Algorithms and Applications for Web-Scale Knowledge Graphs

Marco Ponza

Supervisor
Prof. Paolo Ferragina
Menu

1. **Entity Annotation**
   - The Modeling of Knowledge
   - Terminology
   - The Annotation Pipeline
   - Applications
   - A New Text Representation

2. **Work done in the first year**
   - Entity Relatedness
   - Document Aboutness

3. **Future Work**
1. Entity Annotation
The Modeling of Knowledge

- Classical approaches
  - Document Knowledge = Words
  - Bag-of-words (aka BoW)
  - Vector Space Model (aka VSM) (Salton, 1971)
The Modeling of Knowledge

- Well-known **issues** (Jurafsky, '00)
  - Ambiguity (Polysemy and Synonymy)

**Jaguar**

Jaguar (felin)  or  Jaguar_Cars
The Modeling of Knowledge

- **Well-known issues** (Jurafsky, '00)
  - Ambiguity (Polysemy and Synonymy)
  - Semantic Connections

Barack Obama United States
The Modeling of Knowledge

▷ Well-known issues (Jurafsky, ‘00)
  ○ Ambiguity (Polysemy and Synonymy)
  ○ Semantic Connections

▷ Algorithmic solutions
  ○ Latent Approaches (e.g. LSI/LSA, Word2Vec)
    ■ Unintelligible for humans (Gabrilovich IJCAI ‘07)
  ○ “Knowledge is Power” Hypothesis (Lenat, ‘91; Gabrilovich SIGIR ‘16)
    ■ Semantic and unambiguous concepts
    ■ Depend on the design of Entity Annotators
Entity Annotation

Terminology

▷ Wikipedia Knowledge Graph

▷ Node?
Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).

This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo dɪ ˈsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly Leonardo da Vinci or simply Leonardo (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of paleoanthropology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote". According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".

Born out of wedlock to a notary, Piero da Vinci, and a peasant woman, Caterina, in Vinci in the region of Florence, Leonardo was educated in the studio of the renowned Florentine painter Andrea del Verrocchio. Much of his earlier working life was spent in the service of Ludovico il Moro in Milan. He later worked in Rome, Bologna and Venice, and he spent his last years in France at the home awarded to him by Francis I of France.

Leonardo was, and is, renowned primarily as a painter. Among his works, the Mona Lisa is the most famous and most parodied portrait[8] and The Last Supper the most reproduced religious painting of all time, their fame approached only by Michelangelo's The Creation of Adam. [4]
Entity Annotation

Terminology

▷ Wikipedia Knowledge Graph

▷ Node: Wikipedia Page (Entity)

▷ Link?
Leonardo da Vinci

From Wikipedia, the free encyclopedia

“Da Vinci” redirects here. For other uses, see Da Vinci (disambiguation).

This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnarدو di ˈsɛr ˈpjɛːrə da (ˈv)iŋˈtʃi] (listen)), more commonly Leonardo da Vinci or simply Leonardo (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, botany, astronomy, and cartography. He has been variously called the father of paleontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo da Vinci as the epitome of the "Renaissance Man" or "Genius" and he has become the cultural symbol of the Renaissance.

Invention

From Wikipedia, the free encyclopedia

"Inventor" and "Invented" redirect here. For other uses, see Invention (disambiguation). For more details on inventions throughout history, see Timeline of historic inventions. For the CAD design software, see Autodesk Inventor.

An invention is a unique or novel device, method, composition or process. The invention provides an improvement in engineering and product development process. It may be an improvement upon a machine or an object or a result. An invention that achieves a completely unique function or results may not be obvious and not obvious to others skilled in the same field. An inventor may be taking a big step in creating something.

Some inventions can be patented. A patent legally protects the intellectual property rights of the inventor. The rules and requirements for patenting an invention process of obtaining a patent is often expensive.

Another meaning of invention is cultural invention, which is an innovative and useful concept that expanded on to others.[5] The Institute for Social Inventions collected many such ideas in many important components of artistic and design creativity. Inventions often extend the boundaries of capability.

Astronomy

From Wikipedia, the free encyclopedia

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets, nebulae) and processes (such as supernova explosions, gamma ray bursts, and micrometeorites, meteoroids, meteorites, meteorite, and comets). Astronomers observe and interpret phenomena, both geological and astronomical, and compare them to the organization of the universe. The study of the stars and planets is called observational astronomy, and the study of the nature of light and the behavior of photons is called astrophysics. The practice of astronomical observation is called astrometry, and the study of celestial objects is called astrophysics.

During the 20th century, the field of professional astronomy split into observational and theoretical. Observational astronomy is focused on acquiring data from observations of astronomical objects, which then analyzed. Theoretical astronomy is oriented toward the development of computer or analytical models to explain phenomena. The two fields complement each other, with theoretical astronomy seeking to explain observations behavior using confirmed theoretical results.

Astronomy is one of the few sciences where amateurs can still play an active role, especially if a transient phenomena. Amateur astronomers have made and contributed to many important new results.

Leonardo was, and is, renowned primarily as a painter. Among his works, the Mona Lisa is one of the most famous and most parodied portraits[6] and The Last Supper is the most reproduced painting of all time, their fame approached only by Michelangelo's The Creation of Adam. Leonardo's drawing of the Vitruvian Man is also regarded as a cultural icon, being a symbol of the Renaissance ideal of a man's body in harmony with the universe.
Entity Annotation

Terminology

▷ Wikipedia Knowledge Graph

▷ **Node**: Wikipedia Page (Entity)

▷ **Link**: Wikipedia Hyperlink

Enrich a text $\mathcal{T}$ with proper annotations

Annotation = ($\text{mention}$, $\text{entity}$)
Entity Annotation
The Annotation Pipeline

1. Identify mentions (spots)
2. Retrieve candidate entities

Spotting

Assign the most pertinent entity to each spot

Disambiguation

Remove not pertinent annotations

Pruning

Input Text

Entity Annotator

Annotated Text
Yesterday Maradona won against Mexico.
Entity Annotation
The Annotation Pipeline

Spotting

1. Mention Detection
   - Named Entity Recognition (aka NER)
   - N-gram generation

2. Candidate Generation
   - Gazetteer: \{ mention \rightarrow \text{entities} \}
     - How?
Florence: Leonardo's artistic and social background

Florence at the time of Leonardo's youth was the centre of Christian Humanist thought and culture. Leonardo commenced his apprenticeship with Verrocchio in 1466, the year that Verrocchio's master, the great sculptor Donatello, died. The painter Uccello, whose early experiments with perspective were to influence the development of landscape painting, was a very old man. The painters Piero della Francesca and Filippo Lippi, sculptor Luca della Robbia, and architect and writer Leon Battista Alberti were in their sixties. The successful artists of the next generation were Leonardo's teacher Verrocchio, Antonio del Pollaiuolo and the portrait sculptor Mino da Fiesole, whose lifelike busts give the most reliable likenesses of Lorenzo Medici's father Piero and uncle Giovanni.[53][54][55][56]

Leonardo's youth was spent in a Florence that was ornamented by the works of these artists and by Donatello's contemporaries, Masaccio, whose figurative frescoes were imbued with realism and emotion, and Ghiberti, whose Gates of Paradise gleaming with gold leaf, displayed the art of combining complex figure compositions with detailed architectural backgrounds. Piero della Francesca had made a detailed study of perspective,[57] and was the first painter to make a scientific study of light. These studies and Alberti's treatise De Pictura[58] were to have a profound effect on younger artists and in particular on Leonardo's own observations and artworks.[53][59][60]

Masaccio's "Expulsion from the Garden of Eden" depicting the naked and distraught Adam and Eve created a powerfully expressive image of the human form, cast into three dimensions by the use of light and shade, which was to be developed in the works of Leonardo in a way that was to be influential in the course of painting. The humanist influence of Donatello's "David" can be seen in Leonardo's late paintings, particularly John the Baptist.[53][54]

A prevalent tradition in Florence was the small altarpiece of the Virgin and Child. Many of these were created in tempera or glazed terracotta by the workshops of Filippo Lippi, Verrocchio and the prolific della Robbia family.[53] Leonardo's early Madonnas such as The Madonna with a carnation and the...
Entity Annotation
The Annotation Pipeline

Spotting

1. Mention Detection
   ○ Named Entity Recognition (aka NER)
   ○ N-gram generation

2. Candidate Generation
   ○ Gazetteer: \{ mention → entities \}
     ■ How? Wikipedia anchor texts!
     ■ Ranking (+ Thresholding)
       ● Commonness (Ferragina, CIKM '10; Guo, CIKM '14)
       ● Entity-context Similarity (Zwicklbauer, SIGIR '16)
       ● ...

Entity Annotation
The Annotation Pipeline

Spotting

1. Mention Detection
   ○ Named Entity Recognition (aka NER)
   ○ N-gram generation

2. Candidate Generation
   ○ Gazetteer: \{ mention → entities \}
     ■ How? Wikipedia anchor texts!
     ■ Ranking (+ Thresholding)
       ● Commonness (Ferragina, CIKM '10; Guo, CIKM '14)
       ● Entity-context Similarity (Zwicklbauer, SIGIR '16)
       ● ...

Yesterday Maradona won against Mexico.
Yesterday Maradona won against Mexico.

- Spots have been disambiguated
  - Ambiguous lexical elements (words) are now labeled with unambiguous concepts
- Finally, coherence scores are assigned
Entity Annotation

The Annotation Pipeline

Disambiguation

- LS NED (Cucerzan, ACL ‘07) (Mihalcea, CIKM ‘07)
- TAG-ME (Scaiella, CIKM ‘10)
- DBpediaSpotlight (Mendes, SemSys ‘11)
- fy (Moro, ACL ‘14) (Piccinno, SIGIR ‘14)
- WAT (Nguyen, LDOV ‘14)
- AIDA
- PBoH (Ganea, WWW ‘16)
- DoSeR (Zwicklbauer, SIGIR ‘16)
Entity Annotation
The Annotation Pipeline

Disambiguation

Algorithm: [TAG-ME, WAT] (Scaiella, CIKM ‘10; Piccinno, SIGIR ‘14)

[...] Maradona won against Mexico.

- Voting Scheme
- M&W / Jaccard Relatedness
Entity Annotation
The Annotation Pipeline

Disambiguation

Algorithm: **DoSeR** (Zwicklbauer, SIGIR ’16)

[...] **Maradona** won against **Mexico**.

- Graph of candidates
Entity Annotation
The Annotation Pipeline

Disambiguation

Algorithm: DoSeR (Zwicklbauer, SIGIR '16)

[... ] Maradona won against Mexico.

- Graph of candidates
- Entity2Vec Relatedness
Entity Annotation
The Annotation Pipeline

Disambiguation
Algorithm: DoSeR (Zwicklbauer, SIGIR '16)

... Maradona won against Mexico.

- Graph of candidates
- Entity2Vec Relatedness
- PageRank
Entity Annotation
The Annotation Pipeline

Pruning

Yesterday Maradona won against Mexico.

- Remove not pertinent annotations
- Clear text from erroneous annotations
- Coherence thresholding
Applications

Web Search Results (Gabrilovich, SIGIR '16)
Applications

Web Search Results  (Gabrilovich, SIGIR '16)
Applications

Question Answering  (Gabrilovich, SIGIR '16)
Applications
Implicit Questions  (Gabrilovich, SIGIR ’16)

Condition → What does it mean?  
Symptoms → What do they indicate?
A New Text Representation

- Originally introduced by (Scaiella, WSDM '12)
  - Widely deployed (Dunietz, EACL '14; Schuhmacher, WSDM '14; Ni, WSDM '15), ...

- \textit{Text} = Graph of Entities

- What about...
A New Text Representation

▷ Originally introduced by (Scaiella, WSDM ‘12)
  ○ Widely deployed (Dunietz, EACL ’14; Schuhmacher, WSDM ’14; Ni, WSDM ’15, ...)

▷ Text = Graph of Entities

▷ What about...
  ○ ...edge weights?
  ○ ...node weights?  
    } Work done in the first year
2. Work done in the first year
Entity Relatedness
Entity Relatedness

Compute how much two entities are related.

\[ \text{Relatedness} : \text{Entities} \times \text{Entities} \rightarrow \text{Real} \]

- How much related are...
  - ...Bank with Money?
  - ...Wood with Book?

- Semantic Reasoning:
  - Human: Background Knowledge
  - Machines: Knowledge Graph
Entity Relatedness

(A brief list of) Algorithms and Applications

▷ **Document/Word Similarity**
  ○ **WikiRelate** (Strube, AAI ’06)
  ○ **Explicit Semantic Analysis** (Gabrilovich, IJCAI ’07)
    - WikiWalk (Yeh, ACL ’09)
    - Temporal Semantic Analysis (Radinsky, WWW ’11)
    - Concept Graph Representation (Ni, WSDM ’16)
  ○ **Milne & Witten** (Witten, AAI ’08)
  ○ **Salient Semantic Analysis** (Hassan, AAI ’11)

▷ **Machine Translation** (Agirre, NAACL ’09; Rothe, ACL ’14)

▷ **Document Classification** (Perozzi, WWW ’14; Tang, WWW ’15)

▷ ...
Entity Relatedness

- Two entities are related whether...
  - ...they are described by related texts (Corpus-based)
    - Example: ESA (Gabrilovich, IJCAI '07)
      - Concepts grounded in human cognition
      - Opposite to latent concepts
Entity Relatedness

▷ Two entities are related whether...
  ○ ...they are described by related texts (Corpus-based)
    ■ Example: ESA (Gabrilovich, IJCAI '07)
      ● Concepts grounded in human cognition
      ● Opposite to latent concepts
  ○ ...they are referenced by related entities (Graph-based)
    ■ Example: CoSimRank (Rothe, ACL '14)
Entity Relatedness

CoSimRank (Rothe, ACL ‘14)

- **Graph**-based approach
- Relatedness algorithm for **nodes** in a graph
- Exploits Random Walks
- Algorithm (in brief)
  1. *Sets damping vectors for* $e_1$ *and* $e_2$  \[ e_1, e_2 \in \text{Entities} \]
  2. *Runs an iteration of PageRank*
  3. *Updates relatedness score*
Entity Relatedness
CoSimRank (Rothe, ACL '14)

\[ \text{Relatedness}^0(e_1, e_2) = 0.0 \]
Entity Relatedness

CoSimRank (Rothe, ACL ‘14)

\[
\text{Relatedness}^1(\mathbf{e}_1, \mathbf{e}_2) = 0.16
\]

\[
\begin{array}{c}
p^1(\mathbf{e}_1) \\
0.2 \\
0.4 \\
0.0 \\
0.4 \\
0.0 \\
0.0 \\
0.0 \\
\end{array}
\quad
\begin{array}{c}
p^1(\mathbf{e}_2) \\
0.0 \\
0.0 \\
0.2 \\
0.0 \\
0.4 \\
0.0 \\
0.0 \\
0.0 \\
\end{array}
\]
Entity Relatedness

CoSimRank (Rothe, ACL '14)

\[ \text{Relatedness}^2(e_1, e_2) = 0.33 \]
Entity Relatedness
CoSimRank (Rothe, ACL '14)

$\text{Relatedness}^3(e_1, e_2) = 0.47$
Entity Relatedness

CoSimRank (Rothe, ACL ‘14)

Relatedness$^0(e_1, e_3) = 0.0$
Entity Relatedness

CoSimRank (Rothe, ACL ‘14)

\[
\text{Relatedness}^3(e_1, e_3) = 0.13
\]
Two entities are related whether...
  ○ ...they are described by related texts (Corpus-based)
    ■ Example: ESA (Gabrilovich, IJCAI '07)
      ● Concepts grounded in human cognition
      ● Opposite to latent concepts

  ○ ...they are referenced by related entities (Graph-based)
    ■ Example: CoSimRank (Rothe, ACL '14)

Need of a fair and meaningful comparison
Entity Relatedness
Preliminary Results: The Relatedness Framework

▷ Design algorithms based on
  ○ Set Operations (Milne & Witten, Jaccard, ...)
  ○ Embeddings (Word2Vec, LDA, ...)
  ○ Random Walk (CoSimRank, PPR+Cos)

▷ Preliminary results
  ○ Analyse entity pairs
  ○ Deploy corpus-based algorithms
  ○ A new algorithm: LLP
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW ‘11)

  ○ Clustering algorithm (node labeling)
  ○ Pro: Scales on very large graphs
  ○ Cons: Can generate few big clusters

▷ Layered Label Propagation
▷ Standard Label Propagation with a resolution parameter $\gamma$
  ○ Graph compression
  ○ Algorithm (in brief)
    1. Randomly initialize each node with a label (cluster)
    2. Update label according to a specific rule
      ○ Maximize nonlocal discount (Ronhovde, Phys. Rev. ‘10)
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 1
Step: Initialization

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW ’11)

Round: 1
Step: Initialization

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 1
Step: Updating (1)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 1
Step: Updating (2)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 2
Step: Initialization

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Round: 2
Step: Initialization
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW ’11)

Round: 2
Step: Updating (1)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 2
Step: Updating (2)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation (Boldi, WWW ’11)

Round: 2
Step: Updating (2)
**Entity Relatedness**

A New Algorithm: Layered Label Propagation (Boldi, WWW ’11)

Round: 2  
Step: Updating (2)

<table>
<thead>
<tr>
<th>Node</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Entity Relatedness
A New Algorithm: Layered Label Propagation  (Boldi, WWW ‘11)

Round: 2
Step: Updating (2)

\[\text{Relatedness} = \text{Similarity between signatures}\]
Document Aboutness

Aboutness = Succinct representation of a document’s subject matter (Hutchins, 1977)

▷ Weight information (e.g. entities, words, ...) within a document

Goal
POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don’t look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party’s presidential nomination until after President Bush’s State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

Still, Mr. Obama has been busy telephoning crucial Democrats in Iowa, New Hampshire and other states. There is, of course, only one reason for him to be making such inquiries.

Last week on Capitol Hill, Mr. Obama bumped into Ethel Kennedy, who has been a big admirer. When asked about him, she said, “He can’t run soon enough.”

Mrs. Clinton, meanwhile, plans to announce her decision in the next several weeks, her advisers say. According to several Democrats who have spoken to her, as well as advisers, Mrs. Clinton has given every indication that she is running, short of saying so, and no signals that she is not.

She is making phone calls to Democratic officials, labor leaders and supporters in early nominating states. And she continues to talk to possible consultants and donors, yet she has not made any travel plans to kick off a campaign. JEFF ZELENY and PATRICK HEALY
Document Aboutness

Aboutness = Succinct representation of a document’s subject matter (Hutchins, 1977)

- **Goal**

- **Weight** information (e.g. entities, words, ...) within a document
- Wide range of practical applications:
  - a. Recommendation
  - b. Categorization
  - c. Exploratory search
  - d. Web Ranking
  - e. ...

Document Aboutness

Keyphrase Extraction

Aboutness
- Words
- Proper nouns
- Sentences

Candidate Extraction
- Dictionary
- POS tags

Subject Matter Identification
- Ranking/Classification

Issues
- Interpretation
- Overgeneration
- Infrequency
- Redundancy

Entity Salience

Entities

Entity Annotation

Dependency on KG?
“[...] errors could be addressed using background knowledge.”

(Hasen, ACL ‘14)

“[...] features more directly linked to Wikipedia [...] can provide more focused background information.”

(Dunietz, EACL ‘14)
Document Aboutness

Our Proposal

▷ **Entity Salience Approach**

Document Enrichment

▷ Pos tagging
▷ Mention detection
▷ Dependency parsing
▷ Co-reference resolver.
Document Aboutness
Our Proposal

Entity Salience Approach

- Entity annotation
- Graph of entities
- Relatedness
Document Aboutness
Our Proposal

▷ Entity Salience Approach

Document Enrichment

TextRank

Document Summarizer
Sentence Ranker
Document Aboutness
Our Proposal

▷ Entity Salience Approach

- Document → Feature Generation

  ▷ Entity → Feature Vector
  ▷ Classical syntactic features:
    - Frequency
    - Position
    - ...
  ▷ New syntactic and semantic features:
    - Position-based
    - Dependency-based
    - Relatedness-based
    - Centrality-based
    - ...

TextRank
Document Aboutness

Our Proposal

▷ Entity Salience Approach

Diagram:
- Document
- Document Enrichment
- Feature Generation
- Classification

Salient Entities vs. Non-salient entities

TextRank
Document Aboutness

Our Proposal: Main Contributions

▷ Fully documented system
▷ Public available via Web-API
▷ Improvement of state-of-the-art (Cmu-Google, F1: 61.5)
  ○ New York Times’ dataset (110,000 news, 1.3M entities)
  ○ 62.6 micro-F1 (+1.1%) and 59.5 macro-F1 (+2.5%)
  ○ More robust when entities are not biased at the beginning (+9%)
▷ Deep Feature and Error Analysis
3. Future Work
Future Work

▷ Conclude **Entity Relatedness**
  ○ Finalize experiments
  ○ Related vs Non-related
  ○ Scalability

▷ Improve our **Entity Salience System**
  ○ Deep Learning (i.e. w2v)
  ○ Abstractive Summarization
  ○ Create and test new datasets
  ○ Plug the new TagMe-Wat 2.0 *(Piccinno, 2016)*

▷ **Entity Annotation Improvement**
Thanks!

Any questions?