

Aplicaciones de la Teoría de Redes al Procesamiento del Lenguaje Natural

Alvaro J. Riascos Villegas

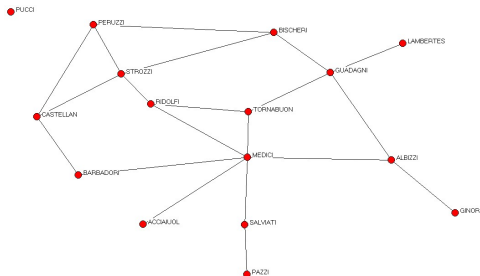
Febrero, 2016

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia
- 3 Redes de Asociaciones
- 4 Aplicaciones
- 5 Redes Semánticas y Sintácticas
- 6 Construcción de Redes

Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Padgett, J.F y C.K. Ansell (1993). Robust action and the rise of the Medici, 1400-1434.
- Red de matrimonios entre familias (cada enlace representa un matrimonio entre miembros de dos familias).



Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Basados en la riqueza y poder político es difícil explicar como los Medici surgieron como una familia tan importante (la familia Strozzi tenía más riqueza y poder político, sin embargo fueron opacados por los Medici).
- La estructura de relaciones puede ser un determinante.
- Si comparamos con cuántas familias se encuentra una familia específica relacionada y comparamos entre ellas, los Medici sobresalen (3 a 2).
- Una relación de cercanía resulta más sugestiva.

Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Sea $P(ij)$ el número de caminos más cortos que conectan una familia i con j . Sea $P_k(ij)$ el número de estos caminos que incluyen a la familia k .
- Por ejemplo si $i = \text{Barbadori}$, $j = \text{Guadagni}$, entonces $P(ij) = 2$. Si $k = \text{Medici}$ entonces $P_k(ij) = 2$ mientras que si $k = \text{Strozzi}$ o Albizzi $P_k(ij)$ es cero o uno respectivamente.

Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Si calculamos una medida de importancia (betweenness Freeman) de cada familia k como:

$$\sum_{i,j:i \neq j, k \in \{i,j\}} \frac{\frac{P_k(ij)}{P(ij)}}{(n-1)(n-2)/2} \quad (1)$$

donde $\frac{P_k(ij)}{P(ij)} = 0$ si no hay caminos entre i y j . El coeficiente $(n-1)(n-2)/2$ es el número máximo de pares de familias que incluirían a la familia k .

Ejemplos de Redes: Familia Medici, Florencia 1400

- Esta medida de poder para los Medici es 0.522. Esto significa que los Medici están en más de la mitad de los caminos más cortos entre todos los caminos más cortos entre cada par de familias.
- Este mismo cálculo para los Strozzi es 0.103. El segundo más alto es los Guadagni que es 0.255.
- En este sentido los Medici estaban mejor posicionados que cualquier otra familia.
- Esta estructura es endógena? Es óptima?

Ejemplos de Redes: Amistades y romances en estudiantes secundaria

- Datos de 90,000 estudiantes de la encuesta Add Health entrevistados en los años 90.
- A los estudiantes se les preguntaba con quién habían tenido relaciones románticas en los últimos seis meses.

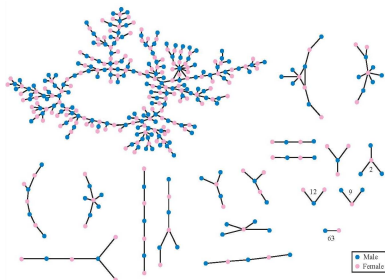


Figure 1.2: A Figure from Bearman, Moody and Stovel [47] based the Add Health Data Set. A Link Denotes a Romantic Relationship, and the Numbers by Some Components Indicate How Many Such Components Appear.

Ejemplos de Redes: Formación aleatoria de redes

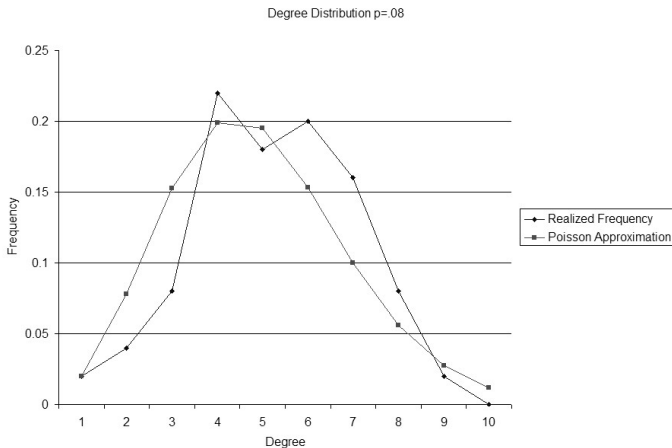


Figure 1.7: Frequency Distribution of a Randomly Generated Network and the Poisson Approximation for a Probability of .08 on each Link

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia**
- 3 Redes de Asociaciones
- 4 Aplicaciones
- 5 Redes Semánticas y Sintácticas
- 6 Construcción de Redes

Redes de Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algún tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o párrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470,000$, $m = 170mm$).

Redes de Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algún tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o párrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470,000$, $m = 170mm$).

Redes de Coocurrencia

- Son redes que se construyen usando como nodos palabras (o solamente las que pertenecen algún tipo de palabra) y creando un enlace entre ellas cuando aparecen juntas en una ventana de N-palabras (puede ser en una frase o párrafo).
- Pueden ser redes dirigidas o no y con pesos (usando la verosimilitud, etc).
- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470,000$, $m = 170mm$).

Redes de Coocurrencia

- La distribución del grado sigue una power law con parámetro $-1,5$, $-2,7$. Esta regularidad empírica sirve para validar una teoría: existe un kernel lexicon del inglés que comparten la mayoría de personas (50m palabras) y una lista de palabras menos usadas.

Redes do Coocurrencia: Ejemplos

- Ferrer-i-Cancho y Sole (2001) muestran que este tipo de redes tienen características de mundos pequeños. Diametro de aprox. 2,65 y coeficiente de clustering 0,5 y distr. libre de escala.
- Usan British National Corpus ($n = 470,000$, $m = 170mm$).
- La distribución del grado sigue una power law con parámetro $-1,5$, $-2,7$. Esta regularidad empírica sirve para validar una teoría: existe un kernel lexicon del inglés que comparten la mayoría de personas (50m palabras) y una lista de palabras menos usadas.

Redes de Coocurrencia: Ejemplos

- Masucci y Rodgers (2006). Palabras adyacentes en la novela de Orwell 1984.
- Encuentran evidencia fuerte de power laws.

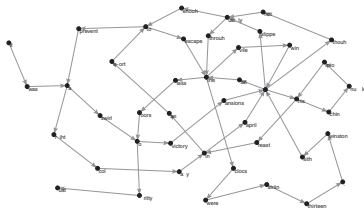


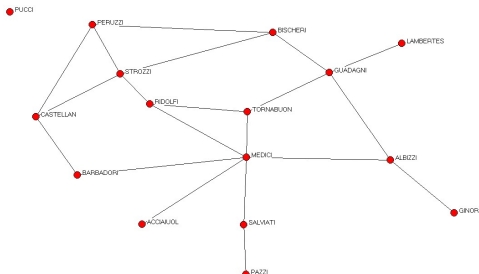
Figure 4.2. The graph corresponding to the beginning of *1984*.

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia
- 3 Redes de Asociaciones**
- 4 Aplicaciones
- 5 Redes Semánticas y Sintácticas
- 6 Construcción de Redes

Ejemplos de Redes: Matrimonios Familia Medici, Florencia 1400

- Padgett, J.F y C.K. Ansell (1993). Robust action and the rise of the Medici, 1400-1434.
- Red de matrimonios entre familias (cada enlace representa un matrimonio entre miembros de dos familias).



- Steyvers y Tenenbaum (2005): Experimento con seis mil personas. Se les dan 500 palabras y ellos declaran si hay alguna asociación entyre pares.

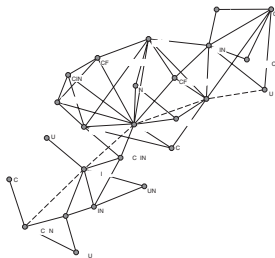


Figure 4.5. Free-word-association network (Steyvers and Tenenbaum 2005b).

- Steyvers y Tenenbaum (2005): Experimento con seis mil personas. Se les dan 500 palabras y ellos declaran si hay alguna asociación entre pares.

Table 4.2. *Properties of semantic networks.*

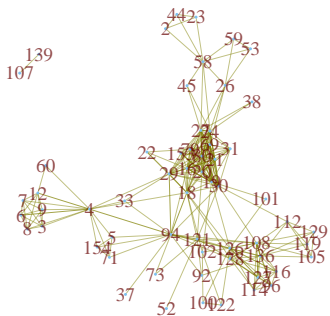
Property	Associative	Roget	WordNet
C	0.186	0.875	0.0265
C_r	4.35×10^{-3}	0.613	1×10^{-4}

The comparison is between the actual clustering coefficient and the clustering coefficient of a random graph of the same size.

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia
- 3 Redes de Asociaciones
- 4 Aplicaciones**
- 5 Redes Semánticas y Sintácticas
- 6 Construcción de Redes

Clasificación de textos: Grafo de Documentos



- An undirected bipartite graph is a triple $G = (D, W, E)$ where D, W are two sets of vertices and E is the set of edges. Take D as the set of documents, W as the set of words they contain and an edge exists if word w occurs in document d . Note that the edges are undirected.
- In this model, there are no edges between words or between documents.
- Dhillon (2001) utiliza un algoritmo de particion de redes bipartitas.
- Tiene la ventaja de que usa información sobre las palabras y documentos para hacer la clasificación

- Detección de plagio.
- The concrete task is to implement a graph edit distance algorithm, which calculates similarity between two graphs.
- The algorithm is based on calculating the number of edit operations needed to transform one graph into another (Riesen and Bunke, 2009).

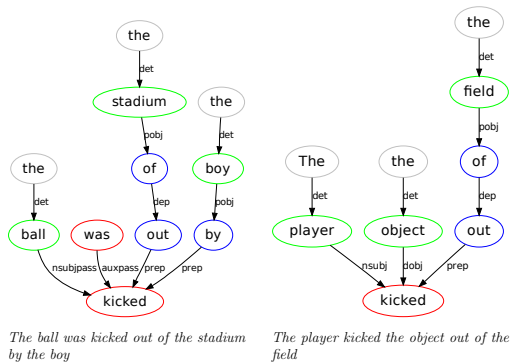


Figure 1.1: Example of dependency graphs

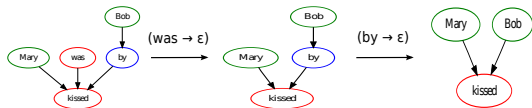


Figure 2.1: An example of edit operations for two graphs

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia
- 3 Redes de Asociaciones
- 4 Aplicaciones
- 5 Redes Semánticas y Sintácticas**
- 6 Construcción de Redes

Redes Semánticas y Sintácticas

- Para la próxima!

Contenido

- 1 Introducción: Teoría de Redes
- 2 Redes de Coocurrencia
- 3 Redes de Asociaciones
- 4 Aplicaciones
- 5 Redes Semánticas y Sintácticas
- 6 Construcción de Redes

Construcción de Redes

- Red de términos.
- Red de documentos.
- Red bipartita de términos y documentos.