Influence Discovery in Semantic Networks: An Initial Approach

Marcello Trovati and Ovidiu Bagdasar
School of Computing and Mathematics
University of Derby
Derby, UK
Email: M.Trovati@derby.ac.uk

***Abstract*—Assessing the influence between concepts, which include people, physical objects, as well as theoretical ideas, plays a crucial role in understanding and discovering knowledge. Despite the huge amount of literature on knowledge discovery in semantic networks, there has been little attempt to fully classify and investigate the influence, which also includes causality, of a semantic entity on another one as dynamical entities. In this paper we will introduce an approach to discover and assess influence among nodes in a semantic network, with the aim to provide a tool to identify its type and direction. Even though this is still being developed, the preliminary evaluation shows promising and interesting results.**

Keywords-Knowledge Discovery; Knowledge Representations; Knowledge acquisition; Algorithms; Semantics

I. Introduction

Relation discovery plays a crucial role in understanding, assessing and predicting how knowledge spreads and evolves [10]. This clearly has a wide set of application in a variety of disciplines, including data and text mining, as well as business intelligence and analytics [3]. In fact, being able to model and analyse the output of data and its meaning is at the heart of the majority of scientific disciplines and applications.

Concepts in sematic networks are connected by edges based on specific lexical and semantic properties. Causal relationships are an example of the above which play an important role in a variety of knowledge discovery tasks with several applications [5]. More specifically, in [9], a method to automatically extract causal relationships to populate Bayesian Networks is introduced, also suggesting that for many applications it is important to consider *influence* rather than causality (see [7] for a detailed discussion).

In this paper we will introduce a method to discover and assess the influence between two semantic concepts. Even though there are a variety of methods and applications in this context (see [10] for more information), the focus is often on data acquisition and aggregation, as well as on specific semantic properties. Clearly, this is a crucial step which has to be thoroughly investigated to overcome its challenges. However, here we will discuss a scalable approach to assess the relevant parameters which determine the way one semantic entity influences another one, assuming we already have the relevant information on the relations defining a semantic network. Our main motivation is based on the expanding need to provide analytics techniques to facilitate the decision making process in an effective and accurate manner. Being able to ascertain the existence and *direction* of an influence from the information described by a semantic network, can certainly contribute to knowledge discovery and be applied to many contexts and scenarios.

In this paper, we will discuss our preliminary findings which are part of ongoing research investigation. More specifically, although the current assumptions and implementation need to be further expanded and elaborated, the initial evaluation shows interesting and promising results.

A. Background

Broadly speaking, semantics, or the science of meaning [6], describes the relationships between concepts within a "language", including among others computer languages, mathematics and science in general, as well as human language. Such semantic relationships naturally create directed networks, called *semantic networks* [6]. Network theory has increasingly attracted much interest from a variety of interdisciplinary research fields, including mathematics, computer science, biology, and the social sciences [8]. In general, networks consist of a collection of nodes, called the *node set* $V = \{v_i\}_{i=1}^n$, which are connected as specified by the *edge set* $E = \{e_{ij}\}_{i \neq j=1}^n$ [1]. Note that we do not allow *self-loops*, that is a single edge starting and ending at the same node. In a *directed network* the direction of the edges is important, which means that $e_{12} \neq e_{21}$. In other words, directed edges are not commutative.

Clearly, most things – if not all – around us are defined by a language which subsequently, can be potentially described by one or more semantic networks. Concepts in a semantic network are typically within a hierarchical structure. For example, the concept of *country* includes sub-concepts such as *people, cities, buildings*, etc. This hierarchy is expressed by the direction of its edges.

Due to the broad definition of semantics, there is a huge variety of examples that fall into the category of semantic networks. Social networks are one of them [4], where concepts correspond to people, whose mutual relationships represent social interactions. Another example of semantics is science itself, where concepts are linked by scientific relationships. For example experimental evidence of a physical interaction between concepts falls into this category.

Understanding the nature of a relation between two semantic objects corresponding to two nodes in a semantic network, has a variety of applications as well as scope for future research. However, this is by no means a trivial task. First of all, what do we mean by “type” of a relation? Usually, there is an implicit simplification by considering only a set of all well defined semantic relations, such as `belongs_to`, `is_part_of`, `is_synonym_of`, etc [2]. Even though it is not always possible to quantitatively and qualitatively assess all the above relations, they can be used to provide a general understanding of the mutual relations between two concepts. As a consequence, it is important to provide effective aggregation algorithms to classify the *overall* influence that one object exerts on another one.

In this paper, influence is defined as the way in which two concepts are linked according to their semantic properties, which include the *direction* and the *type* of such relations.

Any concept is based on hierarchical conceptualisations and abstractions which are typically difficult to fully assess. In fact, if there are certain types of relations between sub-parts of two concepts, then we might not be able to ascertain whether there is any relationship, or influence, between them. Figure 1 depicts an example based on different types of semantic relations where the layers identify the different semantic conceptualisations, namely people, groups of people, and companies. Note that such hierarchy is not just in terms of abstraction, generalisation, or even conceptualisation, but in terms of *specific* relation types. Due to the fact that identifying a general level of a hierarchy order is likely to depend on the context, we assume that we can establish such hierarchy as an initial step of our approach.

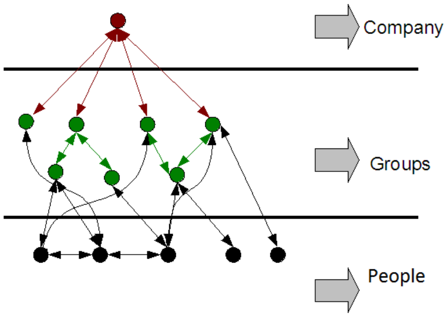

Figure 1. An example of semantic relations among semantic layers and described in Section I

The paper is organised as follows. Section II will introduce the main concepts and results. In Section IV some implementation and evaluation results are discussed, and finally, in Section V focuses on future challenges and research output based on our current results.

II. Description of the Approach

As discussed above, the problem we are aiming to address can be re-phrased as the discovery and assessment of the influence between two nodes. For convenience, we will call such nodes v_A and v_B. Note that unless specified, we will not assume it has a known direction.

Let $G = (V, E)$ be a directed semantic network, where $V = \{v_i\}_{i=1}^n$, and $E = \{e_{ij}\}_{i \neq j=1}^n$ are the node set and edges set respectively. A *path* between two nodes is the set of adjacent edges that join them, and its length is the number of such edges. Let $R(v_A, v_B) \in E$ denote a path between v_A and v_B, such that $R(v_A, v_B) \neq R(v_B, v_A)$. We also assume that G is a dynamical network, that is its nodes and edges might increase or decrease as time progresses. Namely, if we discretise time into equal intervals $t = t_1, t_2, \ldots, t_k, \ldots$, we then write $G_1(V_1, E_1), G_2(V_2, E_2), \ldots, G_k(V_k, E_k), \ldots$ as the corresponding states of the network. Define the set of all directed paths from v_A to v_B as $\mathcal{R}^k(v_A, v_B) = \{R^k(v_A, v_B)\}$, where $R^k(v_A, v_B)$ are the directed paths between v_A and v_B at time $t = t_k$, and let $\mathcal{P}^k$ be the set of all paths in G_k at time $t = t_k$. Let $\tilde{\mathcal{R}}^k(v_A, v_B) = |\{\mathcal{R}^k(v_A, v_B)\}|$, and define

$$p_n(v_A, v_B) = \frac{\tilde{\mathcal{R}}^n(v_A, v_B)}{|\mathcal{P}^n|}. \tag{1}$$

Using a similar approach, we also have

$$p_n(v_B, v_A) = \frac{\tilde{\mathcal{R}}^n(v_B, v_A)}{|\mathcal{P}^n|}. \tag{2}$$

Note that $p_n(v_A, v_B)$ is the probability of choosing a path between v_A and v_B over all the possible paths in the network.

Assume that at time $t = t_{n+1}$ we add or remove $c_{n+1}(v_A, v_B)$ paths from $\mathcal{R}^{n+1}(v_A, v_B)$, and $c_{n+1}(v_B, v_A)$ paths from $\mathcal{R}^{n+1}(v_B, v_A)$. Note that $c_{n+1}(v_A, v_B)$ and $c_{n+1}(v_B, v_A)$ are negative quantities if we remove paths from v_A to v_B or v_B to v_A respectively.

Now we have

$$p_{n+1}(v_A, v_B) = \frac{\tilde{\mathcal{R}}^n(v_A, v_B) + c_{n+1}(v_A, v_B)}{|\mathcal{P}^{n+1}|}, \tag{3}$$

and

$$p_{n+1}(v_B, v_A) = \frac{\tilde{\mathcal{R}}^n(v_B, v_A) + c_{n+1}(v_B, v_A)}{|\mathcal{P}^{n+1}|}. \tag{4}$$

Clearly, we also have

$$|\mathcal{P}^{n+1}| = |\mathcal{P}^n| + c_{n+1}(v_A, v_B) + c_{n+1}(v_B, v_A). \tag{5}$$

Also note that for a large and non-sparse network, we have that $|\mathcal{P}^{n+1}| = |\mathcal{P}^n| + c_{n+1}(v_A, v_B) + c_{n+1}(v_B, v_A) \approx |\mathcal{P}^n|$, if $c_{n+1}(v_A, v_B) + c_{n+1}(v_B, v_A)$ is not too large, and we can use Equations 1 and 2 to find the following ratios

$$\frac{p_{n+1}(v_A, v_B)}{p_n(v_A, v_B)} = 1 + \frac{c_{n+1}(v_B, v_A)}{\tilde{\mathcal{R}}^n(v_A, v_B)}. \tag{6}$$

and

$$\frac{p_{n+1}(v_B, v_A)}{p_n(v_B, v_A)} = 1 + \frac{c_{n+1}(v_B, v_A)}{\tilde{\mathcal{R}}^n(v_B, v_A)}. \tag{7}$$

Let us consider the difference between p_n and p_{n+1}, whilst considering the direction of the paths. Therefore we can define two different quantities, namely

$$\begin{aligned} \Delta_{n+1}(v_A, v_B) = p_{n+1}(v_A, v_B) - p_n(v_A, v_B) = \\ = \frac{\tilde{\mathcal{R}}^n(v_A, v_B) + c_{n+1}(v_A, v_B)}{|\mathcal{P}^{n+1}|} - \frac{\tilde{\mathcal{R}}^n(v_A, v_B)}{|\mathcal{P}^n|} \\ \approx \frac{c_{n+1}(v_A, v_B)}{|\mathcal{P}^n|}, \end{aligned} \tag{8}$$

and

$$\begin{aligned} \Delta_{n+1}(v_B, v_A) = p_{n+1}(v_B, v_A) - p_n(v_B, v_A) = \\ = \frac{\tilde{\mathcal{R}}^n(v_B, v_A) + c_{n+1}(v_B, v_A)}{|\mathcal{P}^{n+1}|} - \frac{\tilde{\mathcal{R}}^n(v_B, v_A)}{|\mathcal{P}^n|} \\ \approx \frac{c_{n+1}(v_B, v_A)}{|\mathcal{P}^n|}. \end{aligned} \tag{9}$$

A. Assessing the Direction of the Influence

In this section, we will discuss a simple approach to determine which direction the influence exhibits. Since we are considering a dynamical network G, the analysis of its behaviour is crucial in understanding the above.
Let $\Delta_{n+1} = \Delta_{n+1}(v_A, v_B) + \Delta_{n+1}(v_B, v_A)$. Call

$$\delta_{n+1}(v_A, v_B) = \frac{\Delta_{n+1}(v_A, v_B)}{\Delta_{n+1}}.$$

Note that if $\delta_{n+1}(v_A, v_B) = 1$, then all the paths are either removed or added from v_A to v_B. The same applies for $\delta_{n+1}(v_B, v_A)$. Let $D = \dfrac{\delta_{n+1}(v_A, v_B)}{\delta_{n+1}(v_B, v_A)}$. Note the following

1) $\delta_{n+1}(v_A, v_B) + \delta_{n+1}(v_B, v_A) = 1$,
2) In order to have a higher proportion of paths with direction from v_A to v_B, at time $t = n + 1$, we need to have the one of the following conditions satisfied
    a) $D > 1$, $\Delta_{n+1}(v_A, v_B) > 0$, $\Delta_{n+1}(v_B, v_A) > 0$;
    b) $\Delta_{n+1}(v_A, v_B) > 0$, and $\Delta_{n+1}(v_B, v_A) < 0$;
    c) $D < 1$, $\Delta_{n+1}(v_A, v_B) < 0$, $\Delta_{n+1}(v_B, v_A) < 0$.

The above conditions can suggest the direction of the influence, namely if they are satisfied, then we assume the influence is from v_A to v_B. This will be used in the evaluation as described in Section IV.

Furthermore, consider the average of all the differences $(\delta_k(v_A, v_B) - \delta_0(v_A, v_B))$, for $k = 1, \ldots n$

$$\begin{aligned} \tilde{\delta}_n(v_A, v_B) = \frac{1}{n}[(\delta_1(v_A, v_B) - \delta_0(v_A, v_B)) + \cdots + \\ +(\delta_n(v_A, v_B) - \delta_0(v_A, v_B))] = \\ = \left(\frac{1}{n}\sum_{i=1}^{n} \delta_i(v_A, v_B)\right) - \delta_0(v_A, v_B). \end{aligned} \tag{10}$$

In other words, $\tilde{\delta}_n(v_A, v_B)$ gives the average of the variations of all the instances $\delta_k(v_A, v_B)$, with respect to $\delta_0(v_A, v_B)$. We can re-write Equation 10 recursively as follows

$$\begin{aligned} & \tilde{\delta}_{n+1}(v_A, v_B) \\ = & \frac{1}{n+1}\left[n\tilde{\delta}_n(v_A, v_B) + \delta_{n+1}(v_A, v_B) - \delta_0(v_A, v_B)\right], \end{aligned} \tag{11}$$

which enables us a continuous assessment of the trend of $\delta_n(v_A, v_B)$, so that we can further understand (and decide) which direction the influence exhibits. In this paper, we have not attempted to provide a definitive and complete set of steps to unequivocally determine the direction of the influence between v_A and v_B. Rather, the above methods only suggests the direction it might have over a number of iterations. Even though this approach has its limitations, mainly due to the over-simplification of the overall semantic network, in Section IV we will discuss a preliminary evaluation which shows the potential of this method.

B. Weighting the Influence

Once the direction of the influence has been determined, assessing its strength is the most crucial part of our approach. The longer the path between two nodes, the less strong the corresponding relation is. As a consequence, we need to introduce a weight which relates to the length of each path between v_A and v_B. This is certainly a well known and exploited property.

However, another important aspect in assessing the strength of a relation is the degree of all the nodes along the paths connecting v_A and v_B. Intuitively, a path with highly connected nodes may suggest that is somehow not so "unique" and its strength dissipates across all the connections. Let $\deg(v_k) = |\{e_{k,z} : e_{k,z} \in E\}|$ be the degree of the node v_k, and $\mathcal{P}_l^n$ be the set of all paths with length l at time $t = t_n$. We can thus re-write (1)

$$p_{n,l}(v_A, v_B) = \lambda_l \frac{\tilde{\mathcal{R}}_l^n(v_A, v_B)}{|\mathcal{P}_l^n|} d_{n,l}(v_A, v_B), \tag{12}$$

where $\tilde{\mathcal{R}}_l^n(v_A, v_B)$ is the set of all directed paths from v_A and v_B with length l, and $\lambda_m \geq \lambda_k$ if $m < k$, and

$$\begin{aligned} & d_{n,l}(v_A, v_B) \\ = & \frac{1}{|\mathcal{P}_l^n|}\sum_{i=1}^{\gamma_{n,l}}\left(\frac{\alpha_{v_{i_0}}}{\deg(v_{i_0})} + \sum_{k=1}^{l-1}\frac{\alpha_{v_{i_k}}}{\deg(v_{i_k}) - 1}\right), \end{aligned} \tag{13}$$

where

- $\gamma_{n,l} = \tilde{\mathcal{R}}_l^n(v_A, v_B) + c_{n,l}(v_A, v_B)$ is the total number of directed paths from v_A to v_B at time $t = t_n$,
- $c_{n,l}(v_A, v_B)$ is the number of newly added, or removed, paths of length l between the two nodes, and
- $\alpha_{v_{i_s}} > 0$ for $s = 0, \ldots, k$ are scaling constants.

We can easily see that

$$0 < d_{n,l}(v_A, v_B) \leq d_{n,l}^{MAX}(v_A, v_B), \tag{14}$$

where $d^{MAX}_{n,l}(v_A, v_B) = \sum_{i=1}^{\gamma_{n,l}} \left(\sum_{k=0}^{l-1} \alpha_{v_{i_k}} \right)$. Similarly, we call $p^{MAX}_{n,l}(v_A, v_B)$ when (12) is evaluated using $d^{MAX}_{n,l}(v_A, v_B)$, rather than $d_{n,l}(v_A, v_B)$.

Note that same reasoning applies to $p_{n,l}(v_B, v_A)$, and Equations 12, 13, and 14 can be modified accordingly. Let

$$W_n = \frac{\sum_l \left(p_{n,l}(v_A, v_B) + p_{n,l}(v_B, v_A)\right)}{\sum_l \left(p^{MAX}_{n,l}(v_A, v_B) + p^{MAX}_{n,l}(v_B, v_A)\right)}. \quad (15)$$

Clearly, $0 < W_n \leq 1$, and in this paper we will use W_n to assess the influence between v_A and v_B at time $t = t_n$, where values close to 1 suggest a strong influence, whereas values near 0 a weak one. Therefore, we assume that if $W_n > 0.5$ and $\tilde{\delta}_n(v_B, v_A) > 0$, then the corresponding influence is regarded as *present*. Otherwise we say it is *absent*. We acknowledge this does not fully encapsulate all the properties of influence, being based on multi-disciplinary issues. However, we believe the above reflects some important general features, which we are planning to expand and investigate in our ongoing research in this field.

III. Preliminary Evaluation

In our initial formulation of the problem under investigation, we have made some perhaps over-simplistic assumptions to initiate a preliminary evaluation. More specifically, we defined two groups of people, group A and group B, with 100 individuals in each of them. We then defined some semantic relations according to the following rules

- Every individual has a `part_of` relation to the group he/she belongs to
- We start off with 20 generic `social_interaction` relations among the members of each group.
- At each time iteration, extra connection among people are either added or removed with a fixed probability p.

The aim is to assess the influence (if any) between group A and group B. We then run 50 iterations and applied our method to assess the properties of the influence between the two nodes. We then manually performed the same task by analysing 5 different randomly generated networks, and evaluated the influence at each iteration. For this preliminary evaluation, we have assumed that λ_l's and α_{v_i}'s in the above equations are all 1.

Table II, compares the manual results with the evaluation carried out with our method. Note that we have only included the direction of the influence if its existence is suggested. We can see that out of the 5 different networks, only the second one has not been identified correctly. Note also for that particular case, $W_{50} = 0.5461$ suggesting that perhaps 0.5 is not the best cut off value for W_n to indicate the existence of an influence.

Table I
Results of the simulation as discussed in Section IV

Network	Paths	D	W
1	1686	0.8036	0.4564
2	1117	0.8707	0.5461
3	3559	10.0431	0.3960
4	1928	0.3271	0.7408
5	36	2.2778	0.6186

Network	$\delta_{50}(v_A, v_B)$	$\delta_{50}(v_B, v_A)$	$\tilde{\delta}_{50}(v_A, v_B)$
1	positive	negative	-2.45
2	positive	positive	34.143
3	positive	negative	14.564
4	negative	positive	-46.997
5	positive	negative	13.109

Table II
Comparison of results from the two evaluations

Network	Influence with Direction	Manual Evaluation
1	Absent	No
2	**Present,** $v_B \rightarrow v_A$	**No**
3	Absent	No
4	Present, $v_B \rightarrow v_A$	Yes, from v_B to v_A
5	Present, $v_A \rightarrow v_B$	Yes, from v_A to v_B

IV. Discussion

The results discussed above clearly show the potential of our method.

At this stage, we have not carried out a full evaluation on real data. However, we have considered a small, but real, data-set based on a similar setting as above. Namely, we had 40 interconnected individuals split into two groups equal size, group A and group B. It was assumed that we had 5 iterations throughout which connections were gradually added according to the topology of the network generated by their connections, so that at the end, all of them were included.

We then applied our method to assess the existence of the influence (if any) between groups A and group B. The outcome was an influence from group B to group A. We then asked each individual to determined whether an influence was actually present, and its direction, to the best of their knowledge. About 61% of people agreed with our result. Again, we assumed that all λ_l's and α_{v_i}'s are all equal to 1. We acknowledge that this is rather an artificial and perhaps inaccurate assumption which might also explain the fact we obtained a precision of 61%.

In future research, we aim to generalise our approach to address the full dynamics of networks. In particular, understanding and predicting how connections change, and how quickly they do so, would enable a better analysis of the system. Furthermore, a full validation will be carried out based on both computer simulations and real data-sets.

V. Conclusion

In this paper we have introduced and discussed a method of analysing and determining the direction and strength of the influence between two nodes in a semantic network. Although there is a wealth of literature on this topic, our motivation was to provide a tool to facilitate the assessment of any influence, including its direction, between semantic entities.

Despite we have only discussed an initial implementation of our method, it is clear that it shows potential in a variety of applications. It is also worth mentioning that we are planning to carry out a full assessment of the computational efficiency of our approach. Preliminary results (not discussed in this paper) appear to support the efficiency of our approach and show clear prospect to be fully developed as a lean and scalable computer system, which can be implemented and applied in a variety of intelligent and knowledge discovery systems.

References

[1] Bollobás B Modern Graph Theory. *Graduate Texts in Mathematics*, Vol. 184, Springer, New York, 1998.

[2] Carnap R Introduction to Semantics: And Formalization of Logic. *Harvard University Press*, Vol. 1. , 1959.

[3] Diehl C P, Namata G, and Getoor L Relationship Identification for Social Network Discovery. *Proceedings of the 22Nd National Conference on Artificial Intelligence*, Vol 1, 2007.

[4] Ebel H, Mielsch L I and Bornholdt S Scale-free Topology of E-mail Networks. *Phys. Rev.* E 66, 035103, 2002.

[5] Girju R and Moldovan D I Text Mining for Causal Relations. *FLAIRS Conference*, 2002.

[6] Lehmann F Semantic Networks in Artificial Intelligence. *Elsevier Science Inc.*, 1992.

[7] Lewis D Causation. *The Journal of Philosophy*, 70(17), 1973

[8] Newman M E J and Park J Why Social Networks Are Different From Other Types of Networks. *Physical Review E.* 68(3) 36122, 2003.

[9] Sanchez-Graillet O and Poesio M Acquiring Bayesian Networks from Text. *LREC*, European Language Resources Association, 2004.

[10] Xiang R, Neville J, and Rogati M Modeling Relationship Strength in Online Social Networks. *Proceedings of the 19th international conference on World wide web (WWW '10)*, 2010.