Entity-based Semantic Association Ranking on the Semantic Web

S Narayana
Gudlavalleru Engineering College Gudlavalleru, Andhra Pradesh, India

S Sivaleela
Gudlavalleru Engineering College Gudlavalleru, Andhra Pradesh, India

A. Govardhan
Professor of CSE & Director of Evaluation, JNTU, Hyderabad

G.P.S. Varma
Professor, S.R.K.R. Engineering College, Bhimavaram

ABSTRACT

The major focus of today's search engines is efficient retrieval of relevant documents from the web. Recently Semantic Web has received greater interest in industry and academia and retrieving relevant information over huge amounts of Semantic Meta data is becoming popular. In particular discovering and ranking complex relationships between two entities over Semantic Meta data became a challenging research topic. Semantic Associations capture complex relationships between two entities in an RDF knowledge base. Given two entities, there exist a huge number of Semantic Associations between entities. Moreover these associations pass through one more intermediate entity. Hence ranking of associations is required in order to get relevant associations. This paper proposes an approach to discover and rank Semantic Associations between two entities based on the user interest. User interest is captured by selecting one or more entities from the user interface. The effectiveness of the ranking method is demonstrated using Spearman Foot rule coefficient. The results show that the proposed ranking is highly correlated with human ranking.

Keywords

Semantic Web, Semantic Association, Complex relationship, RDF, Ontology.

1. INTRODUCTION

The size of the World Wide Web is increasing day by day due to the addition of new web sites every day. As a result, accessing the relevant information from the Web has become intricate and demanding. Traditional search engines such as Google, Alta-Vista, Yahoo etc. have made good progress to locate relevant information from among huge amounts of information on the Web. However these search engines produce relevant information based on keywords or key phrases. As a result they produce too many results which further require investigation to locate relevant information. Sometimes user may be interested in finding what relationships exist between two entities like people, events, and places. Finding such relationships between two entities is more useful in domains such as national security, business intelligence, genetics and pharmacy. For example, the airport security agents may wish to find what relationships exist between two entities P and Q, in order to assess the risk of flight. It is very difficult to find such relationships using traditional search engines. Hence new mechanisms are to be invented to find and rank the relationships between two entities. These complex relationships are called as Semantic Associations

As of now, some effort has been made to find and rank Semantic Associations using ontology and Semantic Web [1] technology. The Semantic Web not only consist resources but also consist heterogeneous relationships that exist between resources. With the amount, size and complexity of ontologies growing rapidly, the number of Semantic Associations between a pair of entities is becoming increasingly overwhelming. Moreover these associations pass through one or more intermediate entities. For example if the user wishes to find Semantic Associations between two entities involving two 'Computer Science Researchers' over the SWETO [8] test-bed, he gets hundreds or thousands of associations. Hence discovering and ranking such associations based on user's interest is needed.

To overcome this problem, Aleman Meza et al.[3] propose a method to rank Semantic Associations using six types of criteria called Subsumption(components that occur at lower level in the hierarchy convey more information than those that occur at upper level), Path length(allows to select longer or shorter paths), Popularity(allows to prefer popular entities or unpopular entities based on number of incoming and outgoing edges), Rarity(allows to prefer rarely occurring or commonly occurring paths), Trust(decides the reliability of the relationship based on its origin) and Context. In this method, context is defined by selecting the region that covers user interested entities from the RDF [6] graph using a touch graph like system. Based on the selection, the weights are calculated and the associations are ranked. But when the size of the RDF graph grows, it is difficult for the user to select interested entities to define the context. This paper proposes a flexible method to select user interested entities to define context. The method operates in two levels; in the first level, associations are generated using Aleman Meza et al. method; in the second level, all the entities that occurs in all the associations are displayed in a list through which user can select his interested entities; based on this selection, associations are ranked.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data model and basic definitions of Semantic Associations. Section 4 explains the proposed method. The experimental results are presented in section 5. Section 6 draws some conclusions and possible future work.

2. RELATED WORK

Several methods have been proposed to discover and rank Semantic Associations. Anyanwu and Sheth et al. [2] propose a method to discover and rank Semantic Associations. In that they used ρ-operator which checks whether an association is possible. If so a traversal is made in the description base. The authors used the notion of context to capture the relevant region(s) which contain potential paths. In addition a user may

assign ranks to important properties in the order of importance. This allows the display of associations with the highest relevance rank first.

Shahdad Shariatmadari et al. [5] propose a technique to find Semantic Associations using semantic similarity. Anyanwu et al. [7] propose a method to rank Semantic Associations. In this method, with the help of a sliding bar user can easily vary the search mode from conventional search mode to discover search mode.

Aleman Meza et al.[3] propose a method to rank Semantic Associations using six types of criteria called Subsumption(components that occur at lower level in the hierarchy convey more information than those that occur at upper level), Path length(allows to select longer or shorter paths), Popularity(allows to prefer popular entities or unpopular entities based on number of incoming and outgoing edges), Rarity(allows to prefer rarely occurring or commonly occurring paths), Trust(decides the reliability of the relationship based on its origin) and Context. This method also ranks Semantic Associations using user preferences such as favor rare or common associations, popular or unpopular associations and shorter or longer associations.

Lee M et al.[11][12] propose a method to rank Semantic Associations based on information theory and spreading activation to expand the semantic network. In this method, the results are provided that are relations between search keyword and other resources in a semantic network.

Viswanathan and Ilango et al. [4] propose a personalization approach for ranking Semantic Associations between two entities. They capture user's interest level in different domains based on their Web browsing history. The value of the user's interest level is stored in a table and based on these values the context weight of the associations is calculated and ranked.

The main difference between the proposed method and other existing methods is that there is less flexibility to select the user interested entities. In the proposed method, more flexibility is given for the user to select interested entities. Associations are passing through these entities are considered more relevant and the others are less relevant.

3. BACKGROUND

3.1 Data Model

The Resource Description Framework (RDF) is a World Wide Web Consortium (W3C) standard for describing Web resources. RDF [6] data model is a directed labeled graph of nodes and edges in which nodes represent resources and edges represent relationships. The edge is labeled with the name of the property and resource is labeled with the URI of the resource. A resource can be an entity or a literal. An RDF statement is a triple consisting of Subject, Predicate, and Object. Subject is connected to Object by the predicate. Object can be another resource or a literal. A special property rdf:typeOf connects resources of the same type. The classes and properties are described in an RDF Schema (RDFS) [8]. RDFS provides the standard vocabulary for RDF.

3.2 Semantic Associations

The complex relationships between two entities are known as Semantic Associations [2]. Most useful Semantic Associations involve some intermediate entities and relationships. To define Semantic Associations, the formalism specified by Anyanwu et al. [2] is followed.

3.2.1 Definition 1 (Semantic Connectivity)

Two entities e1 and en are semantically connected if there exists a sequence e_1, P_1, e_2, P_2, …, e_{n-1}, P_{n-1}, e_n in an RDF graph where e_i ($1 \le i \le n$) are entities and P_j ($1 \le j \le n$) are properties. Fig. 1 shows the semantic connectivity between e_1 and e_n.

3.2.2 Definition 2 (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_{n-1}, P_{n-1}, en and f_1, Q_1, f_2, Q_2, …, f_{n-1}, Q_{n-1}, fn semantically connecting e_1 with en and f_1 with fn respectively, and that for every pair of properties P_i and Q_i, $1 \le i \le n$, either of the following conditions holds; $P_i = Q_i$ or P_i □ Q or Q_i □ P_i (□ means rdf:subPropertyOf), then two paths originating at e_1 and f_1, respectively, are semantically similar.

3.2.3 Definition 3 (Semantic Association)

Two entities e_x and e_y are semantically associated if e_x and e_y are semantically connected or semantically similar.

4. RANKING SEMANTIC ASSOCIATIONS

In the proposed method, ranking of Semantic Associations is performed in two levels. In the first level, associations are ranked using Aleman Meza et al.[3] method which uses six weights to rank Semantic Associations. These are described as follows;

Contex Weight Cp: Consider the scenario in which user is interested in finding Semantic Associations in the domain of 'Computer Science Publication'. Then concepts such as 'Scientific Publication', 'Computer Science Professor' and 'Computer Science Researcher' are considered to be more relevant and the concepts such as 'Financial Organization' and 'Terrorist Organization' are considered to be less relevant. So user is provided to select his interesting regions and based on this associations are ranked.

Subsumption weight Sp: In an RDF graph, entities that occur at lower level in the hierarchy are treated as more specialized entities than the entities that occur at higher levels. Thus, lower level entities convey more meaning. So Associations that consists these entities are more relevant.

Path Length Weight Lp: Some times, user may be interested in finding shorter associations yet in other cases he may wish to find longer associations. So user can determine which association length influence

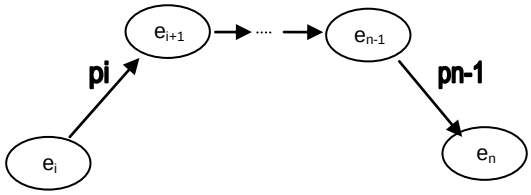

Fig 1: Semantic connectivity between two entities e_1 and e_n

Popularity Weight Pp: Entity popularity is defined based on number of incoming and outgoing edges the entity has. Associations that contain popular entities are considered popular associations. Hence, user has to select whether 'favor more popular associations or favor less popular associations' based on his interest.

Rarity Weight Rp: User may be interested in either rarely occurring events or commonly occurring events. For example, in case money laundering user may be interested in commonly occurring events because money launderers' perform several common transactions to escape from law. So user is allowed to select 'favor rare associations' or 'favor common associations'.

Trust Weight Tp: The entities and relationships in a Semantic Association come from different sources. Some sources may be more trusted and some sources may be less trusted. So trust value is assigned to components in an association based on the source from which it is coming.

The overall Semantic Association, SA, ranking is calculated using the formula as

$$R1_{SA} = k_1 \times C_p + k_2 \times S_p + k_3 \times L_p + k_4 \times P_p + k_5 \times R_p + k_6 \times T_p$$

Where $k_1+k_2+k_3+k_4+k_5<=1$ and is required to fine-tune the ranking of Semantic Associations.

4.1 Ranking Semantic Associations

In the second level, the associations are further ranked based on the entities selected by the user. For example consider the scenario where the user is interested in finding the associations between two persons in the domain of 'Scientific Publication'. Then concepts such as 'Scientific publication', 'Computer Science Professor' and 'Computer Science Researcher' are considered to be more relevant where as concepts such as 'Finance' or 'Financial Organization' would be less relevant. Hence it is possible to capture user's interest through context specification. Context can be specified by selecting the entities from the list in a user interface screen. Thus it is possible to rank the associations based on the entities selected by the user which defines his domain of interest. Figure 2 shows part of an RDF graph. There are three possible associations. The top most association (call it association1) contains three entities 'Country', 'Person', and 'Scientific Publication'. The next association (call it association2) contains two entities 'Person' and 'State'. The third association (call it association3) contains two entities 'Scientific Publication' and 'Computer Science Researcher'. Suppose the user selected two entities such as 'Scientific Publication' and 'Computer Science Researcher'. Then association3 is considered to be more relevant and ranked high as it passes through all the interested entities and the remaining are considered being less relevant and ranked lower.

In the second level of ranking Semantic Associations, SA, following formula is used to rank the Semantic Associations.

$$R2_{SA} = R1_{SA} + \frac{x}{c}$$

Where x denotes number of components of SA passing through the selected entities and c denotes total number of components in SA excluding first and last entities.

5. EXPERIMENTAL RESULTS

To find Semantic Associations, we have used SWETO (Semantic Web Technology Evaluation ontology) [9] test-bed. SWETO captures real world knowledge which includes entities related to cities, states and countries, air ports, companies, banks, terrorist attacks and organizations, persons and researchers, scientific publications, journals, conferences and books. Semantic Associations are generated and ranked using various criteria including favor short or long associations, favor popular or unpopular entities favor rare or common associations and context. In addition, associations are ranked based on the entities selected by the user. The criteria and entities to rank Semantic Associations are selected through user interface. Semantic Associations are ranked by the system as well as manually.

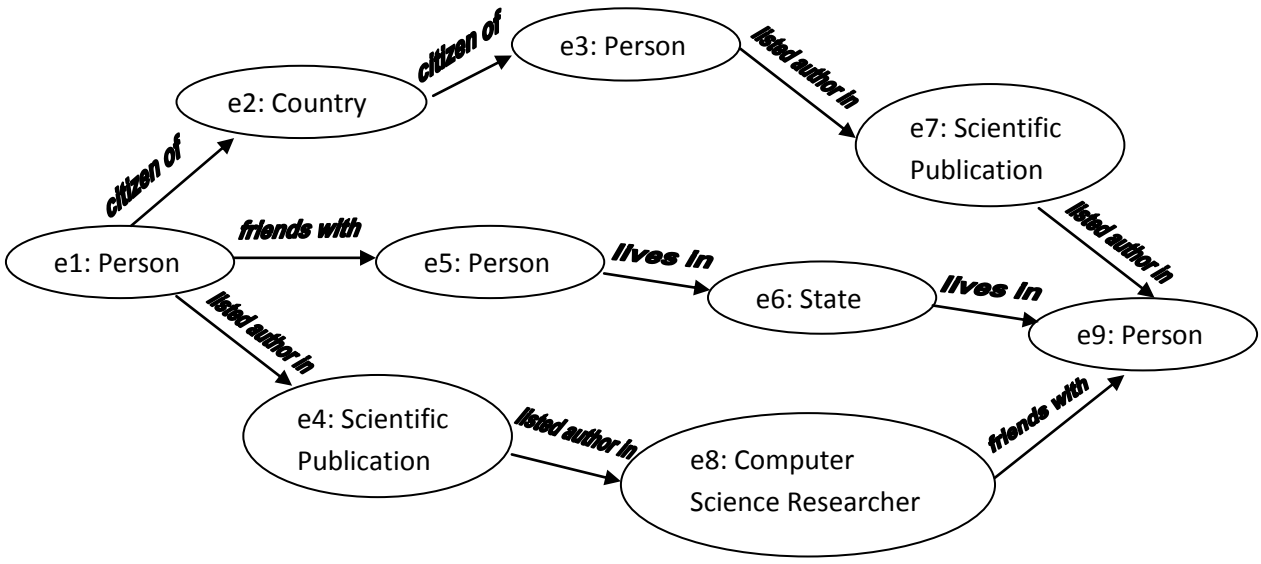

Fig 2: Semantic Associations between two entities e1 and e9 in RDF data

5.1 User Interface

User interface is a web based application using Servlet and Apache Tomcat. Using this interface, user enters two entities between which he wish to find Semantic Associations as shown in figure 3. The system then finds and ranks the Semantic Associations using the criteria as discussed above. Selection of user interested entities is shown in figure 4. Ranked Semantic Association results are shown in Table 1.

5.2 Preliminary Results

To demonstrate the effectiveness of the proposed method, we have compared the system ranking with human ranking between two entities **John Rockefeller** and **Jeb Bush** is shown in figure 5. The x-axis shows ranking of Semantic Associations by the proposed method and y-axis shows user-human ranking which is assigned by the users manually. The Spearman's foot rule [10] distance measure is used to measure the similarity between proposed system ranking and user-human ranking.

Spearman's Foot rule distance measure is given as

$$D(_{system\ ,\ human\ }) = \sum_{i=1}^{n} |R_{isystem} - R_{ihuman}|.$$

Spearman's Foot rule Coefficient

$$C = 1 - \frac{4D}{n^2}$$

Fig 3: User Interface for entering two entities

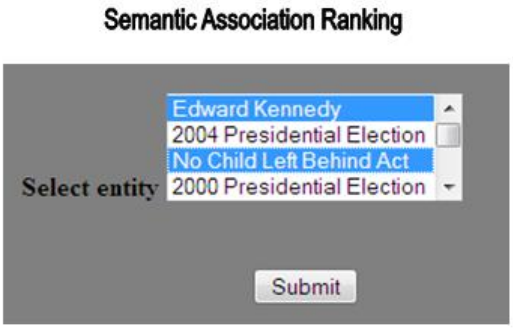

Fig 4: User Interface for selecting entities to define user's interest

Based on the results, the average correlation coefficient between the proposed system ranking and user-human ranking is 0.69. Since the average correlation coefficient between proposed system ranking and user-human ranking is greater than 0.50, the proposed system ranking and user-human ranking are highly correlated. Comparison of correlation between proposed system and other existing methods is shown in figure 6. It shows that correlation of proposed method is higher than the existing methods.

Table 1. Ranking of Semantic Associations

Association	Rank
John Rockefeller - member of - **U.S. Senate** - member of - **Edward Kennedy** - promoted law - **No Child Left Behind Act** - signed by - **George W. Bush** - relative of - **Jeb Bush**	1
John Rockefeller - member of - **Democratic Party** - member of - **Edward Kennedy** - promoted law - **No Child Left Behind Act** - signed by - **George W. Bush** - relative of - **Jeb Bush**	2
John Rockefeller - member of - **U.S. Senate** - member of - **Edward Kennedy** - opposed law - **PATRIOT Act** - signed by - **George W. Bush** - relative of - **Jeb Bush**	3
John Rockefeller - member of - **U.S. Senate** - member of - **Edward Kennedy** - opposed law - **PATRIOT Act** - promoted law - **George W. Bush** - relative of - **Jeb Bush**	4
John Rockefeller - member of - **Democratic Party** - member of - **Edward Kennedy** - opposed law - **PATRIOT Act** - signed by - **George W. Bush** - relative of - **Jeb Bush**	5
John Rockefeller - member of - **Democratic Party** - member of - **Edward Kennedy** - opposed law - **PATRIOT Act** - promoted law - **George W. Bush** - relative of -**Jeb Bush**	6
John Rockefeller -member of- **U.S. Senate** - passed law - **No Child Left Behind Act** - signed by - **George W. Bush** - relative of - **Jeb Bush**	7
John Rockefeller - member of - **U.S. Senate** - member of - **John Kerry** – lost - **2004 Presidential Election**- won - **George W. Bush** - relative of - **Jeb Bush**	8
John Rockefeller- member of - **Democratic Party** - member of - **John Kerry** – lost - **2004 Presidential Election**- won - **George W. Bush** - relative of - **Jeb Bush**	9
John Rockefeller - member of - **Democratic Party** - member of - **Al Gore** – lost - **2000 Presidential Election** – won - **George W. Bush** - relative of - **Jeb Bush**	10

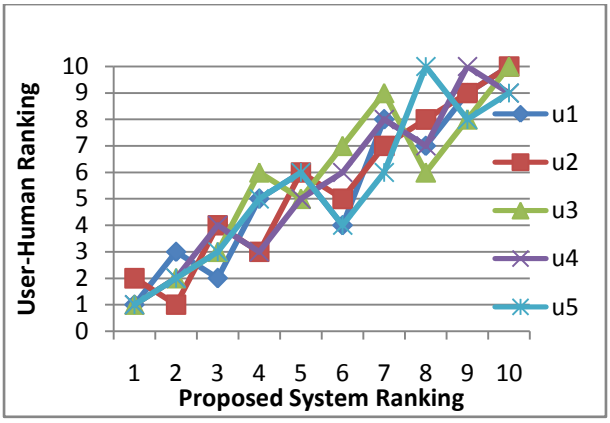

Fig 5: Comparison of human and proposed system ranking

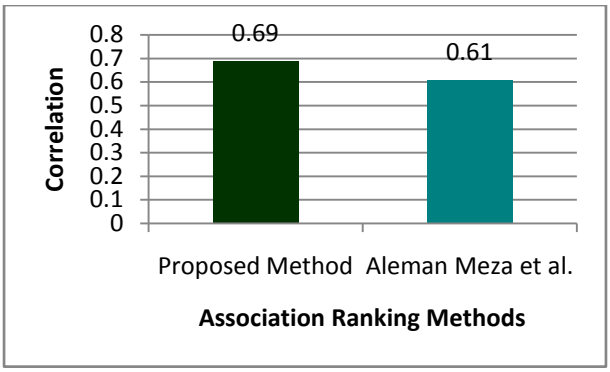

Fig 6: Comparison of Correlation between Proposed and Existing method

6. CONCLUSION

Finding Semantic Associations between two entities is very useful especially in applications such as national security, business intelligence, genetics and pharmaceutical research. Given two entities, there exist a huge number of associations. For discovering and ranking such associations new techniques are required. This paper proposes a technique for efficiently ranking the Semantic Associations based on entity selection. We have compared the ranking of proposed method with the existing methods using Spearman's Foot rule. The average correlation coefficient of proposed method is 0.69 which is greater than other existing methods. It also reveals that proposed system ranking is highly correlated with human ranking. The main limitation of the proposed method is specifying the context and selection of entities from the user interface. Sometimes it is difficult to specify these features. Also when the size of the RDF graph grows it is very difficult to specify these features. So as a future extension, a learning model is to be developed to learn user preferences. Once the model is learnt the user preferences, then Semantic Associations are ranked more easily.

7. REFERENCES

[1] Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web: a new form of web content that is meaningful to computers will unleash a evolution of new possibilities. Sci Am 285(5):34–43. doi:10.1038/scientificamerican1101-34.

[2] Anyanwu, K. Sheth A. ρ-operator: Discovering and Ranking Semantic Associations on the Semantic Web, ACM SIGMOD Record, v. 31 n.4, December 2002.

[3] Aleman-Meza B, Halaschek C, Arpinar IB Sheth A (2005): Ranking Complex Relationships on the Semantic Web. IEEE Internet Computing 9(3); 37-44. Doi:10.1109/MIC.200.63.

[4] V Viswanathan, K Ilango: Ranking semantic relationships between two entities using personalization in contest specification. Informain Sciences, Elsevier, 207 (2012) 35-49

[5] Shariatmadari S, Mamat A, Ibrahim H, Mustapha N (2008) SwSim:Discovering semantic similarity association in semantic web. Pro. Of International Symposium on IT Sim 1-4.

[6] O. Lassila and R. Swick. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation. 1999.

[7] K. Anyanwu, A. Maduko, A. Sheth, SemRank: ranking complex relationship search results on the Semantic Web, in: Proc. of the 14th International World Wide Web Conference, ACM Press, 2005, pp. 117–127.

[8] D. Brickley and R.V. Guha. Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation. 2000.

[9] SWETO: Semantic Web Technology Evaluation Ontology:http:/lsdis.cs.uga.edu/projects/SemDis/Sweto.

[10] P. Diaconis, R. Graham, Spearman's footrule as a measure of disarray, Journal of the Royal Statistical Society Series B 39 (2) (1977) 262–268.

[11] M. Lee, W. Kim, Semantic association search and rank method based on spreading activation for the Semantic Web, in: IEEE International Conference on Industrial Engineering and Engineering Management, 2009, pp. 1523–1527.

[12] M. Lee, W. Kim, S. Park, Searching and ranking method of relevant resources by user intention on the Semantic Web, Expert Systems with Applications 39 (2012) 4111–4121.

[13] S Narayana, A. Govardhan, G.P.S. Varma, Discovering and Ranking Semantic Associations on the Semantic web, International Journal of Computer Science and Management Research, Vol 1 Issue 5 December 2012, pp. 1092-1102.

[14] A. Maedche, S. Staab, N. Stojanovic, R. Studer, Y. Sure, Semantic PortAL-The SEAL approach, in: D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (Eds.), Creating the Semantic Web, MIT Press, MA, Cambridge, 2001.

[15] A. Pretschner, S. Gauch, Ontology based personalized search, in: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, 1999, pp. 391–398.

[16] N. Stojanovic, R. Studer, L. Stojanovic, An approach for the ranking of query results in the Semantic Web, in: Proc. 2nd International Semantic Web Conference, Sanibel Island, Florida, 2003, pp. 500–516.