Web Information Retrieval: Project

Lucia Rodino'
David Ghedalia
Davide Gimondo

Concept

The paper we chose presents a method to find the most influential rock guitarist by applying Google's PageRank algorithm to information extracted from Wikipedia articles. The influence of a guitarist is computed by considering the number of guitarists citing him/her as influence.

The interesting result is the comparison between the rank computed by the algorithm and other ranks drawn up by famous music magazines and experts (in which there are two main processes to create the ranks: music journalists rank their perceived influences, users are asked to vote for their favorite guitarist).

Basically, the experiment consists of building a directed graph where nodes are rock guitarists. There is an outgoing edge from guitarist A to another guitarist B, if guitarist A is influenced by guitarist B.

Kirk Hammet
Joe Satriani

Figure 1: Kirk Hammet is influenced by Joe Satriani

We decided to replicate the experiment with the same methodology, but in a completely different field: philosophy.

The main difference that distinguishes our experiment from the paper's one, is that we did not make use of regular expressions to identify the influenced-influencer pairs.

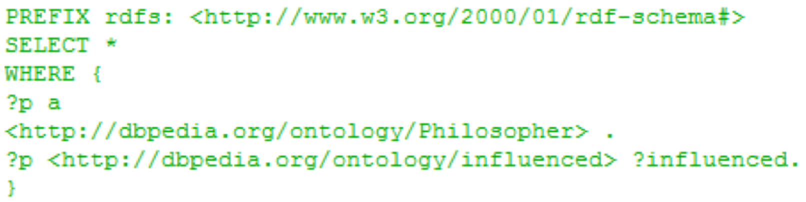
We used two different methods to build the dataset:

- Query DBpedia with SPARQL
- Manual scraping of the data from Wikipedia

In our technical report, these two methods will be treated independently in two sections.

Query DBpedia with SPARQL

The easiest way to query DBpedia is using the SPARQL Explorer for `http://dbpedia.org/sparql` (you can find it at `http://dbpedia.org/snorql/`).



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT *
WHERE {
?p a
<http://dbpedia.org/ontology/Philosopher> .
?p <http://dbpedia.org/ontology/influenced> ?influenced.
}
```

Figure 2: With the query above, we asked the database to provide a list of philosopher pairs influenced-influencer

We included the SPARQLWrapper library (`https://rdflib.github.io/sparqlwrapper/`), which allows us to perform the query and store the result directly in our python script.

With the data stored in this manner, creating the graph is straightforward: it is enough to process every row (each one is constituted of two entries, influenced and influencer) and insert a corresponding edge in the graph.

Manual scraping of the data from Wikipedia

We divided the entire process into three logical phases:

1. Collect in a file all the links to the philosophers that are present in the English version of Wikipedia (the links can be found in the resources section)
2. Check, for each philosopher link, if the relative page contains an Infobox html element including the list of influences and influencers of that given philosopher

3. For philosopher page satisfying this criterion, we write in a file the name of the philosopher, his/her list of influences and influencers

First phase

The English version of Wikipedia contains a list of all the philosophers with an existing article, alphabetically sorted. In the first phase, we simply copy every link in a .txt file.

The total number of philosophers is 1573.

Second phase

Check, for each philosopher link, if the relative page contains an Infobox html element including the list of influences and influencers of that given philosopher.

The total number of philosophers satisfying this requirement is 537.

Third phase

After the second phase, we could have built the graph directly. However, we had to do some manual correction, since there were multiple names associated to the same philosopher. Moreover, some philosophers' influences and influencers also included philosophical schools and/or movements instead of actual philosophers. When the association was obvious, we replaced the name of movement with its major exponent; otherwise, we simply left out the entry. Here, our procedure slightly differs from the one adopted in the original paper: when a guitarist cites a band as an influence, instead of a guitarist, the guitarist with the highest PageRank that has played in the band is chosen).

The manual corrections have been applied to the top-fifty ranked philosophers, otherwise the task would have been too time-consuming.

The final number of philosophers after the third and last phase is 520.

Applying PageRank

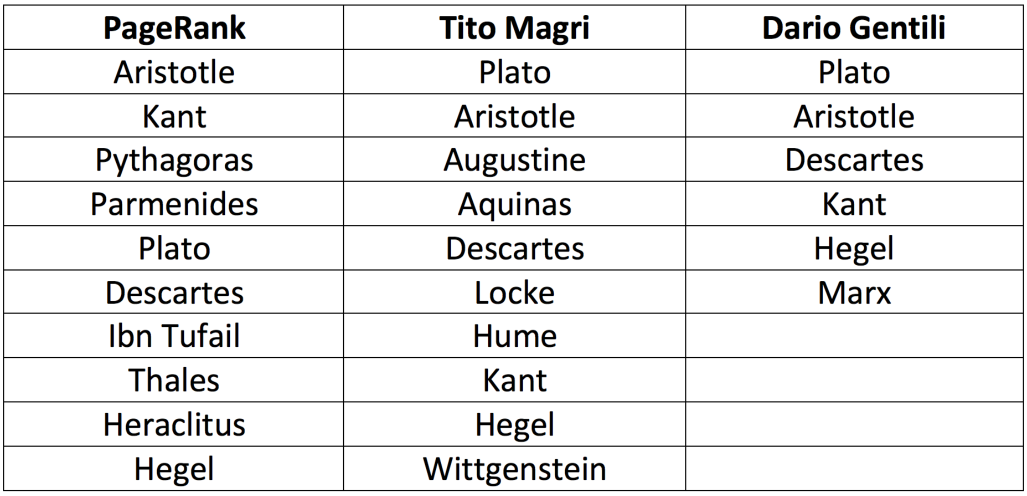
After building the two distinct graphs, we simply applied the PageRank algorithm. To do so, the NetworkX python library has been handy (`https://networkx.github.io/`).

Results

Similarly to the original paper, we wanted to compare the results given by the PageRank algorithm with other qualitative methods. It was not an easy task to do so: unlike rock guitarists, who have been ranked by authoritative sources for the topic in question (Rolling Stone, Guitar Word, Gibson, Telegraph), philosophers have not received the same treatment (several rankings are available on the web, but they are not written by sources that can be considered authoritative).

Therefore, we asked professors of philosophy to draw up an unranked list of who they consider the ten most influential philosophers in history.

In the following table, you can find the comparison of the top ten most influential philosophers determined by our algorithm with Professor Professor Tito Magri (teaching at University ”La Sapienza”) and Dario Gentili (teaching at university ”Roma Tre”).



PageRank	Tito Magri	Dario Gentili
Aristotle	Plato	Plato
Kant	Aristotle	Aristotle
Pythagoras	Augustine	Descartes
Parmenides	Aquinas	Kant
Plato	Descartes	Hegel
Descartes	Locke	Marx
Ibn Tufail	Hume	
Thales	Kant	
Heraclitus	Hegel	
Hegel	Wittgenstein	

Comments

As we can see from the table, the result of PageRank algorithm does not differ too much from the responses of the two philosophy professors (the are five philosophers that appear in all the three rankings); however, the complete ranking computed by PageRank shows a slight bias over the more ancient philosophers.

Applying personalized PageRank

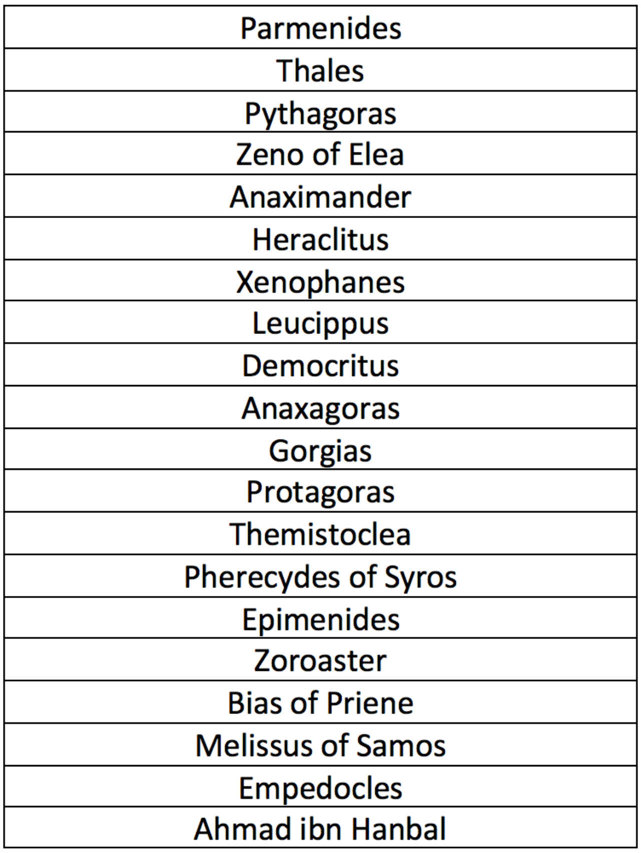
Another experiment was conducted, the topic-specific PageRank. An interest towards Pre-Socratic philosophy has been simulated; to do this, a list from Wikipedia containing almost all pre-Socratic philosophers was consulted (`https://en.wikipedia.org/wiki/Category:Presocratic_philosophers`).

Then, a personalization vector consisting of a dictionary with a key for every graph node representing a Pre-Socratic philosopher and nonzero personalization value for each node has been initialized and added to the inputs of the PageRank method (the nodes not included in the dictionary had zero as personalization value).

The chosen graph was the one resulting from the manual data scraping from Wikipedia.

Results

Here there are the top twenty results of the topic-specific PageRank.



Parmenides
Thales
Pythagoras
Zeno of Elea
Anaximander
Heraclitus
Xenophanes
Leucippus
Democritus
Anaxagoras
Gorgias
Protagoras
Themistoclea
Pherecydes of Syros
Epimenides
Zoroaster
Bias of Priene
Melissus of Samos
Empedocles
Ahmad ibn Hanbal

Comments

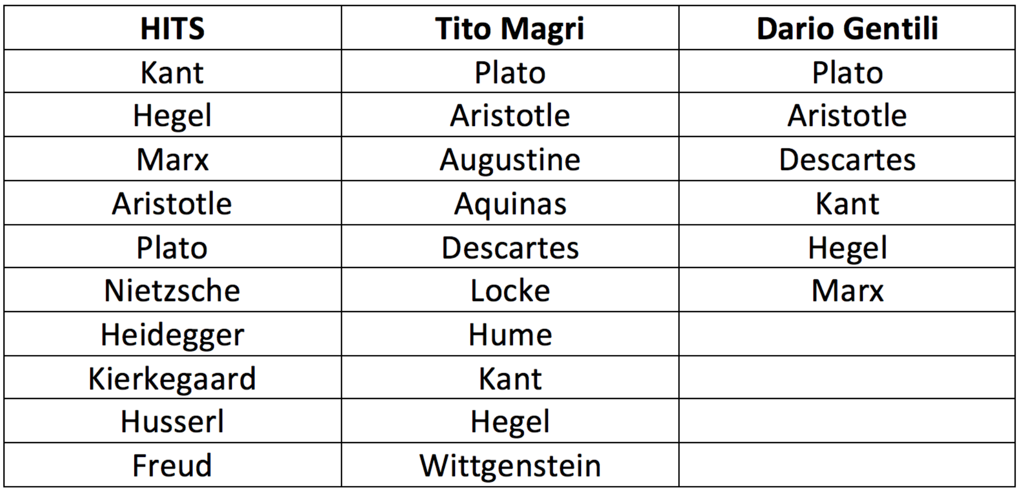
As expected, all the most important Pre-Socratic philosophers appear in the list. However, a more interesting fact is that in the top fifty results there is a considerable prevalence of Islamic and eastern philosophers, who were heavily influenced by their Greek predecessors.

Applying HITS

In the last experiment, the HITS algorithm has been applied. The NetworkX Python library contains the HITS algorithm too. Again, the chosen graph was the one resulting from the manual data scraping from Wikipedia.

Results

In the following table, you can find the comparison of the top ten most influential philosophers determined by HITS (of course, authority ranking has been taken in consideration for this purpose) with Professor Tito Magri (teaching at University ”La Sapienza”) and Dario Gentili (teaching at university ”Roma Tre”).



HITS	Tito Magri	Dario Gentili
Kant	Plato	Plato
Hegel	Aristotle	Aristotle
Marx	Augustine	Descartes
Aristotle	Aquinas	Kant
Plato	Descartes	Hegel
Nietzsche	Locke	Marx
Heidegger	Hume	
Kierkegaard	Kant	
Husserl	Hegel	
Freud	Wittgenstein	

Comments

Compared to the PageRank results, the HITS algorithm returns a top ten (and even more) that includes more recent philosophers in larger number.

Conclusions

The experiment consisted in applying the same procedure descripted in the original paper to philosophers can be considered successful: the discrepancy between the rankings is minimal, especially if the top-twenty list is taken in consideration (instead of the top-ten list only). The complete rankings can be consulted in the Github page. A major difficulty in the experiment has been polishing the data in order to provide reliable results. For this very reason, the dataset obtained with DBpedia was not used in later tests with HITS and personalized PageRank. A greater number of rankings provided by other philosophers would have made the results of the experiment more interesting.

Contacts

Here is the GitHub link of our project: `https://github.com/davegimo/philosophers`