Web Data Mining: PageRank

Chen Hsiao Ting, 2302010238

PageRank, developed by Google co-founders Larry Page and Sergey Brin, is an algorithm for ranking web pages based on inbound link quality and quantity. Higher PageRank scores indicate greater authority. It plays a crucial role in web data mining, improving search results by highlighting high-PageRank pages for enhanced accuracy. Additionally, it's valuable for identifying influential pages in various networks, making it essential for network analysis in fields like social networks and web analytics.

In our assignment, we implemented PageRank by analyzing over one million Wikipedia pages, producing a sorted list of PageRank values for analysis. This report comprises an explanation of the PageRank implementation and a discussion of the results.

1. Indexing

In 'index.py,' an index for every Wikipedia page is created, associating each page's title with a consecutive index. This information is stored in the 'title_to_id_dict' as a dictionary in the format {"title": 1, …}.

Two more dictionaries, 'id_to_title_dict' and 'id_to_outlinks_id_dict,' are defined. The former links an index to a specific page title (e.g., {1: "title", …}), while the latter links an index to a list of indexes, representing all the links within a particular page (e.g., {1: [2, 3, 4], …}). To optimize memory usage, these dictionaries are stored in block-sized disk files.

1.1. Extracting Information

Wikipedia XML is a structured format for storing Wikipedia articles, consisting of a hierarchy with elements like titles, content, and metadata. Understanding Wikipedia XML requires familiarity with its schema, outlining the document structure, and its tags, like '<title>' for article titles and '<text>' for content.

1.2. Title

In Wikipedia XML, a Wikipedia page's title is typically found within the '<title>' element. This allows the use of a regular expression, 're.search(r"<title>([^:]*)</title>", line),' to extract the title from each article. However, title tags can sometimes contain text unrelated to the actual page title, like “Image:”, “Category:” and “File:”. Additionally, some pages may be redirecting, meaning that users accessing the page with the redirect title are automatically directed to the target page with the specified title. Due to these considerations, a title is considered valid only if it exists within the '<title>' tags, lacks “:”, and does not include the '<redirect title>' tag.

1.3. Outlinks

Wikipedia's internal link format features the target page's title enclosed in double square brackets (e.g., '[[Outlink Title]]'). A regular expression, ‘re.findall(r'\[\[[^:\[\]]*\]\]', line)’ is used to retrieve titles of all links within double square brackets on a page. However, some text may contain extra information separated by the symbol ‘|’ (e.g., ‘[[link_title | displayed_text]]’), where the text before ‘|’ indicates the actual target page title, and the text after ‘|’ serves

as a custom display text for the reader. To handle this variation, titles are extracted up to the ' | ' symbol.

2. PageRank

After indexing one million Wikipedia pages, PageRank scores are computed for all the pages in the 'pagerank.py' file. The process begins by loading two dictionaries generated during indexing: 'id_to_title_dict' and 'id_to_outlinks_id_dict'.

2.1. Damping Factor

The algorithm introduces a damping factor (denoted as 'd'), representing the probability that a user follows a link instead of teleporting to a random page. Typically, 'd' is set to 0.85, signifying an 85% probability of following a link and a 15% chance of teleporting.

2.2. Teleport Probability

The teleport probability, calculated as '(1 - d) / len(id_to_title_dict)', represents the likelihood of teleporting to a random page and is distributed evenly across all pages.

2.3. Convergence Threshold

A convergence threshold (epsilon), with a value of 1e-3, is employed to determine when the PageRank algorithm has reached convergence and can terminate.

2.4. PageRank Algorithm

Finally, using the PageRank formula:

```
# PR(A) = (1 - d) / N + d * Σ (PR(Ti) / C(Ti))
pagerank_curr[outlink_id] += d * pagerank[title_id] / outlinks_len
```

where ‘d’ is the damping factor, ‘(1 - d) / N’ is the teleport probability, ‘N’ is the total number of pages, ‘PR(Ti)’ is the PageRank score of a page ‘Ti’ that links to page ‘A’, and ‘C(Ti)’ is the number of outbound links from page ‘Ti’.

The algorithm continues to iterate until it converges, which occurs when the maximum change in PageRank values during an iteration falls below the threshold 'epsilon' (1e-3). This results in a PageRank score for each page. The final scores are then sorted in descending order and saved in an output file in the format 'title \t pr_score'.

3. Result & Discussion

The figure below represents the top 10 Wikipedia pages with the highest PageRank scores from a sample of one million Wikipedia pages.

```
United States Census Bureau   0.002094493413139943
United States     0.0020615532156762545
The New York Times      0.0018613449503613656
```

```
World War II      0.0013197780802870441
United Kingdom    0.000989174460827365
New York City     0.0009661763271006105
London      0.0008025351612722408
Population density      0.0006872968429856856
Germany     0.0006762980248656129
England     0.000673676393959783
```

The findings can be summarized as follows:

1. Page Importance: The “United States Census Bureau”, “United States” and “The New York Times” have the highest PageRank scores, indicating their critical importance and authority.
2. Informativeness: “World War II” and "Population density" rank 4th and 8th, underscoring their substantial relevance and frequent citation, highlighting the significance of these statistical concepts in Wikipedia.
3. Connectivity: Pages like “United Kingdom”, “Germany”, and “England” represent well-connected countries, emphasizing their vital role as points of connection in the Wikipedia network.

The diverse range of topics in the results showcases Wikipedia's vast knowledge spectrum. Authoritative, comprehensive and well-connected pages tend to score higher, serving as hubs for accessing a wide array of information.

4. Limitations

The above PageRank algorithm has several limitations. Firstly, it is computationally intensive, making it impractical for analyzing large datasets and less scalable for extensive web analysis. Furthermore, the algorithm heavily depends on link structures, potentially overlooking other crucial factors like content quality and user engagement, leading to less accurate rankings for less-linked pages. Moreover, PageRank can be manipulated through artificial link creation, and its fixed damping factor may yield inconsistent results. Finally, it doesn't adapt well to dynamic web environments.

To address these limitations, modern PageRank algorithms incorporate machine learning techniques and user behavior analysis to provide more accurate and adaptable rankings.

5. Conclusion

In summary, this report has highlighted the enduring significance of the PageRank algorithm in the realm of web data mining and network analysis. The top pages identified in the sample of one million Wikipedia pages exhibit the algorithm's ability to distinguish content quality and link structure. While the results provide valuable insights into web page importance, the report also highlights the limitations of PageRank, such as its sensitivity to link patterns. As such, future research should explore advanced techniques to address these limitations and enhance the algorithm's applicability in an ever-evolving online landscape.