Information Systems ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Vector-based similarity measurements for historical figures

Yanqing Chen, Bryan Perozzi, Steven Skiena*

Department of Computer Science, Stony Brook University, 100 Nicolls Rd, Stony Brook, NY 11794, United States

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form
25 April 2016
Accepted 1 July 2016

Keywords:
Vector representations
People similarity
Deepwalk

ABSTRACT

Historical interpretation benefits from identifying analogies among famous people: Who are the Lincolns, Einsteins, Hitlers, and Mozarts? As a knowledge source that benefits many applications in language processing and knowledge representation, Wikipedia provides the information we need to make such comparisons. We investigate several approaches to convert the Wikipedia pages of approximately 600,000 historical figures into vector representations to quantify similarity.

On the other hand, Wikipedia pages are assigned to different categories according to their contents as human-annotated labels. A rough similarity estimation could just count the number of shared Wikipedia categories. However, such counting can neither make good similarity quantification (i.e. Is there a difference between those with same number of shared categories?) nor make distinguishable comments on different categories (i.e. Is US Presidents more important than state lawyer when defining similarity?). We use the counting as an indicator to demonstrate high-level agreements of our similarity detection algorithms.

In particular, we investigate four different unsupervised approaches to representing the semantic associations of individuals: (1) TF-IDF, (2) Weighted average of distributed word embedding, (3) LDA Topic analysis and (4) Deepwalk graph embedding from page links. All proved effective, but the Deepwalk embedding yielded an overall accuracy of 88.23% in our evaluation. Combining LDA and Deepwalk yielded even higher performance.

Finally, we demonstrate that our similarity measurements can also be used to recognize the most descriptive Wikipedia categories for historical figures.

We rank the descriptive level of Wikipedia categories according to their categorical coherence, and our ranking yield an overall agreement of 88.27% compared with human crowdsourced data.

1. Introduction

Historical interpretation benefits from identifying analogies among famous people: Who are the Lincolns, Einsteins, Hitlers, and Mozarts of today? Effective analogies should reflect shared personality traits, historical eras, and domains of accomplishment.

Analogies are of course highly subjective, and rest at least partially in the eyes of the beholder: “there are a thousand Hamlets in a thousand people's eyes”. Fig. 1 provides an example, giving reasonable analogies on different aspects of *Isaac Newton*:

We are interested in developing a generalized model to identify analogous figures with a reasonably high similarity level, based on semantics in text and the connections of history. It could be very evocative if correctly identifying analogies like *Martin Luther King* and *Nelson Mandela*; *George Washington* and *Mao Zedong*; *Babe Ruth* and *Sachin Tendulkar*.

In this paper, we propose methods for identifying historical analogies through the large-scale analysis of Wikipedia pages, as well as a reference standard to judge the

* Corresponding author.
E-mail address: skiena@cs.stonybrook.edu (S. Skiena).

http://dx.doi.org/10.1016/j.is.2016.07.001
0306-4379/

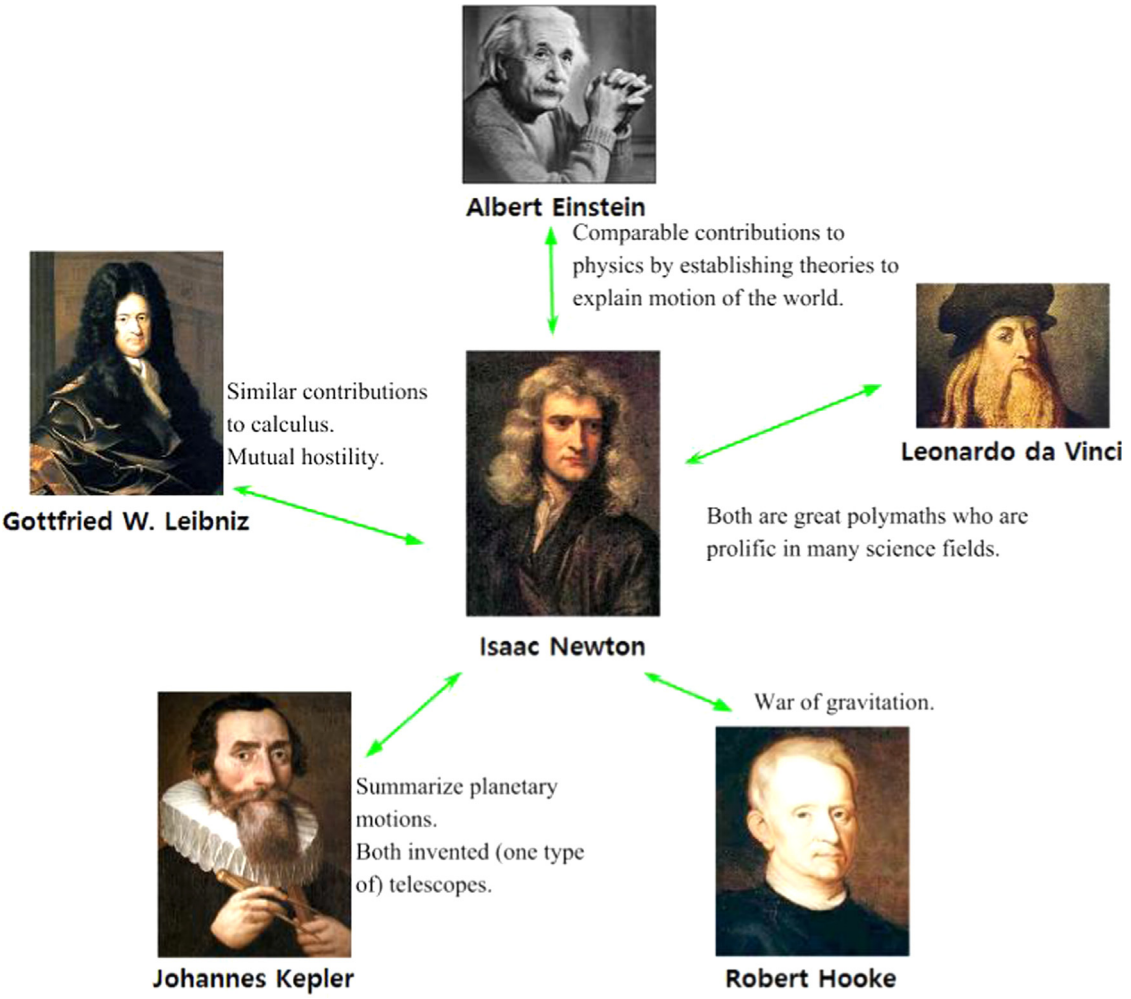

Fig. 1. Analogous historical figures to Isaac Newton, with corresponding explanations of similarity. Analogies are highly subjective, making it impossible to find perfectly fair and objective gold standards.

effectiveness of our methods. The most obvious application of this is in historical interpretation and education, but we believe that the problem runs considerably deeper, since being able to identify similar individuals goes to the heart of algorithms for suggesting friends in social networks, or even matching algorithms pairing up roommates or romantic partners.

Specifically, our work makes the following contributions:

- We propose to use information extracted from Wikipedia categories to be as reference standards to solve this task. Though not perfect, these human-labeled features imply relationships that are shared between similar people. We generated 3,000,000 triples of variable and prescribed difficulty, providing an effective standard to evaluate the performance of our similarity measurement algorithms.
- We investigate four different unsupervised approaches to extract semantic associations from Wikipedia. All proved effective, but our best approach with the graph embedding Deepwalk yielded an overall agreement of 88.23% with human annotated Wikipedia categories. We provide an interactive demonstration of our historical analogies at http://peoplesimilarity.appspot.com/, where you can identify the most similar historical figures to any queried individual.
- We did a careful search to identify the best distance function for each vector model. All metrics yield good results, but highlight different aspects of feature vectors. We also generated a model using linear combination of previously mentioned models to get a better tradeoff between graph structures and text semantics.
- We present a new approach to identifying the most salient categories associated with Wikipedia entities, based on the use of vector representations. We explore several different distance measures and coherence criteria to identify what is best at quantifying descriptive level of Wikipedia categories.
- Through human responses collected from Crowdflower, we verify that our notion of category appropriateness generally jibes with that of human reviewers. Indeed, we have fashioned an iOS game app (FameMatch, available on iTunes) testing how often users agree with our algorithmically selected categorization. We rank

Please cite this article as: Y. Chen, et al., Vector-based similarity measurements for historical figures, Information Systems (2016), http://dx.doi.org/10.1016/j.is.2016.07.001