \tag{2}$$

an n*m matrix is constructed to align the two sequences, whereby the (i^{th}, j^{th}) element of the matrix contains the distance between the two points q_i and c_j. $A_{n\times m} = (d(q_i, c_j))_{n\times m}$ denotes the distance matrix of Q and C; $d(q_i, c_j) = (q_i - c_j)^2$ represents the distance between the corresponding points of the two sequences, This paper considers the Euclidean distance as an example.

Let W be a contiguous set of matrix elements defining a mapping between Q and C, then we have

$$W = \omega_1, \omega_2, \ldots, \omega_k, \ldots, w_K \quad \max(n, m) \le K < n + m - 1 \tag{3}$$

$\omega_k = (d(q_i, c_j))_k$ is the k^{th} element of the path, and the path needs to fulfill the following conditions:

① Boundary: $\omega_1 = (1,1)$, $\omega_K = (n, m)$.This constrains the starting and finishing points in diagonally opposite corner cells of the matrix.

② Continuity: Given $\omega_k = (a, b)$, $\omega_{K-1} = (a', b')$, where $a - a' \le 1$ and $b - b' \le 1$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).

③ Monotonicity: Given$\omega_k = (a, b), \omega_{K-1} = (a', b')$, where $a - a' > 0$ and $b - b'$ >0. This forces the points in W to be monotonic in time.

There are many wrapping paths that satisfy the above conditions. Among all paths, the path that minimizes the value of formula (4) is called the optimal path, and the corresponding distance is the dynamic time-bending distance, DTW (Q, C).

$$DTW(Q, C) = \min\left\{\sqrt{\sum_{k=1}^{K} W_K}\right. \tag{4}$$

The calculation of DTW is a dynamic programming process, where $\gamma(i, j)$ denotes the cumulative distance defined as the sum of distance in the current state (q_i, c_j) and the minimum value of the current cumulative distance.

$$\gamma(i, j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \tag{5}$$

The algorithmic complexity of DTW is O(nm)

Table 1

The number of highly-cited documents in the field of artificial intelligence from 1990 to 201

publication year	Total number	Number of highly cited documents	publication year	Total number	Number of highly cited documents	publication year	Total number	Number of highly cited documents
1990	928	56	1997	11702	217	2004	39395	373
1991	1696	67	1998	14515	232	2005	46030	385
1992	3015	96	1999	11200	218	2006	61109	315
1993	6216	91	2000	16734	281	2007	51350	337
1994	11621	112	2001	22083	307	2008	49142	333
1995	11211	140	2002	33277	320	2009	53259	414
1996	9007	179	2003	34465	374	2010	36644	398

3. Exploratory experiment for sleeping beauty identification based on DTW

Considering some standardized sleeping beauties as samples, we tested and evaluated the applicability of the DTW algorithm. Li identified 10 sleeping beauties that met Raan 's criteria from 21,438 papers published by Nobel Laureates in chemistry, physics, physiology, and medicine, and the publication window was 1901-2012 [23]. We retrieved these 10 sleeping beauties through Web of Science, normalized the annual citation frequency, and formulated them as time series of annual citation points, S1 to S10. These 10 sleeping beauties were then used in preliminary tests to explore the potential of the DTW algorithm in identifying potential sleeping beauties.

First, we chose the time series of the second sleeping beauty (document 2) as the benchmarking standard because it is characterized by a short citation time and a sharp increase in the citation curve. Then, we calculated the DTW-value between the second benchmarking sleeping beauty and the other nine sleeping beauties.

The smaller the DTW-value between the citation frequency distribution (referred to as citation time series) of the two documents, the higher is the similarity between the citation curves of the two documents. Table 2 presents the DTW-value between the second benchmarking sleeping beauty and the others (abbreviated as DTW-value 1). In Table 2, the citation time series of documents 1, 2, and 3 are the closest, and the DTW distance is less than 0.19. The distances between the cited sequences of the sleeping beauty documents 4-10 and 2 are between 0.28 and 0.56. To verify the effectiveness of using DTW-value to identify potential sleeping beauties, we also established a comparative baseline by selecting top 5 highly cited papers that were not sleeping beauties in the AI field, to calculate the average of DTW-values (abbreviated as DTW-value 2) between each of the five highly cited papers and 10 sleeping beauties, as shown in Table 2. An obvious fact is that DTW-value 2 of 10 sleeping beauties is far higher than DTW-value 1, suggesting the effectiveness of DTW methods in identifying potential sleeping beauties based on the benchmarking sleeping beauty.

Table 2

Basic information of 10 sleeping beauties

Document id	author	source title	publication year	total cited frequency	DTW-value 1	DTW-value 2
2	Shockley, W; Queisser, HJ	JOURNAL OF APPLIED	1961	7608	0.000	0.601

		PHYSICS				
1	Langmuir, Irving	JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	1918	14586	0.156	0.529
3	Einstein, A	ANNALEN DER PHYSIK	1905	5480	0.187	0.480
8	Heisenberg, W.; Euler, H.	ZEITSCHRIFT FUR PHYSIK	1936	1972	0.289	0.622
4	Einstein, A	ANNALEN DER PHYSIK	1906	3746	0.333	0.706
9	Purcell, EM; Torrey, HC; Pound, RV	PHYSICAL REVIEW	1946	1588	0.364	0.478
5	Feynman, RP	PHYSICAL REVIEW	1939	2843	0.408	0.584
6	Schroedinger, E.	NATURWISSENSCHAFTEN	1935	2164	0.416	0.523
7	Feynman, RP; Vernon, FL	ANNALS OF PHYSICS	1963	1990	0.523	0.543
10	Staudinger, H; Meyer, J	HELVETICA CHIMICA ACTA	1919	1483	0.562	0.664

4. Verification of DTW method in identifying potential sleeping beauties

A sleeping beauty comprises two different citation periods: sleep and awakening. Therefore, its citation curve has a standard citation distribution and is thus affected less by disciplinary factors, that is, the citation curves of sleeping beauties from different disciplines display a high similarity. Research has shown that the citation characteristics in the field of computer science conform to the first sleep phase and then suddenly enter the awakening phase, which is consistent with the citation curves of most existing research in the field of physics [5]. Therefore, we chose a benchmarking sleeping beauty with the oldest publication year, the longest citation period, and slower citation rate [15]. Subsequently, we measured the DTW-value between the "benchmarking sleeping beauty" and 5245 highly cited papers in the field of artificial intelligence from 1990 to 2010 after normalizing the annual citation frequency curves of all the papers. The seminal study on Brownian motion published by Einstein in 1905 presents a standardized sleeping beauty citation distribution curve with a lengthy sleeping period and a long citation burst after awakening. As shown in Figure 1, this paper is gradually attracting attention after 65 years of publication, and the citation frequency only exceeded 20 in the 66th year after publication. As of 2020, the citation curve of this article had not yet reached its citation peak.

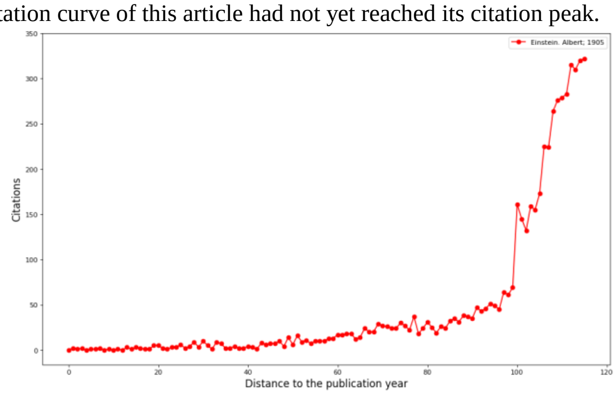

Figure 1: The annual citation frequency curve of a sleeping beauty authored by Einstein in 1905

Using the DTW algorithm, this paper measures the DTW-value between 5245 highly-cited documents in the field of artificial intelligence from 1990 to 2010 and the "benchmarking sleeping beauty" authored by Einstein in 1905. Table 3 calculated and presented the descriptive statistics of the DTW-value of the documents in each publication year. For recent publications, the citation period is shorter; the various statistics of the DTW-value are smaller, such as the average DTW distance of 0.64 in 2010, whereas it was 0.97 in 2001 and 1.35 in 1991. Documents published in 1990 had a long citation duration of 30 years. However, sleeping beauty documents with lower average DTW-value in Table 3 correspond to manuscripts published between 2002 and 2010 with shorter citation durations. This suggests that in identifying sleeping beauties, the DTW-value may be affected by the length of the citation period. Furthermore, the threshold values of TOP 1% and TOP 5% DTW- values of the 5245 highly cited documents in different publication years, were extremely small numbers varying between 0.21 and 0.38. The 80% threshold values of TOP 1% DTW- values published in 21 different years varied between 0.21 and 0.29. This suggests that 1% or 5% of the highly-cited documents in the AI field are extremely close in distance to the "benchmarking sleeping beauty" written by Einstein in 1905.

Table 3

Descriptive statistics of the DTW- value of 5245 highly-cited documents

publication year	mean-value	minimum value	threshold value of TOP 1%	threshold value of TOP 5%	maximum value
1990	1.180	0.311	0.311	0.326	3.428
1991	1.350	0.286	0.286	0.373	3.292
1992	1.128	0.203	0.203	0.296	2.378
1993	1.165	0.309	0.309	0.364	2.324
1994	1.139	0.32 5	0.325	0.375	2.480
1995	1.128	0.213	0.213	0.369	2.244
1996	1.162	0.226	0.226	0.351	2.378
1997	1.085	0.248	0.256	0.337	2.474
1998	1.070	0.224	0.280	0.353	2.263
1999	1.096	0.280	0.300	0.379	2.367
2000	1.080	0.228	0.262	0.370	2.763
2001	0.966	0.221	0.248	0.318	2.362
2002	0.886	0.176	0.230	0.331	2.335
2003	0.881	0.204	0.213	0.325	2.059
2004	0.848	0.203	0.222	0.322	2.505
2005	0.849	0.212	0.226	0.335	2.234
2006	0.786	0.225	0.248	0.313	2.710

2007	0.784	0.206	0.268	0.333	2.822
2008	0.732	0.262	0.284	0.333	3.445
2009	0.689	0.233	0.288	0.353	1.974
2010	0.637	0.244	0.290	0.352	1.696

5. Discussion and Conclusion

The DTW method can identify potentially high-value papers that are very similar to the selected benchmark and closest to high-value citation curves such as sleeping beauties or "highly cited papers." The rapid identification and extensive recommendation of such high-value papers can maximize their scientific and application value. Based on the citation time series of a sleeping beauty published by Einstein in 1905, we measured the DTW-value between 5245 highly cited papers in the field of artificial intelligence, published from 1990-2010. From the empirical results, we can see that the DTW method is very robust, and it can identify potential sleeping beauties with similar citation curves and distance closest to that of the "benchmarking sleeping beauty." Although this study used a sleeping beauty authored by Einstein in 1905, 52 sleeping beauties in the top 1% and 262 sleeping beauties in the top 5% of DTW-values were identified from 5245 highly cited papers.

Many deficiencies remain in this research: (1) We only used the sleeping beauty published by Einstein in 1905 as the "benchmarking sleeping beauty," which is a subjective choice displaying "slow growth in the early stage, and fast growth in the later stage"; (2) The DTW method may not be affected by the length of citation time. The identified sleeping beauties displayed an annual deviation and were mainly distributed after 2000. Some identified sleeping beauties had a short sleep duration, leading to the misidentification of highly-cited papers

6. Acknowledgments

This study was supported by Jiangsu Provincial Social Science Foundation of China (Grant No. 19TQC004) "Research on the stratification and recognition mechanism of literature value in the field of social science"; and supported by sponsored by Qing Lan Project.

7. References

[1] Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. 2015. Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426-7431.

[2] Hu, Z., & Wu, Y. 2014. Regularity in the time-dependent distribution of the percentage of never-cited papers: An empirical pilot study based on the six journals. Journal of informetrics, 8(1), 136-146.

[3] Hu, Z., Wu, Y., & Sun, J. 2018. A quantitative analysis of determinants of non-citation using a panel data model. Scientometrics, 116(2), 843-861.

[4] Hu, X., & Rousseau, R. 2019. Do citation chimeras exist? The case of under-cited influential articles suffering delayed recognition. Journal of the Association for Information Science and Technology, 70(5), 499–508.

[5] Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. 2017. Sleeping beauties in Computer Science: characterization and early identification. Scientometrics, 113(3), 1645-1663.

[6] Van Raan, A. F. 2004. Sleeping beauties in science. Scientometrics, 59(3), 467-472.

[7] Barber, B. 1961. Resistance by Scientists to Scientific Discovery: This source of resistance has yet to be given the scrutiny accorded religious and ideological sources. Science, 134(3479), 596-602.

[8] Garfield, E. 1989. Delayed recognition in scientific discovery-citation frequency-analysis aids the search for case-histories. Current Contents, 23, 3-9.

[9] Wang, J. C., Ma, F. C., Chen, M. J., & Rao, Y. Q. 2012. Why and how can "sleeping beauties" be awakened? Electronic Library, 30(1), 5–18.

[10] Glanzel, W., Schlemmer, B., & Thijs, B. 2003. Better late than never? On the chance to become highly cited only beyond the standard time horizon. Scientometrics, 58(3), 571–586.

[11] Glanzel, W., & Garfield, E. 2004. The myth of delayed recognition. The Scientist, 18(11), 8–9.

[12] Lachance, C., & Larivière, V. 2014. On the citation lifecycle of papers with delayed recognition. Journal of Informetrics, 8(4), 863–872.

[13] Hou J , Yang X .2019. Patent sleeping beauties: evolutionary trajectories and identification methods[J]. Scientometrics, 120(1), 187-215.

[14] Ye, F. Y., & Bornmann, L. 2018. "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle. Journal of the Association for Information Science and Technology, 69(3), 359-367.

[15] Li, J., Shi, D., Zhao, S. X., & Fred, Y. Y. 2014. A study of the "heartbeat spectra" for "sleeping beauties". Journal of informetrics, 8(3), 493-502.

[16] Tsiporkova H E. 2017 . Merging microarray cell synchronization experiments through curve alignment[J]. Bioinformatics, 23(2):e64-70.

[17] Petitjean, F, Inglada, et al. 2012. Satellite Image Time Series Analysis Under Time Warping[J]. IEEE Transactions on Geoscience and Remote Sensing, 50(8), 3081-3095.

[18] Singh G , Bansal D , Sofat S. 2017 . A smartphone based technique to monitor driving behavior using DTW and crowdsensing[J]. Pervasive and Mobile Computing, 40, 56-70

[19] Fang,Y. 2022. Design of Oral English Intelligent Evaluation System Based on DTW Algorithm.MOBILE NETWORKS & APPLICATIONS, 27(4), 1378-1385.

[20] Aach, J., & Church, G. M. 2001. Aligning gene expression time series with time warping algorithms. Bioinformatics, 17(6), 495-508.

[21] Gollmer, K., & Posten, C. 1995. Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses. IFAC Proceedings Volumes, 28(12), 101-106.

[22] Keogh, E., & Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. Knowledge and information systems, 7(3), 358-386.

[23] Li, J., & Shi, D. 2016. Sleeping beauties in genius work: When were they awakened? Journal of the Association for Information Science and Technology, 67(2), 432-440.

[24] Einstein, A. 1905. On the Movement of Small Particles Suspended in Stationary Liquids Required by the Molecular-Kinetic Theory of Heat. Annalen der Physik, 17, 12.