Knowledge Graphs: In Theory and Practice

Nitish Aggarwal, IBM Watson, USA, Sumit Bhatia, IBM Research, India
Saeedeh Shekarpour, Knoesis Research Centre Ohio, USA
Amit Sheth, Knoesis Research Centre Ohio, USA
The material presented in this tutorial represents the personal opinion of the presenters and not of IBM and affiliated organization.
Outline of the tutorial

Part 1: Knowledge Graph Construction
- Introduction
- DBpedia: Knowledge extraction
- Approaches to extend knowledge graph
- Knowledge extraction from scratch

Part 2: Knowledge Graph Analytics
- Finding entities of interest
- Entity exploration
- Upcoming challenges
What is Knowledge Graph

“The Knowledge Graph is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources.”
What is Knowledge Graph

“The Knowledge Graph is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources.”

“A Knowledge graph (i) mainly describes real world entities and interrelations, organized in a graph (ii) defines possible classes and relations of entities in a schema” (iii) allows potentially interrelating arbitrary entities with each other... [Paulheim H.]

“We defines a Knowledge Graph as an RDF graph consists of a set of RDF triples where each RDF triple (s,p,o) is an ordered set of following RDF term ....” [Pujara J. al al.]
What is Knowledge Graph

No single formal definition ...

• Defines real world entities

• Provides relationships between them
What is Knowledge Graph

No single formal definition ... 

• Defines real world entities

• Provides relationships between them

• Contains rules defines through ontologies

• Enable reasoning to infer new knowledge
Why Knowledge Graph

Building an intelligent system that can interact with human, requires knowledge about real world entities.
Building an intelligent system that can interact with human, requires knowledge about real world entities.

• Enhance search results.

• Enhance ad sense.

• Help in language understanding.

• Enables knowledge discovery.
Is there existing knowledge graph ready to use for my application?
DBpedia: Knowledge extraction

New York City

From Wikipedia, the free encyclopedia

"NYC" and "New York, New York" redirect here. For other uses, see New York City (disambiguation); NYC (disambiguation); and New York, New York (disambiguation).

The City of New York, often called New York City or simply New York, is the most populous city in the United States.[9] With an estimated 2016 population of 8,537,673[7] distributed over a land area of about 302.6 square miles (784 km²),[10][11] New York City is also the most densely populated major city in the United States.[12] Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, one of the most populous urban agglomerations in the world.[13][14] With an estimated 23.7 million residents as of 2016,[5] a global power city,[15] New York City has been described as the cultural, financial, and media capital[16][17] of the world.[18][19][20][21][22] and exerts a significant impact upon commerce, transportation, technology, education, politics, and sports. The city’s fast pace[23][24] defines the term New York minute.[25] Home to the headquarters of the United Nations,[26] New York is an important center for international diplomacy.[27]

Situated on one of the world’s largest natural harbors,[28][29] New York City consists of five boroughs, each of which is a separate county of New York State.[30] The five boroughs – Brooklyn, Queens, Manhattan, The Bronx, and Staten Island – were consolidated into a single city in 1898.[31] The city and its metropolitan area constitute the premier gateway for legal immigration to the United States,[32] and as many as 800 languages are spoken in New York.[33][34] making it the most linguistically diverse city in the world.[35][36][37] New York City is home to more than 3.2 million residents born outside the United States,[38] the largest foreign-born population of any city in the world.[39] In 2013, the tri-state New York Metropolitan Statistical Area (MSA) produced a gross metropolitan product (GMP) of nearly US$1.6 trillion.[40] If New York City were a country, it would have the 12th highest GDP in the world.[41]

New York City traces its origins to a trading post founded by colonists of the Dutch Republic in 1624 on Lower Manhattan; the post was named New Amsterdam in 1664.[42] The city and its surroundings came under English control in 1664[43] and were renamed New York after King Charles II of England granted the lands to his brother, the Duke of York.[44] New York served as the United States from 1785 until 1790.[44] It has been the country’s largest city since 1790.[45] The Statue of Liberty greeted millions of immigrants as they came to the Americas by ship in the late 19th and early 20th centuries[46] and is a world symbol of the United States and its ideals of liberty and peace.[47] In the 21st century, New York has emerged as a global node of creativity and entrepreneurship,[48] social tolerance,[49] and environmental sustainability[50][51] and as a symbol of freedom and cultural diversity.[52]

Many districts and landmarks in New York City are well known, and the city received a record 61 million tourists in 2016.[53] hosting three of the world’s ten most visited tourist attractions in 2013.[54] Several sources have ranked New York the most photographed city in the world.[55][56] Times Square, iconic as the world’s "heart,[57]" and its "Crossroads,[58]" is the brightly illuminated hub of the Broadway Theater District,[59] one of the world’s busiest pedestrian intersections,[60][61] and a major center of the world’s entertainment industry.[62] The names of many of the city’s bridges, skyscrapers,[63] and parks are known around the world. Anchored by Wall Street in the Financial District of Lower Manhattan, New York City has been called both the most economically powerful city and the leading financial center of the world.[59][64][65] and the city is home to the world’s two largest stock exchanges by total market capitalization, the New York Stock Exchange and NASDAQ. Manhattan’s real estate market is among the most expensive in the world.[66][67] Manhattan’s Chinatown incorporates the highest concentration of Chinese people in the Western Hemisphere.[71][72] with multiple signature Chinatowns developing across the city.[73][74] providing continuous 24/7 service[75] the New York City Subway is one of the most extensive metro systems worldwide, with 472 stations in operation.[76][77][78] Over 120 colleges and universities are located in New York City, including Columbia University, New York University, and Rockefeller University, which have been ranked among the top universities in the world.[79][80]
The City of New York, often called New York City or simply New York, is the most populous city in the United States.
The City of New York, often called New York City or simply New York, is the most populous city in the United States.

<New York City>, <CityIn> <United States>.

<City Name>, <locatedIn> <Country Name>.
The City of New York, often called New York City or simply New York, is the most populous city in the United States.

Berlin (/ˈbɛrliːn, bərl-/; German: [ˈbɛʁlin] (listen)) is the capital and the largest city of Germany as well as one of its 16 constituent states. With a population of approximately 3.7 million, Berlin is the second most populous city proper in the European Union and the seventh most populous urban area in the European Union. Located in northeastern Germany on the banks of the River Spree, it is the capital of the Berlin Brandenburg Metropolitan Region.

Mumbai (/mʌmˈbaɪ/; also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in India with an estimated city population of 18.4 million. Along with the neighbouring regions of the Mumbai Metropolitan Region, it is second most populous metropolitan area in India, with a population of 21.3 million as of 2016. Mumbai lies on the Konkan on the west coast of India and has a deep natural harbour.
DBpedia: Knowledge extraction

IBM

From Wikipedia, the free encyclopedia
(Redirected from ibm)

For other uses of IBM, see IBM (disambiguation). "Big Blue" redirects here. It is not to be confused with New York Giants. For other uses of Big Blue, see Big Blue (disambiguation).

IBM (International Business Machines Corporation) is an American multinational technology company headquartered in Armonk, New York, United States, with operations in over 170 countries. The company originated in 1911 as the Computing-Tabulating-Recording Company (CTR) and was renamed "International Business Machines" in 1924.

IBM manufactures and markets computers, middleware and software, and offers hosting and consulting services in areas ranging from mainframe computers to nanotechnology. IBM is also a major research organization, holding the record for most patents generated by a business (as of 2017) for 24 consecutive years. Inventions by IBM include the automated teller machine (ATM), the PC, the floppy disk, the hard disk drive, the magnetic stripe card, the relational database, the SQL programming language, the UPC barcode, and dynamic random-access memory (DRAM). The IBM mainframe, exemplified by the System/360, was the dominant computing platform during the 1960s and 1970s.

IBM has continually shifted its business mix by commoditizing markets focusing on higher-value, more profitable markets. This includes spinning off printer manufacturer Lexmark in 1991 and selling off its personal computer (ThinkPad/ThinkCentre) and x86-based server businesses to Lenovo (2005 and 2014, respectively), and acquiring companies such as PwC Consulting (2002), SPSS (2009), and The Weather Company (2016). Also in 2014, IBM announced that it would go "fabless", continuing to design semiconductors, but offloading manufacturing to GlobalFoundries.

Nicknamed Big Blue, IBM is one of 30 companies included in the Dow Jones Industrial Average and one of the world's largest employers, with (as of 2016) nearly 380,000 employees. Known as "IBMers", IBM employees have been awarded five Nobel Prizes, six Turing Awards, ten National Medals of Technology and five National Medals of Science.
DBpedia: Knowledge extraction

IBM

From Wikipedia, the free encyclopedia

Redirected from ibm

For other uses of IBM, see IBM (disambiguation). "Big Blue" redirects here. It is not to be confused with New York Giants. For other uses of Big Blue, see Big Blue (disambiguation).

IBM (International Business Machines Corporation) is an American multinational technology company headquartered in Armonk, New York, United States, with operations in over 170 countries. The company originated in 1911 as the Computing-Tabulating-Recording Company (CTR) and was renamed "International Business Machines" in 1924.

IBM manufactures and markets computers hardware, middleware and software, and offers hosting and consulting services in areas ranging from mainframe computers to nanotechnology. IBM is also a major research organization, holding the record for most patents generated by a business (as of 2017) for 24 consecutive years. Inventions by IBM include the automated teller machine (ATM), the PC, the floppy disk, the hard disk drive, the magnetic stripe card, the relational database, the SQL programming language, the UPC barcode, and dynamic random-access memory (DRAM). The IBM mainframe, exemplified by the System/360, was the dominant computing platform during the 1960s and 1970s.

IBM has continually shifted its business mix by commoditizing markets focusing on higher-value, more profitable markets. This includes spinning off printer manufacturer Lexmark in 1991 and selling off its personal computer (ThinkPad/ThinkCentre) and x86-based server businesses to Lenovo (2005 and 2014, respectively), and acquiring companies such as PwC Consulting (2002), SPSS (2009), and The Weather Company (2016). Also in 2014, IBM announced that it would go "software-defined", continuing to design semiconductors, but offloading manufacturing to GlobalFoundries.

Nicknamed Big Blue, IBM is one of 30 companies included in the Dow Jones Industrial Average and one of the world's largest employers, with (as of 2016) nearly 380,000 employees. Known as "IBMer", IBM employees have been awarded five Nobel Prizes, six Turing Awards, ten National Medals of Technology and five National Medals of Science.

<head entity>, <rel> < tail entity >
DBpedia: Knowledge extraction

IBM

From Wikipedia, the free encyclopedia
(Redirected from ibm)

For other uses of IBM, see IBM (disambiguation). "Big Blue" redirects here. It is not to be confused with New York Giants. For other uses of Big Blue, see Big Blue (disambiguation).

IBM (International Business Machines Corporation) is an American multinational technology company headquartered in Armonk, New York, United States, with operations in over 170 countries. The company originated in 1911 as the Computing-Tabulating-Recording Company (CTR) and was renamed "International Business Machines" in 1924.

IBM manufactures and markets computers, middleware, and software, and offers hosting and consulting services in areas ranging from mainframe computers to nanotechnology. IBM is also a major research organization, holding the record for most patents generated by a business (as of 2017) for 24 consecutive years. Inventions by IBM include the automated teller machine (ATM), the PC, the floppy disk, the hard disk drive, the magnetic stripe card, the relational database, the SQL programming language, the UPC barcode, and dynamic random-access memory (DRAM). The IBM mainframe, exemplified by the System/360, was the dominant computing platform during the 1960s and 1970s.

IBM has continually shifted its business mix by commoditizing markets focusing on higher-value, more profitable markets. This includes spinning off printer manufacturer Lexmark in 1991 and selling off its personal computer (ThinkPad/ThinkCentre) and x86-based server businesses to Lenovo (2005 and 2014, respectively), and acquiring companies such as PwC Consulting (2002), SPSS (2009), and The Weather Company (2016). Also in 2014, IBM announced that it would go "software-defined", continuing to design semiconductors, but offloading manufacturing to GlobalFoundries.

Nicknamed Big Blue, IBM is one of 30 companies included in the Dow Jones Industrial Average and one of the world's largest employers, with (as of 2016) nearly 380,000 employees. Known as "IBMers", IBM employees have been awarded five Nobel Prizes, six Turing Awards, ten National Medals of Technology and five National Medals of Science.

<head entity>, <rel> < tail entity >

Wikipedia Infobox

International Business Machines Corporation

IBM Watson system in 2011

Type
Public

Traded as
NYSE: IBM

Dividends
S&P 100 Component

S&P 500 Component

IBN
US4592001014

Industry
Cloud computing • Cognitive computing

Founded
June 16, 1911; 106 years ago
(as Computing-Tabulating-Recording Company)

Endicott, New York, U.S.[1]

Founder
Charles Ranlett Flint

Headquarters
Armonk, New York, U.S.

Area served
177 countries[2]

Key people
Ginni Rometty (Chairwoman and CEO)

Products
See IBM products

Revenue
$79.919 billion (2016)[3]

Operating income
$13.031 billion (2016)[3]

Net income
$11.912 billion (2016)[3]

Total assets
$71.47 billion (2016)[3]

Total equity
$18.392 billion (2016)[3]

Number of employees
380,000 (2018)[4]

Website
www.ibm.com

In the 1880s, technologies emerged that would ultimately form the core of International Business Machines (IBM). Julius E. Pitman patented the computing scale in 1885.[6] Alexander Dey invented the dial recorder (1889).[7] Herman Hollerith patented the Electric Tabulating Machine[8] and Willard Bundy invented a time clock to record a worker's arrival and departure time on a paper tape in 1889.[9] On June 16, 1911, their four companies were amalgamated in New York State by Charles Ranlett Flint forming a fifth company, the Computing-Tabulating-Recording Company (CTR) based in Endicott, New York.[10] The five companies had 1,300 employees and offices and plants in Endicott and Binghamton, New York; Dayton, Ohio; Detroit, Michigan; Washington, D.C.; and Toronto. They manufactured machinery for sale and lease, ranging from commercial scales and industrial time recorders, meat and cheese slicers, to tabulators and punched cards. Thomas J. Watson, Sr., fired from the National Cash Register Company by John Henry Patterson, called on Flint and, in 1914, was offered CTR.[11] Watson joined CTR as General Manager then, 11 months later, was made President when court cases relating to his time at NCR were resolved.[12] Having learned Patterson's pioneering business practices, Watson proceeded to put the stamp of NCR onto CTR's companies.[13] He implemented sales conventions, "generous sales incentives, a focus on customer service, an insistence on well-groomed, dark-suited salesmen and an evangelical fervor for instilling company pride and loyalty in every worker".[14][15] His favorite slogan, "THINK", became a mantra for each company's employees.[14] During Watson's first four years, revenues reached $9 million and the company's operations expanded to Europe, South America, Asia and Australia.[16] "Watson had never liked the clumsy hyphenated title of the CTR" and in 1924 chose to replace it with the more expansive
DBpedia: Knowledge extraction

Ontology Classes
- owl:Thing
  - Activity (edit)
    - Game (edit)
      - BoardGame (edit)
      - CardGame (edit)
    - Sales (edit)
    - Sport (edit)
      - Athletics (edit)
      - Boxing (edit)
        - BoxingCategory (edit)
        - BoxingStyle (edit)
      - HorseRiding (edit)
      - TeamSport (edit)
        - Soccer (edit)
    - Agent (edit)
      - Deity (edit)
      - Employer (edit)
      - Family (edit)
        - NobleFamily (edit)
      - FictionalCharacter (edit)
        - ComicaCharacter (edit)
        - AnimangaCharacter (edit)
      - DisneyCharacter (edit)
      - MythologicalFigure (edit)
      - NarutoCharacter (edit)
      - SoapCharacter (edit)
    - Organisation (edit)
      - Broadcaster (edit)
        - BroadcastNetwork (edit)
        - RadioStation (edit)
        - TelevisionStation (edit)
    - Company (edit)
      - Bank (edit)
      - Brewery (edit)
      - Caterer (edit)
      - LawFirm (edit)
      - PublicTransitSystem (edit)
        - Airline (edit)
        - BusCompany (edit)
      - Publisher (edit)
      - RecordLabel (edit)
      - Winery (edit)
    - EducationalInstitution (edit)
      - College (edit)
      - Library (edit)
      - School (edit)
      - University (edit)
    - EmployersOrganisation (edit)
    - GeopoliticalOrganisation (edit)
    - GovernmentAgency (edit)
      - GovernmentCabinet (edit)
    - Group (edit)
      - Band (edit)
      - ComedyGroup (edit)
  - Person (edit)
    - Archeologist (edit)
    - Architect (edit)
    - Aristocrat (edit)
    - Artist (edit)
      - Actor (edit)
        - AdultActor (edit)
        - VoiceActor (edit)
      - Comedian (edit)
      - ComicsCreator (edit)
      - Dancer (edit)
      - FashionDesigner (edit)
      - Humorist (edit)
      - MusicalArtist (edit)
        - BackScene (edit)
        - ClassicalMusicArtist (edit)
      - Instrumentalist (edit)
        - Guitarist (edit)
      - MusicDirector (edit)
      - Singer (edit)
      - Painter (edit)
      - Photographer (edit)
      - Sculptor (edit)
    - Astronaut (edit)
    - Athlete (edit)
      - ArcherPlayer (edit)
      - AthleticsPlayer (edit)
DBpedia: Knowledge extraction

Ontology Classes

- owl:Thing
  - Activity (edit)
    - Game (edit)
      - BoardGame (edit)
      - CardGame (edit)
    - Sales (edit)
    - Sport (edit)
      - Athletics (edit)
      - Boxing (edit)
        - BoxingCategory (edit)
        - BoxingStyle (edit)
      - HorseRiding (edit)
      - TeamSport (edit)
        - Soccer (edit)
  - Agent (edit)
    - Deity (edit)
    - Employer (edit)
    - Family (edit)
      - NobleFamily (edit)
    - FictionalCharacter (edit)
      - ComicCharacter (edit)
        - AnimangaCharacter (edit)
      - DisneyCharacter (edit)
      - MythologicalFigure (edit)
      - NarutoCharacter (edit)
    - SoapCharacter (edit)
  - Organisation (edit)
    - Broadcaster (edit)
      - BroadcastNetwork (edit)
      - RadioStation (edit)
      - TelevisionStation (edit)
  - Company (edit)
    - Bank (edit)
    - Brewery (edit)
    - Caterer (edit)
    - LawFirm (edit)
    - PublicTransitSystem (edit)
      - Airline (edit)
      - BusCompany (edit)
    - Publisher (edit)
    - RecordLabel (edit)
    - Winery (edit)
  - EducationalInstitution (edit)
    - College (edit)
    - Library (edit)
    - School (edit)
    - University (edit)
  - EmployersOrganisation (edit)
  - GeopoliticalOrganisation (edit)
  - GovernmentAgency (edit)
    - GovernmentCabinet (edit)
  - Group (edit)
    - Band (edit)
    - ComedyGroup (edit)

Person (edit)

- Archeologist (edit)
- Architect (edit)
- Aristocrat (edit)
- Artist (edit)
  - Actor (edit)
    - AdultActor (edit)
    - VoiceActor (edit)
  - Comedian (edit)
  - ComicsCreator (edit)
  - Dancer (edit)
  - FashionDesigner (edit)
  - Humorist (edit)
  - MusicalArtist (edit)
    - BackScene (edit)
    - ClassicalMusicArtist (edit)
    - Instrumentalist (edit)
      - Guitarist (edit)
      - MusicDirector (edit)
      - Singer (edit)
  - Painter (edit)
  - Photographer (edit)
  - Sculptor (edit)
- Astronaut (edit)
- Athlete (edit)
  - ArcherPlayer (edit)
  - AthleticsPlayer (edit)
# DBpedia: Knowledge extraction

## Properties on Actor:

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>academyAward (edit)</td>
<td>Academy Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>afiAward (edit)</td>
<td>AFI Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>arielAward (edit)</td>
<td>Ariel Award</td>
<td>Actor</td>
<td>Award</td>
</tr>
<tr>
<td>associatedAct (edit)</td>
<td>associated act</td>
<td>Artist</td>
<td>Artist</td>
</tr>
<tr>
<td>baftaAward (edit)</td>
<td>BAFTA Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>cesarAward (edit)</td>
<td>Cesar Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>disciple (edit)</td>
<td>disciple</td>
<td>Artist</td>
<td>Artist</td>
</tr>
<tr>
<td>dutchRKDCode (edit)</td>
<td>Dutch RKD code</td>
<td>Artist</td>
<td>xsd:string</td>
</tr>
<tr>
<td>emmyAward (edit)</td>
<td>Emmy Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>field (edit)</td>
<td>field</td>
<td>Artist</td>
<td>owl:Thing</td>
</tr>
<tr>
<td>filmFareAward (edit)</td>
<td>Film Fare Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>gaudiAward (edit)</td>
<td>Gaudí Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>geminiAward (edit)</td>
<td>Gemini Award</td>
<td>Actor</td>
<td>Award</td>
</tr>
<tr>
<td>goldenCalfAward (edit)</td>
<td>Golden Calf Award</td>
<td>Actor</td>
<td>Award</td>
</tr>
<tr>
<td>goldenGlobeAward (edit)</td>
<td>Golden Globe Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>goldenRaspberryAward (edit)</td>
<td>Golden Raspberry Award</td>
<td>Actor</td>
<td>Award</td>
</tr>
<tr>
<td>goyaAward (edit)</td>
<td>Goya Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>grammyAward (edit)</td>
<td>Grammy Award</td>
<td>Artist</td>
<td>Award</td>
</tr>
<tr>
<td>iftaAward (edit)</td>
<td>IFTA Award</td>
<td>Actor</td>
<td>Award</td>
</tr>
</tbody>
</table>
DBpedia: Knowledge extraction

Ontology (Classes, properties)

Parsers

dbr:IBM dbp:foundedBy dbr:Charles_Ranlett_Flint

dbr:IBM dbp:foundedBy dbr:Charles_Ranlett_Flint

dbr:IBM dbp:foundedBy dbr:Charles_Ranlett_Flint

...............
(Research) problems in knowledge graphs

- Incomplete knowledge
  - Missing entities
  - Missing relations
  - Limited entity and relation types
(Research) problems in knowledge graphs

• Incomplete knowledge
  – Missing entities
  – Missing relations
  – Limited entity and relation types

• Incorrect knowledge
  – Wrong entity label recognition
  – Wrong entity and relation type
  – Wrong facts
Research) problems in knowledge graphs

- Incomplete knowledge
  - Missing entities
  - Missing relations
  - Limited entity and relation types

- Incorrect knowledge
  - Wrong entity label recognition
  - Wrong entity and relation type
  - Wrong facts

- Inconsistency in knowledge
  - Different labels for same entity
  - Merging entities with same labels
Approaches to extend knowledge graphs

• Extracting knowledge from Wikipedia tables
  – Large amount of raw data in form of tables
  – Tables have some implicit structure/patterns
Approaches to extend knowledge graphs

- Extracting knowledge from Wikipedia tables
  - Large amount of raw data in form of tables
  - Tables have some implicit structure/patterns

Wiki:AFC_Ajax containing relations between players, their shirt number, and country
### Approaches to extend knowledge graphs

- <Wiki:AFC_Ajax, dbp:rel, Wiki:Andre_Onana>
- 80% entities in the table have relation dbp:rel with the Wikipedia title entity Wiki_AFC_Ajax
- Other 20% entities are likely to have the same relationship dbp:rel with Wiki_AFC_Ajax

[Munoz E. at al.] Using Linked Data to Mine RDF from Wikipedia's Tables, WSDM 2014

<table>
<thead>
<tr>
<th>No.</th>
<th>Position</th>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Player</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GK</td>
<td>André Onana</td>
<td>20</td>
<td>MF</td>
<td>Lasse Schöne <em>(vice-captain)</em></td>
</tr>
<tr>
<td>3</td>
<td>DF</td>
<td>Joël Veltman <em>(captain)</em></td>
<td>21</td>
<td>MF</td>
<td>Frenkie de Jong</td>
</tr>
<tr>
<td>4</td>
<td>DF</td>
<td>Matthijs de Ligt</td>
<td>22</td>
<td>GK</td>
<td>Benjamin van Leer</td>
</tr>
<tr>
<td>5</td>
<td>DF</td>
<td>Maximilian Wöber</td>
<td>23</td>
<td>MF</td>
<td>Siem de Jong</td>
</tr>
<tr>
<td>6</td>
<td>MF</td>
<td>Donny van de Beek</td>
<td>25</td>
<td>FW</td>
<td>Kasper Dolberg</td>
</tr>
<tr>
<td>7</td>
<td>FW</td>
<td>David Neres</td>
<td>26</td>
<td>DF</td>
<td>Nick Viergever</td>
</tr>
<tr>
<td>8</td>
<td>DF</td>
<td>Daley Sinkgraven</td>
<td>28</td>
<td>DF</td>
<td>Luis Manuel Orejuela</td>
</tr>
<tr>
<td>9</td>
<td>FW</td>
<td>Klaas-Jan Huntelaar</td>
<td>31</td>
<td>GK</td>
<td>Norbert Alblas</td>
</tr>
<tr>
<td>10</td>
<td>MF</td>
<td>Hakim Ziyech <em>(3rd captain)</em></td>
<td>33</td>
<td>GK</td>
<td>Kostas Lamprou</td>
</tr>
<tr>
<td>11</td>
<td>FW</td>
<td>Amin Younes</td>
<td>34</td>
<td>MF</td>
<td>Abdelhak Nouri [110]</td>
</tr>
<tr>
<td>15</td>
<td>MF</td>
<td>Carel Eiting</td>
<td>35</td>
<td>DF</td>
<td>Mitchell Dijks</td>
</tr>
<tr>
<td>17</td>
<td>FW</td>
<td>Václav Černý</td>
<td>42</td>
<td>DF</td>
<td>Deyovaisio Zeefuik</td>
</tr>
<tr>
<td>19</td>
<td>FW</td>
<td>Mateo Cassierra</td>
<td>45</td>
<td>FW</td>
<td>Justin Kluivert</td>
</tr>
</tbody>
</table>
Approaches to extend knowledge graphs

• Features
  – Article features: no. of tables, length
  – Table features: no. of rows, no. of columns
  – Column features: no. of entities in column, potential relations
  – Cell features: no. of entities in a cell, length of cell
  – Many others

• Combines using classification method

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>64.23</td>
<td>70.46</td>
<td>67.20</td>
</tr>
<tr>
<td>SVM</td>
<td>72.43</td>
<td>75.77</td>
<td>74.06</td>
</tr>
<tr>
<td>Logistic</td>
<td>79.62</td>
<td>79.01</td>
<td>79.31</td>
</tr>
</tbody>
</table>

[Munoz E. at al.] Using Linked Data to Mine RDF from Wikipedia's Tables, WSDM 2014
Approaches to extend knowledge graphs

• Features
  – Article features: no. of tables, length
  – Table features: no. of rows, no. of columns

• Rules/heuristics based methods makes mistakes, and hard to create one rule for everyone.

• Even though combining different features achieves 80% accuracy, it introduces 20% noise.

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>64.23</td>
<td>70.46</td>
<td>67.20</td>
</tr>
<tr>
<td>SVM</td>
<td>72.43</td>
<td>75.77</td>
<td>74.06</td>
</tr>
<tr>
<td>Logistic</td>
<td>79.62</td>
<td>79.01</td>
<td>79.31</td>
</tr>
</tbody>
</table>

[Munoz E. at al.] Using Linked Data to Mine RDF from Wikipedia's Tables, WSDM 2014
Table data is limited, we need to go beyond
Approaches to extend knowledge graphs

• Missing entity/literal for a relation
  – “Christopher A. Welty is an American computer scientist, who works at Google Research in NY”
    • <dbr:Chris_Welty> <employedBy> <?>
  
  – "Tom Cruise and Brad Pitt appear in Interview with the Vampire"
    • <dbr:Brad_Pitt> <?> <dbr:Tom_Cruise>
Approaches to extend knowledge graphs

• Missing entity/literal for a relation
  – “Christopher A. Welty is an American computer scientist, who works at Google Research in NY”
    • <dbr:Chris_Welty> <employedBy> <?>

  – "Tom Cruise and Brad Pitt appear in Interview with the Vampire"
    • <dbr:Brad_Pitt> <?> <dbr:Tom_Cruise>

• Knowledge Base Completion
  – Similar to link prediction in social network but a bit more challenging
  – Need to identify relation type in addition to binary output.
Approaches to extend knowledge graphs

• Knowledge Base Completion
  – TransE: learn the entity and relation embeddings by assuming that translation of entity embeddings correspond to their relation embeddings. [Bordes et al. 2013]
  – S + R ≈ T, where <S, R, T>
  – TransH: Learn different entity embedding for different relationships [Wang et al. 2014]
  – TransR: Learn entity and relation embeddings in different space, following by translation perform in relation space. [Lin Y. at el. 2015]
  – Many more methods [Nickel M. at al, 2015]
Knowledge base completion approaches focus on finding missing entities/relations
Need to add new entities from external sources
Need to add new entities from external sources

• Entity recognition in external text resource
  • Many Named Entity Recognition systems

• Link extracted entity to KG or create a new node if it does not have a corresponding entity
  • TAC-KBP (Entity Discovery and Linking task) [Ji H. at el. 2016]
Building knowledge graph such as DBpedia requires lot of manual efforts.
Building knowledge graph such as DBpedia requires lot of manual efforts.

- Many applications require domain/data specific custom knowledge graphs.

- Creating schema with class structure and constraints for each KG is difficult.
How to create a knowledge graph from unstructured text?
Jonathon Watson works at IBM. He has more than 50 patents, and won best inventor award for his invention “Neural Chip by Jon Watson et al.”
Jonathon Watson works at IBM. He has more than 50 patents, and won best inventor award for his invention “Neural Chip by Jon Watson et al.”
Relation extraction

- Supervised methods

Predefined schema (employedBy, bornOn, BirthPlace …)

Training data

Jonathon Watson works at IBM.

Jonathon Watson joined IBM.
Relation extraction

- Supervised methods

Predefined schema (employedBy, bornOn, BirthPlace …)

Training data

Jonathon Watson works at IBM.

Test data

Jonathon Watson is manager at IBM.

Jonathon Watson joined IBM.
Relation extraction

- Supervised methods

Predefined schema (employedBy, bornOn, BirthPlace …)

Training data

employedBy

Jonathon Watson works at IBM.

employedBy

Jonathon Watson joined IBM.

Test data

employedBy

Jonathon Watson is manager at IBM.
Relation extraction

- Supervised methods

  Pros: High accuracy and less noise

  Cons: Hard and expensive to build labeled data
Relation extraction

- Supervised methods
- Distantly supervised methods
Relation extraction

- Supervised methods
- Distantly supervised methods

Training sentences

- Jon Watson works at IBM.
- Jon Watson becomes VP at IBM.
- Michael Decker joins Data Science group at SMU.
- Michael Decker won a national funding award at SMU.
Relation extraction

• Supervised methods
• Distantly supervised methods

Pros: Overcome the effort of labeling data

Cons: Dependency of existing knowledge graph and corresponding text
Relation extraction

- Supervised methods
- Distantly supervised methods
- Unsupervised methods (OpenIE, Universal Schema)

Jonathon Watson works at IBM.

JOIN

Jonathon Watson joined IBM.

work

(ROOT (S (NP (Jon Watson))) (VP (VBZ works) (PP (IN at) (NP IBM))))

join

(ROOT (S (NP (Jon Watson))) (VP (VBD joined) (NP IBM)))
Relation extraction

- Supervised methods
- Distantly supervised methods
- Unsupervised methods (OpenIE, Universal Schema)

Pros: eliminates the effort of labeling data

Cons: Noisy, large number of relations
Relation extraction

- Supervised methods
- Distantly supervised methods
- Unsupervised methods (OpenIE, Universal Schema)

Relation 1: Works employer Company employedBy
Relation 2: livesIn currentCity Country
Relation 3: Vice President executive Board member
### Relation extraction (Universal Schema)

<table>
<thead>
<tr>
<th></th>
<th>employeBy</th>
<th>affiliated</th>
<th>Leader of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jon</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Michael</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Steve</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Joyce</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

- Clustering using vector similarity
- Matrix completion and fill the empty values [Yao L. at el., 2012]
### Entity types identification (Universal Schema)

<table>
<thead>
<tr>
<th></th>
<th>director</th>
<th>musician</th>
<th>actor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jon</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Michael</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steve</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Joyce</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

- Clustering using vector similarity
- Matrix completion and fill the empty values [Yao L. at el., 2012]
Relation extraction in domain

• Supervised methods – Need domain experts to label the data
• Distantly supervised methods – Hard to find corresponding text
• Unsupervised methods (OpenIE, Universal Schema) – Noisy

A 59-year-old African American man with a past medical history of hypertension, benign prostatic hypertrophy, type II diabetes mellitus for the past 15 years, and chronic back pain presents to the hospital with gross hematuria. The patient states that he noticed blood in his urine last night. The patient also reports mild, intermittent flank pain. The patient states that his diabetes and blood pressure are well controlled with medications, and that he has managed his chronic back pain with 2 aspirin per day for the past 4 years. Vital signs are Temp- 98.6°F, BP- 124/82 mm/Hg, pulse- 88/min, and RR- 14/min. Blood work is notable for HbA1C of 6.5%. A pyelogram reveals a ring sign. His current fasting glucose is 140mmol/L. What is the most likely etiology of hematuria in this patient?
Knowledge graphs in domain

- Domain specific entity extraction is more challenging
- Limited relation types
- Less explicit mention of entity and relation types in text
- Creating simple schema requires domain experts
Knowledge graph - Simple

Jonathon Watson works at IBM.

Michael Decker joined IBM.

Michael Decker attends SMU.
Knowledge graph – Simple + Schema

- Jonathon Watson works at IBM.
- Michael Decker joins IBM.
- Michael Decker attended SMU.

Diagram:
- Jonathon Watson is affiliated with IBM.
- Michael Decker is affiliated with SMU.
- SMU is affiliated with IBM.

Jonathon Watson
Michael Decker
SMU
IBM
Jonathon Watson works at IBM.

Michael Decker joined IBM.

Michael Decker attends SMU.
Summary

- Simple knowledge graph works for many applications
- Identify the requirement before finding the solution.
- Many knowledge graphs are publically available
https://www.youtube.com/watch?v=kao05Arliok&feature=youtu.be