A Hybrid Approach for Solving the Semantic Annotation Problem in Semantic Social Networks

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

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

Abstract. In this paper, we propose a hybrid method that gives a solution for the semantic annotation problem. We focus our approach to settle the semantic annotation in social networks. Many approaches use a kind of knowledge representation as taxonomies or ontologies to resolve the annotation problem. Recent works have proposed other probabilistic-based approaches to solve the semantic problem as Bayesian Networks. The nature of the Bayesian learning is given by two phases: the data gathering and the query phase, it can be used to settle the semantic annotation problem viewed as a classification one. This work proposes to combine an ontological approach with a Bayesian learning one applied to give a semantic to publications realized in real time in social networks.

Keywords: Semantic annotation, social networks.

1 Introduction

In this paper it is presented a hybrid method that solves the semantic annotation process in semantic social networks. In the study of the Semantic Web, the aim is to describe the content of annotating resources with unambiguous information to facilitate the exploitation of these resources with software agents [1]. Semantic annotation is the process in which one can relate the Web content with a specific knowledge representation. The method described in this paper consist in use an ontology and a Bayesian Network in order to give a semantic for the textual publications in semantic social networks. The firs method of the proposed strategy work uses an ontology to extract the information of publications in social networks. The second method uses a Bayesian Network to classify publications that can not be annotated by using the ontology method.

In [2], we presented an early approach based in Bayesian Networks to classify publications in social networks. Our new work consist in combine the ontology-based-strategy and the Bayesian-Network-method presented in [2]. The section 2 describes briefly the related work about semantic annotation and ontologies in semantic Web, the section 3 presents the generalities of Bayesian networks, the proposed strategy to solve the semantic annotation problem is discussed in

the section 4, then, the section 5 shows the experiments realized and the results obtained, and finally our conclusions about the results obtained by combining Bayesian Networks method and the Ontology method are remarked in section 6, also the future work to improve our work.

2 Semantic Web, Semantic Annotation and Semantic Social Networks

The Semantic Web can be understood like the idea to bring structure to the meaningful content of Web pages, creating an environment where software agents can suspect user's needs and more[3].

Since the beginning of the century, the Web has been changing into social Web. According to [4] the Web is becoming more and more social, we are now collecting huge amount of knowledge on-line. Semantic Web researchers propose making Web content machine understandable through the use of ontologies, which are commonly shared, explicitly defined, generic conceptualizations [5]. But one of the most important problems we face is the way that make possible that machines can understand the content of the Web, the semantic annotation problem.

According to [1], the goal of semantic annotation is to add comments to Web content so that is becomes machine understandable. Unlike an annotation in the normal sense, which is an unrestricted note, a semantic annotation must be explicit, formal, an unambiguous: explicit makes a semantic annotation publicly accessible, formal makes a semantic annotation publicly agreeable, and unambiguous makes a semantic annotation publicly identifiable. These three properties enable machine understanding, and annotating with respect to an ontology or any classification method makes this possible. A Semantic annotation tool is a kind of software that allows add and manage semantic annotations linked to at least one given documentary resource. In the Semantic Web context, the annotation tool can use an ontology or at least one formal model in order to formalize and organize annotations produced by the restrictions defined in this ontology.

The term semantic social networks was coined independently by Stephen Downes and Marco Neumann in 2004 to describe the application of semantic Web technologies and online social networks [6]. In these sense, in [7] it is proposed a three-layered-model which involves the network between people (social network), the network between the ontologies (ontology network) and a network between concepts occurring in these ontologies. In the semantic social network described in [8], authors use an ontology to represent the knowledge that is used to the annotation process.

3 Ontologies and Bayesian Networks

An ontology is a formal conceptualization of certain domain, the description of its concepts and its relations [9,10]. Ontologies are domain models with special

characteristics that drive to the idea of shared meaning or semantic. Ontologies are expressed with formal languages with a well-defined semantic. They are based in a shared comprehension with the common. In our work we take advantage of the Ontology proposed in [11] in which a scientific domain was proposed and modeled. This Ontology contains 64 classes and 108 terms related to scientific domain, with this ontology, the ontology-based method used in this work describe a semantic for textual publications.

The core of this paper is the use of the Bayesian Networks (BN) to *classify* the social networks publications and in this way to give a semantic of textual publications in a semantic social network. BN are a powerful knowledge representation and reasoning mechanism. Formally, BN are directed acyclic graph (DAG) whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; in this sense, nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values of the node's parent variables and gives the probability of the variable represented by the node [12]. In this work the edges of the DAG represent the most representative terms obtained from the gain information strategy used in [2] for building the BN.

4 The Proposed Hybrid Method

The proposal in this work consist in combine two strategies for solve the problem of semantic annotations in semantic social networks. The strategy proposed in [11] gives an approximation for solve the semantic annotation problem with an ontology-based method, however not all publications can be annotated with this approach, because the information extraction method only can annotate such publications that contain words present in the ontology individuals. Therefore, our proposal consist in use the results given by the ontology-based method as the evidence needed for build a BN with the methodology presented in [2]. Once we have the BN, it can be used to classify publications that can not be annotated with the ontology-based method. With this hybridization, all publications can be associated with a semantic. In Figure 1 it is shown the flow of our strategy.

4.1 Ontology-based Method

The semantic social network presented in [11] used an ontology to describe the semantic of textual publications (sentences) published. When the users create a publication, an annotation tool is invoked in order to check if the ontology contains any concept presented in the published sentence. If so, the annotation system associates the publication with the parent class of the concept found in the ontology structure. Otherwise the publication is not annotated. The Figure 2 shows a part of the taxonomy of the complete ontology used in this work.

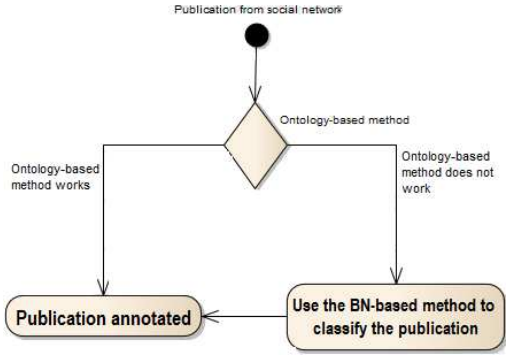


Fig. 1. Strategy flow

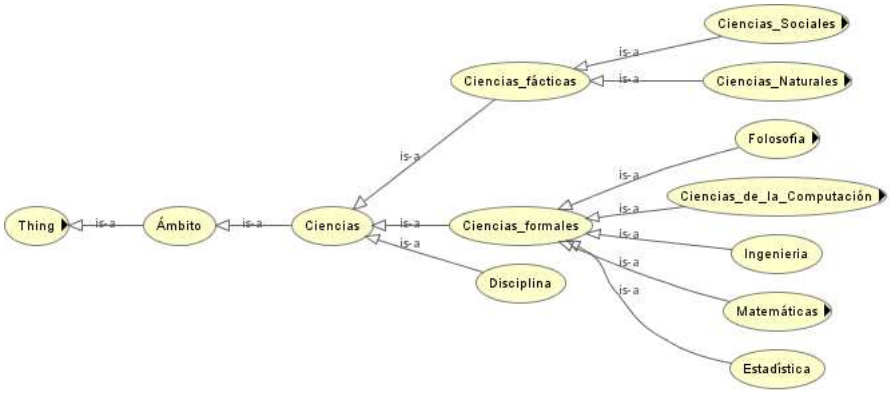


Fig. 2. Part of the ontology taxonomy

4.2 BN-based Method

According with [13], it can be identified three phases in machine learning: data gathering, learning and query phases. In [2] these phases have been implemented as follows.

Data Gathering It has been told that the data gathering must be provide examples, these examples are given by the annotated publications by the ontology-based method, each annotated publication has associated a class. As part of *Data preparation* process, all publications obtained (annotated or not) were subjected to two filters: Terms count and Data binarization. In the first one, a vector is build from the vocabulary, each term in the vocabulary is counted and then an IDF transformation is applied in order to improve the vectorization process. The second one, is a filter applied by using every single publication: each term is converted into binary form depending if each publication contains or not every term in the vocabulary.

Learning Once we have the statistics from annotated publications, it can be possible build the BN. It has been done by using the K2 algorithm proposed in 1992 by Cooper [14]. This method receive the set of the most representative terms and its frequency for building the network. The K2 Algorithm produces a BN in which all terms(nodes) have its respective parents depending of the evidence given by the statistics obtained in the Data gathering phase. A BN as show in Figure 3, is the result of this phase. In this BN, the nodes are considered discrete random variables because they can take two values:1 in case of the sentence that contains the related term to this node, or 0 in otherwise. The BN is produced in a XMLBIF format (XML-based BayesNets Interchange Format) [1].

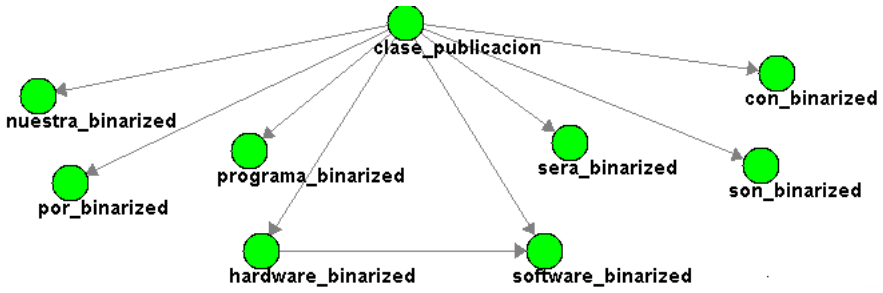


Fig. 3. BN produced by the K2 Algorithm

Query This process is one of the most important contribution in this work. We proposed and implemented a method that *automatically* classify any publication given through a query to the BN generated in the learning phase. The classification algorithm used for classify each incoming publication is the Naïve Bayes Algorithm. Figure 4 resume this phase. First, it is necessary to prepare each publication by removing all Spanish punctuation marks. Then, an evidence must be assigned in the generated BN before, giving the correct value for each node in the BN: if the term represented in the node exists in the incoming publication, the value for this node is 1, otherwise 0. Any value for the node give us enough evidence to obtain a conclusion. The Algorithm 1 explains this phase in detail. The most important aspect in this approach is that the absence (is not present) or the presence (is present) of evidence contributes to the computing of causal probability. With this method any publication can be classified, because always is possible get a trust classification. Finally, when the evidence has been assigned, it can be possible to make an inference through the given evidence. To

[1] The format has been designed primarily by Fabio Cozman, with important contributions from Marek Druzdzel and Daniel Garcia. The XMLBIF format is very simple to understand and can yet represent directed acyclic graphs with probabilistic relations, decision variables and utility values

get the inference needed is necessary to apply the *Variable Elimination Algorithm* to the root of the BN. This query give us the probability of each variable (class), with this information is possible to make a choice about the class of the given publication. The class that contains the maximum probability will be chosen as the class of this publication. This process has been implemented as a Web service. This Web service is invoked by the semantic social network when the ontology-based method does not annotate any incoming publication.

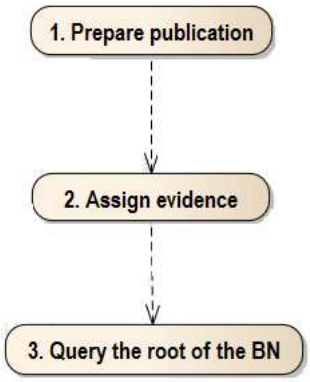


Fig. 4. The Query Process

Algorithm 1 Assign evidence Algorithm

```
1: for all node N ∈ Bayesian Network do
2:     if publication contains the node N then
3:         assign the value "is present" to node N
4:     else
5:         assign the value "is not present" to node N
6:     end if
7: end for
```

5 Experiments and Results

It has been obtained almost one hundred of publications from the social network Moveek [11](the collected data contains publications only in Spanish language).

However, for our study we have been used only the 40% of publications, i.e., the most representative. These annotated publications give us the statistics needed to build correctly the BN. Before applying our annotation strategy, the obtained publications have the distribution shown in Table 1.

Table 1. Ontology-based method accuracy

Annotation success	**Percentage**
Annotated publications	68%
Not Annotated publications	32%

Once that the strategy has been embedded in the semantic social network, it has been taken a sample of new publications in order to proof our strategy. Of course, all publications were annotated by applying the ontology-based method and the BN-based method. The Table 2 summarized the obtained results.

Table 2. Use of annotation methods

Annotation method used	**Percentage**
Ontology-based method	37.5%
BN-based method	62.5%

The results presented in Table 2 show that the new implemented BN-based method is used more times than the Ontology-based method. In fact, the 62.5% of publications could not be annotated without the BN-based method.

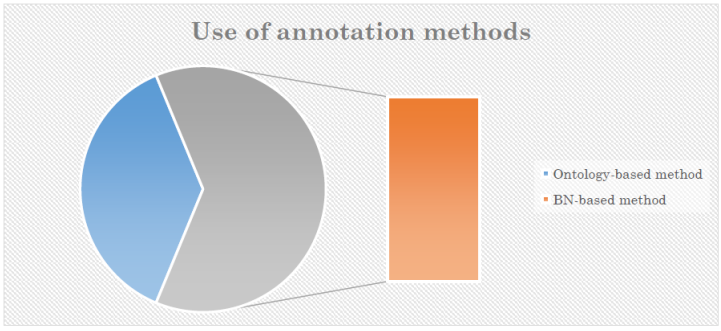
Now, the natural question that one could make is about the accuracy of the classification. For the choose samples, the accuracy of each publication has been verified, in the Table 3 are presented the obtained results.

Table 3. Accuracy method

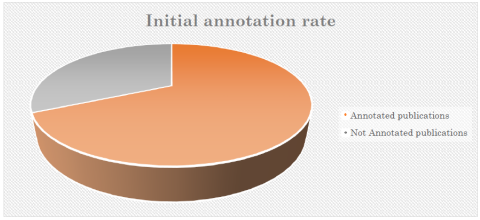
Accuracy of BN-based method	**Percentage**
Correct	60%
Incorrect	40%

We can see that the strategy works well, but it is necesary improve the accuracy percentage of the method. This result has been obtained because the

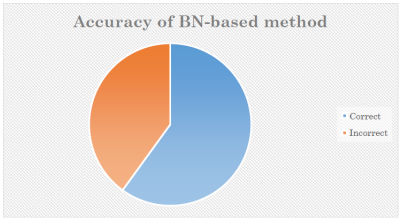
BN used to classify the publications was built with no much examples that could give a better BN structure, i.e., learning Bayesian Networks structure is a NP problem. According with [2], the most examples used to build the BN were part of the *Ciencias Sociales* class. This is the reason of our results, the publications classified incorrectly, were associated to *Ciencias Sociales* class.



(a) Use of annotation methods



(b) Innitial annotation rate



(c) Accuracy of BN-based method

Fig. 5. Results summary

6 Conclusions and Future Work

The strategy applied in this work is based, essentially in the statistics that give evidence to automatize partially the annotation process in a semantic social network, but despite the BN-based method is a very powerful inference mechanism, it depends strongly of enough information to build a correctly BN that can classify the widest possible publications. In our case, we have seen that the BN used to classify need to be modified *continuously* according with new evidence (new publications). This structure modification will be the main topic research for our future work.

The future work consist in creating a software agent that can rebuild *continuously* the BN structure. We know the methodology to build a BN, now, we need to *automatize* this process and make it available to be used by the other

components of the social network platform. In this way, it could be possible improve the BN-based method accuracy.

The rebuild process depends of the accuracy of the classification method, so, it is also necessary develop a mechanism capable of validate the classification by answering directly to social network's user about the accuracy of the classification. With this information could be rebuild the BN when the incorrect classification percentage achieve certain value.

References

1. Prié, Y., Garlatti, S. In: Annotations et métadonnées dans le web sémantique. (2004) Revue I3 Information-Interaction - Intelligence, Numéro Hors-série Web sémantique, 2004, 24 pp.
2. Conde R., J.C., Camarillo R., P., Sánchez L., A.: Clasificación de publicaciones en redes sociales semánticas mediante aprendizaje articial con redes bayesianas. In: Journal Research in Computing Science. (2013) 129–138
3. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284** (2001) 28–37
4. Mika, P.: Social Networks and the Semantic Web. Semantic Web and Beyond, Computing for Human Experience. Springer Science+Business Media, LLC (2007)
5. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5** (1993) 199–220
6. Downes, S.: Semantic networks and social networks. The Learning Organization Journal **12** (2005) 411–417
7. Jung, J.J., Euzenat, J.: Towards semantic social networks. In Franconi, E., Kifer, M., May, W., eds.: ESWC. Volume 4519 of Lecture Notes in Computer Science. Springer (2007) 267–280
8. Camarillo R., P., Sánchez L., A., Nuñez R., D.: Towards a semantic social network. In: IEEE CONIELECOMP 2013. (2013) 74–77
9. Borst, W., Akkermans, J., Top, J.: Engineering ontologies. International Journal of Human-Computer Studies (1997) 365–406
10. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud. **43** (1995) 907–928
11. Camarillo R., P., Sánchez L., A., Nuñez R., D.: Moveek: A semantic social network. In: WILE 2012 (Fifth Workshop on Intelligent Learning Environments). (2012)
12. Ben-Gal, I.E.: Bayesian networks. WWW page (2007) `http://www.eng.tau.ac.il/~bengal/BN.pdf`.
13. Neapolitan, R.: Learning Bayesian networks. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall (2004)
14. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9** (1992) 309–347