Apply Page-Rank on Wikipedia Singers

Daniele Buonadonna, Eleonora Lopez, Umberto di Canito

Contributions

■ **Daniele**: SPARQL queries, Dataset creation (based on queries), first method PageRank

■ **Eleonora**: Named Entity Recognition, second method PageRank

■ **Umberto**: Dataset creation (based on scraping Wikipedia pages), retrieving “influences” section

Reference paper

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 influencer, influencee 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

Muazzam A. Siddiqui

Mining Wikipedia to Rank Rock Guitarists

International Journal of Intelligent Systems and Applications (IJISA),

vol.7, no.12, pp.50-56, 2015. DOI: 10.5815/ijisa.2015.12.05

Reference paper - main ideas

- Find the most influential rock guitarists
- Use quantitative method instead of subjective opinions of people
- Compute graph by analyzing data from the guitarist himself
- Finally, compute PageRank and compare the results to major ranking lists available online

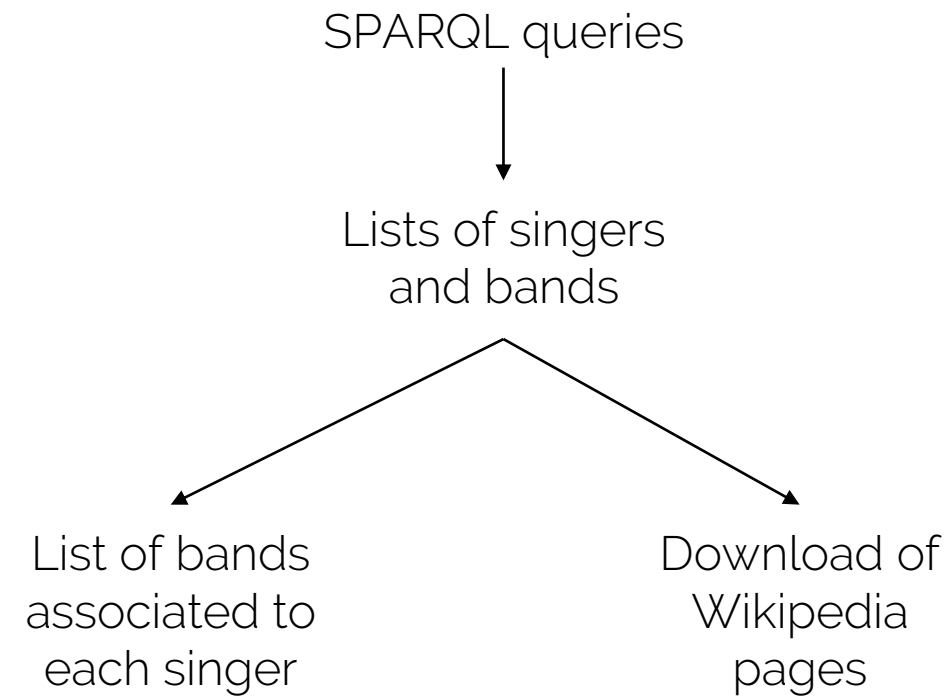
Our project - main ideas

- Replicate the paper by finding the most influential **singers**
- Dataset is computed by SPARQL queries (to retrieve singers' name and bands' name) and by scraping the Wikipedia pages of each singer
- Results are computed using two different methods:
  - Blind meta-data link picking
  - "Influences" section analysis

Dataset construction

- Retrieve all the singers and bands from DBPedia
- Associate each singer to his bands in order to consider the band's influence
- Download each wikipedia page associated to each singer

Python scripts are being used in order to provide the lists and download the pages, resulting in **4581** different **singers** and **2471** different **bands**.



Methods used

1) **Blind meta-data link picking**

- Consider link structure between Wikipedia pages of singers to build directed graph
- **Hypothesis:** if there exists a link from page of singer s1 to page of singer s2 then s1 was influenced by s2

2) **“Influences” section analysis**

- Scraping of Wikipedia singers’ pages to get “influences” section
- Apply NER (Named Entity Recognition) to get singers and bands mentioned in the “influences” section
- **Hypothesis:** if singer s2 is mentioned in the “influence” section of singer s1 then s1 was influenced by s2

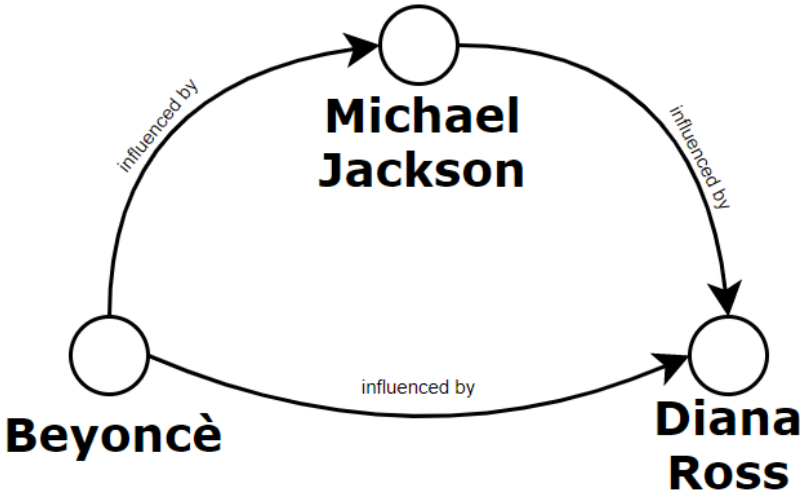
Is the “blind” method reliable?

Compared PageRank result of our method to the result of the reference paper.

Paper	index	Blind method
Jimi Hendrix	1	Jimi Hendrix
Charlie Christian	2	Eric Clapton
Josh White	3	Josh White
Hank Marvin	4	John Lennon
Eric Clapton	5	Elvis Presley
Jimmy Page	6	George Harrison
Django Reinhardt	7	Neil Young
Lonnie Johnson	8	Django Reinhardt
Wes Montgomery	9	Jimmy Page
John Lennon	10	Johnny Cash

The graph

- There is a node for every singer in the list
- There is an edge from singer s1 to singer s2 if s1 was influenced by s2



Other hypotheses

- If a singer is considered to be influenced by a band

then we consider such singer to be influenced by the band's singers

- For the second method, if a singer does not have the "influences" section

then we consider such singer to not have influences

Index	Blind method (1st)	Scraping method (2nd)	The Rolling Stone	Ranker.com	thetoptens.com
1	Elton John	Michael Jackson	Aretha Franklin	Michael Jackson	Michael Jackson
2	Michael Jackson	Aretha Franklin	Elvis Presley	Frank Sinatra	Elvis Presley
3	Elvis Presley	James Brown	John Lennon	Elvis Presley	Freddie Mercury
4	Frank Sinatra	Diana Ross	Marvin Gaye	Aretha Franklin	John Lennon
5	David Bowie	Madonna	James Brown	Adele	Madonna
6	John Lennon	David Bowie	Little Richard	Elton John	David Bowie
7	Eric Clapton	Celine Dion	Mick Jagger	Celine Dion	Paul McCartney
8	Bruce Springsteen	Mariah Carey	Bob Marley	John Lennon	Whitney Houston
9	Madonna	B.B. King	Johnny Cash	David Bowie	Mariah Carey
10	Mick Jagger	John Lennon	David Bowie	Marvin Gaye	Kurt Cobain

Discussion of results

- Results are quite different for each list but compatible
- It is not easy to make a comparison since we do not know the criteria used by the creators of the online lists
- We expected that the scraping method would perform better than the blind method but in the end they both yielded good results
- Elvis Presley appears in all the lists except the scraping method's list, the reason is that the singers influenced by Presley are not many and not so important

Libraries used

Dataset building

- SPARQL Wrapper
  https://pypi.org/project/SPARQLWrapper/
- PyWebCopy 6.0.0
  https://pypi.org/project/pywebcopy/

Defining graph

- HTMLParser
  https://pypi.org/project/HTMLParser/
- StanfordCoreNlp
  https://pypi.org/project/stanfordcorenlp/
- Wikipedia-API 0.5.1
  https://pypi.org/project/Wikipedia-API/

Page Rank computation

- Networkx 2.3
  https://pypi.org/project/networkx/

Thank you