Assigning Global Relevance Scores to DBpedia Facts

Philipp Langer, Patrick Schulze, Stefan George, Matthias Kohnen, Tobias Metzke, Ziawasch Abedjan, and Gjergji Kasneci

Hasso Plattner Institute (HPI)
Potsdam, Germany
first.last@hpi.uni-potsdam.de

***Abstract*—Knowledge bases have become ubiquitous assets in today's Web. They provide access to billions of statements about real-world entities derived from governmental, institutional, product-oriented, bibliographic, bio-chemical, and many other domain-oriented and general-purpose datasets. The sheer amount of statements that can be retrieved for a given entity calls for ranking techniques that return the most salient, i.e., globally relevant, statements as top results.**

In this paper we analyze and compare various strategies for assigning global relevance scores to DBpedia facts with the goal to derive the best one among these strategies. Some of these strategies build on complementary aspects such as frequency and inverse document frequency, yet others combine structural information about the underlying knowledge graph with Web-based co-occurrence statistics for entity pairs. A user evaluation of the discussed approaches has been conducted on the popular DBpedia knowledge base with statistics derived from an indexed version of the ClueWeb09 corpus.

The created dataset can be seen as a strong baseline for comparing entity ranking strategies (especially, in terms of global relevance) and can be used as a building block for developing new ranking and mining techniques on linked data.

I. Introduction

Open structured datasets with factual information from a wide range of domains bear the potential to become prime sources of machine processable data. Frequently, in literature, two key advantages of these datasets over unstructured ones are mentioned [1], [2]:

- They enable search for explicit facts, which can be useful in scholarly and academic (research) applications.
- They enable automated knowledge discovery mechanisms that are based on the combination of explicit facts about people, companies, products, etc. Page-oriented search would involve exhaustive manual inspection of many search results, thus being impractical for these kinds of search needs.

Unfortunately, for a long time, these structured Web assets have been mainly of academic interest, and apart from poof-of-concept search systems, e.g., Wolfram Alpha [1], hakia [2], Google Tables [3], NAGA [3], and other academic search systems over linked data [4], [5], there has been little use for structured knowledge. However, recently, Google introduced an integration of unstructured search results with structured data coming from Google's Knowledge Graph – a knowledge base of facts about popular named entities occurring on the Web. For queries related to such entities, users can choose to inspect the textual snippets or browse through factual data and other related entities. Shortly after the English version, this feature was introduced in other languages as well, suggesting that it has been well received by the users. We think that an important factor to this success is the fact that Google returns a relatively small amount of factual information, typically containing *a handful of salient facts about the query entity*. For example, when searching for Albert Einstein, the user is shown factual information about his birth and death place, alma mater, family, prizes, and other related prominent physicists. The global relevance of the displayed factual information for the query entity plays a crucial role for Google's internal ranking and final presentation of knowledge graph facts. If we extrapolate this idea further and adapt it to the case of open structured Web datasets, a general requirement is to rank facts about query entities in such a way that globally relevant facts appear as top results. For example, when asking for facts about Albert Einstein, DBpedia returns a long list of facts, many of which are either too general (e.g., such as Einstein being a person) or too obscure (e.g., Einstein being a violinist). Similarly, it is difficult to retrieve from DBpedia a list of popular physicists, as the majority of returned results are about commonly less known physicists. For the SPARQL(http://www.w3.org/TR/rdf-sparql-query/) query

```
SELECT Albert_Einstein ?p ?o
  WHERE{Albert_Einstein ?p ?o}
```

(that asks for facts about Albert Einstein) the top-5 DBpedia results are shown in Figure 1.

:Albert_Einstein	rdf:type	owl:Thing
:Albert_Einstein	rdf:type	foaf:Person
:Albert_Einstein	rdf:type	dbpedia:ontology/Person
:Albert_Einstein	rdf:type	dbpedia:ontology/Agent
:Albert_Einstein	rdf:type	dbpedia:ontology/Scientist

Fig. 1. Top-5 facts about Albert Einstein from http://dbpedia.org/snorql/

In a Web application along the lines of Google's knowledge graph, where a succinct set of representative facts about entities is needed, such results would be impractical.

In this paper we focus on two types of SPARQL queries:

- *Subject queries* of the form `SELECT ?s WHERE{?s p o}`, where `?s` denotes the subject variable and `p` and `o` are the given predicate and object, respectively.

[1]http://www.wolframalpha.com/
[2]http://hakia.com/
[3]http://research.google.com/tables

- *Property queries* of the form `SELECT ?p ?o WHERE{s ?p ?o}`, where `s` is a given subject and `?p` and `?o` denote the predicate and object variable, respectively.

We analyze and compare common ranking strategies for these types of queries. Some of these strategies build on complementary aspects such as frequency and inverse document frequency, others combine structural information about the underlying knowledge graph with Web-based co-occurrence statistics for entity pairs. In summary, the contributions presented in this paper are the following:

1) For subject and property queries, we analyze generic information retrieval models for assigning global relevance scores to result facts expressed in the triple formalism of the Resource Description Framework (RDF).
2) We have evaluated the discussed ranking techniques in an extensive user study based on queries to the popular DBpedia dataset.
3) We provide a new annotated version of the DBpedia dataset, with precomputed fact scores according to the best ranking strategy that was identified by the user study. The dataset can be found at:

https://www.hpi.uni-potsdam.de/naumann/sites/dbpedia/

II. Ranking for Subject and Property Queries

In this section, we will highlight three generic ranking mechanisms for subject and property queries. The models are based on common information retrieval ranking techniques for ranking. We adapt these techniques to rank subject and property queries by global relevance and compare them in order to find the most appropriate among them.

The first two mechanisms presented below exploit the structural properties and information-theoretic statistics derived from the DBpedia knowledge graph only (i.e., without taking external sources into account). The third one combines an authority-based measure for subjects with Web-based co-occurrence statistics for subject-object pairs. These mechanisms build on the following definitions.

In order to derive information-theoretic measures from the knowledge base, we need to introduce the notion of documents, based on which frequency, document frequency and other information-theoretic and redundancy-oriented statistics can be estimated.

Definition 1 (Subject, Predicate, and Object Documents): Given a knowledge base G_{KB}, the *document of a subject* s is given by $d_s = \{(s,x,y)|(s,x,y) \in G_{KB}\}$, the *document of a predicate* p by $d_p = \{(x,p,y)|(x,p,y) \in G_{KB}\}$, and the *document of an object* o by $d_o = \{(x,y,o)|(x,y,o) \in G_{KB}\}$.

A. Ranking by Frequency and Document Frequency

This utilizes information-theoretic notions, such as term frequency and document frequency, known from information retrieval, to formalize the prominence and information content of result entities and is similar in spirit to the approach presented in [6] for query relaxation. The same estimations can be used for a global relevance ranking of results to subject and property queries.

For subject queries, this strategy favors subjects that are globally relevant. For property queries, it up-weights facts that are globally relevant and specific to a given query subject.

Before introducing the ranking model for subject and property queries, we present the term and document frequency analogies for subjects and properties. The frequency of a subject s in a predicate document d_p is given by:

$$freq(s;p) = \frac{\#\{s|(s,p,x) \in d_p)\}}{|d_p|} \quad (1)$$

This frequency estimation captures the information content about a subject entity for a given predicate and will be crucial for ranking results to subject queries. The intuition behind this measure is that for n:m relationships, popular entities are expected to have a richer description (e.g., a higher number of *type* triples) than less known ones. Analogously, one can define the frequency parameters $freq(p;s)$, $freq(s;o)$, and $freq(o;s)$.

The document frequency of a subject s is:

$$dfreq(s) = \frac{|\{p|(s,p,*) \in d_s\}| + |\{o|(s,*,o) \in d_s\}|}{|D|} \quad (2)$$

where D is the multiset of all possible documents that can be derived from the knowledge graph (the multiplicity of documents is the result of their construction with respect to s, p, and o). This measure represents the likelihood of randomly selecting an object document or a predicate document that contains the query subject s. Hence, the measure captures the diversity in terms of different predicates and different objects a subjects occurs with; the higher this number, the more popular the subject is expected to be. This is in contrast to the intuition from information retrieval that the higher the document frequency for a term the lower its information content. Thus, for subject queries, we rank the resulting subjects by:

$$Score(s;p,o) = freq(s;p) \cdot freq(s;o) \cdot log(1+dfreq(s)) \quad (3)$$

Note that, because the knowledge base is redundancy-free in terms of facts, $freq(s;o)$ yields similar values for the majority of subject-object pairs. Hence, the crucial ranking parameter in the above equation is $freq(s;p)$.

In contrast to the popularity of subjects, for an object o, it holds that the higher the document frequency of o, the more generic o is; in consequence, the lower is its descriptive power. Hence, for property queries, we rank the resulting predicate-object pairs by:

$$Score(p,o;s) = \frac{freq(p;s) \cdot freq(o;s)}{log(1+dfreq(p)) \cdot log(1+dfreq(o))} \quad (4)$$

where $dfreq(p)$ and $dfreq(o)$ are defined analogously to $dfreq(s)$.

A general drawback of this ranking strategy is that, since it exploits only the knowledge graph (which is free of redundancy in terms of facts), it cannot capture the explicit contextual saliency of a subject with respect to a given predicate and object. Although the ranking of results to subject queries was often satisfactory from a user's perspective, for property queries, the human judges often reported that the ranking

seemed rather arbitrary, see Sec. III. Anecdotically, in Table I, we show the top-5 results for our running-example queries, i.e., the subject query that asks for theoretical physicists and the property query that asks for the properties of Albert Einstein.

TABLE I
RESULTS RETURNED BY THE FREQ- AND DFREQ-BASED RANKING STRATEGY FOR THE SUBJECT QUERY FOR PHYSICISTS (LEFT) AND THE PROPERTY QUERY FOR PROPERTIES OF ALBERT EINSTEIN (RIGHT)

Rank	Subject query results	Property query results
1	Isaac Newton	residence German Empire
2	Albert Einstein	birthPlace Ulm
3	Max Born	academicAdvisor Heinrich Fr...
4	Niels Bohr	academicAdvisor Heinrich Fr...
5	Enrico Fermi	birthPlace German Empire

B. Ranking by Information Diversity

A critical aspect of the previous ranking model was the ranking of results to property queries. Therefore, we further created a probabilistic ranking model, which aims to explicitly capture the importance of a property (i.e., of a predicate and an object) for a given subject in a property query.

For property queries, the probability $P(p, o|s)$ of a predicate and an object given a subject specified in the query can be approximated in two ways:

$$P(p, o|s) = P(p|o, s)P(o|s) \approx P(p|o)P(o|s) \quad (5)$$

$$P(p, o|s) = P(o|p, s)P(p|s) \approx P(o|p)P(p|s) \quad (6)$$

An interpolation between these two approximations yields the final estimation of $P(p, o|s)$ as:

$$P(p, o|s) \approx \alpha P(p|o)P(o|s) + (1 - \alpha)P(o|p)P(p|s) \quad (7)$$

where $0 \leq \alpha \leq 1$ and $P(p|o), P(o|s), P(o|p), P(p|s)$ are estimated through their relative counts in the knowledge base. Empirically, we found that $\alpha = 0.5$ leads to satisfactory rankings from a user's perspective (see also Section III).

Note that the above approximation involves conditional independence assumptions, e.g., in Approximation (5), p is assumed to be independent of s given o. While this assumption is not always true, it enables the estimation of the model parameters from the knowledge graph and leads to decent empirical rankings. It is also important to note that a direct estimation of $P(p, o|s)$ in terms of its maximum likelihood estimation from the knowledge base would not be meaningful given that the knowledge base is free of redundancy.

Table II depicts the top-5 results to the subject query that asks for theoretical physicists and the property query that asks for the properties of Albert Einstein. The ranking differs from I with respect to the property query; this time results such as "Physics", "Jewish", "Scientist" that are more generally associated with Einstein are promoted.

For subject queries, we estimate the probability $P(s|p, o)$ of a result subject given a predicate and an object specified in the query by:

$$P(s|p, o) = \frac{P(s)P(p, o|s)}{P(p, o)} \quad (8)$$

where $P(s)$ is estimated by the relative frequency of the subject s in the knowledge base triples and can be interpreted

TABLE II
RANKING BY INFORMATION DIVERSITY: ON THE LEFT, RESULTS TO THE SUBJECT QUERY PHYSICISTS; ON THE RIGHT, THE RESULTS FOR THE PROPERTY QUERY FOR THE PROPERTIES OF ALBERT EINSTEIN

Rank	Subject query results	Property query results
1	Isaac Newton	field Physics
2	Albert Einstein	fields Physics
3	Max Born	ethnicity Jewish
4	Niels Bohr	type Scientist
5	Enrico Fermi	deathPlace United States

as a prominence prior for the subject entity; the parameter $P(p, o|s)$ is approximated as above (see Approximation (7)). Finally, since $P(p, o)$ is the same for all results, for ranking purposes, it can be omitted from the estimation and we can write:

$$P(s|p, o) \propto P(s)P(p, o|s) \quad (9)$$

Note that directly estimating the maximum likelihood of $P(s|p, o)$ is impossible, since each fact occurs only once in the knowledge base (assuming that it is free of redundancy), hence the above detour through approximations and independence assumptions.

Although, to the best of our knowledge, this ranking model has never been formalized in such a general way in related work, parts of it (e.g., the approximation of $P(p, o|s)$ through $P(o|s)$) can be found in [3] and [7].

In summary, the information diversity model is easy-to-implement and effective, especially when evaluated with respect to the top- and bottom-ranked entities. However, it also bears some limitations, especially with respect to the estimations of the above parameters, which, because of lack of redundancy in DBpedia (and similar knowledge bases), are actually biased underestimations. As we will see later, this problem can be partially alleviated by taking Web-based co-occurrence statistics for subject-object pairs into account. Another problem of derived knowledge bases is that properties of subjects often follow given templates, which is standardized for various entity types. Empirically, for such entities, we found that the ranking by the above model is rather arbitrary.

C. Random Walk Ranking and Web-based Co-occurrence Statistics

The previous strategies aimed at capturing the global relevance of resulting subjects and properties to corresponding queries by relying only on the underlying knowledge graph. Often such relevance scores relate to the general prominence of subjects and objects in the knowledge base in the sense that if a result entity is connected to many other entities or literals, it is expected to be ranked among the top results, regardless of the information given in the query. A common possibility for deriving global prominence scores for subjects and objects from the knowledge graph is to employ a *Random Walk* model along the lines of *PageRank* [8]. Typically, the random walk model works best, when the indegree of a node reflects some kind of *endorsement* for that node; this, however, is not necessarily the case in knowledge graphs. Hence, when adapting the random walk model to knowledge graphs, there are a few limitations that need to be considered: (1) Scores for literals need to be carefully considered, since literals represent leafs in the knowledge graph and would amass most of the

authority from a random walk process. (2) Ranking with respect to predicates is difficult without a major reconciliation of the graph model, in which predicates would be represented as nodes. However, in such a model, predicate nodes would have a much higher indegree than other node types and would therefore amass most of the authority from the random walk process.

The first two problems have also been recognized by [7], where the directed knowledge graph is transformed into a weighted bidirectional graph, with the goal to run a PageRank-like algorithm on it. We follow a simpler strategy, which empirically, as we will see in the evaluation section, yields a highly effective ranking strategy on DBpedia.

To our rescue comes further data provided by DBpedia. Fortunately, the DBpedia knowledge graph provides two types of links that represent entity endorsement to some extent:

1) *Wiki Pagelinks* are derived from the links between Wikipedia articles
2) *Infobox Property Mappings* link Wikipedia infobox properties and labels to entities and literals

To this end, we have implemented a random walk model that also considers these two kinds of links, which can mitigate the two problems mentioned above. Empirically, we found that for subject queries, the random walk model yields highly satisfactory ranking of results (see Section III). For such queries, the ranking captures the general prominence of entities astonishingly well. However, for property queries the ranking was less satisfactory, typically favoring generic and less specific properties. For example, objects with high indegree such as "United States" and "Switzerland" are ranked as top results when asking for the properties of Albert Einstein.

To address this problem, we followed the ranking model proposed in [7] by means of co-occurrence statistics derived from the Web. More specifically, given a subject s, we are interested in those objects o that are generally *associated with* s and yield a high $P(o|s)$. For example, Albert Einstein is generally rather associated with physics than with Switzerland or the United States; hence, we would expect $P(\text{"}\textit{Physics}\text{"}|\text{"}\textit{Albert Einstein}\text{"}) > P(\text{"}\textit{United States}\text{"}|\textit{Albert Einstein}\text{"})$. To estimate such conditionals, we indexed the ClueWeb09 [4] corpus paragraph-wise. Note that this is a much larger corpus than the sample corpora used in [7] to derive the co-occurrence statistics. We constructed 5 different indexes for paragraphs of different length, i.e., paragraphs containing 8, 16, 32, 64, and 128 words (after stop-word removal). The indexes were queried to retrieve the number of co-occurrences for a given subject-object pair. For example, to estimate $P(\text{"}Physics\text{"}|\text{"}Albert\ Einstein\text{"})$, we query the index for all paragraphs that contain the keywords *"Physics Albert Einstein"* and the paragraphs that contain *"Albert Einstein"* only. Note that since predicates are expressed in various ways in natural language text, it is difficult to derive meaningful statistics for the direct estimation of $P(o,p|s)$, hence our approximation through $P(o|s)$:

$$P(o|s) \approx \frac{\#par(o \wedge s)}{\#par(s)} = \frac{\#par(o \wedge s)}{\sum_{o_x} \#par(o_x \wedge s)} \quad (10)$$

[4]http://lemurproject.org/clueweb09/

where $\#par(s)$ denotes the number of paragraphs that contain s and $\#par(o \wedge s)$ the number of paragraphs that contain both, o and s. Equation 10 shows that all that is needed for the above estimation are co-occurrence statistics for subject-object pairs; that is, $\#par(s)$ does not have to be stored explicitly. Analogously, we can use the same co-occurrence statistics to estimate the importance of a subject for a given object, i.e., by the probability $P(s|o)$. In fact, we have evaluated this ranking strategy for subject queries as well (see Section III); however, the best ranking performance was achieved by the combination of the random walk model for subject queries and the co-occurrence statistics for property queries.

To the best of our knowledge none of the prior works that have proposed such Web-based co-occurrence statistics has tried to derive them from a corpus of the size described in this work, nor have they looked into a paragraph-based indexing of the used corpus. [7] has used a page-wise index on Wikipedia articles (back then Wikipedia contained around 3 million English articles). In fact, all English Wikipedia articles are contained in the TREC Category B ClueWeb09 corpus as well. Also, note that indexing a corpus page-wise may lead to biased co-occurrence statistics, since an object and a subject occurring on the same page do not necessarily stand in a relationship. Instead, we have indexed hundreds of millions of paragraphs, with the goal to derive more meaningful and accurate co-occurrence statistics.

In a final, combined ranking model, we use the authority scores derived from the above random walk model to rank the results of subject queries and the co-occurrence statistics to rank the results of property queries. Table III shows the top-ten results for our running-example queries, i.e., the subject query that asks for theoretical physicists and the property query that asks for the properties of Albert Einstein. This anecdotic example shows the impact of the Web-based co-occurrence statistics (computed on paragraphs of length 128), which indeed promote properties such as "$Physics$", "$Physicist$", and "$Scientist$" in the ranking.

TABLE III
COMBINING THE RANDOM WALK MODEL FOR SUBJECT QUERIES (LEFT) WITH THE WEB-BASED CO-OCCURRENCE STATISTICS FOR PROPERTY QUERIES (RIGHT)

Rank	Subject query results	Property query results
1	Albert Einstein	fields Physics
2	Isaac Newton	field Physics
3	Galileo Galilei	deathPlace United States
4	James Clerk Maxwell	placeOfDeath United States
5	Richard Feynman	shortDescription Physicist

III. EXPERIMENTAL EVALUATION

To investigate the effectiveness of the discussed ranking models and the setting of the various parameters we conducted to different studies *User Study I* and *User study II*. In the following, we first describe the quality measures and experimental environment, and then we describe consecutively *User Study I* and *User study II* by describing the corresponding setting and reporting the experimental results.

To measure the quality of each ranking strategy we used the popular measures Mean-Average Precision (MAP) and average Normalized Discounted Cumulative Gain ($aNDCG$)[9].

A. Experimental Data

Our system ranks DBpedia facts as available in the v3.8 release. Using an inverted index created with Apache Lucene v4.0, we derive the Web-based co-occurrence statistics for subject-object pairs. The index is created from the Category B ClueWeb09 corpus, which consists of 50 million English-language pages. The indexing as well as the computation of the result rankings were performed on a server with 4 Intel Xeon E7-8837 CPUs with a total number of 32 physical cores and 256 GB RAM. For the indexing process we divided the ClueWeb09 corpus in 30 equally sized chunks. We pre-computed co-occurrence statistics for all 60 million facts in the DBpedia infobox dataset which took about 5 days. We needed another 6 days for the calculation of all random walk scores on the the graph mentioned in Section II-C. Additionally, we precomputed statistics derived from the DBpedia knowledge graph, such as the number of facts an entity occurs in, the number of properties it has, and the page rank, and stored them into a database.

B. User Study I

In *User Study I*, we evaluated 19 different versions of the presented models, the information diversity model, the frequency & document frequency model, and the Web-based co-occurrence statistics to derive the best parameter configurations (e.g., paragraph length for co-occurrence statistics, various alphas for the information diversity model, etc.). Each version can be viewed as a specific ranking model, and the goal of User Study I was to quickly preselect the four most promising models.

To preselect the most promising models, in User Study I, for each of the 19 specific models, 12 human judges (students from our department) were shown the results to 4 subject queries and 4 property queries. In order to get meaningful evaluations and to make profound judgments possible, in both cases, we selected queries about prominent entities. For each of the 19 anonymized models, the judges were given the task to place each result item (some queries returned more than 100 results) of the 8 result sets in one of the four categories: highly relevant, relevant, less relevant, and irrelevant. To resolve contradictions between multiple evaluations of the same item, result items were mapped to the category that was determined by majority decision on the users' evaluations.

Table IV shows the top-4 results in terms of MAP and $aNDCG$, for subject and property queries, respectively. The best empirical performance was achieved by the random walk model on the DBpedia knowledge graph, the information diversity model, the frequency & document frequency model, and the Web-based co-occurrence statistics with paragraph lengths of 128 words. Note that the latter and the random walk model were evaluated in separation; that is, we used each of the models to rank both, the results of subject and property queries. Overall, the random walk model achieves the highest MAP and $aNDCG$ scores on subject queries, whereas the Web-based co-occurrence statistics achieve the best performance for property queries. The information diversity model considerably outperforms the frequency & document frequency model.

TABLE IV
TOP-4 RANKING MODELS OF USER STUDY I, FOR SUBJECT QUERIES (SQS) AND PROPERTY QUERIES (PQS), RESPECTIVELY

Query type	SQs	PQs	SQs	PQs
Ranking model	MAP	MAP	$aNDCG$	$aNDCG$
Random walk	**0.797**	0.53	**0.9**	0.803
Info. diversity	0.751	0.637	0.88	0.86
Freq. & doc. freq.	0.604	0.54	0.86	0.805
Co-occ. stats (PL128)	0.71	**0.641**	0.878	**0.88**

TABLE V
COMPARISON OF THE TOP-10 RESULTS FOR EACH OF THE 14 SUBJECT QUERIES (SQS) AND 14 PROPERTY QUERIES (PQS).

Ranking strategy	$aNDCG$		
	SQs	PQs	combined
Random walk	0.896	0.797	0.847
Info. diversity	0.507	0.698	0.602
Co-occ. stats (PL128)	0.618	0.832	0.724
RW+PL128	0.896	0.832	**0.864**

C. User study II

To further evaluate these initial findings, we conducted *User study II*. This time, we had 10 human judges who evaluated the three most promising ranking models identified by User Study I as well as the combination of the random walk model (for subject queries) with the Web-based co-occurrence statistics (for property queries).

The evaluation was done based on a side-by-side comparison of the different ranking models for 28 different queries; 14 subject queries and 14 property queries. Again, for the sake of meaningful evaluations and to enable profound judgments (by the users), in addition to the 8 queries from User Study I, we added 10 property queries by randomly selecting 10 entities from 5 different DBpedia categories (i.e., Theoretical Physicists, Grammy Winners, Male "Best Actor" Academy Award Winners, Super Heavyweight Boxers, and U.S. Presidents). Another 10 subject queries were chosen by leveraging the DBpedia ontology class hierarchy. We randomly selected 10 suitable ontology classes. A selected class was deemed suitable if (1) it contained at least 10 entities and (2) it was not too obscure such as, e.g., *AustralianRulesFootballPlayer*. The final query set for this study is shown at https://www.hpi.uni-potsdam.de/naumann/sites/dbpedia/.

For each of the 28 queries, the users were shown the top-10 results generated by different anonymized ranking models. As before, each of the results in a top-10 list, could be labeled as highly relevant, relevant, less relevant, and irrelevant. This labeling allowed us to calculate the $aNDCG$ score for each model. Note that this side-by-side comparison of rankings is used as a standard technique for the evaluation of search engines; it can lead to evaluation results that are quite different from those of separate ranking evaluations, because in such a setting, the user preference about a ranking can change, once other alternatives are made available.

Table V shows the achieved $aNDCG$ scores for the random walk model, the information diversity model, the Web-based co-occurrence statistics (with paragraph length of 128 words), and the combined model that uses the random walk strategy for subject queries and the co-occurrence statistics for property queries (denoted by RW+PL128 in Table V). The combination

of the random walk strategy for subject queries and the co-occurrence statistics for property queries achieves the highest overall $aNDCG$ score. Of course, all the ranking models evaluated here were geared towards the DBpedia dataset. However, we are confident that their generic nature allows their adaption to other similar knowledge bases (e.g., YAGO).

IV. Related Work

In order to enhance the usability of knowledge bases, various approaches to relax and rank query results have been introduced [10], [11], [12], [6], [13], [14], [15], [7], [3]. Kasneci et al. [3] introduced NAGA, a SPARQL-like query language using triple patterns and a language model for computing informativeness of facts and to rank the results. Elbassuoni et al. [12], [6], [13] further elaborated on this strategy by also considering the results of relaxed queries (i.e., if there are too few results for the original query). Anyanwu et. al. [11] present an approach based to rank semantic associations for conventional search as well as discovery search, however they explicitly do not aim at ranking facts by their global relevance.

The MING algorithm [7] introduces an informativeness measure that builds on a natural extension of the random surfer model that underlies PageRank [8]. The extension concerns edge weights (for the knowledge graph) that are based on page-based co-occurrence statistics derived from the Wikipedia corpus. The final scores are used to capture the informativeness of entire subgraphs. Our work differs from all these prior approaches by investigating a wide range of generic ranking strategies and paragraph-based co-occurrence statistics derived from a much larger Web corpus (at least an order of magnitude larger than the corpora used in prior work).

Indeed, the approaches that we have investigated in this work are geared towards the DBpedia dataset (as it is also the case with much of the prior work, e.g., [12], [6], [13], [7], [3], which are geared towards the YAGO [16] dataset); however, all the models presented in this paper are generic enough to be applied to any RDF knowledge base.

Another stream of approaches [17], [18] has investigated probabilistic models for deriving the truthfulness of statements in RDF knowledge bases by aggregating user feedback. Note however that ranking by truthfulness is different from saliency-based ranking.

Entity summarization is also a related field, since in order to summarize an entity the most relevant properties have to be identified. RELIN [19] is an entity summarization approach that is based on the random surfer model. Our experimental evaluation of this model and its comparison with other ranking strategies for global relevance, has shown that a random-surfer-based ranking strategy is highly satisfactory for subject queries. Recently, the new system DIVERSUM [20] was proposed that focuses on diversification in graphical entity summarization. Diversification is indeed a very interesting and useful concept in the context of ranking factual information. However, it is beyond the scope of this work and is part of our future work agenda. With regard to query relaxation, [10], [6], [15] present approaches to improve the usability of knowledge bases. While [15] presents a relaxation operator to logically remove conditions from queries, [10], [6] relax query results using information retrieval techniques. Our approach is orthogonal to query relaxation.

V. Conclusion

In this paper, we investigated a wide range of generic, global-relevance ranking strategies on the DBpedia dataset. By doing so, the work has shed light on the capabilities of information-theoretic, statistical, and random walk models on a coherent dataset of triples (DBpedia). We found that despite the lack of redundancy (with respect to facts), the DBpedia knowledge graph bears enough information to enable information-theoretic and random walk models that translate to highly satisfactory rankings of results, especially for subject queries. For property queries, the ranking could be further improved by taking Web-based co-occurrence statistics into account. Furthermore, we provide our dataset with precomputed co-occurence for DBpedia facts on our website:

https://www.hpi.uni-potsdam.de/naumann/sites/dbpedia/

We hope that the dataset and our analysis will motivate further research on practical search and ranking techniques over RDF data.

References

[1] G. Kasneci, M. Ramanath, F. Suchanek, and G. Weikum, "The yago-naga approach to knowledge discovery," *SIGMOD Record*, vol. 37, no. 4, pp. 41–47, 2009.

[2] G. Weikum, G. Kasneci, M. Ramanath, and F. Suchanek, "Database and information-retrieval methods for knowledge discovery," *Commun. ACM*, vol. 52, no. 4, pp. 56–64, 2009.

[3] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum, "Naga: Searching and ranking knowledge," in *ICDE*, 2008, pp. 953–962.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *Journal of Web Semantics*, vol. 7, pp. 154–165, 2009.

[5] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker, "Sig.ma: Live views on the web of data," *J. Web Sem.*, vol. 8, no. 4, pp. 355–364, 2010.

[6] S. Elbassuoni, M. Ramanath, and G. Weikum, "Query relaxation for entity-relationship search," in *ESWC*, 2011, pp. 62–76.

[7] G. Kasneci, S. Elbassuoni, and G. Weikum, "Ming: mining informative entity relationship subgraphs," in *CIKM*, 2009, pp. 1653–1656.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," *WWW Internet And Web Information Systems*, vol. 54, no. 2, pp. 1–17, 1998.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[10] Z. Abedjan and F. Naumann, "Synonym analysis for predicate expansion," in *ESWC*, 2013, pp. 140–154.

[11] K. Anyanwu, A. Maduko, and A. Sheth, "Semrank: ranking complex relationship search results on the semantic web," in *WWW*, 2005, pp. 117–127.

[12] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum, "Language-model-based ranking for queries on rdf-graphs," in *CIKM*, 2009, pp. 977–986.

[13] S. Elbassuoni, M. Ramanath, and G. Weikum, "Rdf xpress: a flexible expressive rdf search engine," in *SIGIR*, 2012.

[14] T. Franz, A. Schultz, S. Sizov, and S. Staab, "Triplerank: Ranking semantic web data by tensor decomposition," in *ISWC*, 2009, pp. 213–228.

[15] C. A. Hurtado, A. Poulovassilis, and P. T. Wood, "Query relaxation in rdf," *Journal on Data Semantics X*, pp. 31–61, 2008.

[16] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, 2007, pp. 697–706.

[17] G. Kasneci, J. V. Gael, D. Stern, and T. Graepel, "Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise," in *WSDM*, 2011, pp. 465–474.

[18] G. Kasneci, J. Van Gael, R. Herbrich, and T. Graepel, "Bayesian knowledge corroboration with logical rules and user feedback," in *ECML*, 2010, pp. 1–18.

[19] G. Cheng, T. Tran, and Y. Qu, "Relin: relatedness and informativeness-based centrality for entity summarization," in *ISWC*, 2011, pp. 114–129.

[20] M. Sydow, M. Pikua, and R. Schenkel, "The notion of diversity in graphical entity summarisation on semantic knowledge graphs," *Journal of Intell. Inf. Sys.*, pp. 1–41, 2013.