

PageRank on Wikipedia: Towards General Importance Scores for Entities

Andreas Thalhammer and Achim Rettinger

AIFB, Karlsruhe Institute of Technology
{andreas.thalhammer, achim.rettinger}@kit.edu

Abstract. Link analysis methods are used to estimate importance in graph-structured data. In that realm, the PageRank algorithm has been used to analyze directed graphs, in particular the link structure of the Web. Recent developments in information retrieval focus on entities and their relations (i. e. knowledge graph panels). Many entities are documented in the popular knowledge base Wikipedia. The cross-references within Wikipedia exhibit a directed graph structure that is suitable for computing PageRank scores as importance indicators for entities.

In this work, we present different PageRank-based analyses on the link graph of Wikipedia and according experiments. We focus on the question whether some links - based on their position in the article text - can be deemed more important than others. In our variants, we change the probabilistic impact of links in accordance to their position on the page and measure the effects on the output of the PageRank algorithm. We compare the resulting rankings and those of existing systems with page-view-based rankings and provide statistics on the pairwise computed Spearman and Kendall rank correlations.

Keywords: Wikipedia, DBpedia, PageRank, link analysis, page views, rank correlation

1 Introduction

Entities are omnipresent in the landscape of modern information extraction and retrieval. Application areas range from natural language processing over recommender systems to question answering. For many of these application areas it is essential to build on objective importance scores of entities. One of the most successful amongst different methods is the PageRank algorithm [3]. It has been proven to provide objective relevance scores for hyperlinked documents, e. g. in Wikipedia [5,6,9]. Wikipedia serves as a rich source for entities and their descriptions. Its content is currently used by major Web search engine providers as a source for short textual summaries that are presented in knowledge graph panels. In addition, the link structure of Wikipedia has been shown to exhibit the potential to compute meaningful PageRank scores: connected with semantic background information (such as DBpedia [1]) the PageRank scores over the Wikipedia link graph enable rankings of entities of specific types, for example for

Listing 1.1. Example: SPARQL query on DBpedia for retrieving top-10 scientists ordered by PageRank (can be executed at <http://dbpedia.org/sparql>).

```
PREFIX v:<http://purl.org/voc/vrank#>

SELECT ?e ?r
FROM <http://dbpedia.org>
FROM <http://people.aifb.kit.edu/ath/#DBpedia_PageRank>
WHERE {
  ?e rdf:type dbo:Scientist;
  v:hasRank/v:rankValue ?r.
} ORDER BY DESC(?r) LIMIT 10
```

scientists (see Listing 1.1). Although the provided PageRank scores [9] exhibit reasonable output in many cases, they are not always easily explicable. For example, as of DBpedia version 2015-04, “Carl Linnaeus” (512) has a much higher PageRank score than “Charles Darwin” (206) and “Albert Einstein” (184) together in the result of the query in Listing 1.1. The reason is easily identified by examining the articles that link to the article of “Carl Linnaeus”:¹ Most articles use the template Taxobox² that defines the field `binomial_authority`. It becomes evident that the page of “Carl Linnaeus” is linked very often because Linnaeus classified species and gave them a binomial name (cf. [7]). In general, entities from the geographic and biological domains have distinctively higher PageRank scores than most entities from other domains. While, given the high inter-linkage of these domains, this is expected to some degree, articles such as “Bakhsh” (1913), “Provinces_of_Iran” (1810), “Lepidoptera”, (1778), or “Powiat” (1408) are occurring in the top-50 list of all things in Wikipedia, in accordance to DBpedia PageRank 2015-04 [9] (see Table 5). These points lead us to the question whether these rankings can be improved. Unfortunately, this is not a straight forward task as a gold standard is missing and rankings are often subjective.

In this work we investigate on different link extraction³ methods that address the root causes for the effects stated above. We focus on the question whether some links - based on their position in the article text - can be deemed more important than others. In our variants, we change the probabilistic impact of links in accordance to their position on the page and measure the effects on the output of the PageRank algorithm. We compare these variants and the rankings of existing systems with page-view-based rankings and provide statistics on the pairwise computed Spearman and Kendall rank correlations.

¹ Articles that link to “Carl Linnaeus” – https://en.wikipedia.org/wiki/Special:WhatLinksHere/Carl_Linnaeus

² Template:Taxobox – <https://en.wikipedia.org/wiki/Template:Taxobox>

³ With “link extraction” we refer to the process of parsing the wikitext of a Wikipedia article and to correctly identify and filter hyperlinks to other Wikipedia articles.

2 Background

In this section we provide additional background on the used PageRank variants, link extraction from Wikipedia, and redirects in Wikipedia.

2.1 PageRank Variants

The PageRank algorithm follows the idea of a user that browses Web sites by following links in a random fashion (random surfer). For computing PageRank, we use the original PageRank formula [3] and a weighted version [2] that accounts for the position of a link within an article.

- Original PageRank [3] – On the set of Wikipedia articles W , we use individual directed links $link(w_1, w_2)$ with $w_1, w_2 \in W$, in particular the set of pages that link to a page $l(w) = \{w_1 | link(w_1, w)\}$ and the count of outgoing links $c(w) = |\{w_1 | link(w, w_1)\}|$. The PageRank of a page $w_0 \in W$ is computed as follows:

$$pr(w_0) = (1 - d) + d * \sum_{w_n \in l(w_0)} \frac{pr(w_n)}{c(w_n)} \quad (1)$$

- Weighted Links Rank (WLRank) [2] – In order to account for the relative position of a link within an article, we adapt Formula (1) and introduce link weights. The idea is that the random surfer is likely not to follow every link on the page with the same probability but may prefer those that are at the top of a page. The WLRank of a page $w_0 \in W$ is computed as follows:

$$pr(w_0) = (1 - d) + d * \sum_{w_n \in l(w_0)} \frac{pr(w_n) * lw(link(w_n, w_0))}{\sum_{w_m} lw(link(w_n, w_m))} \quad (2)$$

The link weight function lw is defined as follows:

$$lw(link(w_1, w_2)) = 1 - \frac{first_occurrence(link(w_1, w_2), w_1)}{|tokens(w_1)|} \quad (3)$$

For tokenization we are splitting the article text in accordance to white spaces but do not split up links (e.g., `[[brown bear|bears]]` is treated as one token). The token numbering starts from 1, i.e. the first word/link of an article. The method `first_occurrence` returns the token number of the first occurrence of a link within an article.

Both formulas (1) and (2) are iteratively applied until the scores converge. The variable d marks the damping factor: in the random surfer model, it accounts for the possibility of accessing a page via the browser’s address bar instead of accessing it via a link from another page.

For reasons of presentation, we use the non-normalized version of PageRank in both cases. In contrast to the normalized version, the sum of all computed PageRank scores is the number of articles (instead of 1) and, as such, does not reflect a statistical probability distribution. However, normalization does not influence the final ranking and the resulting relations of the scores.

2.2 Wikipedia Link Extraction

In order to create a Wikipedia link graph we need to clarify which types of links are considered. The input for the rankings of [9] is a link graph that is constructed by the DBpedia Extraction Framework⁴ (DEF). The DBpedia extraction is based on Wikipedia database backup dumps⁵ that contain the non-rendered wikitexts of the Wikipedia articles and templates. From these sources, DEF builds a link graph by extracting links of the form `[[article|anchor text]]`. We distinguish two types of links with respect to templates:⁶

1. Links that are defined in the Wikipedia text but do not occur within a template, for example `[[brown bear|bears]]` outside `{{ and }}`.
2. Links that are provided as (a part of) a parameter to the template, for example `[[brown bear|bears]]` inside `{{ and }}`.

DEF considers only these two types of links and not any additional ones that result from the rendering of an article. It also has to be noted that DEF does not consider links from category pages. This mostly affects links to parent categories as the other links that are presented on a rendered category page (i. e. all articles of that category) do not occur in the wikitext. As an effect, the accumulated PageRank of a category page would be transferred almost 1:1 to its parent category. This would lead to a top-100 ranking of things with mostly category pages only. In addition, DEF does not consider links in references (denoted via `<ref>` tags).

In this work, we describe how we performed more general link extraction from Wikipedia. Unfortunately, in this respect, DEF exhibited certain inflexibilities as it processes Wikipedia articles line by line. This made it difficult to regard links in the context of an article as a whole (e. g., in order to determine the relative position of a link). In consequence, we reverse-engineered the link extraction parts of DEF and created the SiteLinkExtractor⁷ tool. The tool enables to execute multiple extraction methods in a single pass over all articles and can also be extended by additional extraction approaches.

2.3 Redirected vs. Unredirected Wikipedia Links

DBpedia offers two types of page link datasets:⁸ one in which the redirects are resolved and one in which they are contained. In principle, also redirect chains of more than one hop are possible but, in Wikipedia, the MediaWiki software is configured not to follow such redirect chains (that are called “double redirect”

⁴ DBpedia Extraction Framework – <https://github.com/dbpedia/extraction-framework/wiki>

⁵ Wikipedia dumps – <http://dumps.wikimedia.org/>

⁶ Template inclusions are marked by double curly brackets, i. e. `{{ and }}`.

⁷ SiteLinkExtractor – <https://github.com/TBritsch/SiteLinkExtractor>

⁸ DBpedia PageLinks – <http://wiki.dbpedia.org/Downloads2015-04>

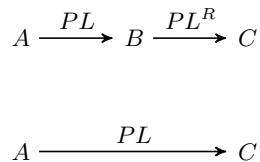


Fig. 1. Transitive resolution of a redirect in Wikipedia. A and C are full articles and B is called a “redirect page”, PL are page links, and PL^R are page links marked as a redirect (e.g. #REDIRECT [[United Kingdom]]). The two page links from A to B and from B to C are replaced by a direct link from A to C .

in Wikipedia)⁹ automatically and various bots are in place to remove them. As such, we can assume that only single-hop redirects are in place. However, as performed by DBpedia, also single-hop redirects can be resolved (see Figure 1). Alternatively, for various applications (especially in NLP) it can make sense to keep redirect pages as redirect pages also have a high number of inlinks in various cases (e.g. “Countries of the world”)¹⁰. However, with reference to Figure 1 and assuming that redirect pages only link to the redirect target, B passes most of its own PageRank score on to C (note that the damping factor is in place).

3 Link Graphs

We implemented five Wikipedia link extraction methods that enable to create different input graphs for the PageRank algorithm. In general we follow the example of DEF and consider type 1 and 2 links for extraction (which form a subset of those that occur in a rendered version of an article). The following extraction methods were implemented:

All Links (ALL) This extractor produces all type 1 and 2 links. This is the reverse-engineered DEF method. It serves as a reference.

Article Text Links (ATL) This measure omits links that occur in text that is provided to Wikipedia templates (i.e. includes type 1 links, omits type 2 links). The relation to ALL is as follows: $ATL \subseteq ALL$.

Article Text Links with Relative Position (ATL-RP) This measure extracts all links from the Wikipedia text (type 1 links) and produces a score for the relative position of each link (see Formula 3). In effect, the link graph ATL-RP is the same as ATL but uses edge weights based on each link’s position.

Abstract Links (ABL) This measure extracts only the links from Wikipedia abstracts. We chose the definition of DBpedia which defines an abstract as

⁹ Wikipedia: Double redirects – https://en.wikipedia.org/wiki/Wikipedia:Double_redirects

¹⁰ Inlinks of “Countries of the world” – https://en.wikipedia.org/wiki/Special:WhatLinksHere/Countries_of_the_world

Table 1. Number of links per link graph. Duplicate links were removed in all graphs (except in ATL-RP where multiple occurrences have different positions).

ALL	ATL	ATL-RP	ABL	TEL
159 398 815	142 305 605	143 056 545	32 887 815	26 460 273

the first complete sentences that accumulate to less than 500 characters.¹¹

This link set is a subset of all type 1 links (in particular: $ABL \subseteq ATL$).

Template Links (TEL) This measure is complementary to ATL and extracts only links from templates (i. e. omits type 1 links, includes type 2 links). The relation to ALL and ATL is as follows: $TEL = ALL \setminus ATL$.

Redirects are not resolved in any of the above methods. We execute the introduced extraction mechanisms on dumps of the English (2015-02-05) and German (2015-02-11) Wikipedia. The respective dates are aligned with the input of DEF with respect to DBpedia version 2015-04.¹² Table 1 provides an overview of the number of extracted links per link graph.

4 Experiments

In our experiments, we first computed PageRank on the introduced link graphs. We then measured the pairwise rank correlations (Spearman’s ρ and Kendall’s τ)¹³ between these rankings and the reference datasets (of which three are also based on PageRank and two are based on page-view data of Wikipedia). With the resulting correlation scores, we investigated on the following hypotheses:

- H1** Links in templates are created in a “please fill out” manner and rather negatively influence on the general salience that PageRank scores should represent.
- H2** Links that are mentioned at the beginning of articles are more often clicked and correlate with the number of page views that the target page receives.
- H3** The practice of resolving redirects does not strongly impact on the final ranking in accordance to PageRank scores.

4.1 PageRank Configuration

We computed PageRank with the following parameters on the introduced link graphs ALL, ATL, ATL-RP, ABL, and TEL: non-normalized, 40 iterations, damping factor 0.85, start value 0.1.

¹¹ DBpedia abstract extraction – <http://git.io/vGZ4J>

¹² DBpedia 2015-04 dump dates – <http://wiki.dbpedia.org/services-resources/datasets/dataset-2015-04/dump-dates-dbpedia-2015-04>

¹³ Both measures have a value range from -1 to 1 and are specifically designed for measuring rank correlation.

4.2 Reference Datasets

We use the following rankings as reference datasets:

DBpedia PageRank (DBP) The scores of DBpedia PageRank [9] are based on the “DBpedia PageLinks” dataset (i. e. Wikipedia PageLinks as extracted by DEF, redirected). The computation was performed with the same configuration as described in Section 4.1. The scores are regularly published as TSV and Turtle files. The Turtle version uses the vRank vocabulary [8]. Since DBpedia version 2015-04, the DBP scores are included in the official DBpedia SPARQL endpoint (cf. Listing 1.1 for an example query). In this work, we use the following versions of DBP scores based on English Wikipedia: 2014, 2015-04.

DBpedia PageRank Unredirected (DBP-U) This dataset is computed in the same way as DBP but uses the “DBpedia PageLinks Unredirected” dataset.¹⁴ As the name suggests, Wikipedia redirects are not resolved in this dataset (see Section 2.3 for more background on redirects in Wikipedia). We use the 2015-04 version of DBP-U.

SubjectiveEye3D (SUB) Paul Houle aggregated the Wikipedia page views of the years 2008 to 2013 with different normalization factors (particularly considering the dimensions articles, language, and time)¹⁵. As such, SubjectiveEye3D reflects the aggregated chance for a page view of a specific article in the interval years 2008 to 2013. However, similar to unnormalized PageRank, the scores need to be interpreted in relation to each other (i. e. the scores do not reflect a proper probability distribution as they do not add up to one).

The Open Wikipedia Ranking - Page Views (TOWR-PV) “The Open Wikipedia Ranking”¹⁶ provides scores on page views. The data is described as “the number of page views in the last year” on the project’s Web site.

The two page-views-based rankings serve as a reference in order to evaluate the different PageRank rankings. We show the amount of entities covered by the PageRank datasets and the entity overlap with the page-view-based rankings in Table 2.

4.3 Results

We used MATLAB for computing the pairwise Spearman’s ρ and Kendall’s τ correlation scores. The Kendall’s τ rank correlation measure has $\mathcal{O}(n^2)$ complexity and takes a significant amount of time for large matrices. In order to speed this up, we sampled the data matrix by a random selection of 1M rows for

¹⁴ DBpedia PageLinks Unredirected – http://downloads.dbpedia.org/2015-04/core-118n/en/page-links-unredirected_en.nt.bz2

¹⁵ SubjectiveEye3D – <https://github.com/paulhoule/telepath/wiki/SubjectiveEye3D>

¹⁶ The Open Wikipedia Ranking – <http://wikirank.di.unimi.it/>

Table 2. Amount of overlapping entities in the final rankings between the PageRank-based measures and the page-view-based ones.

	#entities	\cap SUB (6 211 717 entities)	\cap TOWR-PV (4 853 050 entities)
DBP 2014	19 540 318	5 267 822	4 587 525
DBP 2015-04	20 473 313	5 235 342	4 781 198
DBP-U 2015-04	20 473 371	5 235 319	4 781 198
ALL	18 493 968	4 936 936	4 780 591
ATL	17 846 024	4 936 086	4 779 032
ATL-RP	17 846 024	4 936 086	4 779 032
ABL	12 319 754	4 425 821	4 739 104
TEL	5 028 217	2 913 542	3 320 433

Table 3. Correlation: Spearman’s ρ (the colors are used for better readability and do not comprise additional meaning).

	DBP 2014	DBP 2015-04	DBP-U 2015-04	ALL	ATL	ATL-RP	ABL	TEL	TOWR-PV	SUB
DBP 2014	1.00	0.94	0.72	0.71	0.71	0.66	0.70	0.28	0.64	0.50
DBP 2015-04	0.94	1.00	0.77	0.76	0.76	0.71	0.77	0.16	0.65	0.55
DBP-U 2015-04	0.72	0.77	1.00	1.00	0.99	0.95	0.79	0.34	0.66	0.58
ALL	0.71	0.76	1.00	1.00	0.99	0.95	0.79	0.35	0.66	0.57
ATL	0.71	0.76	0.99	0.99	1.00	0.96	0.80	0.29	0.66	0.55
ATL-RP	0.66	0.71	0.95	0.95	0.96	1.00	0.79	0.31	0.65	0.64
ABL	0.70	0.77	0.79	0.79	0.80	0.79	1.00	0.26	0.50	0.45
TEL	0.28	0.16	0.34	0.35	0.29	0.31	0.26	1.00	0.42	0.41
TOWR-PV	0.64	0.65	0.66	0.66	0.66	0.65	0.50	0.42	1.00	0.86
SUB	0.50	0.55	0.58	0.57	0.55	0.64	0.45	0.41	0.86	1.00

Kendall’s τ . The pairwise correlation scores of ρ and τ are reported in Tables 3 and 4 respectively. The results are generally as expected: For example, the page-view-based rankings correlate strongest with each other. Also DBP-U 2015-04 and ALL have a very strong correlation (these rankings should be equal).

H1 seems to be supported by the data as the TEL PageRank scores correlate worst with any other ranking. However, ATL does not correlate better with SUB and TOWR-PV than ALL. This indicates that the reason for the bad correlation might not be due to the “bad semantics of links in the infobox”. With random samples on ATL - which produced similar results - we found that the computed PageRank values of TEL are mostly affected by the low total link count (see Table 1). With respect to the initial example, the PageRank score of “Carl Linnaeus” is reduced to 217 in ATL. However, a general better performance of ATL is not noticeable with respect to the comparison to SUB and TOWR-PV. We assume that PageRank on DBpedia’s RDF data results in similar scores as TEL as DBpedia [1] extracts its semantic relations mostly from Wikipedia’s infoboxes.

Indicators for **H2** are the scores of ABL and ATL-RP. However, similar to TEL, ABL does not produce enough links for a strong ranking. ATL-RP, in

Table 4. Correlation: Kendall’s τ on a sample of 1 000 000 (the colors are used for better readability and do not comprise additional meaning).

	DBP 2014	DBP 2015-04	DBP-U 2015-04	ALL	ATL	ATL-RP	ABL	TEL	TOWR-PV	SUB
DBP 2014	1.00	0.86	0.65	0.64	0.64	0.57	0.60	0.20	0.47	0.35
DBP 2015-04	0.86	1.00	0.76	0.74	0.73	0.63	0.69	0.11	0.48	0.39
DBP-U 2015-04	0.65	0.76	1.00	0.99	0.95	0.84	0.68	0.25	0.48	0.41
ALL	0.64	0.74	0.99	1.00	0.95	0.84	0.68	0.25	0.48	0.40
ATL	0.64	0.73	0.95	0.95	1.00	0.86	0.69	0.20	0.48	0.39
ATL-RP	0.57	0.63	0.84	0.84	0.86	1.00	0.69	0.22	0.47	0.46
ABL	0.60	0.69	0.68	0.68	0.69	0.69	1.00	0.19	0.37	0.33
TEL	0.20	0.11	0.25	0.25	0.20	0.22	0.19	1.00	0.30	0.29
TOWR-PV	0.47	0.48	0.48	0.48	0.48	0.47	0.37	0.30	1.00	0.70
SUB	0.35	0.39	0.41	0.40	0.39	0.46	0.33	0.29	0.70	1.00

contrast, produces the strongest correlation with SUB. This is an indication that - indeed - articles that are linked at the beginning of a page are more often clicked. This is supported by related findings where actual HTTP referrer data was analyzed [4].

With respect to **H3**, we expected DBP-U 2015-04 and DBP 2015-04 to correlate much stronger but DEF does not implement the full workflow of Figure 1: although it introduces a link $A \rightarrow C$ and removes the link $A \rightarrow B$, it does not remove the link $B \rightarrow C$. As such, the article B occurs in the final entity set with the lowest PageRank score of 0.15 (as it has no incoming links). In contrast, these pages often accumulate PageRank scores of 1000 and above in the unredirected datasets. If B would not occur in the final ranking of DBP 2015-04, it would not be considered by the rank correlation measures. This explains the comparatively weak correlation between the redirected and unredirected datasets.

4.4 Conclusions

Whether links from templates are excluded or included in the input link graph does not impact strongly on the quality of rankings produced by PageRank. WLRank on articles produces best results with respect to the correlation to page-view-based rankings. In general, although there is a correlation, we assume that link and page-view-based rankings are complementary. This is supported by Table 5 which contains the top-50 scores of SUB, DBP 2015-04, and ATL-RP: The PageRank-based measures are strongly influenced by articles that relate to locations (e. g., countries, languages, etc.) as they are highly interlinked and referenced by a very high fraction of Wikipedia articles. In contrast, the page-view-based ranking of SubjectiveEye3D covers topics that are frequently accessed and mostly relate to pop culture and important historical figures or events. We assume that a strong and more objective ranking of entities is probably achieved by combining link-structure and page-view-based rankings on Wikipedia. In general, and especially for applications that deal with NLP, we recommend to use the unredirected version of DBpedia PageRank.

Table 5. The top-50 rankings of SubjectiveEye3D (< 0.3, above are: Wiki, HTTP 404, Main Page, How, SDSS), DBP 2015-04, and ATL-RP.

	SUB	DBP 2015-04	ATL-RP
1	YouTube	Category:Living people	United States
2	Searching	United States	World War II
3	Facebook	List of sovereign states	France
4	United States	Animal	United Kingdom
5	Undefined	France	Race and ethnicity in the United States Census
6	Lists of deaths by year	United Kingdom	Germany
7	Wikipedia	World War II	Canada
8	The Beatles	Germany	Association football
9	Barack Obama	Canada	Iran
10	Web search engine	India	India
11	Google	Iran	England
12	Michael Jackson	Association football	Latin
13	Sex	England	Australia
14	Lady Gaga	Australia	Russia
15	World War II	Arthropod	China
16	United Kingdom	Insect	Italy
17	Eminem	Russia	Japan
18	Lil Wayne	Japan	Village
19	Adolf Hitler	China	Moth
20	India	Italy	World War I
21	Justin Bieber	English language	Romanize
22	How I Met Your Mother	Poland	Spain
23	The Big Bang Theory	London	Romanization
24	World War I	Spain	Europe
25	Miley Cyrus	New York City	Romania
26	Glee (TV series)	Catholic Church	Soviet Union
27	Favicon	World War I	London
28	Canada	Bakhsh	English language
29	Sex position	Latin	Poland
30	Kim Kardashian	Village	New York City
31	Australia	Counties of Iran	Catholic Church
32	Rihanna	Provinces of Iran	Brazil
33	Steve Jobs	Lepidoptera	Netherlands
34	Selena Gomez	California	Greek language
35	Internet Movie Database	Brazil	Category:Unprintworthy redirects
36	Sexual intercourse	Romania	Scotland
37	Harry Potter	Europe	Sweden
38	Japan	Soviet Union	California
39	New York City	Chordate	Species
40	Human penis size	Netherlands	French language
41	Germany	New York	Mexico
42	Masturbation	Administrative divisions of Iran	Genus
43	September 11 attacks	Iran Standard Time	United States Census Bureau
44	Game of Thrones	Mexico	Turkey
45	Tupac Shakur	Voivodeship (Poland)	New Zealand
46	1	Sweden	Census
47	Naruto	Powiat	Middle Ages
48	Vagina	Gmina	Paris
49	Pornography	Moth	Communes of France
50	House (TV series)	Departments of France	Switzerland

5 Related Work

This work is influenced and motivated by an initial experiment that was performed by Paul Houle: In the Github project documentation of SubjectiveEye3D he reports about Spearman and Kendall rank correlations between SubjectiveEye3D and DBpedia PageRank [9]. The results are similar to our computations. The normalization that has been carried out on the SUB scores mitigates the effect of single peaks and makes an important contribution towards providing objective relevance scores. The work of Eom et al. [5] investigates on the difference between 24 language editions of Wikipedia with PageRank, 2DRank, and CheiRank rankings. The analysis focuses on the rankings of the top-100 persons in each language edition. We consider this analysis as seminal work for investigation on mining cultural differences with Wikipedia rankings. This is an interesting topic as different cultures use the same Wikipedia language edition (e. g., United Kingdom and the United States). Similarly, the work of Lages et al. provide rankings of universities of the world in [6]. Again, 24 language editions were analyzed with PageRank, 2DRank, and CheiRank. PageRank is shown to be efficient in producing similar rankings like the “Academic Ranking of World Universities (ARWU)” (that is provided yearly by the Shanghai Jiao Tong University). In a recent work, Dimitrov et al. introduce a study on the link traversal behavior of users within Wikipedia with respect to the positions of the followed links. Similar to our finding, the authors conclude that a great fraction of clicked links can be found in the top part of the articles.

Comparing ranks on Wikipedia is an important topic and with our contribution we want to emphasize the need for considering the signals “link graph” and “page views” in combination.

6 Summary & Future Work

In this work, we compared different input graphs for the PageRank algorithm, the impact on the scores, and the correlation to page-view-based rankings. The main findings can be summarized as follows:

1. Removing template links has no general influence on the PageRank scores.
2. The results of WLRank with respect to the relative position of a link indicate a better correlation to page-view-based rankings than other PageRank methods.
3. If redirects are resolved, it should be done in a complete manner as otherwise entities get assigned artificially low scores. We recommend using a unredirected dataset for applications in the NLP context.

Currently, we use the link datasets and the PageRank scores in our work on entity summarization [10,11]. However, there are many applications that can make use of objective rankings of entities. As such, we plan to investigate further on the combination of page-view-based rankings and link-based ones. In effect, for humans, rankings of entities are subjective and it is a hard task to approximate “a general notion of importance”.

Acknowledgement. The authors would like to thank Thimo Britsch for his contributions on the first versions of the SiteLinkExtractor tool. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346 and by the German Federal Ministry of Education and Research (BMBF) within the Software Campus project “SumOn” (grant no. 01IS12051).

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Busan, Korea, November 11-15, 2007*. Springer Berlin Heidelberg, 2007.
2. R. Baeza-Yates and E. Davis. Web Page Ranking Using Link Attributes. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, WWW Alt. '04*, pages 328–329, New York, NY, USA, 2004. ACM.
3. S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7, WWW7*, pages 107–117. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 1998.
4. D. Dimitrov, P. Singer, F. Lemmerich, and M. Strohmaier. Visual Positions of Links and Clicks on Wikipedia. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*, pages 27–28. International World Wide Web Conferences Steering Committee, 2016.
5. Y.-H. Eom, P. Aragn, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky. Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. *PLoS ONE*, 10(3):1–27, Mar 2015.
6. Lages, José, Patt, Antoine, and Shepelyansky, Dima L. Wikipedia Ranking of World Universities. *Eur. Phys. J. B*, 89(3):69, Mar 2016.
7. Linné, Carl von and Salvius, Lars. *Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.*, volume v.1. Holmiae :Impensis Direct. Laurentii Salvii, 1758.
8. A. Roa-Valverde, A. Thalhammer, I. Toma, and M.-A. Sicilia. Towards a formal model for sharing and reusing ranking computations. In *Proceedings of the 6th International WS on Ranking in Databases in conjunction with VLDB 2012*, 2012.
9. A. Thalhammer. DBpedia PageRank dataset. Downloaded from http://people.aifb.kit.edu/ath#DBpedia_PageRank, 2016.
10. A. Thalhammer, N. Lasierra, and A. Rettinger. LinkSUM: Using Link Analysis to Summarize Entity Data. In *Proceedings of the 16th International Conference on Web Engineering (ICWE 2016)*. To appear, 2016.
11. A. Thalhammer and A. Rettinger. Browsing DBpedia Entities with Summaries. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 511–515. Springer, 2014.