USING SEMANTIC ANNOTATIONS OF WEB SERVICES FOR ANALYZING INFORMATION DIFFUSION IN THE DEEP WEB

Shahab Mokarizadeh[1], Peep Küngas[2] and Mihhail Matskin[1]
[1]*Royal Institute of Technology, Stockholm, Sweden*
[2]*University of Tartu, Tartu, Estonia*

Keywords: Semantic Web Services, Information Diffusion, Deep Web, Linked Services, Web Services Network Analysis.

Abstract: Since Web services represent a fragment of the Deep Web, Web service interface descriptions reflect the content types available in the Deep Web. Therefore semantic annotations of these Web service interfaces, after using them to link services to services networks, allow analysis of the structure of the Deep Web. In this work, we investigate information diffusion, as one of highlighted Deep Web research directions, among networks of Web services. We present a model for analyzing information diffusion between both individual service providers and entire service industries. The proposed model is evaluated based on set of public Web services interface description harvested from public registries. The model indicates high potential of the proposed method in understanding the hidden structure of the Deep Web and interactions between individual service providers or service industries.

1 INTRODUCTION

Web services represent a fragment of the Deep Web (Bergman, 2001) since they facilitate access to data, which is neither visible to search engines nor directly explorable. Semantic annotations of Web service interfaces not only make the services searchable by their thematic content but also allow, after using the annotations to link the services into services networks, analysis of the underlying deep web content. In this work, we investigate information diffusion, as one of highlighted Deep Web research directions (Geller et al., 2008), among networks of Web services. Information diffusion is defined as the communication of knowledge over time among members of a social system (Shi et al., 2009). The respective studies (Cha et al., 2009; Teng et al., 2009; Shi et al., 2009) in the context of biosphere, microblogs, and publication citation have turned to be useful for revealing intrinsic properties of particular real world phenomena. Similarly, the services that are published in the Web not only offer capabilities but also indirectly exploit the content and data published by other Web services. This creates a kind of conceptual ecology of knowledge where information is shared and flows along input and output parameters of service operations. Our hypothesis is that analysis of information diffusion in Web services networks can reveal intrinsic properties of underlying Web services. An example of such properties is the hidden reality of how Web services in different service commodities have been designed from information exchange perspective.

This paper presents a model for analyzing information diffusion among commodities of Web services given the network of Web services. The proposed approach relies on a set of semantically annotated and categorized web services to first construct a Web services network, then transform it into a category (commodity) network, and finally compute a diffusion matrix. The diffusion matrix captures the volume of potential information flow between Web services categories. The volume of information flow reflects collaboration between different service industries. The proposed approach is evaluated on set of public Web services (in WSDL interfaces) exposed by the major service industries. From semantic deep web perspective, our work follows deep web service annotation approach to access deep web content and it addresses deep web data fusion issue according to Geller et al. (Geller et al., 2008).

The rest of this paper is organized as follows. In Section 2 we introduce the foundations of Web service categorization, semantic annotation and network formation. In Section 3 we outline our model for analyzing information diffusion in Web service net-

Mokarizadeh S., Küngas P. and Matskin M..
USING SEMANTIC ANNOTATIONS OF WEB SERVICES FOR ANALYZING INFORMATION DIFFUSION IN THE DEEP WEB.
DOI: 10.5220/0003931801100115
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 110-115
ISBN: 978-989-8565-08-2

works. Section 4 describes our experimental settings and analyses the results. Finally, conclusions and future work are presented in Section 5.

2 PRELIMINARIES

We define information diffusion in terms of information flow from output parameter(s) of a Web service operation to input parameter(s) of other Web service operations in a Web services network. To this end, we first categorize and semantically annotate the Web services under examination. Web service matchmaking is the next step which leads to construction of a Web service networks. Finally, we apply our information diffusion discovery model to estimate the information flow in the network.

2.1 Web Service Categorization

In Web service categorization step, we assign each individual Web service to its corresponding categories. A category describes a general kind of a service that is provided, for example "banking service" and "weather service" (Heß and Kushmerick, 2003). In the context of this paper we are only interested in categorizing Web services at higher category levels (e.g. "E-Commerce", "Weather", etc.) rather than at lower levels (e.g. "search for a flight", "get temperature"). For instance, *Logistics* category in our categorization scheme includes any Web service whose operations are related in some way to transportation or postal services such as *DHL Service and Fedex Notification Service*. In this regard, our categorization scheme is similar to the approach exploited by Heß and Kushmerick (Heß and Kushmerick, 2003) and Crasso et al. (Crasso et al., 2008). We assume that there exists a set $D = \{d_1, d_2, ..., d_n\}$ of Web service categories where no structural relationship (e.g. taxonomic) is assumed among members of D. It should be noted that a Web service can be associated with multiple categories.

2.2 Semantic Annotation and Web Service Matching

In this work we only require annotation of basic elements of Web service operation input and output parameters. These element names are either WSDL message part names or XML schema leaf element names. The reason is that the actual pieces of information, exchanged between services, are encoded with these basic elements. The extracted terms are ingredients of our previously developed ontology learning component (Mokarizadeh et al., 2010) to generate a reference domain ontology. The reference ontology is formally presented as $C = \{c_1, c_2...\}$, where c_i represents an element in a reference ontology. In our reference ontology, concepts are inter-related through additional ontological relations (Mokarizadeh et al., 2010).

Semantic annotations of Web services are exploited in order to find semantic matching between inputs and outputs of services. As the annotated elements (i.e. *terms*) are in fact instances in the generated reference ontology, the instance matching process is used to find ontological relationships between those instances. We employ a rule-based instance matching method that has been already described and evaluated in our previous work (Mokarizadeh et al., 2011). The matching component takes as input a pair of instances and produces a correspondence element. Each correspondence element implies whether a semantic relation holds between the two given instances, according to a particular matching rule. The presence of such semantic relation means that the underlying output and input elements of Web service operation parameters can be matched. The implicit assumption here is that matching process is only performed between pair of elements where one of them represents an output element of a Web service operation and the second one depicts an input element of another Web service operation. The results of matching process is exploited in Web service network formation which will be discussed in next sections.

2.3 Web Services Network Models

We distinguish Annotated, Semantic and Category representations of Web service networks derived from semantically annotated Web services.

Annotated Web Service Model. This network captures main elements of WSDL descriptions as nodes and edges of a directed graph. The graph is further enriched with references to ontology elements and category labels. A node P_i in this model refers to input and output parameters (i.e. the WSDL message part names and XSD schema leaf element names) of Web service operations. Every node is annotated with: 1) a semantic label C_i that points to an ontology element in reference ontology C, and 2) category label D_i that refers to the affiliated category in category list D. Finally, nodes are connected by respective Web service operations represented as directed edges from nodes representing input elements towards the nodes depicting the output elements. In fact, an instance of this network model is nothing more than a collection of discreet graphs constructed to facilitate understanding of the subsequent network transformation mechanisms.

An illustrative example of this network model is shown on the left side of Figure 1. Accordingly, the network is formed by two web services (WS_1 and WS_2), each of which consists of one operation (OP_1 and OP_2 respectively). The services are classified under category labels D_1 and D_2. Basic elements are denoted by nodes $P_1 - P_5$ and annotated with concepts $C_1 - C_4$. Moreover, the assigned category to each Web service WSDL description is propagated to their WSDL elements (not shown in Figure 1 for the sake of readability).

Semantic Network Model. A Semantic network is a loop-free directed graph and it is the semantically unified representation of the underlying annotated Web service network. A directed edge in this graph shows direct dependency between source and target node such that the concept represented by target node is produced by a service operation only if the required concepts, represented by source nodes, is given. A semantic node C_i in this model refers to a semantic concept. This concept could represents unification of one or several ontological concepts in ontology C. Every semantic node is further associated with category vector $\overrightarrow{Q}$ denoting the weight of the semantic node in different categories wrt its relative occurrence in the categories.

Category Network Model. This model represents a directed graph and it is used to capture the category view of the underlying Web services network. Node D_i in this model represents an individual Web service category while edges are denoting inter-category relationships (e.g. direction of information flow). Moreover, edges are labeled with weights expressing the volume of information flowing from source to target node. Unlike the previous models, self-loops are permitted in this model.

3 INFORMATION DIFFUSION IN WEB SERVICES NETWORKS

3.1 Web Services Network Formation

Applying of the proposed semantic annotation and matching methods results in emergence of the respective annotated Web services network. This network is the main input for construction of other two types of networks—semantic and category networks. Transformation mechanisms to create instances of semantic and category networks are elaborated in the rest of the section.

3.1.1 Semantic Network Formation

Transformation of an annotated Web services network to a semantic network starts by replacing the nodes with corresponding ontological concepts. Lets consider again the example of the annotated network in Figure 1. In this transformation process, the input parameters P_1 and P_4 are replaced with ontological concepts C_1 and C_3 respectively while C_2 and C_4 substitute the output parameters P_2, P_3, P_5 in a similar manner. Part *(b)* in Figure 1 shows the transformed network. Next, we exploit the results of match-making process to unify the concepts representing matched output and input elements. This potentially results in emergence of new nodes with unified concept labels. Every emerging semantic node also inherits the incoming and outgoing edges of the parent matching nodes as well. Lets consider the set $\{\langle C_1, C_3\rangle, \langle C_2, C_4\rangle\}$ as the only possible matching cases in the previous example. Thus as a result of unification, we will have a graph with source node $C_{1,3}$ and target node $C_{2,4}$ and three directed edges from $C_{1,3}$ to $C_{2,4}$. Next redundant edges are eliminated, so that there will be only one edge connecting two nodes. Part *(c)* of Figure 1 illustrates the result of this transformation. Meanwhile the associated categories of the nodes in the respective annotated network are propagated to corresponding semantic nodes. Each node in the semantic network might be associated to several categories. We model the affiliated categories of semantic node C_u as a normalized category vector $\overrightarrow{Q_u} = \{q_1, q_2, ..., q_n\}$, where every item q_s represents the weight of concept C_u in the category $D_s \in D$. The concept weights are calculated as follows:

$$q_s = \frac{frequency\ of\ C_u\ in\ D_s}{\sum_{i=1}^{n} frequency\ of\ C_u\ in\ D_i} \qquad (1)$$

where n refers to the size of category set D. Returning back to network presented in part *(c)* in Figure 1, both semantic nodes $C_{1,3}$ and $C_{2,4}$ are associated with both D_1 and D_2 as the result of weight propagation. The normalized category vector for $C_{1,3}$ according to (1) is $\overrightarrow{Q_{1,3}} = \{q_1 = 0.5, q_2 = 0.5\}$ and for semantic node $C_{2,4}$ is $\overrightarrow{Q_{2,4}} = \{q_1 = 0.67, q_2 = 0.33\}$.

3.1.2 Category Network Formation

Transformation of a semantic network into a category network starts with replacing semantic nodes with their affiliated category labels. Meanwhile, we propagate the category weights from semantic nodes to the corresponding edges. The category propagation mechanism works as follows. Let us assume that there exists a directed edge (C_u, C_v) in the semantic

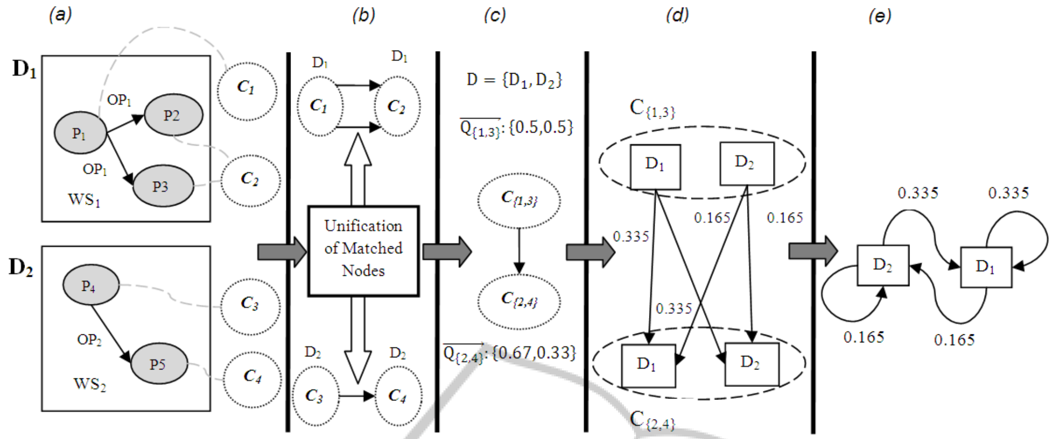

Figure 1: Transformation of annotated Web services to category network. *(a)*:Annotated network, *(c)*: Semantic network, *(e)*: Category network, *(b)* and *(d)*: Intermediate networks.

network. In addition let us also assume that C_u is affiliated with category D_s with weight $q_{u,s}$ and similarly, C_v is associated to category D_t with weight $q_{v,t}$. By replacing the semantic nodes with respective categories, we obtain the partial category weight for directed edge (D_s,D_t) as follows:

$$\omega_{u,v(D_s,D_t)} = q_{u,s} * q_{v,t} \qquad (2)$$

We refer to $\omega_{u,v(D_s,D_t)}$ as partial weight since the transformation may result in appearance of multiple edges between the same pair of category nodes. In the last step, we restructure the network so that identical nodes (i.e. nodes with the same labels) are unified. Consequently, the category weights of identical edges (i.e those having the same source and target nodes) are augmented into one single representative edge and weight vector. In other words, for every directed edge (D_s,D_t) in the category network, the actual weight is computed as follows:

$$W_{(D_s,D_t)} = \sum_{\forall\, edge\,(u,v)\; in\, Semantic\, Network} \omega_{u,v(D_s,D_t)} \qquad (3)$$

To illustrate application of the preceding, let us consider again the semantic network in part *(c)* of Figure 1. The result of first stage of transformation is depicted in the graph shown in part *(d)* of Figure 1. Accordingly, the semantic nodes $C_{1,3}$ and $C_{2,4}$ are replaced with their associated category labels D_1 and D_2. As the category weights of semantic nodes are available in $\overrightarrow{Q_1}$ and $\overrightarrow{Q_2}$, we apply (2), which results in the following edge weights: $\omega_{(D_1,D_1)} = 0.5 * 0.67$, $\omega_{(D_1,D_2)} = 0.5 * 0.33$, $\omega_{(D_2,D_1)} = 0.5 * 0.67$ and $\omega_{(D_2,D_2)} = 0.5 * 0.33$. Next, by unifying the identical edges and augmenting the category weights,the Category network presented in *(e)* of Figure 1 is constructed. Since the transformation resulted only one instance for each category edge, the actual weight for each edge will be equal to the partial weight. Part *(e)* of Figure 1 illustrates the final constructed Category network.

3.2 Measuring Information Diffusion

In order to measure density of information flow between different Web service categories, we adopt the approach exploited by Shi et al. (Shi et al., 2009) in the context of analyzing information diffusion in citation networks. We regard category weights as diffused information volume from source toward target category nodes. In order to make the information flow between different categories in one scale and make comparison, we follow Z-score normalization principles. To this end, we first compute the sum of all weights for all outgoing edges from each category in the network and populate a matrix A with these values. We then normalize (i.e. divide) the volume (i.e. sum) of weighted edges between any pair of nodes by the rate we would expect if the volume of weights of incoming and outgoing edges were the same.

Let us assume that $W_{(D_i,D_j)}$ is the actual weight of edge (D_i,D_j) obtained by utilization of (3), $W_{i*} = \sum_j W_{(D_i,D_j)}$ is the sum of all weights of all links from category i, $W_{*j} = \sum_i W_{(D_i,D_j)}$ is the sum of all weights of all links to category j and $W = \sum_{i,j} W_{(D_i,D_j)}$ is the sum of all weights of all links in matrix A. Then the expected volume of weights, assuming indifference to ones in their own category and others, from category i to category j is $E[W_{ij}] = W_{i*} \times W_{*j}/W$.

We define the category weight as a Z-score that measures standard deviations with respect to expected W_{ij}. Here we have learned that $W \gg W_{i*}$ and $W \gg W_{*j}$, hence we approximate the standard deviation by $\sqrt{E[W_{ij}]}$. In this way, for every entry in matrix A, we obtain a normalized value, which we refer to as *diffusion weight* (ϕ):

$$\phi_{ij} = (W_{ij} - \frac{W_{i*} \times W_{*j}}{W}) / \sqrt{\frac{W_{i*} \times W_{*j}}{W}} \qquad (4)$$

A high proximity (ϕ_{ij}) between categories i and j reveals a strong tendency for semantic concepts associated to category i to be resulted from invocation of services which take semantic input concepts associated to category j.

4 EXPERIMENTAL SETTINGS AND RESULTS

4.1 Data

We evaluated the proposed approach for measuring information flow in a collection of public Web services from different categories. This collection consists of around 30000 Web services' descriptions in WSDL language and they have been harvested from different public repositories during the period of 2005–2011. From this set of descriptions, we manually identified the categories of 1107 Web services according to the classification made by *SOA Trader* website [1]. We acknowledge that we haven't done any evaluation over the accuracy of this categorization. The extracted categories (26 items) together with the quantity of Web services in each category are summarized in Table 1. Additionally, each category is associated to an identifier. This identifier allows to locate each category in the computed information flow matrix presented in Figure 2.

In order to facilitate creation of semantic networks, we extracted top 30000 most recurrent terms (XSD schema leaf element names or WSDL message part names) from all WSDL documents in our dataset. This limit was mainly set to reduce the amount of computational resources needed to perform the experiments and to make evaluation process a manageable task for a human expert. This collection of most frequent terms was first syntactically normalized and processed. Next a reference ontology is automatically generated based on the mechanism explained earlier in Section 2.2. The generated ontology then is used to semantically annotate input and output parameters of Web service operations. The ontology embodies 11610 ontological concepts and it annotates around 66% of entire targeted WSDL elements. Next, we exploited the result of match-making mechanism to automatically discover matching Web service elements. Based on previous evaluation results (Mokarizadeh et al., 2011), our annotation and matching mechanism can achieve the accuracy of around 27% in terms of F-measure metric. The F-measure is defined as the weighted harmonic mean of precision and recall.

The result of Web service match-making (i.e. the correspondence elements) provides ingredients for generating the semantic network and category network formation. The general characteristics of all three types of networks (annotated, semantic and category) are shown in Table 2.

Table 1: The number of global Web services in each category.

Index	*Category*	*#Size*	*Index*	*Category*	*#Size*
1	Travel	46	14	Weather	125
2	B2B	21	15	Business	8
3	E-Health	1	16	Finance	159
4	Statistics	4	17	Interoperability	3
5	Communication	154	18	Location	33
6	Human Resources	5	19	Science	4
7	News	74	20	E-Commerce	113
8	Utilities	21	21	Security	1
9	Data	5	22	Logistics	19
10	Test	11	23	Bioinformatics	227
11	Dictionaries	6	24	GIS	16
12	Contacts	6	25	Internet	19
13	Entertainment	5	26	Industry	4

Table 2: General characteristics of exploited networks.

Network Type	*#Nodes*	*#Edges*	*Average Out Degree*
Annotated	8062	302065	37.47
Semantic	4050	157411	38.87
Category	26	588	22.62

4.2 Results

By applying (4) to the resulted category networks, we obtain a diffusion weight matrix visualized at Figure 2. The row and column numbers in the matrix are indexes to locate the corresponding category names in Table 1. The accumulated density in the diagonal line of the matrix reveals that some communities in this collection mainly provide input for their own services and consume mostly the information provided by the same community. This is because Web services in these communities exploit frequently domain-specific concepts as input and output parameters. We refer to this behavioral model as *self-referential* pattern. However, only small number of communities, namely *B2B*, *Communication*, *Business*, *Finance* and *Location* exhibit noticeable self-referential behavior.

Based on the entries in the matrix, the smallest information flow volume belongs to *E-Commerce* community. This is because the concepts representing the output parameters of services in this community is rarely appearing as input parameters of services operating in other communities. Moreover, it can be seen that this community follows also the self-referential pattern. Hence, it can be inferred that *E-Commerce* is

[1] http://www.soatrader.com/web-services/

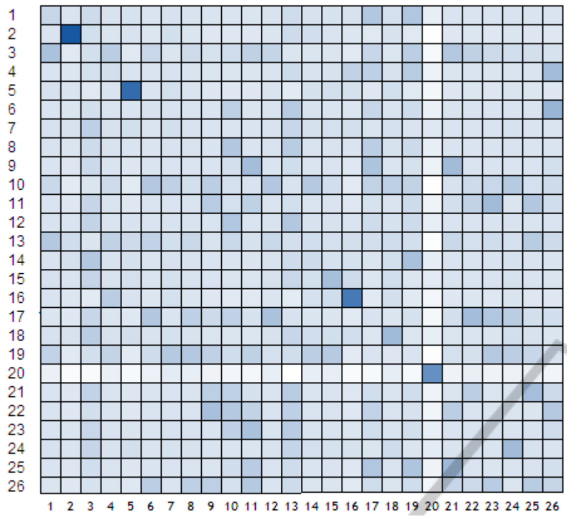

Figure 2: Visualization of matrices of category weights between different communities of Web services. Each entry is shaded according to a normalized Z-score representing whether the density of information flow is higher or less than expected at random. Darker shading indicates higher Z-scores. The diagonal line represents information flow within same category.

the most isolated community as it receives and delivers the least amount of information compared to other communities. From another perspective, isolated communities are potential candidates for developing new value-added services. By this, we mean services that can make a bridge between the isolated communities and the rest of the world, provided that logically developing such a service is meaningful and brings business value for either of parties. The aforementioned heuristics are quite compatible with the analytical rules suggested by Cui et al. (Cui et al., 2009) for pinpointing service composition opportunities in a large-scale Web services network.

The implicit assumption in the aforementioned analysis is that the utilized annotation and matchmaking scheme determines (sufficiently) accurate semantics of parameters and performs precise matching. The imperfection or bias in the annotation scheme or match-making approaches potentially leads to significant deviation from actual results which could even falsify our current results.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a model for using semantic annotations of Web service interface descriptions to measure information diffusion among categories of Web services. The experimental results demonstrate that the proposed model can be effectively used to reason about information diffusion patterns between categories of Deep Web resources, more specifically between public Web services. The main priority of our future work is targeted towards increasing the quality (both semantic annotation and categorization) of evaluated dataset to analyze further the identified patterns.

REFERENCES

Bergman, M. K. (2001). The deep web: Surfacing hidden value. *World Wide Web Internet And Web Information Systems*, 7(1):1–17.

Cha, M., Mislove, A., and Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World Wide Web*, WWW '09, pages 721–730, USA. ACM.

Crasso, M., Zunino, A., and Campo, M. (2008). Awsc: An approach to web service classification based on machine learning techniques. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 12(37):25–36.

Cui, L. Y., Kumara, S., Yoo, J. J.-W., and Cavdur, F. (2009). Large-scale network decomposition and mathematical programming based web service composition. In *Proceedings of the 2009 IEEE Conference on Commerce and Enterprise Computing*, pages 511–514.

Geller, J., Chun, S. A., and Jung, Y. (2008). Toward the semantic deep web. *Computer*, 41(9):95 –97.

Heß, A. and Kushmerick, N. (2003). Learning to attach semantic metadata to web services. In *ISWC2003*, pages 258–273. Springer.

Mokarizadeh, S., Küngas, P., and Matskin, M. (2010). Ontology learning for cost-effective large-scale semantic annotation of web service interfaces. In *EKAW*, pages 401–410.

Mokarizadeh, S., Küngas, P., and Matskin, M. (2011). Evaluation of a semi-automated semantic annotation approach for bootstrapping the analysis of large-scale web service networks. In *Web Intelligence and Intelligent Agent Technology*, pages 388–395. IEEE/WIC/ACM.

Shi, X., Tseng, B. L., and Adamic, L. A. (2009). Information diffusion in computer science citation networks. *CoRR*, abs/0905.2636.

Teng, W.-G., Pai, W.-M., and Chen, K.-C. (2009). Exploring information diffusion patterns with social relationships in the blogosphere. In *ICCI '09*, pages 422–427.