# PageRank on Wikipedia: Towards General Importance Scores for Entities

Andreas Thalhammer(✉) and Achim Rettinger

AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany
{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 context/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 context/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 [4]. It has been proven to provide objective relevance scores for hyperlinked documents, for example in Wikipedia [6,8,11]. 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 computed on the Wikipedia link graph enable rankings of entities

H. Sack et al. (Eds.): ESWC 2016 Satellite Events, LNCS 9989, pp. 227–240, 2016.
DOI: 10.1007/978-3-319-47602-5_41

**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
```

of specific types, for example for scientists (see Listing 1.1). Although the provided PageRank scores [11] 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 Taxobox[2] 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 (see [9]). In general, entities that help to structure 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, according to our computations articles such as "Bakhsh" (1914), "Powiat" (1408), "Chordate" (1527), and "Lepidoptera" (1778) are occurring in the top-50 list of all things in Wikipedia (see Table 5, column "DBP 2015-04"). These observations led 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 extraction[3] methods that aim to address the root causes of the observed effects. We focus on the question whether some links—based on their context/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 context/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 [4] and a weighted version [2] that accounts for the position of a link within an article.

– Original PageRank [4] – 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 out-going links $c(w) = |\{w_1 | link(w, w_1)\}|$. The PageRank of a page $w_0 \in W$ is computed as follows:

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

– 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:

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

The link weight function $lw$ is defined as follows:

$$lw(link(w_1, w_2)) = 1 - \frac{first\_occurrence(link(w_1, w_2), w_1)}{|tokens(w_1)|} \tag{3}$$

In order to form a correct probability model, the individual link weight is normalized in accordance to the link weights of all outgoing links of a page in Formula 2. If we set the link weight of every incoming link to the same value (e.g., 1) we obtain the original PageRank formula (see Formula 1). The used helper functions of Formula 3 can be described as follows:

- $first\_occurrence(link(w_1, w_2), w_1)$ – the token number of the first occurrence of a $link(w_1, w_2)$ at the respective Wikipedia page $w_1$. The token numbering starts at 1 (i.e., the first word/link in the wikitext).
- $tokens(w_1)$ – the total number of tokens of the Wikipedia page $w_1$. Tokenization is performed as follows: we split the article text in accordance to white spaces but do not split up links (e.g., `[[brown bear|bears]]` is treated as one token).

Both Formulas (1) and (2) are iteratively applied until the scores converge. The variable $d$ is called "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 of the PageRank scores does not influence the final ranking (i.e., the resulting ordering relation between the Wikipedia articles does not change).

## 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 [11] is a link graph that is constructed by the DBpedia Extraction Framework[4] (DEF). The DBpedia extraction is based on Wikipedia database backup dumps[5] 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 between 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 and 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 (i.e., links to all articles of that category) are presented only in the rendered version of the category page (i.e., they 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 SiteLinkExtractor[7] tool. The tool enables to

[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.

$$A \xrightarrow{PL} B \xrightarrow{PL^R} C$$

$$A \xrightarrow{\quad\quad PL \quad\quad} C$$

**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 $C$.

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" in Wikipedia)[9] automatically and various bots are in place to remove them. Therefore, we assume that only single-hop redirects are in place. However, as performed by DBpedia, also single-hop redirects can be resolved (see Fig. 1). Alternatively, for various applications (especially in NLP) it can make sense to keep redirect pages as they also have a high number of inlinks in various cases (e.g., "Countries of the world")[10]. In that case, with reference to Fig. 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). Thus, we assume that the PageRank score of pages of type $C$ is not heavily influenced by resolving/not resolving redirects.

## 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:

[8] DBpedia PageLinks – http://wiki.dbpedia.org/Downloads2015-04.

[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.

**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 (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 fact, 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 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 (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 executed the introduced extraction mechanisms on a dump of the English Wikipedia of February 5, 2015. This date is in line 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.

**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 |

## 4 Experiments

In our experiments, we first computed PageRank on the introduced link graphs. We then measured the pairwise rank correlations (Spearman's $\rho$ and Kendall's $\tau$)[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:

[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 $[-1, 1]$ and are specifically designed for measuring correlations between ranked lists.

**H1** Links in templates are created in a "please fill out this form" manner and rather negatively influence the general estimate of salience that PageRank scores should represent.

**H2** Links that are mentioned at the beginning of articles are more often clicked and, therefore, the ATL-RP and ABL rankings correlate stronger with the page-view-based rankings.

**H3** The practice of resolving redirects does not strongly impact the final ranking (in accordance to PageRank scores) as redirect pages pass most of their score on to the respective target page.

### 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.

### 4.2 Reference Datasets

We use the following rankings as reference datasets:

**DBpedia PageRank (DBP)** The scores of DBpedia PageRank [11] 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 Sect. 4.1. The scores are regularly published as TSV and Turtle files. The Turtle version uses the vRank vocabulary [10]. Since DBpedia version 2015-04, the DBP scores are included in the official DBpedia SPARQL endpoint (see Listing 1.1 for an example query). In this work, we use the following versions of DBP scores based on English Wikipedia: DBpedia 3.8, 3.9, 2014, and 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 Sect. 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).

[14] DBpedia PageLinks Unredirected – http://downloads.dbpedia.org/2015-04/core-i18n/en/page-links-unredirected_en.nt.bz2.

[15] SubjectiveEye3D – https://github.com/paulhoule/telepath/wiki/SubjectiveEye3D.

**The Open Wikipedia Ranking (TOWR)** The TOWR project is maintained by the Laboratory for Web Algorithmics of the Università degli Studi di Milano. It provides Wikipedia rankings in accordance to different ranking methods in a Web interface[16] for direct comparison. They provide the following measures:[17]

**TOWR-PR** PageRank computed on the Wikipedia link graph with the parallel Gauß-Seidel method [7] of the LAW[18] library.

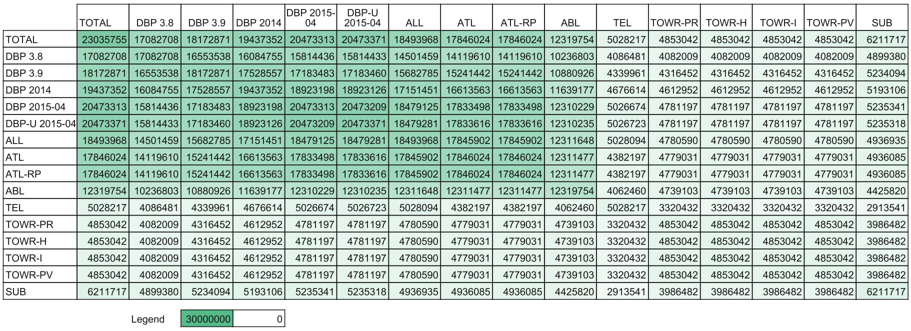
**TOWR-H** Harmonic centrality as introduced in [3] computed on the Wikipedia link graph.

**TOWR-I** Indegree, ranks Wikipedia pages in accordance to their number of incoming links.

**TOWR-PV** Page views, ranks Wikipedia pages in accordance to "the number of page views in the last year"[19].

The two page-views-based rankings (i.e., SUB and TOWR-PV) serve as a reference in order to compare the different graph-based rankings. We show the mutual overlap of entities covered by the individual rankings in Table 2.

**Table 2.** Number of mutually covered entities (the colors are used for better readability and comprise no further meaning).



| | TOTAL | DBP 3.8 | DBP 3.9 | DBP 2014 | DBP 2015-04 | DBP-U 2015-04 | ALL | ATL | ATL-RP | ABL | TEL | TOWR-PR | TOWR-H | TOWR-I | TOWR-PV | SUB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TOTAL | 23035755 | 17082708 | 18172871 | 19437352 | 20473313 | 20473371 | 18493968 | 17846024 | 17846024 | 12319754 | 5028217 | 4853042 | 4853042 | 4853042 | 4853042 | 6211717 |
| DBP 3.8 | 17082708 | 17082708 | 16553538 | 16084755 | 15814436 | 15814433 | 14501459 | 14119610 | 14119610 | 10236803 | 4086481 | 4082009 | 4082009 | 4082009 | 4082009 | 4899380 |
| DBP 3.9 | 18172871 | 16553538 | 18172871 | 17528557 | 17183483 | 17183460 | 15682785 | 15241442 | 15241442 | 10880926 | 4339961 | 4316452 | 4316452 | 4316452 | 4316452 | 5234094 |
| DBP 2014 | 19437352 | 16084755 | 17528557 | 19437352 | 18923198 | 18923126 | 17151451 | 16613563 | 16613563 | 11639177 | 4676614 | 4612952 | 4612952 | 4612952 | 4612952 | 5193106 |
| DBP 2015-04 | 20473313 | 15814436 | 17183483 | 18923198 | 20473313 | 20473209 | 18479125 | 17833498 | 17833498 | 12310229 | 5026674 | 4781197 | 4781197 | 4781197 | 4781197 | 5235341 |
| DBP-U 2015-04 | 20473371 | 15814433 | 17183460 | 18923126 | 20473209 | 20473371 | 18479281 | 17833616 | 17833616 | 12310235 | 5026723 | 4781197 | 4781197 | 4781197 | 4781197 | 5235318 |
| ALL | 18493968 | 14501459 | 15682785 | 17151451 | 18479125 | 18479281 | 18493968 | 17845902 | 17845902 | 12311648 | 5028094 | 4780590 | 4780590 | 4780590 | 4780590 | 4936935 |
| ATL | 17846024 | 14119610 | 15241442 | 16613563 | 17833498 | 17833616 | 17845902 | 17846024 | 17846024 | 12311477 | 4382197 | 4779031 | 4779031 | 4779031 | 4779031 | 4936085 |
| ATL-RP | 17846024 | 14119610 | 15241442 | 16613563 | 17833498 | 17833616 | 17845902 | 17846024 | 17846024 | 12311477 | 4382197 | 4779031 | 4779031 | 4779031 | 4779031 | 4936085 |
| ABL | 12319754 | 10236803 | 10880926 | 11639177 | 12310229 | 12310235 | 12311648 | 12311477 | 12311477 | 12319754 | 4062460 | 4739103 | 4739103 | 4739103 | 4739103 | 4425820 |
| TEL | 5028217 | 4086481 | 4339961 | 4676614 | 5026674 | 5026723 | 5028094 | 4382197 | 4382197 | 4062460 | 5028217 | 3320432 | 3320432 | 3320432 | 3320432 | 2913541 |
| TOWR-PR | 4853042 | 4082009 | 4316452 | 4612952 | 4781197 | 4781197 | 4780590 | 4779031 | 4779031 | 4739103 | 3320432 | 4853042 | 4853042 | 4853042 | 4853042 | 3986482 |
| TOWR-H | 4853042 | 4082009 | 4316452 | 4612952 | 4781197 | 4781197 | 4780590 | 4779031 | 4779031 | 4739103 | 3320432 | 4853042 | 4853042 | 4853042 | 4853042 | 3986482 |
| TOWR-I | 4853042 | 4082009 | 4316452 | 4612952 | 4781197 | 4781197 | 4780590 | 4779031 | 4779031 | 4739103 | 3320432 | 4853042 | 4853042 | 4853042 | 4853042 | 3986482 |
| TOWR-PV | 4853042 | 4082009 | 4316452 | 4612952 | 4781197 | 4781197 | 4780590 | 4779031 | 4779031 | 4739103 | 3320432 | 4853042 | 4853042 | 4853042 | 4853042 | 3986482 |
| SUB | 6211717 | 4899380 | 5234094 | 5193106 | 5235341 | 5235318 | 4936935 | 4936085 | 4936085 | 4425820 | 2913541 | 3986482 | 3986482 | 3986482 | 3986482 | 6211717 |

| Legend | 🟩 30000000 | 0 |
|---|---|---|

## 4.3 Results

We used MATLAB for computing the pairwise Spearman's $\rho$ and Kendall's $\tau$ correlation scores. The Kendall's $\tau$ 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

[16] The Open Wikipedia Ranking – http://wikirank.di.unimi.it/.

[17] For their 2015 edition (that we analyze), the link-graph-based measures are applied on an English Wikipedia extract of April 3, 2015. Links in infoboxes were not considered.

[18] LAW – http://law.di.unimi.it/.

[19] Source: http://wikirank-2015.di.unimi.it/more.html.

Kendall's $\tau$. The pairwise correlation scores of $\rho$ and $\tau$ 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. The DDP rankings correlate strongest with the respective neighboring DBP versions. Also DBP-U 2015-04 and ALL have a very strong correlation (these rankings should be equal).

**Table 3.** Correlation: Spearman's $\rho$ (the colors are used for better readability and comprise no further meaning).

| | DBP 3.8 | DBP 3.9 | DBP 2014 | DBP 2015-04 | DBP-U 2015-04 | ALL | ATL | ATL-RP | ABL | TEL | TOWR-PR | TOWR-H | TOWR-I | TOWR-PV | SUB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DBP 3.8 | 1.000 | 0.965 | 0.930 | 0.885 | 0.696 | 0.689 | 0.692 | 0.646 | 0.672 | 0.295 | 0.832 | 0.736 | 0.777 | 0.624 | 0.541 |
| DBP 3.9 | 0.965 | 1.000 | 0.960 | 0.910 | 0.707 | 0.699 | 0.701 | 0.653 | 0.685 | 0.289 | 0.872 | 0.768 | 0.810 | 0.638 | 0.537 |
| DBP 2014 | 0.930 | 0.960 | 1.000 | 0.941 | 0.719 | 0.709 | 0.712 | 0.661 | 0.700 | 0.278 | 0.904 | 0.796 | 0.836 | 0.648 | 0.502 |
| DBP 2015-04 | 0.885 | 0.910 | 0.941 | 1.000 | 0.771 | 0.756 | 0.758 | 0.708 | 0.770 | 0.164 | 0.772 | 0.697 | 0.723 | 0.654 | 0.551 |
| DBP-U 2015-04 | 0.696 | 0.707 | 0.719 | 0.771 | 1.000 | 1.000 | 0.985 | 0.945 | 0.792 | 0.344 | 0.773 | 0.695 | 0.726 | 0.657 | 0.582 |
| ALL | 0.689 | 0.699 | 0.709 | 0.756 | 1.000 | 1.000 | 0.985 | 0.945 | 0.788 | 0.346 | 0.782 | 0.707 | 0.731 | 0.661 | 0.565 |
| ATL | 0.692 | 0.701 | 0.712 | 0.758 | 0.985 | 0.985 | 1.000 | 0.958 | 0.797 | 0.294 | 0.792 | 0.711 | 0.732 | 0.658 | 0.551 |
| ATL-RP | 0.646 | 0.653 | 0.661 | 0.708 | 0.945 | 0.945 | 0.958 | 1.000 | 0.794 | 0.315 | 0.794 | 0.714 | 0.736 | 0.646 | 0.642 |
| ABL | 0.672 | 0.685 | 0.700 | 0.770 | 0.792 | 0.788 | 0.797 | 0.794 | 1.000 | 0.263 | 0.542 | 0.441 | 0.535 | 0.499 | 0.455 |
| TEL | 0.295 | 0.289 | 0.278 | 0.164 | 0.344 | 0.346 | 0.294 | 0.315 | 0.263 | 1.000 | 0.487 | 0.425 | 0.522 | 0.419 | 0.407 |
| TOWR-PR | 0.832 | 0.872 | 0.904 | 0.772 | 0.773 | 0.782 | 0.792 | 0.794 | 0.542 | 0.487 | 1.000 | 0.859 | 0.889 | 0.645 | 0.593 |
| TOWR-H | 0.736 | 0.768 | 0.796 | 0.697 | 0.695 | 0.707 | 0.711 | 0.714 | 0.441 | 0.425 | 0.859 | 1.000 | 0.809 | 0.677 | 0.614 |
| TOWR-I | 0.777 | 0.810 | 0.836 | 0.723 | 0.726 | 0.731 | 0.732 | 0.736 | 0.535 | 0.522 | 0.889 | 0.809 | 1.000 | 0.668 | 0.616 |
| TOWR-PV | 0.624 | 0.638 | 0.648 | 0.654 | 0.657 | 0.661 | 0.658 | 0.646 | 0.499 | 0.419 | 0.645 | 0.677 | 0.668 | 1.000 | 0.857 |
| SUB | 0.541 | 0.537 | 0.502 | 0.551 | 0.582 | 0.565 | 0.551 | 0.642 | 0.455 | 0.407 | 0.593 | 0.614 | 0.616 | 0.857 | 1.000 |

| Legend |
|---|
| 🟩 1.000 |
| 🟧 0.500 |
| 🟥 0.000 |

**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, this subjective perception of improvement can not be generalized (with the used measures).

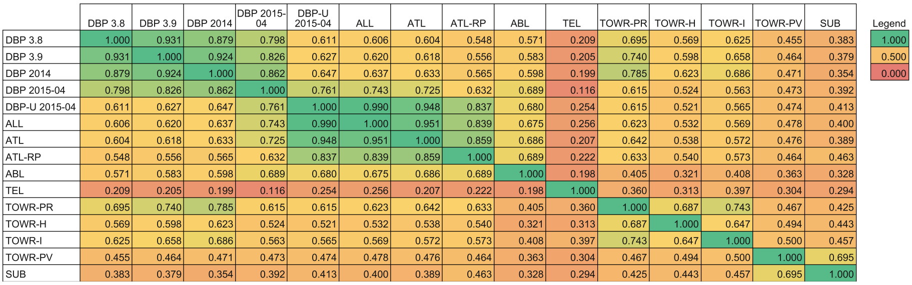
On a side note: we assume that computing PageRank on DBpedia's RDF data would produce similar scores as TEL because DBpedia extracts its semantic relations mostly from Wikipedia's infoboxes.

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 contrast, produces the strongest correlation with SUB. The improvement of ATL-RP comparred to ATL is clearly visible. 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 of Dimitrov et al. [5] where actual HTTP referrer data was analyzed.

With respect to **H3**, we expected DBP-U 2015-04 and DBP 2015-04 to correlate much stronger than the results suggest. As a reason, we found that DEF does not implement the full workflow of Fig. 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, in DBP-U 2015-04, these pages often accumulate PageRank scores of 1000 and above. 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.

**Further Observations.** Another surprising result is the rather weak correlation of TOWR-PR with all the other PageRank-based rankings. As the Wikipedia dump date of DBpedia 2015-04 (that we also used for our measures, see Sect. 3) is only two months apart from the dump date used by TOWR, we expected much stronger correlations here. This is amplified by the observation that TOWR-PR correlates stronger with older DBP versions. However, Table 2 already suggests a clear difference with respect to the number of covered entities. Therefore, we assume that the preprocessing of the link graph performed by TOWR induces this bias. This is also supported by the strong correlations between the link-graph-based TOWR measures (i.e., TOWR-PR, TOWR-H, and TOWR-I) visible in Tables 3 and 4.

**Table 4.** Correlation: Kendall's $\tau$ on a sample of 1 000 000 (the colors are used for better readability and comprise no further meaning).



| | DBP 3.8 | DBP 3.9 | DBP 2014 | DBP 2015-04 | DBP-U 2015-04 | ALL | ATL | ATL-RP | ABL | TEL | TOWR-PR | TOWR-H | TOWR-I | TOWR-PV | SUB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DBP 3.8 | 1.000 | 0.931 | 0.879 | 0.798 | 0.611 | 0.606 | 0.604 | 0.548 | 0.571 | 0.209 | 0.695 | 0.569 | 0.625 | 0.455 | 0.383 |
| DBP 3.9 | 0.931 | 1.000 | 0.924 | 0.826 | 0.627 | 0.620 | 0.618 | 0.556 | 0.583 | 0.205 | 0.740 | 0.598 | 0.658 | 0.464 | 0.379 |
| DBP 2014 | 0.879 | 0.924 | 1.000 | 0.862 | 0.647 | 0.637 | 0.633 | 0.565 | 0.598 | 0.199 | 0.785 | 0.623 | 0.686 | 0.471 | 0.354 |
| DBP 2015-04 | 0.798 | 0.826 | 0.862 | 1.000 | 0.761 | 0.743 | 0.725 | 0.632 | 0.689 | 0.116 | 0.615 | 0.524 | 0.563 | 0.473 | 0.392 |
| DBP-U 2015-04 | 0.611 | 0.627 | 0.647 | 0.761 | 1.000 | 0.990 | 0.948 | 0.837 | 0.680 | 0.254 | 0.615 | 0.521 | 0.565 | 0.474 | 0.413 |
| ALL | 0.606 | 0.620 | 0.637 | 0.743 | 0.990 | 1.000 | 0.951 | 0.839 | 0.675 | 0.256 | 0.623 | 0.532 | 0.569 | 0.478 | 0.400 |
| ATL | 0.604 | 0.618 | 0.633 | 0.725 | 0.948 | 0.951 | 1.000 | 0.859 | 0.686 | 0.207 | 0.642 | 0.538 | 0.572 | 0.476 | 0.389 |
| ATL-RP | 0.548 | 0.556 | 0.565 | 0.632 | 0.837 | 0.839 | 0.859 | 1.000 | 0.689 | 0.222 | 0.633 | 0.540 | 0.573 | 0.464 | 0.463 |
| ABL | 0.571 | 0.583 | 0.598 | 0.689 | 0.680 | 0.675 | 0.686 | 0.689 | 1.000 | 0.198 | 0.405 | 0.321 | 0.408 | 0.363 | 0.328 |
| TEL | 0.209 | 0.205 | 0.199 | 0.116 | 0.254 | 0.256 | 0.207 | 0.222 | 0.198 | 1.000 | 0.360 | 0.313 | 0.397 | 0.304 | 0.294 |
| TOWR-PR | 0.695 | 0.740 | 0.785 | 0.615 | 0.615 | 0.623 | 0.642 | 0.633 | 0.405 | 0.360 | 1.000 | 0.687 | 0.743 | 0.467 | 0.425 |
| TOWR-H | 0.569 | 0.598 | 0.623 | 0.524 | 0.521 | 0.532 | 0.538 | 0.540 | 0.321 | 0.313 | 0.687 | 1.000 | 0.647 | 0.494 | 0.443 |
| TOWR-I | 0.625 | 0.658 | 0.686 | 0.563 | 0.565 | 0.569 | 0.572 | 0.573 | 0.408 | 0.397 | 0.743 | 0.647 | 1.000 | 0.500 | 0.457 |
| TOWR-PV | 0.455 | 0.464 | 0.471 | 0.473 | 0.474 | 0.478 | 0.476 | 0.464 | 0.363 | 0.304 | 0.467 | 0.494 | 0.500 | 1.000 | 0.695 |
| SUB | 0.383 | 0.379 | 0.354 | 0.392 | 0.413 | 0.400 | 0.389 | 0.463 | 0.328 | 0.294 | 0.425 | 0.443 | 0.457 | 0.695 | 1.000 |

| Legend |
|---|
| 🟩 1.000 |
| 🟧 0.500 |
| 🟥 0.000 |

In addition to ATL-RP, also the link-graph-based TOWR measures exhibit a stronger correlation with SUB than the other PageRank-based measures. However, with respect to Table 2 it becomes clear that their overlap with SUB is 949 603 entities less than the one of ATL-RP (or −19% relative to the overlap of ATL-RP and SUB). With this difference, the correlation scores are not directly comparable.

## 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 strong 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 most likely achieved by combining link-structure and page-view-based rankings on Wikipedia. For applications that deal with NLP, we recommend to use the unredirected version of DBpedia PageRank.

## 5 Related Work

There are two common types of Wikipedia rankings: one is based on measures on the link graph, the other is based on consumption (e.g., page views). In the following, we briefly introduce the state of the art in both Wikipedia ranking methods.

**Measures on the Wikipedia link graph:** The work of Eom et al. [6] 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 often use the same language edition of Wikipedia (e.g., United Kingdom and the United States use English). Similarly, the work of Lages et al. provide rankings of universities of the world in [8]. 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). The Open Wikipedia Ranking (TOWR) also applies different graph measures on the Wikipedia link graph (see Sect. 4.2).

The above approaches vary the applied graph measures (PageRank, 2DRank, CheiRank, indegree, harmonic centrality) but do not vary the link extraction methods. In this paper, we experiment with both, different input graphs and a combination of a new weighted input graph and WLRank.

**Wikipedia consumption patterns:** The official page view statistics of various Wikipedia projects are publicly available as dumps[20] or as a Web API[21]. Our work on this paper was mainly influenced and motivated by an initial experiment that was performed by Paul Houle: in the Github project documentation of SubjectiveEye3D (see Sect. 4.2 for more details on SubjectiveEye3D), he reports

[20] Page view statistics for Wikimedia projects – https://dumps.wikimedia.org/other/pagecounts-raw/.

[21] Wikipedia Pageview API – https://wikitech.wikimedia.org/wiki/Analytics/PageviewAPI.

**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 |

about Spearman and Kendall rank correlations between SubjectiveEye3D and our published PageRank computations [11].[22] His results are similar to our computations. 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 [5]. 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 features “link graph” and “page views” in combination.

## 6 Summary and 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 an unredirected dataset for applications in the NLP context.

Currently, we use the link datasets and the PageRank scores in our work on entity summarization [12,13]. However, there are many applications that can make use of objective rankings of entities. Therefore, we plan to investigate further on the combination of page-view-based rankings and link-graph-based ones. In effect, for humans, rankings of entities are subjective and it is a hard task to approximate “a general notion of importance”.

**Acknowledgments.** The authors would like to thank Thimo Britsch for his contributions on the first versions of the SiteLinkExtractor tool. They also would like to thank Paul Houle and Sebastiano Vigna for their pointers and insights. 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).

[22] Paul Houle on the correlation between DBP and SUB – https://github.com/paulhoule/telepath/wiki/Correlation-of-Subjective-Importance-Scores.

## References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
2. Baeza-Yates, R., Davis E.: Web page ranking using link attributes. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Amp; Posters, WWW Alt. 2004, pp. 328–329. ACM, New York (2004)
3. Boldi, P., Vigna, S.: Axioms for centrality. Internet Math. **10**(3–4), 222–262 (2014)
4. Brin, S., Page, L.: The Anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International Conference on World Wide Web 7, pp. 107–117. Elsevier Science Publishers B. V, Amsterdam (1998)
5. Dimitrov, D., Singer, P., Lemmerich, F., Strohmaier, M.: Visual positions of links and clicks on Wikipedia. In: Proceedings of the 25th International Conference Companion on World Wide Web, WWW 2016 Companion, pp. 27–28. International World Wide Web Conferences Steering Committee (2016)
6. Eom, Y.-H., Aragn, P., Laniado, D., Kaltenbrunner, A., Vigna, S., Shepelyansky, D.L.: Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. PLoS ONE **10**(3), 1–27 (2015)
7. Kohlschütter, C., Chirita, P.-A., Nejdl, W.: Efficient parallel computation of pagerank. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 241–252. Springer, Heidelberg (2006)
8. Lages, J., Patt, A., Shepelyansky, D.L.: Wikipedia ranking of world universities. Eur. Phys. J. B **89**(3), 69 (2016)
9. von Linné, C., Salvius, L., Linnaei, C.: Systema naturae per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis., volume v. 1. Impensis Direct. Laurentii Salvii, Holmiae (1758)
10. Roa-Valverde, A., Thalhammer, A., Toma, I., Sicilia, M.-A.: 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)
11. Thalhammer, A.: DBpedia pagerank dataset (2016). http://people.aifb.kit.edu/ath#DBpedia_PageRank
12. Thalhammer, A., Lasierra, N., Rettinger, A.: LinkSUM: using link analysis to summarize entity data. In: Bozzon, A., Cudré-Mauroux, P., Pautasso, C. (eds.) ICWE 2016. LNCS, vol. 9671, pp. 244–261. Springer, Heidelberg (2016). doi:10.1007/978-3-319-38791-8_14
13. Thalhammer, A., Rettinger, A.: Browsing DBpedia entities with summaries. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8798, pp. 511–515. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11955-7_76