Clasificación de publicaciones en redes sociales semánticas mediante aprendizaje artificial con redes Bayesianas

J. Carlos Conde-Ramírez, Pablo Camarillo-Ramírez y Abraham Sánchez-López

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla, México
{juanc.conde,pablo.camarillo,asanchez}@cs.buap.mx

Resumen El estudio de las redes sociales actualmente esta relacionado con el desarrollo científico de la Web semántica, dada la gran cantidad de información no estructurada presente en las redes sociales. Ésta información debe ser entendida y clasificada para que los administradores de las redes sociales puedan establecer políticas adecuadas que mejoren la experiencia del usuario. En este trabajo proponemos una metodología para clasificar publicaciones realizadas en una red social semántica desarrollada por los autores. Esta metodología de clasificación esta basada en el uso de las redes Bayesianas y en una ontología de dominio para realizar el proceso de anotación semántica de manera más precisa y en forma automática.

Palabras clave: Redes Bayesianas, Web semántica, redes sociales, ontología

1. Introducción

Nuestro estudio parte de la necesidad de mejorar el proceso de anotación semántica de nuestra red social *Moveek* descrito en [1]. El proceso de anotación se explicará con mayor detalle en la sección 2, pero en esencia es el proceso mediante el cual una computadora puede "entender" el contenido que tiene almacenado y asignarle un significado a partir de una base de conocimiento, en este caso una ontología. En este trabajo se presenta una forma de mejorar los resultados obtenidos en el proceso de anotación. Por la naturaleza del proceso de anotación descrito en [1], no todas las publicaciones realizadas en nuestra red social se anotan, ya que la anotación está limitada a la presencia de ciertos términos en el texto que está asociado a la publicación. La manera que proponemos una mejora de este proceso de anotación es el uso de las redes Bayesianas, que nos permitirán clasificar las publicaciones realizadas en nuestra red social a partir de la evidencia y la estadística de la frecuencia de ciertos términos. Con ello se habrá mejorado significativamente la clasificación y por consecuencia el respectivo índice de anotación semántica, haciendo uso de la extracción de la información que depende de la ontología.

En la sección 2 se presentan los preliminares teóricos necesarios para comprender el concepto de la Web semántica y las redes Bayesianas. En la sección 3 se presentan los módulos que se desarrollaron para realizar los experimentos que nos permitieron observar el rendimiento de nuestra propuesta. En la sección 4 se presenta la metodología que proponemos para incrementar el índice de publicaciones anotadas, así como la forma en la que se construye la red Bayesiana utilizada para realizar los experimentos de clasificación. En la sección 5 se presentan los experimentos que se realizaron para obtener el nuevo índice de anotación de este novedoso método de clasificación. En la sección 6 se presentan los resultados del rendimiento de nuestra propuesta para clasificar publicaciones y con ello reforzar el proceso de anotación semántica con ontologías. Finalmente, en la sección 7 se documentan las observaciones y conclusiones que se obtuvieron al desarrollar esta investigación.

2. Marco teórico

A continuación se describen los conceptos más importantes que definen la estructura de esta investigación, como lo son: la Web semántica, la anotación semántica, la ontología y las redes Bayesianas.

Web semántica. La idea es agrupar la información de manera útil y comprensible para la computadora. Por lo tanto uno de los objetivos de la Web semántica es afinar la búsqueda en Internet mediante el uso de metadatos. En este trabajo esos metadatos están contenidos en una ontología, de esta forma, nuestra propuesta es clasificar de manera autónoma la mayor cantidad posible de publicaciones.

Anotación semántica. El término anotación se refiere a una nota, una crítica, una explicación o un comentario. Es decir, escribimos una nota sobre un tema o bien lo criticamos, explicamos o comentamos. Una anotación por sí misma no tiene sentido, está siempre asociada al objeto que ha sido anotado, es por esto que las anotaciones se consideran como *metadatos*. Es importante en este caso precisar el significado del recurso documental; este puede corresponder a un documento completo o bien solamente a un fragmento de este.

Ontología. Una ontología es una conceptualización formal de un dominio, la descripción de sus conceptos y sus relaciones [2,3]. Son modelos de dominio con dos características especiales que conducen a la noción de significado o semántica compartidos donde; las ontologías son expresadas en lenguajes formales con una semántica bien definida, y se basan en una comprensión compartida con la comunidad.

En este estudio, se ha propuesto el desarrollo de una ontología que modele el conocimiento científico que se publica en la red social semántica desarrollada para este estudio[4]. En la sección 3 se presenta la ontología OntoScience como la base de conocimiento, utilizada para la red social semántica que presentamos.

Redes Bayesianas. En inglés *Bayesian Networks* (BN) son una poderosa representación del conocimiento y de los mecanismos de razonamiento. Una red Bayesiana es un modelo probabilístico multivariado, es decir, se vale de métodos

estadísticos para determinar la contribución de varios factores obteniendo un evento simple como resultado. Los eventos y relaciones causales son representados matemáticamente mediante probabilidades condicionales que involucran variables aleatorias que son representadas mediante un grafo dirigido el cual indica la influencia de los factores de forma explícita y que permite obtener la distribución de probabilidad conjunta correspondiente.

La identificación de eventos independientes, por definición, facilita el cálculo de ciertas probabilidades y por lo tanto contribuye a llegar más rápido a una conclusión. Las propiedades gráficas de d-separación[1] se corresponden con las propiedades de independencia en el espacio probabilista asociado. Existe una propiedad fundamental que permite limitar los cálculos de probabilidades, demostrada por Verma y Pear en 1988, que afirma que"Si X y Y son d-separados por Z, entonces X y Y son independientes dado Z"[5].

$$< X|Z|Y > \Rightarrow P(X|Z,Y) = P(X|Z)$$

En otras palabras, el bloqueo de información descrito en los grafos de causalidades, es válido también en la representación probabilista subyacente. Por lo tanto la traducción de un grafo causal a un espacio probabilista conduce a resultados consistentes con el razonamiento intuitivo inferido del grafo.

En el *aprendizaje artificial* la recolección de datos generalmente involucra la recopilación de casos, ejemplos, o instancias de objetos de diferentes tipos o categorías (clases), de modo que para el paso de aprendizaje artificial un modelo de estos datos es creado de forma que después pueda ser utilizado para identificar grupos o similitudes en los datos (aprendizaje no supervisado) o predecir la clase de nuevos objetos (aprendizaje supervisado). De la misma forma una BN puede ser creada automáticamente (aprendizaje) usando datos estadísticos (ejemplos). Existen dos tipos de aprendizaje [6]: *Aprendizaje de parámetros* que dada la estructura de una red (grafo), encontrar el mejor conjunto de parámetros (probabilidades condicionales) para considerar los datos observados y *Aprendizaje de estructura* que sin ninguna hipótesis sobre la estructura de la red, buscar aquella que represente lo mejor posible los datos observados una vez que ya se han proporcionado los mejores parámetros. Así, dados los valores de un subconjunto de variables (evidencia) una BN puede calcular las probabilidades de otro conjunto de variables (variables de consulta).

3. Trabajo relacionado

En los últimos años la Web semántica se ha convertido en un área de investigación indispensable para la búsqueda y recuperación de información. En particular artículos como [7,8,9,10] proponen herramientas o metodologías para la extracción y conceptualización de datos en redes sociales.

[1] El concepto de la d-separación permite precisar en qué condiciones una información puede tratarse localmente, sin perturbar el conjunto del grafo. Por lo que la mejor interpretación es el *bloqueo*.

En [11] Q. Rajput y S. Haider proponen el *framework* de notación semántica *BNOSA* para la extracción de información relevante a partir de datos sin estructura, sin una gramática y prácticamente sin coherencia. El conjunto de corpus utilizados corresponden a paginas Web de compra-venta. BNOSA consta principalmente de dos fases para llevar a cabo su propósito. En la primera fase utiliza una ontología previamente definida para extraer los datos y conceptualizar el dominio del problema con la ayuda de *context words* y tipos de datos ya definidos. La segunda fase utiliza los valores *bien definidos* de los atributos producidos en la primera fase como evidencia dura para poder resolver conflictos; como corregir valores duplicados o incluso predecir valores perdidos. Para este fin la segunda fase hace uso una red Bayesiana correspondiente a la ontología definida. Así como esta, existen algunas otras referencias, no menos importantes, sobre métodos de aprendizaje automático en redes sociales como en [12,13,14].

Por otra parte, en investigaciones anteriores se han empleado redes Bayesianas en tareas de *clasificación de documentos*; haciendo un análisis estadístico y utilizando un algoritmo para generar la estructura de la red Bayesiana que resuelva la clasificación de nuevos documentos. Para lo cual debe establecerse cierta relación semántica entre el contenido del documento y su clase. El propósito es observar la clasificación con el modelo de clasificación construido por Naïve Bayes e investigar como el modelo de clasificación con redes Bayesianas mejora la precisión de clasificación generando una estructura más compleja que representa mejor las probabilidades causales. La clasificación es basada en una simple inferencia, es decir, todas las variables excepto la de clase son conocidas. Para comprobar la efectividad de la red se utiliza JavaBayes.

Como ya se mencionó, esta investigación se basa en la red social descrita en [4] la cual utiliza un motor de extracción de información basado en una ontología para realizar el proceso de anotación semántica. El desarrollo de la ontología llamada *OntoScience* permite tener una base de clasificación y por lo tanto de anotación para las publicaciones realizada en nuestra red social. Dicha ontología se realizó tomando en cuenta un dominio científico para que la clasificación de las publicaciones fuera asignada a alguna ciencia y de esta forma, en el proceso de anotación, se tomara esa ciencia como significado de la publicación. En este trabajo utilizamos una muestra de las publicaciones realizadas en esta red social para probar la metodología descrita en la siguiente sección.

4. Metodología propuesta

La metodología propuesta en este trabajo consiste en la creación de una red Bayesiana y su uso para la clasificación de publicaciones realizadas en la red social empleada para este estudio e incrementar el índice de anotación semántica. En la Figura 1 se muestra un esquema general de la metodología para la creación de la red Bayesiana.

Cada una de las siguientes actividades corresponde a un paso en la generación de la red Bayesiana de la Figura 1.

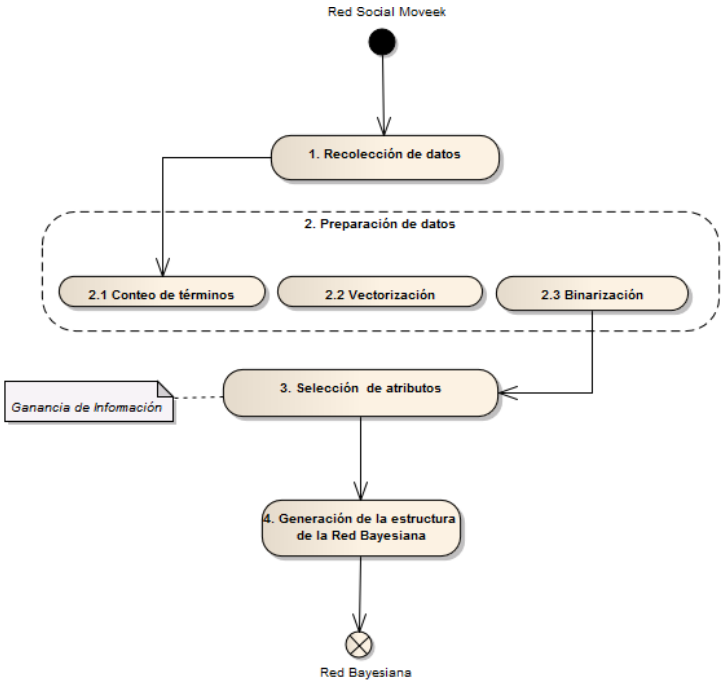

Figura 1. Metodología para la generación de la red Bayesiana

4.1. Recolección de datos

Concentrar todas las publicaciones de la red social, una por renglón, eliminando acentos y signos de puntuación, con la intención de homogeneizar el conjunto de palabras que más adelante serán utilizadas como atributos a seleccionar.

4.2. Preparación de los dados

Una vez concentradas las publicaciones se define el formato del archivo para que este describa explícitamente:

- Nombre del usuario que realizó la publicación (*usuario_publicacion*).
- Cuerpo de la publicación, que corresponde a todas las palabras utilizadas (*contenido_publicacion*).
- Clasificación obtenida por la ontología (*clase_publicacion*).

Con los datos formateados, se continua con el procesamiento de los datos para obtener los atributos o palabras más representativos de nuestro conjunto de datos a través del estudio estadístico pertinente. Esto con la intensión de generar la estructura de la red Bayesiana que mejor represente los datos. Este procesamiento de datos consta de los siguientes pasos:

1. **Contabilizar términos.**

   Para lo cual se necesita convertir los datos STRING a NOMINALES y posteriormente VECTORIZAR el resultado. Para vectorización, y por las características de nuestra muestra, fue necesario aplicar una transformación IDF que es una medida para saber si un término es "común" o es "raro" en los documentos. Es obtenida realizando el siguiente cálculo:

   $$f_{ij} log(\frac{documentos}{documentosquecontieneneltermino})$$

   Donde f_{ij} es la frecuencia del termino i en el documento j.

2. **Binarizar conjunto de datos.** Con esto los valores de los atributos se convierten en binarios a partir del atributo definido "clase". Por lo tanto, el nuevo valor de un atributo binarizado será cero sólo cuando su valor original sea exactamente cero. Es decir, aparece o no aparece en la publicación.

4.3. Selección de atributos

Para el proceso de selección de atributos se aplica el método de Ganancia de Información para evaluar un atributo con respecto a su clase. Para predecir la precisión se utiliza el método de Variación Cruzada[2]; es importante en la vigilancia contra pruebas de hipótesis sugeridas por los datos, especialmente donde más pruebas son difíciles, costosas o imposibles de recolectar [15].

4.4. Generación de la estructura de la red Bayesiana

Como resultado de los procesos antes descritos se obtiene la estructura de la red Bayesiana, la cual nos permitirá clasificar cualquier publicación dada alguna clase de la ontología. A continuación se muestra un breve ejemplo de la estructura en XML que tiene la red Bayesiana empleada para obtener la clasificación de una publicación.

```
<BIF VERSION="0.3">
<NETWORK>
  <NAME>FASE2A</NAME>
  <VARIABLE TYPE="nature">
        <NAME>su_binarized</NAME>
        <OUTCOME>v0</OUTCOME> <OUTCOME>v1</OUTCOME>
        <PROPERTY>position = (100,84)</PROPERTY>
  </VARIABLE>
  ...
  <DEFINITION>
        <FOR>linguista_binarized</FOR>
        <GIVEN>clase_publicacion</GIVEN> <GIVEN>con_binarized</GIVEN>
        <TABLE> 0.83333 0.16666 0.5 0.5 0.9375 0.0625 0.5 0.5</TABLE>
  </DEFINITION>
...
</NETWORK>
```

[2] Es un método estadístico que evalúa cómo los resultados de un análisis estadístico se generalizan. Utilizado en aplicaciones de predicción, obtiene la precisión un modelo predictivo que se llevará a la práctica.

5. Experimentos realizados

En la fase de aprendizaje *no supervisado*, el análisis estadístico se realizó incrementando el número de atributos seleccionados con el fin de observar el comportamiento de la clasificación con Naïve Bayes y Redes Bayesianas. Inicialmente se utilizaron los algoritmos ya mencionados sin considerar la separación estructural de nuestra ontología; donde las superclases A y B (ciencias fácticas y ciencias formales respectivamente) contienen sus propias subclases. En otras palabras, se consideran sólo a las subclases.

Tabla 1. Estadísticas de clasificación de instancias utilizando validación cruzada.

# atrib.	atributos o términos	% precisión General	% precisión Fase 1	% precisión Fase 2-CA	% precisión Fase 2-CB
1	1	28.57	66.67	81.82	90.0
2	1,2	57.14	66.67	72.73	70.0
3	1,2,3	38.10	71.43	90.91	80.0
4	1,2,3,5	57.14	80.95	90.91	90.0
5	1,2,3,5,7	57.14	80.95	81.82	80.0
6	1,2,3,5,7,9	61.91	76.19	81.82	80.0
7	1,2,3,5,7,9,15	66.67	80.95	81.82	80.0
8	1,2,3,5,7,9,15,20	57.14	80.95	81.82	80.0
9	1,2,3,5,7,9,15,20,25	57.14	80.95	81.82	80.0
10	1,2,3,5,7,9,15,20,25,35	52.38	80.95	72.73	80.0
11	1,2,3,5,7,9,15,20,25,35,45	52.38	80.95	81.82	80.0
12	1,2,3,5,7,9,15,20,25,35,45,55	47.62	76.19	81.82	80.0
13	1,2,3,5,7,9,15,20,25,35,45,55,105	52.38	71.43	72.73	80.0
14	1,2,3,5,7,9,15,20,25,35,45,55,105,155	57.14	71.43	72.73	80.0

En la Tabla 1 se observa que la clasificación en "General", tiene menor precisión que la obtenida por nuestra propuesta (columnas 4-6). Incluso tomando 7 atributos (precisión máxima de 66.67 %) y aplicando el algoritmo de redes Bayesianas el porcentaje desciende a 61.91 %. Con los resultados en "Fase 1" se observa que con 6 atributos la precisión de clasificación es de 76.19 %. Aplicando el algoritmo de redes Bayesianas a los mismos atributos la precisión sube a 80.95 %. Para los resultados en "Fase 2-CA" se observa que con 6 atributos la precisión de clasificación es de 81.82 %, pero aplicando el algoritmo de redes Bayesianas sube a 90.91 %. Por su parte en "Fase 2-CB" los resultados muestran que con 5 atributos la precisión de clasificación es de 80.0 %, pero aplicando el algoritmo de redes Bayesianas la precisión sube a 90 %.

Nótese que el algoritmo de Redes Bayesianas generalizado no siempre mejora los resultados, por lo que se tomó como referencia el algoritmo de Naïve Bayes. Sin embargo el resultado obtenido con el algoritmo de Redes Bayesianas siempre es aproximado al obtenido con Naïve Bayes.

6. Resultados obtenidos

De acuerdo al comportamiento de los resultados observados en Figura 2, se seleccionaron los atributos de la Tabla 2 para generar los modelos de las redes Bayesianas para la primera y segunda fase.

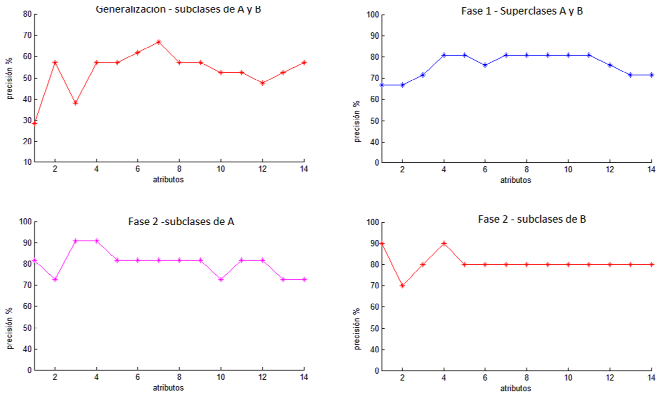

Figura 2. Comportamientos de los resultados obtenidos en las estadísticas.

Una vez generadas las estructuras de las redes Bayesianas que se muestran en Figura 3, lo que resta es asignar evidencia correspondiente para que se calculen las probabilidades relacionadas y así obtener un valor probabilístico de clasificación.

Tabla 2. Atributos seleccionados a partir de la precisión obtenida con el algoritmo de Redes Bayesianas.

Fase de Clasificación	Atributos o Términos	Precisión
Primera, superclases A y B	*nuestra, por, hardware, software, programa, sera, son, con*	85.7143 %
Segunda, subclases de A	*su, con, se, sus, linguista*	90.9091 %
Segunda, subclases de B	*sera, virus, problema, mientras, existencia*	90 %

La asignación se hace definiendo la presencia o ausencia de cada atributo de la red, representados por los nodos hoja del grafo. En la práctica, cuando un comentario es publicado en *Moveek* y no puede ser clasificado por la ontología, pasa a ser analizado de forma probabilista por las redes Bayesianas. En la primera fase se define a que superclase pertenece. Por ejemplo el comentario:

"programa avanzado de estudios en seguridad informática Versión Julio 2009 El Arte de la Guerra nos enseña que no debemos depender de la posibilidad de que el enemigo no venga sino que debemos estar siempre listos a recibirlo No debemos depender de la posibilidad de que el enemigo no nos ataque sino del hecho de que logramos que nuestra posición sea inatacable"

Las únicas palabras que aparecen son "nuestra" y "programa", estableciendo esta evidencia se obtiene una probabilidad de 0.9795 de pertenencia a la superclase B (Ciencias Formales), lo cual es correcto. En la segunda fase se observa que ninguna palabra de la publicación aparece sobre la red c) de la Figura 3, por lo tanto la evidencia para cada nodo es de ausencia (valor = falso ó 0).

Sin embargo se obtiene una probabilidad de 0.9385 de pertenecer a la clase B1 (Ciencias de la Computación), lo cual es evidentemente cierto.

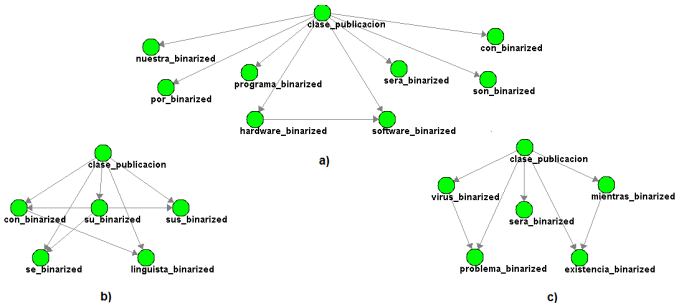

Figura 3. Redes Bayesianas generadas. a)1a. Fase, b)2a. Fase - subclases de A, c)2a. Fase subclases de B

Lo más importante de este enfoque es que tanto la ausencia como la presencia de evidencia contribuyen al cálculo de las probabilidades causales. Esto garantiza la clasificación puesto que siempre se obtiene un valor que define la clasificación confiablemente.

7. Conclusiones y trabajo futuro

Durante el desarrollo de este trabajo de investigación nos pudimos dar cuenta que la tarea de anotación semántica requiere de más de un mecanismo de anotación para incrementar el índice de anotación de nuestra red social. Dado que la metodología propuesta incrementa el número de publicaciones clasificadas en la red social, se cumplió con el objetivo. Sin embargo, los métodos empleados para realizar esta clasificación requieren de una evidencia significativa, es decir, de un número considerable de publicaciones para que los valores inferidos por la red Bayesiana producida sean mucho más confiables.

Por otro lado, éste estudio nos permitió notar que por la naturaleza de la red social, las publicaciones contenidas en dicha red tienen características diversas. Esto hace que el corpus, necesario para el estudio estadístico y la producción de la red Bayesiana, necesite un tratamiento previo para equilibrar el tamaño de las publicaciones agregadas. Por ende, uno de los aspectos que se tomarán en cuenta para mejorar nuestra metodología será establecer un criterio para considerar sólo aquellas publicaciones que permitan obtener un corpus más equilibrado y así mejorar la precisión de clasificación.

Con la metodología propuesta en este trabajo hemos obtenido un proceso de anotación semántica más robusto y exacto, lo cual es un avance significativo para obtener redes sociales completamente semánticas. De esta forma, los administradores de las redes sociales tendrán la oportunidad de conocer mejor a

su audiencia y establecer mecanismos para mejorar la experiencia diaria de los usuarios, entre otras aplicaciones que conlleva lograr construir una red social semántica.

Referencias

1. Pablo, C.R., Abraham, S.L., David, N.R.: Towards a semantic social network. In: IEEE CONIELECOMP 2013. (2013) 74–77
2. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud. **43**(5-6) (December 1995) 907–928
3. Borst, W., Akkermans, J., Top, J.: Engineering ontologies. International Journal of Human-Computer Studies (1997) 365–406
4. Pablo, C.R., Abraham, S.L., David, N.R.: Moveek: A semantic social network. In: WILE 2012 (Fifth Workshop on Intelligent Learning Environments). (2012)
5. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988)
6. Neapolitan, R.: Learning Bayesian networks. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall (2004)
7. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Web Semant. **3**(2-3) (October 2005) 211–223
8. Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M.: Polyphonet: An advanced social network extraction system from the web. Web Semant. **5**(4) (December 2007) 262–278
9. Garcia-Castro, A., Labarga, A., Garcia, L., Giraldo, O., Montaña, C., Bateman, J.A.: Invited paper: Semantic web and social web heading towards living documents in the life sciences. Web Semant. **8**(2-3) (July 2010) 155–162
10. Carminati, B., Ferrari, E., Heatherly, R., Kantarcioglu, M., Thuraisingham, B.: Semantic web-based social network access control. Computers & Security **30**(2-3) (March 2011) 108–115
11. Rajput, Q., Haider, S.: BNOSA: A bayesian network and ontology based semantic annotation framework. J. Web Sem. **9**(2) (2011) 99–112
12. Andrea, E., Fabrizio, S.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). (2006) 417–422
13. Grigori, S., Sabino, M.J., Francisco, V.J., Alexander, G., Noé, C.S., Francisco, V., Ismael, D.R., Sergio, S.G., Alejandro, T., Juan, G.: Empirical study of machine learning based approach for opinion mining in tweets. In Batyrshin, I., González Mendoza, M., eds.: Advances in Artificial Intelligence. Volume 7629 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013)
14. Liu, B.: Sentiment analysis and subjectivity. In Indurkhya, N., Damerau, F., eds.: Handbook of Natural Language Processing, Second Edition. (2010)
15. Walpole, R.: Probabilidad y estadística para ingeniería y ciencias. Pearson Educación (2007)