An Empirical Evaluation of Techniques for Ranking Semantic Associations

Gong Cheng, *Member, IEEE*, Fei Shao, and Yuzhong Qu

Abstract—Searching for associations between entities is needed in many domains like national security and bioinformatics. In recent years, it has been facilitated by the emergence of graph-structured semantic data on the Web, which offers structured semantic associations more explicit than those hiding in unstructured text for computers to discover. The increasing volume of semantic data often produces excessively many semantic associations, and requires ranking techniques to identify the more important ones for users. Despite the fruitful theoretical research on innovative ranking techniques, there is a lack of comprehensive empirical evaluation of these techniques. In this article, we carry out an extensive evaluation of eight techniques for ranking semantic associations, including two novel ones we propose. The practical effectiveness of these techniques is assessed based on 1,200 ground-truth rankings created by 30 human experts for real-life semantic associations and the explanations given by the experts. Our findings also suggest a number of directions in improving existing techniques and developing novel techniques for future work.

Index Terms—Semantic association, ranking, semantic data, entity homogeneity, relation heterogeneity

1 Introduction

The Web is evolving from a web of documents into a web of data. RDFa, Microdata, Embedded JSON-LD, and Microformats data has been embedded in 38 percent of webpages [33]. In the meantime, thousands of datasets are accessible as Linked Data [1]. This enormous amount of machine-readable, triple-structured *semantic data* has enabled increasingly many intelligent applications. Among others, modern Web search engines have leveraged semantic data to improve the accuracy of search and to enhance their search results with direct and rich answers [13].

For instance, *searching for associations* (aka *relationships*) between a set of two or more entities is a common type of information needs to be satisfied by search engines, and has found applications in many and various domains such as national security [5] and bioinformatics [24]. Early search engines not possessing semantic data have to analyze documents and discover associations hiding in the text [22], and thus inevitably suffer from imprecision and incompleteness. With triple-structured semantic data (e.g., Fig. 1) which can also be represented as an entity-relation graph (e.g., Fig. 2), it would be straightforward to find and return paths and subgraphs that connect user-specified entities, as done by the recent dedicated search engines including RelFinder [16] and Explass [11]. Such graph-structured *semantic associations* (e.g., Fig. 3) are relatively easy to be found, compared with associations hiding in unstructured text.

The increasing volume of semantic data can offer excessively many semantic associations, which are not equally important to users. Research efforts have focused on *ranking techniques*, to show the more important semantic associations earlier. Among others, data-centric techniques rank semantic associations by exploiting various aspects of semantic data [3], [4], [8], [17], [20], [26], [28]; user-centric techniques solicit users' preference [11], [16], [31], [35]. Despite the fruitful theoretical research on innovative ranking techniques, there is a lack of comprehensive empirical evaluation of them. Existing evaluation efforts reported in the literature have the following limitations. First, most techniques [3], [4], [17], [26] are developed to rank semantic associations between exactly two entities, which are usually path-structured. The effectiveness of these techniques for more general graph-structured semantic associations between more than two entities (e.g., Fig. 3) is still open. Second, although some techniques can handle semantic associations between more than two entities, their empirical evaluation is unfortunately limited to a case study [8], [28] or a comparison between only a few techniques [20], possibly because an extensive evaluation is expensive due to the involvement of human experts.

In consideration of the above limitations of existing evaluation efforts, in this work, we carry out an extensive empirical evaluation of eight data-centric techniques for ranking semantic associations, in order to answer the research question *which techniques for ranking semantic associations are practically effective*. Our contribution is threefold.

- We survey existing techniques for ranking semantic associations, and propose two novel techniques based on the heterogeneity or homogeneity of the constituents of a semantic association.
- We evaluate the effectiveness of eight data-centric techniques based on ground-truth rankings created

- *The authors are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.*
E-mail: {gcheng, yzqu}@nju.edu.cn, fshao@smail.nju.edu.cn.

Manuscript received 20 Feb. 2017; revised 22 June 2017; accepted 31 July 2017. Date of publication 4 Aug. 2017; date of current version 4 Oct. 2017.
(Corresponding author: Gong Cheng.)
Recommended for acceptance by M. Sanderson.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2735970

<Websoft, TYPE, Institute> <Qu, name, "Yuzhong Qu"> <Qu, TYPE, Professor> <Qu, worksAt, Websoft> <Qu, knows, Liu> <Liu, TYPE, Engineer>	<Cheng, TYPE, Person> <Cheng, worksAt, Websoft> <Cheng, knows, Liu> <Hu, TYPE, Person> <Hu, worksAt, Websoft> <Hu, knows, Liu>

Fig. 1. A running example of semantic data.

by human experts for real-life semantic associations. The ground-truth rankings are published for reuse.
- We analyze the explanations given by the experts for their rankings, to justify our evaluation results and to inspire future research.

The remainder of this article is organized as follows. Section 2 gives some preliminaries. Section 3 reviews existing techniques for ranking semantic associations and proposes two novel techniques. Section 4 describes the design of our empirical evaluation. Section 5 reports the results of the evaluation. Finally, Section 6 concludes the paper with future work.

2 Preliminaries

Semantic data usually adopts a data model called *entity-property-value triples*, or *triples* for brevity. Fig. 1 shows a set of triples as a running example of semantic data. Entities can be of any types of things, such as `Websoft` which is an institute and `Qu` who is a professor in Fig. 1. Properties are divided into three categories: `TYPE`, attributes, and relations, which have different kinds of values.

- `TYPE` is a property whose values are classes, i.e., entity types, such as `Professor` in Fig. 1.
- Attributes have primitive data values (aka literals) as values, e.g., integers, strings. For instance, `name` in Fig. 1 is an attribute with a string value in quotes.
- Relations have entities as values, such as `worksAt` in Fig. 1.

A set of entity-property-value triples can be represented as a directed graph, of which vertices represent entities and property values, and arcs represent properties. Fig. 2 depicts the triples in Fig. 1 as a graph. Research on semantic association commonly ignores classes and literals. The remaining subgraph induced by entities and relations is called an *entity-relation graph*, as illustrated in Fig. 2.

Given a *query* consisting of a set of *query entities* that a user is interested in, which are vertices of an entity-relation graph, a *semantic association* is a connected subgraph that contains all the query entities. In addition, researchers have placed various structural constraints on such a subgraph, leading to a number of definitions of semantic association that differ slightly from each other. Among others, in [18], a semantic association is a tree-structured subgraph that contains all the query entities. In [10], a semantic association is a minimal subgraph that contains all the query entities and is connected, which implicitly requires a tree structure, and further, requires the leaves to only come from query entities. Fig. 3 illustrates such a semantic association between `Cheng`, `Hu`, and `Qu`, which is a subgraph of the entity-relation graph in Fig. 2. In [8], [20], [28], a semantic association is not necessarily tree-structured, but is constrained to contain all the query entities and a limited number of other entities. In this work, we follow the definition given in [10].

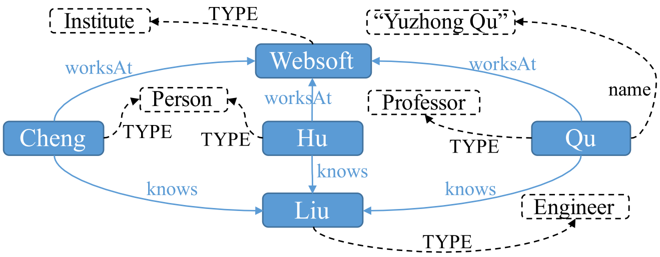


Fig. 2. A graph representation of our running example of semantic data. Blue solid vertices (i.e., entities) and arcs (i.e., relations) induce an entity-relation subgraph.

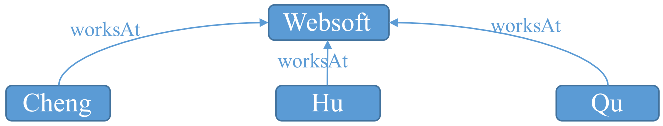


Fig. 3. A semantic association between `Cheng`, `Hu`, and `Qu`.

Formally, we deal with a finite directed unweighted entity-relation graph $G = \langle E, A, R, l \rangle$ where

- E is a set of entities as vertices,
- A is a set of arcs, each arc $a \in A$ directed from its tail vertex $t(a) \in E$ to its head vertex $h(a) \in E$,
- R is a set of relations, and
- $l : A \mapsto R$ labels each arc $a \in A$ with a relation $l(a) \in R$.

Given a query, i.e., a non-empty set of query entities $Q \subseteq E$, a semantic association $x = \langle E_x, A_x \rangle$ is a subgraph of G consisting of vertices E_x and arcs A_x that satisfies the following conditions.

- Its vertices contain all the query entities, i.e., $Q \subseteq E_x$,
- it is connected, and
- it is minimal, i.e., none of its proper subgraphs also satisfies the above two conditions.

The diameter of x, denoted by $diam(x)$, is the greatest distance between any pair of vertices in x. For instance, the diameter of the semantic association in Fig. 3 is 2.

The schema of G is an ontology. Let C be the set of all classes in the ontology, which form a subsumptive hierarchy representing the is-a-subclass-of relationships, as illustrated in Fig. 4. For each entity $e \in E$, let $T(e) \subseteq C$ be e's types.

3 Techniques for Ranking Semantic Associations

In Section 3.1, we will briefly review existing techniques for ranking semantic associations. Then in Section 3.2, we will present a detailed implementation of eight techniques,

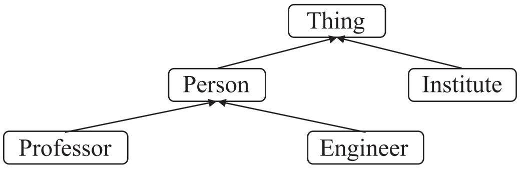


Fig. 4. A running example of class hierarchy.

including two novel techniques proposed in this work, to be evaluated in our empirical study.

3.1 Literature Review

We divide known techniques for ranking semantic associations mainly into two categories: data-centric techniques and user-centric techniques. *Data-centric* techniques analyze various aspects of semantic data. *User-centric* techniques focus on users' preference. In addition, considering that semantic associations may overlap with each other, *diversity-based reranking* can improve the quality of the top-ranked results. Finally, we will discuss the research in *keyword search on graph data*, which also handles the problem of ranking semantic associations.

3.1.1 Data-Centric Techniques

Techniques in the literature have mainly exploited five aspects of semantic data for ranking semantic associations: size, frequency, centrality, informativeness, and specificness.

Size. The size of a semantic association, such as the number of its constituent arcs, is a commonly used feature. A small semantic association usually represents a simple, strong relationship between entities, whereas a large semantic association may reveal a complex, hidden relationship that attracts users in certain applications. Therefore, when ranking semantic associations, it may be inadvisable to give priority consistently to small semantic associations nor consistently to large ones. Instead, in [2], [3], [6], users are invited to explicitly give their personal preference. Alternatively, to void excessive user interaction, users' implicit preference can be captured by machine learning techniques [9].

Frequency. Frequently seen information is usually believed to be important. In [18], the importance of a semantic association is derived from the frequency of occurrence of its constituent entities and relations in semantic data, i.e., the number of triples in which an entity or a relation occurs. A different view is expressed in [17], which regards infrequent things as important due to their exclusivity. Specifically, in a semantic association, a relation between two entities is more important if each of the two entities is connected through this property to fewer other entities. In [3], [9], users are, either explicitly or implicitly, responsible for making decisions on whether to give priority to frequency or infrequency.

Centrality. Graph centrality can be regarded as a generalization of frequency in the context of graphs, and has been used to assess the importance of entities in semantic associations. Among others, degree centrality is simple but effective [3], [6], [9]. The degree of an entity (which is a vertex) in semantic data is the number of arcs incident to it or, in other words, its frequency of occurrence as an endpoint of arcs. A more complex centrality is PageRank [25] used in [8]. PageRank uses a model of a random surfer who continuously walks in the graph, either moving from a vertex to a random neighbor or jumping to a random vertex. The probability that the surfer is located at a vertex after a sufficiently large number of steps is defined as the centrality of the vertex. Interestingly, although PageRank and degree look very different, actually they are strongly correlated with each other [12]. A variant of PageRank is called random walk with restart (RWR) [20], [28], in which the surfer can only jump to query entities. Therefore, RWR is biased towards entities that are close to and densely connected to query entities.

Informativeness. Informativeness measures the amount of information contained in the constituent entities and relations of a semantic association [4], [26]. Based on information theory, the occurrence of a relation in a triple is treated as a probabilistic event. Then, the informativeness of a relation, i.e., the information content associated with the event of its occurrence, is measured by self-information, which is defined as the negative logarithm of the probability of this event. Probability is estimated using relative frequency observed in semantic data. In other words, a relation that occurs less frequently will, once observed, carry a larger amount of information, and thus is more important. This way of measuring informativeness is similar to the notion of rarity defined in [3].

It is worth noting that although frequency and informativeness are opposite measures, they not necessarily conflict with each other and can be jointly used, e.g., giving priority to frequent entities but infrequent relations [3].

Specificness. Specificness provides a different implementation of informativeness, based on not semantic data but its schema, i.e., an ontology. It considers ontological semantics, unlike all the above-mentioned techniques measuring the structure or statistics of data. Specifically, relations in an ontology often form a subsumption hierarchy, and a relation that is deeper in the hierarchy has a more specific meaning and thus carries a larger amount of information [2]. The specificness of the constituent entities of a semantic association can be analogously computed, according to the depth of their types in the class hierarchy [2], [3].

3.1.2 User-Centric Techniques

Some of the above-mentioned data-centric techniques allow users to affect the results of ranking by soliciting their preference about certain aspects of semantic data. More general user-centric techniques in the literature have taken a variety of users' preference into account.

Among others, semantic association search engines compute the relevance of semantic associations to a query submitted by a user, e.g., the degree of the match between query keywords and the relations in semantic associations [4]. When a query is given as a set of entities, the importance of a semantic association can be derived from the closeness between its constituent entities and the query entities. Closeness is measured by the number of short paths [8] or by RWR [20], [28].

In [11], [16], [30], [35], a user can further specify, in addition to a query, some constraints such as a set of intermediate entities, relations, or keywords that should or should not appear in returned semantic associations. In [2], [3], [32], a user can provide a set of interesting entities or relations that are preferred (but not required) to be seen in returned semantic associations. In [31], users' interests are captured by analyzing their Web browsing history.

3.1.3 Diversity-Based Reranking

Top-ranked semantic associations produced by the above-mentioned techniques may not comprise the best results because the information they provide can overlap with each other and thus exhibit redundancy. To improve the diversity of the results, in [26], semantic associations are reranked, only allowing the top-ranked results to have a limited overlap. The overlap between two semantic associations is defined as the Jaccard similarity between their sets of constituent relations. In [27], a framework and a number of algorithms are proposed to directly search for diversified top-ranked results.

3.1.4 Keyword Search on Graph Data

Numerous research efforts have been directed at keyword search on graph data [7], [15], [19], [23], [34]. An answer to a keyword query is a connected subgraph matching all the query keywords, e.g., containing at least one vertex matching each keyword. Such subgraphs and semantic associations are very similar, and the techniques for ranking them are alike. For instance, the size of an answer is considered in [23], [34]. The degree and PageRank centrality of vertices are used in [7] and [34], respectively. The relevance of an answer to a query, i.e., how well the query keywords match the text in the answer, is measured in [23], [34].

3.2 Implementation

We implement eight data-centric techniques for ranking semantic associations, to be evaluated in our empirical study. Six of them are popular techniques reviewed in Section 3.1.1 or their variants, covering all the five aspects of semantic data considered by existing research. In addition, we propose two novel techniques based on the heterogeneity or homogeneity of the constituents of a semantic association. Other techniques, in particular user-centric techniques reviewed in Section 3.1.2, will be evaluated in future work.

Size. All the techniques exploiting the size of a semantic association reviewed in Section 3.1.1 handle exactly two query entities [2], [3], [6], [9]. In that specific case, semantic associations are restricted to paths, and the size of a semantic association is measured by the length of the path. Here, more generally, given two or more query entities, we measure the size of a semantic association x by its diameter

$$Size(x) = diam(x)\,, \tag{1}$$

which is a natural generalization of the length of a path.

Frequency. Following [9], [17], we measure the frequency of occurrence of the constituent relations of a semantic association x. Specifically, for each arc $a \in A_x$ labeled with $l(a) \in R$, the outgoing relative frequency of $l(a)$ is defined as the proportion of arcs labeled with $l(a)$ in the outgoing arcs of $t(a)$ in the entity-relation graph

$$rf_{out}(a) = \frac{|\{a' \in A : t(a') = t(a) \text{ and } l(a') = l(a)\}|}{|\{a' \in A : t(a') = t(a)\}|}\,. \tag{2}$$

The incoming relative frequency of $l(a)$ is defined analogously

$$rf_{in}(a) = \frac{|\{a' \in A : h(a') = h(a) \text{ and } l(a') = l(a)\}|}{|\{a' \in A : h(a') = h(a)\}|}\,. \tag{3}$$

Finally, we calculate the aggregate outgoing and incoming relative frequency of all the constituent relations of x

$$Freq(x) = \frac{1}{|A_x|}\sum_{a \in A_x} \frac{rf_{out}(a) + rf_{in}(a)}{2}\,. \tag{4}$$

Centrality. Among the centrality techniques reviewed in Section 3.1.1, we implement the degree centrality because it is the most widely used in the literature [3], [6], [9] and has proved to be correlated with some other centrality techniques including PageRank [12]. We calculate the average vertex degree of a semantic association x

$$Centr(x) = \frac{1}{|E_x \setminus Q|}\sum_{e \in (E_x \setminus Q)} |\{a \in A : t(a) = e \text{ or } h(a) = e\}|\,, \tag{5}$$

in which query entities are excluded. Including or excluding query entities would not change the relative value of *Centr* because they contribute equally to all the semantic associations; we choose to exclude query entities just in order to prevent their equal contribution from smoothing the results of *Centr*. Note that $Centr(x)$ would be undefined if $E_x = Q$. However, that did not happen in our experiments.

Informativeness. Following [4], [26], we measure the informativeness of the constituent relations of a semantic association x using the technique reviewed in Section 3.1.1. A relation that less frequently occurs would be more informative. Formally, for each arc $a \in A_x$ labeled with $l(a) \in R$, the informativeness of $l(a)$ is measured by its self-information, which is defined as the negative logarithm of its relative frequency of occurrence in the entity-relation graph, divided by its maximum possible value for normalization purposes

$$si_r(a) = \frac{-\log \frac{|\{a' \in A : l(a') = l(a)\}|}{|A|}}{-\log \frac{1}{|A|}}\,. \tag{6}$$

Then we calculate the average informativeness of all the constituent relations of x

$$RInf(x) = \frac{1}{|A_x|}\sum_{a \in A_x} si_r(a)\,. \tag{7}$$

Analogously, we measure the informativeness of the constituent entities of x based on their types. An entity having a type that less frequently occurs would be more informative. Specifically, for each entity $e \in E_x$, the informativeness of e is defined as the maximum of the negative logarithm of the relative frequency of occurrence of its types (i.e., $T(e)$) in the entity-relation graph, divided by its maximum possible value for normalization purposes

$$si_c(e) = \frac{\max_{c \in T(e)} -\log \frac{|\{e' \in E : c \in T(e')\}|}{|E|}}{-\log \frac{1}{|E|}}\,. \tag{8}$$

Then we calculate the average informativeness of all the constituent entities of x, excluding query entities

$$EInf(x) = \frac{1}{|E_x \setminus Q|}\sum_{e \in (E_x \setminus Q)} si_c(e)\,. \tag{9}$$

Specificness. Following [2], [3], we measure the specificness of the constituent entities of a semantic association x using the technique reviewed in Section 3.1.1. An entity having a type that is deeper in the class hierarchy would be more specific. Specifically, for each class $c \in C$, let $depth(c)$ be its depth in the class hierarchy, and let D be the depth of the entire class hierarchy, i.e., the maximum of the depth of all the classes in the hierarchy. For each entity $e \in E_x$, the specificness of e is defined as the maximum of the relative depth of its types (i.e., $T(e)$)

$$rd(e) = \max_{c \in T(e)} \frac{depth(c)}{D} . \tag{10}$$

Then we calculate the average specificness of all the constituent entities of x, excluding query entities

$$Spec(x) = \frac{1}{|E_x \setminus Q|} \sum_{e \in (E_x \setminus Q)} rd(e) . \tag{11}$$

Heterogeneity or Homogeneity. The above-mentioned techniques for ranking semantic associations generally measure some kind of importance for each individual entity or relation in a semantic association, and then aggregate the results by taking the mean value. However, they fail to take the interdependence of entities or relations into account. In this regard, we propose two novel techniques measuring the collective importance of the constituents of a semantic association as a whole. We consider interdependence from the aspect of semantic agreement between constituents, and measure the heterogeneity or homogeneity of their types, i.e., the diversity or uniformity of the types of the constituent relations or entities of a semantic association x. We define the *relation heterogeneity* of x as the proportion of distinct relations in the constituent relations of x

$$RHet(x) = \frac{|\{l(a) : a \in A_x\}|}{|A_x|} . \tag{12}$$

Entities are more complex to handle than relations because an entity can have more than one type. So for each pair of entities in x, we calculate the Jaccard similarity between their types. Then we define the *entity homogeneity* of x as the average similarity of all pairs of the constituent entities of x

$$EHom(x) = \frac{1}{\binom{|E_x|}{2}} \sum_{\substack{e_i, e_j \in E_x \\ e_i \neq e_j}} \frac{|T(e_i) \cap T(e_j)|}{|T(e_i) \cup T(e_j)|} . \tag{13}$$

Conclusion. To summarize, the eight techniques to be evaluated are

- size, i.e., $Size$ defined in Eq. (1),
- (relation) frequency, i.e., $Freq$ defined in Eq. (4),
- (entity) centrality, i.e., $Centr$ defined in Eq. (5),
- relation informativeness, i.e., $RInf$ defined in Eq. (7),
- entity informativeness, i.e., $EInf$ defined in Eq. (9),
- (entity) specificness, i.e., $Spec$ defined in Eq. (11),
- relation heterogeneity, i.e., $RHet$ defined in Eq. (12), and
- entity homogeneity, i.e., $EHom$ defined in Eq. (13).

4 Evaluation Design

To evaluate the eight techniques for ranking semantic associations, we intended to invite human experts to create ground-truth rankings of semantic associations, and then measure the degree of agreement between the ranking generated by each technique and the ground-truth ranking. Considering that ranking a large set of graph-structured semantic associations would be a complex, difficult task for humans, we invited experts to rank only two semantic associations at a time, i.e., to make pairwise comparisons. In particular, to isolate the effects caused by different techniques, we ensured that each pair of semantic associations to be compared received considerably different scores from only one technique but (nearly) equal scores from all the other seven techniques. Then we would be able to evaluate the effectiveness of each separate technique.

All the results of comparisons made by the experts are available at: ws.nju.edu.cn/association/rankeval2017/.

4.1 Dataset

Our evaluation was performed on the DBpedia datasets [21] (version 2015-10), which provided encyclopedic semantic data extracted from Wikipedia and had been widely used in the studies of semantic association search and ranking [10], [11], [16], [17], [18], [26].

Specifically, an entity-relation graph was derived from the Mappingbased Objects dataset, which consisted of high-quality relations between entities extracted from Wikipedia infoboxes. The relation `dbo:type` was excluded because it actually represented the property `TYPE` but not a relation between entities. The remaining entity-relation graph comprised 5,356,354 vertices and 17,494,749 arcs.

The schema of the entity-relation graph, including a class hierarchy, was obtained from the DBpedia Ontology. Entities' types were obtained from the Instance Types dataset, the Instance Types Transitive dataset, and the relation `dbo:type` in the Mappingbased Objects dataset. Each entity had at least one type, namely `owl:Thing`. The average number of types of an entity was 7.23. Some entity types were defined somewhere other than the DBpedia Ontology (e.g., FOAF). They were excluded in our implementation of the technique $Spec$ defined in Eqs. (10) and (11), because otherwise, the depths of classes in different class hierarchies would be on different scales, leading to unfair comparisons.

4.2 Queries

Because we intended to invite human experts to rank semantic associations, we needed to use queries that could result in sufficiently many semantic associations to be ranked, and were fairly familiar to the public to ensure high-quality judgments. To that end, we constructed queries as follows based on QALD-5 [29] and Google's Knowledge Graph [13].

QALD-5, an evaluation campaign on question answering over Linked Data, provided a set of 300 factoid questions about common entities (e.g., Berlin, Michael Jordan), expressed in both natural language and SPARQL [14] for training purposes. From those SPARQL queries, we identified a total of 250 distinct entities that appeared in our entity-relation graph derived from DBpedia, called *seed entities*. For each seed entity, we submitted its name (e.g., Michael Jordan) as a keyword query to the Google search engine. That might trigger Google's Knowledge Graph to return a set of entities that "people also search for" (e.g., Kobe Bryant, LeBron James). If so, we would try to find the

TABLE 1
Statistics About Queries and Query Results

Statistic	Number of entities in a query			Total
	2	3	4	
Number of initial queries	125	125	125	375
Number of queries retrieving at least one semantic association	114	100	79	293
Average number of semantic associations retrieved per query	22,201	316,142	443,373	
Largest number of semantic associations retrieved by a query	502,028	8.59M	5.56M	
Number of queries retrieving 1–500,000 semantic associations	114	92	66	272

corresponding entities in our entity-relation graph, called *related entities*. Finally, for 125 seed entities out of 250, we successfully found at least 4 related entities for each of them. Those seed and related entities were used to construct queries as follows.

Initially, based on each of the 125 seed entities and its 4 or more related entities, we constructed 3 queries, each consisting of 2, 3, or 4 entities including the seed entity and respectively 1, 2, or 3 entities randomly selected from its related entities. For each of the above 375 queries, we leveraged the algorithm proposed in [10] to search our entity-relation graph for semantic associations connecting all the query entities and having a diameter of 4 or smaller; as reported in [10], larger values of diameter would require searching almost the entire entity-relation graph and find too many semantic associations to fit in memory, due to the small-world effect observed on DBpedia.

As shown in Table 1, at the one end, 293 queries out of 375 retrieved at least one semantic association; the remaining queries which retrieved an empty set of semantic associations were excluded from the subsequent experiments. At the other end, some queries retrieved several million semantic associations, which would be too many to handle in our experiments. Therefore, we excluded 21 queries that retrieved more than 500,000 semantic associations. Finally, the remaining 272 queries were used in the subsequent experiments.

4.3 Semantic Associations

The 272 queries retrieved a total of 8.40M semantic associations. Not all of them would be used in the evaluation. Each *qualified pair of semantic associations* to be compared by human experts was required to satisfy three conditions.

- The two semantic associations were retrieved by the same query.
- They received considerably different scores, namely having a difference larger than 20 percent, from one technique out of eight, denoted by $Tech$. The pair was thus called $Tech$-centered, e.g., $Size$-centered or $Freq$-centered; $Tech$ was called the *centric technique* of the pair.
- They received (nearly) equal scores, namely having a difference not larger than 5 percent, from all the other seven techniques.

TABLE 2
Number of Qualified Pairs of Semantic Associations

Centric technique	Number of entities in a query			Total
	2	3	4	
$Size$-centered	165	13,054	11,670	24,889
$Freq$-centered	424.43M	690.81M	218.68M	1,33B
$Centr$-centered	465.74M	474.08M	216.40M	1.16B
$RInf$-centered	90.78M	48.32M	4.38M	143.48M
$EInf$-centered	41.74M	8.81M	11.17M	61.73M
$Spec$-centered	31,676	41,868	105,657	179,201
$RHet$-centered	311.01M	1.47B	1.37B	3.15B
$EHom$-centered	23.33M	30.37M	373.13M	426.83M

As shown in Table 2, for each centric technique, at least 24,889 qualified pairs of semantic associations could be found, providing us with considerably many choices for the evaluation. Besides, they were retrieved by a diversity of queries, as shown in Table 3.

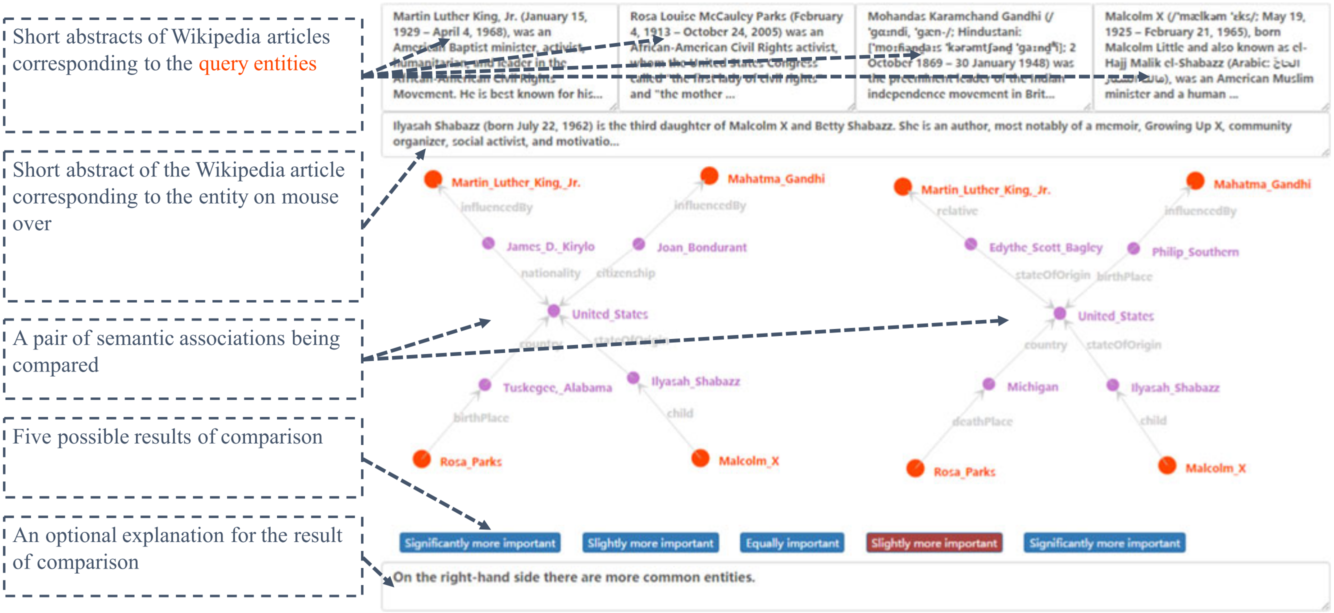
Considering the workload of the human experts that we could afford, a total of 240 qualified pairs of semantic associations were selected and used in the evaluation. Specifically, for each of the 24 combinations of centric technique and number of entities in a query (i.e., 2, 3, or 4), 10 qualified pairs of semantic associations were selected. To maximize the diversity of the semantic associations used in the evaluation, the 10 pairs of semantic associations were required to be retrieved by different, randomly chosen queries. That was possible because, according to Table 3, for every combination of centric technique and number of entities in a query, at least 11 queries contributed qualified pairs of semantic associations. Further, among all the qualified pairs of semantic associations retrieved by a chosen query, the pair of semantic associations having the largest difference in their scores received from their centric technique was selected, in order to maximize the difference characterized by the centric technique so that the difference would be more likely to be identified by the experts.

4.4 Human Experts

We invited 30 students majoring in computer science to make pairwise comparisons between semantic associations. They consisted of 21 male and 9 female students, including 4 PhD students, 21 master students, and 5 undergraduate students. According to their responses to a pre-experiment questionnaire, 11 of them were familiar with semantic data,

TABLE 3
Number of Queries Contributing Qualified Pairs of Semantic Associations

Centric technique	Number of entities in a query			Total
	2	3	4	
$Size$-centered	11	25	16	52
$Freq$-centered	102	81	61	244
$Centr$-centered	97	73	55	225
$RInf$-centered	81	55	37	173
$EInf$-centered	68	40	29	137
$Spec$-centered	29	17	14	60
$RHet$-centered	101	82	63	246
$EHom$-centered	69	52	40	161



Short abstracts of Wikipedia articles corresponding to the query entities

Short abstract of the Wikipedia article corresponding to the entity on mouse over

A pair of semantic associations being compared

Five possible results of comparison

An optional explanation for the result of comparison

Martin Luther King, Jr. (January 15, 1929 – April 4, 1968), was an American Baptist minister, activist, humanitarian, and leader in the African-American Civil Rights Movement. He is best known for his...

Rosa Louise McCauley Parks (February 4, 1913 – October 24, 2005) was an African-American Civil Rights activist, whom the United States Congress called "the first lady of civil rights" and "the mother ...

Mohandas Karamchand Gandhi (/ˈɡɑːndi, ˈɡæn-/; Hindustani: [ˈmoːɦənd̪aːs ˈkərəmtʃənd̪ ˈɡaːnd̪ʱi]; 2 October 1869 – 30 January 1948) was the preeminent leader of the Indian independence movement in Brit...

Malcolm X (/ˈmælkəm ˈɛks/; May 19, 1925 – February 21, 1965), born Malcolm Little and also known as el-Hajj Malik el-Shabazz (Arabic: الحاج مالك الشباز), was an American Muslim minister and a human ...

Ilyasah Shabazz (born July 22, 1962) is the third daughter of Malcolm X and Betty Shabazz. She is an author, most notably of a memoir, Growing Up X, community organizer, social activist, and motivatio...

On the right-hand side there are more common entities.

Fig. 5. A screenshot of the Web-based user interface for comparing a pair of semantic associations and optionally giving an explanation.

13 were not familiar with but were aware of semantic data, and 6 were unfamiliar with semantic data.

4.5 Process

Each of the 240 selected pairs of semantic associations was assigned to 5 human experts, to be independently compared by them; that is, a total of 1,200 pairwise comparisons were made. Each of the 30 experts made 40 comparisons out of 1,200, consisting of 5 pairs for each of the 8 centric techniques. The 40 pairs assigned to each expert were served in random order.

To compare a pair of semantic associations, experts interacted with a Web-based user interface shown in Fig. 5. The two semantic associations to be compared were visualized as node-link diagrams, and the two diagrams were arranged in random order. In each diagram, query entities were in red and the other entities were in purple. The expert could click on an entity to open a new window showing the corresponding DBpedia page, to obtain more information about that entity. To quickly acquaint the expert with each query entity, the short abstract of the corresponding Wikipedia article was directly provided at the top of the user interface. In addition, when the expert hovered the pointer over an entity other than query entities, the short abstract of the corresponding Wikipedia article would also be shown on the user interface. Finally, the expert was required to respond which of the two semantic associations was significantly or slightly more important (which was encouraged), or to respond that they were equally important (which was discouraged); the expert was instructed to consider a semantic association more important if it was believed to be favored by more users of semantic association search. In addition, the expert was encouraged but not mandated to give an explanation for the result of comparison.

4.6 Metrics

We aimed to measure the degree of agreement between the ranking generated by each technique and the ground-truth ranking created by human experts. To that end, in a pair of semantic associations, let x_H be the one receiving a higher score from the centric technique of the pair, and let x_L be the one receiving a lower score. As shown in Table 4, the results of comparisons made by the experts were quantified in the range of $[-1, 1]$ in two modes:

- *fine-grained mode*, which distinguished "slightly more important" from "significantly more important", and
- *coarse-grained mode*, which ignored such difference.

In both modes, on the pairs of semantic associations for a centric technique, if the mean of the quantified results of comparison were close to 1 or -1, the technique would be considered effective. For instance, if the mean obtained for the *Size* technique were close to -1, it would indicate that the experts generally preferred semantic associations having a smaller size. To test the statistical significance of that result, we carried out the one-sample t-test of the null hypothesis that the mean was equal to 0.

5 Evaluation Results

5.1 Agreement between Human Experts

First of all, to assess the agreement between human experts on each pair of semantic associations, we calculated the largest

TABLE 4
Quantification of the Results of Comparison

	Fine-grained mode	Coarse-grained mode
x_H is significantly more important.	1.0	1.0
x_H is sightly more important.	0.5	1.0
x_H and x_L are equally important.	0.0	0.0
x_L is slightly more important.	-0.5	-1.0
x_L is significantly more important.	-1.0	-1.0

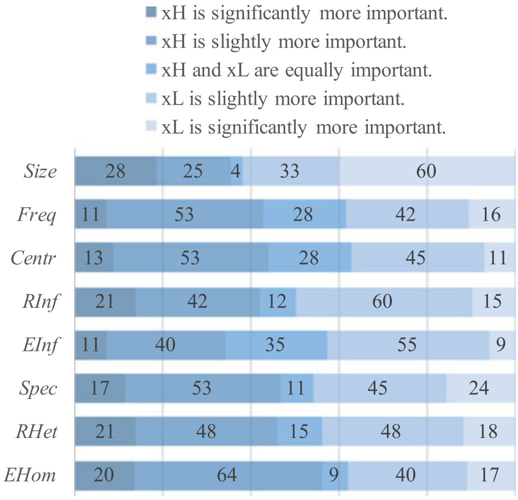


Fig. 6. Overall distribution of the results of comparison.

percentage of the 5 quantified results of comparisons that agreed with each other in the coarse-grained mode. The mean percentage over all the 240 pairs was 68 percent; i.e., out of the 5 human experts, there were an average of 3.4 human experts in the majority having the same opinion, showing a considerable degree of agreement between human experts.

5.2 Overall Results

For each of the eight centric techniques, the distribution of the 150 results of comparisons made on 30 pairs of semantic associations is shown in Fig. 6. The mean and standard deviation of the 150 quantified results are shown in Tables 5 and 6, in the fine-grained mode and in the coarse-grained mode, respectively.

For six centric techniques, namely *Freq*, *Centr*, $RInf$, $EInf$, *Spec*, and $RHet$, the mean was in the range of $[-0.037, 0.040]$ in the fine-grained mode and in the range of $[-0.087, 0.067]$ in the coarse-grained mode, being very close to 0. It showed that human experts expressed largely conflicting views about the relative importance of x_H and x_L. Actually, neither of them was preferred in more than 50 percent of the comparisons. Therefore, those six techniques seemed not *generally* effective indicators of the importance of semantic association considered by the experts.

The mean obtained on *EHom*-centered pairs of semantic associations was 0.100 in the fine-grained mode, being notably large though not significantly different from 0 according to the t-test at the significance level of 0.05. In the coarse-grained mode, the mean increased to 0.180 and was significantly different from 0, indicating that the experts generally preferred semantic associations consisting of entities having relatively similar types (i.e., x_H). Specifically, such semantic associations were preferred in 84 comparisons out of 150 (56 percent).

On *Size*-centered pairs of semantic associations, the mean was -0.240 in the fine-grained mode and -0.267 in the coarse-grained mode, both of which were significantly different from 0, indicating that semantic associations having a relatively small size (i.e., x_L) were generally considered more important by the experts. Specifically, in 93 comparisons out of 150 (62 percent), smaller semantic associations were preferred.

Table 7 shows a breakdown of the Mean values in Table 6 by the number of entities in a query. For *Size* and *EHom*, negative and positive values were consistently obtained under different numbers of entities, respectively, indicating their effectiveness, whereas for the other techniques, a mix of positive and negative values was observed.

To summarize, the above results suggested that, in general, size (*Size*) and entity homogeneity (*EHom*) effectively predicted the importance of semantic association considered by the experts. In particular, small semantic associations consisting of entities having similar types were preferred by the experts in most of the comparisons they made, such as the one illustrated in Fig. 7a compared with the one in Fig. 7b sampled from our data. However, as to the other six techniques, it would be too early to make a general conclusion that they were not effective. It was possible that x_H was consistently preferred by *some* experts, whereas x_L was preferred by the others. For instance, as to

TABLE 5
Overall Quantified Results of Comparison in the Fine-Grained Mode

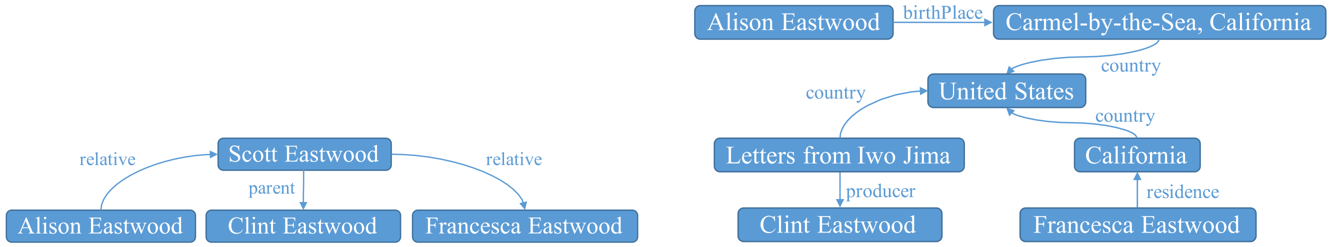
Centric technique	Mean	Standard deviation	t-test (p-value)
Size	**-0.240**	0.794	**0.000**
Freq	0.003	0.584	0.944
Centr	0.040	0.569	0.391
$RInf$	-0.020	0.642	0.703
$EInf$	-0.037	0.541	0.407
Spec	-0.020	0.663	0.712
$RHet$	0.020	0.650	0.707
EHom	**0.100**	0.642	0.059

TABLE 6
Overall Quantified Results of Comparison in the Coarse-Grained Mode

Centric technique	Mean	Standard deviation	t-test (p-value)
Size	**-0.267**	0.953	**0.001**
Freq	0.040	0.904	0.589
Centr	0.067	0.902	0.367
$RInf$	-0.080	0.959	0.309
$EInf$	-0.087	0.874	0.227
Spec	0.007	0.966	0.933
$RHet$	0.020	0.952	0.797
EHom	**0.180**	0.956	**0.022**

TABLE 7
Breakdown of Mean in Table 6 by the Number of Entities in a Query

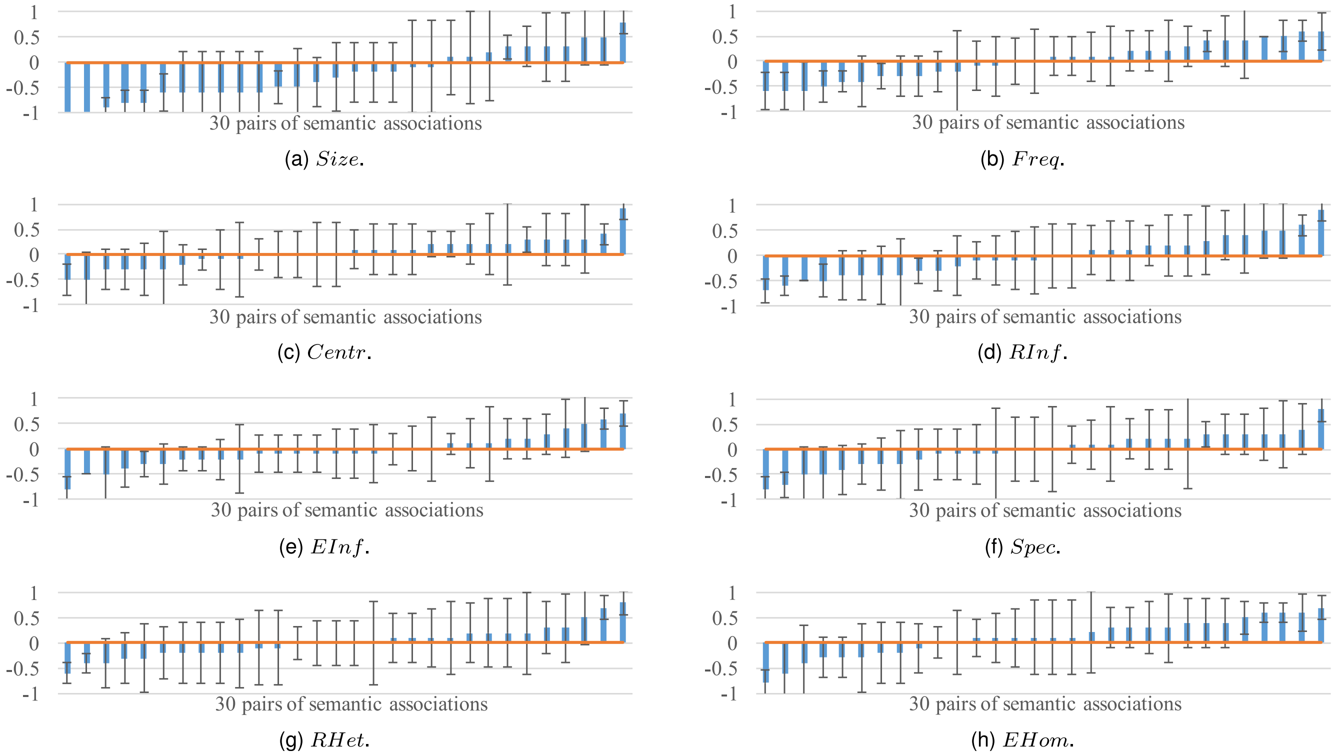
Centric technique	Number of entities in a query		
	2	3	4
Size-centered	-0.56	-0.08	-0.16
Freq-centered	-0.10	0.18	0.04
Centr-centered	0.16	0.20	-0.16
$RInf$-centered	-0.14	-0.10	0.00
$EInf$-centered	-0.12	0.06	-0.20
Spec-centered	0.02	0.08	-0.08
$RHet$-centered	0.06	-0.06	0.06
EHom-centered	0.22	0.22	0.10



(a) A small semantic association consisting of entities having similar types.

(b) A large semantic association consisting of entities having diverse types.

Fig. 7. Two semantic associations between `Alison Eastwood`, `Clint Eastwood`, and `Francesca Eastwood`.



(a) *Size*.

(b) *Freq*.

(c) *Centr*.

(d) *RInf*.

(e) *EInf*.

(f) *Spec*.

(g) *RHet*.

(h) *EHom*.

Fig. 8. Mean (bars) and standard deviation (lines) of the quantified results of comparisons made on each pair of semantic associations in the fine-grained mode.

the *RHet* technique, it was possible that half of the experts believed that semantic associations consisting of diverse relations were important, whereas the other half preferred uniform relations. In that case, the mean of the quantified results of comparison would be close to 0, but still, such a technique was *partially effective* because, as discussed in Section 3, there were strategies (e.g., permitting user customization) for deciding on whether to rank semantic associations in ascending or descending order of their scores received from a technique. Therefore, later in Section 5.4, a more detailed user-wise analysis was performed to identify partially effective techniques.

5.3 Query-Wise Results

To further explain the effectiveness of each technique, a more detailed query-wise analysis is performed to aggregate the quantified results of comparison *by each individual pair of semantic associations*. Specifically, for each of the eight centric techniques, the mean and standard deviation of the 5 quantified results of comparisons made on each of the 30 pairs of semantic associations are shown in Figs. 8a, 8b, 8c, 8d, 8e, 8f, 8g, and 8h, all in the fine-grained mode. The number and percentage of pairs of semantic associations receiving different levels of consistent results of comparison are summarized in Table 8. The results in the coarse-grained mode are similar, so that we only present summarized

TABLE 8
Number and Percentage of Pairs of Semantic Associations Receiving Different Levels of Consistent Results of Comparison in the Fine-Grained Mode

Centric technique	$\lvert\text{Mean}\rvert \geq 0.3$		$\lvert\text{Mean}\rvert \geq 0.5$	
	≤ -0.3	≥ 0.3	≤ -0.5	≥ 0.5
Size	**15 (50%)**	**7 (23%)**	**13 (43%)**	**3 (10%)**
Freq	9 (30%)	8 (27%)	4 (13%)	4 (13%)
Centr	6 (20%)	6 (20%)	2 (7%)	1 (3%)
RInf	10 (33%)	7 (23%)	4 (13%)	4 (13%)
EInf	6 (20%)	5 (17%)	3 (10%)	3 (10%)
Spec	8 (27%)	7 (23%)	4 (13%)	1 (3%)
RHet	5 (17%)	5 (17%)	1 (3%)	3 (10%)
EHom	**6 (20%)**	**12 (40%)**	**2 (7%)**	**5 (17%)**

TABLE 9
Number and Percentage of Pairs of Semantic Associations Receiving Different Levels of Consistent Results of Comparison in the Coarse-Grained Mode

Centric technique	$\lvert\text{Mean}\rvert \geq 0.3$		$\lvert\text{Mean}\rvert \geq 0.5$	
	≤ -0.3	≥ 0.3	≤ -0.5	≥ 0.5
Size	16 (53%)	6 (20%)	16 (53%)	6 (20%)
Freq	10 (33%)	9 (30%)	7 (23%)	8 (27%)
Centr	8 (27%)	9 (30%)	4 (13%)	4 (13%)
RInf	10 (33%)	6 (20%)	9 (30%)	5 (17%)
EInf	10 (33%)	7 (23%)	6 (20%)	4 (13%)
Spec	7 (23%)	7 (23%)	5 (17%)	5 (17%)
RHet	6 (20%)	8 (27%)	4 (13%)	5 (17%)
EHom	5 (17%)	12 (40%)	5 (17%)	11 (37%)

results in Table 9 due to space limitations. In the following, we will focus on the results in the fine-grained mode.

Out of the 30 *Size*-centered pairs of semantic associations, 15 pairs (50 percent) received a mean value below -0.3, being more than twice as many as those receiving a mean value above 0.3; the difference was even larger at the level of -0.5 and 0.5, namely 13 pairs (43 percent) versus 3 pairs (10 percent). In accordance with the conclusion in Section 5.2, it showed again that relatively small semantic associations were generally preferred by the human experts in much more cases. Analogously, out of the 30 *EHom*-centered pairs of semantic associations, 12 pairs (40 percent) received a mean value above 0.3, being twice as many as those receiving a mean value below -0.3, indicating again that semantic associations consisting of entities having relatively similar types were generally preferred in much more cases. The above results further demonstrated the general effectiveness of *Size* and *EHom*.

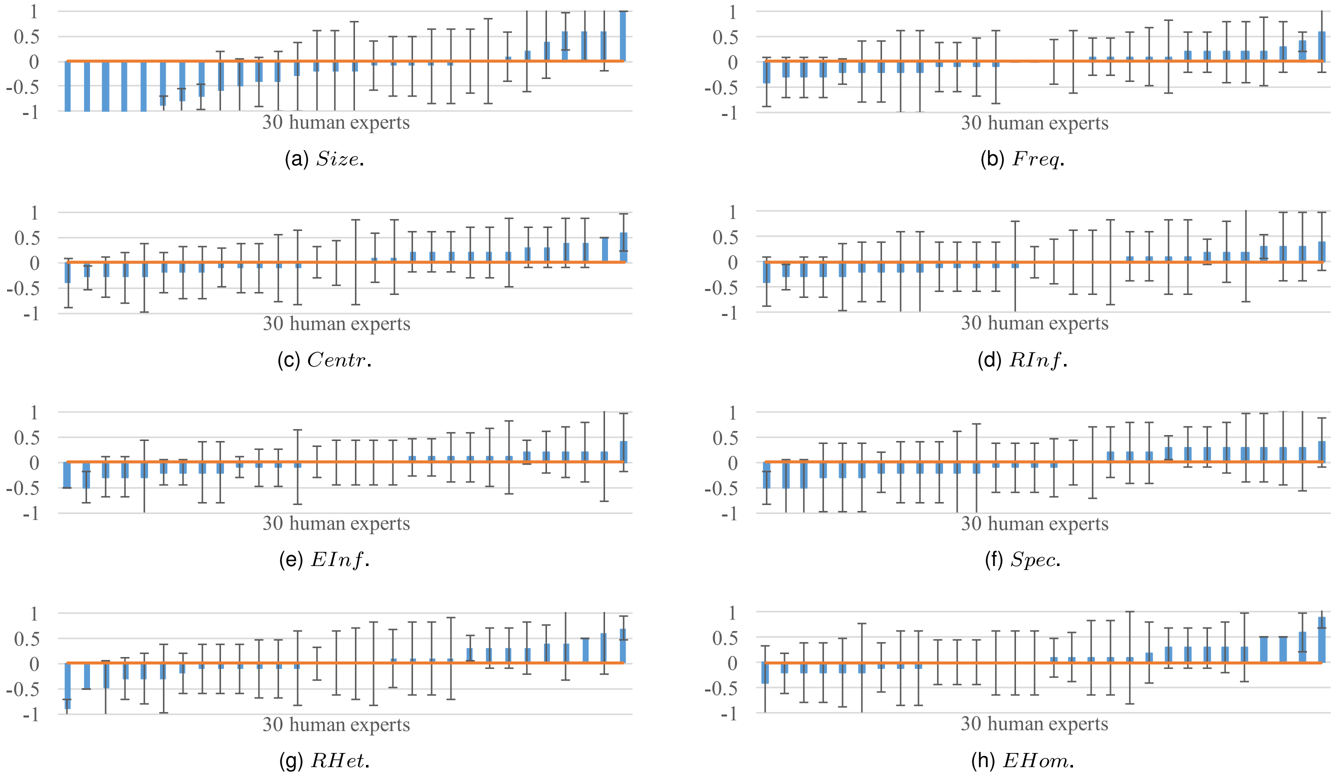
For the other six centric techniques, the number of pairs of semantic associations receiving a mean value below -0.3 was comparable to that above 0.3, further demonstrating that those techniques were not generally effective.

5.4 User-Wise Results

The aim of user-wise analysis is to measure the agreement between the ranking generated by each technique and the ground-truth ranking created *by each individual human expert*. It is possible that although the experts generally have conflicting views about the relative importance of x_H and x_L, each individual expert holds a consistent view in the comparisons made by herself/himself. If so, such a *partially effective* technique would be useful in a number of applications, as discussed at the end of Section 5.2.

For each of the eight centric techniques, the mean and standard deviation of the five quantified results of comparisons made by each of the 30 experts are shown in Figs. 9a, 9b, 9c, 9d, 9e, 9f, 9g, and 9h, all in the fine-grained mode. The number and percentage of experts responding different levels of consistent results of comparison are summarized in Table 10. The results in the coarse-grained mode are similar, so that we only present summarized results in Table 11 due to space limitations. In the following, we will focus on the results in the fine-grained mode.

For three centric techniques, namely *Freq*, *RInf*, and *EInf*, more than two thirds of the experts responded a mean value in the range of $(-0.3, 0.3)$; 93 percent or more of the experts responded a mean value in the range of $(-0.5, 0.5)$. It showed that most experts expressed an ambiguous view about the relative importance of x_H and x_L.



(a) *Size*.

(b) *Freq*.

(c) *Centr*.

(d) *RInf*.

(e) *EInf*.

(f) *Spec*.

(g) *RHet*.

(h) *EHom*.

Fig. 9. Mean (bars) and standard deviation (lines) of the quantified results of comparisons made by each human expert in the fine-grained mode.

TABLE 10
Number and Percentage of Human Experts Responding Different Levels of Consistent Results of Comparison in the Fine-Grained Mode

Centric technique	$\lvert\text{Mean}\rvert \geq 0.3$		$\lvert\text{Mean}\rvert \geq 0.5$		p-value < 0.05
	≤ -0.3	≥ 0.3	≤ -0.5	≥ 0.5	
Size	**13 (43%)**	**5 (17%)**	**10 (33%)**	**4 (13%)**	**10 (33%)**
Freq	4 (13%)	3 (10%)	0 (0%)	1 (3%)	1 (3%)
Centr	**5 (17%)**	**6 (20%)**	0 (0%)	2 (7%)	2 (7%)
RInf	5 (17%)	4 (13%)	0 (0%)	0 (0%)	0 (0%)
EInf	5 (17%)	1 (3%)	2 (7%)	0 (0%)	2 (7%)
Spec	**6 (20%)**	**9 (30%)**	**3 (10%)**	**0 (0%)**	1 (3%)
RHet	**6 (20%)**	**9 (30%)**	**3 (10%)**	**3 (10%)**	**4 (13%)**
EHom	**1 (3%)**	**9 (30%)**	**0 (0%)**	**4 (13%)**	**4 (13%)**

TABLE 11
Number and Percentage of Human Experts Responding Different Levels of Consistent Results of Comparison in the Coarse-Grained Mode

Centric technique	$\lvert\text{Mean}\rvert \geq 0.3$		$\lvert\text{Mean}\rvert \geq 0.5$		p-value < 0.05
	≤ -0.3	≥ 0.3	≤ -0.5	≥ 0.5	
Size	13 (43%)	5 (17%)	11 (37%)	5 (17%)	10 (33%)
Freq	5 (17%)	8 (27%)	3 (10%)	3 (10%)	1 (3%)
Centr	5 (17%)	9 (30%)	4 (13%)	6 (20%)	2 (7%)
RInf	6 (20%)	3 (10%)	6 (20%)	1 (3%)	0 (0%)
EInf	8 (27%)	4 (13%)	5 (17%)	0 (0%)	2 (7%)
Spec	5 (17%)	8 (27%)	4 (13%)	6 (20%)	1 (3%)
RHet	7 (23%)	8 (27%)	5 (17%)	7 (23%)	4 (13%)
EHom	3 (10%)	9 (30%)	1 (3%)	8 (27%)	4 (13%)

Therefore, those three techniques seemed not effective even in respect of individual experts.

For *Size*, *Centr*, *Spec*, *RHet*, and *EHom*, at least one third of the experts responded a mean value above 0.3 or below -0.3. Among them, for *Size*, *Spec*, *RHet*, and *EHom*, at least 10 percent of the experts responded a mean value above 0.5 or below -0.5. In particular, for *Size*, *RHet*, and *EHom*, at least 10 percent of the experts responded a mean value significantly different from 0 according to the t-test at the significance level of 0.05, indicating that a notable portion of experts expressed a consistent view about the relative importance of x_H and x_L in their *own* results of comparison, regardless of the disagreement between them. In Sections 5.2 and 5.3, *Size* and *EHom* were shown to be generally effective. So here, relation heterogeneity (*RHet*) was a newly identified partially effective technique. Specifically, on the one hand, considerable preference to semantic associations consisting of relatively uniform relations (i.e., $\text{Mean} \leq -0.3$) was found on 6 experts (20 percent), including 3 (10 percent) having strong preference (i.e., $\text{Mean} \leq -0.5$); on the other hand, semantic associations consisting of relatively diverse relations were also considerably preferred (i.e., $\text{Mean} \geq 0.3$) by 9 experts (30 percent), including 3 (10 percent) having strong preference (i.e., $\text{Mean} \geq 0.5$) as well.

5.5 User Feedback

Table 12 summarizes the explanations for the results of comparison given by human experts.

The *Size* technique calculated the diameter of a semantic association and characterized its graph structure, which could be directly observed by the experts. On *Size*-centered pairs of semantic associations, 25 experts out of 30 (83 percent) explicitly or implicitly mentioned in at least one explanation that they preferred semantic associations connecting query entities via relatively short paths (i.e., x_L), and none of the experts expressed opposite views. The

TABLE 12
Explanations Given by Human Experts

Centric technique	Samples of explanations supporting x_H	Samples of explanations supporting x_L
Size	Not found.	Query entities are connected via fewer arcs. Query entities are connected via fewer vertices. Query entities are more closely connected. It is more compact. **— from 25 experts (83%)**
Freq and inverse *RInf*	Its constituent relations are more common. Its constituent relations are more often used. It is easier to understand common relations. — from 5 experts (17%)	Its constituent relations are more informative. Its constituent relations are more significant. — from 11 experts (37%)
Centr	Its intermediate entities are more famous. — from 11 experts (37%)	London (in x_H) is too common to interest. Los Angeles (in x_L) is more precise than California (in x_H). — from 9 experts (30%)
EInf	Its intermediate entities are more informative. Its intermediate entities are more specific. — from 2 experts (7%)	Some intermediate entities are rare. Some intermediate entities are not familiar to me. — from 5 experts (17%)
Spec	Its constituent entities are more specific. — from 2 experts (7%)	A location (in x_L) is more important than a specific hospital (in x_H). — from 1 expert (3%)
RHet	It contains more diverse relations. — from 6 experts (20%)	Its constituent relations are the same. It is easier to understand uniform relations. **— from 15 experts (50%)**
EHom	Its constituent entities have the same type. Its intermediate entities have the same type. **— from 17 experts (57%)**	Not found.

diameter of a semantic association exactly constrained the maximum distance between query entities, so that the $Size$ technique was proved generally effective in Section 5.2.

The $EHom$ technique measured the uniformity of the types of the constituent entities of a semantic association, which could also be easily observed by the experts. On $EHom$-centered pairs of semantic associations, 17 experts (57 percent) explicitly supported such uniformity (i.e., x_H) in their explanations, and none of the experts expressed opposite views, so that the $EHom$ technique was proved generally effective in Section 5.2. In addition, a number of experts expressed their preference to semantic associations consisting of similar entities in a more general sense; having the same type was just one kind of similarity. Those experts also favored entities having thematically related types such as "singer" and "genre", or entities sharing a particular property value such as "scientists studying in the same field" or "people living at the same time". That would inspire us to extend our implementation of entity homogeneity in future work, to *support different similarity measures*.

The $RHet$ technique measured the diversity of the types of the constituent relations of a semantic association, which was also an observable feature. On $RHet$-centered pairs of semantic associations, 15 experts (50 percent) explicitly preferred uniform relations (i.e., x_L) in their explanations, whereas six experts (20 percent) favored diverse relations (i.e., x_H) in some of their explanations. Such conflicting views accorded well with the partial effectiveness of $RHet$ shown in Section 5.4. Moreover, 2 experts (7 percent) were found to hold conflicting views in their own explanations; on some pairs they preferred uniform relations but on some other pairs they preferred diverse relations. In future work, It would be interesting to extensively *explore the usage of relation heterogeneity in different contexts*.

Different from the above techniques whose effects were easily observable and were indeed mentioned by the majority of experts, $Centr$, $EInf$, and $Spec$ measured the importance of the constituent entities of a semantic association by calculating their degree centrality, the rarity of their types, and the depth of their types in the class hierarchy, respectively, all of which were unlikely to be directly perceived by the experts and were mentioned by only a minority of experts in their explanations. Specifically, on $Centr$-centered pairs of semantic associations, 11 experts (37 percent) noticed the popularity of entities and favored famous entities which in the meantime had a large degree (i.e., x_H), whereas 9 experts (30 percent) disliked very common or less precise entities despite their large degree. On $EInf$-centered pairs of semantic associations, 2 experts (7 percent) preferred informative and specific entities (i.e., x_H), whereas 5 experts (17 percent) were not willing to see rare or unfamiliar entities which had uncommon types. On $Spec$-centered pairs of semantic associations, the results were similar. To conclude, the experts generally favored *famous but not overly common* and *specific but not overly rare* entities. It explained why $Centr$, $EInf$, and $Spec$ seemed not effective in the previous sections, and revealed the key to making them effective in future work: *finding just the right intermediate levels of centrality, informativeness, and specificness*.

The remaining two techniques, $Freq$ and $RInf$, were opposite measures calculating the frequency and infrequency of the constituent relations of a semantic association, respectively. On $Freq$- and $RInf$-centered pairs of semantic associations, 5 experts (17 percent) preferred common relations which also occurred frequently in the data, whereas 11 experts (37 percent) favored infrequent and thus informative relations in some of their explanations. In the previous sections, $Freq$ and $RInf$ seemed not effective, which was not surprising because the explanations also did not give us clues about the circumstances in which frequency or infrequency should dominate. Therefore, we could not foresee any ways of making them effective in future work.

Besides the above explanations regarding the eight techniques considered in our evaluation, we identified two clusters of explanations that would possibly lead us to novel techniques in future work. Both of them were related to graph structure: 10 experts (33 percent) mentioned in their explanations that they preferred *symmetric graph structures*; 8 experts (27 percent) said they were willing to see *arcs directed in the same direction*. However, it remained to be determined whether such preference was specific to the method of visualizing semantic associations in our evaluation, namely drawing graph-structured semantic associations as node-link diagrams.

6 Conclusions and Future Work

We have conducted an extensive empirical evaluation of eight data-centric techniques for ranking semantic associations, including two novel ones proposed in this work measuring the heterogeneity or homogeneity of the constituents of a semantic association. The results of 1,200 comparisons of 240 pairs of real-life semantic associations made by 30 human experts, as well as the explanations given by the experts, demonstrate the practical effectiveness of two techniques, namely size and entity homogeneity. The experts generally prefer small semantic associations consisting of entities having similar types. Besides, relation heterogeneity is proved partially effective. Semantic associations consisting of uniform relations and those consisting of diverse ones are both preferred by a notable portion of the experts. The practical effectiveness of the other five techniques is open.

We have published all the results of comparisons made by the experts, which can be reused in future research to analyze users' preference and to evaluate novel techniques. However, we remind that our selection of semantic associations is biased towards the techniques we evaluate.

Our findings also suggest several directions in improving existing techniques and developing novel techniques as future work. First, the centrality, informativeness, and specificness of the constituent entities of a semantic association seem not effective in their current forms because the experts generally favor famous but not overly common, and specific but not overly rare entities. Therefore, it is possible to make these techniques effective by finding just the right intermediate levels of centrality, informativeness, and specificness. Second, entity homogeneity can be improved by using other measures of similarity between entities, in addition to the type-based measure used in our implementation. Third, the symmetry of graph structure and the direction of arcs are two factors that have not been considered prior to our evaluation but are mentioned by many experts in their explanations. It would be interesting to determine whether they are generally effective indicators of the importance of semantic association or are only specific to certain methods of visualization like node-link diagrams in our evaluation.

Although our empirical evaluation is one of the most extensive ones to date, we are aware that it has the following limitations. First, the experts make pairwise comparisons so that they can accomplish it with ease, but in the future we will try to solicit list-wise ground-truth rankings, to enable the use of more straightforward metrics for assessing the quality of a ranking. Second, the queries are constructed to contain common entities so that they can result in sufficiently many semantic associations to be ranked and are also familiar to the experts to ensure high-quality judgments. However, it would be interesting to also consider long-tail entities in future work. Third, the focus of this work is an evaluation of individual data-centric techniques, which will be extended in future work to include user-centric techniques as well as combinations of techniques. Fourth, we plan to involve more datasets and more diverse participants in the evaluation.

Acknowledgments

This work was supported in part by the NSFC under Grant 61572247, in part by the 863 Program under Grant 2015AA015406, in part by the NSFC under Grant 61370019, and in part by the Qing Lan Project. The authors would like to thank all the participants in the experiments.

References

[1] A. Abele, J. McCrae, P. Buitelaar, A. Jentzsch, and R. Cyganiak, "Linking open data cloud diagram 2017, " Jan. 2017. [Online]. Available: http://lod-cloud.net/

[2] B. Aleman-Meza, C. Halaschek, I. Arpinar, and A. Sheth, "Context-aware semantic association ranking," in *Proc. 1st Int. Conf. Semantic Web Databases*, Sep. 2003, pp. 24–41.

[3] B. Aleman-Meza, C. Halaschek-Wiener, I. Arpinar, C. Ramakrishnan, and A. Sheth, "Ranking complex relationships on the Semantic Web," *IEEE Internet Comput.*, vol. 9, no. 3, pp. 37–44, May/Jun. 2005, doi: 10.1109/MIC.2005.63.

[4] K. Anyanwu, A. Maduko, and A. Sheth, "SemRank: Ranking complex relationship search results on the Semantic Web," in *Proc. 14th Int. Conf. World Wide Web*, May 2005, pp. 117–127, doi: 10.1145/1060745.1060766.

[5] A. Sheth, et al., "Semantic association identification and knowledge discovery for national security applications," *J. Database Manage.*, vol. 16, no. 1, 2005, Art. no. 21, doi: 10.4018/jdm.2005010103.

[6] P. Barnaghi and S. Abdul Kareem, "A context-aware ranking method for the complex relationships on the Semantic Web," in *Proc. 6th Int. Conf. Adv. Language Process. Web Inf. Technol.*, Aug. 2007, pp. 129–134, doi: 10.1109/ALPIT.2007.70.

[7] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, "Keyword searching and browsing in databases using BANKS," in *Proc. 18th Int. Conf. Data Eng.*, Feb./Mar. 2002, pp. 431–440, doi: 10.1109/ICDE.2002.994756.

[8] C. Chen, G. Wang, H. Liu, J. Xin, and Y. Yuan, "SISP: A new framework for searching the informative subgraph based on PSO," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2011, pp. 453–462, doi: 10.1145/2063576.2063645.

[9] N. Chen and V. Prasanna, "Learning to rank complex semantic relationships," *Int. J. Semantic Web Inf. Syst.*, vol. 8, no. 4, pp. 1–19, Oct. 2012, doi: 10.4018/jswis.2012100101.

[10] G. Cheng, D. Liu, and Y. Qu, "Efficient algorithms for association finding and frequent association pattern mining," in *Proc. 15th Int. Semantic Web Conf.*, Oct. 2016, pp. 119–134, doi: 10.1007/978–3-319-46523-4_8.

[11] G. Cheng, Y. Zhang, and Y. Qu, "Explass: Exploring associations between entities via top-K ontological patterns and facets," in *Proc. 13th Int. Semantic Web Conf.*, Oct. 2014, pp. 422–437, doi: 10.1007/978–3-319-11915-1_27.

[12] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, "Approximating PageRank from in-degree," in *Proc. 4th Int. Workshop Algorithms Models Web-Graph*, Nov./Dec. 2006, pp. 59–71, doi: 10.1007/978–3-540-78808-9_6.

[13] Google Official Blog, "Introducing the knowledge graph: Things, not strings," May 2012. [Online]. Available: http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

[14] S. Harris and A. Seaborne, "SPARQL 1.1 query language," Mar. 2013. [Online]. Available: http://www.w3.org/TR/sparql11-query/

[15] H. He, H. Wang, J. Yang, and P. Yu, "BLINKS: Ranked keyword searches on graphs," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2007, pp. 305–316, doi: 10.1145/1247480.1247516.

[16] P. Heim, S. Lohmann, and T. Stegemann, "Interactive relationship discovery via the Semantic Web," in *Proc. 7th Extended Semantic Web Conf.*, May/Jun. 2010, pp. 303–317, doi: 10.1007/978–3-642-13486-9_21.

[17] I. Hulpuş, N. Prangnawarat, and C. Hayes, "Path-based semantic relatedness on linked data and its use to word and entity disambiguation," in *Proc. 14th Int. Semantic Web Conf.*, Oct. 2015, pp. 442–457, doi: 10.1007/978–3-319-25007-6_26.

[18] X. Jiang, X. Zhang, W. Gui, F. Gao, P. Wang, and F. Zhou, "Summarizing semantic associations based on focused association graph," in *Proc. 8th Int. Conf. Adv. Data Mining Appl.*, Dec. 2012, pp. 564–576, doi: 10.1007/978–3-642-35527-1_47.

[19] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar, "Bidirectional expansion for keyword search on graph databases," in *Proc. 31st Int. Conf. Very Large Data Bases*, Aug./Sep. 2005, pp. 505–516.

[20] G. Kasneci, S. Elbassuoni, and G. Weikum, "MING: Mining informative entity relationship subgraphs," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, Nov. 2009, pp. 1653–1656, doi: 10.1145/1645953.1646196.

[21] J. Lehmann, et al., "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web J.*, vol. 6, no. 2, pp. 167–195, 2015, doi: 10.3233/SW-140134.

[22] G. Luo, C. Tang, and Y.-l. Tian, "Answering relationship queries on the Web," in *Proc. 16th Int. Conf. World Wide Web*, May 2007, pp. 561–570, doi: 10.1145/1242572.1242648.

[23] Y. Luo, W. Wang, X. Lin, X. Zhou, J. Wang, and K. Li, "SPARK2: Top-k keyword query in relational databases," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 12, pp. 1763–1780, Dec. 2011, doi: 10.1109/TKDE.2011.60.

[24] Y. Makita, et al., "PosMed: Ranking genes and bioresources based on Semantic Web association study," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W109–W114, Jun. 2013, doi: 10.1093/nar/gkt474.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 422, Nov. 1999.

[26] G. Pirrò, "Explaining and suggesting relatedness in knowledge graphs," in *Proc. 14th Int. Semantic Web Conf.*, Oct. 2015, pp. 622–639, doi: 10.1007/978–3-319-25007-6_36.

[27] L. Qin, J. Yu, and L. Chang, "Diversifying top-K results," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1124–1135, Jul. 2012, doi: 10.14778/2350229.2350233.

[28] H. Tong and C. Faloutsos, "Center-piece subgraphs: Problem definition and fast solutions," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2006, pp. 404–413, doi: 10.1145/1150402.1150448.

[29] C. Unger, et al., "Question answering over linked data (QALD-5)," in *Proc. 6th Conf. Labs Eval. Forum*, Sep. 2015.

[30] V. Viswanathan and K. Ilango, "Finding relevant semantic association paths through user-specific intermediate entities," *Human-Centric Comput. Inf. Sci.*, vol. 2, no. 1, Dec. 2012, Art. no. 9, doi: 10.1186/2192–1962-2-9.

[31] V. Viswanathan and K. Ilango, "Ranking semantic relationships between two entities using personalization in context specification," *Inf. Sci.*, vol. 207, pp. 35–49, Nov. 2012, doi: 10.1016/j.ins.2012.04.024.

[32] V. Viswanathan and K. Ilango, "Ranking semantic associations between two entities—extended model," in *Proc. 4th Asian Conf. Intell. Inf. Database Syst.*, Mar. 2012, pp. 152–162, doi: 10.1007/978–3-642-28493-9_17.

[33] Web Data Commons, "RDFa, Microdata, Embedded JSON-LD, and Microformats Data Sets - October 2016," Jan. 2017. [Online]. Available: http://webdatacommons.org/structureddata/2016-10/stats/stats.html

[34] M. Yang, B. Ding, S. Chaudhuri, and K. Chakrabarti, "Finding patterns in a knowledge base using keywords to compose table answers," *Proc. VLDB Endowment*, vol. 7, no. 14, pp. 1809–1820, Oct. 2014, doi: 10.14778/2733085.2733088.

[35] M. Zhou, Y. Pan, and Y. Wu, "Conkar: Constraint keyword-based association discovery," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2011, pp. 2553–2556, doi: 10.1145/2063576.2064017.

Gong Cheng received the PhD degree in computer software and theory from Southeast University, in 2010. He is an associate professor in the Department of Computer Science and Technology, Nanjing University. His current research interests include semantic search, data summarization, and question answering. He has published more than 20 papers in major venues in these areas such as WWW, IJCAI, and ISWC. He is a member of the IEEE.

Fei Shao received the BS degree in computer science and technology from Nanjing University, in 2016. She is working toward the master's degree in the Department of Computer Science and Technology, Nanjing University. Her current research interests include semantic search and question answering.

Yuzhong Qu received the BS degree in mathematics from Fudan University, in 1985, the MS degree in mathematics from Fudan University, in 1988, and the PhD degree in computer software from Nanjing University, in 1995. He is a professor in the Department of Computer Science and Technology, Nanjing University. His research interests include Semantic Web, question answering, and novel software technology for the Web. He has published more than 80 papers in major venues in these areas such as WWW, ISWC, and the *Journal of Web Semantics*.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.