NUIST >>>

南京信息工程大学
Nanjing University of Information Science & Technology

Identifying Potential Sleeping Beauties Based on Dynamic Time Warping Algorithm and Citation Curve Benchmarking

Reporter: Yu Chen Date:2023.06.26
Authors: Zewen Hu, Yu Chen & Jingjing Cui

明 德 格 物　　立 己 达 人

CONTENTS

CONTENTS

Research background

- 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 .
- 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 .
- Most academic databases or platforms 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 identifying and recommending "sleeping beauties" would have significant application value.

Research background

- Since Van Raan proposed the concept of sleeping beauty , a series of quantitative studies on the identification and application of sleeping beauties were implemented and published.
- (1) Identifying sleeping beauties through curve fitting.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 based on subjective indicators. Which is variable and greatly influenced by interference factors and subjective perception of scholars.
- (3) Identifying sleeping beauties by objective indicators. It 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.

Research questions and Objectives

- Research questions:
  - Whether there have a new method to efficiently identify sleeping beauties?
  - What methods can be used to accurately identify standardized sleeping beauties combining the advantages of curving fitting and indicators-based methods and overcome their disadvantages?
  - Which method can deliver the overall best outcome?
- Objectives:
  - Designing and implementing a new Dynamic time warping (DTW) method to more efficiently identify sleeping beauties based on “benchmarking sleeping beauty” citation curve .
  - Improving the Dynamic time warping (DTW) method to more accurately identify standardized sleeping beauties.
  - Comparing the performance of identifying sleeping beauties between DTW method and quadratic function fitting method.

DATA

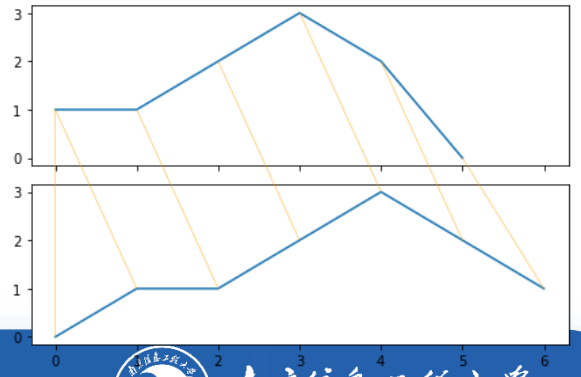
- The data from the Web of Science database top 1% of the highly cited papers (5425 articles) between 1990 and 2010 in the field of artificial intelligence.

- The number of highly cited as well as total number of papers in the field of artificial intelligence during the indicated 21 years.

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

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.
- 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.
- For any given “benchmarking sleeping beauty” citation curve, the DTW algorithm can effectively and accurately identify potential SBs that conform to the “benchmarking sleeping beauty” citation curve by calculating the closest DTW distance to the benchmarking citation curve.

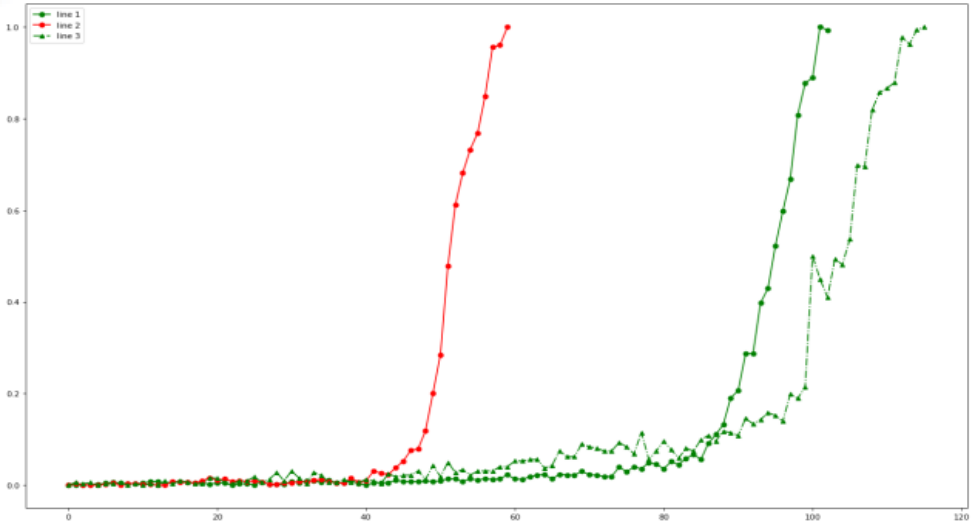


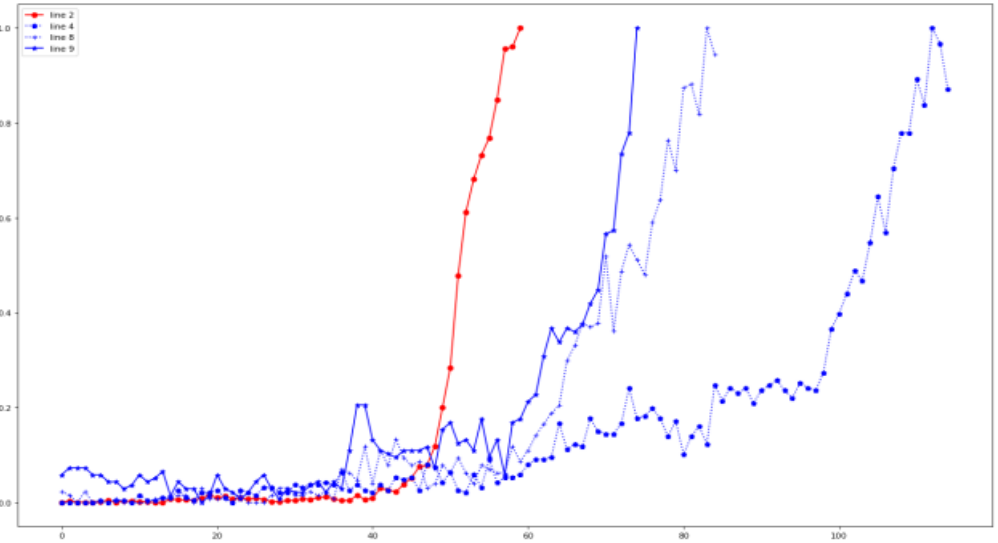
Exploratory experiment for sleeping beauty identification based on DTW

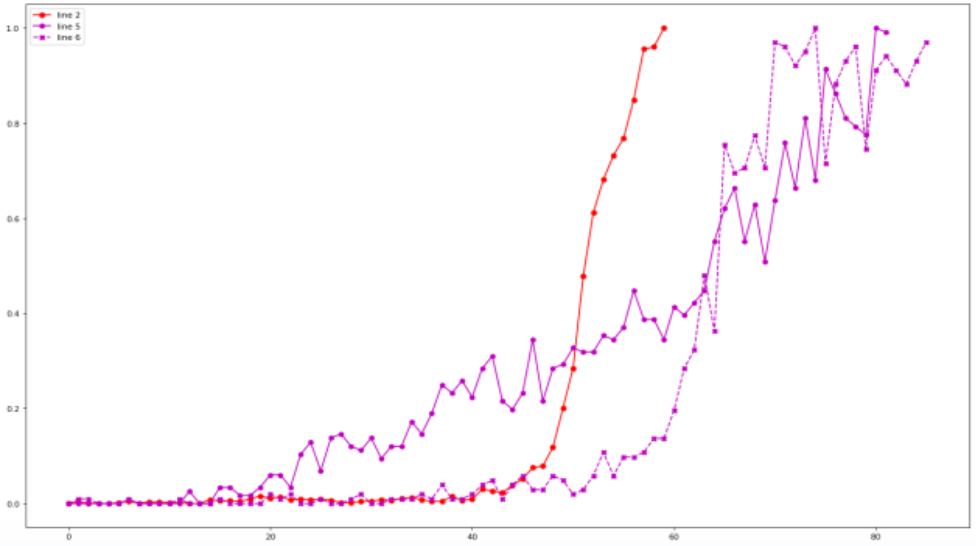
- We used 10 sleeping beauties identified by Li (2016) in preliminary tests to explore the potential of the DTW algorithm in identifying potential sleeping beauties.
- We chose the time series of the second sleeping beauty (document 2) as the benchmarking standard.
- Then, calculated the DTW-value between the benchmarking sleeping beauty and the other nine sleeping beauties.
- Finally, established a comparative baseline by selecting top 5 highly cited papers in the AI field, to calculate the average of DTW-values between each of the five highly cited papers and 10 sleeping beauties.

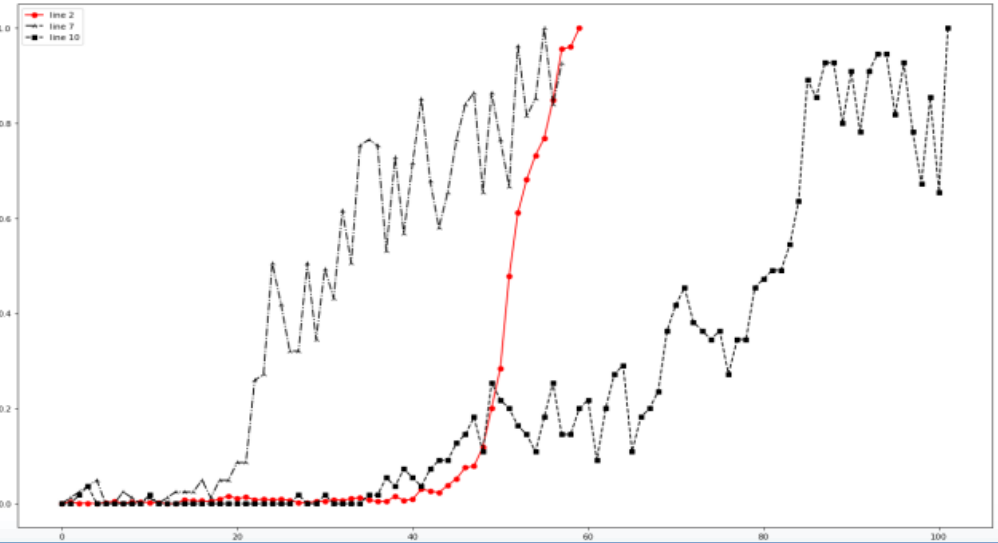
document id	author	source title	publication year	total cited frequency	DTW-value 1	DTW-value 2
2	Shockley, W; Queisser, HJ	JOURNAL OF APPLIED PHYSICS	1961	7608	0.000	0.601
1	Langmuir, Irving	JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	1918	14586	0.156	0.529
3	Einstein, A	ANNALEN DER PHYSIK	1905	5480	0.187	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

Exploratory experiment for sleeping beauty identification based on DTW



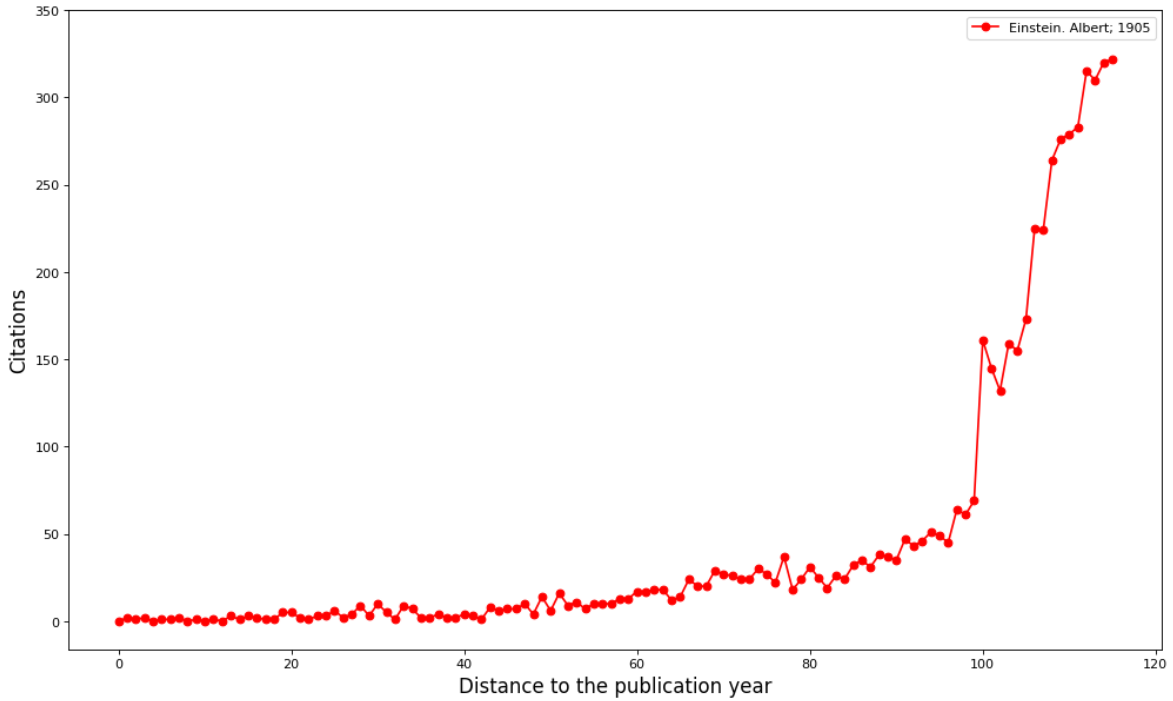






Verification of DTW method in identifying potential sleeping beauties

- We chose a benchmarking sleeping beauty with the oldest publication year, the longest citation period, and slower citation rate.
- Subsequently, we measured the DTW-value between the “benchmarking sleeping beauty” citation curve 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.



Verification of DTW method in identifying potential sleeping beauties

- We 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.
- The DTW-value may be affected by the length of the citation period.
- 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.

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

Verification of DTW method in identifying potential sleeping beauties

publication year	1%	5%	PSB5%	publication year	1%	5%	PSB5%	publication year	1%	5%	PSB5%
1990	0	2	0.76%	1997	3	10	3.82%	2004	7	24	9.16%
1991	0	2	0.76%	1998	1	9	3.44%	2005	5	20	7.63%
1992	1	4	1.53%	1999	0	5	1.91%	2006	4	28	10.69%
1993	0	2	0.76%	2000	2	9	3.44%	2007	2	19	7.25%
1994	0	1	0.38%	2001	9	19	7.25%	2008	1	19	7.25%
1995	2	6	2.29%	2002	5	19	7.25%	2009	2	13	4.96%
1996	1	7	2.67%	2003	5	30	11.45%	2010	2	14	5.34%

- Table shows the annual distribution of the number of documents for 1% and 5% thresholds of the DTW-value, sorted from small to large.
- The number of sleeping beauties in the top 1% and 5% of the DTW-values are 52 and 262, respectively.
- The number of sleeping beauty are increasing over time, especially from 2000 to 2010.
- The percentage of sleeping beauties at 5% threshold among the highly cited papers considered (abbreviated as PSB5%) were between 1% and 11%

Verification of DTW method in identifying potential sleeping beauties

- The DTW algorithm is not affected by the length of citation and sleep periods. The DTW method can effectively identify potential sleeping beauties with shapes close to that of the "benchmarking sleeping beauty."
- However, an obvious fact is verified in the Table that the DTW algorithm can identify not only some standardized sleeping beauties with sleep periods of more than four years, but also many sleeping beauties with extremely short sleep periods of 1- 4 years or false sleeping beauties (i.e. highly cited papers) without sleeping duration but similar citation curves.

publication year	5%	0 years	1-4 years	More than 4 years	publication year	5%	0 years	1-4 years	More than 4 years	publication year	5%	0 years	1-4 years	More than 4 years
1990	2	0	1	1	1997	10	4	1	5	2004	24	5	15	4
1991	2	1	1	0	1998	9	3	5	1	2005	20	4	16	0
1992	4	0	2	2	1999	5	0	5	0	2006	28	9	15	4
1993	2	0	1	1	2000	9	4	3	2	2007	19	5	13	1
1994	1	0	1	0	2001	19	8	9	2	2008	19	8	10	1
1995	6	0	3	3	2002	19	4	11	4	2009	13	4	9	0
1996	7	4	2	1	2003	30	7	17	6	2010	14	7	7	0

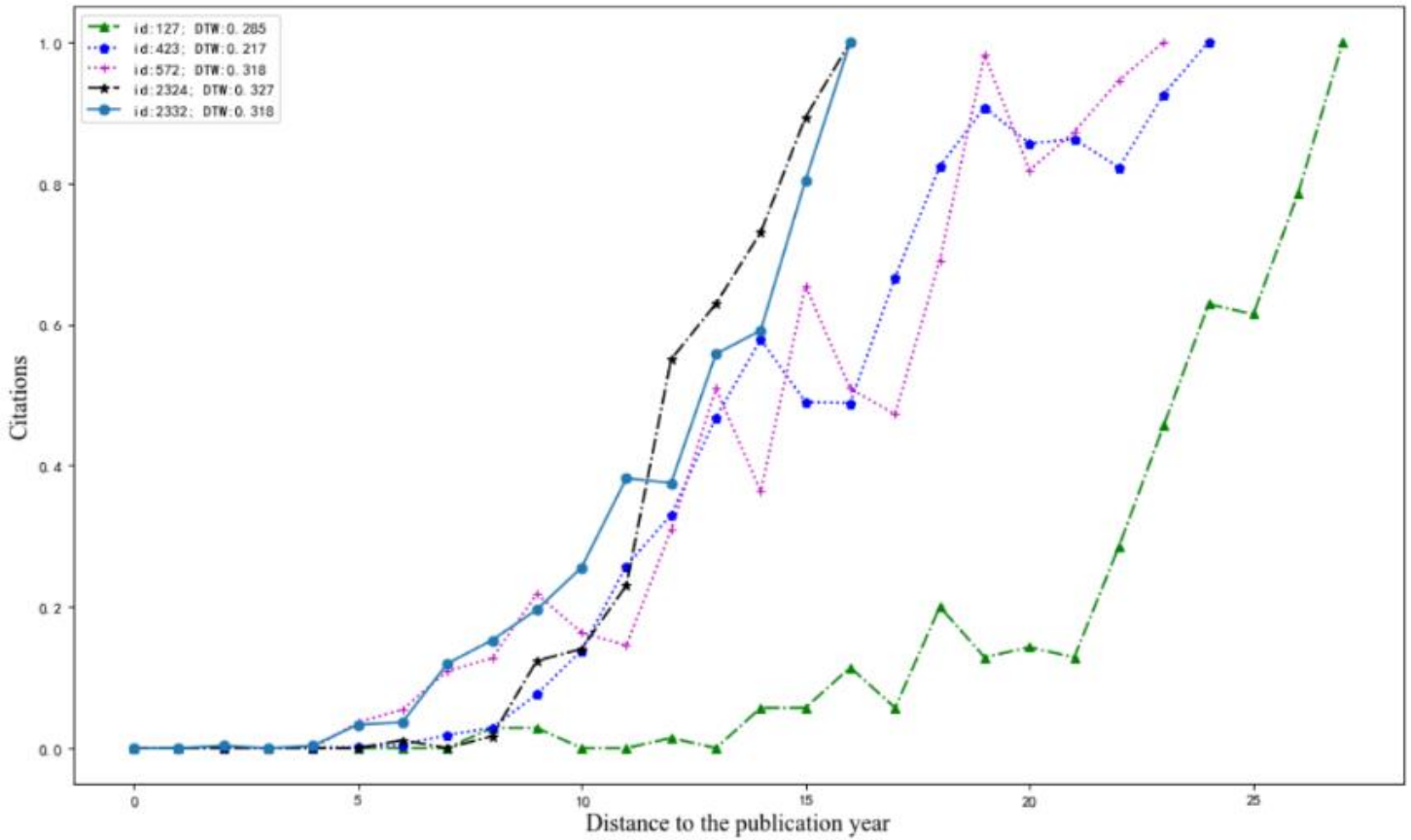
Experiments to improve the DTW algorithm

- Utilized the three indicators defined by Van Raan for sleeping beauties to set a DTW-value threshold for each year and identify a standardized sleeping beauty with the help of sleep length.
- **DTW* algorithm:** (1) Screening out papers with top 5% DTW-value; (2) Calculating the sleep length of these papers and selecting those with sleep time of no less than 5 years.
- The annual distribution of the 38 sleeping beauties identified by the new DTW* scheme more concentrated from 1992-2006.
- The average sleeping time of 38 sleeping beauties is 7.05years, and the longest is 18 years.

year	1%	5%	DTW*	year	1%	5%	DTW*	year	1%	5%	DTW*
1990	0	2	1	1997	3	10	5	2004	7	24	4
1991	0	2	0	1998	1	9	1	2005	5	20	0
1992	1	4	2	1999	0	5	0	2006	4	28	4
1993	0	2	1	2000	2	9	2	2007	2	19	1
1994	0	1	0	2001	9	19	2	2008	1	19	1
1995	2	6	3	2002	5	19	4	2009	2	13	0
1996	1	7	1	2003	5	30	6	2010	2	14	0

Experiments to improve the DTW algorithm

- The annual citation frequency distribution of the five sleeping beauties identified by DTW* with smaller DTW-values between 0.21 and 0.33.
- The sleeping beauties identified by the new DTW*scheme are the sleeping beauties with high standardized sleeping beauty citation curves.
- These five sleeping beauties have long sleep durations ranging from 6 to 18 years.



Effectiveness analysis of the DTW algorithm in identifying sleeping beauties

- Compared the differences in the recognition results of the DTW algorithm and the quadratic function fitting method.
- Table shows the descriptive statistics of the DTW-value and R2-value of 5245 highly cited papers.
- The higher negative correlation coefficient of 0.69 between DTW- and R2-values of 5245 highly cited papers verifies the effectiveness of the DTW algorithm in recognizing sleeping beauties with similar citation distribution curves.

	mean-value	standard deviation	minimum value	25%	median	75%	maximum
DTW	0.903	0.460	0.176	0.541	0.809	1.170	3.445
R^2	0.714	0.201	0.007	0. 876	0.757	0. 590	0.996

Effectiveness analysis of the DTW algorithm in identifying sleeping beauties

- **Quadratic function fitting:** limiting the R2 to greater than 0.876 of the 25% threshold value, thereby identifying 46 sleeping beauties from 5245 highly cited documents in the artificial intelligence field.
- The number of identified documents by the quadratic function fitting method and the top 1% DTW-value were relatively small, mainly because these two methods require a relatively standard citation distribution curve.
- The sleeping beauties with the optimal value identified by DTW and the quadratic function fitting method were mostly published after 2000.

year	1%	5%	DTW*	R^2	year	1%	5%	DTW*	R^2	year	1%	5%	DTW*	R^2
1990	0	2	1	0	1997	3	10	5	2	2004	7	24	4	3
1991	0	2	0	0	1998	1	9	1	2	2005	5	20	0	2
1992	1	4	2	0	1999	0	5	0	1	2006	4	28	4	5
1993	0	2	1	1	2000	2	9	2	2	2007	2	19	1	0
1994	0	1	0	1	2001	9	19	2	6	2008	1	19	1	4
1995	2	6	3	3	2002	5	19	4	4	2009	2	13	0	2
1996	1	7	1	3	2003	5	30	6	5	2010	2	14	0	0

Effectiveness analysis of the DTW algorithm in identifying sleeping beauties

- To verify that the results identified by the DTW algorithm are consistent with the sleeping beauties identified by the quadratic function fitting, the DTW-value of the 10 sleeping beauties with the top R2 are shown in Table.
- The same sleeping beauties with higher R2 and smaller DTW-value as well as extremely similar the “benchmarking sleeping beauty” citation curve.
- Identify some sleeping beauties with shorter sleeping periods or some highly cited papers without sleeping periods.

id	year	DTW	R^2	sleeping time
1697	2001	0.699	0.984605692	2
3080	2004	0.327	0.978363218	1
576	1996	0.371	0.968104878	8
3456	2006	0.456	0.968104878	0
3457	2006	0.573	0.968583331	5
2335	2003	0.482	0.966131727	5
425	1995	0.374	0.964887857	3
3465	2006	0.273	0.962386031	1
582	1996	0.563	0.960368301	1
3489	2006	0.361	0.960884988	3

Effectiveness analysis of the DTW algorithm in identifying sleeping beauties

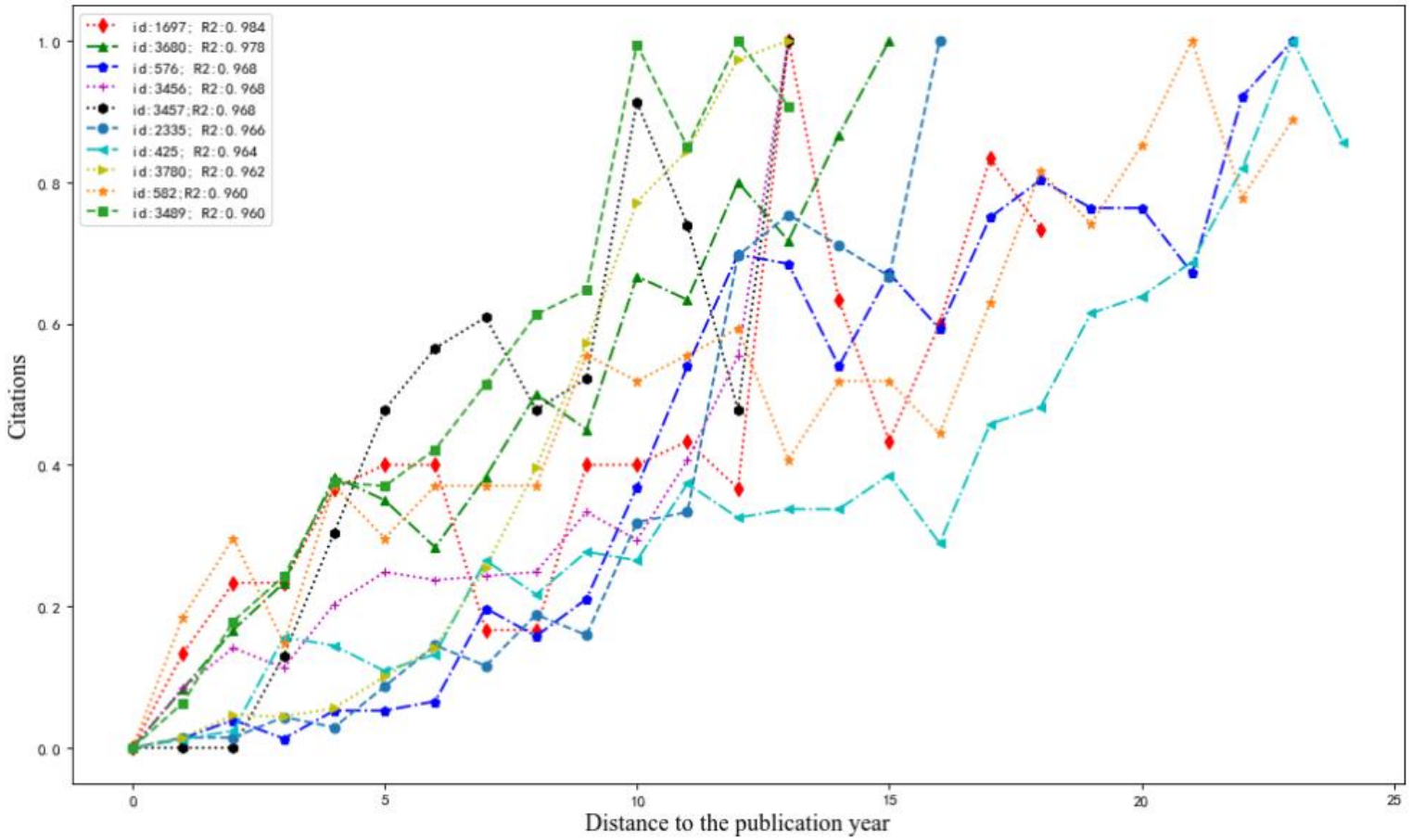
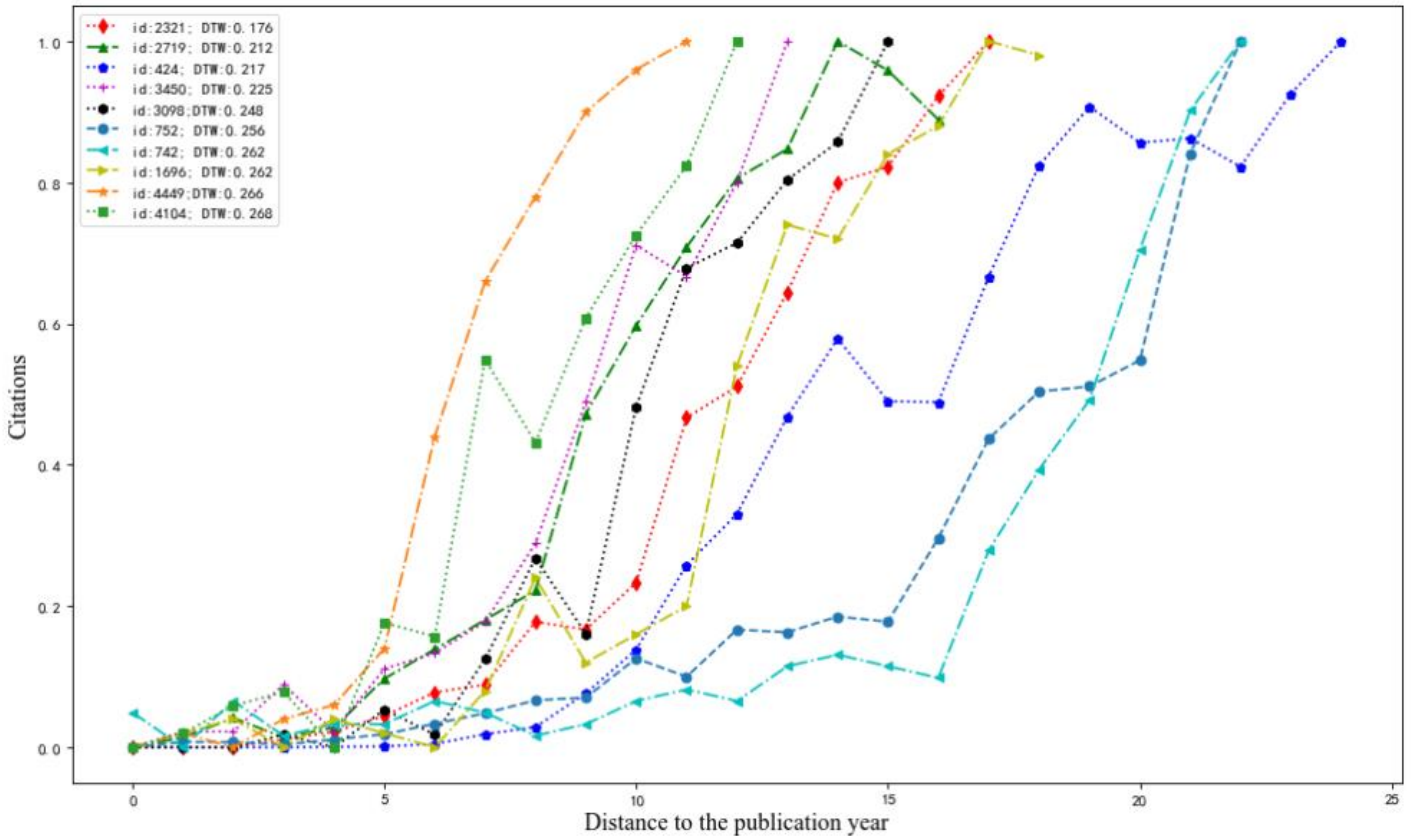


Figure shows the citation and fitting curves of the top 10 sleeping beauties sorted by R2.

Effectiveness analysis of the DTW algorithm in identifying sleeping beauties

- The 10 sleeping beauties with Top DTW-values identified by the new DTW* scheme, exhibited standardized sleeping beauty citation curves with long sleep periods of between five and eight years and shapes closer to that of the “benchmarking sleeping beauty.”
- DTW* algorithm has higher recognition accuracy in identifying all standardized sleeping beauties from the massive documents that fully conform to the shape of “benchmarking sleeping beauty” citation curve.



Summary

(1) 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.” However, this method may identify some non-standardized sleeping beauties with extremely short sleep periods.

(2) In terms of the annual distribution of the sleeping beauties, 52 and 262 sleeping beauties (the top 1% and 5% of the DTW-value) were distributed from 2000 to 2010, showing an increasing trend with years, while the 38 sleeping beauties identified by the DTW* method were distributed from 1995 to 2006, suggesting longer citation periods and sleeping times.

(3) The new DTW* scheme was superior to the DTW method, and the quadratic function fitting method not only inherited the higher efficiency and robustness of the DTW method, but also avoided the shortcomings of the DTW and the quadratic function fitting methods in identifying some highly cited papers without sleeping periods.

南京信息工程大学
Nanjing University of Information Science & Technology

Thanks for you listening !

明 德 格 物　　立 己 达 人