Predicting Importance of Historical Persons using 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 contains a lot of contemporary as well as history-related information, and given its vast coverage and richness, it can be used to rank entities in a variety of different ways. In this work, we are interested in utilizing Wikipedia for judging historical person's importance. Based on the two well-known lists of the most important people in the last millennium, we look closely into factors that determine significance of historical persons. We predict person's importance using six classifiers equipped with features derived from link structure, visit logs and article content.

Keywords

Wikipedia, historical analysis, person influence

1. INTRODUCTION

Wikipedia provides the wealth of information about the present as well as about the past. It contains numerous biographies of historical persons, descriptions of past events, histories of places and so on. It is not surprising that Wikipedia is an important source of history-related knowledge for many Web users. For example, the study of search tactics in the context of temporal information retrieval elucidated that users tend to primarily rely on Wikipedia when looking for past-related information [10].

Given its large coverage and reasonable quality [4], it is possible to utilize Wikipedia for automatically extracting or deriving history-related knowledge (e.g., [13,14,15]). In this work, we study the significance of persons described in Wikipedia articles. Studies of persons' importance and influence help to understand their role in history and allow comparing them with others. *Comparative history* [6] is especially relevant in this regard as a branch of historical studies concerned with contrastive exploration of events and entities from different times and places. Motivated by a broad objective of supporting and complementing historical studies by computational approaches, we propose to estimate person's importance using Wikipedia-derived features, especially, features based on the link structure of Wikipedia. In particular, we are interested in two research questions in this study:

Q1. *Can we predict historical person's importance using data derived from Wikipedia?*

Q2. *What are the features that make a person significant?*

We utilize two well-known lists of the most important persons in the past that were compiled by professionals [3,7]. Based on these data we investigate whether it is possible to successfully use Wikipedia for historical significance prediction and we look into the components that make a person important and memorable. Specifically, we study which features of Wikipedia articles about a historical person are predictive for considering the person as one of the top most influential individuals in the history using range of different classifiers.

2. RELATED WORK

Wikipedia with its huge amount of collectively created content remains an attractive data source for researchers who increasingly utilize it for numerous knowledge intensive tasks (e.g., [1,8,9,13,14,15]). In this work we focus on a particular subset of such researches, that is, on works devoted to the derivation of history-related knowledge such as comparative analysis of historical persons' significance. Eom *et al*. [2] analyzed the hyperlink networks of 24 Wikipedia language editions and automatically extracted the top 100 historical figures for each language edition. Thanks to this, they could investigate spatial, temporal, and gender distributions of persons with respect to their cultural origins. Skiena and Ward [13] proposed ranking historical people using *PageRank* algorithm applied on the hyperlink graph of Wikipedia, where nodes represent person-related articles. To complement their analysis they also looked into the occurrences of person names in Google Books dataset [11][1]. Takahashi *et al.* [14] measured the historical influence of persons in Wikipedia through the spatio-temporal analysis based on the modified PageRank algorithm [12].

Although, in this work we also predict the importance of persons, our methodology is different than ones used in [2,13,14]. Unlike those researches, we employ a classification approach based on the data compiled by professionals. In addition, we also try to understand which particular features are useful to determine the historical significance in the context of Wikipedia.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
 ISBN 978-1-4503-4073-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983871

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

3. DATASET CREATION

For creating the dataset to be used for feature construction we have downloaded the English Wikipedia dump provided by Wikimedia foundation[2]. DBpedia ontology datasets (PersonData ontology class) [1] were used for selecting Wikipedia articles about persons. We next extracted the core content of articles with BeautifulSoup library[3] excluding lists and common footers.

We then collected hyperlinks using *Yago2* [8] and merged redirecting nodes as well as we removed self-links by excluding links with identical origin and destination. Birth and death dates of persons were also obtained from Yago2. Every date was converted to the decade-level granularity for simplifying computations. In case when a person lacked her birth or death date, the missing date was substituted by the most probable birth or death decade following the procedure used in [9]. In particular, the most probable decade was estimated based on all the persons who were born or who died at the same decade as the one of the target person. Persons without both the dates recorded were necessarily removed. After retaining only people born in and after the 11th century, the final dataset contained 459,991 persons.

4. PREDICTING INFLUENCE

In this section we describe features used for classification and the motivation behind their choice.

4.1 Features

4.1.1 Time-Invariant Importance

A common way to measure person importance is to analyze how it is connected with others. We thus create a directed graph, $G(V,E)$, where V is the set of nodes (or Wikipedia articles) representing persons and E is the set of linking them edges. An edge e_{ij} ($e_{ij} \in E$) from a node v_i to a node v_j ($v_i, v_j \in V$) indicates the presence of a hypertext link in v_i that leads to v_j. We next estimate a node's prestige in G by applying a centrality measure based on the iterative score computation using the random walk theory [11]:

$$\boldsymbol{R} = (1-\alpha)\boldsymbol{M}\times\boldsymbol{R} + \alpha\left[\frac{1}{|V|}\right]_{|V|\times 1} \quad (1)$$

$\boldsymbol{R}$ is a vector containing node scores, $\boldsymbol{M}$ is an aperiodic transition matrix, while α is a decay factor equal to 0.15. Note that since G contains all the nodes (persons) in our dataset, the importance of each node is computed involving persons who lived at different time periods. Hence, we call this measure the *time-invariant person importance*.

4.1.2 Contemporary Importance

We hypothesize that important persons should not only be in general well-connected with other people, but, in many cases, they should have also high prestige within the "social networks" of their times. To measure such *contemporary importance* we simulate a historical social network in a unit time period t_l as a graph, $G^l(V^l,E^l)$. G^l is composed of the set of nodes V^l denoting Wikipedia articles about persons who lived at t_l, and of the edge set E^l including the hypertext links between these articles. Using the century granularity, t_l represents a single century. Eleven *temporal social networks* are then created for the entire time period of analysis, $T=(t_1,\ldots,t_l,\ldots,t_{11})$. Note that a person v_i with its lifespan denoted by $\tau(v_i)$ is assigned to a given century t_l if the midpoint of her life is contained in this century ($2|\tau(v_i) \cap t_l| > |\tau(v_i)|$). In other words, a person belongs to the social network of a given century t_l if the larger part of her life was at t_l.

We compute the node importance using Eq. 1 calculated this time however on each historical social network, G^l. Note that the computed score indicates how prominent a person was among the people of her century, while the score described in Sec. 4.1.1 measures person's prestige among all the people in the dataset, irrespectively of time.

To ensure the across-century comparability, the score of each node in a given century is normalized by dividing it by the total sum of the scores of all the people assigned to that century.

4.1.3 Current Importance

Many influential historical persons are still remembered and are often referred to due to their legacy and perceived significance. We thus quantify to what extent a given historical person is connected with the present times. In particular, following [9], we approximate such connection by estimating the connectivity of historical persons with the "present" persons. As present persons we consider people who were alive during any year included in the period Θ = [1970s, 2000s]. Note that this period can be arbitrarily extended or shortened to represent the notion of the present time.

The past-to-present connectivity measure we want to compute should reflect the closeness of nodes denoting historical persons to the nodes which correspond to the present persons. To compute it, we bias the random walk on G using the static score distribution vector $\boldsymbol{d}$ (see Eq. 2) in a similar way to the one employed in *TrustRank* [5] algorithm used for filtering spam Web pages. We call it the *current importance of persons*.

$$\boldsymbol{R} = (1-\alpha)\boldsymbol{M}\times\boldsymbol{R} + \alpha\boldsymbol{d}$$
$$\boldsymbol{d} = \begin{cases} 1/|d| & if\ \tau(v_i) \cap \Theta \neq \emptyset \\ 0 & if\ \tau(v_i) \cap \Theta = \emptyset \end{cases} \quad (2)$$

4.1.4 Popularity

We consider the view frequency of an article as an effective measure of attention and interest of public in the person described by this article. It is reasonable to assume that many important historical persons are quite well-known and the information on them is often checked and referred to. To quantify the popularity of past persons, we utilize access logs made available by the Wikipedia Foundation[4]. Specifically, we use the number of accesses to all the Wikipedia articles within our dataset that took place during 5 years from January 2009 to December 2013.

4.1.5 Consistency of Popularity

The fact that a person-related article has been viewed continuously with a stable frequency, or, rather, some rare events (e.g., anniversaries, commemorations) triggered sudden rises of its popularity over brief time periods could help to tell if a person is "continuously important" or not. We then measure the consistency of person's popularity using the article view distribution:

$$C(v_i) = -\sum_{j=1}^{m} P_j(v_i) log P_j(v_i) \quad (3)$$

$C(v_i)$ is the entropy of view distribution where $P_j(v_i)$ denotes the normalized access count of an article v_i during j-th month based on the total access data of the article from Jan. 2009 to Dec. 2013.

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

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

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

4.1.6 Content Features

We next include few simple features based on articles and their content.

4.1.6.1 Article Length

We hypothesize that articles about significant persons should have on average more content than ones on less significant individuals. Indeed, the content of articles on salient persons often includes persons' detailed biographies, circumstances of their lives, descriptions of achievements, contributions and relations to other people or to contemporary events, as based on multiple studies by historians, and other professionals. We then use the length of the article content measured in bytes as another feature.

4.1.6.2 In-degree

The number of links pointing to a person is also considered as a simple measure of node importance.

4.1.6.3 Out-degree

Correspondingly, it is reasonable to assume that important persons could have many outgoing links to other persons, similarly to the above assumption regarding the content's length. We use then both the in- and out-degrees as additional classification features.

4.1.7 Distance from Present

We suspect a potential effect of time such as "*the longer time ago a person lived, the more probable is that she or he is considered influential*". This could be supported by the reasoning that the person's importance could arise due to its memory being cherished and, perhaps, also gradually strengthened over long time (e.g., few centuries). To capture the influence of time, we then calculate the number of decades that passed from the decade representing the mid-life point of a person until now.

4.2 Classification Setup

We make use of the two well-known lists of the 100 most influential persons in history: one by M.H. Hart and one compiled by the Life magazine [3]. The former was originally constructed in 1992 forming the basis of the book "*The 100: A Ranking of the Most Influential Persons in History*" [7] which was translated into 15 languages and sold in over half a million copies. The list by the Life magazine was created in 1988 and is available online[5]. After removing persons born before the 11th century we had 55 persons for the Hart's list and 80 for the Life list.

We employ several classifiers using Scikit-learn[6] library with 10-fold cross validation for determining whether a given person belongs or not to the aforementioned lists of important persons. All the classification features listed in Sec. 4.1 have been standardized by subtracting their means and dividing by their standard deviations.

To create instances of the negative class we decided not to select persons randomly as it would overestimate the classification results due to the inherent bias towards more recent decades. That is, randomly selected persons would likely be the ones from the last two centuries due to the present-related bias of the distribution of person-related articles in Wikipedia [9,13,14]. In addition, the random selection would result in picking many less famous persons due to the power low characteristics of Wikipedia articles (e.g., *In-degree*, *Time-Invariant Importance*) making then the classification task relatively easy. Instead, we have selected instances of the negative class by employing the following procedure. For each classification feature (except for the *Distance from Present*) we have selected from each century 100 persons having the top value of this feature, making sure they are not present in either of the above-mentioned lists of the important historical persons. We then removed duplicate persons. The resulting set contains 3,553 people who are likely to be somehow important and famous due to the high values of at least one of their classification features. Note that this set is also balanced over time. We then randomly selected the negative instances for training and testing, repeating each time during the 10-fold cross validation procedure.

5. Results

Table 1 presents the classification results using different classifiers: SVM [15] with linear kernel (SVM Lin.), SVM with RBF kernel (SVM RBF), Naïve Bayes (NB), Nearest Neighbors (NN), Decision Tree (Dec. Tree) and Random Forrest (Ran. Forr.). We can see that, in general, the prediction task is quite successful and it is possible to determine with a reasonably good precision whether a person is or is not in the lists of top influential people. It is also easier to successfully classify instances based on the Life's than based on the Hart's list (likely, due to more persons in the former). An additional observation is that Decision Tree performs best on the later, while the Nearest Neighbors method is the top-performing one on the former.

In Tables 2-3 we show the classification results obtained by using SVM with linear kernel when a given feature is used alone. Additionally, in Tables 4-5 we include the results when a given feature is removed. Due to space limitation we could not show the results by all the classifiers, and, instead, we chose ones given by SVM equipped with linear kernel for its popularity and the frequent usage in text classification research.

Looking at the results we can observe that, somewhat unsurprisingly, *Popularity* used alone is quite a strong predictor of significance (F_1-score of 0.727 and 0.731 for the Hart's and Life's list, respectively). However, it is not sufficient as evidenced by contrasting the results of Tables 2 and 3 with the overall performance shown in Table 1. The entropy of the view distribution (*Consistency of Popularity*) has however much smaller impact (F_1-score of 0.286 and 0.558, respectively), which means that the consistency of the access over time matters less than the total sum of views. Interestingly, the ablation of this feature results in the largest performance drop for the Life list. Also, notably, just only the knowledge of the *Article Length* or *Out-degree* helps to classify persons from the Hart's list with reasonably good success.

When it comes to the link-related metrics such as *Time-Invariant, Contemporary* and *Current Importance* we can observe that *Current Importance* achieves on average better results than the other two link structure dependent measures. This suggests that the connectivity with the present may be a stronger discriminative signal for ranking historical persons than the overall connectivity measure within the whole social graph (*Time-Invariant Importance*) or the one within the people living at the same century (*Contemporary Importance*). The last feature seems to perform worse than the other two which implies that some important historical persons may not be well-known or "connected" within their times. Alternatively, perhaps, our memory and knowledge of their lives might be insufficient.

Overall, the results indicate that to a good extent it is possible to recreate professional judgments by employing automatic means based only on the Wikipedia-derived data. We also notice that using *Time-Invariant Importance* alone, as suggested in the prior literature, is generally insufficient. Finally, while many

[5] http://www.tostepharmd.net/hissoc/top100people.html

[6] http://scikit-learn.org/stable/index.html

observations are consistent for both the used lists, still some results are different (e.g., the effect of *Article Length* and *Distance from Present*) suggesting that it is not easy to definitely reject or accept some of the features.

Table 1 Classification results of precision (P), recall (R) and F1-score (F_1) using the Hart's (H) and Life's (L) lists.

Classifier	P (H)	R (H)	F_1 (H)	P (L)	R (L)	F_1(L)
SVM Lin.	*0.816*	*0.804*	*0.803*	*0.879*	*0.863*	*0.860*
SVM RBF	0.714	0.688	0.685	0.850	0.798	0.788
NB	0.735	0.693	0.683	0.743	0.676	0.653
NN	0.804	0.789	0.784	**0.932**	**0.931**	**0.931**
Dec. Tree	**0.818**	**0.813**	**0.815**	0.901	0.897	0.898
Ran. Forr.	0.810	0.799	0.801	0.915	0.912	0.912

Table 2 Classification results on the Hart's list with SVM (linear kernel) for each feature used alone.

Feature used	Precision	Recall	F_1
Time-Invariant Importance	0.593	0.520	0.437
Current Importance	0.635	0.543	0.476
Contemporary Importance	0.361	0.457	0.300
Article length	0.749	0.710	0.701
In-degree	0.714	0.608	0.572
Out-degree	0.702	0.658	0.644
Popularity	**0.802**	**0.741**	**0.727**
Consistency of Popularity	0.237	0.449	0.286
Dist. from Present	0.648	0.635	0.629

Table 3 Classification results on the Life's list with SVM (linear kernel) for each feature used alone.

Feature used	Precision	Recall	F_1
Time-Invariant Importance	0.685	0.602	0.558
Current Importance	0.738	0.604	0.539
Contemporary Importance	0.531	0.484	0.336
Article length	0.633	0.551	0.489
In-degree	0.731	0.661	0.636
Out-degree	0.550	0.512	0.440
Popularity	**0.812**	**0.744**	**0.731**
Consistency of Popularity	0.685	0.602	0.558
Dist. from Present	0.669	0.646	0.637

Table 4 Classification results on the Hart's list using SVM (linear kernel) with a given feature ablated.

Feature removed	Precision	Recall	F_1
Time-Invariant Importance	0.821	0.812	0.812
Current Importance	0.817	0.806	0.806
Contemporary Importance	0.835	0.824	0.825
Article length	0.788	0.764	0.763
In-degree	0.795	0.781	0.781
Out-degree	0.767	0.726	**0.718**
Popularity	**0.741**	**0.724**	0.724
Consistency of Popularity	0.816	0.803	0.804
Dist. from Present	0.830	0.819	0.818

Table 5 Classification results on the Life's list using SVM (linear kernel) with a given feature ablated.

Feature removed	Precision	Recall	F_1
Time-Invariant Importance	0.882	0.867	0.864
Current Importance	0.878	0.861	0.857
Contemporary Importance	0.881	0.865	0.862
Article length	0.875	0.858	0.856
In-degree	0.872	0.849	0.845
Out-degree	0.875	0.859	0.855
Popularity	0.847	0.830	0.825
Consistency of Popularity	**0.844**	**0.814**	**0.809**
Dist. from Present	0.857	0.822	0.817

6. CONCLUSIONS & FUTURE WORK

Data mining approaches for historical studies have recently gained much interest. In this paper we harness Wikipedia for measuring historical person's significance. In particular, we investigate methods for across-time comparison of personal importance. Based on the two popular lists of significant persons compiled by professionals we study to what extent it is possible to determine person's importance and what are its decisive factors.

In future, we plan to use textual features of articles and include different data sources than Wikipedia. We also want to focus on other entities besides persons such as events or places and on measuring influence with richer set of constraints. The latter could involve (a) **topics** such as science, literature, art (e.g., the most important persons in physics), (b) **spatial areas** such as Asia, Germany, Paris (e.g., the most important persons in Europe) and (c) **time frames** such as middle ages, 1980s, 2013 (e.g., the most important persons in the 19th century), as well as **their combinations** (e.g., the most important persons in physics in Europe in the 19th century).

7. ACKNOWLEDGMENTS

This research was supported in part by MEXT Grants-in-Aid for Scientific Research (15K12158, 15H01718) and 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".

8. REFERENCES

[1] 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*, Springer, 722–735, 2007

[2] 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

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

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

[5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam with Trustrank. In *VLDB2004*, 576-587

[6] C. J. Halperin *et al*. Comparative History in Theory and Practice: A discussion. *The American Historical Review*, 87(1):123–143, 1982

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

[8] J. Hoffart *et al*. YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *WWW '11*, 229-232

[9] A. Jatowt, D. Kawai, K. Tanaka: Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. In *JCDL2016*, 17-26

[10] H. Joho, A. Jatowt and R. Blanco. Temporal Information Searching Behaviour and Strategies. *Inf. Process. Manag*. 51(6): 834-850, 2015

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

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford University*, January 29, 1998

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

[14] 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 *HT '11*. ACM, New York, NY, USA, 83-92

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

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