Wikipedia link graph as a knowledge base for identifying connections between concepts

Marek Ciglan

Dep. of Computer and Information Science
NTNU, Trondheim, Norway.
marek.ciglan@idi.ntnu.no
Outline

- Motivation – finding interesting connections in Wikipedia link graph
- Why straightforward approach does not work
- Approach
  - Spreading activation over weighted link graph
- Application – WikiPop
- Possible extensions & Future work
  - Construction of concept network from document collections
  - Graph mining algorithms for mining concept networks
Motivation

- Wikipedia – valuable source of the data for research in computer science
  - Natural language processing
  - Enriching information retrieval systems
  - Ontology building
- Unique structure:
  - Each article describes a concept
  - Links defines relations between concepts
  - Articles organized in category system
Motivation

- Using Wikipedia link graph, we want to answer question:
  - what is the connection between two or more concepts?
  - what is common and what connects given input concepts
Motivation

- Using Wikipedia link graph, we want to answer question:
  - what is the connection between two or more concepts?
  - what is common and what connects given input concepts
- Straightforward solution:
  - Take Wikipedia link graph
  - Find shortest paths between input nodes
Shortest paths example

Iggy Pop

Joy Division
## Shortest paths example

<table>
<thead>
<tr>
<th>Iggy Pop</th>
<th>Joy Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980 in music</td>
<td></td>
</tr>
<tr>
<td>24 Hour Party People</td>
<td></td>
</tr>
<tr>
<td>Punk rock</td>
<td></td>
</tr>
<tr>
<td>Ian Curtis</td>
<td></td>
</tr>
<tr>
<td>Trainspotting (soundtrack)</td>
<td></td>
</tr>
<tr>
<td>Timeline of punk rock</td>
<td></td>
</tr>
<tr>
<td>Rock music</td>
<td></td>
</tr>
<tr>
<td>Control (2007 film)</td>
<td></td>
</tr>
<tr>
<td>David Bowie</td>
<td></td>
</tr>
<tr>
<td>Rage (TV program)</td>
<td></td>
</tr>
<tr>
<td>List of NME covers</td>
<td></td>
</tr>
<tr>
<td>List of SingStar titles</td>
<td></td>
</tr>
<tr>
<td>So It Goes (TV series)</td>
<td></td>
</tr>
<tr>
<td>Gothic rock</td>
<td></td>
</tr>
<tr>
<td>Biographical film</td>
<td></td>
</tr>
<tr>
<td>RCA Records</td>
<td></td>
</tr>
<tr>
<td>Virgin Records</td>
<td></td>
</tr>
<tr>
<td>List of bands named after other performers' songs</td>
<td></td>
</tr>
<tr>
<td>List of lead vocalists</td>
<td></td>
</tr>
<tr>
<td>Bass guitar</td>
<td></td>
</tr>
<tr>
<td>Perry Farrell</td>
<td></td>
</tr>
<tr>
<td>Australian rock</td>
<td></td>
</tr>
<tr>
<td>Industrial music</td>
<td></td>
</tr>
<tr>
<td>Red Hot Chili Peppers</td>
<td></td>
</tr>
<tr>
<td>Noise rock</td>
<td></td>
</tr>
<tr>
<td>New Wave music</td>
<td></td>
</tr>
</tbody>
</table>
### Shortest paths example

1. **1980 in music**
   - 24 Hour Party People
   - Punk rock
   - Ian Curtis
   - Trains
   - Timeline
   - Rock music
   - Control
   - David
   - Rage
   - List of songs
   - So It Goes
   - Gothic
   - Biographical film
   - RCA Records
   - Virgin Records
   - List of bands named after other performers' songs
   - List of lead vocalists
   - Bass guitar
   - Perry Farrell
   - Australian rock
   - Industrial music
   - Red Hot Chili Peppers
   - Noise rock
   - New Wave music

### Problems:
- sheer number of results
- no way to rank paths
- semantic inconsistency of the connections
Shortest paths example

Problems:
- sheer number of results
- no way to rank paths
- semantic inconsistency of the connections

Iggy Pop -> Rock and roll
Rock and roll -> Pet Shop Boys

List of SingStar titles -> Pet Shop Boys
Joy Division -> List of SingStar titles
Properties of Small-World nets

- Power law degree distribution
Properties of Small-World nets

- Power law degree distribution
Properties of Small-World nets

- Power law degree distribution
Properties of Small-World nets

- Power law degree distribution
- Short mean-shortest path length
  - e.g. 6 degrees of separation of Kevin Bacon
Properties of Small-World nets

- Power law degree distribution
- Short mean-shortest path length
  - e.g. 6 degrees of separation of Kevin Bacon
- High clustering coefficient

\[ C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E. \]
Properties of Small-World nets

Fig. Source: https://elgg.leeds.ac.uk/soc6tjw/weblog/15764.html
Spreading activation

- Designed for searching semantic and association networks
- Nodes – represent objects
- Edges – represent relations between objects

Algorithm (short, informal description)
- Set initial activation value for input node(s)
- Spread activation along weighted/typed edges from activated nodes
- Utilize breadth first expansion
- Output: nodes with high activation values > threshold
Spreading activation
Spreading activation

Initial node
Spreading activation

1st iteration of activation spreading
Spreading activation

1st iteration of activation spreading
Spreading activation

2\textsuperscript{st} iteration of activation spreading
Spreading activation

2\textsuperscript{nd} iteration of activation spreading
Activation vector spreading (AVS)

- Standard SA – activation is a number
- AVS – activation vector instead of a single activation value
- Each element of activation vector represents amount of activation received from different initial nodes
Activation vector spreading (AVS)
Activation vector spreading (AVS)
Activation vector spreading (AVS)
Spreading activation example

Iggy Pop

Joy Division
## Spreading activation example

<table>
<thead>
<tr>
<th>Iggy Pop</th>
<th>Joy Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Bowie</td>
<td>United States</td>
</tr>
<tr>
<td>Punk rock</td>
<td>Punk rock</td>
</tr>
<tr>
<td>Rock music</td>
<td>Rock music</td>
</tr>
<tr>
<td>The Stooges</td>
<td>The Stooges</td>
</tr>
<tr>
<td>Allmusic</td>
<td>Allmusic</td>
</tr>
<tr>
<td>Rolling Stone</td>
<td>Rolling Stone</td>
</tr>
<tr>
<td>Ramones</td>
<td>Ramones</td>
</tr>
<tr>
<td>Bass guitar</td>
<td>Bass guitar</td>
</tr>
<tr>
<td>Singing</td>
<td>Singing</td>
</tr>
<tr>
<td>London</td>
<td>London</td>
</tr>
<tr>
<td>Alternative rock</td>
<td>Alternative rock</td>
</tr>
<tr>
<td>U2</td>
<td>U2</td>
</tr>
<tr>
<td>Guitar</td>
<td>Guitar</td>
</tr>
<tr>
<td>Sex Pistols</td>
<td>Sex Pistols</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>United Kingdom</td>
</tr>
</tbody>
</table>
Activation vector spreading (AVS)

- Problem:
  - How to weight links between Wikipedia articles?

- When using constant edge weights:
  - Top scoring results contain connecting nodes of a very general nature
  - Obvious answers, not of high interest for user
  - E.g.:
    - Q: What is the connection between Peter Sellers and Monty Python?
Edge weighting – CatTax

- Use the semantics describing nodes to weight the strength of an edge connecting them
- We use Wikipedia categories assigned to an article as a description of the node’s content
- Weight edge according to the similarity of nodes categories

- Challenges:
  - How to adapt semantic relatedness measures proposed for lexical semantic networks to the Wikipedia category graph
  - How to clean the Wikipedia category graph
Edge weighting – Indegree Square Ratio (ISR)

- Minimize the appearance of very general connections
- How to identify whether an article is more general than other
- Very general concepts tend to have a high indegree (number of incoming links)

\[
ISR(e(i, j)) = \begin{cases} 
\frac{\text{indegree}(i)^2}{\text{indegree}(j)^2} & \text{iff } \frac{\text{indegree}(i)^2}{\text{indegree}(j)^2} < 1 \\
1 & \text{iff } \frac{\text{indegree}(i)^2}{\text{indegree}(j)^2} \geq 1
\end{cases}
\]
SA on weighted link graph
Evaluation

- Identification of interesting connections in Wikipedia is a recommendation system
- User evaluation
  - Users were comparing three weighting methods
    - ISR
    - CatTax
    - Random Weighting
Progress: 0/8

Evaluate relation sets between:

Michael Jackson & Pink (singer)

Chose the most obvious set.

Result No: 0
Pink (singer) -> The Brit Awards
| Brit Awards | 2003 Best International Female Artist | Won

The Brit Awards -> Michael Jackson
*1989: [[Phil Collins]], [[Michael Jackson]], [[Fairground Attraction]] and [[Tracy Chapman]] win 2 awards.

Result No: 1
Michael Jackson -> Mysa

Pink (singer) -> Mysa
*In 2001, they recorded a cover of [[Lil' Kim]]'s 1997 single "Lady Marmalade" with [[Christina Aguilera]], rapper Lil' Kim and [[Mysa]] for the soundtrack of the film "Wu-Tang: An American Saga".

Result No: 2
Michael Jackson -> Justin Timberlake

Pink (singer) -> Justin Timberlake
| Pink (singer) | > | Justin Timberlake | Pink collaborated with several other artists in 2006 and 2007, when she opened for [[Justin Timberlake]] on the American leg of his [[FutureSex/LoveShow]] Tour.

Result No: 3
Michael Jackson -> Janet Jackson
| Michael Jackson | -> | Janet Jackson | Jackson had three sisters, [[Rebbie Jackson|Rebbie]], [[LaToya Jackson|LaToya]], and [[Janet Jackson]], and five brothers, [[Marlon Jackson|Marlon]], and [[Tito Jackson|Tito]]] (Trevor), [[Elvis Presley|Elvis]], [[Genevieve Jackson|Geneviève]], [[Mariah Jackson|Mariah]], and [[Kandy Jackson|Kandy]]] (Kandy). refname = "Nelson George overview 2004" > George, p. 20</ref>

Janet Jackson -> Christina Aguilera
| Janet Jackson | -> | Christina Aguilera | In 2009, they collaborated with her good friend [[Lisa Marie Presley]] on the track "Thinkin' Bout You". (Nelson George overview 2004) > "Now What"

Result No: 4
Michael Jackson -> Smooth Criminal
| Michael Jackson | -> | Smooth Criminal | "Smooth Criminal" was an experience of innocence, rape and likely murder.

Smooth Criminal -> Pink (singer)
| after = "Get the Party Started!" > by [[Pink (singer)|Pink]]

Result No: 5
Michael Jackson -> Dallas Austin
| Michael Jackson | -> | Dallas Austin | "Dallas Austin" (born [[December 20]], [[1970]]) is an American [[songwriter]], [[producer]], [[US singer]], [[US record producer]], [[US musician]], [[US composer]], [[US songwriter]]. Some of his most notable clients include [[Michael Jackson]], [[TLC (band)|TLC]], [[Beyoncé]], [[Jay Z]], [[Pink (singer)|Pink]], [[Monica (entertainer)|Monica]], [[Madonna (entertainer)|Madonna]], [[Gwen Stefani]], [[Shirley Bassey]], [[Anastacia]], [[Japanese singer]], [[Namie Amuro]], [[Kiss from the Moon]], [[The Human League]], [[The Human League]], [[The Human League]].

Pink (singer) -> United States
| Pink (singer) | -> | United States | Born = [[Birth date and age|1979]] [[United States]], [[Pennsylvania]], [[Downtown]], [[Doylestown]], [[United States]].

Result No: 6
Michael Jackson -> Lisa Marie Presley
| Michael Jackson | -> | Lisa Marie Presley | in 2009, they collaborated with her good friend [[Lisa Marie Presley]] on the track "Thinkin' Bout You". (Nelson George overview 2004) > "Now What"

Pink (singer) -> Pop music
| Pink (singer) | -> | Pop music | "Rockin' Around the Christmas Tree" is a [[Christmas song]].

Result No: 7
Michael Jackson -> Billboard Hot 100
| Michael Jackson | -> | Billboard Hot 100 | "Billboard Hot 100" and "Billboard Hot 100" chart reflected the music industry's top 100 songs.

Pink (singer) -> Billboard Hot 100
| Pink (singer) | -> | Billboard Hot 100 | It became Pink's first single to miss the top forty on "Billboard Hot 100".
Evaluation

• Performed user evaluation
  • Results showed strong correlation of the user voting and used methods.
  • ISR method has been identified as the best method out of the three compared, while Random Weighting was almost in all the cases voted as the most obvious one.
Application - WikiPop

- Exploit Wikipedia page view statistics (number of visits per day) to detect trends in public interest
- Build service that detects events from the page views data
**Uptrends on English Wikipedia today**

1. **Ghost in the Shell: Stand Alone Complex +17,719%**
   
   Ghost in the Shell: Stand Alone Complex (攻殻機動隊 STAND ALONE COMPLEX, Kōkaku Kidōtai Sutando Arōn Konpurekkusu?, "Mobile Armoured Riot Police: Stand Alone Complex") is a 26 episode Japanese animated television series produced by Production I.G and based on ...

2. **Dave Niehaus +205,244%**
   
   Dave Niehaus (February 19, 1935 - November 10, 2010) was an American sportscaster. He was the lead play-by-play announcer for the American League's Seattle Mariners from their inaugural season in 1977 until his death in 2010. He was inducted into the Base...

   **Popular related pages:** Rick Rizzs +38,500%, The Double (Seattle Mariners) +38,000%, 1995 Seattle Mariners season +6,300%, List of Seattle Mariners broadcasters +14,900%, 1995 American League Division Series +2,489%

3. **Lorem ipsum +2,892%**
   
   In publishing and graphic design, lorem ipsum is placeholder text (filler text) commonly used to demonstrate the graphic elements of a document or visual presentation, such as font, typography, and layout. The lorem ipsum text is typically a section of a ...

4. **Python (film) +30,076%**
   
   While out at a local swimming hole the next day, John Cooper (Rosanoff), his girlfriend Kristen (Barron), his best friend Tommy (Wheaton), and Tommy's girlfriend Theresa (Mornell), find a pet Burmese python belonging to one of the dead girls, Lisa. Deputy...
Application - WikiPop

- Exploit Wikipedia page view statistics (number of visits per day) to detect trends in public interest
- Build service that detects events from the page views data
- Allow personalization:
  - By defining context (set of Wikipedia concepts)
  - Identify topics with increased popularity related to a given context
Valid inputs for the Wikipedia concepts fields are the titles of Wikipedia articles.

Few examples: Halloween, Michael Jackson, Nick Cave, Iggy Pop, Science Fiction

Base Distance: 2
Relatedness: Highest
Use incoming links: 0
Min Page Views: 1000

WikiPop

Grunge
Go

Spencer Elden
Grunge -> Nevermind
Nevermind -> Spencer Elden

search on web: >>

Foo Fighters
- Dave Grohl
- Pat Smear

Grunge -> Nirvana (band)
Nirvana (band) -> Foo Fighters

search on web: >>
Application - WikiPop

- Assign weight to edges in Wikipedia link graph – combination of two strategies:
  - Assign lower weight to links to 'hub' nodes (with large indegrees)
  - Assess relatedness of two concepts as a correlation of the page views during popularity peaks; e.g.
Term Association Networks

- Wikipedia link graph = concept network
- Can we build a concept network from a collection of documents?
Term Association Networks

- Wikipedia link graph = concept network
- Can we build a concept network from a collection of documents?

Diagram:

1. Document collection
2. Concept extraction
3. Concept association rules mining
4. Association Network construction
Term Association Networks

- Wikipedia link graph = concept network
- Can we build a concept network from a collection of documents?

Simple association rules: A ->B (p%)
Term Association Networks

Degree distribution:

Confidence int. measure:

Lift interestingness Measure:
Future work on concept networks mining

- Use of graph mining algorithms on concept networks:
  - Community detection
  - Link analysis (e.g. HITS, PageRank)
Future work on concept networks mining

HITS on term association networks

<table>
<thead>
<tr>
<th>VLDB09 data set:</th>
<th>WSDM10 data set:</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>RESULT</td>
</tr>
<tr>
<td>QUERY</td>
<td>MODEL</td>
</tr>
<tr>
<td>RESULT</td>
<td>DATA</td>
</tr>
<tr>
<td>ALGORITHM</td>
<td>NUMBER</td>
</tr>
<tr>
<td>NUMBER</td>
<td>QUERY</td>
</tr>
<tr>
<td>2V TIME</td>
<td>SEARCH</td>
</tr>
<tr>
<td>PROCESSING</td>
<td>INFORMATION</td>
</tr>
<tr>
<td>FIGURE</td>
<td>USER</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>USE</td>
</tr>
<tr>
<td>SECTION</td>
<td>ALGORITHM</td>
</tr>
<tr>
<td>DATABASE</td>
<td>WEB</td>
</tr>
<tr>
<td>EXAMPLE</td>
<td>WORK</td>
</tr>
<tr>
<td>VALUE</td>
<td>TIME</td>
</tr>
<tr>
<td>USE</td>
<td>APPROACH</td>
</tr>
<tr>
<td>CASE</td>
<td>SECTION</td>
</tr>
<tr>
<td>APPROACH</td>
<td>PAGE</td>
</tr>
<tr>
<td>ORDER</td>
<td>VALUE</td>
</tr>
<tr>
<td>MODEL</td>
<td>EXAMPLE</td>
</tr>
<tr>
<td>INFORMATION</td>
<td>DOCUMENT</td>
</tr>
</tbody>
</table>
Future work on concept networks mining

- Community detection in Wikipedia link graph
- Community containing concept 'Community Structure':
Future work on concept networks mining

- Community detection in concept network mined from personal paper collection
- Community containing concept 'Community Structure'
Thank you for your attention.