Relación contextual de palabras en libros de Shakespeare usando mapas auto-organizados

Luis Alfredo Moctezuma[1], Jessica López[1], Caleb Jiménez[1], Maya Carrillo[1], Luis Enrique Colmenares-Guillen[1], J. Guadalupe Ramos[2]

[1] Benemérita Universidad Autónoma de Puebla, Puebla, México

[2] Instituto Tecnológico de La Piedad, La Piedad, Michoacán, México

luisalfredomoctezum@gmail.com,{cmaya,lecolme}@cs.buap.mx, {acissejol, calebji,joguar}@hotmail.com

Resumen. El significado de las palabras puede encontrarse por el contexto en el cual estas ocurren. Los contextos de las palabras pueden ser reflejados y agrupados mediante una red SOM (*Self- Organizing Map*). En los experimentos que se muestran en el presente trabajo, se aprecian los diferentes grupos de palabras, obtenidos al capturar información de contexto, empleando representaciones vectoriales de palabras, en tres obras de Shakespeare: Macbeth, Julio Cesar y Hamlet. Los resultados obtenidos muestran que los SOM funcionan de manera adecuada para identificar los diferentes conceptos que los autores manejan en sus obras.

Palabras clave: Redes neuronales, mapas auto-organizados, relación contextual, Shakespeare.

1. Introducción

El procesamiento de lenguaje natural (PLN) se ocupa del reconocimiento y utilización de la información expresada en lenguaje humano para ser empleada en sistemas computacionales. En su estudio intervienen diferentes disciplinas tales como: lingüística, inteligencia artificial, filosofía, matemáticas, psicología, y ciencia cognitiva. En PLN para analizar los textos generalmente se consideran cuatro niveles de análisis: léxico, sintáctico, semántico y pragmático. Si bien la mayoría de la técnicas están basadas en análisis léxico y sintáctico, es evidente la necesidad de comprender el significado de los textos para lograr elevar el desempeño de diversas tareas de PLN como análisis de sentimiento, la generación automática de reportes, recuperación de información, búsqueda de respuestas, por mencionar algunas. Sin embargo, las técnicas de representación del significado no han obtenido los resultados deseados, y numerosas cuestiones continúan sin encontrar soluciones satisfactorias.

Definir qué es el significado no es una tarea sencilla, y puede dar lugar a diversas interpretaciones. Es posible distinguir entre significado independiente y significado dependiente del contexto. El primero, tratado por la semántica, hace referencia al significado que las palabras tienen por sí mismas sin considerar el significado adquirido según el uso en una determinada circunstancia. Por otra parte, el componente significativo de una frase asociado a las circunstancias en que ésta se da, es estudiado por la pragmática y conocido como significado dependiente del contexto.

Un mapa auto-organizado (SOM) es una herramienta que analiza datos en muchas dimensiones con relaciones complejas entre ellos y los presenta en una visualización sencilla en sólo dos dimensiones. La propiedad más importante de una SOM es que preserva las propiedades topológicas de los datos. Dichos mapas pueden utilizarse para visualizar roles contextuales de las palabras, es decir similitud en su uso en contextos cortos formados por las palabas adyacentes, este es el objetivo del presente trabajo tomando como fuente obras de Shakespeare.

En diferentes aplicaciones, los datos de entrada a un SOM, son numéricos. Sin embargo al trabajar con palabras y considerar su contexto, el orden en el que aparecen dichas palabras es importante y no basta con representarlas por un número, como por ejemplo: frecuencia de términos o frecuencia inversa. En este trabajo, cada palabra se representó como un vector cuyas entradas representan una secuencia única de unos y ceros. El contexto de una palabra *a* fue capturado sumando los vectores únicos de las palabras adyacentes con las que aparecía a lo largo del texto, considerando una ventana simétrica de dimensión uno. Los resultados obtenidos muestran que la similitud entre palabras puede definirse indirectamente, capturando su significado en función del contexto en el que aparecen (palabras vecinas).

Este artículo está organizado de la siguiente manera en la sección 2 se introducen algunos trabajos relacionados, en la sección 3 se define lo que es un SOM y el procedimiento seguido para el aprendizaje de la red, en la 4 se describe la representación utilizada para las palabras, en la 5 las condiciones de operación del SOM, en la sección 6 se presentan los resultados y finalmente en la 7 las conclusiones.

2. Trabajo relacionado

En [1] se utilizar un SOM para organizar las palabras en categorías gramaticales representadas en una matriz bidimensional. La similitud de las categorías se refleja en función de su distancia sobre la matriz. Este tipo de mapa de categorías de palabras, puede utilizarse en aplicaciones para analizar colecciones grandes de documentos. En [3] se identifican categorías gramaticales empleando un corpus en chino, se emplean vectores de contexto de dimensión 650 y para capturar la información de contexto ventanas simétricas de dimensión 2.

3. Mapas auto-organizados

Un SOM es un tipo de red neuronal, útil para tareas de agrupamiento y auto-organización de grandes cantidades de datos de manera eficiente. T. Kohonen [8] los

presentó por primera vez en 1982. La principal característica de los SOM es que preservan las relaciones topológicas. Su finalidad es descubrir la estructura subyacente de los datos introducidos en él. Este tipo de red consiste de un conjunto de neuronas sobre una cuadricula de $N = \{n_1, n_2, \dots, n_k\}$ que se conectan de manera idéntica a la entrada *X*. La localización de cada neurona sobre la cuadricula está representada por el valor de entrada asociado. Para determinar si una neurona *i* está cerca de una neurona *j*, se calcula la distancia euclidiana, generalmente:

$$d(n_i, n_j) = || r_i - r_j ||.$$

En el proceso de entrenamiento de la SOM las neuronas interactúan entre ellas, la relación entre ellas está regulada por una *función de vecindad* $H(|| r_i - r_j ||)$ que mide la intensidad de la relación entre neuronas, la interacción o intercambio de información es más fuerte cuando la distancia entre neuronas es pequeña. Se tiene que *H* tiene un parámetro *p* el cual representa el radio de la vecindad de las que serán consideradas neuronas cercanas. Una neurona es vecina de otra si la *función de vecindad* es pequeña de acuerdo a la función Gaussiana.

Para las neuronas que son vecinas se realiza un proceso de cooperación y competencia para determinar que neurona es la ganadora y modificar los pesos propios y de neuronas vecinas calculadas con la función de vecindad. Para esta fase puede utilizarse el modelo de *sombrero Mexicano*. Todas las neuronas tienen cierta influencia sobre sus neuronas vecinas, esto es por la *función de vecindad*.

En el proceso de entrenamiento se utiliza un conjunto finito de datos de entrada $X = \{x_0, \dots, x_{m-1}\} \subset R^n$. Lo que busca la red es encontrar neuronas que tengas pesos similares e ir modificándolos en cada iteración, para que si las neuronas están cerca según la distancia euclidiana se junten o separen. Como se especifica en los siguientes pasos:

1. Cada nodo se inicializa con un peso (aleatorio) (W)
2. Se selecciona al azar un vector del conjunto de entrenamiento.
3. Se calcula el nodo de la red que tiene el peso más similar al vector anterior. Para ello, simplemente se calculan las distancias euclidianas entre los vectores W de cada nodo y el vector de entrenamiento.
4. Se calcula el radio de la función de vecindad. Este radio comenzará siendo grande (como para cubrir la red completa) y se va reduciendo en cada iteración.
5. Cada nodo en el radio de la vecindad ajusta su peso para parecerse al vector de entrenamiento seleccionado en el paso 2, de forma que los nodos que son vecinos se vean más modificados siguiendo la siguiente formula.

$$W_j(n+1) = W_j(n) + \wedge_{ij}(n)\eta(n)\Big(X(n) - W_j(n)\Big),$$

donde η es la taza de aprendizaje y $\wedge_{ij}(n) = e^{\left[\frac{-d^2_{ij}}{2\sigma^2(n)}\right]}$ es la función gaussiana que calcula el radio de la vecindad, d es la distancia entre las neuronas y σ disminuye en cada iteración.

6. Repetir desde el paso 2 (el número de iteraciones que se considere necesario).

Las SOM se han utilizado en diversos trabajos por ejemplo para simular la adquisición del lenguaje [4], visualizar agrupamiento [5], su aplicación en procesamiento de lenguaje natural puede comprobarse en [6, 7].

4. Representación vectorial

En el presente trabajo, se analizaron tres libros de William Shakespeare en idioma inglés: Macbeth, Julio Cesar y Hamlet. El número de palabras analizadas fueron 67,805, obteniendo un vocabulario de 13,118 palabras. El vocabulario es poco común, ya que por el año en que fueron escritos, se usan palabras en diferentes idiomas dentro de conversaciones.

Cada uno de los tres libros se analizó por separado. Se eliminaron marcas de puntuación y caracteres especiales, todas las letras en mayúsculas fueron sustituidas por su correspondiente letra minúscula. Los artículos y preposiciones también fueron eliminados de los textos, así como las palabras de frecuencia menor a 3.

Para una palabra *a* que denominaremos clave, el contexto fue capturado considerando la palabra que la precede y sucede, así se formaron tramas de la forma ("predecesor", "clave", "sucesor"). Cada palabra fue representada con una sucesión de 24 dígitos binarios únicos. Para capturar el contexto se crearon vectores de dimensión 72. En los primeros 24 dígitos se almaceno la suma vectorial de todas las representaciones de las palabras que precedían a la palabra clave en el texto, y en los últimos 24 la suma vectorial de todas las palabras que sucedían a la palabra clave. Los 24 dígitos intermedios representaron las diferentes palabras del vocabulario. En la Tabla 1 se puede apreciar un ejemplo de los valores únicos en los vectores para cada una de las palabras, la parte inicial y final a ceros antes de iniciar la captura del contexto.

Tabla 1. Vector de ejemplo con 8 valores binarios creado para cada palabra clave.

Palabra	Clave
reason	00000000 00010000 00000000
beare	00000000 01101000 00000000
heart	00000000 00010100 00000000
roome	00000000 00110001 00000000

La Tabla 2 presenta un ejemplo de dos contextos para la palabra clave *beare*, donde se tienen los predecesores diferentes. Como las palabras tienen representaciones únicas asociadas, se utilizan dichas representaciones para formar los contextos sumando la representación de los antecesores y sucesores a *beare*, Tablas 3 y 4.

Tabla 2. Contextos para la palabra *beare*.

Predecesor	Clave	Sucesor
reason	beare	heart
roome	beare	reason

Tabla 3. Vector de contexto para la palabra "beare", sumando predecesor y sucesor, las 24 posiciones mostradas constituyen ahora la representación de beare.

Predecesor	Clave	Sucesor	Palabra
00010000	01101000	00010100	beare
00110001	01101000	00010000	beare

Tabla 4. Vectores únicos de cada palabra a ocho dígitos, sin presentar los ocho ceros anteriores y posteriores.

Predecesor	Clave	Sucesor	Palabra
01000001	11010001	00100100	beare

Una vez obtenidos los vectores de contexto para las palabras del vocabulario, estos fueron la entrada al SOM, a continuación se describe el proceso seguido para obtener el agrupamiento contextual de las palabras.

5. Proceso de aprendizaje de la red neuronal

Los vectores de contexto fueron usados como entrada en la red neuronal. Las SOM utilizada fue de dimensiones 7 por 9 neuronas, cada neurona representa una colonia o cluster de palabras. Este paso permitió encontrar la relación contextual de las palabras clave de los vectores de contexto. Para etiquetar las palabras en la red se usaron las técnicas presentadas por Teuvo Kohonen en [3]. La representación de las etiquetas en la red fue realizada con 794 palabras obtenidas de cada libro. El algoritmo se ejecutó tres veces, una vez por cada libro.

En un SOM al cabo de suficientes iteraciones, el espacio de datos de entrada es cubierto por el mapa y cada dato del espacio multidimensional de entrada tiene una proyección en el espacio bidimensional de salida, es decir, cada palabra puede ser ubicada en una celda del mapa y en celdas vecinas palabras con vectores similares, como puede observarse en las Figura 1,2, y 3.

6. Resultados

En Fig. 1 se pueden apreciar los resultados de la obra Macbeth de Shakespeare. En el cluster de la parte superior izquierda de la figura se observa que la red asoció palabras que tienen que ver con el cuerpo humano, por ejemplo: *sangre*, *ojos*, y *corazón*. En la parte central superior de la imagen el agrupamiento se realizó a partir de diferentes roles de las personas en la sociedad: *esposo, persona*, *amigo*, *hijo, hermano, papá*. El tercer agrupamiento de la red SOM, está directamente relacionado con muerte y dolor, siendo algunas de las palabras encontradas las siguientes: *muerto, muerte, noche, peligro, trueno*, *tirano*, entre otras. El cuarto agrupamiento corresponde a palabras de valentía y lealtad, por ejemplo: *verdad, honor, espíritu, fortuna,* entre otras. Por último, el quinto agrupamiento para esta obra es un conjunto

de palabras de enfrentamientos: *traición*, *antorcha*, *espada*, *contender*, *fantasma*, *traidor*, etc.

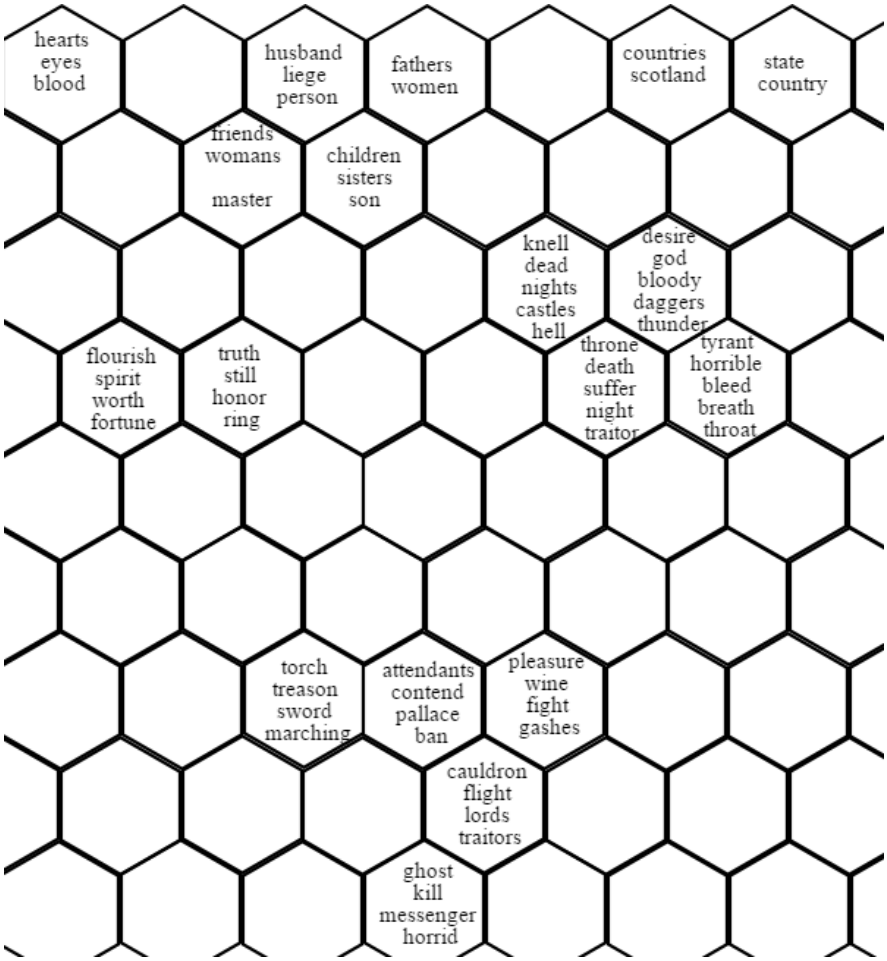

Fig. 1. Resultados del agrupamiento de la red SOM para la obra Macbeth de Shakespeare.

La Fig. 2 muestra el agrupamiento que la red SOM obtuvo de la obra Julio Cesar de Shakespeare. El primer grupo localizado en la parte superior izquierda de la imagen reconoce palabras relacionadas con enfrentamientos, por ejemplo: *pelea*, *débil*, *conspiración*, *soportar*, *cabezas*, *gloria*, entre otras.

Otro grupo, localizado en la parte superior izquierda de la imagen, contiene palabras de independencia, tal es el caso de: *honorabilidad*, *poder*, *libertad*, *conquista*, *fortuna*, *ambición* y *nobleza*. Un tercer grupo alberga palabras de guerra: *corazón*, *honor*, *sangre*, *muerte*, *enojo*, *peligro*, *culpa*. Un cuarto agrupamiento es de palabras directamente relacionadas con los humanos, por ejemplo: *hombres*, *hermano*, *soldado*, *hijo*, *mujer*, *mujeres*, *esposa*, *gente*, *hombre*, *rey*, *señor*, etc.

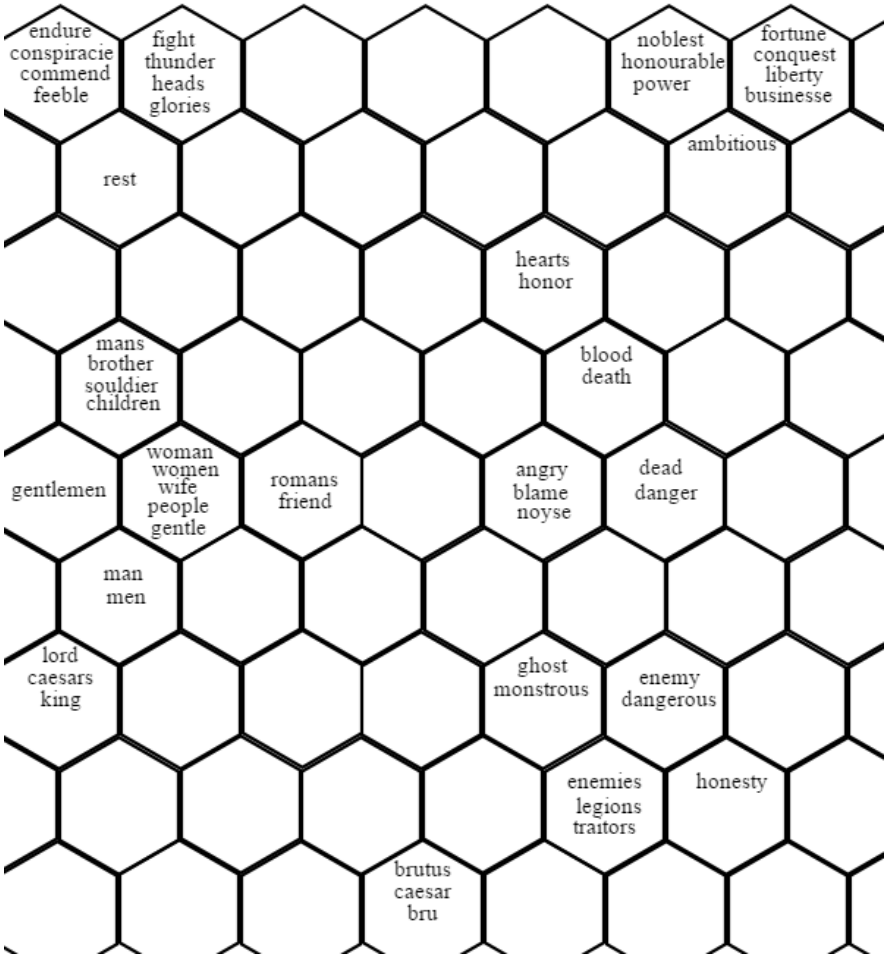

Fig. 2. Resultados del agrupamiento de la red SOM para la obra Julio Cesar de Shakespeare.

En la Fig 3 se muestra el agrupamiento realizado por la red SOM en la obra literaria Hamlet de Shakespeare. Uno de los agrupamientos identificados para esta obra fueron palabras de adjetivos, por ejemplo: *sonriente*, *amable*, *libre*, *humilde*, *extraño*, etc. Un segundo agrupamiento se refiere a palabras que tienen que ver con los humanos: *hermano*, *hermana*, *esposa*, *marido*, *hijo*, *hombre*, *rey*, *criatura*, *joven*, *chicas*, *rey*, *señores*, entre otras. El tercer agrupamiento de la imagen mostrada refleja palabras asociadas con terminología policiaca: *ley*, *crimen*, *causa*, *fuerza*, *violencia*.

7. Conclusiones

El análisis de las obras de Shakespeare, tratando de identificar los principales conceptos empleados, utilizando mapas auto-organizado resultó una técnica adecuada.

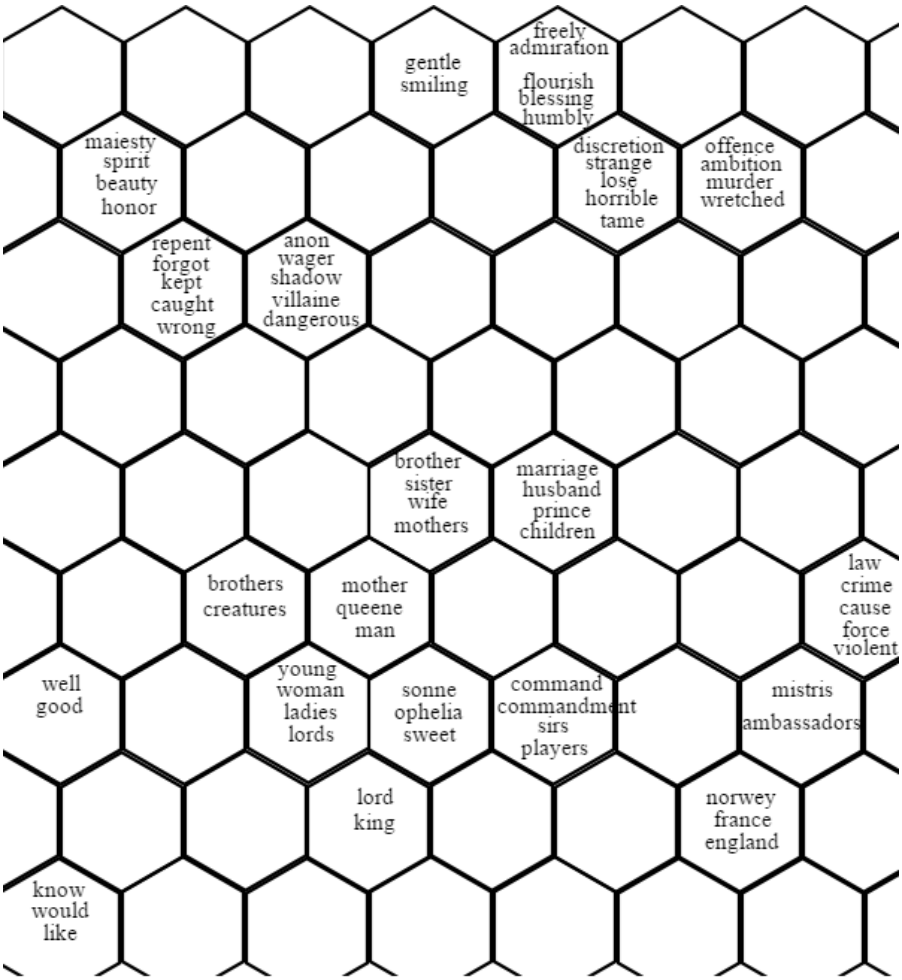

Fig. 3. Resultados del agrupamiento de la red SOM para la obra Hamlet de Shakespeare.

Construir modelos simbólicos explícitos como el análisis del contexto es más fácil para los lingüistas y analistas de obras, debido a que pueden saber de manera fácil y rápida, que temas son los abordados por un determinado autor, cual es la tendencia de palabras, y también las categorías de palabras que existen en un conjunto de obras literarias, sin embargo, esto no es fácil de automatizar. Con los experimentos presentados comprobamos que los SOM pueden utilizarse con este fin.

En el presente experimento se analizaron 3 obras de Shakespeare, como trabajo a futuro se propone analizar todas las obras de Shakespeare y encontrar la tendencia de las palabras que utiliza en sus obras, el uso de algunas palabras por su contexto en la narrativa de cada una de las obras, y además analizar un corpus más grande de otros autores de la época para obtener más información de tipo de narración de ese tiempo. Con esto posiblemente podamos aplicar la misma técnica desarrollada en tareas de atribución de autoría.

Referencias

1. Honkela, T.: Self-Organizing Maps of Words for Natural Language Processing Applications (2012)
2. Shakespeare, W.: Internet Shakespeare Editions. [Online] http://internetshakespeare.uvic.ca/Library (2013)
3. Kohonen, T.: Matlab implementations and applications of the self organizing map. Helsinki, Finland (2014)
4. Li, P., Zhao, X.: Self-organizing map models of language acquisition. Frontiers in psychology 4 (2013)
5. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11(3), pp. 586–600 (2000)
6. Miikkulainen, R.: Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory. MIT press (1993)
7. Scholtes, J.C.: Kohonen Feature Maps in Natural Language Processing (1991)
8. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43 (1982)
9. Krista, L.: Self-organized Maps of Documents Collections: A new Approach to Interactive Exploration. Association for the advancement of artificial intelligence (1996)