Metodología basada en grafos para la identificación de perfiles de usuario

Patricia Espinoza, Darnes Vilariño, David Pinto,
Josefa Somodevilla, Mireya Tovar

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
México

{patricia.efong,mariajsomodevilla}@gmail.mx, {darnes,dpinto,mtovar}@cs.buap.mx

Resumen. En la presente investigación se propone un modelo para la identificación de perfiles de usuario, a través de la creación y análisis de un grafo de co-ocurrencia. Se utilizan 4 corpus en Inglés: de Blogs, de Redes sociales, de Críticas y de Twitter y 2 corpus en Español: de Blogs y de Críticas para el desarrollo de los grafos. Para la creación y extracción de la información del grafo se han utilizado las herramientas NetworkX[1] (creación del grafo) y Gephi[2] (extracción de características del grafo). En general el corpus de Blogs en el idioma Español fue el que presentó los mejores resultados.

Palabras clave: Perfil de usuario, grafos de co-ocurrencia, medidas de centralidad.

1. Introducción

En el mundo actual, se generan contenidos electrónicos de todo tipo, todos los días. Blogs, Twitter, Facebook, son algunas de las plataformas mas comunes para compartir textos de algún tema en particular. Suponiendo que alguien quisiera analizar esos textos para determinar alguna característica en particular o común entre ellos, sería casi imposible, debido al volumen de información que existe actualmente. Por esta necesidad de automatización surgen tareas enfocadas al análisis de los textos, la que se aborda en esta investigación es la tarea de identificar el perfil de un autor de manera automática.

Dicha tarea, se basa en encontrar patrones de escritura entre diferentes grupos, los cuales pueden incluir el género, la edad, el lenguaje nativo y la nacionalidad, entre otras cosas. Esta tarea ha ganado gran relevancia debido a las aplicaciones que se le pueden dar, por ejemplo en análisis forenses, en seguridad y hasta en mercadotecnia.

El enfoque principal de esta investigación, es determinar correctamente el género (female, male) y el rango de edad (18-24, 25-34, 35-49, 50-64, 65+) del

[1] https://networkx.github.io/
[2] http://gephi.github.io/

autor de un documento dado. Para cumplir este objetivo se desarrolló un modelo de aprendizaje automático a partir del análisis de grafos de co-ocurrencia que permite encontrar aspectos relevantes de cada documento.

Los documentos son extraídos de 4 corpus en Inglés: de Blogs, de Redes sociales, de Críticas y de Twitter y 2 corpus en Español: de Blogs y de Críticas. Dichos corpus fueron obtenidos de la conferencia internacional PAN 2014[3].

La estructura del artículo es la siguiente. En la sección 2 se presentan los trabajos desarrollados en la literatura con respecto al uso de grafos para diferentes problemas de clasificación. La sección 3 presenta la descripción del el modelo de clasificación. La discusión acerca de los resultados obtenidos se presenta en la sección 5. Finalmente la conclusión del presente trabajo de investigación se realiza en la sección 6.

2. Estado del arte

Se realizó un estudio sobre los trabajos desarrollados en esta área, enfatizando sus avances y el tipo de diseño que se utiliza al momento de crear los grafos en diferentes tareas, así como sus aportaciones científicas, encontrando el siguiente panorama general:

En el trabajo desarrollado por [7], lo que se busca es realizar consultas sobre una base de datos de grafos indexados, para esto, la representación de los grafos se hace por medio de un código o *canonical label* al que llaman *DFS Code*, si dos grafos son iguales entonces comparten el mismo código. Dicho código es generado al realizar una búsqueda en profundidad en el grafo.

De igual forma en [4] proponen un método para representar una imagen de manera formal, la cual consiste en un conjunto de objetos con propiedades y relaciones. Se busca hacer la representación a través de un grafo etiquetado dirigido, el problema que se aborda es el de cuales propiedades seleccionar para la construcción del grafo. En esta aproximación los objetos son representados por los nodos, y las relaciones y propiedades son las aristas.

En [1] el objetivo de los autores es diseñar un motor de búsqueda que haga uso de la estructura de los hiperenlaces de la Web para encontrar sitios web de interés. Este motor de búsqueda es capaz de encontrar no solo palabras clave o de algún tema en particular, si no que puede buscar un hiperenlace con una estructura deseada. En ese grafo cada URL representa un vértice etiquetado como '_page_', las aristas están etiquetadas como '_hyperlink_' y apuntan de una URL padre a un URL hijo. También se hace un análisis del texto de cada página, se eliminan signos de puntuación, palabras cerradas, etiquetas HTML y todas las palabras restantes se agregan al grafo como un nodo nuevo etiquetado con la palabra y se relacionan con la página correspondiente (nodo '_page_') por medio de una arista etiquetada como '_word_'.

Otro trabajo que utiliza grafos para representar la información es presentado en [2], donde el problema a resolver es la correferencia de entidades. Una entidad

[3] http://pan.webis.de/

es un objeto o un conjunto de objetos del mundo real y una mención es una referencia textual a una entidad. El objetivo de este trabajo es identificar a que entidad hace referencia una mención, para esto utilizan una representación del espacio de correferencia mediante un grafo no dirigido, en donde los nodos representan todas las menciones en el texto y las aristas relacionan a los nodos que se refieren a la misma entidad. Cada arista tiene un peso asignado, el cual representa el grado de confianza de correferencia entre esos nodos.

En los trabajos [3] y [6] también se busca resolver el problema de correferencia. Ambos de igual forma que en el trabajo anterior, crean un grafo donde los nodos son las menciones y las aristas modelan una relación entre esas menciones. Cada arista tiene un peso asignado y en cada trabajo se utiliza un método específico para calcular ese peso.

Por último en [5] se busca hacer un anális del significado de un texto mediante una representación de ese texto en un grafo dirigido, en el cual las palabras del texto se representan por los nodos y las relaciones entre las palabras se representan por las aristas. Un punto interesante de este trabajo es que se crean aristas entre las palabras que están directamente conectadas (una detrás de otra), pero también se conectan palabras que están separadas por un número de palabras definido, para que las palabras que son usadas dentro de un mismo contexto estén conectadas.

Las investigaciones revisadas demuestran que la representación de los textos mediante grafos se está utilizando en la actualidad, ya que existen diversas herramientas que permiten crear grafos con un gran número de nodos y aristas, además de que los grafos logran representar de manera correcta diferentes niveles del lenguaje. Todos estos trabajos nos sirven como referencia y línea base para crear un modelo efectivo, pero es importante destacar que no importa el modelo que se esté evaluando, siempre va a ser más simple detectar el género, que la edad, pues los hombres y las mujeres escriben o se interesan por temas diferentes independientemente de la edad que tienen. Un aspecto importante a estudiar es la técnica de clasificación que se debe usar y su comportamiento frente a los modelos en los que se aplique.

3. Metodología

A continuación se presenta la metodología desarrollada.

3.1. Preprocesamiento del corpus

Debido a que el corpus con el que se trabaja es descargado directamente de la página del PAN, es necesario varias operaciones antes de trabajar con él, algunas de ellas son:

1. Separar el corpus por autor.
2. Separar el corpus por género.
3. Sustituir los símbolos HTML que pueda contener el texto, por su equivalente en utf8.

Para el último punto se desarrolló un diccionario de símbolos HTML, el proceso se puede observar en la figura 1.

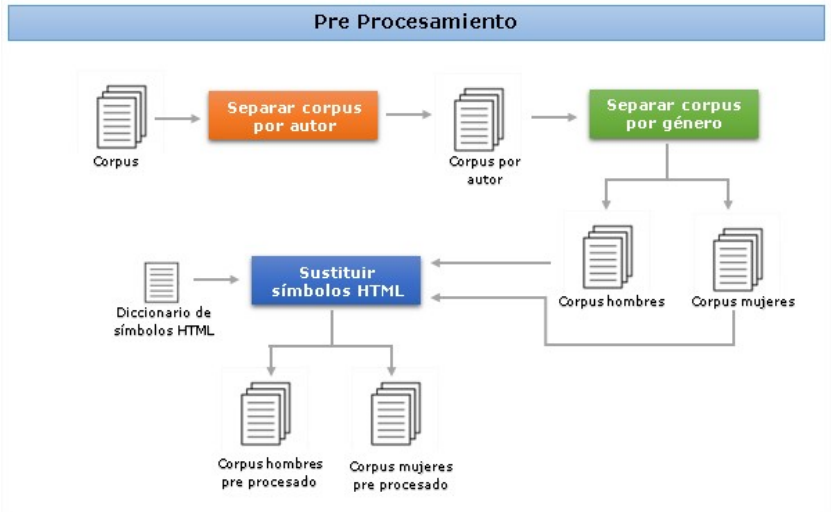

Fig. 1. Preprocesamiento estándar del corpus.

El preprocesamiento estándar se hace para limpiar el corpus de manera general, posteriormente se remueven del corpus las **palabras cerradas**(artículos, conjunciones, verbos auxiliares, etc) , ya que son las que se utilizan con más frecuencia, pero en realidad no aportan significado o no cambian el contexto del texto. La detección de las palabras cerradas se hace a partir de un diccionario en Inglés y otro en Español.

En el tercer paso se sustituyen las palabras restantes en el texto por su correspondiente lema, esto se hace con el objetivo de simplificar y hacer más eficientes los procesos posteriores, ya que se reduce la complejidad de la red resultante, disminuyendo el tamaño del vocabulario. Para realizar este proceso se utilizó la función parse[4] que viene dentro de la librería de Clips utilizada en aproximaciones anteriores.

Como último paso se eliminan los signos de puntuación, los números y se lleva todo el texto a minúsculas (lo que evita que una misma palabra sea considerada como dos palabras diferentes). Todo este proceso se refleja en la figura 2.

Un fragmento del texto resultante se puede observar a continuación:

```
currently see wave idea datum center throw traditional model
datum center management air accelerate demand process datum
storage capacity globally come together environmental demand
create area.
```

[4] http://www.clips.ua.ac.be/pages/pattern-en

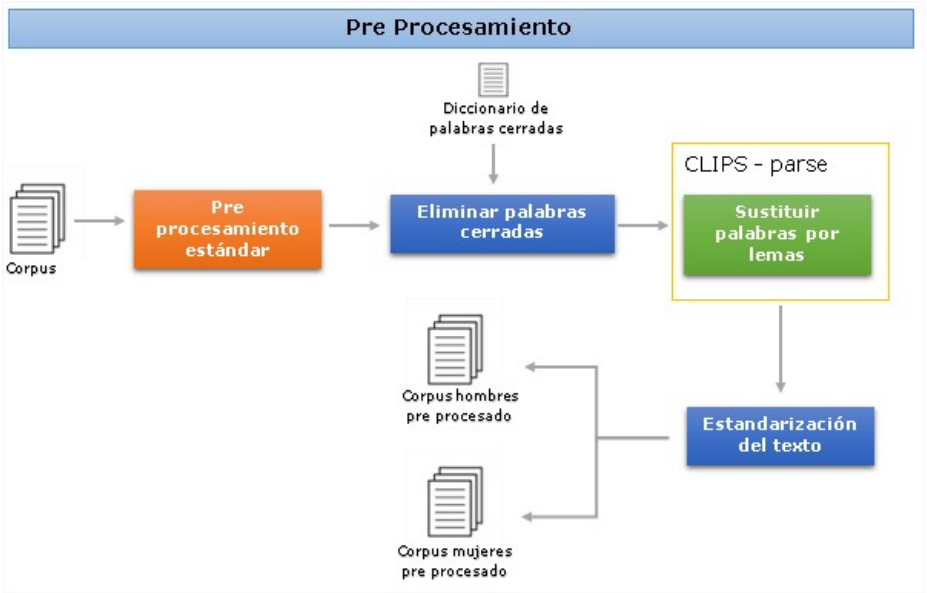

Fig. 2. Preprocesamiento del texto para la creación del grafo.

3.2. Creación del grafo

Despúes de realizar el preprocesamiento de los corpus, el siguiente paso es usar el texto resultante para crear un grafo de co-ocurrencia. Este tipo de grafos se ha convertido en una de las formas más simples y efectivas de representar las relaciones entre las palabras, ya que su implementación es muy fácil de realizar.

Se dice que dos palabras co-ocurren si entre ellas se encuentra un número fijo de palabras, a esto se le llama ventana. En este caso se utilizaron dos tipos de ventanas: una para relacionar los términos que están uno junto al otro (ventana de 0), y otra para relacionar palabras dentro de una ventana igual a 3. El objetivo de la segunda ventana es el de reforzar la relación entre palabras que ocurren en contextos similares.

Formalmente, un grafo de co-ocurrencia dirigido G es un par ordenado G=(V,E), donde:

- V : Conjunto de vértices o nodos los cuales representan las palabras del texto.

- E : Conjunto de pares ordenados de elementos de V que representan la relación entre estos nodos:

En la figura 3 se puede observar el grafo para la siguiente oración: *"currently see wave idea datum center throw traditional model datum center management*

air accelerate demand process datum storage capacity globally come together environmental demand create area"; Se muestran las relaciones que se crean entre las palabras no secuenciales.

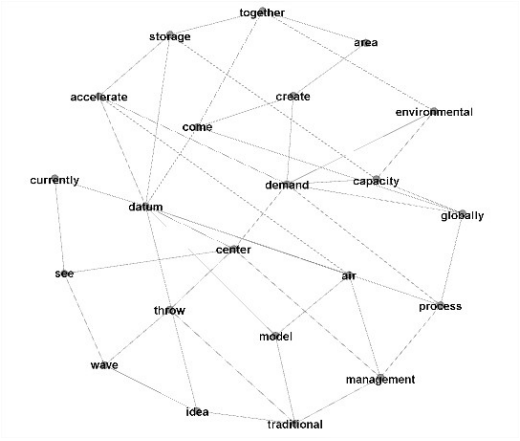

Fig. 3. Grafo de co-ocurrencia.

El proceso para la creación del grafo se puede observar en la figura 4. Se desarrolló un grafo por género {female, male}, se separó el corpus por grupos de edad y se creó un grafo por cada grupo de edad, este proceso se realizó por cada corpus en Inglés y en Español. Al final se obtuvo un total de **72** grafos, los cuales se guardan en un formato xml, para posteriormente crear una representación visual del mismo por medio de Gephi y calcular las medidas de centralidad deseadas.

3.3. Extracción de las carácteristicas del grafo

El desarrollo de grafos de co-ocurrencia permite extraer las palabras relevantes dentro del texto, por medio de las medidas de centralidad y de modularidad, disponibles dentro de la herramienta de análisis de grafos Gephi. Estas medidas se explican a continuación:

- **Interconectividad (Betweenness centrality)**: es un indicador de la centralidad de un nodo dentro de la red. Es igual al número de veces que se pasa por ese nodo para llegar a otros nodos en el grafo o dicho de otra forma, es el número de veces que un nodo aparece al calcular el camino más corto de los otros nodos en la red. Los nodos con una interconectividad alta se pueden decir que son los que tienen mayor influencia dentro de la red, ya que son capaces de representar el contexto en el que se encuentra una cierta palabra.

En la figura 5 se tiene un ejemplo de esta medida, utilizando el mismo grafo de la sección anterior, pero ahora el tamaño de los nodos está dado por el grado

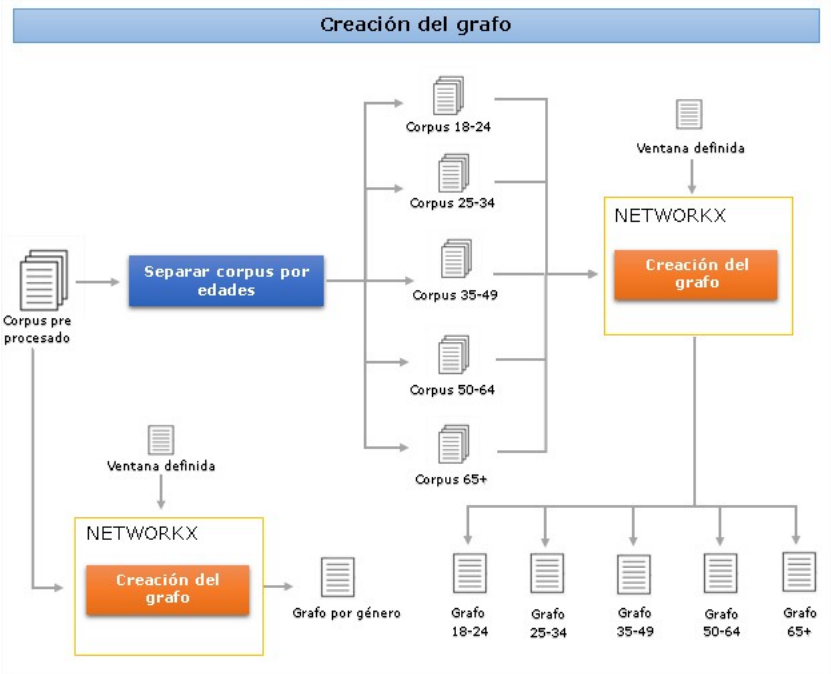

Fig. 4. Creación del grafo.

de interconectividad, fácilmente se puede observar que los más grandes son los más interconectados ya que conectan los dos extremos del grafo.

- **Modularidad (Modularity)**: Mide la fuerza con la que se divide una red en módulos (grupos, clusters o comunidades). Los grafos con gran modularidad tienen conexiones densas entre los nodos que se encuentran en el mismo módulo y conexiones escasas entre nodos de otros módulos. Para esta tarea, nos ayuda a encontrar palabras que se relacionan en torno a un tema dado.

Siguiendo el mismo ejemplo, en la figura 6 se muestran por color los clusters en los que se agrupan los nodos del grafo, el tamaño de los nodos está dado por la medida anterior.

El proceso para el análisis del grafo se puede observar en la figura 7. Se recibe el grafo en formato XML y se calcula el grado de interconectividad entre los nodos, esta herramienta permite observar visualmente los nodos con mayor interconectividad del grafo, ya que pueden ser filtrarlos por tamaño y color.

La segunda medida que se calcula es la modularidad, para que agrupe los nodos por comunidades y se puedan distinguir cada comunidad con un color. Al final lo que interesa es obtener una lista de palabras, en donde cada palabra tenga 2 medidas, el grado de interconectividad y la comunidad a la que pertenece.

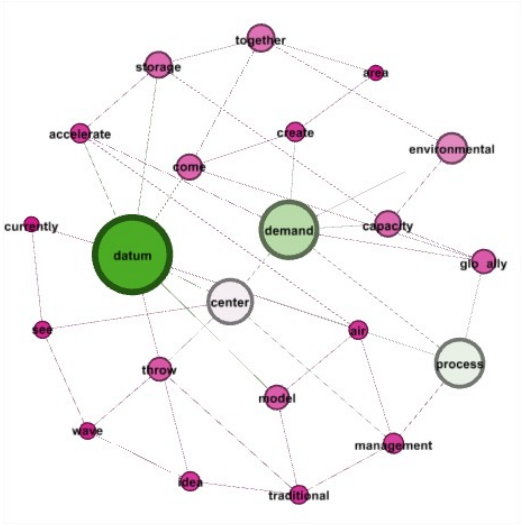

Fig. 5. Ejemplo de interconectividad.

3.4. Proceso de clasificación

Se desarrolló un modelo supervisado el cual se puede observar en la figura 8. Como primer paso se tiene el preprocesamiento, que se realiza para preparar los corpus, posteriormente la creación y análisis del grafo. Después se seleccionan las características o palabras relevantes para ese corpus y esa clase, y se realiza un conteo de las veces que aparece cada palabra en cada documento. También se utilizan todas las comunidades resultantes del análisis y cada vez que se cuenta una palabra, se incrementa el valor de la comunidad o comunidades a las que pertenece.

Se genera un vector por cada documento, donde la longitud de éste es igual al número de palabras elegidas más el número de comunidades . Cada posición del vector corresponde al número de veces que aparece esa palabra en el documento y en el caso de las comunidades, corresponde al número de palabras que pertenecen a esa comunidad en el documento. El atributo clasificador corresponderá al género del autor. Una vez que se han construido los vectores se utiliza como clasificador a la máquina de soporte vectorial para crear el *Modelo de clasificación por género*.

Posteriormente se separan por género los vectores y se les asigna el atributo clasificador correspondiente al rango de edad del autor. Aquí se crean dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre*. Para que a cada modelo solo entren vectores que correspondan a ese género.

En la fase de pruebas se realiza el mismo proceso para crear los vectores con los documentos de prueba que son evaluados con los modelos construidos.

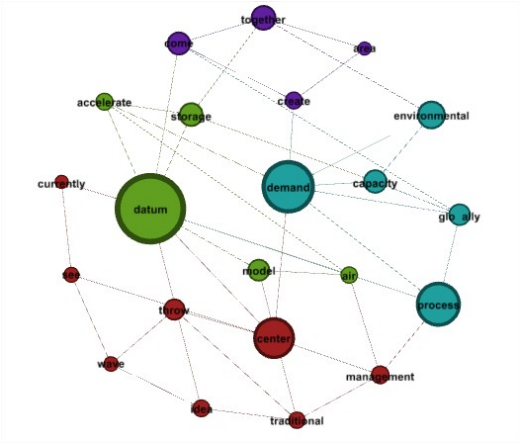

Fig. 6. Ejemplo de modularidad.

4. Descripción de los experimentos

Para estos experimentos se tomaron varios conjuntos de palabras para analizar el comportamiento del clasificador, para cada clase {female, male} de cada corpus y cada experimento se probó por documento y por autor. Con estos conjuntos de palabras de crearon los modelos para clasificar los documentos por género y por edad, los experimentos se explican en detalle a continuación:

- **Experimento 1**: Se escogieron las 1000 palabras con mayor interconectividad (véase el Capítulo 3.3) de cada clase {female, male}.
- **Experimento 2**: Se tomaron todas las palabras del vocabulario de cada clase, excluyendo las que tienen una interconectividad igual a cero.
- **Experimento 3**: Se excluyeron las que tienen una interconectividad igual a cero. Se dividió el total de palabras entre 2 y se tomó mil palabras arriba de la mitad y mil palabras abajo de la mitad, un total de 2000 palabras por clase.
- **Experimento 4**: Se excluyeron las que tienen una interconectividad igual a cero. Se calculó el promedio de la interconectividad de cada palabra y se tomó mil palabras arriba del promedio y mil palabras abajo del promedio, un total de 2000 palabras por clase.

Los experimentos 3 y 4 se realizaron con la hipótesis de que las palabras con mediana interconectividad serían más representativas de su clase, ya que se repitieron menos las palabras entre las clases, a comparación de los experimentos anteriores.

Por último se realizaron 2 experimentos más, pero ahora específicamente para crear un modelo para calcular la edad de los autores de los documentos. Para esto se crearon 10 grafos adicionales por cada corpus, se tienen dos clases para el género {female, male} y 5 clases para la edad {18-24, 25-34, 35-49,

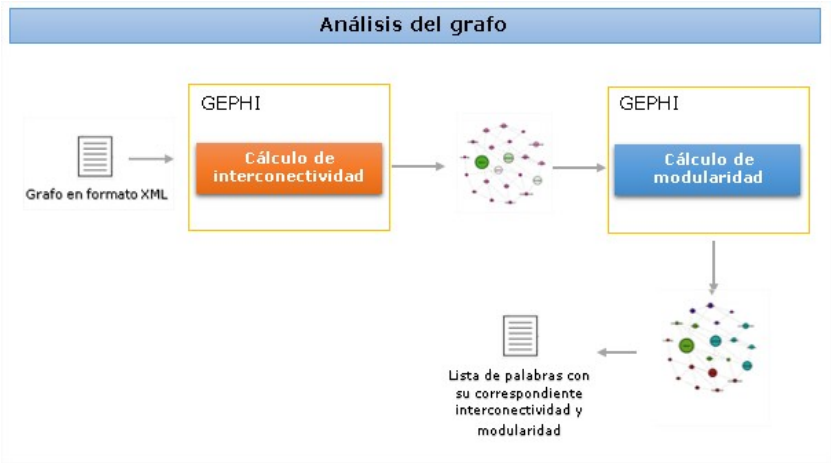

Fig. 7. Análisis del grafo.

50-64, 65+}, posteriormente se creó un grafo por cada clase género-edad (female-18-24, female-25-34, etc). Obteniendo como resultado 5 conjuntos de palabras con su respectiva interconectividad por cada género, para entrenar cada modelo edadHombre y edadMujer (véase 3.4) se utilizaron instancias correspondientes al género del modelo que se entrenó.

- Experimento 5: Se escogieron las 1000 palabras con mayor interconectividad de cada clase (female-18-24, female-25-34, etc), con un total de 5000 palabras para cada modelo.
- Experimento 6: Se escogieron las 1000 palabras con mayor interconectividad de cada clase como en el experimento anterior, pero se observó que las clases que más se confunden entre ellas son: 25-34, 35-49 y 50-64. Debido a esto se decidió tomar las siguientes mil palabras con mayor interconectividad de estas clases en particular, 1000 palabras para las clases 18-24 y 65 y 2000 palabras para las clases mencionadas anteriormente dio un total 8000 palabras para cada modelo.

A continuación se muestran los resultados de los experimentos para cada corpus, en **negritas** los mejores resultados de cada tipo de experimento (por autor o por documento). Debido a que los Experimentos 5 y 6 se diseñaron para calcular la edad, no aplican las pruebas sobre el corpus por género, esto se indica con N/A.

5. Resultados

En el siguiente cuadro se muestra un resumen con los mejores resultados de los experimentos de cada corpus, se muestra en **negritas** los mejores resultados

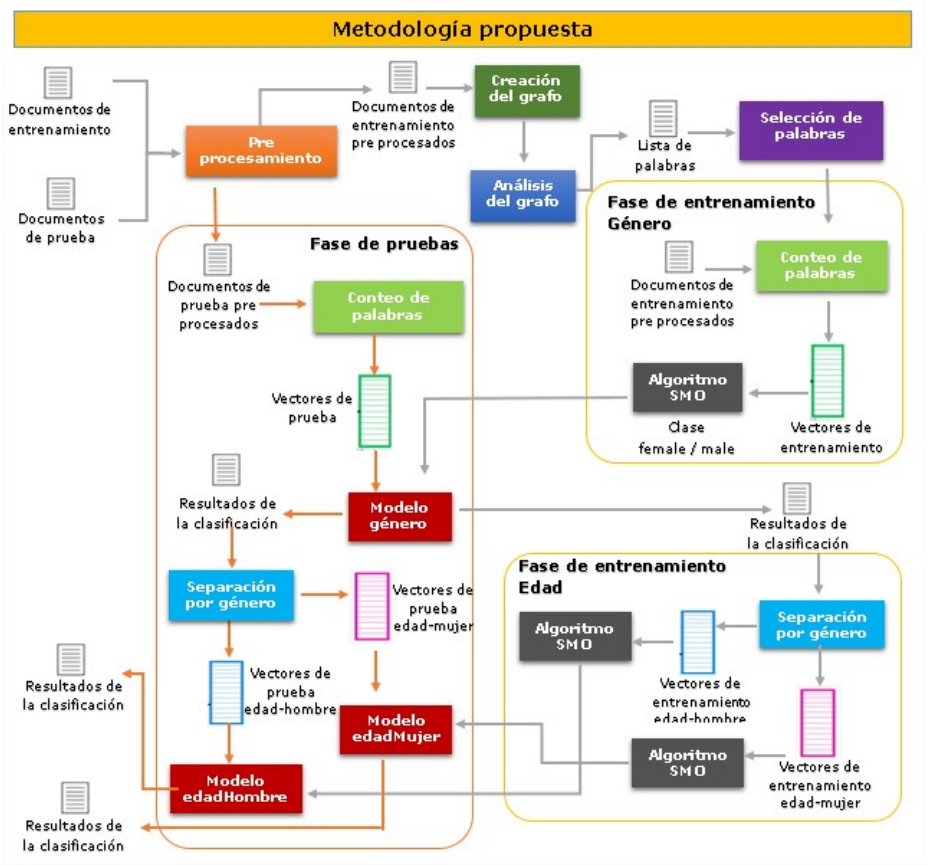

Fig. 8. Metodología para el modelo creado a partir de Gephi.

por idioma. La clasificación se realizó con el algoritmo máquinas de soporte vectorial (SMO) implementado en weka.

Se puede observar en el cuadro 1 que los mejores resultados los obtuvieron los corpus de blogs, pero en general el corpus de blogs en español fue el que presentó los mejores resultados, tanto para el género como para la edad. Otro detalle importante a resaltar es que para la edad, el experimento con mejor desempeño para casi todos los corpus fue el número 6 y para el caso del género fueron el 2 y el 1.

6. Conclusiones

Se desarrolló un modelo para la detección del perfil de un autor (género y edad) mediante grafos de co-ocurrencia. Se pudo observar que el comportamiento

Tabla 1. Resumen de la primera aproximación para ambos idiomas.

Número de características	Tipo de clasificación	Tipo de organización	Presición
INGLÉS			
Blogs			
Experimento 2	Por género	Por documento	**80.76**
Experimento 6	Por edad (mujeres)	Por documento	**67.58**
Experimento 6	Por edad (hombres)	Por documento	**73.36**
Reviews			
Experimento 1	Por género	Por autor	66.82
Experimento 5	Por edad (mujeres)	Por autor	33.89
Experimento 6	Por edad (hombres)	Por documento	31.63
Socialmedia			
Experimento 1	Por género	Por documento	62.39
Experimento 6	Por edad (mujeres)	Por documento	57.67
Experimento 6	Por edad (hombres)	Por documento	56.85
Twitter			
Experimento 4	Por género	Por autor	72.1
Experimento 1	Por edad (mujeres)	Por documento	61.47
Experimento 5	Por edad (hombres)	Por documento	70.61
ESPAÑOL			
Blogs			
Experimento 2	Por género	Por documento	**84.79**
Experimento 6	Por edad (mujeres)	Por Documento	**74.92**
Experimento 2	Por edad (hombres)	Por Documento	**84.24**
Socialmedia			
Experimento 1	Por género	Por autor	63.67
Experimento 5	Por edad (mujeres)	Por autor	47.64
Experimento 6	Por edad (hombres)	Por documento	41.36

del modelo fue un poco diferente para ambos idiomas, superando los resultados de blogs en Español al de blogs en Inglés, siendo lo opuesto en el caso de socialmedia.

En el caso de los experimentos se puede concluir que la idea de realizar grafos por edad (Experimento 5 y 6) fue buena, ya que estos fueron los que mejor resultados brindaron para casi todos los casos. Y para el caso del género, las mil palabras con mayor interconectividad (Experimento 1) fueron las que mejor comportamiento tuvieron. Con los resultados obtenidos se puede afirmar que la confeccíon de grafos de co-ocurrencia para seleccionar caraterísticas para construir modelos de clasificación permite el desarrollo de modelos efectivos, ya que en la mayoría de los experimentos superan las precisiones reportadas en la literatura.

Referencias

1. Cook, D., Manocha, N., Holder, L.B.: Using a graph-based data mining system to perform web search|. International Journal of Pattern Recognition and Artificial Intelligence 17(705) (2003)
2. Cristina, N., Gabriel, N.: Bestcut: a graph algorithm for coreference resolution. In: EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 275–283 (2006)
3. Jie, C., Michael, S.: End-to-end coreference resolution via hypergraph partitioning. In: COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics. pp. 143–151 (2010)
4. Krahmer, E., Verleg, A., Erk, S.: Graph-based generation of referring. In: Computational Linguistics archive. pp. 53–72 (2003)
5. Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. Nodus Labs (2011)
6. Vincent, N.: Graph-cut-based anaphoricity determination for coreference resolution. In: NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 575–583 (2009)
7. Yan, X., Yu, P.S., Han, J.: Graph indexing: A frequent structure-based approach. In: SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data. pp. 335–346 (2004)