UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

gesis
Leibniz-Institut
für Sozialwissenschaften

Attention Dynamics of Scientists on the Web

Masterarbeit

zur Erlangung des Grades einer Master of Science (M.Sc.)
im Studiengang Web Science

vorgelegt von
Tatiana Sennikova
Matrikelnummer: 214202789

Erstgutachter:	JProf. Dr. Claudia Wagner GESIS – Leibniz-Institut für Sozialwissenschaften
Zweitgutachter:	Dr. Fariba Karimi GESIS – Leibniz-Institut für Sozialwissenschaften

Koblenz, 13. November 2016

Erklärung

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe und die Arbeit von mir vorher nicht in einem anderen Prüfungsverfahren eingereicht wurde.

	Ja	Nein
Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	☐	☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	☐	☐
Der Text dieser Arbeit ist unter einer Creative Commons Lizenz (CC BY-SA 4.0) verfügbar.	☐	☐
Der Quellcode ist unter einer MIT Lizenz verfügbar.	☐	☐
Die erhobenen Daten sind unter einer Creative Commons Lizenz (CC BY-SA 4.0) verfügbar.	☐	☐

. . . . . .

(Ort, Datum) (Unterschrift)

Acknowledgements

I would like to express my gratitude to both Claudia Wagner and Fariba Karimi for sharing their expertise, immense knowledge, and for the challenging questions which incented me to widen my research from various perspectives. Their guidance did not only result in important contributions to this thesis, but also kept me motivated and interested in the scientific research process. I am also grateful to Anna Samoilenko who provided insight and expertise that greatly assisted the research. Additionally, I would like to thank GESIS - Leibnitz Institute for the Social Sciences for financially supporting the research. I wish to express my sincere thanks to Alex Druk for providing the Wikipedia page views statistics. Without this data, my research would not be feasible. At last, I would like to thank my family and friends who supported me during writing this thesis. This accomplishment would not have been possible without them.

Zusammenfassung

Diese Arbeit betrachtet die Online-Aufmerksamkeit gegenüber Forschern und deren Forschungsthemen. Die enthaltenen Studien vergleichen die Aufmerksamkeitsdynamiken gegenüber Gewinnern wichtiger Forschungspreise mit Forschern die keinen Preis erhalten haben. Web-Signale wie Wikipedia Seitenaufrufe, Editierungen von Wikipedia-Artikeln und Google Trends wurden als Proxy für Online-Aufmerksamkeit verwendet. Dabei wurde herausgefunden, dass Wikipedia-Artikel über die Forschungsthemen von Gewinnern zeitnahe zum Artikel über den Gewinner erstellt wurden. Eine mögliche Erklärung hierfür könnte sein, dass die Forschungsthemen in einer engeren Beziehung zu den Gewinnern stehen. Dies würde die These unterstützen, dass Gewinner ihr Forschungsgebiet eingeführt haben. Zusätzlich wuchs die Online-Aufmerksamkeit gegenüber den Forschungsthemen von Gewinnern nach dem Tag an dem der Artikel über den Forscher erstellt wurde. Daraus kann abgeleitet werden, dass Themen von Gewinnern beliebter sind als die Themen von Forschern die keinen Preis erhalten haben. Des Weiteren wurde gezeigt, dass Gewinner des Nobelpreises vor der Verkündung weniger Online-Aufmerksamkeit erhalten als die Liste von Nominierten basierend auf den Thomson Reuters Citation Laureates. Ferner sank die Beliebtheit gegenüber der Preisträger schneller als gegenüber Forschern die keinen Preis erhalten haben. Zuletzt wurde demonstriert, dass eine Vorhersage der Gewinner basierend auf Aufmerksamkeitsdynamiken gegenüber Forschern problematisch ist.

Abstract

This thesis analyzes the online attention towards scientists and their research topics. The studies compare the attention dynamics towards the winners of important scientific prizes with scientists who did not receive a prize. Web signals such as Wikipedia page views, Wikipedia edits, and Google Trends were used as a proxy for online attention. One study focused on the time between the creation of the article about a scientist and their research topics. It was discovered that articles about research topics were created closer to the articles of prize winners than to scientists who did not receive a prize. One possible explanation could be that the research topics are more closely related to the scientist who got an award. This supports that scientists who received the prize introduced the topics to the public. Another study considered the public attention trends towards the related research topics before and after a page of a scientist was created. It was observed that after a page about a scientist was created, research topics of prize winners received more attention than the topics of scientists who did not receive a prize. Furthermore, it was demonstrated that Nobel Prize winners get a lower amount of attention before receiving the prize than the potential nominees from the list of Citation Laureates of Thompson Reuters. Also, their popularity is going down faster after receiving it. It was also shown that it is difficult to predict the prize winners based on the attention dynamics towards them.

Contents

List of Abbreviations

IMDb Internet Movie Database
CCF Cross-Correlation Coefficient
MK Mann-Kendall
DFT Discrete Fourier Transformation
DWT Discrete Wavelet Transformation
PAA Piecewise Aggregation Approximation
SVD Single Value Decomposition
SAX Symbolic Aggregate approXimation
HMM Hidden Markov Models
DTW Dynamic Time Warping
SpADe Spatial Assembling Distance
SUSh Scalable U-Shapelet
BOP Bag Of Patterns

1. Introduction

1.1. Motivation

Over the last years, the estimation of the scientific contribution of scholars got a lot of attention from the academic community [Aks06; HLN11; MM75]. Funding agencies, prize awarding committees and a variety of non-scientific organizations use the information about the academic impact to finance the most promising researches, to honor a scientist with a prize, or to hire the best experts in a field. For many years, the citation index has been used as an instrument to estimate the academic contribution of a scientist. However, many people would agree that the references listed in the bibliography of a paper gradually differ in their contribution to this paper [HLN11]. Therefore, there is an inherited bias in the citation counting method. At the same time, a tremendous growth of science over the last years led to a fragmentation of disciplines into small specialties [GW10]. It obstructs the usage of the citation index as an instrument for the scientific contribution evaluation [GW10]. Therefore, altmetrics reflecting the academic impact are needed. Recent research in the field of computational social science [SY14] seeks to solve this issue. The authors studied whether metrics derived from Wikipedia articles correlate with the academic notability. Despite the fact that they did not find a statistically significant correlation, they shed a light on the public perception of scholars. Nevertheless, the evaluation of a scientific contribution remains a challenging task.

This study addresses the problem of estimating the academic and social impact of scientists by analyzing the online attention towards them. Even though one could agree that a higher public attention towards a scientist proclaims his importance for the society, it does not always concur with his academic success. The way how Nobel Prizes were awarded in the last few years illustrates this contradiction.

In 2016, several months before the Nobel Prize announcement, people started using the hashtag #NobelforVeraRubin in their posts in social networks. This social action aimed to draw attention to the fact that the accomplishments of the American astronomer Vera Rubin, who provided evidence for the existence of dark matter, are not yet recognized by the Nobel committee. Further, many scientists agreed that her discoveries deserved to be honored by the prize. Moreover, in 2008 she was already named by Thomson Reuters as a candidate to win the Nobel Prize, based on her citation index. Nevertheless, in 2008 the Nobel Prize was divided between the Japanese physicists Yoichiro Nambu, Makoto Kobayashi, and Toshihide Maskawa. When examining the page view statistics of the Wikipedia articles about Vera Rubin, Yoichiro Nambu, Makoto Kobayashi, and Toshihide Maskawa before and after the prize announcement, one could observe that before the prize announcement, Vera Rubin got more attention from the public than the scientists who finally received the prize Figure 1.1. In

the subsequent years (2009-2015), the prize winners were more popular, but in 2016 the public drew its attention back towards Vera Rubin. This example shows that high public attention as well as a high citation index do not always lead to a scientific recognition. Therefore, addressing the question how academic impact, scientific success, and public attention relate to each other could bring new insights into methods of scientific contribution evaluation.

1.2. Approach and Research Questions

This work addresses the general question about the relation between scientific achievements and the online attention towards scientists. For this, I will consider the following questions:

1. Is the success of a scientist determined by the field he or she is working in or was the popularity of the field influenced by the scientist?
2. How does the public react to the success of a scientist?
3. Can we predict the future success of a scientist based on the dynamics of the public attention towards him/her?

The methods presented in the thesis attempt to answer these questions by examining different Web signals that reflect the public attention towards scientists and their research topics. The following Web signals are used:

- Wikipedia page views and edits of articles about scientists and their research topics.
- Google Trends statistics about the number of search requests containing the name of scientists or their research topics.

The choice of Wikipedia is driven by the fact that it takes the 5th place of the most visited web sites according to the Alexa ranking[1]. The advantages of using Wikipedia data for the analysis of a public behavior are caused by its nature. Since the content of Wikipedia can be freely created and edited by anyone, it rather represents the wisdom of the crowds than personal opinions. Moreover, it is often the first source for getting a quick introductory information about a topic [Wel+10]. Therefore, the audience of Wikipedia is diverse and reflects the real population of Internet users.

This thesis analyses the relation between the career success of scientists and the online attention towards them. For simplicity, the success of a scientist is defined as the fact of being honored by one of the following awards: Nobel Prize, Thomson Reuters Citation Laureates, Fields Medal, Abel Prize, International Prize for Biology, Turing Award, and IEEE Medal of Honor. The methods applied in this research are based on the time series analysis of Web signals. It compares the attention dynamics towards prize winners with a baseline that consists of notable scientists who worked at the same time and in the same fields as the prize winners but did not receive a prize[2].

[1] http://www.alexa.com/topsites (accessed Nov. 05, 2016)

[2] The code and datasets that were used for the thesis are available on https://github.com/tsennikova/scientists-analysis (accessed Nov. 13, 2016)

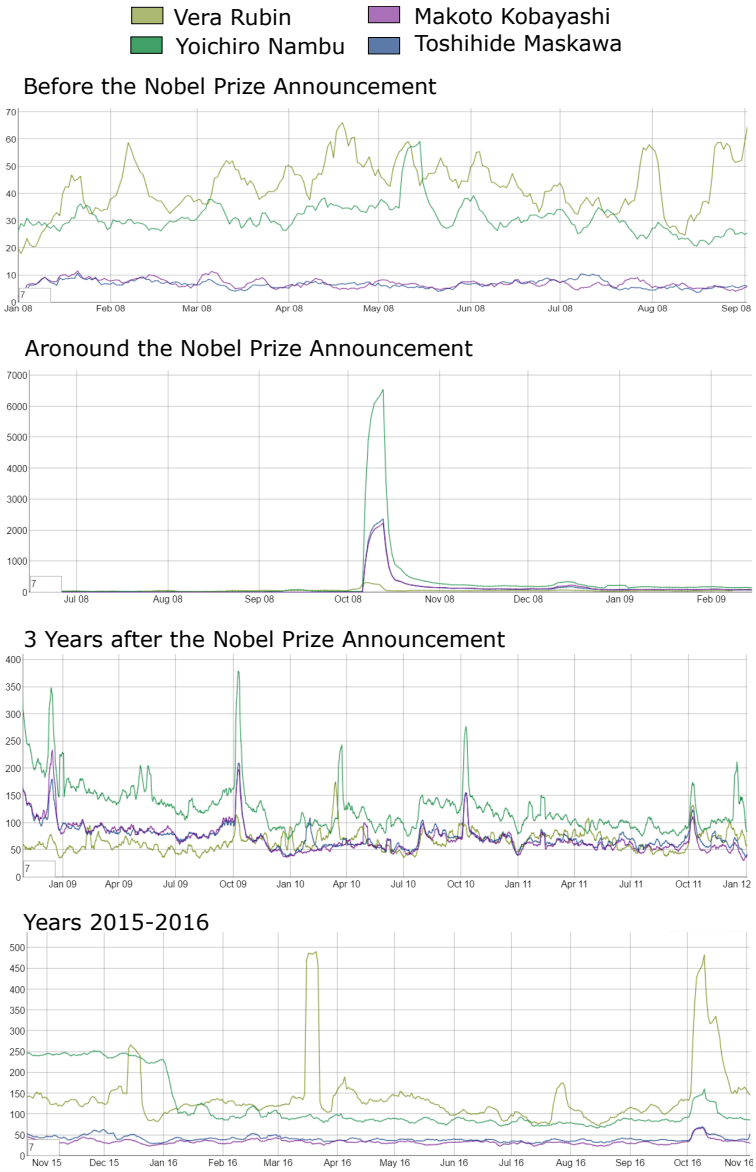


Figure 1.1.: Wikipedia page views statistics of the articles about Vera Rubin, Yoichiro Nambu, Makoto Kobayashi, and Toshihide Maskawa. The figure shows that before the prize announcement, Vera Rubin, who was not honored with the Nobel Prize, got more attention from the public than the scientists who received the prize. Around the time of the announcement of the award, the attention peaks towards the Nobel Prize winners, but in the subsequent years the attention towards them declines. In 2015-2016, the public attention towards Vera Rubin exceeds the attention towards the Nobel Prize winners of 2008 again. This example demonstrates that the public attention do not always concur with the scientific recognition.

1.3. Overview

The remaining part of this thesis is structured as follows: Chapter 2 discusses the related studies in success prediction and attention dynamics on the Web. Chapter 3 describes the used datasets and methods of data collection. Chapter 4 provides a theoretical background on time series analysis and motivates the choice of the applied methods. Chapter 5 and Chapter 6 describe the applied techniques and the results of the analysis. Chapter 7 concludes the thesis.

2. Related work

This chapter gives an overview of the related studies. It consist of two sections. In the first section, the researches that aim to predict the success of a person are reviewed. The second section gives an overview of the studies related to the analysis of the public attention on the Web.

2.1. Prize Predictions

The studies related to the prize prediction use the same approaches as the success prediction. The fields that received the most attention from the academic community are Oscar winners prediction and Nobel Prize winners prediction.

2.1.1. Oscar Winners Prediction

Every year, the Academy of Motion Picture Arts and Sciences awards films, filmmakers, actors, artists, and technicians for the cinematic achievements in the film industry. An Oscar nomination and win boosts the box-office revenue of the film and increases the public attention. In this section I will review several studies aimed to predict Oscar winners based on the information available on the Internet including the dynamics of the public attention towards Oscar winners and nominees.

In his research about the prediction of the Oscar winners, Pardoe [Par07] used discrete choice model to provide predictions about the four major Oscar categories for each year from 1938 to 2006. He demonstrated that due to more available information, the prediction accuracy for the years 1977 to 2006 has improved from 69% to 79%. He also showed that it is more difficult to predict Best Actress than Best Actor or Best Director. Moreover, winning another award in the past improves the probability to get an Oscar in the future.

In another research, Kaplan [Kap06] used logistic regression to analyze 40 years of Best Picture nominations. He used a number of parameters including genre, length of movie, previous nominees, and personnel for the prediction model. He showed that an experienced crew, a proficient director who won The Directors Guild of America Award, a famous actor and an epic genre biography of the movie gives a 99,7% chance to win the Best Picture Oscar.

Krauss et al. [Kra+08] applied methods of social network analysis and web data mining to run a model for forecasting movie success. The paper presents an analysis of forum discussions on the Internet Movie Database (IMDb) through the prism of “intensity” and “positivity”. Krauss and others observed a correlation between the Academy Awards presentation and the communication about the movie on the IMDb forum. They found out that a higher movie success correlates with a higher communication intensity and positivity of the discussion. At

the same time there is a 88% correlation between the intensity of discussion and its positivity. Thus, Krauss and others showed that estimating the level of positivity of the forum discussion is enough to predict an Oscar win or nominee.

2.1.2. Nobel Prize Winners Prediction

This research addresses the following question: "Can we predict the future success of a scientist based on the dynamics of the public attention towards him?". I am approaching this problem by analyzing the attention dynamics towards successful scientists. Thus, it is important to look into the methods applied to predict the prizes that are awarded for academical achievements. The Nobel Prize is the award that receives most attention from researchers. This section presents studies that aim to identify possible Nobel Prize winners.

One of the first works aimed to answer the question if Nobel Prize winners could be predicted was done by Garfield and Malin [GM68] in 1968. The authors claimed not to predict, but rather to identify the group of possible candidates analyzing the citation statistics. Garfield and Malin use Science Citation Index to rank top 1% of scientist who should be considered as the Nobel Prize nominees. In the research the first author method of the citations counting was used. The following factors were considered: how much time passed since the scientist's most cited paper, the field in which he worked, and if it was theoretic or experimental work. It was found that Nobel Prize winners have the citation counts approximately fifty times higher than the average scientist. Nobel Prize winners publish frequently, they have at least one key paper that have high continuous impact, and they are cited over a long period of time before and after the prize.

Later on Ashton and Oppenheim [AO78] carried a study to test the hypothesis of Garfield and Malin that citation counting can be used to predict the Nobel Prize winners. They discovered that the method that counts the citations of the articles where the scientist is not a first author has higher prediction power than first author method. It happens due to the fact that most of noticeable scientists stop publishing papers as a first authors after they recieved a public recognition.

In contrast with the studies conducted by Garfield, Malin, Ashton, and Oppenheim, more recent research carried by Gingras and Wallace [GW10], showed that due to the rapid growth of science and the fragmentation of disciplines into many small specialties, simple citation counting has lost its predictive power. Gingras and Wallace examined evolution of the profiles of Nobel Prize winners in chemistry and physics between 1901 and 2007. The analysis was performed over three distinct periods (1901 – 1945, 1946 – 1970, 1971 – 2007). It showed that in the first period, the peak of centrality and citation occurs the same year as the award. The second period showed that as the scientific community became more fragmented, concentration of attention around the potential winners and other scientists distributed more evenly. After 1970 the centrality and citation of Nobel Prize winners and nominees are distributed nearly uniform. In the third period there is no clear evidence of an important peak in citation or centrality before the Prize was awarded. Therefore, the predictive power of bibliometric measures has lost its power.

Karazija and Momkauskaite [KM04] studied the Nobel laureates in physics and their topics. They examined the following characteristics: winners and losers, distribution of

the candidates over different countries, nationalism and internationalism in the nominations, gender inequality of the candidates, etc. They found that there are few fields which got more awards than others. They are: elementary particles, nuclear physics and atomic physics. The authors showed that in the second half of the 20th century the discoveries are more often made by the groups of scientists, than before. The authors demonstrated that the probability to obtain the prize for the theorist is larger than for the experimenter. In relation to the age of the Nobel laureats, Karazija and Momkauskait revealed that the age of the winners at the moment of their main discoveries varies from 22 till 62 years and the time-interval between the discovery and its recognition by the Nobel committee is around 15 years. Regarding to the distribution of the male and female candidates, proportion of women among the proposed candidates is around 1%.

2.2. Attention Dynamics

The studies about the attention dynamics can be divided into two groups. The first group of studies examines how the collective attention influences the creation and consumption of information on the Web. The second investigates the connection between the collective attention on the Web and real-world events.

Information Creation and Consumption

Lehmann et al. [Leh+14] analyzed biography articles in Wikipedia to understand how the public attention influences the creation of a new information. The authors analysed the public attention through the prism of monthly views, time spent on the page and the number of pages views during the visit. They studied creation of a new information by examining the length of the article and the number of edits. The analysis showed that the most read articles do not necessary correspond to the frequently edited. Despite the fact that the most edited articles tend to be long and have a better quality, article quality does not drive popularity.

Later on Ciampaglia, Flammini, and Menczer [CFM15] studied whether the creation of a new knowledge precedes or follows its demand, based on the page views statistics of the Wikipedia articles. The authors found that in most of the cases demand of the information precedes its supply. It was shown that the Wikipedia article gets more attention when it is created shortly before or after the attention peak to the corresponding topic.

Szabo and Huberman [SH10] presented a method for predicting the long–term popularity of online content based on the early measurements of user access. They collected data from Digg (http://digg.com) and YouTube (http://youtube.com) to predict the amount of attention that particular article or video will get over time. As user attention focuses on the content with the extreme regularity, it is possible to predict long-term popularity of content based on the early attention patterns. The authors discovered that while Digg stories gain attention quickly (about a day), YouTube videos keep attracting views over their lifetimes. Nevertheless, the general rule that the more popular is content at the beginning, the more popular it will be later on is true for both cases. However, Digg allows to make more accurate predictions, since the attention raises here more quickly. Observing social networking of Digg, Szabo

and Huberman found that users tend to pay attention to the same content as their peers. Nevertheless, after content was promoted, the social network does not affect the users choice.

Prediction of the Real-World Events

In their work, Gruhl et al. [Gru+05] studied the relation between online content and customer behavior such as purchase decisions. They analyzed around half a million sales for 2,340 books over a period of four months. The authors demonstrated that the time series analysis of blog postings and web discussions represent an early indicator of "real-world" behavior. Finally, they showed that the volume of blog postings can be used to predict spikes in actual consumer purchase decisions at the online retailer Amazon.

Asur and Huberman [AH10] analyzed whether the dynamics of the public attention on the Web can predict real-world outcomes. They used Twitter to forecast box office revenue of the movies. The research showed that the rate at which movie tweets are generated can be used to build a powerful model for predicting a movie box office revenue. Moreover, this prediction is consistently better than Hollywood Stock Exchange. They showed that correlation between the average tweet rate per hour and the box office revenue is 90%.

Another research aimed to investigate the connection between the collective attention on the Web and the real-world events was conducted by Goel et al. [Goe+10]. In the first part of the research the authors showed that the online search statistics can anticipate a consumer behavior. In the second part, Goel et al. compared search-based models with the models based on the public available data. Despite the fact that previous analysis showed that the search data is a good predictor of future outcomes, alternative information sources often perform equally good or even better. In the third part of the research the authors reexamined the utility of the search data in monitoring influenza caseloads. The research demonstrated that in the flu monitoring as well as in some other domains, search data are comparable in utility to the alternative information sources such as public reports of flu caseloads provided by the Centers for Disease Control and Prevention.

3. Data collection

The following Web signals were used to analyze the online attention towards prominent scientists and their research topics:

- Wikipedia page views and edits of articles about scientists and their research topics.
- Google Trends statistics about the number of search requests containing the name of scientists or their research topics.

These Web signals were collected for the following datasets:

- Dataset of prize winners (or seed dataset)
- Dataset of baseline scientists (see Section 3.2)
- Dataset of research topics of the prize winners
- Dataset of research topics of the baseline scientists

The methodology of the data collection and the datasets themselves are described in the following. The sizes of the datasets are presented in Appendix A.2.

3.1. Prize Winners

I manually collected a list of scientists who received one of the following awards: Nobel Prize, Thomson Reuters Citation Laureates, Fields Medal, Abel Prize, International Prize for Biology, Turing Award, IEEE Medal of Honor. The data were collected for the research fields of biology, chemistry, computer science, economics, mathematics, physics, and medicine. I collected the gender of the scientists, year of the award, date when the Wikipedia article was created, and the scientific field of the scientist. This information was collected for scientists who received an award during the period between 2008 and 2015. The choice of the time period is due to the fact that page view statistics of the Wikipedia articles are not available for the years before 2008. Overall, the dataset of the prize winners contains 262 unique scientists. In the following, this dataset is referred to as the seed dataset. The distribution of the scientists between different disciplines, prizes, and the number of male and female prize winners are presented in Appendix A.4, Appendix A.5, Appendix A.3.

3.2. Baseline Scientists

For the baseline I use a list of highly cited scientists who worked in the same scientific fields and in the same time period as the prize winners, but who did not receive an award. To collect this data, the list of Highly Cited Scientists provided by Thomson Reuters was used[1]. The list consists of highly cited scientists between 2001 and 2015. The larger time period was chosen because of the time gap between the initial discovery and a potential prize win. The scientists who have already received an award in the past were removed from the list. Then, 262 scientists with the same distribution between scientific fields as in the seed dataset were randomly picked.

3.3. Scientific Topics

In this work, incoming and outcoming Wikipedia links of the article about the scientist are considered to determine the topics of a scientist. The data collection was done in three steps. First, all incoming and outcoming links from the Wikipedia article about the scientist were collected. Then, a list of categories that the links belong to was retrieved. On the third step, I filtered this list using a set of stop words. The set of stop words contains the words that are related to a person, institution, or geographical location. This way, the articles from irrelevant categories were eliminated. I used a general rule "if the link belongs to the category which name contains a word from the stop list, then this link should be removed from the list of topics". The evaluation of the filtering algorithm was performed on 10 randomly picked articles about scientists. Overall, 590 links were examined. The results of the evaluation are presented in Table 3.1.

3.4. Wikipedia Pageviews

I received the Wikipedia page view statistics from the project Wiki Trends [2]. Wiki Trends provides aggregated page view statistics for English Wikipedia articles. The data is based on the number of the daily visits to English Wikipedia articles consisting of the Wikipedia pages and all redirects to them. The data of this project was derived from Wikimedia data dumps[3]. In order to eliminate influences of daily and seasonal patterns on article's page views, the data is normalized as follows:

$$Vnorm_i = \frac{V_i * max(M)}{M_i} \tag{3.1}$$

Here, V_i is the number of daily visits to a certain article, M_i is the number of Wikipedia Main Page views for the same day, and $max(M)$ is the maximum number of Wikpedia Main Page views.

[1] http://hcr.stateofinnovation.thomsonreuters.com/page/archives (accessed Jul. 28, 2016)
[2] http://www.wikipediatrends.com/ (accessed Jul. 28, 2016)
[3] https://dumps.wikimedia.org/ (accessed Jul. 28, 2016)

		Prediction: True	Prediction: False
Actual	True	112	13
	False	9	457

Table 3.1.: Evaluation of the filtering algorithm. The algorithm eliminates links that are not related to scientific topics. For the evaluation, 10 Wikipedia articles were randomly picked. In total, 590 links from these articles were evaluated. The overall accuracy of the method is 0.96, precision is 0.93, and recall is 0.9.

3.5. Wikipedia Edits

I collected the daily number of edits of the Wikipedia articles about scientists and their research topics. This data was collected via the Wikipedia API[4]. The edits made by bots were filtered. For this filtering, a list of Wikipedia bots[5] was used. Repeated edits from the same person during one day were eliminated.

3.6. Google Trends

The Google trends data was collected in two steps. First, I queried scientist names and related topics. If there is an exact match between the query and the Google search term, then I retrieved a relative number of the Web searches (see Equation (3.2)). If an exact match was not found, then I collected data for the most relevant search term suggested by Google. While the data about Wikipedia page views and edits were collected on a daily basis, Google trends provides the data only on a weakly basis. The collected data is worldwide and covers the period between January 2004 and May 2016. It is important to note that Google does not provide raw data about the number of searches performed by users. It shows the total number of searches for a term relative to the total number of search requests sent to Google over time.

[4] https://en.wikipedia.org/wiki/Special:ApiSandbox#action=query&format=json (accessed Nov. 09, 2016)

[5] https://en.wikipedia.org/w/index.php?title=Category:All_Wikipedia_bots&from=D (accessed Jul. 28, 2016)

The values are scaled on a range of 0 to 100 based on the topic's proportion to all searches on all topics [6]:

$$G_i(x) = \frac{100 * x_i}{T_i * M_n} \tag{3.2}$$

Here, x_i is the number of searches that include the keyword of interest at the time interval i, T_i is the total number of searches at this interval, n is the number of time intervals i, and $M_n = max_{i=1...n} \frac{x_i}{T_i}$ [Ask15].

[6] https://support.google.com/trends/answer/4365533?hl=en (accessed Nov. 02, 2016)

4. Theoretical background

This chapter provides a theoretical background on time series analysis and motivates the choice of the applied methods. First, the definition of the time series should be given:

Definition 1: A time series is a set of real values $T = t_1, t_2, ..., t_n$ ordered in time.

I continue by describing the correlation analysis between the datasets of different Web signals.

4.1. Time Series Correlation Analysis

Correlation is an essential feature of the time series analysis. It measures the linear dependence between two points of two different time series observed at different times. In given research correlation analysis is used to explore the relations between different Web signals. The linear relationship between them is expected. In most of the cases, correlation analysis should be performed over the stationary time series.

Definition 2: A strictly stationary time series is a time series where given t_1,..., t_l the joint statistical distribution of X_{t1},..., X_{tl} is the same as the joint statistical distribution of $X_{t1+\tau}, ..., X_{tl+\tau}$ for all τ and l, where τ is a time shift and l is a number of observations.[Nas10].

Another words, probability distribution of the stochastic process X_t is constant under a shift in time. However, the definition of strictly stationary time series is too strict for the real life data, therefore the weak definition of stationarity is frequently used.

Definition 3: A weakly stationary time series is a time series that has finite variance and its mean does not depend on time. [Nas10]

All the time series that do not satisfy these requirements are considered to be non stationary. Most of the real life time series are non stationary. In given work the stationarity of the time series is ensured by dataset normalization (see Chapter 3) and by applying moving average.

Definition 4: A moving average is formed by computing the mean of the time series values t_1,..., t_i over a specific number of periods.

Moving average helps to filter the "noise" and smooth the seasonality peaks. After removing seasonality and trends from the time series, the correlation analysis can be performed.

Definition 5: : Correlation is a linear measure of similarity between two signals. Cross-correlation is a generalization of the correlation measure as it takes into account the lag of one signal relatively to the other. If $lag = 0$, then the correlation is equal to the cross-correlation.

In this study cross-correlation coefficient (CCF) defined as the Pearson product-moment correlation coefficient for lagged time series. This method of the correlation analysis is not new, but it is computationally cheap, non-parametric and showed good results over time.

Therefore, it is reasonable to use it for research, where correlation between the time series is not in itself a subject of study. I describe how the CCF was calculated in Section 5.1.

4.2. Time Series Trend Analysis

There are no proven techniques to identify the trend components in the time series data. However, to find a monotonic trend one of the statistical tests can be applied. The first step in the process of the trend identification is smoothing. I use moving average to smooth the time series. Then Mann-Kendall (MK) test is used to determine a monotonic trend.

4.2.1. Mann-Kendall Test

Mann [Man45] suggested to use the test for significance of Kendall's tau to determine the monotonic trend of the signal. In general, the MK test checks whether values tend to increase or decrease with the time. The test is non parametric and it does not require the normality assumption. The assumptions behind the MK test are:

- When no trend is present, the measurements are independent and identically distributed.
- The sample collection provides representative observations of the underlying populations over time.

I describe the steps of MK test in Section 5.3.

4.3. Time Series Clustering

There are two key aspects to achieve accurate and efficient clustering of the time series: data representation methods and similarity measures [Din+08]. In this section I compare existing representation methods, similarity measures and clustering approaches that could be applied to the time series.

4.3.1. Representation Methods

Time series are essentially high dimensional data that often suffer from the noise and outliers. The processing of raw time series is expensive in terms of computational power and storage costs. Therefore, many representation methods have been investigated. Esling and Agon [EA12] describeàq three main categories of the data representation: nondata adaptive, data adaptive, and model based. In this chapter I will follow the same taxonomy.

Nondata Adaptive Representation

In nondata adaptive representation, parameters of the transformation do not depend on the shape of the data and stay the same for every time series. The most common representations of this type are: Discrete Fourier Transformation (DFT), Discrete Wavelet Transformation (DWT), and Piecewise Aggregation Approximation (PAA) [Din+08].

DFT can be used to decompose the data into simpler pieces, it is a frequency domain representation of the data. DFT maps n features from initial time series into points of N-dimensional feature space [FRM94]. One of the problems of DFT is that it does not provide the frequencies location in time. Therefore, many researchers found DWT to be more effective representation [CF99; PM02; CFY03]. DWT represents time series as a wavelet series that are discretely sampled from the initial data. The advantage of using DWT over DFT is its temporal resolution. Another words, DWT is able to give the locations of frequencies in time. In general, if Fourier coefficients represent a global view to the data, wavelet coefficients represent local subsections of the data.

PAA have been proposed by Keogh et al. [Keo+01] as more flexible and powerful alternative of DWT. First, the data is divided into m equal-sized subsequences. Then the mean is calculated for each subsequence and a vector of the mean-values becomes the dimensionality reduced representation of time series. The advantage of PAA over DWT is possibility to handle queries of an arbitrary length, faster computational speed, and the ability to use different distance measures.

Data Adaptive Representation

In data adaptive representation methods parameters of the transformation can vary depending on the nature of the available data. Almost all the nondata adaptive techniques can become data adaptive by adding data-sensitive selection step [EA12]. The category of data adaptive representations includes the following methods Single Value Decomposition (SVD), Symbolic Aggregate approXimation (SAX), representation by shapelets and many others.

SVD allows to reduce a set of time series of length n to a set of points in an N-dimensional space. However, SVD requires eigenvalues computation for large matrices which has $O(Sn^2)$ complexity. Therefore, applying it even for moderate size datasets is not practical [Cha+02].

SAX was introduced by Lin et al. [Lin+03] in 2003. It is a new symbolic representation of time series that based on the same idea as PAA. In contrast to PAA that construct just an approximate representation of the time series, SAX representation takes into account the underlying data and adjusts its parameters in order to minimize the global reconstruction error. SAX representation reduces dimensionality of the time series by creating a sequences of short words. SAX enables to run data mining algorithms such as clustering, classification, indexing, and anomaly detection. The results produced by these algorithms over the time series represented as the SAX words are identical to the algorithms that operate on the raw data. This representation technique is superior to other representations, including PAA [Lin+03; SRT07].

Another method of the time series discretization similar to SAX has been proposed by Bagnall, Janacek, and Zhang [BJZ03]. The method discretizes the time series to a binary string using the median as the threshold. The authors showed that discretizing data to above and below the median can diminish negative effect of the outliers without reducing the accuracy of clustering algorithms when there are no outliers, and significantly increase accuracy when outliers are more likely.

Finally, a high promising method of time series representation was proposed by Ye and Keogh [YK11]. They introduced a new time series primitive, time series shapelet. Time

series shapelet is a time series subsequence that maximally represent the data. The goal of the shapelet is to split the dataset of time series into two groups, based on their similarity towards the shapelet. Good shapelet is a subsequence that from the one hand is not a part of majority of time series in the dataset, from another it is not too unique. This property ensures a good separation power. The idea of the method is based on the assumption that most of the data are irrelevant for the clustering [UBK15]. Therefore, selective ignoring of the data helps to mitigate sensitivity to noise, outliers and other irrelevant information. It has been shown that shapelets are much more expressive in terms of representation power and could be significantly more accurate and robust for many data mining algorithms [YK11]. Shapelets represent the local features of the time series, whereas most of the other time series representations consider global features. Ye and Keogh demonstrated that the shapelets representation can provide accurate, interpretable, and much faster classification and clustering in a wide variety of domains. The disadvantage of this method is the time consuming shapelets discovery. The best known running time for the shapelets discovery algorithms is $O(n^2m^3)$, where n is the number of time series in the dataset and m is the length of the longest time series [RK13].

Model Based

Model based approach of the time series representation grounds on the assumption that a particular time series is generated by an underlying model. Hence, the task of the data representation reduces to the task of finding the parameters of this model. Two time series are considered to be similar if they have been generated by models with the same parameters [EA12]. The most well-known models are ARIMA models and Hidden Markov Models (HMM).

ARIMA model captures stochastic properties of the time series. It assumes that every time series consists from the following components a trend, a cycle, a stochastic resistance component, and a random element. The goal of ARIMA representation is to find these parameters for each time series [KGP01].

In HMM the system is assumed to be a set of unobserved states, each of which has a probability distribution over the possible outcome that being observed. A transition matrix specifies the probability to move from one state to another in each point of time. One of the differences between ARIMA and HMM is that ARIMA requires fitting the model to each data before clustering, whereas HMM involves forming the cluster models on each iteration of the clustering algorithm [BJZ03].

4.3.2. Similarity Measures

According to Esling and Agon [EA12],) similarity measures should satisfy the following requirements:

1. It should guarantee a correct recognition of similar objects, even when they are not identic mathematically

2. It should concord with the human intuition

3. It should reflect notable features in a local and global views
4. It should be universal
5. It should be resistant to the transformations

Ding et al. [Din+08] distinguish four categories of the similarity measures lock-step measures, elastic measures, threshold-based measures, and pattern-based measures.

Lock-Step Measures

Lock-step similarity measures are the measures that compare the i-th point of one time series to the i-th point of another time-series [Din+08]. The classical example of the lock-step measure is Lp norms measures, such as the Euclidean distance. The Euclidean distance has been the most widely used similarity measure for the time series. Besides being the most intuitive measure it has several other advantages. It has linear complexity of evaluating, it is easy to implement, and it is parameter-free. Despite the fact that some researches claim its poor performance due to the warping issue and sensitivity to the noise and outliers [AO01], recent researches [Din+08; SK08] showed that the Euclidean distance is superior for the large datasets as in this case the probability to find almost exact match is high.

Elastic Measures

Elastic measures suppose to be a solution to the time warping and sensitivity issues of the Euclidean distance. This type of the measures provides comparison of one-to-many and one-to-none points [Din+08]. The most well-known elastic measure is Dynamic Time Warping (DTW) [BC94]. DTW allows to match different sections of the time series by its stretching or compression. However, it has been shown that computational time of DTW is relatively high $O(n^2)$, it could cause a problem for the large datasets. Another disadvantage of the algorithm is its approximate nature that does not guarantee finding of the optimal solution [EA12]. In the research Ding et al. [Din+08] showed that DTW is superior over the Euclidean distance especially for the small datasets. However, the Euclidean distance is faster, more straightforward and shows almost the same accuracy as DTW on the large datasets.

Threshold-Based Measures

The idea of the threshold-based measures, such as TQuEST, based on the idea of using a threshold τ, to transform the time series into a sequence of threshold-crossing time intervals. Each interval could be represented as a point in a two-dimensional space, where the dimensions are starting and ending time of the interval. The similarity in this case is defined with one of the Lp norms measures applied to the interval points. However, the experiments held by Ding et al. [Din+08] showed that performance of TQuEST is worse than Euclidean and DTW distances.

Pattern-Based Measures

The algorithm of the pattern-based measures, such as Spatial Assembling Distance (SpADe), finds patterns in the data by shifting and scaling the time series along the temporal and amplitude dimensions. SpADe requires a number of parameters to be tuned. For example temporal and amplitude scale factors, pattern length, size of the sliding window etc. Experiments showed that accuracy of SpADe is close to the Euclidian distance, but lower than DTW [Din+08].

4.3.3. Clustering Algorithms

"The goal of clustering is to identify structure in unlabeled data, by organizing data into homogeneous groups where the within-group-object similarity is minimized and the between-group-object dissimilarity is maximized" [Lia05].

In contrast to the static data, the time series feature's values are changing over time. Clustering of the time series differs from the clustering of the static data in the method how the similarity between the objects is computed. Once the distance measure is defined, it is possible to adapt any general-purpose clustering algorithm for the task of the time series clustering [EA12]. The following general-purpose clustering algorithms have been applied in the previous studies: hierarchical clustering, k-means, and self-organizing maps [Lia05]. It is important to say that, clustering algorithms could be performed over any type of the time series representation.

Hierarchical Clustering

Hierarchical clustering groups the time series objects in a tree of clusters. Two types of hierarchical clustering algorithm could be distinguished: agglomerative (bottom-up) and divisive (top-down). Hierarchical clustering often suffers from its inflexibility, once the merge operation has been executed, it cannot be modified. If elastic distance measure, such as DTW was chosen, hierarchical clustering is suitable for the time series of unequal length [Lia05].

K-Means

The idea behind k-means algorithm is to group the objects into k clusters in which each object belongs to the cluster with the nearest mean. The algorithm of data partition is iterative. It starts with the arbitrary chosen initial cluster centers. After that two main steps of the algorithm are performing: distribution of objects among clusters and updating clusters center. In a classical k-means algorithm each object belongs to only one class. However, there are modifications where each object could belong to the several classes. This algorithm calls fuzzy c-means, it helps to overcome the disadvantage of a classical k-means, that fail to find clusters with complex shapes. K-means algorithm as well as its modifications works better with the time series of equal length. The reason is the usage of the Euclidean distance as a similarity measure. Moreover, the concept of cluster centers becomes unclear, when the cluster contains time series of unequal length [Lia05].

Self-Organizing Maps

Self-organizing map is a class of neural network that is trained using unsupervised learning. The training process is initialized by assigning the random values to the weight vectors of network neurons. After the initialization the algorithm iterates over the three main steps: presentation of randomly chosen input vector, evaluation, and updating of the vector weights [Lia05]. The similarity measure that is usually used is the Euclidean distance. Since the self-organizing maps behave in a way similar to k-means, it does not work well with time series of unequal length.

5. Methodology

Time series analysis is an area of research that has attracted a lot of attention from the academic community in recent years [EA12; Lia05; YK11; UBK15]. Most of the research is focused on the developing new clustering techniques and distance measurements. The most well-known methods were introduced in Chapter 4. In this chapter the methodology of the applied techniques will be described.

5.1. Time Series Correlation Analysis

To estimate the correlation between two time series I use the CCF that stands for cross-correlation coefficient and is defined as a function of the time lag of one time series relatively to the another [Joh07]. The values of the CCF vary between -1 and +1. To assess how much the time series T_1 agrees with the time series T_2, T_1 should be shifted along T_2 in order to find the minimum distance between the time series [Joh07]:

$$\rho_{X,Y}(\tau) = \frac{E[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)]}{\sigma_X \sigma_Y} \tag{5.1}$$

Here, X and Y are two time series, τ is a time lag, μ_X and μ_Y are the mean values of the time series X and Y, σ_X and σ_Y are the values of variance of the time series X and Y. For example, Figure 5.1 shows Google Trends and Wikipedia page views time series of the economist Toni Atkinson, who was considered as a candidate to win the Nobel Prize by Thomson Reuters in 2012. The Web signals correlate with $CCF = 0.56$. The correlation between two datasets D_1 and D_2 can be find as [Joh07]

$$\rho_{D_1,D_2}(\tau) = \frac{\| \sum_{i=1}^{n} \rho_{X_i,Y_i}(\tau) \|}{n} \tag{5.2}$$

where n is a number of time series objects in the datasets D_1 and D_2.

For example, to calculate the correlation between the datasets of Google Trends and Wikipedia Views one should calculate the correlation between Google Trends and Wikiedia Views of each scientist and then take the average of the absolute values of the CCF.

5.2. Time Series Time Lag Analysis

In order to understand the attention dynamics towards scientific topics the time lag analysis was performed. First I collected the dates when the Wikipedia articles about the scientists and related topics were created. Then I calculate the time difference between the creation date of the article about the scientist and the article about related topics articles. The time difference

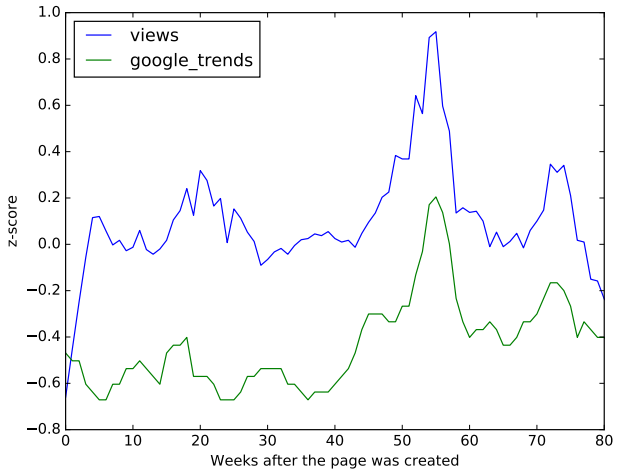


Figure 5.1.: The figure shows the time series of Wikipedia page views of the article about Toni Atkinson and Google Trends search statistics of the number of queries containing his name. The cross-correlation coefficient between these Web signals is equal to 0.56. The figure demonstrates that the time series have similar patterns, but different amplitudes.

is positive if the topic article was created after the article about the scientist, and it is negative in the opposite case. Then I assume that all scientist articles were created on day 0 and count how many topic articles have time lag of 1, 2, 3 etc. days. At the end I create a list of pairs "time lag – number of articles create" that I will use for the analysis.

For example, the Wikipedia article about the biologist Akiko Iwasaki is linked with the topic articles from Table 5.1, and the Wikipedia article about the chemist Richard Holm is linked with the topic articles from Table 5.2. After the time lags between the creation date of the topics and the scientists articles were calculated the list of pairs "time lag – number of articles created" can be obtained see Table 5.3. I use this list for the analysis.

Article Title	Day of the Creation	Time Lag
Akiko Iwasaki	2015-03-30	0
Immunobiology	2005-04-09	-3642
Molecular biology	2001-07-30	-4991
T cell immunity	2015-04-20	21
Commensal bacteria	2005-11-11	-3426

Table 5.1.: The table demonstrates an example how the time lags between the article about the scientist and articles about the related topics are calculated. It shows the time lags between the day when the article about the biologist Akiko Iwasaki and articles about his related topics were created. It was assumed that all scientist articles were created on day 0. One can see that the article "T cell immunity" was created 21 day after the article "Akiko Iwasaki", whereas the articles "Immunobiology", "Molecular biology", and "Commensal bacteria" were created before the article "Akiko Iwasaki".

Article Title	Day of the Creation	Time Lag
Richard Holm	2007-10-18	0
Chemical synthesis	2002-09-13	-1861
Iron-sulfur cluster	2002-09-30	-1844
Carbon monoxide dehydrogenase	2007-11-08	21
Biomimetic synthesis	2013-11-24	2229

Table 5.2.: The table demonstrates an example how the time lags between the article about the scientist and articles about the related topics are calculated. It shows the time lags between the day when the article about the chemist Richard Holm and articles about the related topics were created. It was assumed that all scientist articles were created on day 0. One can see that the articles "Carbon monoxide dehydrogenase" and article "Biomimetic synthesis" were created after the article "Richard Holm", whereas the articles "Chemical synthesis" and "Iron-sulfur cluster" were created before the article "Richard Holm".

Time Lag	Number of Articles Created With This Time Lag
-4991	1
-3642	1
-3426	1
-1861	1
-1844	1
21	2
2229	1

Table 5.3.: The table shows how the list of pairs "time lag – number of articles created with this time lag" is formed. The list is based on the time lags calculated in the previous step that were presented in Table 5.1 and Table 5.2. One can see that there are two articles that were created with the time lag equal to 21. These articles are the article "T cell immunity" that was created 21 day after the article "Akiko Iwasaki" and article "Carbon monoxide dehydrogenase" that was created 21 day after the article "Richard Holm". This list is used to analyze how many topic articles were created with each time lag.

5.3. Time Series Trends Analysis

I applied the trend analysis in order to understand whether there is a rising or falling trend in the attention dynamics towards the scientists and scientific topics. For this purpose the MK test was used. The steps of the test are described in [Gil87]. I will briefly summarize them in the following:

1. For time series $T = t_1, t_2, ..., t_n$ the sign of all $\frac{n(n-1)}{2}$ possible differences $t_j - t_k$, where $j < k$ should be determined.

2. $sign(t_j - t_k)$ is an indicator function that takes values -1, 0, or 1 according to the sign of $t_j - t_k$.
   - $sign(t_j - t_k) = 1 \quad \text{if} \quad t_j - t_k > 0$
   - $sign(t_j - t_k) = -1 \quad \text{if} \quad t_j - t_k < 0$
   - $sign(t_j - t_k) = 0 \quad \text{if} \quad t_j - t_k = 0$

3. Compute the number of positive differences minus the number of negative differences:

$$P = \sum_{k-1}^{n-1} \sum_{j-k+1}^{n} sgn(x_j - x_k) \tag{5.3}$$

4. Compute the variance of P:

$$VAR(P) = \frac{1}{18}[n(n-1)(2n+5) - \sum_{(p-1)}^{g} t_p(t_p - 1)(2t_p + 5)] \tag{5.4}$$

Here, g is the number of groups that consist from the elements that are equal to each other (tied groups), and t_p is the number of the elements in the p-th tied group. For example, in the time series $\{1, 5, 7, 5, 5, 7, 8, 1\}$ there are $g = 3$, $t_1 = 2$ for the element "1", $t_2 = 3$ for the element "5", and $t_3 = 2$ for the element "7".

5. Compute the MK test statistic:

$$\begin{aligned} Z_{MK} &= \frac{P-1}{\sqrt{VAR(P)}} \quad \text{if} \quad P > 0 \\ Z_{MK} &= 0 \quad \text{if} \quad P = 0 \\ Z_{MK} &= \frac{P+1}{\sqrt{VAR(P)}} \quad \text{if} \quad S < 0 \end{aligned} \tag{5.5}$$

The trend is positive for $Z_{MK} > 0$ and negative for $Z_{MK} < 0$. If $Z_{MK} = 0$ then the data do not have a monotonic trend.

5.4. Time Series Clustering Analysis

In this work, two clustering algorithms were implemented. Here I briefly describe both algorithms. More details about the work of the algorithms are given in Section 5.4.3 and Section 5.4.4.

The first implemented algorithm was introduced by Ulanova, Begum, and Keogh [UBK15]. The method is called SUSh (Scalable U-Shapelet) clustering. It considers u-shapelets which are time series subsequences that maximally represent the data (the formal definition of the u-shapelets is given with Equation (5.11)). The idea behind the algorithm is based on the assumption that most of the data in time series are useless for clustering. SUSh clustering focuses on the local parts of time series while ignoring the global shape.

The second implemented algorithm is called Bag of Patterns (BOP) clustering. The algorithm was proposed by Lin and Li [LL09]. It is based on the opposite assumption to SUSh clustering. It states that local patterns of time series and their order are meaningless for a clustering. However, the global shape of time series is important.

In the next section, I will define the key terms and notations that are used to describe the two clustering algorithms.

5.4.1. Definitions

Most of the definitions that are used in this section are adapted from Ulanova, Begum, and Keogh [UBK15], Zakaria, Mueen, and Keogh [ZMK12], Lin et al. [Lin+07], and Lin and Li [LL09]. In this work, time series data about Wikipedia page views, Wikipedia edits, and the number of Google searches of prominent scientists and related research topics are analyzed. In Chapter 4, a definition to the time series data was given. Now, the dataset of time series objects should be defined.

Definition 6: A dataset of time series objects is a dataset $D = T_1, T_2, ..., T_n$, where T_i is a time series object.

Definition 7: A subsequence $S_{k,l}$ is a set l of continues real values from a time series T_i that satisfies a condition: $1 \leq l \leq n$ and $1 \leq k \leq n$, where l is a length of the subsequence and k is a starting position in time series T_i.

As it was mentioned in Section 4.3, there are two key aspects to achieve accurate and efficient clustering of the time series: distance measure and data representation methods. First I will define the distance measures that were used.

Definition 8: The subsequence distance is a minimum distance between subsequence S of length l and time series subsequence $T_{i,l}$. In other words, in order to find the alignment with the minimum distance, the subsequence S should be slided against the time series T. In the research the Euclidian distance is used. It is defined as:

$$dist(S, T_{i,l}) = \sqrt{\frac{1}{l}\sum_{k=1}^{l}(S_k - T_{i+k})^2} \tag{5.6}$$

In order to make the distance measure sustainable to time warping and differences in scale, I perform a z-normalization of the time series before computing the distance:

$$z_i = \frac{t_i - \mu}{\sigma} \tag{5.7}$$

Here, t_i is a single point in a time series T. μ and σ are mean and variance of T. Thus, the subsequence distance is defined as:

$$sdist(S, T) = min_{1 \leq i \leq n-l} dist(S, T_{i,l}) \tag{5.8}$$

where $1 \leq l \leq n$.

After defining the distance measures, the data representation method will be introduced. In this thesis I use the SAX representation since it previously showed good results for clustering time series [UBK15; LL09].

5.4.2. SAX Representation

The idea of the SAX representation is based on the algorithms and data structures that are used for text analysis. Text retrieval algorithms received a lot of attention during the last decades. They are well studied and have shown good results for a number of data mining tasks such as indexing, classification, and clustering. In general, the SAX representation of the time series is a discretization and numerosity reduction technique that converts raw time series into a symbolic view. The SAX representation of a time series is space and time efficient since it allows distance measures that lower bound the distance measures which are defined on the original series [Lin+07].

To convert raw time series data into SAX representation a sliding window is used. It extracts every possible subsequence of length n. In this work I use $n = 360$, what approximately equal to the length of a year and allows a better reflection of the attention dynamics, including

β_i \ α	3	4	5
β_1	-0.43	-0.67	-0.84
β_2	0.43	0	-0.25
β_3		0.67	0.25
β_4			0.84

Table 5.4.: A lookup table of breakpoints for the size of the alphabet from 3 to 5. The breakpoints in the table divide the Gaussian distribution on the equally probable pieces. This table is used to obtain a symbolic representation of the time series.

seasonality. The choice of n can also be explained by the fact that in order to make the code more clear n should be dividable by the length of the SAX word w. In this research $w = 9$ (I explain the choice below), therefore it is reasonable to use $n = 360$, instead $n = 365$.

I perform the z-normalization of each subsequence and then transform the data using the Piecewise Aggregation Approximation (PAA) [Cha+02]. The idea of PAA is intuitive. To reduce the dimensionality of a time series from m to w dimensions, the data should be divided into w equal size subsets. Then the mean of each subset should be calculated and a vector of mean values becomes the data reduced representation. Formally it can be written in the following way: For time series T of the length n, the PAA representation is $\grave{T} = \grave{t}_1, ...\grave{t}_w$, where

$$\grave{t}_i = \frac{w}{m} \sum_{j=\frac{m}{w}(i-1)+1}^{\frac{m}{w}i} \grave{t}_j \tag{5.9}$$

Here, w is the number of PAA segments that represent time series T. Further it will represent a word in the SAX series. The choice of w depends on the data. Lin and Li [LL09] claim that time series with smooth patterns can be described with a small w. On the other hand, time series with rapidly changing patterns can be better described with greater w. Since the data in the research represent the attention dynamics towards scientists before and after getting an award, the patterns tend to change fast around the winning date. Therefore, based on an empirical evaluation, I use $w = 9$. Bigger word sizes did not improve the clustering performance but dramatically increased the calculation time. Smaller word sizes showed a worse clustering performance.

After the time series are transformed using PAA, a symbolic representation can be obtained. It is done by using a breakpoints table. The breakpoints in the breakpoints table are defined in such a way that they represent equally probable pieces of the Gaussian distribution. The breakpoints can be looked up in a statistical table. For example, Table 5.4 gives breakpoints for the size of the alphabet from 3 to 5.

All PAA values that are below the smallest breakpoint are mapped to the symbol “a”, all the values that are greater or equal to smallest breakpoint but less than second smallest breakpoint are mapped to “b” etc. The idea of the mapping is illustrated on the Figure 5.2.

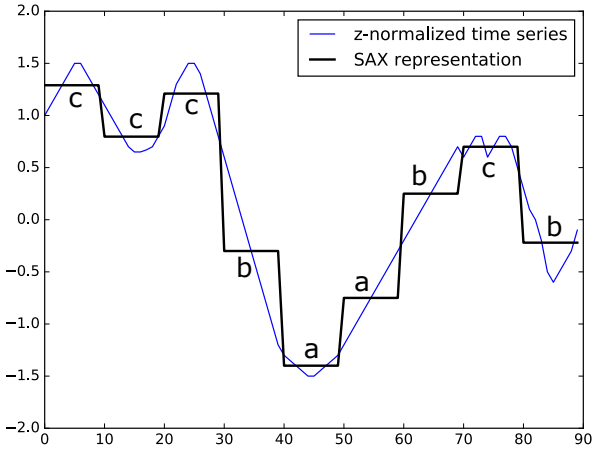


Figure 5.2.: SAX representation of the time series with the alphabet size $A = 3$ and length of the SAX word $w = 9$. The blue line on the plot denotes the raw time series. The black line stands for the time series transformed using the Piecewise Aggregation Approximation. The values of the Piecewise Aggregation Approximation are the mean values of the subsequences extracted from the raw time series. The length of each subsequence is $w = 9$. Based on the values obtained from the breakpoints table the values of the Piecewise Aggregation Approximation transformed to the symbols from a to c. In this way the raw time series with the length equal to 90 was transformed to the SAX word "cccbaabcb" with the length equal to 9.

Formally it can be written in the following way: The mapping from a PAA representation $\grave{T} = \grave{t}_1, ..., \grave{t}_w$ to the SAX word $\tilde{T} = \tilde{t}_1, ..., \tilde{t}_w$ using alphabet $A = \alpha_1, \alpha_2, ..., \alpha_l$ is obtained with

$$\tilde{t}_i = \alpha_j, \quad \text{if} \quad \beta_{j-1} \leq \grave{t}_i \leq \beta_j \tag{5.10}$$

where β_j is a breakpoint [LL09].

Lin and Li [LL09] argue that size of A does not have a significant impact on the performance of the clustering algorithm applied after. In our research the alphabet size is equal to 4.

After obtaining a set of strings corresponding to a subsequence in the time series, one can observe that there are consecutive subsequences that are mapped to the same string. In this case I store only the first occurrence of the string and ignore the repetitions until I meet a string that is different. This way a numerosity reduction is achieved.

5.4.3. SUSh Clustering

In this subsection I give details about the SUSh clustering. It was shown by Ulanova, Begum, and Keogh [UBK15] that SUSh clustering outperforms earlier existing algorithms because of the following reasons. First, the u-shapelet technique allows to consider only relevant subsequences of the time series data while ignoring the noise. Second, it is defined for time series of different lengths. Third, u-shapelets could provide additional insights into the data. An additional advantage of u-shapelets clustering is the possibility to assign "non-class" label to the time series objects. The authors point out that real-word datasets often contain data that could not be separated into clusters. Attempts to label them could lead to poor clustering results.

The steps of the SUSh clustering are following:

1. Convert raw time series into SAX representation (see Section 5.4.2)
2. Reduce u-shapelets candidates with a low separation power (see Random Masking Algorithm in Section 5.4.3)
3. Cluster the time series objects in the dataset based on its $sdist$ to the u-shapele candidate (see Clustering in Section 5.4.3)

First, a definition to the core concept of the SUSh clustering – unsupervised shapelet (u-shapelet) – should be given.

Definition 9: An unsupervised shapelet $\grave{S}$ is a subsequence of a time series T for which the $sdist$ between $\grave{S}$ and any time series from subset of time series D_A is significantly smaller than $sdist$ between $\grave{S}$ and the rest of time series D_B in the dataset D.

$$sdist(\grave{S}, D_A) << sdist(\grave{S}, D_B) \tag{5.11}$$

It can be said that u-shapelet $\grave{S}$ has a separation power. It can split the time series dataset D into two groups based on the distance to $\grave{S}$. The vector that contains all the subsequence distances $sdist(\grave{S}, T_i)$ from u-shapelet $\grave{S}$ to time series T_i of the dataset D calls an orderline.

If an u-shapelet candidate does not satisfy the rule 5.11, then it has a low separation power. It happens because of several reasons. First, the u-shapelet could be a part of majority of the

times series in the dataset. By analogy to text retrieval such a u-shapelet could be called a stop word like “the” or “a”. These words are included in every document and therefore they are useless for clustering. Another example of a bad u-shapelet is a subsequence that is too unique. Such a u-shapelet candidate is too rare to be useful for separation.

In order to estimate the separation power of a u-shapelet candidate a gap score is used. It is defined as:

$$gap = \mu_B - \sigma_B - (\mu_A + \sigma_A) \tag{5.12}$$

where μ_A and μ_B represent $mean(sdist(\grave{S}, D_A))$ and $mean(sdist(\grave{S}, D_B))$, while σ_A and σ_B represent $std(sdist(\grave{S}, D_A))$ and $std(sdist(\grave{S}, D_B))$ [UBK15]. When D_A or D_B contain insignificant number of objects, the gap score is set to zero.

The searching algorithm of the best u-shapelet can be seen as a greedy search which is aimed to maximize the gap between time series subsets D_A and D_B of the dataset D. However, to calculate the gap scores for each u-shapelet candidate, the Euclidean distance needs to be calculated $O(NM^2 \log M)$ times [UBK15]. Here, N is the number of time series in the dataset D and M is the average length of the time series. To optimize the computation time, one of the discretization techniques should be applied to the time series. In this research the SAX representation was used (see Section 5.4.2)).

Random Masking Algorithm

After converting the time series into its SAX representation, one can see that depending on the alphabet size, length of the SAX word, and the size of the sliding window cause a situation in which two identical time series subsequences are mapped to similar, but not necessary identical strings. As was discussed above, the bottle-neck of the algorithm is the gap-score computation that has to be calculated for each u-shapelet candidate. In the SAX representation, every SAX word is a u-shapelet candidate. Therefore, it is important to reduce the number of candidates for which the gap score should be calculated. Since similar u-shapelets will have similar separation power, similar words should be removed from the SAX representation. I achieve this by applying the random masking algorithm.

First, recall the parameters I use for clustering. I use a sliding window with the length 360, alphabet size of $A = 4$ and a SAX word length of $w = 9$. By applying the random masking algorithm I aim to find similar words. In this case it is the words that differ in three or less symbols. I perform 10 rounds of random masking, where I count how many time series share a masked SAX word with each u-shapelet candidate. Then I filter out all the candidates that share the same mask with too many or too few time series, as they do not satisfy the requirements for the good u-shapelet candidate.

Suppose there is a u-shapelet candidate $s_1 = \{a, a, a, b, b, c, c, b, a\}$ and the SAX word $w_1 = \{a, a, b, b, c, c, c, b, a\}$. The algorithm randomly eliminates three symbols. After applying random masking, the masked u-shapelet candidate is $s_1' = \{a, a, *, b, *, c, c, *, a\}$ and the SAX word is $w_1' = \{a, a, *, b, *, c, c, *, a\}$. One can see that the candidate and the word share the same masked signature. It means that they have a collision. It is expected that the most similar subsequences will collide more often.

After I applied 10 rounds of random masking, I obtained a vector of the numbers of collisions for each candidate after each round. Ulanova, Begum, and Keogh [UBK15] demonstrate that the variance of the number of collisions for a u-shapelet candidate is a good predictor of its gap score. U-shapelet candidates that have low variance are more likely to have higher gap scores. Therefore, in order to check the most promising candidates first, I sort the list of u-shapelet candidates by the variance of the number of collisions. Moreover, Ulanova, Begum, and Keogh [UBK15] prove that it is sufficient to compute the gap score for less than 1% of the candidates to find u-shapelets that are a good enough for a separation.

Clustering

The purpose of a u-shapelet is to separate time series objects in the dataset by its $sdist$ to this u-shapelet. This purpose defines the clustering algorithm that is applied in this work. I iteratively split the dataset of time series with each u-shapelet, considering the subset D_A to be a separate cluster. Then I remove D_A from the rest of the data and repeat splitting with the next u-shapelet. The stop criteria of the algorithm is the decline of the gap score. As soon as the gap-score of the next u-shapelet becomes less than half of the gap score of the first used u-shapelet, I stop the clustering algorithm and assign the label "no-class" to the rest of the data.

5.4.4. Bag Of Patterns Clustering

In this subsection I describe BOP clustering. BOP clustering is inspired by the bag of words representation of textual data. When comparing two strings, one can use a string edit distance such as the Levenshtein distance to estimate the strings similarity. When comparing two documents, one does not compare them word by word. Instead, the documents should be compared on a higher level that can capture its structure and semantics. The BOP approach considers each time series as a document. Each document can be represented as a vector in the vector space. Each dimension of the vector corresponds to a word from the vocabulary. The value of each component reflects the frequency of the given word in the document.

The steps of the BOP clustering are the following:

1. Convert the time series into sets of SAX words (see Section 5.4.2)
2. Construct a pattern vocabulary
3. Construct a word-sequence (bag of pattern) matrix using the SAX words

   Formally, in the bag of patterns matrix M, each row i denotes a SAX word from the vocabulary, each column j denotes a time series from the dataset, and each value M_{ij} stores the frequency of word i occurring in time series j.
4. Run any clustering algorithm that is applicable to text retrieval

   In this work I use hierarchical agglomerative clustering with the Euclidean distance as a similarity measure. As a linkage criterion I use the variance of the clusters that are being merged.

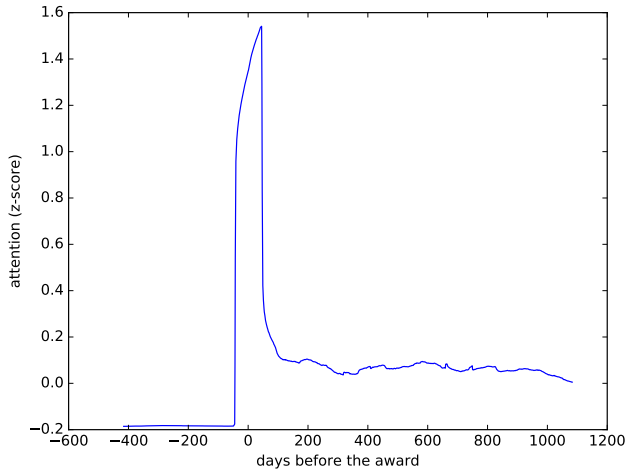


Figure 5.3.: Attention pattern in the manually labeled cluster. This pattern is used to evaluate the clustering algorithms. During the evaluation I examine how good the clustering algorithm can group time series with this pattern in one cluster.

5.4.5. Evaluation

Since there is no ground truth available for the given dataset, a manual labeling is necessary. Due to the fact that clustering of the time series data is a complex task even for a human, the manual data labeling is a challenging problem. Despite this challenge, one dominant cluster from all the time series data was distinguished. This cluster contains patterns similar to the pattern shown in Figure 5.3. Table Table 5.5 shows how good the SUSh algorithm can group time series with this pattern in one cluster. For the evaluation, the time series of scientists' Wikipedia page views were used. As one can see, the performance of the SUSh algorithm for clustering time series of Wikipedia page views of scientists is poor. The accuracy of the algorithm is 0.6 and the F-score is 0.59 which is close to a random labeling. One reason for the poor performance of the algorithm could be the general assumption behind the algorithm. As was mentioned above, SUSh clustering ignores the global shape of the time series and focuses on the local subsequences. This approach works good for the datasets that were used in the previous researches [UBK15; ZMK12] since they have a similar global shape. They used, for example, data sets of heart rate time series or time series of electricity consumption. In this thesis, the dynamics of attention reflected by Wikipedia Views and Google Trends is very different for every scientist. Therefore, ignoring the global shape of the time series is not appropriate.

As an alternative to SUSh clustering algorithm, the BOP clustering algorithm was used. As it was said above, the assumption for this algorithm is the opposite of the SUSh clustering. It considers the global shape of time series instead of local subsequences. The performance of the BOP clustering was evaluated with the same method as the performance of the SUSh clustering. The results of the evaluation are presented in the Table 5.6. One can see that

		Clustered	
		True	**False**
Actual	**True**	79	36
	False	68	79

Table 5.5.: Evaluation of the SUSh clustering algorithm. The accuracy of the clustering is 0.6, precision is 0.53, recall is 0.68, and F-score is 0.59.

		Clustered	
		True	**False**
Actual	**True**	92	23
	False	35	112

Table 5.6.: Evaluation of the BOP clustering algorithm. The accuracy of the clustering is 0.78, precision is 0.72, recall is 0.8 and F-score is 0.82.

accuracy of the clustering improved from 0.6 to 0.78 and F-score from 0.59 to 0.78. Recall that these numbers reflect how good the algorithm can group time series with a pattern similar to Figure 5.3 in an individual cluster. In the following I use BOP clustering as a method to cluster Wikipedia page views, Wikipedia edits, and Google Trends time series.

6. Results

In this chapter I present the results of the analysis of attention dynamics towards prominent scientists and their related topics.

6.1. Agreement Between the Different Web Signals

In order to understand if the different Web signals reveal similar information, I performed a correlation analysis of the time series of different Web signals. The methodology of the analysis is described in Section 5.1. The results of the analysis are presented in Table 6.1.

Table 6.1 shows that the highest correlation is observed inside the seed dataset between the Web signals Wikipedia page views and Google Trends. The CCF between them is 0.516. As time series with the CCF of around 0.5 have similar patterns along the timeline, this correlation coefficient can be interpreted as high. Figure 5.1 illustrates the example of the time series that correlate with the CCF of 0.56. An explanation of a high correlation between the Wikipedia page views and Google Trends is intuitive. Since Google ranks Wikipedia articles high in the search results, people who search for a general information about the scientist will likely visit the Wikipedia article about him. At the same time, the correlation between the topic datasets for those two signals is much lower. That could be explained by a high diversity of the topics and the fact that the attention towards a scientific concepts is driven by many factors. When people are searching for a person on Google, they are more likely to look for general information about him. On the other hand, when somebody is searching for a scientific topic, the demanded information could be more specific than the limited description from the Wikipedia article. For example, an international conference in molecular biology could increase the amount of search requests to Google for a particular topic, but not necessary will increase the number of Wikipedia page views of the article about this topic, as the participants of the conference would probably search for a more specific information than Wikipedia could provide.

The dataset of Wikipedia edits is sparse and does not incorporate a lot of information about the public attention. Therefore, Web signals correlate with the Wikipedia edits weakly. Nevertheless, there is an exception. The CCF between the Wikipedia views and Wikipedia edits in the seed dataset is 0.482. However, for baseline data it is just 0.246. Such a big difference between seed and baseline datasets could be explained by the nature of the seed data. When the scientist got the prize, his biography article should be updated. Therefore, an increase of the article edits could be observed after the scientist got the prize. For example, Figure 6.1 shows the cross correlation between the Wikipedia views and Wikipedia edits of the article about John Tate, American mathematician who received the Abel Prize in 2010. The CCF in thos case is 0.975. This relation cannot be observed in the baseline data, as there

Datasets	**Cross-Correlation Coefficient**	
	Seed	Baseline
Datasets of the Scientists		
Wikipedia Views / Google Trends	0.516	0.505
Wikipedia Views / Wikipedia Edits	0.482	0.246
Wikipedia Edits / Google Trends	0.234	0.165
Datasets of the Topics		
Wikipedia Views / Google Trends	0.305	0.301
Wikipedia Views / Wikipedia Edits	0.139	0.137
Wikipedia Edits / Google Trends	0.227	0.134

Table 6.1.: The table presents the absolute values of averaged cross-correlation coefficient between the following Web signals: Wikipedia page views, Wikipedia edits, and Google Trends. The seed dataset of the scientists contains the Web signals of the scientists who received a prize. The baseline dataset of the scientists contains the Web signals of the scientists who worked at the same time and in the same field as the prize winners, but who did not receive an award. The datasets of the topics contain the research topics of the prize winners or the baseline scientists. The table shows that there is a high correlation between Wikipedia page views and Google Trends for the seed and baseline datasets. The correlation between Wikipedia page views and Wikipedia edits is high for the seed dataset of scientists.

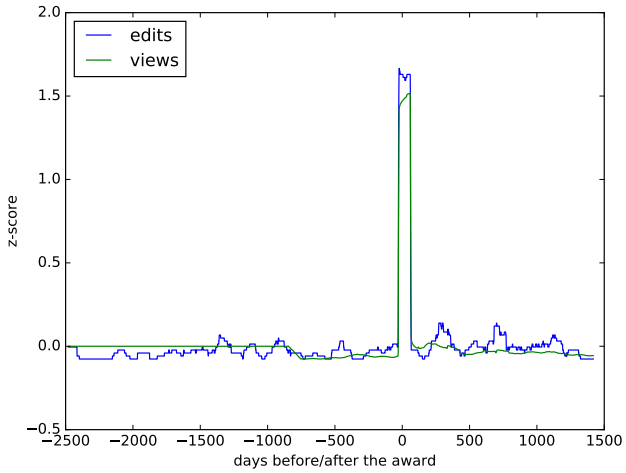


Figure 6.1.: The figure shows the time series of the Wikipedia page views and Wikipedia edits of the article about the mathematician John Tate. One can see, that the number of views and edits peaks after the Abel Prize was received. The cross-correlation coefficient between the Web signals is 0.975.

is no event that would drive the increase of views and edits at the same time. Yet, events such as a death of the scientist could trigger such a behavior. However, they are not frequently observed in the analyzed dataset.

6.2. Topic Analysis

To study the differences of attention dynamics towards topics of scientists that got an award compared to topics of the ones who did not, the attention patterns towards the scientific topics were analyzed. In Section 3.3, the topics were defined as the incoming or outcoming Wikipedia links of the Wikipedia article about the scientist. To ensure that the links are related to the scientific topics, they were automatically filtered. The algorithm of the filtering and its evaluation is described in Section 3.3. For the topic analysis I compared the time intervals between the dates when the articles about the scientists and the articles about the related topics were created. The steps of the analysis are described in Section 5.2. The results of the analysis are presented in Figure 6.2 and Figure 6.3.

Figure 6.2 shows the probability distribution of the amount of topic articles created at each time interval. The zero point on the x-axis refers to the day when the article about the scientist was created. If the topic article was created before the article about the scientist, then it lays in the negative interval of x-axis. If it was created after the article about the scientist, then it lays in the positive interval. The values on the y-axis are normalized by the probability density (i.e. the integral of the histogram sums up to 1).

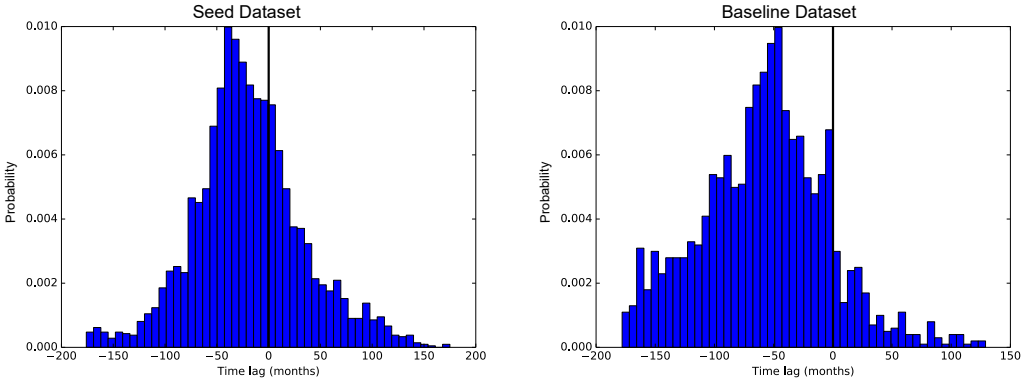


Figure 6.2.: The figure shows the probability distribution of a topic article to be created at each time interval before or after the creation of the article about a scientist. The seed dataset contains the prize winners and their research topics. The baseline dataset contains the scientists who did not recieve a prize and their research topics. The zero point on the x-axis corresponds to the day when the article about the scientist was created. If the topic article was created before the article about the scientist, then it lays in the negative interval of x-axis. If it was created after the article about the scientist, then it lays in the positive interval. One can see that the highest probability of the topic article to be created is around 50 weeks before the article about the scientist. However, for the seed dataset the topic articles were created closer to the 0 point in comparison to the baseline dataset.

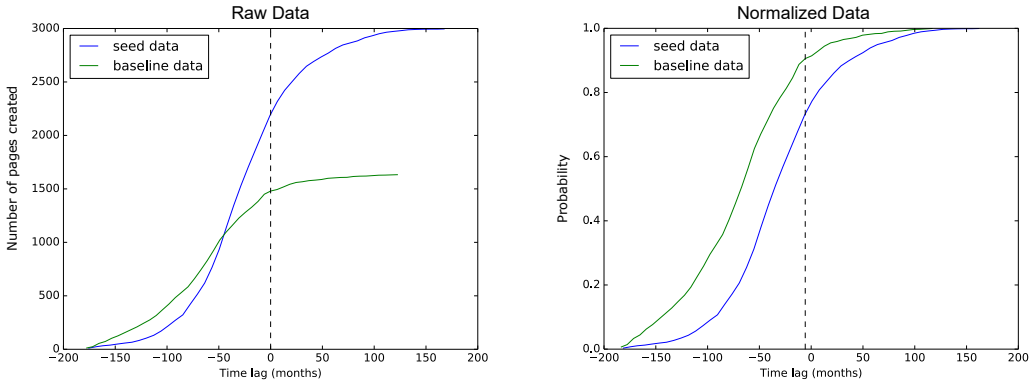


Figure 6.3.: Cumulative probability distribution of the amount of topic articles created at each time interval before or after the creation of the article about a scientist. The left plot presents raw data. On the right plot the data is normalized by the total number of the topic articles. The seed dataset contains the prize winners and their research topics. The baseline dataset contains the scientists who did not receive a prize and their research topics. The zero point on the x-axis corresponds to the day when the article about the scientist was created. If the topic article was created before the article about the scientist, then it lays in the negative interval of x-axis. If it was created after the article about the scientist, then it lays in the positive interval. The cumulative distributions show that the probability distribution of the baseline data grows faster before the 0 point. That means that there is a higher probability for the topic articles to be created before the scientist article. Moreover, around the 0 point, the baseline cumulative distribution is relatively flat, whereas the seed cumulative distribution changes rapidly. Thereby, one could conclude that the topic articles from the seed dataset were created around the date when the article about the scientist was created.

For example, the highest probability of the topic article to be created is around 50 weeks before the article about the scientist. This is true for both datasets. However, for the seed dataset the topic articles were created closer to the 0 point in comparison to the baseline dataset. Most of the topic articles from the baseline dataset were created around 50 weeks before the article about the scientist and there are almost no topic articles created after the scientist article. The cumulative distributions that presented on Figure 6.3 show that the probability distribution of the baseline data grows faster before the 0 point. That means that there is a higher probability for the topic articles to be created before the scientist article. Around 90% of the topic articles from the baseline dataset were created before the article about the scientist. At the same time, the cumulative probability distribution for the seed dataset grows more slowly in the beginning. Only 75% of topics articles were created before the 0 point. Moreover, the slopes of the curves are different. Around the 0 point, the baseline cumulative distribution is relatively flat, whereas the seed cumulative distribution changes rapidly. Thereby, one can conclude that the topic articles from the seed dataset were created around the date when the article about the scientist was created. The possible interpretation is that the topics are more closely related to the scientist who got an award. This supports that scientists who got the prize introduced the topics to the public.

Considering these two patterns, it is interesting to compare the public attention towards the topics from these two groups. Since there is no "winning date" for the baseline data, it is reasonable to compare the attention dynamics towards the topics before and after the article about the scientist was created, as this event exists in the both datasets. For this, a trend analysis was performed. The methodology of the trend analysis is described in Section 5.3.

Figure 6.4 presents the attention trends towards the topics before and after the page about the scientist was created. I used the Wikiedia page views dataset as a Web signal that reflects the public attention. Figure 6.4 shows that there are similar attention trends towards topics from the seed and baseline datasets for the period before the article about the scientist was created. After the page about the scientist was created, the attention towards the topics from the seed dataset has more often a growing trend than the topics from the baseline. 54% of the topics from the seed dataset have a decreasing trend and 40% increasing. At the same time, 62% of the topics from the baseline dataset have a decreasing trend and 30% an increasing. It could be explained by the fact that in the seed dataset, the period after the creation of the article about the scientist includes the day when the scientist got the award. The popularity of the topic was affected by the popularity of the scientist. Therefore, the uprising trend for the scientific topics from the seed dataset can be observed more often than for the topics from the baseline dataset. Thus, one could conclude that the topics of the scientists who got the award are more interesting for the public than the topics of the scientist who did not get the prize.

6.3. Attention Dynamics towards Prominent Scientists

To understand how the attention dynamics vary between scientists from the different disciplines and prizes, a clustering analysis was performed. I analyzed the Wikipedia page views statistics of the articles about the scientists who got an award. The analyzed period includes Wikipedia page views statistics between 3 years before and 1 year after the scientists got the

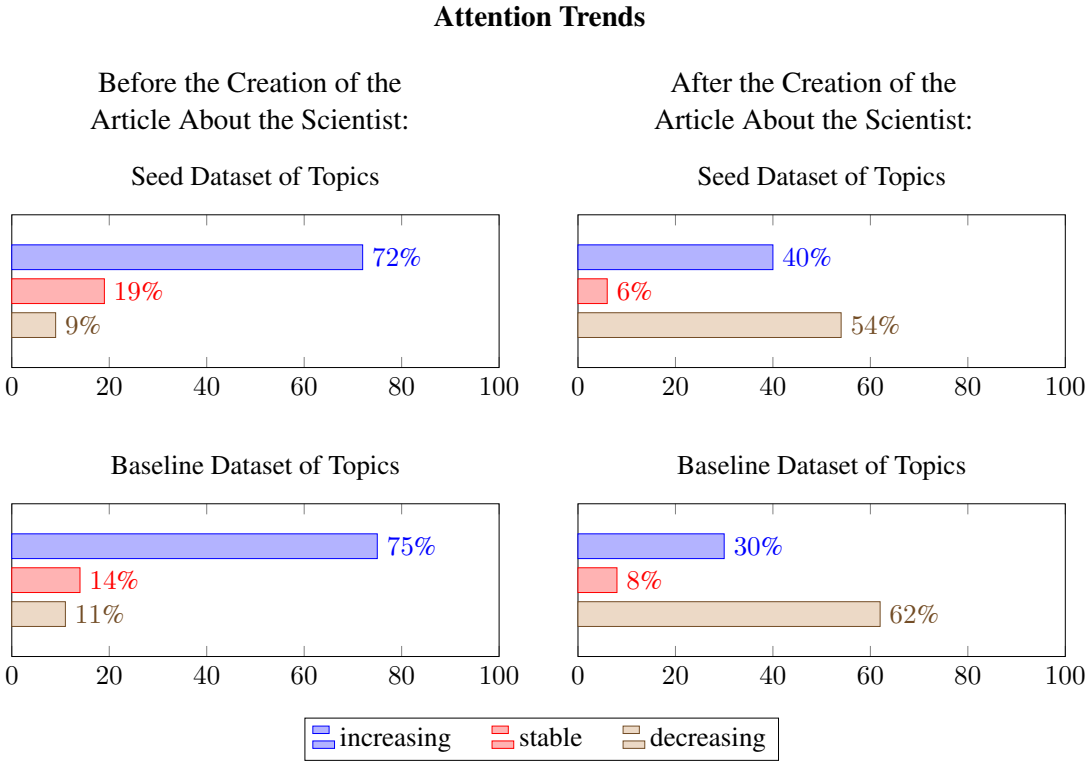


Figure 6.4.: The figure shows the attention trends towards the scientific topics before and after the page about the scientist was created. The attention is reflected by the Wikipedia page views of the articles about the topics. The seed dataset of topics consists of the research topics of the prize winners. The baseline dataset of topics consists of the research topics of the scientists who did not receive a prize. One can see that there are similar attention trends towards topics from the seed and baseline datasets for the period before the article about the scientist was created. After the page about the scientist was created, the attention towards the topics from the seed dataset has more often a growing trend than the topics from the baseline. It could be explained by the fact that in the seed dataset, the popularity of the topic was affected by the popularity of the scientist.

Cluster	Size
Cluster 1	91
Cluster 2	113
Cluster 3	58

Table 6.2.: The size of the clusters. The clustering was performed over the piecewise normalized time series of Wikipedia page views about the scientists.

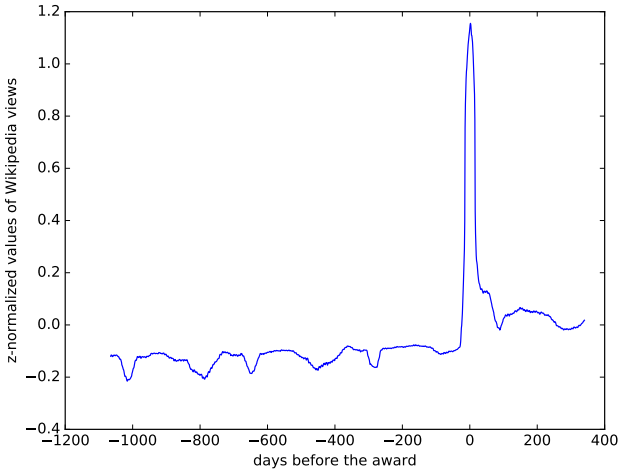


Figure 6.5.: Attention pattern inside the first cluster. The plot presents the mode of z-normalized time series from the first cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views declines. The attention pattern of the time series in the first cluster is similar to the pattern inside the second cluster.

award. For the analysis, the BOP clustering algorithm was used. The methodology of the BOP clustering is described in Section 5.4.4. A preliminary analysis showed that the method of the data normalization strongly affects the clustering results. First, I will present the results for the algorithm that was applied on the piecewise normalized data.

6.3.1. Time Series Clustering Using the Piecewise Normalization

For the piecewise normalization I used Equation (5.7) by modifying it such that μ and σ are the mean and variance of subsequence $S \subset T$ where S is a sliding window and its length is 360. Based on the parameters I used (see Appendix A.6), the data was separated into three clusters. The size of the clusters is presented in Table 6.2 and the attention patterns inside the clusters in Figure 6.5, Figure 6.6, and Figure 6.7.

One can see that the patterns of the three clusters are similar. There is an attention spike around the winning date after which the number of the Wikipedia page views declines to

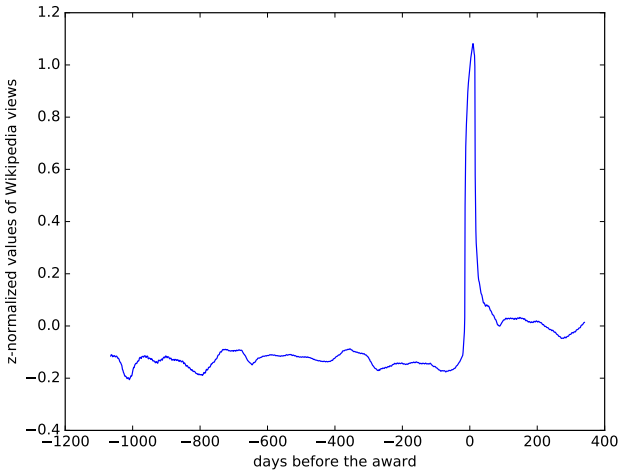


Figure 6.6.: Attention pattern inside the second cluster. The plot presents the mode of z-normalized time series from the second cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views declines. The attention pattern of the time series in the second cluster is similar to the pattern inside the first cluster.

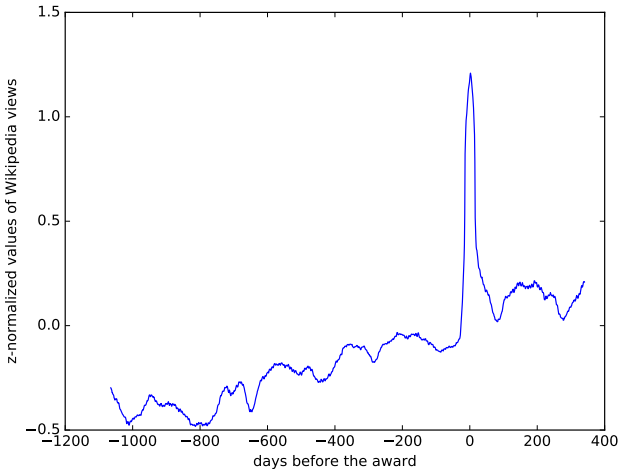
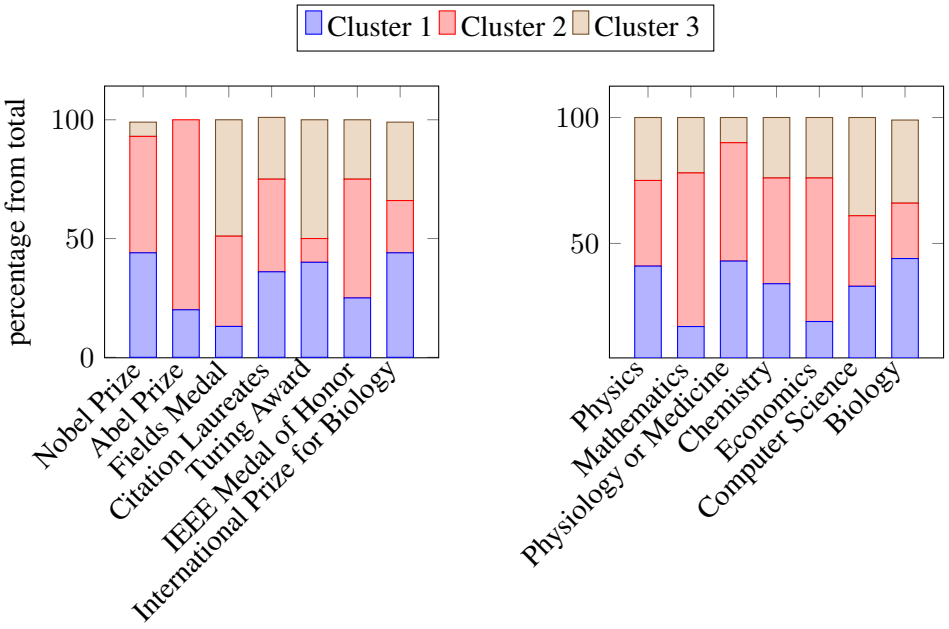


Figure 6.7.: Attention pattern inside the third cluster. The plot presents the mode of z-normalized time series from the third cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views continues its grows with the same dynamics as before a prize was awarded. The attention pattern of the time series in this cluster differs from the patterns inside the first and the second clusters.



(a) Distribution of prize winners between the clusters. The numbers are calculated as a percentage from the total number of scientists who got each award. The algorithm separated almost all the Abel Prize winners in the second cluster. Nevertheless the result is not representative due to the small number of Abel Prize winners in the dataset (see Appendix A.5).

(b) Distribution of disciplines between the clusters. The numbers are calculated as a percentage from the total number of scientists from each discipline. One can see that scientists from the different disciplines distributed between the clusters evenly. The algorithm could not find specific attention patterns towards the scientists from the different disciplines.

Figure 6.8.: Distribution of scientists from each discipline or prize in the clusters. The time series were normalized piecewise. The clustering algorithm did not separate the scientists from the different disciplines or prizes into individual clusters.

its initial values. The exception is the Cluster 3, where the attention towards the scientists continues its growth after the winning date. Figure 6.8 shows that the clustering algorithm did not separate the scientists from the different disciplines or prizes into individual clusters. Nevertheless, it can be observed that most of the Nobel Prize and Abel Prize winners are segregated into the Cluster 2. However, since this cluster is the largest cluster, the probability of a random scientist belonging to the Cluster 2 is the highest. Moreover, since the distribution of the scientists between the different prizes and fields is not even, it is important for the analysis to look at the total number of the prize winners for each award and discipline that we have in our dataset. For example, the Turing Award and the Fields Medal were received only by a small number of scientists. The number of the scientists in the different disciplines is shown in Appendix A.4 and the number of awards in Appendix A.5.

Cluster	Size
Cluster 1	96
Cluster 2	20
Cluster 3	146

Table 6.3.: The size of the clusters. The clustering was performed over the time series of Wikipedia page views about the scientists. The time series were fully normalized, without using sliding window.

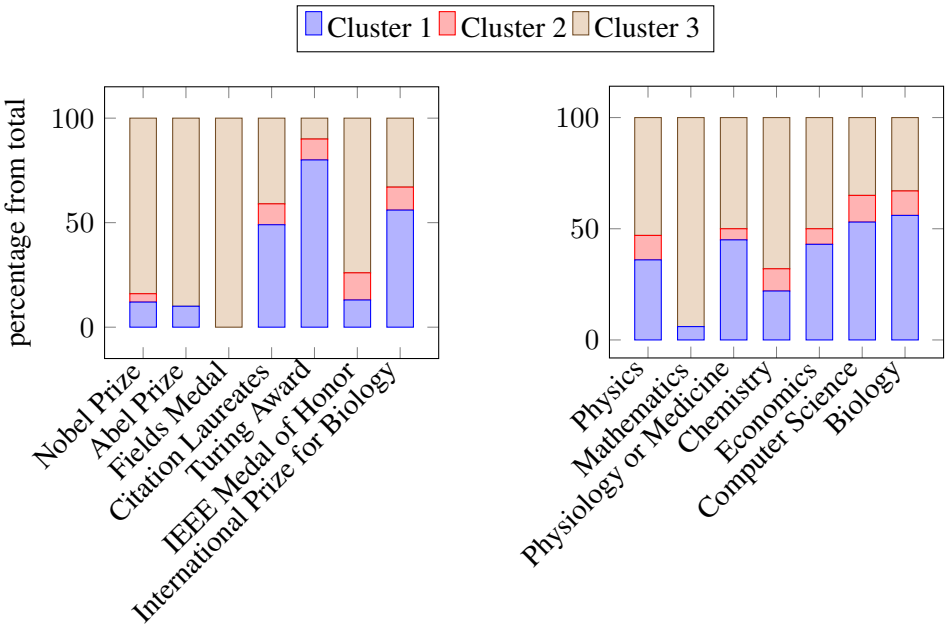
6.3.2. Time Series Clustering Using the Full Time Series Normalization

As stated above, the time series normalization strongly affects the clustering results. This can be explained by the fact that due to the attention spike around the winning day, the mean values calculated over the subsequences of the time series differ dramatically from the mean value of the full time series. Therefore, I performed the clustering analysis over the time series that were normalized without the usage of the sliding window Equation (5.7). The cluster sizes are presented in Table 6.3. Figure 6.9 demonstrates the distribution of scientists from different disciplines or who received different prizes.

Figure 6.9 shows that Nobel Prize, Abel Prize, and Fields Medal winners were grouped into Cluster 3. However, there is no clear separation of the disciplines between the clusters. Figure 6.10, Figure 6.11, and Figure 6.12 show the patterns inside the clusters. One can see that Cluster 1 and Cluster 2 have similar attention patterns, whereas cluster 3 has a smoother pattern. In Cluster 3, the attention dynamics before the award date do not fluctuate as strong as in Cluster 1 or Cluster 2. Moreover, there is no increasing trend of attention before the award was announced. Therefore, it could be assumed that Nobel Prize, Abel Prize, and Fields Medal winners do not get a lot of attention before the award and lose their popularity very fast. However, Figure 6.12 shows the normalized attention dynamics. Therefore, it is not clear if a smoother pattern can be observed due to the difference in the attention dynamics, or due to the normalization that made the time series before the spike look smoother.

To check this hypothesis, an additional analysis needed. In order to compare the attention dynamics before and after the scientist got the award, a trend analysis was performed. The advantage of the trend analysis is that it does not require the data to be normalized. This way, the previously observed effect of the normalization can be avoided.

Firstly, the group of scientists for the analysis should be defined. During the clustering analysis, Nobel Prize, Abel Prize, and Fields Medal winners were grouped into the individual cluster. There are 77 Nobel Prize winners, 10 Abel Prize winners, and 8 Fields Medal winners in the original dataset (see Appendix A.4). Based on the majority of the Nobel Prize winners in the cluster, it is reasonable to focus the analysis on them. During the trend analysis I compare the attention dynamics towards the Nobel Prize winners with the scientists who were nominated to the Nobel Prize, but did not win. To define the Nobel Prize nominees I used the Citation Laureates of Thompson Reuters. The purpose of the Citation Laureates of Thompson Reuters is to predict the Nobel Prize winners based on the citation analysis. The list of the Citation Laureates is created annually for each of the four disciplines: physics, chemistry,



(a) Distribution of prize winners between the clusters. The numbers are calculated as a percentage from the total number of scientists who got each award. One can see, that Nobe Prize winners, Abel Prize winners, and Fields Medal winners were grouped into an individual cluster.

(b) Distribution of disciplines between the clusters. The numbers are calculated as a percentage from the total number of scientists from each discipline. One can see, that mathematicians were grouped in the second cluster.

Figure 6.9.: Distribution of prize winners between the clusters. A full time series normalization without using sliding window was performed before the clustering.

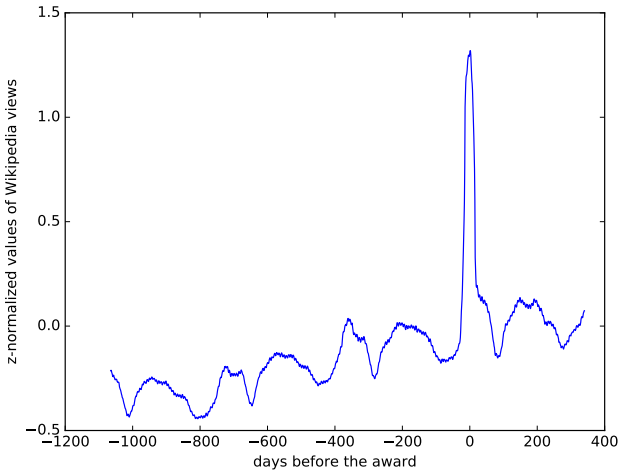


Figure 6.10.: Attention pattern inside the first cluster. The plot presents the mode of z-normalized time series from the first cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views continues its growth with the same dynamics as before a prize was awarded.

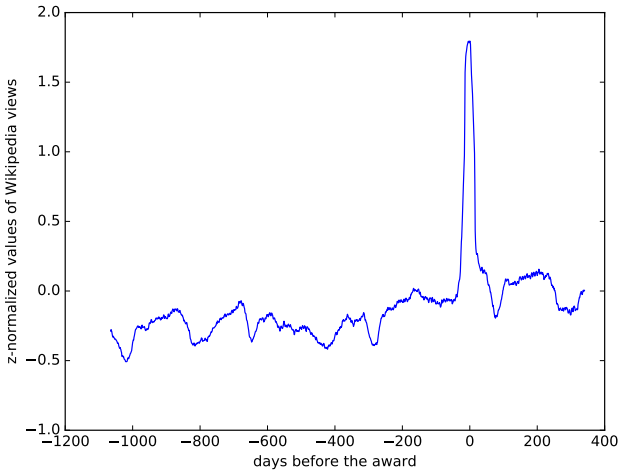


Figure 6.11.: Attention pattern inside the second cluster. The plot presents the mode of z-normalized time series from the second cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views keeps the same dynamics as before a prize was awarded.

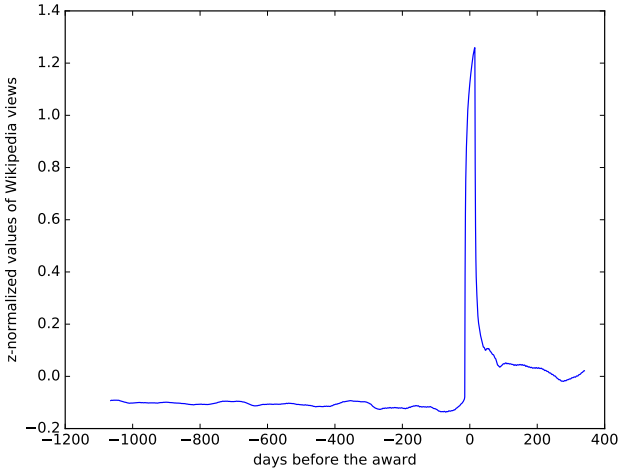


Figure 6.12.: Attention pattern inside the third cluster. The plot presents the mode of z-normalized time series from the third cluster. There is an attention spike around the winning date after which the number of the Wikipedia page views goes down. The attention pattern inside the third cluster is more smooth than in the first and second clusters. Additional analysis needed to investigate whether a smoother pattern can be observed due to the difference in the attention dynamics, or due to the normalization that made the time series before the spike look smoother.

medicine, and economics. This way, the analyzed dataset consist of the Nobel Prize winners and candidates for each year between 2008 and 2015 in the four disciplines listed above. To eliminate the effect of the attention spike around the day of the Nobel Prize announcement, the period between one week before and one month after the announcement was ignored. Apart from that, the period from three years before until three years after the winning day was considered. The methodology of the trend analysis is described in Section 5.3. The aim of the trend analysis is to observe whether there is an increasing, decreasing, or no trend in the period before and after the award was announced. The results of the analysis are presented in Figure 6.13. It shows that scientists in the group of Nobel candidates show more often an increasing trend in attention before the prize was awarded in comparison to the actual Nobel Prize winners. 73% of the Nobel Prize candidates and 58% of the Nobel Prize winners have an increasing attention trend before the award was announced. Moreover, 29% of the Nobel Prize winners have stable attention trend before the award, against only 18% of the candidates. After the prize was awarded, the group of Nobel Prize winners more often has a decreasing trend than the Nobel candidates. In the period from 1 month after the Nobel Prize announcement until 3 years after, 71% of Nobel Prize winners and 44% candidates have a decreasing trend. Therefor, Nobel candidates more often show increasing attention dynamics after the award announcement. 35% of the candidates and 22% of the actual prize winners have an increasing trend after the award was announced. The results suggest that the decision of the Nobel committee does not necessarily reflect the current attention towards the scientist. This way, the hypothesis that in most of the cases Nobel Prize winners were not expected by the public and lose their popularity after getting the award can be confirmed.

6.4. Prize Prediction

Earlier I performed the clustering analysis of the time series for the period between three years before and one year after the prize announcement. This way, Nobel Prize, Abel Prize, and Fields Medal winners were grouped into a single cluster. Now I want to check if it is possible to group the prize winners into an individual cluster based on the attention dynamics before the award was announced. For this goal I performed a clustering over the dataset of Nobel Prize winners and candidates for the period between three years before and one week before the award announcement. The evaluation of the clustering is presented in Table 6.4.

The goal of this clustering analysis is to predict whether a scientist will get the award based on the attention he got from the public. Table 6.4 shows that the clustering algorithm separated the data into 2 clusters with the sizes 177 and 37. Despite the high recall, the precision, accuracy and F-score are low. The evaluation shows that it is difficult to predict the Nobel Prize winners based on the attention dynamics reflected by the Wikipedia page views of the articles about the scientists.

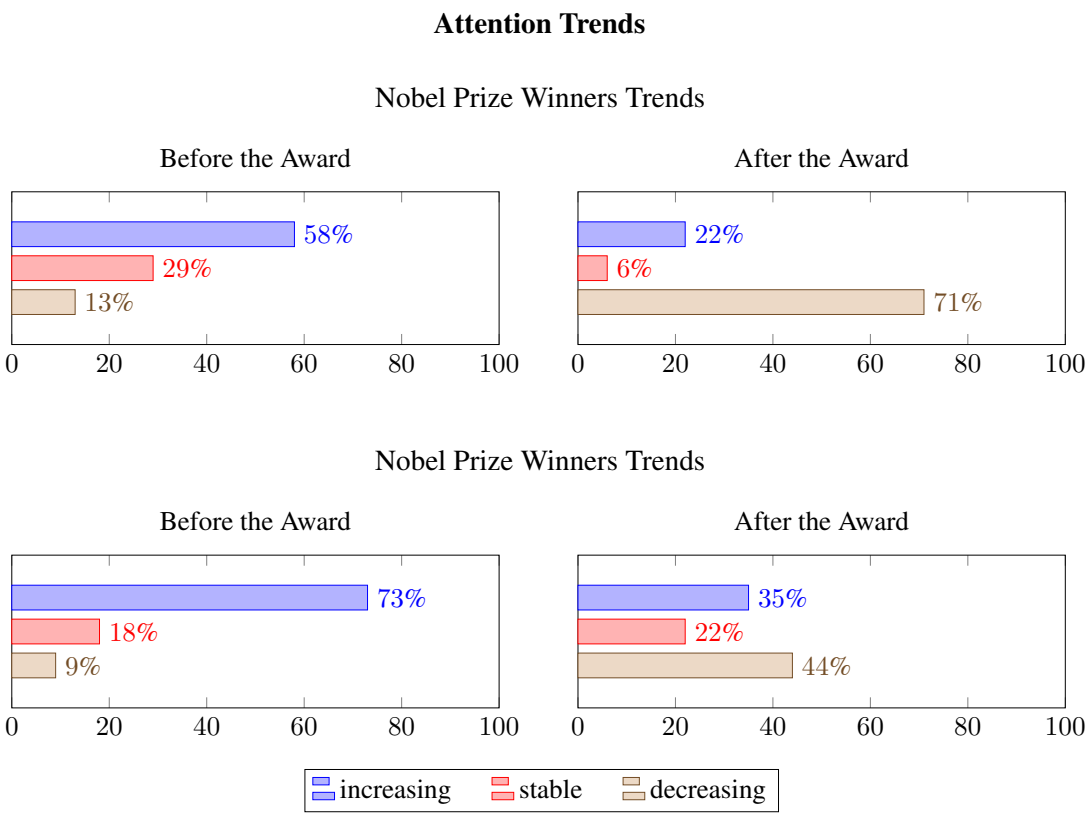


Figure 6.13.: Attention trends before and after the Nobel Prize announcement inside the groups of the Nobel Prize candidates and winners. The Nobel Prize candidates are represented by the list of Citation Laureates of Thompson Reuters. The period between one week before and one month after the announcement was ignored. Apart from that, the period from three years before until three years after the winning day was considered. One can see, that the group of Nobel candidates more often demonstrates an increasing trend in attention before the prize was awarded in comparison to the actual Nobel Prize winners. After the prize was awarded, the group of Nobel Prize winners more often has a decreasing trend than the Nobel candidates. Nobel candidates more often show increasing attention dynamics after the award announcement. Therefore, one can say that, the decision of the Nobel committee does not necessarily reflect the current attention towards the scientist.

		Clustered	
		True	**False**
Actual	**True**	76	1
	False	101	36

Table 6.4.: The table shows the evaluation of the clustering performed over the time series of Nobel candidates and winners before the award announcement. The Nobel Prize candidates are represented by the list of Citation Laureates of Thompson Reuters. The accuracy of the clustering is 0.52, precision is 0.43, recall is 0.99 and F-score is 0.6. The evaluation shows that it is difficult to predict the Nobel Prize winners based on the attention dynamics reflected by the Wikipedia page views of the articles about the scientists.

7. Conclusion

7.1. Discussion

This research studied to what extent scientific achievements are reflected in online attention towards the scientists. The following questions were addressed:

1. Is the success of a scientist determined by the field he or she is working in or was the popularity of the field influenced by the scientist?

   To answer this question, I performed an analysis of the time lag between the creation date of the Wikipedia article about the scientist and the Wikipedia articles about the topics he was working on (see Section 5.2). The analysis showed that the topic articles of the prize winners were created around the date when the article about the scientist was created. The findings suggest that the research topics of the prize winners are more related to them than topics of the scientists who did not receive an award. Extending this idea, it could be said that scientists who received a prize introduced their research topics to the public.

   A trend analysis of the public attention towards the research topics of prize winners and scientists who did not receive a prize showed that the topics of scientists who received an award are more interesting for the public than the topics of scientists who did not receive a prize (see Section 6.2).

2. How does the public react to the success of a scientist?

   A trend analysis of the attention dynamics towards Nobel Prize winners and nominees showed that Nobel Prize winners get a lower amount of attention before receiving the prize than nominees. Moreover, after receiving the award, the attention towards Nobel Prize winners is going down faster then the attention towards the nominees (see Section 6.3). This means that from one hand, Nobel Prize winners were not expected by the public and from another hand, the public looses its interest in the prize winners after the the Nobel Prize announcement.

3. Can we predict the future success of a scientist based on the dynamics of the public attention towards him/her?

   To answer this question, a clustering analysis of Nobel Prize winners and possible candidates suggested by Thomson Reuters Citation Laureates was performed (see Section 6.4). The analysis showed a high recall of 0.99. Yet, the accuracy of the clustering is only 0.52 and the F-score is 0.6. This shows that it is difficult to predict the prize winners based on the collected Web signals and the clustering algorithms that were applied.

The main contribution of the research is twofold. First, the research revealed the interrelation between the success of a scientist and success of the field he is working in. The presented methods can be generalized to investigate how an information network reacts on an event and how the attention spreads from the original subject to the related topics.

Second, this study presents the relations between the online attention towards scientist and his scientific achievements. Nowadays, the citation analysis is losing its power as an instrument of evaluating the scientific contribution of scholars [GW10]. The previous research [SY14; GW10] raised a question on which metrics to use to estimate the contribution of the scientist. Therefore, new metrics to justify the scientific and social value of the scientist are needed. From this perspective, the online attention can be seen as an indicator of the academic visibility and can be used by the general public and funding agencies to estimate the scientific and social contribution of the scientist.

7.2. Limitations

First, the research is limited by the choice of Web signals that were analyzed. Recall that Wikipedia page views, Wikipedia edits, and Google Trends were chosen as the Web signals that reflect the dynamics of a public attention. The analysis discovered that the dataset of Google Trends strongly correlates with the dataset of Wikipedia page views. At the same time, the dataset of Wikipedia edits can not be used for the analysis due to its sparsity. Thus, the dataset of Wikipedia page views was primarily used for the analysis. However, this dataset contains a bias based on the fact that the viewers of Wikipedia do not represent the true population. This problem is inherent to the audience of any online media and could be tackled either by a careful choice of the analyzed media or by a high diversity.

Another source of a bias could result from the research methods that were used. For example, the clustering algorithms that were used in this study are very sensitive to the way of data normalization. This problem was described in Section 6.3 and addressed by using different normalization approaches. Nevertheless, it could be appropriate to use other clustering algorithms that are less sensitive to the time series normalization.

The chosen evaluation technique of the clustering algorithm could be another source of a bias. The evaluation method was described in Section 5.4.5. Recall that a manual labeling of the data was performed to obtain a ground truth. Therefore, an estimation of the clustering quality fully relies on the quality of the manual labeling, which could be biased due to the human factor. Nevertheless, since the evaluation of the different clustering algorithms in itself is not a subject of the study, the choice of the evaluation technique can be considered as reasonable.

The data that was analyzed in this research is restricted by the group of successful scientists. The success of the scientists is defined as a fact of being honored with one of the prizes listed in Section 3.1. Nevertheless, one can argue that other measurements of academic success are more relevant. One example is the citation index that was often used in previous research to estimate the academic impact of a scientist [GM68; AO78; GW10; KM04].

Finally, the analyzed data was restricted by the group of scientists and their topics. However, it was not studied whether the results of the analysis could be generalized to the other profes-

sions. In order to generalize the results, the research should be extended to the representatives of the other trades.

7.3. Future Work

One of the possible directions of a future work is the analysis of different information shocks. This research showed that the events such as the Nobel Prize announcement cause an information shock associated with a rapid increase of a public attention towards the scientist who received it and his academic field. The future research can study how different types of information shocks influence the network of related topics.

Alternatively, one can study how attention dynamics before the information shock influences the future attention and how the collective memory shapes the future perception.

Further research is necessary to investigate whether it is possible to predict the future success of a scientist based on the dynamics of the public attention towards him. For this goal, supervised learning techniques could be applied.

Finally, alternative signals of public attention such as the number of tweets can be introduced to the research. This way, the analyzed population gets more diverse and the results of the research are more general.

Appendices

A. Appendix

A.1. Summarization of the Notation

D	A dataset of the time series objects $D = T_1, T_2, \ldots, T_n$
T	A time series $T = t_1, t_2, ..., t_n$
$S_{j,l}$	A subsequence of time series T_i, where l is a length of the subsequence and j is a starting position in time series T_i
$\grave{S}$	An unsupervised shapelet
$\grave{T}$	A PAA representation of time series $\grave{T} = \grave{t}_1, ...\grave{t}_w$
w	A SAX word
A	A SAX alphabet
$\tilde{T}$	A SAX representation of time series $\tilde{T} = \tilde{t}_1, ..., \tilde{t}_w$
$\rho_{X,Y}$	A cross-correlation coefficient between the time series X and Y

A.2. Size of the Datasets

	Baseline	**Seed**
Scientists	262	262
Topics	1070	1911

A.3. Gender Distribution of the Scientists in the Dataset of Prize Winners

Gender	**Number of Scientists**
Male	242
Female	20

A.4. Distribution of the Scientists between the Different Disciplines

Discipline	**Number of Scientists**
Physics	57
Mathematics	18
Physiology and Medicine	58
Chemistry	50
Economics	54
Computer Science	18
Biology	9

A.5. Distribution of the Scientists between the Different Prizes

Name of the Prize	**Number of Scientists**
Nobel Prize	77
Abel Prize	10
Fields Medal	8
Citation Laureates	163
Turing Award	10
IEEE Medal of Honor	8
International Prize for Biology	9

A.6. Parameters of the Clustering Algorithm

Notation	**Meaning**	**Value**
n	Size of the sliding window	360
w	Length of the SAX-word	9
A	Alphabet size	4

References

[AH10] Sitaram Asur and Bernardo A Huberman. „Predicting the future with social media“. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Volume 1. IEEE. 2010, pages 492–499 (cited on page 8).

[Aks06] Dag W Aksnes. „Citation rates and perceptions of scientific contribution“. In: *Journal of the American Society for Information Science and Technology* 57.2 (2006), pages 169–185 (cited on page 1).

[AO01] Cláudia M Antunes and Arlindo L Oliveira. „Temporal data mining: An overview“. In: *KDD workshop on temporal data mining*. Volume 1. 2001, page 13 (cited on page 17).

[AO78] Susan V Ashton and Charles Oppenheim. „A method of predicting Nobel prizewinners in chemistry“. In: *Social Studies of Science* 8.3 (1978), pages 341–348 (cited on pages 6, 52).

[Ask15] Nikos Askitas. „Trend-Spotting in the Housing Market“. In: (2015) (cited on page 12).

[BC94] Donald J Berndt and James Clifford. „Using Dynamic Time Warping to Find Patterns in Time Series.“ In: *KDD workshop*. Volume 10. 16. Seattle, WA. 1994, pages 359–370 (cited on page 17).

[BJZ03] AJ Bagnall, GJ Janacek, and M Zhang. „Clustering time series from mixture polynomial models with discretised data“. In: (2003) (cited on pages 15, 16).

[CF99] Kin-Pong Chan and Ada Wai-Chee Fu. „Efficient time series matching by wavelets“. In: *Data Engineering, 1999. Proceedings., 15th International Conference on*. IEEE. 1999, pages 126–133 (cited on page 15).

[CFM15] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. „The production of information in the attention economy“. In: *Scientific reports* 5 (2015) (cited on page 7).

[CFY03] FK-P Chan, AW-C Fu, and Clement Yu. „Haar wavelets for efficient similarity search of time-series: with and without time warping“. In: *IEEE Transactions on knowledge and data engineering* 15.3 (2003), pages 686–705 (cited on page 15).

[Cha+02] Kaushik Chakrabarti et al. „Locally adaptive dimensionality reduction for indexing large time series databases“. In: *ACM Transactions on Database Systems (TODS)* 27.2 (2002), pages 188–228 (cited on pages 15, 26).

[Din+08] Hui Ding et al. „Querying and mining of time series data: experimental comparison of representations and distance measures“. In: *Proceedings of the VLDB Endowment* 1.2 (2008), pages 1542–1552 (cited on pages 14, 17, 18).

[EA12] Philippe Esling and Carlos Agon. „Time-series data mining“. In: *ACM Computing Surveys (CSUR)* 45.1 (2012), page 12 (cited on pages 14–18, 20).

[FRM94] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. *Fast subsequence matching in time-series databases*. Volume 23. 2. ACM, 1994 (cited on page 15).

[Gil87] Richard O. Gilbert. *Statistical methods for environmental pollution monitoring*. John Wiley & Sons, 1987 (cited on page 23).

[GM68] Eugene Garfield and Morton V Malin. „Can Nobel Prize winners be predicted“. In: *135th meetings of the American Association for the Advancement of Science, Dallas, TX*. Citeseer. 1968 (cited on pages 6, 52).

[Goe+10] Sharad Goel et al. „Predicting consumer behavior with Web search“. In: *Proceedings of the National academy of sciences* 107.41 (2010), pages 17486–17490 (cited on page 8).

[Gru+05] Daniel Gruhl et al. „The predictive power of online chatter“. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pages 78–87 (cited on page 8).

[GW10] Yves Gingras and Matthew L Wallace. „Why it has become more difficult to predict Nobel Prize winners: a bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901–2007)“. In: *Scientometrics* 82.2 (2010), pages 401–412 (cited on pages 1, 6, 52).

[HLN11] Wen-Ru Hou, Ming Li, and Deng-Ke Niu. „Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution“. In: *BioEssays* 33.10 (2011), pages 724–727 (cited on page 1).

[Joh07] Soren Johansen. „Correlation, regression, and cointegration of nonstationary economic time series“. In: *CREATES Research Paper* 2007-35 (2007) (cited on page 20).

[Kap06] David Kaplan. „And the Oscar goes to... a logistic regression model for predicting Academy Award results“. In: *Journal of Applied Economics & Policy* 25.1 (2006), page 23 (cited on page 5).

[Keo+01] Eamonn Keogh et al. „Dimensionality reduction for fast similarity search in large time series databases“. In: *Knowledge and information Systems* 3.3 (2001), pages 263–286 (cited on page 15).

[KGP01] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta. „Distance measures for effective clustering of ARIMA time-series“. In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE. 2001, pages 273–280 (cited on page 16).

[KM04] Romualdas Karazija and Alina Momkauskaite. „The Nobel prize in physics-regularities and tendencies“. In: *Scientometrics* 61.2 (2004), pages 191–205 (cited on pages 6, 52).

[Kra+08] Jonas Krauss et al. „Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis.“ In: *ECIS*. 2008, pages 2026–2037 (cited on page 5).

[Leh+14] Janette Lehmann et al. „Reader preferences and behavior on Wikipedia“. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM. 2014, pages 88–97 (cited on page 7).

[Lia05] T Warren Liao. „Clustering of time series data—a survey“. In: *Pattern recognition* 38.11 (2005), pages 1857–1874 (cited on pages 18–20).

[Lin+03] Jessica Lin et al. „A symbolic representation of time series, with implications for streaming algorithms“. In: *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM. 2003, pages 2–11 (cited on page 15).

[Lin+07] Jessica Lin et al. „Experiencing SAX: a novel symbolic representation of time series“. In: *Data Mining and knowledge discovery* 15.2 (2007), pages 107–144 (cited on pages 24, 25).

[LL09] Jessica Lin and Yuan Li. „Finding structural similarity in time series data using bag-of-patterns representation“. In: *International Conference on Scientific and Statistical Database Management*. Springer. 2009, pages 461–477 (cited on pages 24–26, 28).

[Man45] HB Mann. *Non-Parametric Tests against Trend. Econmetrica, 13, 245-259*. 1945 (cited on page 14).

[MM75] Michael J Moravcsik and Poovanalingam Murugesan. „Some results on the function and quality of citations“. In: *Social studies of science* 5.1 (1975), pages 86–92 (cited on page 1).

[Nas10] Guy Nason. *Wavelet methods in statistics with R*. Springer Science & Business Media, 2010 (cited on page 13).

[Par07] Iain Pardoe. „Predicting Oscar winners“. In: *Significance* 4.4 (2007), pages 168–173 (cited on page 5).

[PM02] Ivan Popivanov and Renee J Miller. „Similarity search over time-series data using wavelets“. In: *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE. 2002, pages 212–221 (cited on page 15).

[RK13] Thanawin Rakthanmanon and Eamonn Keogh. „Fast shapelets: A scalable algorithm for discovering time series shapelets“. In: *Proceedings of the 13th SIAM international conference on data mining*. SIAM. 2013, pages 668–676 (cited on page 16).

[SH10] Gabor Szabo and Bernardo A Huberman. „Predicting the popularity of online content“. In: *Communications of the ACM* 53.8 (2010), pages 80–88 (cited on page 7).

[SK08] Jin Shieh and Eamonn Keogh. „i SAX: indexing and mining terabyte sized time series“. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pages 623–631 (cited on page 17).

[SRT07] Thomas Stiefmeier, Daniel Roggen, and Gerhard Tröster. „Gestures are strings: efficient online gesture spotting and classification using string matching“. In: *Proceedings of the ICST 2nd international conference on Body area networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2007, page 16 (cited on page 15).

[SY14] Anna Samoilenko and Taha Yasseri. „The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics“. In: *EPJ Data Science* 3.1 (2014), pages 1–11 (cited on pages 1, 52).

[UBK15] Liudmila Ulanova, Nurjahan Begum, and Eamonn Keogh. „Scalable clustering of time series with U-shapelets“. In: *SIAM international conference on data mining (SDM 2015)*. SIAM. 2015 (cited on pages 16, 20, 24, 25, 28–31).

[Wel+10] Katrin Weller et al. „Social software in academia: Three studies on users' acceptance of Web 2.0 Services“. In: (2010) (cited on page 2).

[YK11] Lexiang Ye and Eamonn Keogh. „Time series shapelets: a novel technique that allows accurate, interpretable and fast classification“. In: *Data mining and knowledge discovery* 22.1-2 (2011), pages 149–182 (cited on pages 15, 16, 20).

[ZMK12] Jesin Zakaria, Abdullah Mueen, and Eamonn Keogh. „Clustering time series using unsupervised-shapelets“. In: *2012 IEEE 12th International Conference on Data Mining*. IEEE. 2012, pages 785–794 (cited on pages 24, 31).