Similarity for Natural Semantic Networks

Francisco Torres Sara E. Garza

Abstract

A natural semantic network (NSN) represents the knowledge of a group of persons with respect to a particular topic. NSN comparison would allow to discover how close one group is to the other in terms of expertise in the topic— for example, how close apprentices are to experts or students to teachers. We propose to conceive natural semantic networks as weighted bipartite graphs and to extract feature vectors from these graphs for calculating similarity between pairs of networks. By comparing a set of networks from different topics, we show the approach is feasible.

1 Introduction

Knowledge representation structures implicit information and turns it into a valuable asset. The representation technique of *natural semantic networks* or NSN's, specifically, reflects the knowledge of a population with respect to a topic or domain by gathering responses from a sample group. When the group profiles are distinct, the responses and resulting networks tend to be different; measuring similarity between NSN's not only would allow to quantify this difference, but would eventually lead to the assessment of the knowledge level of a given group with respect to the experts of the domain (e.g. how close an apprentice is to a master in the field or an undergraduate to a teacher or graduate student). We propose an approach for NSN similarity calculation that is based on graph theory and document similarity; this approach, which considers both content and structure from the network, views the NSN as a bipartite graph and extracts a weighted feature vector for comparison.

The rest of this paper is organized as follows: Section 2 offers pertinent background and Section 3 briefly describes related work; our approach is explained in Section 4 and results are provided in Section 5. Section 6, finally, offers closing remarks and future work.

2 Background

This section introduces necessary vocabulary, notation, and formulas for natural semantic networks, bipartite graphs, and document similarity.

2.1 Natural semantic networks

Natural semantic networks (NSN's), introduced by Figueroa et al. [6], study long-term memory by gathering a socio-cognitive perspective on a given topic. To generate a natural semantic network, a set P of *participants* (20-40) is given a set C of *target concepts* (6-10). For every $c \in C$, each participant must provide a set of individual words that come to mind when c is presented; these words are known as *definers*. The participant must also score each definer (using a scale 1-10) according to its importance within the target concept. Let us formally denote the score of participant p for definer d_i in concept c_k as $sc_k^i(p) \in \{1 \dots 10\}$.

The total score of a definer within a given concept is known as its m-value; given d_i and c_k, this value is calculated as $m_k^i = \sum_{p \in P} sc_k^i(p)$. The ten definers with the highest m-value make up a concept's *SAM group*, where "SAM" stands for "Semantic Analysis of M-value" [7]. Let us note that a definer can be in more than one SAM group; it is thus possible to have not a single but a set of m-values for a particular definer. This also gives rise to another important metric: the f-value of a definer. The f-value is simply the number of times that the definer appears in the network. For d_i, we denote this value as f_i. A fragment of an NSN is shown in Table 1.

Table 1: Fragment of two SAM groups in a natural semantic network.

Ecology		
F	**Definer**	**M**
1	Recycle	50
2	***Nature***	30
2	***Animals***	20
1	Plants	10

Environment		
F	**Definer**	**M**
2	***Nature***	100
2	***Animals***	70
1	Water	60
1	Trees	50

2.2 Bipartite graphs

The mathematical representation for a network is a *graph*. A graph $G = (V, E)$ consists of a set V of entities known as *vertices* and a set E of connections known as *edges*. If the edges are assigned numerical weights, the graph is said to be *weighted*. A *bipartite graph*[1] is graph whose vertex set V is divided into two disjoint subsets V_1 and V_2 and whose set E only contains edges that join vertices from different subsets. A classical example of a bipartite graph is the *actor-movie* network, where the vertex subsets are conformed by actors and movies, and where each edge indicates an actor participating in a movie [8].

From a bipartite graph, it is possible to extract two *projections* or unipartite graphs (e.g. a projection where only movies are vertices and edges represent common actors between them). Formally, in a projection $G_P = (V_P, E_P)$ of a bipartite graph $G_B = (V_B, E_B)$ where $V_P \subset V_B$,

[1] Let us note that any graph with a single vertex set is called *unipartite* or *monopartite*.

$$E_P = \{\{u, v\} : (u, v \in V_P) \wedge (\{u, w\}, \{v, w\} \in E_B) \wedge (w \in V_B \setminus V_P)\}.$$

2.3 Document similarity with the vector space model

The *vector space model* of information retrieval views a document as a *bag of words* where order is not important and extracts a *weight vector* from this bag; each vector's length is equal to the size of the document collection's vocabulary (unique words), and each weight represents the importance of a particular vocabulary word in the document (0 if the word is not present). A common metric for calculating similarity between document vectors is the *cosine similarity* [1]:

$$\text{cosim}(a, b) \quad = \quad \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \times \left|\vec{b}\right|}, \tag{1}$$

where a and b are the documents, $\vec{a}$ and $\vec{b}$ are the vectors, and $w_{i,a}$ and $w_{i,b}$ are the weights for word i in, respectively, $\vec{a}$ and $\vec{b}$. A similarity of 0 indicates that the documents have no common words and a similarity of 1 indicates that the documents are identical.

3 Related work

Network comparison is inherently related to *graph matching* [3], which can be *exact* or *inexact*. While the first addresses problems related to *graph isomorphisms* (detecting if two graphs are equal), the second attempts to provide the number of operations needed to turn one graph into another (*graph edit distance*) or a degree of resemblance between graphs (*graph similarity*). Our work and related work fall into this last category.

The works by Dehmmer and Emmert [5] and Qureshi et al. [9] both extract feature vectors for calculating graph similarity; while the former utilizes vertex degree (i.e. the number of conected edges), the latter uses statistical and symbolic features for object recognition. Meanwhile, the approach by Champin and Solnon [4] first obtains different mappings for the pair of graphs and then computes similarity with a psychologically-sustained metric. With regard to semantic data similarity, Bergmann and Gil [2] focus on semantic workflow retrieval by building graphs with different types of vertices and edges; on the other hand, Sanchez et al. [10] compare the NSN's of two distinct groups by means of an index that calculates the ratio of common edges with respect to the total amount possible (similar to the Jaccard index).

4 Measuring similarity for natural semantic networks

Our approach consists of calculating NSN pairwise similarity by compacting the networks into weighted feature vectors and obtaining cosine similarity for these vectors. Each feature is given either by a vertex or an edge of the networks, and each weight represents the importance of that feature. Because the nucleus of an NSN is given by its definers (target concepts are usually fixed along networks for the same topic), we represent the NSN as a graph where each vertex is a definer and each edge is the similarity or closeness between a pair of these. To determine which definers are related, as well as their closeness, we consider that the definer graph is a projection from a concept-definer *weighted bipartite graph*. In this other graph, there exists an edge between a concept and a definer when the latter belongs to the SAM group of the former; the weight of the edge is simply the m-value of the definer in that group.

In the definer projection, there is an edge between definers if these are found together in one or more SAM groups. To calculate edge weights, we assume that definers are closer or more similar to each other if the difference in their m-values is small. As a result, we first compute the *relative difference* between definers d_a and d_b for the SAM group of a concept c as

$$\delta_r(m_c^a, m_c^b) = \left| \frac{m_c^a - m_c^b}{m_c^{\max} - m_c^{\min}} \right|, \tag{2}$$

where $m_c^{\max}$ and $m_c^{\min}$ are, respectively, the maximum and minimum m-values of the group. Since the difference between definers is actually a *distance*, we obtain relative similarity by taking the complement of $\delta_r(m_c^a, m_c^b)$:

$$\mathrm{sim}(m_c^a, m_c^b) = 1 - \delta_r(m_c^a, m_c^b). \tag{3}$$

Also, because one same pair of definers can appear in several groups, we calculate the overall similarity between d_a and d_b as the average of their relative similarities in the set $C_{a,b} \subseteq C$ of SAM groups that contains both of them. An edge weight $w_{a,b}$ is, therefore, calculated with

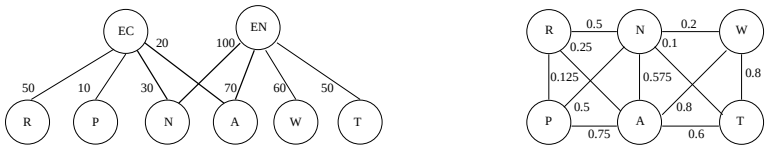
$$w_{a,b} = \frac{\sum_{c \in C_{a,b}} \mathrm{sim}(m_c^a, m_c^b)}{|C_{a,b}|}. \tag{4}$$

Since a weight of 0 typically indicates the absence of an edge, we set $\mathrm{sim}(m_c^a, m_c^b)$ as half of the second lowest similarity in c's group when the numerator of Eq. 2 is $m_c^{\max} - m_c^{\min}$. To illustrate these calculations, an example of the bipartite and definer graphs (extracted from Table 1) is given by Figure 1.

Every edge weight of the definer graph will also become a weight that corresponds to an edge feature in the NSN's feature vector. Regarding vertex features, the weight is given by the relative f-value of the definer, denoted as ϕ_a for d_a:

$$\phi_a = \frac{f_a}{f_{\max}}, \tag{5}$$

where $f_{\max}$ is the highest f-value found in the network.



(a) Bipartite view of the NSN. (b) Definer projection.

Figure 1: NSN as a graph. Note that EN and EC are concepts, while R, N, A, P, W, and T are definers.

5 Results

With the intent of showing how the proposed approach handles objects that are expected to be similar (networks from the same topic) and objects that are expected to be dissimilar (networks from different topics), we built a *similarity matrix* with a set of natural semantic networks from different topics; these networks were made available by a research group at the authors' university [11, 12]. The four topics covered by these networks are: *ecology* (ec1-ec6), *sentimental relationships* (lov1-lov4), *ethics* (eth1, eth2), and *scientific skills* (sk). The resulting matrix is depicted in Figure 2, where networks from the same topic were placed adjacent to each other (i.e. in blocks).

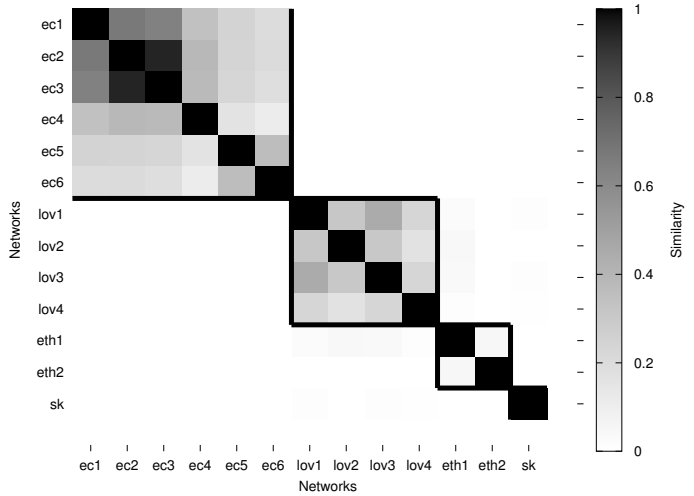


Figure 2: Similarity matrix.

We can clearly appreciate in the matrix the expected block-diagonal pattern, which indicates that similarity within the same topic (0.23 on average) is higher than similarity between different topics (0.005 on average).

6 Conclusions and Future work

We have presented an approach for measuring similarity between natural semantic networks. The approach, which uses both content and structure, views each network as a concept-definer bipartite graph and extracts the definer projection from this graph to create a weighted feature vector; vectors are compared using cosine similarity. Future work includes the use of *fuzzy graphs* for visualizing specific differences between the networks.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. ACM Press, New York, NY, USA, 1999.

[2] Ralph Bergmann and Yolanda Gil. Retrieval of semantic workflows with knowledge intensive similarity measures. In *Case-Based Reasoning Research and Development*, pages 17–31. Springer, 2011.

[3] Horst Bunke. Graph matching: Theoretical foundations, algorithms, and applications. In *Proc. Vision Interface*, volume 2000, pages 82–88, 2000.

[4] Pierre-Antoine Champin and Christine Solnon. Measuring the similarity of labeled graphs. In *Case-Based Reasoning Research and Development*, pages 80–95. Springer, 2003. URL `http://liris.cnrs.fr/csolnon/publications/ICCBR03.pdf`. 86.

[5] Matthias Dehmer and Frank Emmert-Streib. Comparing large graphs efficiently by margins of feature vectors. *Applied mathematics and computation*, 188(2):1699–1710, 2007.

[6] J Figueroa, González G., and V. Solís. Una aproximación al problema del significado: las redes semánticas. *Revista latinoamericana de psicología*, 13 (3):447–458, 1981.

[7] Ernesto O Lopez and John Theios. Semantic analyzer of schemata organization (saso). *Behavior Research Methods, Instruments, & Computers*, 24 (2):277–285, 1992.

[8] M. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[9] Rashid Jalal Qureshi, J Ramel, Hubert Cardot, and Prachi Mukherji. Combination of symbolic and statistical features for symbols recognition. In *Signal Processing, Communications and Networking, 2007. ICSCN'07. International Conference on*, pages 477–482. IEEE, 2007.

[10] Martha Patricia Sánchez Miranda, Arturo de la Garza González, and Ernesto Octavio López Ramirez. Simulaciones computacionales sobre cuestiones ambientales en dos grupos de contraste. *Liberabit*, 19(2):223–233, 2013.

[11] Francisco Torres and R. López. Foraging information in webpages through meaning. *International Journal of Good Conscience*, 5(2):308–323, 2010. ISSN 1870-557X.

[12] María Elena Urdiales. *Sobre el esquema relacional de pareja en jovenes y adultos del área metropolitana de Monterrey*. PhD thesis, Universidad Autónoma de Nuevo León, 2009. URL `http://uanl.vtlseurope.com/lib/item?id=chamo:224600&theme=UANL`.