Unsupervised Semantic Field Analysis
by example of the Heat Equation

Special Topic on Networks

Candidate Number: 1072462

Abstract

This work will attempt to

1 Motivation

For non-periodic problem settings, Chebyshev series are a fantastic choice (Biemann 2006).

2 Introduction

Let $\mathbb{N} = \mathbb{Z}^+$ denote the positive integers and $N_0 := \{0\} \cup \mathbb{N}$ the nonnegative integers.

The methods we will discuss to identify semantic fields will be based on graph clustering algorithms applied to a text corpus word connectedness / neighbourhood network. As we will discuss later, different notions of connectedness can give us different insight into the structure of a natural language. We will focus our attention on methods for undirected graphs, "graphs without direction" (cf. Definition 2.1).

2.1 Definition: Undirected Graph

A graph $G = (V, E)$ with vertices V and edges $E \subseteq V \times V$ is undirected if and only if $(v_i, v_j) \in E \Rightarrow (v_j, v_i) \in E \quad \forall\, v_i, v_j \in V$.

Vertices are also often referred to as *nodes*. Every graph G is uniquely described by its adjacency matrix $A \in \{0,1\}^{n \times n}$ (Definition 2.2), which allows us to talk about "linear algebra" of graphs.

2.2 Definition: Adjacency Matrix

Let $A \in \{0,1\}^{n \times n}$ denote the symmetric adjacency matrix of an undirected graph $G = (V, E)$. Its entries are given by $a_{ij} = \{A\}_{ij} = \mathbb{1}_{(v_i, v_j) \in E}$, so $a_{ij} = 1$ if vertex v_i is connected to v_j and 0 otherwise.

By construction, $A = A^T$ is symmetric and has all-0s in the diagonal, a definition that corresponds to the fact that you cannot be friends with yourself in a social network.

Further let $m := |E|$ and $n := |V|$ denote the number of edges and vertices, respectively. The degree d_i of a vertex $v_i \in V$ is defined by the number of edges connecting to it, so

$$d_i := \deg(v_i) = \Big| \{(v_j, v_k) \in E \mid v_j = v_i\} \Big|,$$

for an undirected graph $G = (V, E)$. The handshaking lemma (Lemma 2.1) tells us an important fact useful for normalisation.

2.1 Lemma: Handshaking

For every finite, undirected graph $G = (V, E)$ the individual vertex degrees sum up to exactly twice the number of edges, so

$$\sum_{i=1}^{n} d_i = \sum_{v \in V} \deg(v) = 2m\,.$$

The individual vertex degrees can be summarised in the so-called *degree matrix* $D := \operatorname{diag}(d_1, ..., d_n)$, $D \in \mathbb{N}_0^{n \times n}$. The graph *Laplacian* is defined by $L := D - A$.

Given a graph, we are interested in performing graph clustering, also referred to as community detection or graph partitioning, the goal of which is to obtain a set of mutually exclusive clusters $C_i \subseteq V$ (cf. Definition 2.3).

2.3 Definition: Graph Clustering

Let $C = \{C_i \subseteq V\}_{i=1\ldots n_C}$ denote a clustering of $G = (V, E)$ into $n_C \in \mathbb{N}$ clusters where $C_i \cap C_j = \{\}\ \forall\, i, j \in \{1, ..., n_C\}$ and $\bigcup_{i=1}^{n_C} C_i = V$. Let $s_i \in \{1, ..., n_C\}$ denote the assigned cluster of vertex $v_i \in V$.

These clusterings may be better or worse depending on the context, but a generally solid measure of "clustering goodness" is *modularity* (Definition 2.4).

2.4 Definition: Modularity

For a given undirected graph $G = (V, E)$ and clustering C, let

$$Q := \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(s_i, s_j)\,,$$

with $\delta(\cdot, \cdot)$ the Kronecker delta indicating whether two vertices v_i and v_j belong to the same cluster (Grindrod and Lambiotte 2022).

Modularity is a measure of the quality of a clustering (also referred to as a partitioning) of G. It can also be written as

$$Q = \frac{1}{2m} \sum_{c=1}^{n_C} \left[\sum_{v_i, v_j \in C_i} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \right]\,,$$

which might make its purpose a bit clearer.

3 Clustering Algorithms

3.1 Girvan-Newman

3.2 Louvain

Grindrod and Lambiotte 2022. Blondel et al. 2008.

3.3 Chinese Whispers

The *Chinese Whispers* algorithm due to Biemann 2006

```
Input: an undirected graph G = (V, E).
Output: a graph clustering C = {C_i}_{i=1,...,n_C} into n_C classes.

Initialise the algorithm with n_C = n classes, one per vertex.
while there are changes or the iteration maximum is reached, do
  for v_i in shuffle(V), do
    Set s_i = 3
  end
end
```

Corresponds to an agent-based simulation of a social network Biemann 2006.

3.4 Watset

Ustalov et al. 2019.

3.5 Spectral Clustering Methods

Fortunato 2010.

3.6 Fiedler

Fortunato 2010.

4 Author analysis?

Authors whose works circulate around these semantic fields: bla bla, maybe not that interesting

5 Discussion and Outlook

Bla

References

Biemann, Chris (June 2006). 'Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems'. In: *TextGraphs: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, pp. 73–80. DOI: 10.5555/1654758.1654774.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre (Oct. 2008). 'Fast unfolding of communities in large networks'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Fortunato, Santo (2010). 'Community detection in graphs'. In: *Physics Reports* 486.3, pp. 75–174. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2009.11.002.

Grindrod, Peter and Renaud Lambiotte (5th Nov. 2022). *C5.4 Networks*. Lecture Notes for the course.

Ustalov, Dmitry, Alexander Panchenko, Chris Biemann and Simone Paolo Ponzetto (Sept. 2019). 'Watset: Local-Global Graph Clustering with Applications in Sense and Frame Induction'. In: *Computational Linguistics* 45.3, pp. 423–479. ISSN: 0891-2017. DOI: 10.1162/coli_a_00354.

A Appendix Things?