Information Retrieval and Web Search

PageRank for Summarization and Keyword Extraction

Instructor: Rada Mihalcea
PageRank

- Usually applied on directed graphs
  - From a given vertex, the walker selects at random one of the out-edges
- Given $G = (V,E)$ a directed graph with vertices $V$ and edges $E$
  - $\text{In}(V_i) =$ predecessors of $V_i$
  - $\text{Out}(V_i) =$ successors of $V_i$

$$S(V_i) = \frac{(1-d)}{N} + d \sum_{j \in \text{In}(V_i)} \frac{1}{|\text{Out}(V_j)|} S(V_j)$$

$d$ – damping factor $\in [0,1]$ (usually 0.85)
PageRank

- Applied also to undirected graphs
  - From a given vertex, the walker selects at random one of the incident edges
- Adapted to weighted graphs
  - From a given vertex, the walker selects at random one of the out (directed graphs) or incident (undirected graphs) edges, with higher probability of selecting edges with higher weight

\[
WS(V_i) = \frac{(1-d)}{N} + d \sum_{j \in \text{In}(V_i)} \frac{W_{ji}}{\sum_{V_k \in \text{Out}(V_j)} W_{jk}} WS(V_j)
\]
Convergence

- Convergence
  - Error below a small threshold value (e.g. 0.0001)
    \[
    | S(V_i)^k - S(V_i) | \approx | S(V_i)^{k+1} - S(V_i)^k |
    \]

- Text-based graphs – convergence usually achieved after 20-40 iterations
  - more iterations for undirected graphs
Convergence

(250 vertices, 250 edges)
PageRank for Text Processing

- Suitable for text processing tasks where a ranking over "cognitive units" is required
  - “cognitive unit” = text unit that conveys information
  - words, phrases, sentences, documents, etc.

- Steps:
  1. Model text as a graph
  2. Run graph-based ranking algorithm to convergence
  3. Use scores attached to vertices for application-specific decisions
Text Summarization

1. Build the graph:
   - Sentences in a text = vertices
   - Similarity between sentences = weighted edges

   Model the cohesion of text using intersentential similarity

2. Run random walk algorithm:
   - keep top N ranked sentences
   - sentences most “recommended” by other sentences
Underlying Idea: A Process of Recommendation

- A sentence that addresses certain concepts in a text gives the reader a recommendation to refer to other sentences in the text that address the same concepts.

- Text knitting (Hobbs 1974)
  - repetition in text “knits the discourse together”

- Text cohesion (Halliday & Hasan 1979)
Sentence Similarity

- Inter-sentential relationships
  - weighted edges
- Count number of common concepts
- Normalize with the length of the sentence

\[
Sim(S_1, S_2) = \frac{|\{w_k \mid w_k \in S_1 \land w_k \in S_2\}|}{\log(|S_1|) + \log(|S_2|)}
\]

- Other similarity metrics are also possible:
  - Longest common subsequence
  - String kernels, etc.
Graph Structure

- Undirected
  - No direction established between sentences in the text
  - A sentence can “recommend” sentences that precede or follow in the text
- Directed forward
  - A sentence “recommends” only sentences that follow in the text
  - Seems more appropriate for movie reviews, stories, etc.
- Directed backward
  - A sentence “recommends” only sentences that precede in the text
  - More appropriate for news articles
Hurricane Gilbert Heads Toward Dominican Coast

By RUDDY GONZALEZ

SANTO DOMINGO, Dominican Republic (AP) — Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night, the National Hurricane Center reported. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.

There were no reports of casualties.

Residents returned home, happy to find little damage from 80 mph winds and sheets of rain.

The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.

Cabral said residents of the province of Barahona should closely follow Gilbert's movement.

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.

Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.

The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's south coast.

There were no reports of casualties.

San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.

On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

Residents returned home, happy to find little damage from 80 mph winds and sheets of rain.

Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.

The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.
An Example
An Example

Automatic summary
Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The National Hurricane Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's coast.

Reference summary I
Hurricane Gilbert swept toward the Dominican Republic Sunday with sustained winds of 75 mph gusting to 92 mph. Civil Defense Director Eugenio Cabral alerted the country's heavily populated south coast and cautioned that even though there is no need for alarm, residents should closely follow Gilbert's movements. The U.S. Weather Service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Gilbert brought coastal flooding to Puerto Rico's south coast on Saturday. There have been no reports of casualties. Meanwhile, Hurricane Florence, the second hurricane of this storm season, was downgraded to a tropical storm.

Reference summary II
Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.
Evaluation

- Task-based evaluation: automatic text summarization
  - Single document summarization
    - 100-word summaries
  - Multiple document summarization
    - 100-word multi-doc summaries
    - clusters of ~10 documents

- Automatic evaluation with ROUGE (Lin & Hovy 2003)
  - n-gram based evaluations
    - unigrams found to have the highest correlations with human judgment
  - no stopwords, stemming
Evaluation

- Data from DUC (Document Understanding Conference)
  - DUC 2002
  - 567 single documents
  - 59 clusters of related documents
- Summarization of 100 articles in the TeMario data set
  - Brazilian Portuguese news articles
    - Jornal de Brasil, Folha de Sao Paulo
  - (Pardo and Rino 2003)
Results: Single Document Summarization

- Single-doc summaries for 567 documents (DUC 2002)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Undirected</th>
<th>Dir.forward</th>
<th>Dir.backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.4904</td>
<td>0.4202</td>
<td><strong>0.5008</strong></td>
</tr>
<tr>
<td>HITS&lt;sub&gt;A&lt;/sub&gt;</td>
<td>0.4912</td>
<td>0.4584</td>
<td><strong>0.5023</strong></td>
</tr>
<tr>
<td>HITS&lt;sub&gt;H&lt;/sub&gt;</td>
<td>0.4912</td>
<td>0.5023</td>
<td>0.4584</td>
</tr>
</tbody>
</table>

Top 5 systems (DUC 2002)

<table>
<thead>
<tr>
<th>S27</th>
<th>S31</th>
<th>S28</th>
<th>S21</th>
<th>S29</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5011</td>
<td>0.4914</td>
<td>0.4890</td>
<td>0.4869</td>
<td>0.4681</td>
<td>0.4799</td>
</tr>
</tbody>
</table>
Results: Single Document Summarization

- Summarization of Portuguese articles
- Test the language independent aspect
  - No resources required other than the text itself
- Summarization of 100 articles in the TeMario data set

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Undirected</th>
<th>Forward</th>
<th>Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>HITS (^w_A)</td>
<td>0.4814</td>
<td>0.4834</td>
<td>0.5002</td>
</tr>
<tr>
<td>HITS (^w_H)</td>
<td>0.4814</td>
<td>0.5002</td>
<td>0.4834</td>
</tr>
<tr>
<td>PR (^w_W)</td>
<td>0.4939</td>
<td>0.4574</td>
<td>0.5121</td>
</tr>
</tbody>
</table>

- Baseline: 0.4963
Results: Multiple Document Summarization

- Cascaded summarization ("meta" summarizer)
  - Use best single document summarization algorithms
    - PageRank (Undirected / Directed Backward)
    - HITS$\_A$ (Undirected / Directed Backward)
  - 100-word single document summaries
  - 100-word "summary of summaries"

- Avoid sentence redundancy:
  - set max threshold on sentence similarity (0.5)

- Evaluation:
  - build summaries for 59 clusters of ~10 documents
  - baseline: first sentence in each document
## Results: Multiple Document Summarization

Multi-doc summaries for 59 clusters (DUC 2002)

### “Meta” Algorithm – Graph

<table>
<thead>
<tr>
<th>Algorithm – Graph</th>
<th>PageRank-U</th>
<th>PageRank-DB</th>
<th>HITS\textsubscript{A}-U</th>
<th>HITS\textsubscript{A}-DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank-U</td>
<td>0.3552</td>
<td>0.3499</td>
<td>0.3456</td>
<td>0.3465</td>
</tr>
<tr>
<td>PageRank-DB</td>
<td>0.3502</td>
<td>0.3448</td>
<td>0.3519</td>
<td>0.3439</td>
</tr>
<tr>
<td>HITS\textsubscript{A}-U</td>
<td>0.3368</td>
<td>0.3259</td>
<td>0.3212</td>
<td>0.3423</td>
</tr>
<tr>
<td>HITS\textsubscript{A}-DB</td>
<td>0.3572</td>
<td>0.3520</td>
<td>0.3462</td>
<td>0.3473</td>
</tr>
</tbody>
</table>

### Top 5 systems (DUC 2002)

<table>
<thead>
<tr>
<th></th>
<th>S26</th>
<th>S19</th>
<th>S29</th>
<th>S25</th>
<th>S20</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>0.3578</td>
<td>0.3447</td>
<td>0.3264</td>
<td>0.3056</td>
<td>0.3047</td>
<td>0.2932</td>
</tr>
</tbody>
</table>
Keyword Extraction

- Identify important words in a text
- Keywords useful for
  - Automatic indexing
  - Terminology extraction
  - Back-of-the-book indexing
  - Within other applications: Information Retrieval, Text Summarization, Word Sense Disambiguation
- Previous work
  - mostly supervised learning
PageRank for Keyword Extraction

- Store words in vertices
- Use co-occurrence to draw edges
- Rank graph vertices across the entire text
- Pick top N as keywords
An Example

Compatibility of systems of linear constraints over the set of natural numbers
Criteria of compatibility of a system of linear Diophantine equations, strict
inequations, and nonstrict inequations are considered. Upper bounds for
components of a minimal set of solutions and algorithms of construction of
minimal generating sets of solutions for all types of systems are given.
These criteria and the corresponding algorithms for constructing a minimal
supporting set of solutions can be used in solving all the considered types of
systems and systems of mixed types.

Keywords by TextRank: linear constraints, linear diophantine equations,
natural numbers, non-strict inequations, strict inequations, upper bounds
Keywords by human annotators: linear constraints, linear diophantine
equations, non-strict inequations, set of natural numbers, strict inequations,
upper bounds
Evaluation

- **Data:**
  - 500 INSPEC abstracts, previously used in keyphrase extraction [Hulth 2003]

- **Settings:**
  - nouns and adjectives
  - select top N/3

- **Previous work**
  - mostly supervised learning
  - [Hulth 2003]
  - training/development/test : 1000/500/500 abstracts

<table>
<thead>
<tr>
<th>Method</th>
<th>Assigned Total</th>
<th>Assigned Mean</th>
<th>Correct Total</th>
<th>Correct Mean</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextRank</td>
<td>6,784</td>
<td>13.7</td>
<td>2,116</td>
<td>4.2</td>
<td>31.2</td>
<td>43.1</td>
<td>36.2</td>
</tr>
<tr>
<td>Ngram with tag</td>
<td>7,815</td>
<td>15.6</td>
<td>1,973</td>
<td>3.9</td>
<td>25.2</td>
<td>51.7</td>
<td>33.9</td>
</tr>
<tr>
<td>NP-chunks with tag</td>
<td>4,788</td>
<td>9.6</td>
<td>1,421</td>
<td>2.8</td>
<td>29.7</td>
<td>37.2</td>
<td>33</td>
</tr>
<tr>
<td>Pattern with tag</td>
<td>7,012</td>
<td>14.0</td>
<td>1,523</td>
<td>3.1</td>
<td>21.7</td>
<td>39.9</td>
<td>28.1</td>
</tr>
</tbody>
</table>