Artículo de Revisión

Citación: Niño, Mónica y Ordóñez, Sonia. "Identificación de relaciones de pertenencia entre los nodos de una red social". En: *Ingeniería, Vol. 18, Num. 1, págs. 50-64.*

Identificación de relaciones entre los nodos de una red social

Identification of node relationships in a social network

Mónica Andrea Niño Barón
Estudiante Maestría
Universidad Distrital
Grupo GESDATOS
ninoandrea75@hotmail.com

Sonia Ordóñez Salinas
Docente
Universidad Distrital
Directora Grupo GESDATOS
sordonez@udistrital.edu.co

Resumen

El presente artículo realiza una revisión del tema, representación y clasificación de de relaciones de pertenencia entre los nodos de una red social. Para ello, se abordan aspectos sobre Procesamiento de Lenguaje Natural, Minería de Texto, Recuperación de Información y Entidades Nombradas. Se hace una descripción de cada una de ellas y se referencian y discuten trabajos académicos destacados que se han desarrollado en dicho tema.

Palabras clave: Red social, nodos, pertenencia, Procesamiento de Lenguaje Natural, entidades nombradas.

Abstract

In this paper a review is conduced about representation and classification of membership among nodes belonging to a social network. For this purpose, topics such as Natural Language Processing, Text Mining, Information Retrieval and Named Entities are considered description and survey of outstanding approaches is carry out in each topic.

Key words: social network, nodes, membership, Natural Language Processing, named entities.

1. Introducción

Los servicios de redes sociales involucran a millones de usuarios en línea. Aunque el concepto de una red social se ha establecido desde comienzos del siglo 20 [1], tan solo en los últimos años se ha comenzado a masificar su uso en internet gracias a servicios como *Facebook*, T*witter* o *Linkedin*. Su número de visitantes únicos en Latinoamérica pasó de 53.248.000 en 2008 [2] a 139.000.000 en 2013[3, 4].

La información que puede extraerse de dichos servicios presenta características importantes para los investigadores de diferentes áreas, y por

Fecha recibido: abr. 6/2013
Fecha modificado: jun. 5/2013
Fecha aceptado: jun. 14/2013

lo tanto, su crecimiento también permite contar con repositorios de información cada vez más grandes como los de las universidades de Arizona [5] y Stanford [6], el proyecto Ancora [7], el American National Corpus [8] y Molinero en Español [9], entre otros. La cantidad de información hace que la calidad de la misma pueda disminuirse, y en ocasiones, los márgenes de error no sean los adecuados.

La resolución de preguntas en estas grandes colecciones concernientes a ubicar una persona como miembro de una comunidad, debe considerar por un lado el uso de técnicas para reconocimiento de entidades nombradas (todo término o palabra que corresponde a un nombre propio que identifica una entidad del mundo real y no se encuentra en los diccionarios [10]), y por el otro, técnicas para establecer si dichas entidades pertenecen a la comunidad o más genéricamente a un dominio de investigación, o descartarlas en su defecto. A esta temáticamente se le podría denominar ***búsqueda de relaciones de pertenencia***.

En el presente artículo se presentan los avances en dicho tema. Para ello, se ha dividido el trabajo en tres secciones: En la primera sección se explicará el marco de las relaciones de pertenencia y las entidades nombradas dentro del Procesamiento de Lenguaje Natural (PLN). En una segunda sección, se presenta el estado del arte sobre la identificación de relaciones de pertenencia en redes sociales, el modelo matemático con que se estudia el tema y las métricas propuestas. Y en la última sección se presentan las herramientas para el apoyo en tareas de PLN. El artículo finaliza con las conclusiones del estudio.

2. Procesamiento de Leguaje Natural

El término se utiliza para describir la función que cumplen los sistemas de hardware y de software dedicados a analizar el lenguaje hablado o escrito. [11].

Dentro de las técnicas asociadas al procesamiento de lenguaje natural se encuentran la minería de texto y la extracción de texto e información, que se definen a continuación:

2.1 Minería de texto

La minería de texto es el proceso de extraer patrones interesantes a partir de grandes colecciones de textos para descubrir conocimiento [12]. Es también el descubrimiento de reglas de asociación importantes dentro de un corpus de texto [13]. Usualmente se utilizan en corpus de gran tamaño, como es el caso de la información obtenida de servicios de redes sociales, u otros servicios soportados en internet.

En relación al tema de estudio, aparece en el año 2004, en las memorias de IEEE/WIC/ACM International Conference, un estudio sobre inferencias en conversaciones de

chat [14] utilizando técnicas de minería de texto, con el objeto de determinar el tema principal en una sala de chat. En dicho estudio se observa un análisis completo de la situación, incluyendo manejo de ruido, expresiones concisas y dinámicas, texto cambiante, etc. Los autores utilizan el modelo de clasificación de texto SVM (Máquina de Vectores de Soporte [15]) dejando planteada la posibilidad de manejar los participantes del chat como una red social.

Posteriormente en el año 2006, Feldman publica un libro [16] sobre minería de texto con una recopilación de trabajos, técnicas y herramientas para minería de texto. Sin embargo, en él todavía no se contemplaban aplicaciones a dominios de redes sociales.

A partir del 2008, con el auge de las redes sociales en Internet como *Facebook* y *Twitter*, se proponen estudios sobre técnicas de minería de texto para extraer información en dichas redes. En la mayoría de ellos se acude a una copia de los datos fuera de línea (descargados de diferentes servicios de redes sociales), se procesan y analizan dichos datos, y en ocasiones se ofrecen herramientas para su visualización. Hay un patrón en tales trabajos y es la aceptación de las técnicas y métricas propuestas por X. Chunyan [17] quince años atrás, sobre el manejo de redes sociales con base en grafos. Hui [18] por ejemplo, propone utilizar técnicas de minería de texto para extraer grupos con intereses comunes dentro de una red de blogs. Fard [19] utiliza el mismo concepto para descubrir grupos criminales. Nieto [20] presenta un trabajo similar en español (entendiendo que cada lenguaje tiene sus propias reglas y los trabajos difieren en metodología, algoritmos y enfoque). Todos ellos comienzan a utilizar datos expuestos en los servicios de redes sociales, con las técnicas de Waserman, pero incorporando mejoras para cada lenguaje y escenario.

En el año 2010, se presenta un libro con casos de estudio sobre minería aplicada a redes sociales [21]. El trabajo se basa en las memorias y artículos presentados en el MSN-DS 2009 de Atenas, Grecia (*Minning Social Networks for Decision Supports Workshop*), considerado el primer evento internacional sobre minería de texto aplicada a las redes sociales. El libro muestra el uso del análisis de redes sociales para el descubrimiento de patrones sobre ellas mismas.

A comienzos de 2010 aparecen trabajos con técnicas variadas [21, 22]. Wu [23] por ejemplo propone un método de detección de patrones de texto a través de la utilización de un "multígrafo dirigido", previamente enunciado por Feldman [16]; adicionalmente se desarrolla un algoritmo basado en la distancia de las palabras clave; el método se aplica en la detección de plagio de documentos y correos electrónicos fraudulentos escritos por la mima persona, obteniendo resultados destacados.

Recientemente, Rusell [24] planteó en 2011 un trabajo para extraer información de redes sociales, basado en etiquetas de HTML5, con javascript; dicho trabajo fue validado y extendido por M. Fire [25] en 2013, soportado con tecnología Ajax.

2.1.1 Preprocesamiento de texto

El preprocesamiento de texto es una de las tareas iniciales y primordiales en las actividades de procesamiento del lenguaje natural; sus actividades más importantes son: eliminación del ruido o eliminación de palabras no relevantes [26], identificación de la raíz de las palabras (llamado "lematización" o "stemming") y la ruptura del texto en párrafos y palabras, lo cual es llamado tokenizacion [27]. Manning [28, 29] sintetiza diferentes técnicas para cada una de dichas actividades.

El preprocesamiento de texto suele cambiar según el lenguaje. En algunos de ellos como el inglés, las reglas para conjugación de verbos son simples pero los verbos en ocasiones pueden actuar como sustantivos; en español es común evitar el sujeto, o utilizar verbos auxiliares, las excepciones a las reglas del lenguajes son muy extensas y muchas palabras suelen ser nombres propios, verbos y sujetos a la vez. En alemán, la mayoría de las palabras son a su vez compuestas por otras, algo que no es usual en otros lenguajes. Otros de ellos utilizan incluso códigos diferentes, lo cual hace a cada uno especial. Igualmente, el dominio del problema (el campo al cual se aplicará la técnica) puede variar. Por ejemplo, el vocabulario especializado de una u otra profesión hacen variar la técnica entre sí.

Hay pocas coincidencias entre trabajos, pero es posible determinar algunas tendencias. En cuanto a lematización en español e inglés se suele utilizar el algoritmo de Porter, basado en las reglas comunes del lenguaje y condicionales simples. Para la tokenización en español e inglés se suele separar por palabras o por frases sencillas, y para la lematización se suele utilizar tesauros y ontologías.

2.1.2 Entidades nombradas (NE)

En 2005 hubo una serie de trabajos como el de Peng [30] para extraer entidades nombradas en un corpus grande de texto, el de Tomita [31], que presenta un algoritmo para reconocimiento de entidades nombradas con un sistema jerárquico de palabras, el trabajo de Diamantaras [32] que utiliza un sistema de clasificación "linear binario" para determinar entidades nombradas en un texto, y el de Pu-Jen [33], que propone un método no supervisado para detectar entidades nombradas basado en el análisis del documento completo.

En 2008, Todorovic [34] desarrolla un algoritmo basado en cadenas ocultas de Markov para el reconocimiento de entidades nombradas de tipo persona, localización y organización; en 2009, Jianhan [35] , se presenta un trabajo sobre un *framework* para reconocimiento de entidades nombradas escalable diseñado para Web.

Appice en el 2010 [36] publica un trabajo que abarca extracción de entidades nombradas en literatura biomédica. Es importante resaltarlo debido a que cada posible diccionario técnico representa diferentes retos en un determinado algoritmo.

En el 2011 Jung [37] propuso un método para la identificación de entidades nombradas en microtextos, partiendo de la agrupación de los mismos y ya que la red social se encuentra automáticamente construida, los vínculos sociales son utilizados como base para extraer de una mejor manera las entidades nombradas. También Bollegala [38] realiza estudios sobre la identificación de entidades nombradas en microtextos o textos breves que se publican en redes sociales como Twitter y Facebook. Otros estudios como el de Tkatchenko [39] propone un enfoque semi-supervisado para la construcción de conjuntos de entrenamiento para la clasificación de entidades nombradas. Para su desarrollo se usó una taxonomía de entidades nombradas llamada BBN [40] , un umbral de al menos 40 artículos de Wikipedia, y un subconjunto de las 400 palabras en minúscula más frecuentes, del corpus Reuters.

Otros trabajos relacionados con entidades nombradas se refieren a la búsqueda e identificación de seudónimos de personas en la Web [38], mediante la extracción de patrones léxicos y su posterior clasificación.

2.1.3 Extracción y representación de relaciones de asociación.

La tesis doctoral de Jimenez Ruiz [41] desarrolla un estudio sobre un modelo formal para la extracción de reglas difusas, plantean varias propuestas para la extracción del conocimiento de bases de datos, utilizando teorías de subconjuntos difusos. Proponen nuevas reglas para la descripción de la relación entre dos conjuntos de ítems; por último plantean un procedimiento para la extracción de reglas de asociación.

Existe otro estudio realizado por Uday Kiran [42], sobre la extracción de reglas de asociación soportado en elementos diferentes o "raros" (el autor se basa en la teoría de que en cualquier dominio de problema los casos comunes son los más sencillos de analizar pero aportan menor precisión, mientras que las excepciones inciden más en la precisión de un caso de uso, es decir, que se le debe asignar mayor ponderación a las palabras o frases menos comunes en una colección [24, 43]). El autor asigna un peso a cada palabra o frase mediante su frecuencia en la colección; calcula la cantidad de veces que aparece cada palabra en su documento y multiplica dicho valor por una frecuencia inversa, que incluya el resto de palabras de la colección para darle mayor relevancia a las palabras cuando sean más relevantes por fuera del documento. Esta técnica es conocida como TF-IDF [44, 45].

2.1.4 Clasificación de información

Dentro de los algoritmos de clasificación supervisada, para texto, se encuentran las SVM (Máquinas de Vectores de Soporte, método lineal basado en la maximización de la separación entre dos clases distintas de vectores proyectados en espacios de mayor dimensionalidad [15]), ANN (Redes Neuronales Artificiales [46], que son algoritmos que simulan

el comportamiento neuronal, en los cuales aparecen estructuras de información llamada nodos o neuronas, enlazadas entre sí por medio conexiones), Regresión Logística [47] (técnica en la cual los valores de clasificación se seleccionan por aproximación a partir de una serie de funciones dependientes del modelo), Naive-Bayes (basado en inferencias probabilísticas de la teoría de Bayes) [16], KNN(Clasificación por similitud con los vecinos más cercanos [48]) y árboles de decisión [15] (construidos con técnicas de teoría de la información y análisis de entropía de los datos). Cada uno de ellos presenta características diferentes de rendimiento según su uso [15].

Por otra parte se encuentran los grafos conceptuales [49] (como formalismo estructurado y clasificado como estructura conceptual) permiten representar conocimiento a través de aristas entre dos tipos de nodos: conceptos y relaciones, palabras y símbolos propios de la lógica matemática y del lenguaje natural.

2.1.5 Relaciones entre entidades nombradas

Existen diferentes tipos de relaciones entre miembros de un mismo conjunto y entre los conjuntos en sí mismos. Cuando se tiene una muestra o corpus con una serie de entidades nombradas, es deseable determinar cuál (o cuáles) debe actuar como conjuntos, y cuáles deben actuar como objetos pertenecientes a dicho conjunto.

Existen trabajos al respecto, como el desarrollado en 2004 por Hasegawa [50], que definía cinco pasos en el proceso de asociación entre entidades nombradas: marcar las entidades nombradas (en inglés se utiliza el término *tagging*), identificar casos recurrentes, medir la cantidad de similitudes dentro del contexto, hacer conjuntos de pares y etiquetar cada conjunto de pares; en 2005 Cheng [33] utiliza el mismo concepto de pares de entidades nombradas, pero ponderando cada par según la similitud en todo el corpus (asignando un valor de cero cuando no hay suficiente similitud y un valor alto en la medida en que dicho par de entidades nombradas fuese más encontrado dentro de los documentos). Al obtener un alto grado de precisión, es un modelo que se utiliza actualmente. En 2007, Hirano [51] propone adicionar un mecanismo de aprendizaje supervisado al proceso, el cual mejora en un 4.4% la precisión. En 2011, Tkatchenko [39] propone utilizar un clasificador con aprendizaje semi-supervisado basado en SVM para establecer relaciones entre entidades nombradas dentro de *Wikipedia*. El trabajo ofrece niveles de precisión cercanos a 1 (100%) al aplicar el clasificador sobre 18 clases, lo cual es un resultado destacable.

2.2 Recuperación de información

A nivel internacional, hay trabajos, en el que se hace estudios comparativos entre métodos para extracción de atributos de personas en la Web (en inglés) [52]. En este estudio, se realizan entre otros experimentos, la recuperación "superficial", en el que se extraen etiquetas de marcado HTML y se intenta obtener entidades nombradas desde allí; se

compara contra una búsqueda en profundidad. La comparación entre los métodos se hace en términos de las medidas de desempeño *Recall* y *Precision*.

Douglas y Gong [53], plantean una técnica de desambiguación de nombres por medio de *Hierachical Clustering* (Agrupación Jerárquica), que utiliza la combinación de algunos métodos planteados por Artiles [54], Manning [28], Schütze [29] y Elmacioglu [55], con la técnica de aprendizaje SVM (Cristianini [56]).

En 2010 se reportan casos de estudio [23, 57, 58] sobre la recuperación de información en redes sociales, mezclando técnicas ya conocidas pero sin planteamientos especiales. En 2011 continuó la misma tendencia [59, 60].

En cuanto a aplicaciones en español, existen avances en Extracción de Información basado en técnicas específicas, como el planteado por Ropero [61] (Método general de Extracción de información basado en el uso de Lógica Borrosa: aplicación en portales Web).

3. Identificación de relaciones en redes sociales

Jamail [62] define una red social como una estructura social entre actores en su mayoría personas u organizaciones. Dicha estructura se basa en vínculos sociales, económicos, y de cualquier otra índole.

3.1 Modelo computacional

En el desarrollo de modelos computacionales de redes sociales, recientes estudios se han enfocado en diseño y elaboración de servicios que permitan enriquecer con diferentes componentes tecnológicos, como celulares, GPS, etc. El proyecto realizado por Jung-Tae [63], pretende ofrecer formas más inteligentes y activas para intercambios en el sistema de información. Está compuesto por un integrador de servicios sociales, ubicación y una ontología social, ayudando a la interpretación semántica de los usuarios y su información de manera casi automática. Como herramientas para la representación de la ontología se utilizó XFN (*XHTML Friends Network*) y OWL (*Ontology Web Language*) para la modificación de relaciones. También en el 2010, se publica un libro escrito por Furht [57] en el que se observan las tecnologías y aplicaciones para el manejo de redes sociales, se presentan casos de estudio y se estudian las tendencias a nivel operativo de las diferentes herramientas.

3.2 Métricas para establecer relaciones en una red social

En 2003, James Moody de la Universidad de Ohio [64], comienza a analizar las redes sociales desde el punto de vista jerárquico, manejando matemáticamente el tema como árboles. En este estudio, más social y antropológico, se establece además el concepto de

cohesión estructural, según la capacidad de extraer miembros del grupo sin que el grupo se divida. En el estudio se presenta un algoritmo simple para determinar dicha medida.

Ulrik Bandes [65] habla de herramientas de análisis para fenómenos sociales y establece la centralidad como una de las herramientas, basada en la ruta más corta definida en teoría de grafos. Newman [66] establece la centralidad (intermediación, cercanía o grado) que se ha utilizado para evaluar las redes sociales desde un modelo matemático. Define, basado en la teoría de grafos, el nodo más importante (central) según la suma de sus conexiones.

Se ha avanzado poco en buscar nuevas métricas o algoritmos para perfeccionar las anteriores. Las teorías de grafos han sido aceptadas como herramientas de análisis, y los algoritmos que se utilizan tienen más de 20 años de utilización y aceptación. Por tanto, se entienden como aceptadas las siguientes métricas para establecer y cuantificar relaciones en una red social, entendida como un grafo:

Centralidad (intermediación, cercanía, grado): La centralidad es un atributo estructural de los nodos en una red. Se trata de un valor asignado al nodo debido a su posición estructural en la red. Por ejemplo en un grafo en forma de estrella, el nodo central ocupa un valor máximo de centralidad, mientras que los nodos de las puntas ocupan un valor de centralidad inferior.[67]

Conectividad (puente). Un "punto de quiebre" de un grafo es un conjunto de aristas que, al ser removidas, dejan el grafo inconexo. Un puente es un punto de quiebre de una arista [68].

Coeficiente de agrupamiento: El coeficiente de agrupamiento (*clustering coefficient*) de un vértice en un grafo es la medida para cuantificar el nivel de interconexión de dicho nodo con sus vecinos [69]

Densidad: La densidad de un grafo define la cantidad de posibles aristas (relaciones) utilizadas en él. Técnicamente, un grafo puede tener una densidad entre 0 (cuando ningún nodo está conectado) y 1 (cuando todos los nodos están conectados con todos los demás nodos). Es decir, que un grafo denso es un grafo en el que el número de aristas está cercano al número máximo de aristas. Lo opuesto, un grafo con solo algunas aristas, es un grafo disperso [10] .

3.3 Estructuras para la representación de redes sociales

El estudio de las redes sociales ha dado origen al diseño de modelos que permitan representar de alguna manera el conocimiento y las relaciones entre ellas.

Para la representación de las redes sociales generalmente se utilizan métodos formales (matemáticas y grafos), ya que permiten representar las descripciones de las redes de for-

ma compacta y sistemática, así mismo utilizar herramientas informáticas para el análisis de la red de datos [62]. Igualmente, para la representación de las relaciones en las redes sociales, las herramientas más conocidas son los grafos y las matrices.

3.3.1 Grafos

Junhua [70] define un grafo como un conjunto de nodos (actores) y un conjunto de líneas (relaciones) que conectan los nodos.

Según Jamali [62] el uso de los grafos en redes sociales pequeñas es muy útil, mientras que en redes grandes su lectura e interpretación se hace compleja. Ramanathan [71] considera que las los grafos son útiles para la representación binaria de las relaciones entre nodos, pero no cuando existe propiedades de grupo diferentes.

3.3.2 Estructuras vectoriales y modelos estadísticos

Shoaib [72] desarrolló una ontología que representa la personalización en sitio de las redes sociales, permite el almacenamiento de usuarios, grupos, mensajes, perfil de los acontecimientos en formato de máquina para poder procesarlos después. Utiliza fuentes externas a la red social para la personalización del sitio de los usuarios de la red y cargue de perfiles, los cuales están almacenados de manera semántica. Este estudio se ha hecho solo para la academia.

3.3.3 Herramientas para el apoyo en tareas del procesamiento de Lenguaje natural

Cuando se realizan tareas de procesamiento de lenguaje natural, se dedica un tiempo considerable en la selección y evaluación de herramientas para la realización de dichas tareas. Lobur [73] realizó una evaluación de *Natural Language Toolkit* y su utilización en el campo educativo en el curso de lingüística computacional de la *Lviv Polytechnic National University*, concluyendo que una alta proporción de la muestra de estudiantes, que no contaban con conocimientos en programación, adquirieron habilidades en el Procesamiento de Lenguaje Natural.

En la Universidad Nacional de Colombia se desarrolló una guía metodológica para la selección de técnicas de depuración de datos [74]. Dicha guía está justificada sobre los siguientes postulados: ninguna métrica es adaptable a todos los conjuntos de datos, es poco probable que se resuelva pronto la pregunta de cuál de los métodos, debe utilizarse para una determinada tarea de depuración. La tarea de depuración de datos, es altamente dependiente de los datos y no es evidente que exista una técnica que domine a todas las demás en todos los conjuntos de datos. Y define dentro de la metodología técnicas para detección de duplicados, corrección de valores faltantes y detección de valores atípicos.

4. Conclusiones

Para el descubrimiento de reglas de asociación en un texto, se deben cumplir con las siguientes etapas: preprocesamiento de texto, identificación de entidades nombradas, representación de relaciones de asociación, clasificación de la información y establecimiento de relaciones de pertenecía entre las entidades nombradas. Así mismo si se desean identificar relaciones entre los miembros o nodos de las redes sociales, se requiere que la información o el texto a analizar este estructurado o tenga una representación formal. Esto puede ser una fuente de trabajos futuros, ya que se hace importante explorar técnicas en cualquier etapa de las mencionadas anteriormente, enmarcándolo en un idioma determinado y en un dominio de problema específico.

En lo que respecta a la representación de relaciones en redes sociales, se puede realizar utilizando diferentes estructuras, como son: grafos, estructuras vectoriales y modelos estadísticos, aunque se encuentra abierta la posibilidad de representar relaciones con diferentes técnicas.

Son muchos los trabajos encontrados en el ámbito del procesamiento de lenguaje natural aplicado a las redes sociales; es un tema que está en vigor y amerita profundizarlo. Así mismo en la práctica es posible encontrar aplicabilidad o cualquier avance en ramas como el mercadeo (inteligencia de negocios, fuerza de ventas, tendencias y preferencias), en educación (metodologías basadas en tendencias, establecimiento de nuevas técnicas) y seguridad, este último siendo el más estudiado hasta ahora.

Referencias bibliográficas

[1] G. Simmel, "Conflict," in *Conflict and the Web of Group Affiliations*, T. F. Press, Ed., ed Glencoe, IL: The Free Press, 1908.

[2] A. Lipsman, "Social Networking Explodes Worldwide as Sites Increase their Focus on Cultural Relevance," comScore2008.

[3] A. L. Zain, "Futuro digital Latinoamérica 2013: El estado actual de la industria digital y las tendencias que están modelando el futuro," ComScore2013.

[4] J. Seguic, "El Crecimiento de Redes Sociales en América Latina: La Influencia de Los Medios Sociales en el Escenario Digital de América Latina. Septiembre 2011.," ComScore2011.

[5] Z. R. a. L. H. (2009, 04/07/2013). *Social Computing Data Repository at {ASU}.* Available: http://socialcomputing.asu.edu

[6] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 548-556.

[7] C.-C. d. L. i. Computació. (04/07/2013). *AnCora.* Available: http://clic.ub.edu/corpus/es

[8] N. Ide and C. Macleod, "The american national corpus: A standardized resource of american english," in *Proceedings of Corpus Linguistics 2001,* 2001.

[9] M. A. Molinero, B. Sagot, and L. Nicolas, "A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe," in *Proceedings of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, 2009.

[10] B. Preiss, *Data Structures and Algorithms with Object-Oriented Design Patterns in C++*: John Wiley & Sons, 1998.

[11] P. Jackson and I. Moulinier, *Natural language processing for online applications: text retrieval, extraction and categorization*: John Benjamins Pub., 2007.

[12] A. H. Tan, "Text Mining: promises and challenges," *South East Asia Regional Computer Confederation, Sigapore,* 1999.

[13] M. Delgado, N. Marin, D. Sanchez, and M. A. Vila, "Fuzzy association rules: general model and applications," *Fuzzy Systems, IEEE Transactions on,* vol. 11, pp. 214-225, 2003.

[14] V. H. Tuulos and H. Tirri, "Combining Topic Models and Social Networks for Chat Data Mining," in *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, 2004, pp. 206-213.

[15] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, 2006.

[16] R. S. Feldman, J, *The Text Mining Handbook*. New York: Cambridge University Press, 2006.

[17] X. Chunyan, "Human-machine interface evaluation method based on grey interval relation membership degree," in *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, 2010, pp. 288-291.

[18] W. Hui-Ju, I. H. Ting, and W. Kai-Yu, "Combining Social Network Analysis and Web Mining Techniques to Discover Interest Groups in the Blogspace," in *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, 2009, pp. 1180-1183.

[19] A. M. Fard and M. Ester, "Collaborative Mining in Multiple Social Networks Data for Criminal Group Discovery," in *Computational Science and Engineering, 2009. CSE '09. International Conference on*, 2009, pp. 582-587.

[20] A. Nieto Martín and M. Maroto Calatayud, "Redes sociales en internet y" data mining" en la prospección e investigación de comportamientos delictivos," in *Derecho y redes sociales,* 2010, pp. 207-258.

[21] I.-H. Ting, H.-J. Wu, and T.-H. Ho, *Mining and Analyzing Social Networks*: Springer Publishing Company, Incorporated, 2010.

[22] G. Xu, Y. Zhang, and L. Li, *Web Mining and Social Networking: Techniques and Applications:* Springer-Verlag New York, Inc., 2010.

[23] Q. Wu, E. Fuller, and C.-Q. Zhang, "Graph Model for Pattern Recognition in Text

Mining and Analyzing Social Networks." vol. 288, I. H. Ting, H.-J. Wu, and T.-H. Ho, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 1-20.

[24] M. Russell. (2011). *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites.* Available: http://books.google.com.co/books?id=SYM1lrQdrdsC&printsec=frontcover&dq=Mining+the+Social+Web:+Analyzing+Data+from+Facebook,+Twitter,+LinkedIn&hl=es&ei=S-FRTo2ZHMHr0gGWl82sBw&sa=X&oi=book_result&ct=result&resnum=1&ved=0CC0Q6AEwAA#v=onepage&q&f=false

[25] M. Fire, R. Puzis, and Y. Elovici, "Organization Mining Using Online Social Networks," *arXiv preprint arXiv:1303.3741,* 2013.

[26] Facebook. (2012, 20/02/2012). *Facebook developers*. Available: http://developers.facebook.com/

[27] Twitter. (2012, 20/02/2012). *Twitter developers.* Available: https://dev.twitter.com/

[28] C. Manning, Raghavan, P., Schütze, H, *An introduction to information retrieval,* 2009.

[29] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*: MIT Press, 1999.

[30] L. Peng, W. Xiao Long, G. Yi, and Z. Yu Ming, "Extracting answers to natural language questions from large-scale corpus," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, 2005, pp. 690-694.

[31] T. Tomita, Y. Okimoto, H. Yamamoto, and Y. Sagisaka, "Speech recognition of a named entity," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on,* 2005, pp. 1057-1060.

[32] K. I. Diamantaras, I. Michailidis, and S. Vasiliadis, "A Very Fast and Efficient Linear Classification Algorithm," in *Machine Learning for Signal Processing, 2005 IEEE Workshop on,* 2005, pp. 93-98.

[33] C. Pu-Jen, C. Hsin-Chen, P. Yi-Cheng, and C. Lee-Feng, "Annotating text segments in documents for search," in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, 2005, pp. 317-320.

[34] B. T. Todorovic, S. R. Rancic, I. M. Markovic, E. H. Mulalic, and V. M. Ilic, "Named entity recognition and classification using context Hidden Markov Model," in *Neural Network Applications in Electrical Engineering, 2008. NEUREL 2008. 9th Symposium on*, 2008, pp. 43-46.

[35] Z. Jianhan, "An adaptive approach for web scale named entity recognition," in *Web Society, 2009. SWS '09. 1st IEEE Symposium on*, 2009, pp. 41-46.

[36] A. Appice, M. Ceci, and C. Loglisci, "Discovering Informative Syntactic Relationships between Named Entities in Biomedical Literature," in *Advances in Databases Knowledge and Data Applications (DBKDA), 2010 Second International Conference on*, 2010, pp. 120-125.

[37] J. J. Jung, "Towards Named Entity Recognition Method for Microtexts in Online Social Networks: A Case Study of Twitter," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, 2011, pp. 563-564.

[38] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Automatic Discovery of Personal Name Aliases from the Web," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 23, pp. 831-844, 2011.

[39] M. Tkatchenko, A. Ulanov, and A. Simanovsky, "Classifying Wikipedia entities into fine-grained classes," in *Data Engineering Workshops (ICDEW), 2011 IEEE 27th International Conference on,* 2011, pp. 212-217.

[40] J. Y. Zhang, Y. (2003) Robustness of Regularized Linear Classification Methods in Text Categorization. *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval. ACM Press*. 190–197.

[41] M. D. Jiménez Ruiz "Modelado formal para la representación y evaluación de reglas de asociación," E.T.S. de ingeniería informática y telecomunicación Tesis Doctoral, Departamento de Ciencias de la comunicación e inteligencia artificial Universidad de Granada, Granada, 2010.

[42] R. Uday Kiran and P. Krishna Re, "An improved multiple minimum support based approach to mine rare association rules," in *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, 2009, pp. 340-347.

[43] M. Kumar and R. Vig, "Focused Crawling Based Upon TF-IDF Semantics and Hub Score Learning," *Journal of Emerging Technologies in Web Intelligence,* vol. 5, pp. 70-77, 2013.

[44] L. Hyeokju, H. Joon, and K. Sung-Ryul, "Implementation of a Large-Scalable Social Data Analysis System Based on MapReduce," in *Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on,* 2011, pp. 228-233.

[45] W. Wenxian, C. Xingshu, Z. Yongbin, W. Haizhou, and D. Zongkun, "A Focused Crawler Based on Naive Bayes Classifier," in *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on,* 2010, pp. 517-521.

[46] M. Kudelka, V. Snasel, Z. Horak, and A. E. Hassanien, "Web Communities Defined by Web Page Content," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on,* 2008, pp. 385-389.

[47] A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," in *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, 2010, pp. 193-198.

[48] Z. Lijuan, W. Linshuang, G. Xuebin, and S. Qian, "A clustering-Based KNN improved algorithm CLKNN for text classification," in *Informatics in Control, Automation and Robotics (CAR), 2010 2nd International Asia Conference on,* 2010, pp. 212-215.

[49] J. M. Exposto, J. , A. Pina, A. Alves, and J. Rufino. (2005, Geographical Partition for Distributed Web Crawling. *GIR '05: Proc. of the Geographic Information Retrieval,* 55–60.

[50] T. Hasegawa, S. Sekine, and R. Grishman, "Discovering relations among named entities from large corpora," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, p. 415.

[51] T. Hirano, Y. Matsuo, and G. Kikui, "Detecting semantic relations between named entities in text using contextual features," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions,* 2007, pp. 157-160.

[52] M. Lan, "Which Who are They? People Attribute Extraction and Disambiguation in Web Search Results," *ACM09,* 2009.

[53] J. O. Gong, Douglas, "Determine the Entity Number in Hierarchical Clustering for Web Personal Name Disambiguation," *WWW2009,* 2009.

[54] J. Artiles, J. Gonzalo, and S. Sekine, "The SemEval-2007 WePS evaluation: establishing a benchmark for the web people search task," presented at the Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, 2007.

[55] E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan, and D. Lee, "PSNUS: web people name disambiguation by simple clustering with rich features," presented at the Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, 2007.

[56] N. S.-T. Cristianini, John, *An introduction to support Vector Machines: and other kernel-based learning methods*: Cambridge University Press, 2000.

[57] B. Furht, *Handbook of Social Network Technologies and Applications*: Springer-Verlag New York, Inc, 2010.

[58] C. Tobar, A. Germer, J. Adán-Coello, and R. de Freitas, "Retrieving Wiki Content Using an Ontology

Mining and Analyzing Social Networks." vol. 288, I. H. Ting, H.-J. Wu, and T.-H. Ho, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 21-33.

[59] M. Forestier, J. Velcin, and D. Zighed, "Extracting Social Networks to Understand Interaction," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, 2011, pp. 213-219.

[60] P. Bogdanov, N. D. Larusso, and A. Singh, "Towards Community Discovery in Signed Collaborative Interaction Networks," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010, pp. 288-295.

[61] J. Ropero, "Método general de Extracción de información basado en el uso de Lógica Borrosa. Aplicación en portales web," Tesis doctoral, Ingeniería Informática, Universidad de Sevilla, España, 2009.

[62] A. Manchego, "Sistema de información de detección de plagio en documentos digitales usando el método Document Fingerprinting," 2009.

[63] K. Jung-Tae, L. Jong-Hoon, L. Hoon-Ki, and P. Eui-Hyun, "Design and Implementation of the Location-Based Personalized Social Media Service," in *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, 2010, pp. 116-121.

[64] J. Moody and D. R. White, "Structural cohesion and embeddedness: A hierarchical concept of social groups," *American Sociological Review,* vol. 68, pp. 103-127, 2003.

[65] U. Bandes, "A Faster Algorithm for Betweenness Centrality," *Journal of Mathematical Sociology,* vol. 25, pp. 163-177, 2001.

[66] M. E. J. Newman, "Mathematics of networks," *The New Palgrave Encyclopedia of Economics,* vol. 2nd edition, 2007.

[67] A. Bavelas, *A mathematical model for group structures* vol. 7: Massachusetts Institute of Technology, 1948.

[68] I. Pohl, *An algorithm for finding bridges and its extensions:* Stanford Linear Accelerator Center. Computation Group, 1968.

[69] D. J. W. y. S. Strogatz, *Collective dynamics of 'small-world' networks,* 1998.

[70] D. Junhua, I. Cruz, and L. ChengCheng, "A formal model for building a social network," in *Service Operations, Logistics, and Informatics (SOLI), 2011 IEEE International Conference on*, 2011, pp. 237-242.

[71] K. M. M. Risvik, R, "Search Engines and Web Dynamics," *Computer Networks,* vol. 39, pp. 289-302, 2002.

[72] M. Shoaib and A. Basharat, "Ontology based knowledge representation and semantic profiling in personalized semantic social networking framework," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, 2010, pp. 95-99.

[73] M. Lobur, A. Romanyuk, and M. Romanyshyn, "Using NLTK for educational and scientific purposes," in *CAD Systems in Microelectronics (CADSM), 2011 11th International Conference The Experience of Designing and Application of*, 2011, pp. 426-428.

[74] I. Amón, "Guia metodológica para la selección de técnicas de depuración de datos," Tesis de Maestría, Facultad de Minas, Escuela de Sistemas, Universidad Nacional de Colombia, Medellín, 2010.

Mónica Andrea Niño Barón

Ingeniera de Sistemas Especialista en gestión de Sistemas y Tecnologías de la Información, estudiante MSc.. Ciencias de la Información y las Comunicaciones y miembro del grupo de investigación GESDATOS en la Universidad Distrital Francisco José de Caldas en Bogotá. e-mail: ninoandrea75@hotmail.com

Sonia Ordoñez Salinas

Docente de la Universidad.Distrital, Facultad de Ingeniería. Estadística de la Universidad Nacional. Ingeniera de Sistemas de la Universidad Distrital. Especialista en Teleinformática, Universidad Distrital. Magíster en Ingeniería de Sistemas, Universidad Nacional. Doctor en Ingeniería deSistemas y Computación, Universidad Nacional. Grupo de Investigación Gesdatos U.D. sordonez@udistrital.edu.co