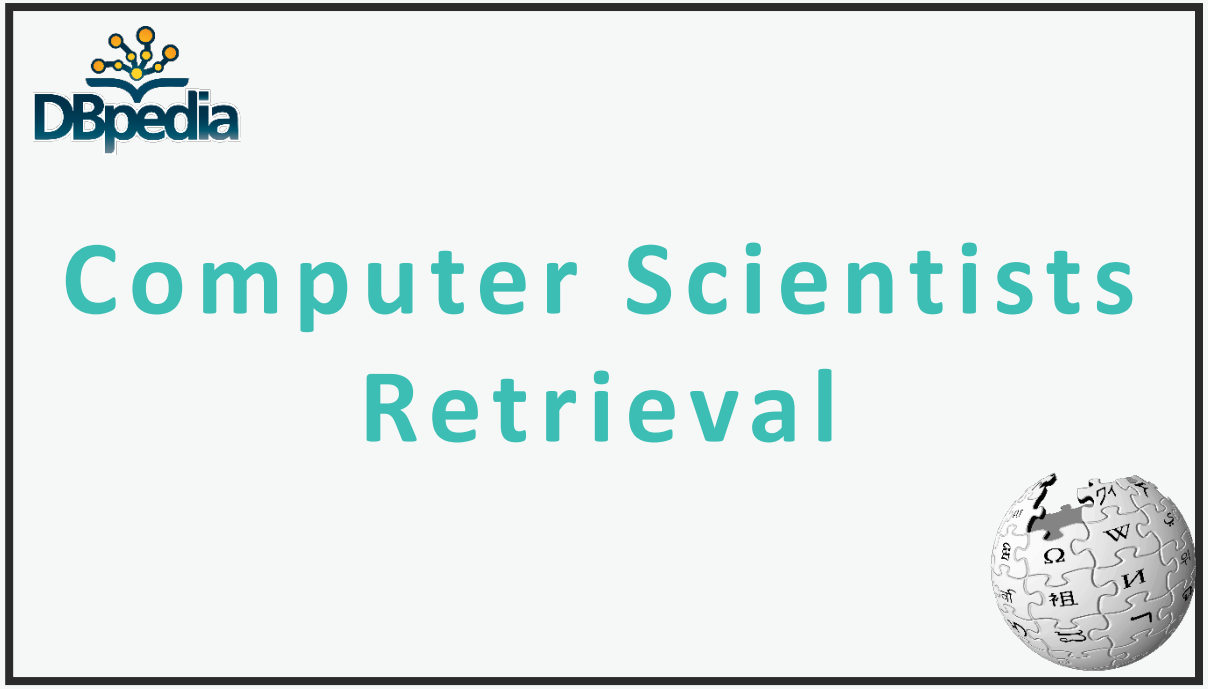

DBpedia
Computer Scientists
Retrieval

Our Team

Luca Tomei	1759275
Daniele Iacomini	1706790
Andrea Aurizi	1706890

Outline of Talk

Motivation & Related Work

Methodology & Experimental Design

Results

Conclusions and Future Works

I.J. Intelligent Systems and Applications, 2015, 12, 50-56
Published Online November 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.12.05

Mining Wikipedia to Rank Rock Guitarists

Muazzam A. Siddiqui
Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
Email: maasiddiqui@kau.edu.sa

Abstract—We present a method to find the most influential rock guitarist by applying Google PageRank algorithm to information extracted from Wikipedia articles. The influence of a guitarist was estimated by the number of guitarists citing him/her as an influence and the influence of the latter. We extracted this who-influenced-whom data from the Wikipedia biographies and converted them to a directed graph where a node represented a guitarist and an edge between two nodes indicated the influence of one guitarist over the other. Next we used Google PageRank algorithm to rank the guitarists. The results are most interesting and provide a quantitative foundation to the idea that most of the contemporary rock guitarists are influenced by early blues guitarists. Although no direct comparison exist, the list was still validated against a number of other best-of lists available online and found to be mostly compatible.

Index Terms—Wikipedia mining, PageRank for people, information extraction, text mining, music data mining.

I. INTRODUCTION

Music artists are ranked based upon a variety of criteria such as their popularity, skill level, album sales etc. These ranks are important to the artists themselves as they result into an increased fan base and popularity, and to the fans, as the latter would like to see their favorite musicians at the top spots. Like other musicians, guitarists are ranked based upon their creativity, skill level at the instrument and their influence over other guitarists as well as the genre as a whole. A number of such *best-of* lists are available on the Internet. These lists are primarily generated through crowdsourcing where fans vote for their favorite artists and/or compiled by subject matter experts such as music journalists, critics or guitarists themselves. These lists have always been controversial and a source of argument among fans when they do not find their favorite artist in the position they were expecting them to be. In this paper we combined techniques from information extraction and graph mining to find the most influential rock guitarists. The influence of a guitarist was computed by considering the number of guitarists citing him/her as an influence and, in turn, their own influences. This information about influences is available in the biographical sketches on Wikipedia of these guitarists. The Wikipedia page for most of the guitarists lists the guitarists who influenced their playing. The information is usually available within the article in an unstructured form such as *X cites* X_1, X_2, ..., X_n *as influences*. We extracted this information from the Wikipedia pages, identified the influencer and the influencee and converted this to a directed graph where nodes represented guitarists and edges represented the influence relationship. The presented work makes two main contributions:

1. Using a quantitative method to find the most influential guitarist
2. Estimation of influence from the guitarist community itself, instead of fans

It should be noted that our method finds the most influential guitarists and not the best guitarist. The latter would require measurement of different performance indicators. Another important point to note is that the current work includes the guitarist articles in English Wikipedia only, but the techniques presented here can be easily modified to incorporate Wikipedia articles in other languages and other categories such as influential philosophers, musicians etc.

This paper is organized as follows. A review of related work is presented in section II. Section III describes the corpus creation process from Wikipedia. Extraction of influencee, influencer pairs is described in section IV. Section V briefly describes PageRank and its usage to rank guitarists. Results are presented in section VI.

II. RELATED WORK

A number of magazines related to music or otherwise have published their own lists of best guitarist. These include Rolling Stone, Time, Telegraph, Esquire, Guitar World, Revolver Mag etc. These lists are essentially generated manually using one or a combination of the following methods:

1. Music journalists rank the guitarists based upon their perceived influences
2. Users are asked to vote for their favorite guitarist
3. Guitarists are asked to vote for their favorite

A. The Lists

A brief overview of these lists is provided below. A comparison of results will be provided in the later section of this paper.

1) Music Expert Compilation

Motivation & Related Work:
Ranking Guitarists

Apply Google PageRank to Wikipedia guitarists articles to rank them based on influence they had on each other

Key Points

1) Collecting data about guitarists and their influences from Wikipedia;
2) Converting this data into a directed graph, where a node represents a guitarist, and an edge from A to B represents the influence from A to B;
3) Applying Google PageRank algorithm to rank the guitarists.

Goals

1) Using a quantitative method to find the most influential guitarist
2) Estimation of influence from the guitarist community itself, instead of fans

Methodology - Our Approach

- Replicate the paper by finding the most influential **computer scientists**

- Dataset is computed by SPARQL queries, to retrieve computer scientist influenced and influencers, and by scraping the Wikipedia pages of each computer scientist

- Results are computed using two different methods:
    - Blind meta-data link picking
    - “Influences” section analysis

Methodology:

DBpedia Dataset Construction

- Retrieve all the computer scientists from DBpedia

SPARQL queries

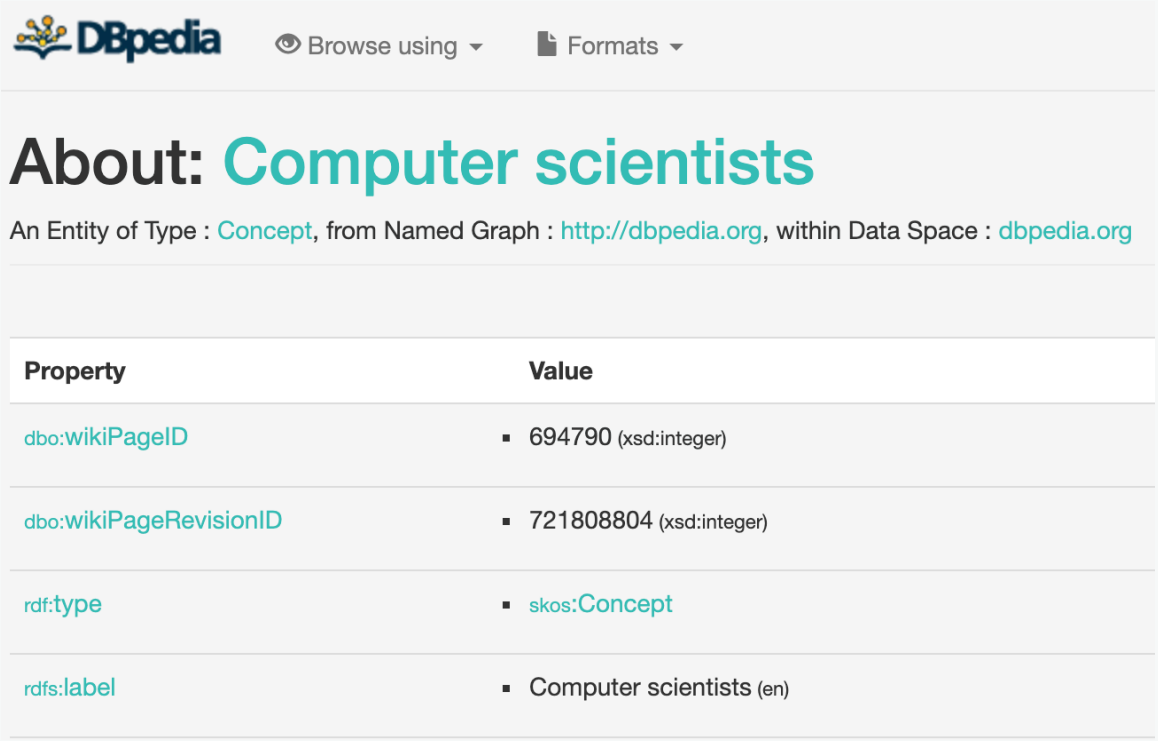
↓

Lists of computer scientist

↓

Few results

- DBpedia returns to the Wikipedia page



As the figure above shows, DBpedia does not retrieve results for computer scientists, because there is no dataset on them.

In DBpedia the majority of computer scientists is assigned to a generic class *dbpedia:Person* and not distinguished from other people.

Methodology:

Manual Wikipedia scraping

01

Collect all C.S. Links

Collect in a file all the links of computer scientists present on Wikipedia

02

Check biography table

Check whether the relative page contains information about his influences

03

Computing Pagerank

Quantify the importance of computer scientists

04

C.S. Fields Classification

Classification of the various fields of study of a computer scientist

Methodology:

Wikipedia Dataset Construction

- Retrieve all the computer scientists from Wikipedia
- Associate each computer scientist in order to consider his influences
- Download each Wikipedia page associated to each computer scientist

Python scripts are being used in order to provide the lists and download the pages, resulting in **509** different **computer scientist**.

Lists of computer scientists

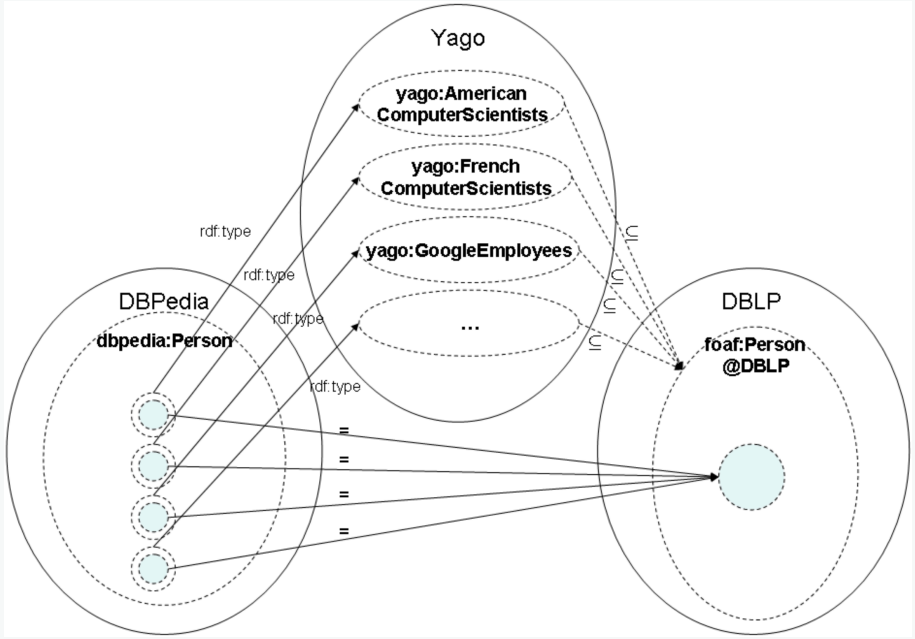
↓

Retrieve Computer Scientists Wikipedia pages

Applying Google PageRank algorithm to rank computer scientists.

Computer Scientist branches classification

VS



Methodology – Methods Used

1) Blind meta-data link picking

- Consider link structure between Wikipedia pages of computer scientists to build directed graph

- **Assumption :** if there exists a link from page of computer scientist s_1 to page of computer scientist s_2 then s_1 was influenced by s_2

2) "Influences" section analysis

- Scraping of Wikipedia computer scientists' pages to get "influences" section

- **Assumption:** if computer scientist s_2 is mentioned in the "influence" section of computer scientist s_1 then s_1 was influenced by s_2

Results & Personalized Pagerank

Donald_Knuth	0.08372
Rudy_Rucker	0.07940
Fred_Brooks	0.07389
Adi_Shamir	0.06984
Douglas_Engelbart	0.06941
Allen_Newell	0.06934
Stephen_Wolfram	0.06704
Alan_Kay	0.06534
Niklaus_Wirth	0.06421
Alan_Perlis	0.06368
E._Allen_Emerson	0.06327
Herbert_A._Simon	0.06236
Dennis_Ritchie	0.06233
Edmund_M._Clarke	0.06215
Amir_Pnueli	0.06165
Dana_Scott	0.06021
John_McCarthy	0.05920
John_Cocke	0.05905
Barbara_Liskov	0.05878
Charles_Bachman	0.05854

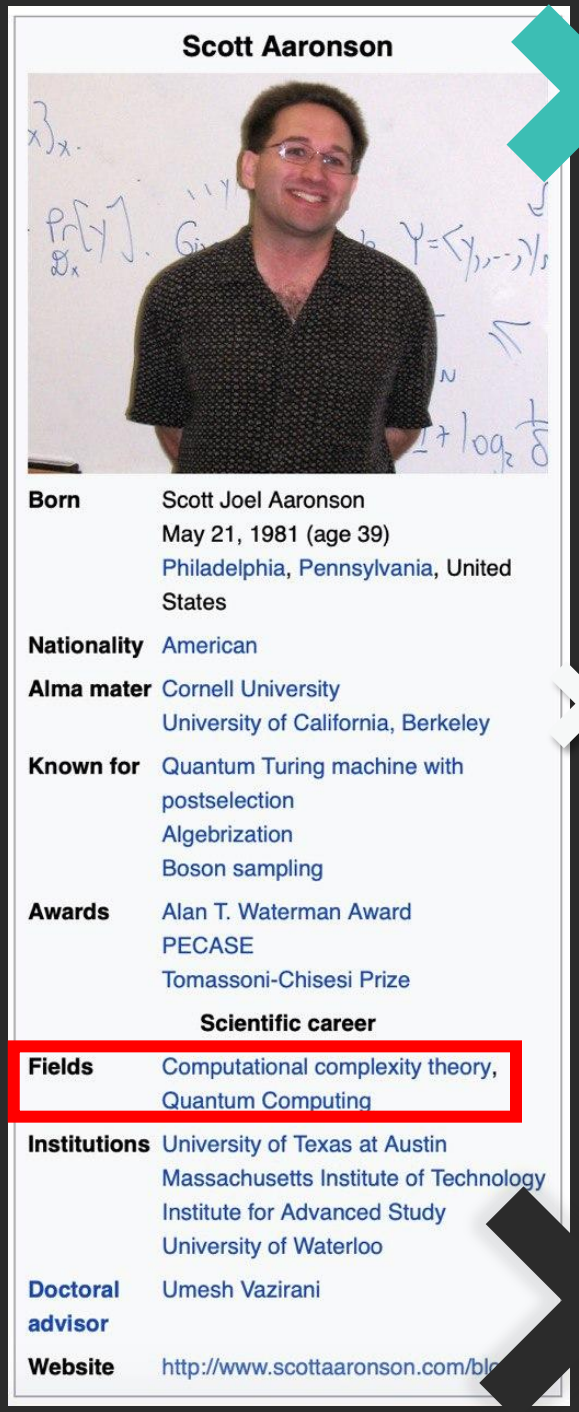
Personalized Pagerank

Donald_Knuth	0.01353
Vint_Cerf	0.01129
Niklaus_Wirth	0.01079
Fred_Brooks	0.01050
Silvio_Micali	0.01032
Shafi_Goldwasser	0.01005
Marvin_Minsky	0.00987
Douglas_Engelbart	0.00981
Leslie_Valiant	0.00957
Allen_Newell	0.00956
Adi_Shamir	0.00918
Alan_Kay	0.00916
John_McCarthy	0.00911
Herbert_A._Simon	0.00910
Robert_Tarjan	0.00909
John_Hopcroft	0.00898
Richard_Karp	0.00891
Dennis_Ritchie	0.00890
Stephen_Wolfram	0.00882
John_Cocke	0.00861

NetworkX Pagerank

Robert_Tarjan	0.01675
Donald_Knuth	0.01673
Adi_Shamir	0.01672
Michael_O._Rabin	0.01672
E._Allen_Emerson	0.01671
Ron_Rivest	0.01671
Leonard_Adleman	0.01671
Edmund_M._Clarke	0.01671
Allen_Newell	0.01667
John_McCarthy	0.01666
Herbert_A._Simon	0.01665
Silvio_Micali	0.01664
Shafi_Goldwasser	0.01662
Richard_Karp	0.01661
John_Cocke	0.01661
Niklaus_Wirth	0.01654
Fred_Brooks	0.01650
Douglas_Engelbart	0.01649
Vint_Cerf	0.01648
Leslie_Valiant	0.01647

NetworkX Hits



Another Approach:

Computer Scientists Branches

A further experiment was to try to draw up a ranking of the best branches of study carried out by these people.

In evaluating the fields that can be used, it has been verified that the majority of computer scientists present in the famous Wikipedia *infobox table* a field called *Field* which is right for us: it contains every category of study carried out from the person being examined.

Applying Pagerank and HITS on Branches

What We Can Do

Compare the results given by the Pagerank algorithm with other qualitative methods

Computer science	0.04921
Mathematics	0.02152
Artificial intelligence	0.01889
Human-computer interaction	0.01111
Theoretical computer science	0.00921
Logic	0.00838
Semantic web	0.00812
Machine learning	0.00812
Robotics	0.00786
Electrical engineering	0.00776
Entrepreneur	0.00673
Statistics	0.00673
Computer engineering	0.00673
Computational information systems	0.00673
Cognitive science	0.00658
Cryptography	0.00622
Parallel computing	0.00622
Computer graphics	0.00622
Operating systems	0.00622
Engineering	0.00596

NetworkX Pagerank

Computer science	0.32612
Mathematics	0.0905
Artificial intelligence	0.06685
Logic	0.03428
Electrical engineering	0.02833
Human-computer interaction	0.02135
Cognitive psychology	0.01891
Internet	0.01787
Cryptography	0.01739
Engineering	0.01721
Parallel computing	0.01688
Computer engineering	0.01667
Cognitive science	0.01277
Machine learning	0.012120
Theoretical biology	0.01152
Cryptanalysis	0.01152
Complex systems	0.0111
Political science	0.01052
Economics	0.01052
Biology	0.01038

NetworkX HITS

Conclusions

- Successful appliance of procedures descripted in the original paper to computer scientists
- The discrepancy between the rankings is minimal
- Major difficulty in the experiment: polishing the data in order to provide reliable results

Future Works

- University Computer Scientists attended
- University Computer Scientists has worked as a teacher or researcher
- Best university ranking

References

[1] Mining Wikipedia to Rank Rock Guitarists, http://www.mecs-press.org/ijisa/ijisa-v7-n12/IJISA-V7-N12-5.pdf

[2] Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution, 2009 Springer-Verlag Berlin Heidelberg, http://oro.open.ac.uk/23438/5/23438.pdf

Our Team

Luca Tomei	1759275
Daniele Iacomini	1706790
Andrea Aurizi	1706890

Thanks