Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia

Adam Jatowt, Daisuke Kawai and Katsumi Tanaka

Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
{adam, tanaka}@dl.kuis.kyoto-u.ac.jp
daisuke@gauge.scphys.kyoto-u.ac.jp

ABSTRACT
Wikipedia is the result of a collaborative effort aiming to represent human knowledge and to make it accessible for everyone. As such it contains lots of contemporary as well as history-related information. This research looks into historical data available in Wikipedia to explore its various time-related characteristics. In particular, we study Wikipedia articles on historical persons. Our analysis sheds new light on the characteristics of information about historical persons in Wikipedia and quantifies user interest in such data. We use signals derived from the hyperlink structure of Wikipedia as well as from article view logs and we overlay them over temporal dimension to understand relations between time, link structure and article popularity. In the latter part of the paper, we also demonstrate different ways for estimating person importance based on the temporal aspects of the link structure.

Categories and Subject Descriptors
H.3.m [**Information Systems**]: Information Storage and Retrieval: Miscellaneous

General Terms
Measurement, Human Factors.

Keywords
Wikipedia, historical analysis, digital history, social networks, temporal link analysis

1. INTRODUCTION
History plays significant roles in our society by giving account of the past, explaining the present and offering lessons for the future. It helps to create meaning, coherence, orientation as well as settles the foundations of nations, our identities and memories, etc. As such, the history is one of the fundamental subjects taught from elementary schools onwards. The field of history science has recently started to benefit from the advances in computer science and information technologies [3,10,24,30,40,47,53], much like social sciences have been fostered by the advent of *computational social science* [35]. *Digital history* (aka. *computational history* or *histoinformatics*) has emerged as a subset of Digital Humanities that utilizes automatic approaches to process, organize, make sense of historical data and to verify or validate historical hypotheses. The growing interest in the application of computational approaches to the history science is also evidenced by dedicated interdisciplinary events (e.g., [28,36]).

Source criticism takes prime position in the history science [5]. The credibility, coverage, origin and other characteristics of sources are usually carefully scrutinized before the start of research. Although typically, primary sources are the main interest of historical analysis, secondary sources are also common subject to investigation. Wikipedia as the largest base of collaboratively created knowledge, is naturally one of them. Despite initial wave of criticism, it has been increasingly used in humanities including the history and memory science (e.g., [7,10,31,36,45,47]). For example, the president of the American Historical Association W. Cronon in the association's publication "Perspectives on History" [7] has recently called historians for embracing Wikipedia in education and research and for actively contributing to make it even better.

The importance of Wikipedia from the viewpoint of history science is due to its powerful educational impact. Wikipedia is hugely popular with nearly 500 million unique visitors each month[1] and constitutes a crucial source of history-related knowledge for majority of users. Typically, users refer to it as a starting point (or springboard) in their search for past-related information. For example, based on controlled user studies conducted in 2014, Wikipedia has been found to be the most frequently visited website for searchers seeking historical knowledge or wishing to corroborate historical facts [29]. Furthermore, Wikipedia and derived from it datasets (e.g., DBpedia[2] [2] or Yago2 [23]) are being commonly used as bases for many knowledge intensive processing tasks (e.g., [13,14,16,25,30,31,32,41,47,48,54]), some of which explicitly focus on historical data (e.g., [16,25,30,31,47]).

Prior studies investigated numerous aspects of Wikipedia including the process of its creation and evolution, the credibility and coverage of its content, controversy, collaboration or lack of it and so on [18,23,33,45]. However, dedicated analysis of history-related content as well as its broad temporal aspects of Wikipedia

JCDL '16, June 19-23, 2016, Newark, NJ, USA
 ISBN 978-1-4503-4229-2/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2910896.2910911

[1] Wikipedia is the 7th most visited website globally by Alexa ranking (25/1/2016) http://www.alexa.com/siteinfo/wikipedia.org

[2] http://dbpedia.org

articles has not been done so far. In this paper we analyze Wikipedia in order to understand the way in which history is recorded, organized and remembered and through this to better inform future digital history studies. We focus on a particular entity type, persons. Persons are the essence of the history and constitute the large fraction of Wikipedia content. They can be also easily positioned on timeline (provided their birth and death dates are known) unlike other types of entities, e.g., locations, ideas or concepts for which temporal boundaries are harder to be determined. We quantitatively study multiple aspects of historical persons such as the amount of recorded entities per each decade, the link distribution of different cohorts segmented based on their lifetimes and the characteristics of temporal snapshots of social networks. Since the history tends to be defined as an "unending dialogue between the present and the past" [5], we also analyze the connectivity between the present and the past as well as the distribution of viewing frequency of Wikipedia pages on people from different eras.

In particular, a series of questions guide our study:

Q1. *How many historical persons are described in Wikipedia? How much content is available about them?*
Q2. *How are historical persons connected in Wikipedia? What is the effect of time on the link structure and on the overall connectivity within Wikipedia?*
Q3. *How much does user interest in history change with regards to the distance in the past? Is there any correlation between the time when a person lived and its current popularity?*
Q4. *How strongly is the content on the past persons connected to the one on present persons?*
Q5. *How can we estimate historical person's importance using Wikipedia link structure?*

To answer the above questions, we assume a novel analysis style that organizes Wikipedia articles chronologically by the valid time of entities and which associates link-based metrics with time. Note that, unlike this work, prior studies focused mainly on the creation time of Wikipedia content (i.e., on the way in which users collaboratively create content over time or on the recency of information).

To sum up, we exhaustively investigate the characteristics of historical persons described in Wikipedia. We analyze the way in which they are described, the way in which they are inter-connected as well as the extent to which the information on them is accessed by Wikipedia visitors. We then demonstrate the temporal orientation of links to show that link distribution depends not only on semantics but also on time. In addition, we discuss several centrality measures on social historical graph.

The remainder of this paper is structured as follows. In the next section we provide an overview of the related work. Section 3 introduces the dataset and its preprocessing. The next section is the main part of the paper giving the analysis results. Section 5 contains the summary of findings and general discussion. Finally, we conclude the paper and outline the future work in the last section.

2. RELATED WORK

2.1 Wikipedia Analysis and Use

Wikipedia with its vast amount of user-generated content is a goldmine of knowledge both for average readers and for researchers who increasingly use it for many knowledge intensive tasks (e.g., [13,14,16,25,30,31,32,41,47,48,54]). It has been applied to various research areas in computer science ranging from natural language processing, information retrieval, information extraction, ontology construction, etc.

Wikipedia articles have been reported to have, in general, sufficient accuracy [18]. Within the history realm, a recent essay [45] by American historian Rosenzweig have found that mistakes in Wikipedia are as equally common as in reputable sources, and that serious mistakes are typically corrected within hours. Yet, one valid complaint found is the bias of its writers who predominantly are English-speaking males from Western culture. Another shortcoming is that Wikipedia "summarizes and reports the conventional and accepted wisdom on a topic, but does not break new ground." In other words, it lacks original research.

Many works in digital humanities utilize Wikipedia or derived from it knowledge bases. For example, Huet *et al*. [25] studied appearances of historical persons in the past editions of Le Monde based on their attributes derived from the Wikipedia to portray trends in popularity of different professions and the rise of the importance of women in French society. Garcia-Fernandez *et al*. [16] automatically determined publication dates of documents based on a range of linguistic features, one of which is the appearance of historical persons' names in text. The information on the lifespan of detected persons was collected from Wikipedia and used as additional signal for estimating document age. Eom *et al*. [10] studied the hyperlink networks of 24 Wikipedia language editions and automatically extracted the top 100 historical figures for each Wikipedia edition in order to investigate their spatial, temporal, and gender distributions with respect to their cultural origins. Skiena and Ward [47] ranked historical people using PageRank algorithm applied on the hyperlink graph consisting of person pages in Wikipedia. They also used the appearance statistics of person names in Google Books dataset[3]. Takahashi *et al.* [46] estimated influence of historical persons in unsupervised way based on spatio-temporal analysis and the adaptation of PageRank algorithm [44] using Wikipedia link structure. Other examples of using history-related data in Wikipedia can be found in [39] and [50]. Given the popularity of Wikipedia we believe it is important to undertake deeper studies of its history-related content. However, as far as we know, only one work tried to quantify the amount of historical data in Wikipedia. Kittur *et al.* [33] found through sampling that in 2009 about 11% content was strictly devoted to history and the content has grown 143% from 2006 to 2008.

In this work we assume a novel objective. Given the wealth of data on the past in Wikipedia and its frequent usage in education as well as in research, we look closely how the link structure, and the strength of remembering are related to the time periods of historical entities. To the best of our knowledge, this is the first Wikipedia study that considers this kind of temporal analysis of Wikipedia articles.

2.2 Collective Memory Studies

The concept of collective memory (social memory) popularized by Halbwachs [20,22] defines the collective view of society on the past. Collective memory is often contrasted with the concept of *collective amnesia* defined by Jacoby [26] as forceful or unconscious suppressions of memories, especially, those related to disgraceful or inconvenient events. In a similar fashion to personal memory [9], the social memory is known to decrease along time and to be subject to temporal variations following the occurrence of memory triggers such as sudden events or anniversaries [3,31]. Studies of collective memory can help us to understand the

[3] https://books.google.com/ngrams/datasets

mechanisms of forgetting and remembering as well as can explain the role of history in our lives. In addition, they have direct implications on the archival selection by memory institutions such as national or dedicated archives [30]. Traditionally, the research on collective memory has been based on small-scale investigations of personal accounts. Relatively few works have been carried out that use computational approaches for quantifying the characteristics of social memory over large text datasets. Cook *et al*. [6] investigated the decay of fame over time on the basis of the collection of news articles that covers 20th century. In [3] we studied memory decay and the way in which past years are remembered using the dataset of English news articles about different countries spanning 90 years. In another work [27] we have also analyzed the way in which users refer to the time in Twitter in order to measure *collective temporal attention* towards the past and the future.

Ferron and Massa [11] and Kanhabua *et al*. [31] proposed to treat Wikipedia as a global memory space. Differently to our work they focused on memory triggers that cause forgotten or poorly remembered events to be brought back into social attention. Anniversaries are natural examples of memory triggers. In another case, current events may also serve as triggers of the memories of similar, past events. Ferron and Massa studied also the way in which memory forms by analyzing the collaboration dynamics of Wikipedia contributors who edit articles on tragic events such as acts of terrorism (e.g., World Trade Center collapse) or natural disasters (e.g., Katrina Hurricane). Our work can be seen as complementary to that of Ferron and Massa [11] and of Kanhabua *et al*. [31].

3. DATA PREPARATION

3.1 Data Collection

We used the English Wikipedia dump provided by Wikimedia foundation[4]. To collect Wikipedia pages about persons we utilized DBpedia ontology datasets (PersonData ontology class) [2]. To capture core article content we used the BeautifulSoup library[5] excluding lists as well as footers under commonly used footer titles: '`See also`', '`References`', '`External links`' and '`Notes`'.

We then collected hyperlinks using Yago2 [23]. Based on the collected links, we could create directed graph, $G(V,E)$, where V is the set of nodes representing persons and E is the set of edges connecting them. An edge e_{ij} from a node v_i to node v_j indicates the presence of a hypertext link in v_i that leads to v_j.

To solve the problem of redirects, nodes redirecting to other pages within Wikipedia were merged with their targets. In addition, self-loops (self-links) were removed by excluding links with identical origin and destination.

3.2 Attribute Assignment

The information on the birth and death of persons has been obtained from Yago2. While many nodes in our dataset have complete attributes, certain fraction lacked either birth or death dates, while some had neither of them. In Fig. 1 we show the rate of persons without the birth date (green line) and the rate of persons without the death date (red line). The former measures the percentage of persons that died at a given decade who lack their birth date, while the latter shows the percentage of persons born at a given decade whose death date is not known. We counted only persons for which at least one of the dates (birth or death) is known (if both are unknown it is, of course, difficult to assign a person to timeline). The high rate of persons without known death dates in the current and in the last century is not surprising as many are still alive. On the other hand, interestingly, we notice that the rate of persons without the known birth date in the past centuries is higher than the one of persons without the death rate. This may be attributed to the lack of efficient demographics recording (e.g., civil registry) and archiving tools or systems in the past [51]. A death of a person, especially a prominent one, was likely noticed and recorded. Yet, his/her birth related information may not always have been known, unless the person was born to a well-known or noble family. We also notice that the probability of a person article to lack her birth date is higher, the longer time ago the person lived.

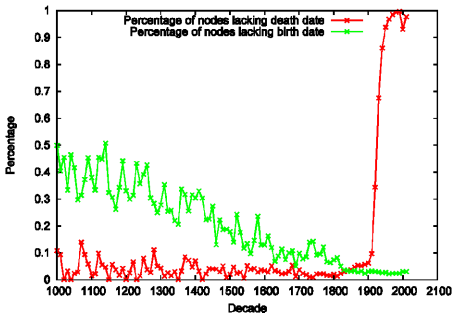

Figure 1 Rate of persons lacking birth or death date per decade.

Note that excluding persons lacking the attribute values would significantly decrease the amount of data at very distant decades, for which, the data is already sparse. We then inferred missing attributes for persons who lack either birth or death dates[6] after first mapping each known date to its decade for minimizing error. Based on the nodes with the complete set of attribute values (i.e., persons with known both the birth and death dates), we first computed: the *mean death date* for people born at a given decade and the *mean birth date* for people who died at a given decade. In result, each birth decade d_b was associated with the most probable death decade computed over the people born at d_b. Similarly, each death decade d_d was associated with the most probable birth decade calculated over the people who died at d_d.

People born in the 20th and 21st centuries who lack their death dates were treated differently. Many of them are still alive so assigning their death decades requires a forecasting procedure to avoid underestimation. We estimated their probable death decades by the least square error method trained on all the persons born after year 200 and before year 1900. The forecasting is reliable when we look at the part of the plot from year 200 until 1900 as shown in Fig. 2. It portrays the average death date for people born at a given birth decade based on the nodes which have complete set of attributes. We can observe a strong linearity for most of the time period except for the two noisy first centuries. After removing the data from the first two and the last two centuries, the fitted linear trend line was: $y = 1.003x + 54.61$ ($R^2 = 0.9173$).

We then assigned the most probable birth and death decades for the nodes that lacked either of the attributes. In the remaining of this paper we will focus on persons born during and after 11th century onwards. The total number of nodes we use in the analysis is 459,991.

[4] https://dumps.wikimedia.org/enwiki

[5] https://pypi.python.org/pypi/beautifulsoup4

[6] Nodes that lacked both dates were removed.

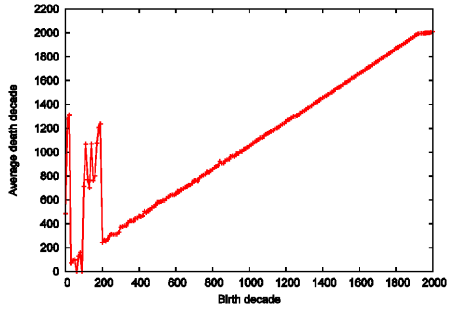

Figure 2 Average death decade for persons at different time.

4. ANALYSIS

4.1 Content Analysis

We first look into the average number of persons per decade. Fig. 3 shows on the log scale the counts of persons alive at past decades. A person is associated with a given decade if her or his lifetime overlaps with the decade. We observe a strong increase in the number of persons that have their Wikipedia articles, the closer to the present time. The plot features close to exponential character and could be actually well-approximated by the three straight lines (with corresponding epochs: 1000-1399, 1400-1699, 1700-1990) each having higher slope value than the previous one.

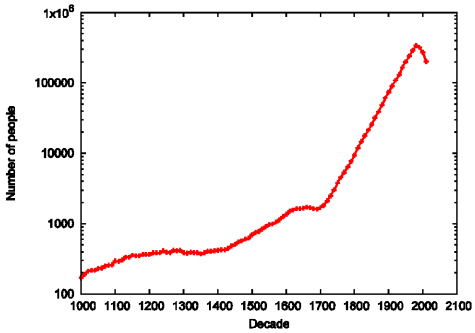

Figure 3 Number of persons per decade in log scale.

The increasing number of people at more recent decades can be of course explained by the demographic trend of rapid population increase over recent times [34]. It also aligns with an intuitive hypothesis that the amount of "remembered" data decreases exponentially with the time elapsed. Previous study on news article collections has demonstrated similar exponential decrease in the strength of remembering of past years [3].

We next examine the effect of time on the amount of content within the Wikipedia articles. According to an intuitive hypothesis, the more time ago a person lived, the less information should be available about her or him, and, hence, the article about the person should be, on average, shorter than ones on more recent people. To the contrary when looking at Fig. 4, which plots the average article length for each decade, it becomes apparent that rather opposite happens.

The mean lengths of articles of people in the 19th and 20th centuries are on average shorter than of persons in the previous centuries. This is likely due to the fact that many less famous persons who lived in the recent past are recorded in Wikipedia. For such persons, Wikipedia contributors may have difficulty to find enough verifiable and informative content, or they may be simply less interested and motivated to contribute.

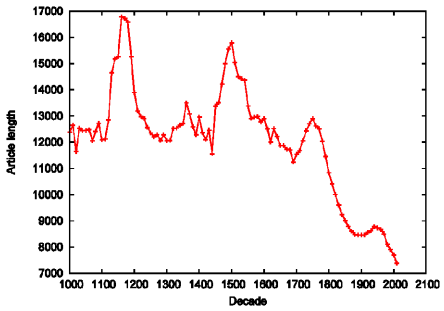

Figure 4 Average article length per decade.

4.2 Connectivity Analysis

We now look into the connectivity aspects of articles to investigate temporal differences of links. First, we examine the change in the number of links in relation to time. Fig. 5 contrasts the average in- and out-degrees with the time when persons lived. While certain fluctuations can be observed, on average, the mean numbers of incoming (red line) and outgoing (green line) links are decreasing, the more recently a person lived. This suggests that Wikipedia pages about more recent persons are, on average, connected less strongly with other persons than the pages about more distant persons. We also observe that in- and out-degree values tend to correlate over time.

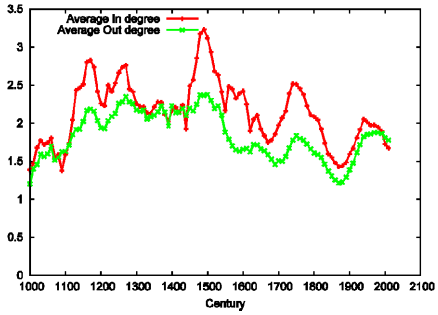

Figure 5 Average in-link and out-link degrees in the past.

To investigate more the decrease in the link rate for the recent persons we show in the upper graph plot of Fig. 6 the ratio of nodes that have at least one incoming link and the ratio of nodes with at least one outgoing link. A person is considered to live in a given century if most of her life span occurred at that century (i.e., the century contains the midpoint of the person's life). We notice that there are relatively more persons without any in- or out-links leading to other persons in more recent centuries. This may be again due to many Wikipedia pages about less known persons in the recent times. Note that, even if a page may have few links to other persons, it still can link to other types of Wikipedia articles that we do not consider in this study (e.g., locations, events or concepts) or to persons who did not live in the second millennium.

We think that, ideally, a person should be well-connected to the social context of its time, that is, to other relevant, contemporary persons. By this visitors could receive contextual information for obtaining more organized and structured view of a person. So weaker connectivity means less chances to understand a target person as well as her context and to discover other related persons. According to the theory of *structuralism* [49], the meaning of concepts resides in the relationships with other concepts. Thus, concepts or entities considered alone may be difficult to be understood and should rather be viewed within their context. Similar idea should apply also to Wikipedia entities. A possible remedy could be adjusting Wikipedia's editing policies and

guidelines to put more emphasis on sufficient "grounding" of described persons.

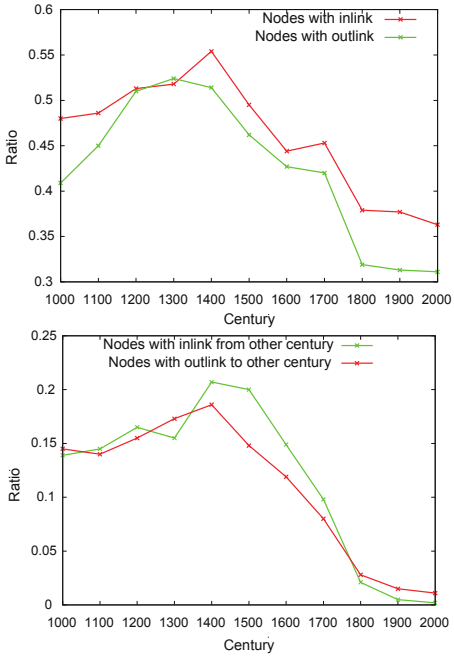

Figure 6 Ratios of persons with at least one in- or out-link from/to any century (top) and from/to centuries different than the one of the target person (bottom).

We next analyze *across-time connectivity* of historical persons. Links between persons from other historical centuries are less likely to indicate physical relationships and actual interaction between the persons. Instead, they tend to be the artefact of historical comparison, family lineage, or point to the start/continuation/end of some processes, etc. In the lower graph of Fig. 6 we show the ratio of nodes that have at least one out-going link to or one in-coming link from someone who lived in another century. We can observe that the persons in the last two centuries have on average less connectivity with the people outside their centuries, than the persons at earlier centuries. This means that Wikipedia pages on persons from the distant past tend to have more *across-century connectivity* than the pages about persons living more recently.

To better portray the inter-century linkage we next plot the aggregate temporal orientation of links in Fig. 7. It displays the rate of links coming from the past persons (blue color), the contemporary persons (green) and the future persons (red) at every decade. Two persons are considered contemporary if their lifetimes have non-empty overlap. Looking at Fig. 7 we notice that on average few links tend to originate from persons living in the past, while most of the links are from the contemporary people. This means that whenever a page has an in-link there is high probability that that link comes from a contemporary person. Note that naturally, the amount of links from the "future persons" decreases the closer to the latest decade.

Figures 6 and 7 do not inform about the distance between linking persons. We then measure the average distance between connected persons and superimpose it over time. The distance is expressed as the number of years that separate the origin and the target of every link. The calculation is done as follows. For each decade *d* we first collect all persons who lived in that decade. Then, for each person *p* alive at *d* we collect all its in-links and compute the distance between *d* and the decade from which each such link originates. The latter is represented as the mean decade of the link's origin (mean lifetime point of the person that links to *p*). Finally, we compute the average distance for all the in-links of all the persons living in *d* to portray how far the people alive at *d* are linked from. Fig. 8 shows the results. Interestingly, we notice relatively large distance for people living in the distant past, and, a smaller average distance for more recent persons. This confirms the higher across-time connectivity of past persons.

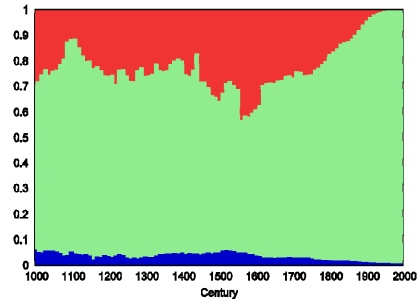

Figure 7 Rate of in-links from past persons (blue), contemporary persons (green) and future persons (red).

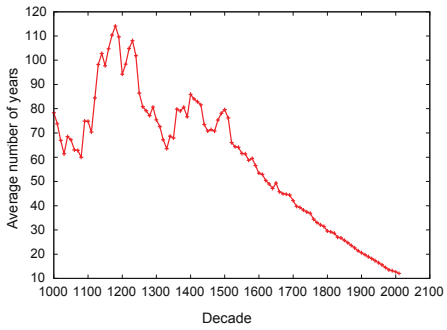

Figure 8 Average distance between link source and link target.

In the next two graphs (Fig. 9) we provide a more detailed view of the relative link distances, separately, for the in- and out-links. For all the people associated with a given century we show the average distribution of their link distances computed as the number of years between the midpoints of connected persons. In particular, the upper graph of Fig. 9 gives the plot for in-links and the lower one for out-links. Note that the midpoints of target persons are always positioned at point 0. The small peaks on the right hand side of the upper graph (the graph showing in-links) are due to links from persons alive at present times. Such peaks are less pronounced at the lower graph of Fig. 9 (the graph showing out-links) suggesting rather weak across-time reciprocity.

The connectivity analysis shown in this section gives rise for forming the hypothesis of a temporal version of social homophily [38] (or "temporal homophily"): *A person tends to be linked more with persons around its time than with persons from distant time.*

This suggests the possibility of automatically dating persons or entities – an important task considering that many entities lack such temporal metadata (as partially shown in the Sec. 3). This is despite the fact that such information is necessary for various processing tasks that harness Wikipedia (e.g., [16,25]). Detecting (or supporting the detection of) an entity's time period could be done by link analysis in a similar way as the one in which approximate page timestamp is gauged by analyzing the timestamps of its neighborhood [43] (i.e., pages linking to the target page).

The above hypothesis has also potential to impact approaches which utilize Wikipedia link structure for entity-to-entity relationship analysis [13,14,32,41,48,54]. The difference between the *activity times* of connected nodes could be taken into consideration when evaluating the relations' strengths or when detecting the topics of such relations.

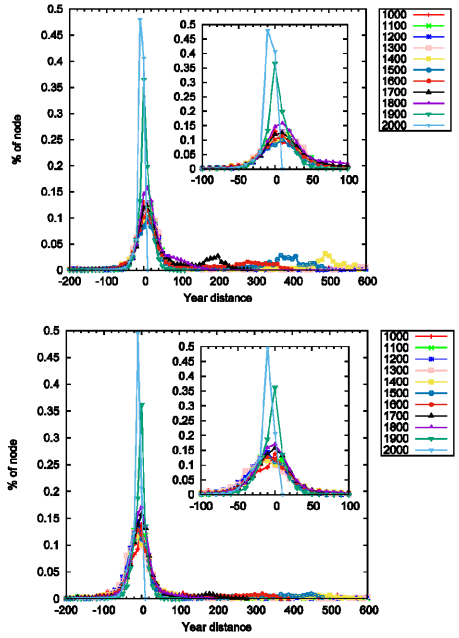

Figure 9 Average in-link (top) and out-link (bottom) distances. The horizontal axis denotes the distance between the linked pages (negative values mean links from/to the past person).

4.3 Historical Social Networks

Studies of social networks are nowadays common due to the popularity of SNSs and the social Web. Analogously, re-creating social networks of the past is also appealing [8]. However, constructing social networks as they existed in the past is inherently difficult and, likely, will never be completely possible given fragmentary data from the past. Instead, as a simple approximation, we can study networks formed by the hyperlink structure between the Wikipedia articles about historical persons.

We define a historical social network in a unit time period t_i as a graph, $G_i(V_i,E_i)$, composed of Wikipedia articles on persons that lived at t_i treated as the set of graph nodes, V_i, and the links between them considered as the edges, E_i. We adopt here the century granularity, hence, t_i represents here a single century. A series of temporal social networks (one for each century) is then created for the entire time period of analysis, $T=(t_1,...,t_i,...,t_n)$. Note that a person is assigned to a given century if the midpoint of his/her life is included in that century. In Fig. 10 we demonstrate the networks for a few selected centuries visualized with the ARF presentation layout [17]. ARF belongs to the class of force-directed graph layouts and is characterized by a circular shape. It is easy to read thanks to the fact that the layout displays as much symmetry as possible and that single nodes are pushed to the outer edges of the circle.

A common way for analyzing social networks is to estimate node importance or prestige by applying centrality measures. We thus first compute node importance using the well-known *PageRank* [44] algorithm in each historical social network, G_i.

Unlike previous works [10,47] in which PageRank is calculated on the entire social graph G, we compute it separately for each social network, G_i. To distinguish between these two approaches, we will call the random walk computation on a historical social network, *Century PageRank*, while the one on the entire graph, *Global PageRank*. Century PageRank score indicates how prominent a person is among people living in her century, while PageRank score measures person's prestige among all the persons in the Wikipedia social graph (or at least in our dataset), irrespectively of time. In Fig. 11 we plot the Pearson Correlation Coefficient between the Century PageRank and Global PageRank scores for each different century. Although the correlation is positive we can see that the rankings based on the two scores are not exactly same, especially, for centuries before the 17th century.

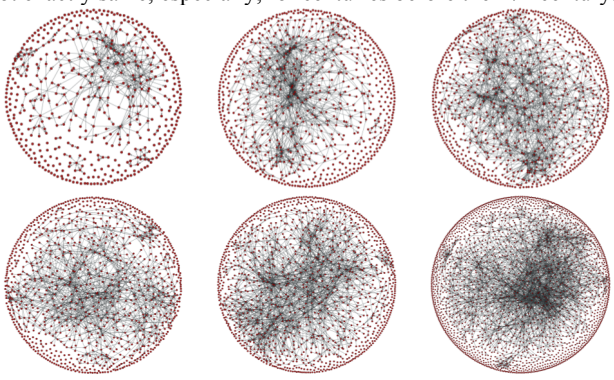

Figure 10 Social networks in the 11th, 12th, 13th, 14th, 15th and 16th centuries (from top left to bottom right).

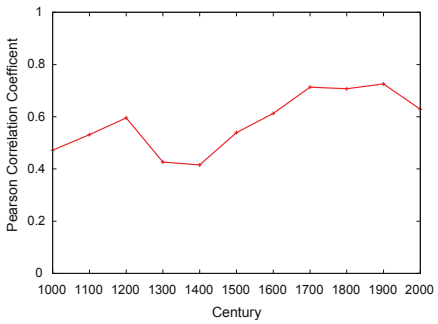

Figure 11 Pearson Correlation Coefficient between Global PageRank and Century PageRank.

The cumulative plot of Century PageRank for people living in different centuries is shown in Fig. 12. As it can be seen, the distributions of the scores are in general quite uniform in each century.

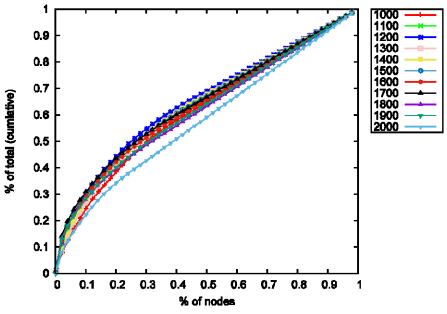

Figure 12 Cumulative plots of Century PageRank distributions in each century. Horizontal axis represents nodes ordered by their Century PageRank scores.

4.4 Remembering Past

We analyze in this section the degree to which past persons are connected with the present and the strength with which they are remembered. We consider the "present-to-past" connectivity as one measure of utility of the history. In fact, the role of history is to teach lessons, and the usefulness of the past accounts relies on how much they can serve the current society [15]. We propose two approaches in this paper: one based on the link analysis and the other based on the view logs. They are described below.

4.4.1 Present-to-Past Connectivity

The first measure quantifies how much a historical person is linked to the present time. In particular, we estimate the connectivity of historical persons with the "present" persons. As present persons we consider people alive during any of the four decades (1970s – 2000s). Note that this period can be arbitrarily chosen. The present-to-past connectivity measure should represent the closeness of nodes denoting historical persons to the nodes corresponding to the present persons. We propose to apply Biased PageRank similar to the concept of TrustRank [19] on graph *G* where the random walk is biased to the present persons. We call it a *Present-Biased PageRank*.

Fig. 13 shows the average Present-Biased PageRank scores obtained by averaging the scores for people alive at a given decade in the past. The graph can be interpreted as the relation strength between the persons from a given past decade and the present persons. As it can be seen, the rate drastically decreases from the 20th century backwards in time to, more or less, stabilize after 1900s. Notably, people around 15th and 16th centuries seem to be connected bit more to the present. This observation aligns with the relatively higher number of links to such people from "future persons" as shown in Fig. 7.

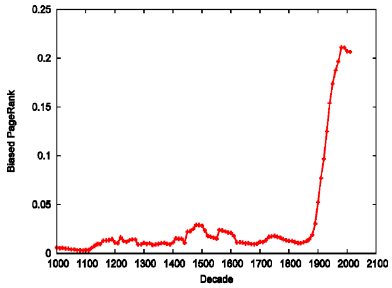

Figure 13 Average values of Present-Biased PageRank per decade.

To understand how Present-Biased PageRank scores distribute in each century we show cumulative plots for each century in Fig. 14. The plots for more distant centuries reveal long tail distributions in which few persons have very high scores while the rest of people have small scores. It suggests a winner-takes-all situation in the past centuries. Only few selected nodes from each century are strongly connected with the present. On the other hand, the last two centuries have close to linear cumulative distribution of scores. When comparing Fig. 14 with Fig. 12 we conclude that the Century PageRank scores for the far away centuries (distant past) are distributed more evenly than ones of Present-Biased PageRank.

We next show in Fig. 15 the Pearson Correlation Coefficient between the scores by Present-Biased PageRank and those by Century PageRank. We can see that prominent people at a given century are not necessarily strongly connected to the present. The correlation for persons living at distant centuries is low indicating quite weak connectivity of the top prominent persons in those centuries to the people living at the present times.

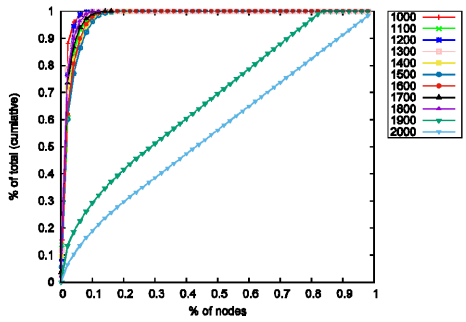

Figure 14 Cumulative plots of Present-Biased PageRank distributions in each century. Horizontal axis represents the percent of nodes ordered by their Biased PageRank scores.

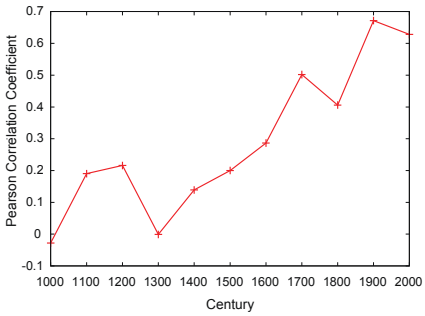

Figure 15 Pearson Correlation Coefficient between Present-Biased PageRank and Century PageRank.

4.4.2 View Frequency

We next analyze the distribution of visitor views in different times. We consider the viewing frequency as a measure of attention and interest in history and the evidence that the past matters. To quantify the popularity of past persons, we make use of the view logs available from the Wikipedia Foundation[7]. We count accesses to every page in our dataset that took place during 5 years long time period from the start of January 2009 to the end of December 2013. The upper plot of Fig. 16 shows in a log scale the average view count of persons alive in a given past decade. We also list the top viewed person for each century in Table 1.

Looking at the average number of views for persons from different centuries we can conclude that the interest to the past persons remains quite strong. We see that although, the average viewership of historical persons differs across centuries, it does not depend on the time segments in a simple way. For example, persons active in 15th and 16th centuries gather the highest attention of visitors. For a comparison we also show the total view count per time at the bottom plot of Fig. 16.

In Fig. 17 we portray changes in the average view rate over the 5 years' long time period (monthly granularity) for which we collected view logs. We can see that the viewership does not remain stable over the viewing time. After examining the peaks, we have found that many can be explained by anniversaries or sudden discoveries related to the past persons. For example, while Shakespeare is an unquestionable "king" of the 16th century (see the last column of Table 1), he has been "dethroned" in February 2013 when Google search engine commemorated the 540th

[7] https://dumps.wikimedia.org/other/pagecounts-raw/

anniversary of the birthday of Polish astronomer Copernicus with a related doodle [37]. Similarly, a doodle for the 374th anniversary of the birthday of Danish anatomist, Nicolas Steno's caused a spike in the line for the 17th century on January 2012 [42]. During the same month when Copernicus became the most often viewed person of the 16th century's cohort, Richard III of England "won" the first place within the 15th century cohort (ahead of the usual winner: Leonardo da Vinci) following the remarkable discovery of his remains in Leicester, England [4]. Another example illustrates a very rare case related to past prophecies. The peak on December 2012 within the aggregated view rate of the 16th segment is due to the frequent visits of Nostradamus's (Michel de Nostredame) page, presumably, in association with the alleged Mayan Prophecy.

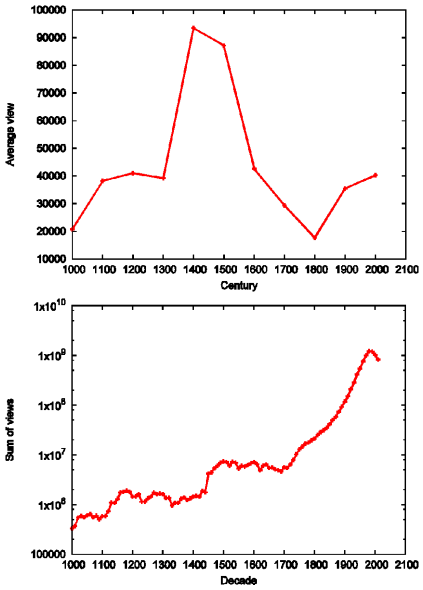

Figure 16 Average views for persons living in a given past century (left) and the total sum of views on log scale (right).

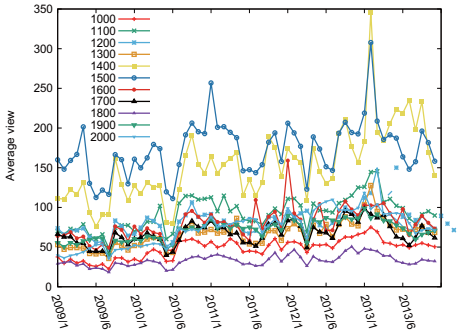

Figure 17 Average views for persons over the viewing time.

Finally, we analyze the cumulative view distribution per century in Fig. 18. In each century the visitors' attention is quite skewed and there are rather few persons whose pages are accessed very frequently, while many pages are visited rarely.

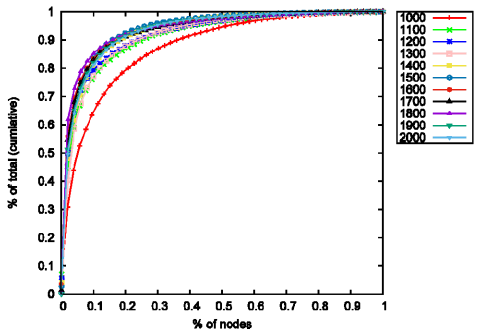

Figure 18 Cumulative plots of visitor views among persons in different centuries.

5. DISCUSSION

In this section we first summarize the main findings of this study and then we provide additional discussion.

- The number of persons recorded in Wikipedia depends on the time when they lived. There appears to be close to exponential growth in the number of person-related articles along with the increase in time.
- The average length of articles about past persons is longer than those about more recent persons.
- The average number of in- and out-links to other persons decreases along with time (from past towards the present).
- Across-time connectivity varies with the time of person's life. Pages on more historical persons tend to be more connected to the people who lived in other centuries.
- For any person, there are more links originating from people who lived in later centuries than from ones who lived in the previous centuries. Temporal orientation of links is then skewed towards the future.
- The link distribution does not solely depend on the semantics, but time plays certain role here, too. In every century there are more links to/from contemporary persons than to/from persons who lived in other times. The distance between the origin and target of the links becomes on average shorter for more recent persons.
- The view log study shows that, in aggregate, there are few views to pages about persons from the distant past. Yet, on average, the articles on the past persons are more frequently accessed than the ones on the current persons.
- The view rate of historical persons is not stable over time. Occasional peaks in the view frequency happen due to anniversaries or other events related to the memory or changes in our knowledge about historical persons.
- Only few persons are strongly "remembered" from the past as quantified by the distributions of Present-Biased PageRank scores and view counts of their pages.
- PageRank on the global social network can be complemented by other time-based centrality measures such as Century PageRank and Present-Biased PageRank. These metrics are correlated, yet, they are not equal.

Table 1 Top ranked persons by *Global PageRank*, *Century PageRank*, *Present-Biased PageRank* and by View Frequency.

Century	*Global PageRank*	*Century PageRank*	*Present-Biased PageRank*	View Frequency
11th	William the Conqueror	William the Conqueror	William the Conqueror	William the Conqueror
12th	Genghis Khan	Saladin	Genghis Khan	Genghis Khan
13th	William Wallace	Thomas Aquinas	Thomas Aquinas	William Wallace
14th	Geoffrey Chaucer	Petrarch	Hafez	Geoffrey Chaucer
15th	Leonardo da Vinci	Joan of Arc	Leonardo da Vinci	Leonardo da Vinci
16th	William Shakespeare	Philip II of Spain	William Shakespeare	William Shakespeare
17th	Isaac Newton	Rembrandt	Rembrandt	Isaac Newton
18th	George Washington	George Washington	George Washington	George Washington
19th	Abraham Lincoln	Abraham Lincoln	Abraham Lincoln	Abraham Lincoln
20th	Michael Jackson	Bill Clinton	Barack Obama	Michael Jackson

Link structure is often used for quantifying relationships on Wikipedia. Thus, properly understanding the role of time and temporal attributes of links ("temporal link signature") can help to improve the results. While there have been studies on the temporal evolution of Wikipedia (e.g., the evolution of Wiki Graph), more investigation should be done on the temporal scope of Wikipedia articles and on time effect on their interconnectivity. In addition, automatic means of time-scoping Wikipedia articles could be proposed using link structure (see Sec. 4.2).

Historical knowledge is especially useful when it strongly relates to the present. Thus computing the importance of historical entities should consider the extent to which the entities are useful for present users. There can be several ways to quantify the past-present relations. We have suggested two such ways in Sec. 4.4.

For improving the usefulness of its articles, Wikimedia Foundation could encourage contributors who edit pages on past entities to try to add more links related to the present time to better explain their roles and importance. At the same time, such entities should not be disconnected from their contemporary context (e.g., social context in the past). An interesting idea would be to propose automatic construction of summaries (e.g., in the form of term clouds) to portray a person's relation to both the current as well as to its contemporary time.

Cultural memory is often categorized into two modes [1]: *passive* (aka. "canon") and *active* (aka. "archive"). The latter represents what is visible to public, while the former comprises what is not "on display". Both the view frequency and the past-to-present connectivity could be regarded as signals useful to distinguish the passive from active memory. This could have implications on archival and preservation decisions [30].

Lastly, additional studies are needed for re-constructing actual social networks in the past. Wikipedia can however provide a foundation for such networks. In addition, link structure based metrics of importance and prestige such as ones listed in this paper and others [8] should be contrasted against the lists of top influential or important persons in the past [12,21] which are manually compiled by professionals.

6. CONCLUSIONS

Studies of the history and the collective memories are important due to the significance of history and its role in our society. At the same time, since Wikipedia constitutes the main source of historical information for online users and for many knowledge processing tasks, the in-depth analysis of its content is needed. The objective of this paper is to help better understand the characteristics of historical data in Wikipedia through applying a novel kind of study. We think that this study and similar ones could support better design of any systems that utilize historical data in Wikipedia, especially, ones that use information on persons or their social networks. Also, we hope our work can contribute to the collective memory studies.

Several avenues of future work emerge from this research. We first plan to focus on other entities such as events or places. The difficulty here lies in estimating their temporal attributes to position them in time. Second, it is appealing to compare multiple language editions of Wikipedia for the amount and the focus of the historical knowledge they hold. Finally, the comparison of Wikipedia with historical textbooks could shed more light on the coverage and correctness of contained history-related information.

7. ACKNOWLEDGMENTS

This research was supported in part by the Japan Science and Technology Agency (JST) research promotion program Presto/Sakigake: "Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents" and by Grant-in-Aid for Scientific Research (No. 15H01718) from MEXT of Japan.

8. REFERENCES

[1] A. Assmann. *Introduction to Cultural Studies*. Schmidt Erich Verlag, 2008 (in German).

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *ISWC'07/ASWC'07*, 722–735. Springer, 2007.

[3] C.-M. Au Yeung, and A. Jatowt. Studying how the Past is Remembered: Towards Computational History through Large Scale Text Mining. In *CIKM 2011*, pp. 1231-1240, 2011.

[4] J. F. Burns. Bones Under Parking Lot Belonged to Richard III, 2/2013, http://www.nytimes.com/2013/02/05/world/europe/richard-the-third-bones.html?_r=1

[5] E.H. Carr. *What is History?* Penguin, London, 1961.

[6] J. Cook, A. Das Sarma, A. Fabrikant, and A. Tomkins. Your Two Weeks of Fame and Your Grandmother's. In *WWW 2012*. ACM, New York, NY, USA, 919-928, 2012.

[7] W. Cronon. Scholarly Authority in a Wikified World. *Perspectives in History*, 2012.

[8] M. Düring. Can Network Analysis Reveal Importance? Degree Centrality and Leaders in the EU Integration Process. *Social Informatics*. Springer International Publishing, 2014. 314-318.

[9] H. Ebbinghaus. *Memory: A Contribution to Experimental Psychology*. 1913.

[10] Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, D. L. Shepelyansky. Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions, *PLoS ONE* 10(3), 2014.

[11] M. Ferron and P. Massa. Collective Memory Building in Wikipedia: the Case of North African Uprisings. In *WikiSym '11*. ACM, New York, NY, USA, 114-123, 2011.

[12] R. Friedman. The Life Millennium: *The 100 Most Important Events and People of the Past 1000 Years*, Bulfinch P., 1998.

[13] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. IJCAI 2007*, pp. 1606–1611, 2007.

[14] E. Gabrilovich, *et al*. Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI 2006*.

[15] H.-G. Gadamer. *Truth and Method*. London: Sheed and Ward, 1975.

[16] A. Garcia-Fernandez, A.-L. Ligozat, M. Dinarelli and D. Bernhard. When was it Written? Automatically Determining Publication Dates. In *SPIRE 2011*, 2011.

[17] M. Geipel. Self-Organization Applied to Dynamic Network Layout, *International Journal of Modern Physics C* vol. 18, no. 10 (2007), pp. 1537-1549.

[18] J. Giles. Internet Encyclopaedias Go Head to Head, *Nature* 438, 900-901, 2005.

[19] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam with Trustrank. In *VLDB 2004*, 576-587, 2004.

[20] M. Halbwachs. *La Mémoire Collective*. Les Presses universitaires de France, (in French) 1950.

[21] M.H. Hart. *The 100: A Ranking of the Most Influential Persons in History*. Citadel; Revised edition (June 1, 2000)

[22] C. Hoerl and T. McCormack. *Time and Memory: Issues in Philosophy and Psychology*. No.1. 2001.

[23] J. Hoffart et al. YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *WWW 2011*, 229-232, 2011.

[24] L. Hoffmann. Looking Back at Big Data, *Communications of the ACM*, Vol.56 Issue.4, pp.21-23, 2013.

[25] T. Huet, J. Biega, F. Suchanek Mining History with Le Monde, In *AKBC 2013 workshop at CIKM2013*, 2013.

[26] R. Jacoby. Social Amnesia: *A Critique of Contemporary Psychology*. 1997.

[27] A. Jatowt, E. Antoine, Y. Kawai, T. Akiyama. Mapping Temporal Horizons, Analysis of Collective Future and Past Related Attention in Microblogging, In *WWW 2015*, 484-494, 2015.

[28] A. Jatowt, G. Dias, M. Düring and A. van Den Bosch. The HistoInformatics2014 Workshop, *Socinfo2014 Workshop Proceedings*, Springer LNCS 8852, 2014.

[29] H. Joho, A. Jatowt, and R. Blanco. Temporal Information Searching Behaviour and Tactics, *Information Processing and Management Journal*, Elsevier 51(6), 834-850, 2015.

[30] N. Kanhabua, C. Niederée, and W. Siberski. Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study. In *iPres 2013*.

[31] N. Kanhabua, T. N. Nguyen, C. Niederée. What Triggers Human Remembering of Events? A Large-scale Analysis of Catalysts for Collective Memory in Wikipedia. In *JCDL 2014*, 341-350, 2014.

[32] D. Kinzler. WikiSense — Mining the Wiki. In *Proceedings of Wikimania 2005*, *The First International Wikimedia Conference*. Wikimedia Foundation, 2005.

[33] N. Kittur, E. H. Chi, and B. Suh. What's in Wikipedia?: Mapping Topics and Conflict using Socially Annotated Category Structure. *In CHI '09*, 1509-1512, 2009.

[34] M. Kremer. Population Growth and Technological Change: One Million B.C. to 1990, *Quarterly Journal of Economics*, Oxford Journals, pp.681-716, 1993.

[35] D. Lazer et al. Computational Social Science, *Science*, 2009, 721-723.

[36] P. Lendvai and K. Zervanou. *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2013)* at ACL'13, 2013.

[37] T. Malik. Google Doodle Honors 16th Century Astronomer Nicolaus Copernicus, February 19, 2013, http://www.space.com/19868-nicolaus-copernicus-google-doodle.html

[38] M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*. 27:415–444, 2001.

[39] O. Medelyan, D. Milne, C. Legg, and Ian H. Witten. Mining Meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.* 67, 9 (2009), 716-754.

[40] J.-B. Michel *et al*. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176-182, 2011.

[41] D. Milne, O. Medelyan, and I. H. Witten. Mining Domain-specific Thesauri from Wikipedia: A Case Study, In *WI'06*, pp. 442–448, 2006.

[42] Nicolas Steno Google doodle marks his 374th birth anniversary, 2/2012, http://www.theguardian.com/technology/2012/jan/11/nicolas-steno-google-doodle

[43] S. Nunes, C. Ribeiro, and G. David. Using Neighbors to Date Web Documents. In *Proceedings of the WIDM'07 Workshop associated to CIKM'07*, 129-136, 2007.

[44] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford University*, 1998.

[45] R. Rosenzweig. Can History be Open Source? Wikipedia and the Future of the Past. *The Journal of Amer. History* 93:1, 2006 117-46

[46] Y. Takahashi, H. Ohshima, M. Yamamoto, H. Iwasaki, S. Oyama, and K. Tanaka. Evaluating Significance of Historical Entities based on Tempo-spatial Impacts Analysis using Wikipedia Link Structure. In *Proceedings of HT '11*. ACM, New York, NY, USA, 83-92, 2011.

[47] S. Skiena and C. B. Ward. *Who's Bigger, Where Historical Figures Really Rank*. Cambridge University Press, 2014.

[48] M. Strube and S. Ponzetto. WikiRelate! Computing Semantic Relatedness using Wikipedia. *In AAAI-06*, 1419–1424, 2006.

[49] J. Sturrock. *Structuralism and since: from Lévi Strauss to Derrida*, Introduction. 1979.

[50] S. Whiting, J.M. Jose and O. Alonso. Wikipedia as a Time Machine. In *TempWeb'14 at WWW2014*, 857-861, 2014.

[51] T. Wood. *An Introduction to Civil Registration*. Federation of Family History Societies (Publications) 1994.

[52] V. Vapnik. *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[53] G. Zaagsma. On Digital History. *BMGN – Low Countries Historical Review*, 128(4):3–29, 2013.

[54] X. Zhang, Y. Asano and M. Yoshikawa. Mining Knowledge on Relationships between Objects from the Web. *IEICE Transactions 97-D(1)*: 77-88 (2014).