Ranking Semantic Associations between Two Entities – Extended Model

V. Viswanathan[1] and Ilango Krishnamurthi[2]

[1] Department of Computer Applications
[2] Department of Computer Science and Engineering
Sri Krishna College of Engineering and Technology, Coimbatore-641008,Tamil Nadu, India
{visuskcet,ilango.krishnamurthi}@gmail.com

Abstract. Semantic association is a set of relationships between two entities in knowledge base represented as graph paths consisting of a sequence of links. The number of relationships between entities in a knowledge base might be much greater than the number of entities. So, ranking the relationship paths is required to find the relevant relationships with respect to the user's domain of interest. In some situations, user may expect the semantic relationships with respect to specific domain closer to any one of these entities. Consider the example for finding the semantic association between the person X and person Y. If the user has already known something about the person X such as person X may be associated with financial activities or scientific research etc., then the user wants to focus on finding and ranking the relationship between two persons in which the users' context is closer to person X. In many of the existing systems, there is no consideration given into context closeness during ranking process. In this paper, we present an approach which allows the extraction of semantic associations between two entities depending on the choice of the user in which the context is closer to left or right entity. The average correlation coefficient between proposed ranking and human ranking is 0.70. We compare the results of our proposed method with other existing methods. It explains that the proposed ranking is highly correlated with human ranking. According to our experiments, the proposed system provides the highest precision rate in ranking the semantic association paths.

Keywords: Semantic Web, Semantic Association, Complex relationship, RDF, RDF Schema.

1 Introduction

Information retrieval over semantic metadata has received a great amount of interest in both industry and academia. The semantic web contains not only resources but also includes the heterogeneous relationships among them. In the current generation technologies of search engine, it is very difficult to find the relationships between entities. For example, 'how two entities are related?' is the most crucial question. Discovering relevant sequences of relationships between two entities answers this question. Semantic association represents a direct or indirect relationship between two

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 152–162, 2012.

entities. Different entities may be related in multiple ways. For example, finding the semantic association between two persons 'X' and 'Y' who belong to film industry and they are involved in financial activities and Politics. There may be multiple paths between two entities that involve more intermediate entities that cover multiple domains. Discovering and ranking such relations based on user's interest is required. To combat this problem, Anyanwu,K et al., propose to rank semantic association [3] using six types of metrics called Subsumption (how much meaning a semantic association conveys depending on the places of its components in the RDF), Path Length (that allows preference of either immediate or distant relationships), Popularity (number of incoming and outgoing edges), Rarity (rarely occurring entity), Trust (determining how reliable a relationship is according to its origin) and context weight. In this method, path weights are calculated using these parameter values and then ranked according to their total weights. Semantic Association path with more weight is ranked first.

Suppose user knows some information about person 'X' or person 'Y' such as person 'X' involved in more financial activities or person 'Y' involved in politics etc. In such cases, user may be interested in finding that types of relationships between 'X' and 'Y' and it should be ranked first.

Consider the example for the relationships between person '**John**' and movie **'Slumdog_millionaire**' in an RDF.

John *– edit_music* ***- Human_Movie*** *– support_fund -* ***Tara_Funding_Agencies*** *– support_fund -* ***Ragam_Music*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***
John *– member_of* ***- Tara_Funding_Agency*** *- supports_fund -* ***Human_Movie*** *– associated_with -* ***Ragam_Music*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***
John *– edits_music –* ***Human_Movie*** *– associated_with -* ***Ragam_Music*** *– support_fund -****Tara_Funding_Agencies*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***

From the above example, the same set of components(properties and entities) may be scattered over the path in different possible combination. If the user is interested in Music and Finance, Anyanwu,K et al., method produce the same weights for all paths. In this method, while calculating each metric, the component values are calculated independently and then summed up. Sometimes, all the paths are having equal weights are ranked arbitrarily. In that case, user has to go through this subset of paths to find the relevant paths. Suppose user may be interested in finding and ranking relevant association according to his domain of interest which is closer to either left entity (John) or right entity (Slumdog millionaire), we have to rank these paths according to user's interest. In the existing systems, users may select the choice for favoring long path or favoring short path, favoring popularity or favoring unpopular or favoring rarity, but there are no ways to select the choice for context closeness. In the proposed method, we find and rank the semantic association paths between two entities according to the users' needs with context closeness.

The organization of this paper is as follows: In Section 2, we present an overview of the background and basic definitions of semantic association. In Section 3, an overview of some related works in the area of semantic association is given. Section 4, explains how the semantic association paths weight is evaluated and ranked. Experimental evaluation of the proposed approach is explained in Section 5. Section 6, summarizes the contribution and states the possible future work.

2 Background

The Resource Description Framework (RDF)[9][10] data model provides a framework to capture the meaning of an entity by specifying how it relates to other entities. In RDF model, concepts of entities are linked together with relations (properties). The properties are denoted by arcs and labeled with the relation name. The definition of the RDF graph is as follows:

Definition 1(RDF graph): RDF graph is a directed labeled graph to represent the relationship between entities.

Semantic associations are complex relationships between resource entities [4]. Most of the useful semantic associations involve some intermediate entities and relations. It helps the user to see the connection between different people, places and events. Semantic associations are based on concepts such as semantic connectivity and semantic similarity. To describe the semantic connection between two entities in domain RDF, we introduce some definition[3]:

Definition 2(Semantic Connectivity): Two entities e_1 and e_n are semantically connected if there exists a sequence e_1, P_1, e_2,P_2, e_3,$P_{3.....}$ e_{n-1},P_{n-1}, e_n in an RDF graph where e_i ($1\leq i \leq n$) are entities and P_j ($1\leq j< n$) are properties.

Definition 3 (Semantic Similarity): Two entities e_1 and f_1 are semantically similar if there exist two semantic paths e_1,P_1, e_2,$P_{2,}$ e_3,$P_{3.....}$ e_{n-1},$P_{n-1,}$ e_n and f_1,Q_1, f_2,$Q_{2,}$ f_3,$Q_{3.....}$ f_{n-1},$Q_{n-1,}$ f_n semantically connecting e_1 with e_n and f_1 with f_n respectively, and that for every pair of properties P_i and Q_i, $1\leq i<n$, either of the following conditions holds: $P_i = Q_i$ or $P_i \subseteq Qi$ or $Qi \subseteq Pi$ ($\subseteq$ means rdf:subPropertyOf), then two paths originating at e_1 and f_1, respectively, are semantically similar.

Definition 4 (Semantic Association): Two entities e_x and e_y are semantically associated if e_x and e_y are semantically connected or semantically similar.

3 Related Work

Several techniques have been proposed related to ranking of semantic associations. Some of them are summarized below:

For ranking the results of complex relationship searches on the semantic web, Anyanwu et al.[3] present a flexible approach called SemRank. In this method, with the help of a sliding bar user can easily vary their search mode from conventional search mode to discovery search mode.

Shahdad Shariatmadari et al.[14] present a method for finding semantic association based on the concept semantic similarity. ρ-operator [4] is used for discovering semantic similarities and graph similarity approach [15] is used to rank the similarity. The similarity between two paths will be calculated based on the degree of similarity of the nodes and edges using subsumption function proposed by Aleman-Meza [2]. The ranking approach proposed by Anyanwu,K. et.al [4] considers 'context' based on value assignments for different ontologies.

Aleman-Meza et al.[2] discuss a framework that uses ranking techniques to identify more interesting and more relevant semantic associations and define a ranking formula that considers Subsumption Weight S_P (how much meaning a semantic association conveys depending on the places of its components in the RDF), Path Length Weight L_P (that allows preference of either immediate or distant relationships), Popularity P_P (number of incoming and outgoing edges), Rarity R_P (rarely occurring entity) and Trust Weight T_P (determining how reliable a relationship is according to its origin) and context weight, for assessing the effectiveness of the ranking scheme. In this method '*user defined weight*' is assigned for each 'context regions' specified by the user and it is used to calculate the context weight. The ranking results depend on the criteria defined by the user.

Lee, M et al.[11] propose a semantic search methodology for measuring the information content of a semantic association that consists of resources and properties based on information theory and expanding the semantic network based on spreading activation. In this method, they provide search results that are connected and ordered relations between search keyword and other resources as link of relation on semantic network.

Dong, X., et al.[7] present a prototype system called Chem2Bio2RDF Dashboard for automatic collecting semantic association within the systems chemical biology space and apply a series of ranking metrics called Quality, Specificity and Distinctiveness to select the most relevant association.

To discover semantic associations between linked data, Vidal,M et al.[17] propose an authority-flow based ranking technique that is able to assign high scores to terms that correspond to potential discoveries, and to efficiently identify these highly scored terms. They also propose an approximate solution named graph-sampling. This technique samples events in a Bayesian network that models the topology of the data connections and it estimates ranking score that measure how important and relevant are the associations between two terms.

Major difference between our approach and other existing methods is that there is no facility provided to find the semantic association paths with user interested components are closer to either left or right entity. In our method, we have considered this feature in selecting choice for user's context which may be closer to either left entity or right entity. So, when paths are to be ranked, according to the users' choice, the discovery process finds semantic association paths with more context weighted entities are closer to either left or right entity is considered as highly relevant in ranking, while others are ranked lower.

4 Calculating Context Weight

Context Weight is one of the semantic metrics which is used to determine the relevancy based on a user specific view. Consider the scenario in which someone is interested in discovering how two persons are related to each other in the domain of “Funding Company”. Concepts such as “Finance” or “Financial organization” would be most relevant, whereas something like “Music Company” would be less meaningful. So it is possible to capture a user’s interest through a Context Specification through user interface screen. Thus, using the context specified, it is possible to rank a path according to its relevance with a user’s domain of interest. Fig.1 illustrates various paths containing different domain entities in the RDF. The top most path (call it Path1) contains one Financial entity and one Music entity. The next path (call it Path 2) contains one Financial entity and other one not in the any domain category. The third path (call it Path 3) contain two Music entities. We assume that there are two users, user1 is more interested in Music domain and user2 is more interested in Financial domain. The expected ranking of these three paths for the user1 would be Path 3, Path 1, and Path 2 and for the user2 would be Path 2, Path1, and Path3.

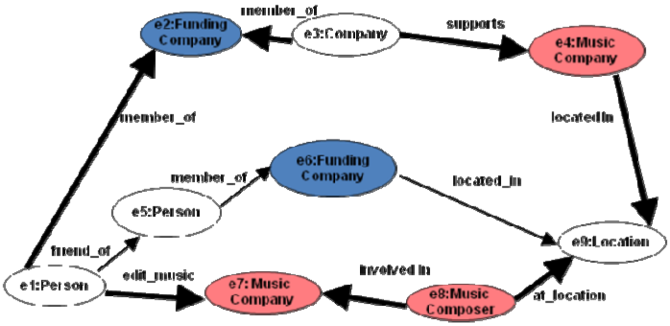

Fig. 1. Paths between two entities person and location in the RDF

From the Fig.1, paths may pass through more than one domain specified by the user. So the component (entities and properties) of the path passing through the region is multiplied by the corresponding weight of the region. Some of the components may not pass through any of the user specified region. These components may be irrelevant to the user. So we have to exclude the component for calculating context weight of the path. Given this background Aleman-Meza et.al[2] defined a formula to calculate the context weight of a given path P as follows:

$$C_p = \frac{1}{|c|}\left(\left(\sum_{i=1}^{region}\left(r_i\left(\sum c \in R_i\right)\right)\right) X \left(1 - \frac{\#c \notin R}{|c|}\right)\right) \qquad (1)$$

Here r_i is the user assigned weight of the region R_i and |c| is the total number of components in the path (excluding the start and end entities). To best of our knowledge Aleman-Meza et.al[2] proposed a method for ranking the semantic association using various criteria such as *Path length, Subsumption, Context, Popularity, Rarity and Trust* to get the relevancy.

Suppose the sub graph in an RDF contains paths between two entities and all the paths having same number of intermediate entities and are scattered in different position. In such case, Aleman-Meza et.al[2] method rank these paths arbitrarily, because of same context values. So, we have to rank this subset according to the users' interest. This can be achieved by using context closeness in ranking. We have defined the context closeness as follows:

Definition 5 (Context closeness): There exists a sequence e_1,P_1,e_2,P_2,e_3,P_3..... e_{n-1}, P_{n-1},e_n in an RDF graph where e_i ($1\leq i \leq n$) are entities and P_j ($1\leq j< n$) are properties, and sum of context weight of user interested entities in the first half of the sequence is greater than second half of the sequence, we can say that the context is closer to left entity e_1; otherwise, context is closer to right entity e_n.

To calculate the *context value* based on the choice of the user selection in context closer, the formula (1) has been modified as follows:

$$C_i = cv_i \sum c \in D_i \tag{2}$$

$$MC_p = \frac{1}{|c|}\left(\sum_{i=1}^{|c|/2} C_i + 0.1 \sum_{(|c|/2)+1}^{n} C_i\right) X\left(1 - \frac{\#c \notin D}{|c|}\right) \tag{3}$$

$$MC_p = \frac{1}{|c|}\left(0.1\sum_{i=1}^{|c|/2} C_i + \sum_{(|c|/2)+1}^{n} C_i\right) X\left(1 - \frac{\#c \notin D}{|c|}\right) \tag{4}$$

Where |c| is the total no of components in the path (excluding the start and end entities). The formula (3) and (4) are used to calculate the context weights closer to left entity and right entity respectively. The context weight MC_p is used as a parameter to calculate the weight of the semantic association paths.

4.1 Ranking the Semantic Association Paths

Our approach defines a path rank as a function of various intermediate weights. These are described as follows:

***Subsumption weight* Sp:** In RDF, entities that are in the lower hierarchy can be considered to be more specialized instances of those further up in the hierarchy. Thus, lower entities have more specific meaning. So, high relevance can be assigned based on subsumption.

Path Length Weight L_P**:** In some queries, a user may be interested in the shortest paths. This may infer a strong relationship between two entities. In some cases a user may wish to find indirect or longer paths. Hence, the user can determine which association length influence.

Popularity Weight P_P**:** The number of incoming and outgoing relationships of entities called popular entities. Path contains highly popular entities may be more relevant. Hence, the user has to select either 'favor more popular associations' or 'favor less popular associations' based on their need.

Rarity Weight R_P**:** Sometimes rarely occurring events are considered to be more interesting than commonly occurring ones. Depending on the requirements, user has to select 'favor rare associations' or 'favor more common associations.[1][12].

Trust Weight T_P**:** Various entities and their relationships in a semantic association originate from different sources. Some of these sources may be more trusted than others. Thus, trust values need to be assigned to the meta-data extracted depending on its source.

We calculate Subsumption Weight S_P, Path Length Weight L_P, Popularity P_P, Rarity R_P and Trust Weight T_P [1] along with user-specific context weight MC_P. These weights are used to determine the path relevancy. So, all the intermediate weights are added to calculate the rank of each path.
Overall association Rank is calculated using the criteria as

$$W_p = k_1 \, x \, S_p + k_2 \, x \, L_p + k_3 \, x \, MC_p + k_4 \, x \, T_p + k_5 \, x \, P_p + k_6 \, x \, R_p \qquad (5)$$

where $k_i(1 \le i \le 6)$ are preference weights and $\sum k_i = 1$. The resulting paths are ranked based on the users' domain of interest. Depending on the requirements, users can also change the preference weights to fine-tune the ranking criteria. In our experiments, we have given high weights to context component and use the other ranking components as secondary criteria.

5 Experimental Evaluation

For finding semantic association paths, we have used an RDF consisting of 52 classes, 70 properties and 3000 entities covering various domains such as Music, Finance, Terrorism and Sports etc. To test the performance of our system, we have selected 40 pairs of entities in the RDF. Semantic association paths has been generated and ranked under the various criteria such as favor short association or favor long association, favor popularity entities or favor unpopular entities, favor rarity, context closer to right entity or context closer to left entity. Criteria have been selected through user interface. Semantic association paths ranking has been done by the above users through the system as well as manually.

5.1 Preliminary Results

To demonstrate our ranking scheme's effectiveness, Fig. 2 shows comparison of human and proposed system ranking results between the entity sets (**Entity1:** John and **Entity2:** Slumdog millionaire). Here 'John' is entity under the class 'Music director' and 'Slumdog millionaire' is the entity under the class 'Movie'. The x-axis represents semantic associations rank first, second and so on according to the proposed system results. The y-axis represents user-human ranking which is assigned manually by the users. We used the Spearman's footrule [6] distance as the measure of similarity between proposed system ranking and user-human ranking using the formula given below:

Spearman's Foot rule distance

$$D_{(system,human)} = \sum_{i=1}^{n} \left| R_{i_{system}} - R_{i_{human}} \right| \tag{6}$$

$$\text{Spearman's Foot rule Coefficient } C = 1 - \frac{4D}{n^2} \tag{7}$$

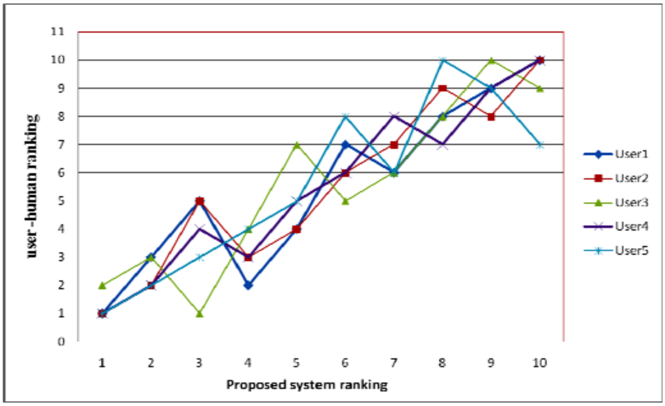

Fig. 2. Comparison of human and proposed system ranking with top-k results

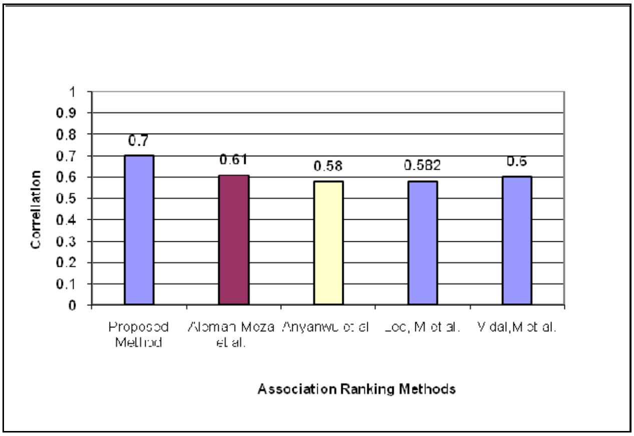

Fig. 3. Comparison of correlation

According to our experiments, the average correlation coefficients between proposed system ranking and user-human‘s ranking is 0.70. . Since the average correlation coefficient is greater than 0.5, the proposed system’s ranking and user-human ranking are highly correlated. Fig. 3 shows the comparison of Correlation between human rankings with proposed system and other existing association ranking methods. It explains that correlation between human ranking and our proposed approach is higher than other existing methods. We have evaluated the precision rate from the top-k semantic association paths from the ranked results for the proposed system and other existing methods. Precision represents the fraction of the relevant paths from top-k semantic association paths. Fig. 4 shows the comparison of precision rate of proposed method with existing methods. Irrespective of ‘k’ value, precision rate will increase or decrease. Among the five methods which show the same phenomenon. But,the method which we have adapted is more significant and provides high precision rate.

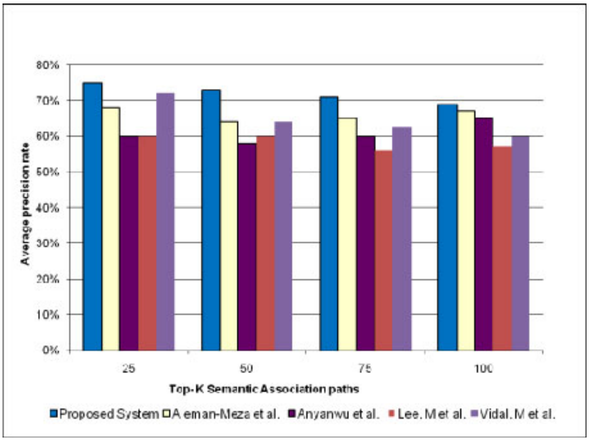

Fig. 4. Comparison of precision rate

6 Conclusion

Semantic data contains entities and heterogeneous relationships among them. The number of relationships between entities might be much greater than the number of entities. Ranking these relationship paths are required to find the relevant relationships between entities with respect to user’s domain of interest. Sometimes users’ may expect the relationships between two entities in which his/her context is closer to any one of the end points either left entity or right entity. The proposed method, find and rank the semantic association paths with respect to specific domain components which is closer to either left entity or right entity. We compare our proposed method with existing methods through spearman correlation coefficient and precision rate. The average correlation coefficient between proposed system ranking, Aleman-Meza et al., Anyanwu et al., Lee,M. and Vidal,M., with human ranking are 0.70, 0.61,0.58,0.582 and 0.6 respectively. It explains that our proposed system ranking is highly correlated with human ranking. According to our experiments as measure of precision rate, we can conclude that our proposed system achieves high precision rate with top-k ranking than others. In future, we plan to generate the semantic web usage ontology from web usage information of each user and which may be used to get personalized semantic associations ranking.

References

1. Anderson, R., Khattak, A.: The Use of Information Retrieval Techniques for Intrusion Detection. In: Proc. 1st Int'l Workshop Recent Advances in Intrusion Detection (1998)
2. Aleman-Meza, B., Halaschek, C., Arpinar, I.B., Sheth, A.: Ranking Complex Relationships on the Semantic Web. IEEE Internet Computing 9(3), 37–44 (2005)
3. Anyanwu, K., Maduko, A., Sheth, A.: SemRank: Ranking Complex Relationship Search Results on the Semantic Web. In: Proc. of the 14th Int'l World Wide Web Conference, pp. 117–127. ACM Press (2005)
4. Anyanwu, K., Sheth, A.: ρ-Queries: Enabling Querying for Semantic Associations on the Semantic web. In: Proc. of the 12th Int'l World Wide Web Conference, pp. 690–699 (2003)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 285(5), 34–43 (2001)
6. Diaconis, P., Graham, R.: Spearman's Footrule as a Measure of Disarray. J. Royal Statistical Soc. Series B 39(2), 262–268 (1977)
7. Dong, X., Ding, Y., Wang, H., Chen, B., Wild, D.J.: Ranking semantic associations in systems chemical biology space. In: 19th Int'l World Wide Web Conference, FWCS 2010, Raleigh, NC (2010)
8. Jiang, X., Tan, A.: Learning and inferencing in user ontology for personalized Semantic Web search. Information Sciences 179, 2794–2808 (2009)
9. Lassila, O., Swick, R.R.: Resource Description Framework(RDF) Model and syntax specification, W3C Recommendation (1999)
10. Brickley, D., Guha, R.V.: Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation (2000)
11. Lee, M., Kim, W.: Semantic Association Search and Rank Method based on Spreading Activation for the Semantic Web. In: IEEE Int'l Conference on Industrial Engineering and Engineering Management, pp. 1523–1527 (2009)
12. Lin, S., Chalupsky, H.: Unsupervised Link Discovery in Multi-Relational Data via Rarity Analysis. In: Proc. 3rd IEEE Int'l Conf. Data Mining, pp. 171–178. IEEE CS Press (2003)
13. Nazir, S., Afzal, M.T., Qadir, M.A., Maurer, H.A.: CAOWL-SAI: Context aware OWL based semantic association inference. In: NCM 2010, pp. 746–751 (2010)
14. Shariatmadari, S., Mamat, A., Ibrahim, H., Mustapha, N.: SwSim: Discovering semantic similarity association in semantic web. In: Proc. of the Int'l. Symposium on IT Sim 2008, pp.1–4 (2008)

15. Sokolsky, O., Kannan, S., Lee, I.: Simulation-Based Graph Similarity. In: Hermanns, H. (ed.) TACAS 2006. LNCS, vol. 3920, pp. 426–440. Springer, Heidelberg (2006)
16. Stojanovic, N., Mädche, A., Staab, S., Studer, R., Sure, Y.: SEAL – A Framework for Developing SEmantic PortALs. In: K-CAP 2001 – Proceedings of the First Int'l. ACM Conference on Knowledge Capture, Victoria, B.C., Canada (2001)
17. Vidal, M., Rashid, L., Ibabez, L., Rivera, J., Rodrogiez, H., Ruckhaus, E.: A Ranking-Based Approach to Discover Semantic Association Between Linked' Data. In: The 2nd Int'l Workshop on Inductive Reasoning and Machine Learning for the Semantic Web, pp. 18–29 (2010)