Publiziert in: Alt et al. (Hrsg.), Tagungsband 15. Interuniversitäres Doktorandenseminar Wirtschaftsinformatik der Universitäten Chemnitz, Dresden, Freiberg, Halle-Wittenberg, Jena und Leipzig, Leipzig, 2011, S. 11-21.

Using Semantic Web Technologies for Classification Analysis in Social Networks

Marek Opuszko

Friedrich-Schiller-Universität Jena, Carl-Zeiss-Straße 3, 07743 Jena

marek.opuszko@uni-jena.de

Abstract: The Semantic Web enables people and computers to interact and exchange information. Based on Semantic Web technologies, different machine learning applications have been designed. Particularly to emphasize is the possibility to create complex metadata descriptions for any problem domain, based on pre-defined ontologies. In this paper we evaluate the use of a semantic similarity measure based on pre-defined ontologies as an input for a classification analysis. A link prediction between actors of a social network is performed, which could serve as a recommendation system. We measure the prediction performance based on an ontology-based metadata modeling as well as a feature vector modeling. The findings demonstrate that the prediction accuracy based on ontology-based metadata is comparable to traditional approaches and shows that data mining using ontology-based metadata can be considered as a very promising approach.

1 Introduction

The vision of the Semantic Web, as coined by [Berners et al. 2001], is a common framework in which data is stored and shared in a machine-processable way[1]. The content of the Semantic Web is represented by formal ontologies, providing shared conceptualizations of specific domains [Gruber 1993]. Emerged as an extension from the WWW, this technology provides a universally usable tool to model any problem domain. Although the Semantic Web focuses on data stored on the web, this also implies the simultaneous representation of network and feature vector data (in other words, a flat file) in one consistent data representation in the context of (social) network analysis. Different standards for semantic description have been introduced in the past, including different levels of expressiveness. These standards offer a formal way to specify shared vocabularies that are used to create statements about resources. Possible standards are the Resource Description Framework (RDF) [Klyne et al. 2004], the Resource Description Framework Schema RDF(S) [Brickley et al. 2004] and the Web Ontology Language (OWL) [Smith et al. 2004]. We will rely on the RDF standard in this paper.

Soon after its development, the concept of the Semantic Web led to the emergence of new disciplines, for instance Semantic Web Mining [Berendt et al. 2002, 264]. Furthermore, the possibility to model any problem domain in a flexible, formalized manner, quickly gained attention in the data mining and machine learning community. A short overview of ontology languages for the semantic web to represent knowledge gives [Pulido et al. 2006, 489]. Researchers started exploring how traditional machine learning

[1] http://www.w3.org/2001/sw/

techniques may be applied to Semantic Web data [Delteil et al. 2001], [Emde et al. 1996, 122]. The machine learning community has developed a variety of algorithms for different problem domains such as clustering, classification and pattern recognition. Still, the incorporation of all background knowledge available, with other words, the best data representation, is a major issue in data mining [Han et al. 2006, 36]. Semantic Web technologies, in this particular area, promise to enhance the incorporation of all knowledge available. Besides, a growing number of recent decision-making problems have to take into consideration very different kinds of data at once, i.e. (social) network data [Gloor et al 2009, 215].

Nevertheless, Semantic Web data is basically a graph containing all information, whereas traditional machine learning algorithms usually process data in the form of feature vector data, stored in an n-by-p data matrix containing n objects and p variables. In recent research, two approaches have been introduced to exploit the wealth of machine learning algorithms available to process Semantic Web data. One possibility is to preprocess and transform the data to work with traditional methods, i.e. Instance Extraction. Instance extraction is, however, a non-trivial process and requires a lot of domain knowledge. In an RDF graph, for example, all data is interconnected and all relations can be made explicit. [Grimnes et al. 2008, 303] describe the process as the extraction of the relevant subgraph to a single resource. So the question of relevance in the problem domain has to be answered. Another approach is to change existing algorithms to work on graph-based relational data or, if that is not possible, to create new ones. This approach is in fact of growing interest as new data sources in the form of linked data increasingly become available[2]. [Huang et al. 2009] defined a statistical learning framework based on a relational graphical model on which machine learning is possible. Nevertheless, both approaches are costly and non-trivial. Moreover, we can observe that an increasing number of organizations and companies store data in a graph-based relational manner, so a direct utilization of the data is strongly demanded.

Researchers developed methods to measure the similarity between any two objects in ontology-based metadata to later serve as an input for data mining algorithms e.g. hierarchical clustering. [Maedche et al. 2002, 348] introduced a promising approach by defining a set of similarity measures to compare ontology-based metadata. A generalized framework based on the work of Maedche and Zacharias has been introduced by [Lula et al. 2008]. [Grimnes et al. 2008, 303] investigated on Instance Based Clustering of Semantic Web resources and showed that the ontology-based distance measure introduced by [Maedche et al. 2002, 348] performs well for cluster analysis in comparison to other distance measures, such as Feature Vector Distance Measure and Graph Based Distance Measure. The overall performance is, however, hard to evaluate as it differs according to the quality measure Grimne and colleagues used. Beside the work of [Maedche et al. 2002, 348], other efforts have been made in this area ([Bisson 1995, 236], [Emde et al. 1996, 122]), excluding the use of ontological background knowledge. Nonetheless, the recent development in this field, as just mentioned, shows the potential of this approach. Based on these findings, we will test the semantic similarity measure in a classification analysis and measure its performance. A classification analysis has the significant advantage that the prediction accuracy can be measured directly, thus allowing precise conclusions. After [Maedche et al. 2002, 348] solved one important challenge for performing data mining on Semantic Web data, the question remains what drawbacks exist in comparison to classical methods that work on feature vector

[2] See http://linkeddata.org/ for a very interesting project on that issue.

data. Disadvantages of the universal modeling feature, that Semantic Web technologies offer could be loss of accuracy and increased computational time. Furthermore, if the data has to be preprocessed to be handled by traditional methods, additional costly effort is required to transform the data into an adequate format. Besides, new sources of error usually arise when doing so. Moreover, the methods have been mainly tested on data from the Semantic Web background and instances have been extracted from Semantic Web data, whereas this paper follows the opposite approach.
Yet, many other contexts exist where Semantic Web modeling could enhance the performance of traditional methods. One example is Social Networks respectively Social Network Analysis (SNA). Social Media datasets in particular often comprise both (social) network and ego-centered (feature vector) data. Usually both data are treated separately, facing a possible loss of background knowledge leading to a potential lack of analysis accuracy. We want to point out that in this paper we will only focus on the issue of predictive accuracy of the ontology-based similarity measure. We will investigate the usefulness of the ontology-based approach as a source for data mining in the context of SNA in comparison to data modeled as traditional feature vector data. Two real-world datasets comprising both social network and ego-centered data are used. We will further perform a link prediction analysis on both datasets. Link prediction is an important task in network science and has numerous applications in various fields. In the context of social media, link prediction could be used to suggest possible acquaintances in social networks, suggest products for advertisement or propose information artifacts to social media users. [Kautz et al. 1997, 63], for instance, investigated in social networks on how to find companions, assistants or colleagues. However, as link prediction is merely an exemplary task to complete in this paper, we will not illustrate further issues of link prediction here and refer to [Lichtenwalter et al. 2010, 243] for a good overview of the recent development in this field. We will use the baseline predictor *common neighbors* as reference for a state of the art link prediction method.
In detail, the task we will examine in this paper is to predict a relation between two arbitrary actors of a social network. The prediction will be based on the similarity between those two actors. We will compare the prediction on one hand based on the data modeled as RDF-based metadata using the ontology-based similarity measure proposed by [Maedche et al. 2002, 348] and on the other hand based on a traditional feature vector modeling. The main goal of the paper is to assess the applicability of Semantic Web technologies to support SNA on complex datasets.

2 Dataset

We will use two real world datasets to evaluate our approach. Both datasets comprise network as well as ego centered data of two different online communities. In order to assess the quality of an analysis based on technologies derived from the Semantic Web, we firstly created a flat file (feature vector) representation of both datasets. Secondly, datasets in a RDF representation are created including the modeling of ontologies adequate to the specific problem domains. We will extract a social network from both datasets. These social networks will be used to perform a link prediction between actors of the social networks. The link prediction will rely on semantic similarity measures as well as a traditional feature vector.

The first dataset comprises data of a German beach-volleyball community. The Saxonian Beach Volleyball League provided us with anonymized data of 2262 Players and 359 hobby tournaments in the years 2002 to 2010, gathered from their community web-

site's database. The data for each player comprised gender, geographical location, points achieved in last season and age. Each tournament is characterized by the geographical location, the skill level and the date. Figure 1 shows the resulting ontology including some sample data. The ontologies were created using the FOAF[3] and WSG 84[4] vocabulary to model players' attributes and the relations between them. The figure depicts a sample snapshot of two players and two tournaments and their relations. As seen in Figure 1, players attend tournaments and therefore form a two-mode social network through their attendance. The feature vector, i.e. ego-centered, data has been extracted directly from the player table and includes: *gender, age, geographical location (latitude, longitude), points achieved in last season* and *overall points*.

As the players form a social network through their participation in tournaments, we transformed the two-mode network into an undirected one-mode network, only considering the players' relationships if they have co-attended a tournament. We will use this player network to evaluate the prediction. The relation between players is weighted with the number of co-attendances. Players without any relation were excluded. The resulting network included 2,175 vertices and 105,922 edges. The network can be characterized as a dense network in the context of social networks with a density of .044 and an average degree of 97.39, meaning that on average, every player has a relation[5] with nearly 100 other players. An average local clustering coefficient of .727 and a diameter of 5 characterize the network as a small world network as stated by [Watts et al. 1998, 440]. Furthermore, the degree distribution among all vertices depicts a scale-free network, following [Barabási et al. 1999, 509].

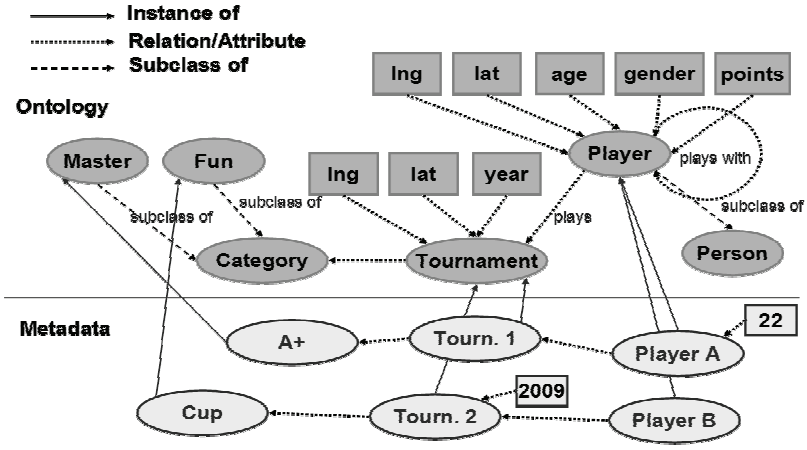


Figure 1: Ontology and example metadata of the dataset used for classification analysis

The second dataset has been extracted from a survey among 692 Facebook users of university freshman, four months after the semester start. We asked the students to grant us access to their personal profiles including their friend list, group memberships, event attendances and *like* relations. Facebook users form social relations through becoming a *friend*. To become friends in this context not necessarily means a strong social relationship. The term *acquaintance* is surely more adequate here. Users on Face-

[3] http://xmlns.com/foaf/0.1/
[4] http://www.w3.org/2003/01/geo/wgs84_pos
[5] The term *relation* in this context does not necessarily mean a social relation. Players in this particular network are related, if they participate in the same tournaments.

book can further join groups where users with common interest can interchange information. Users on Facebook are further able to create and attend events. Typical events are music concerts or cultural events of any kind as well as personal events like private birthday parties. A very interesting concept is the possibility for every Facebook user to "like" any object (referenced through an URL) on the web. Furthermore, the *like* function can be read in two directions. Users show with this function that they actually like something, e.g. movie stars, music bands or writers and books, on the other hand, users simultaneously share the object they like with their friends or with other Facebook users. The second functionality may be interpreted as a recommendation rather than an actual evaluation. As a summary, Facebook likes can be read both, as a set of personal attitudes and interests and also as information that users share among each other. The underlying idea of the decision support system in this work is that users who share similar groups, events and likes, are more likely to be connected as Facebook friends.

Similar to dataset one, users have ego-centered data "gender", "hometown", "timezone", "actual location", "country" and also relational data like friend relations or memberships in groups and items users like on Facebook. Groups and Likes are further characterized by variables like a unique URL or a name. Likes and Groups further may be attached to a certain category whereas categories can be recursively related to other categories. The categories are derived from the Facebook website and reach from root categories like "Interests", „Music" or "TV" to sub granular categories like "Local Businesses". Dataset two comprises 45,108 relations to 24,648 unique Likes. There are further 6,894 memberships in 5,993 unique groups. The social network is formed by the friend relation on Facebook. Based on this user-to-user relation, a social network is created. The network shows 3,028 connections between the 692 users, leading to an average vertex degree (number of friends) of 8.75. The resulting network is less dense than the network in dataset one, showing a higher network diameter with a value of 13, a lower clustering coefficient .349 and a lower general density of .013.

An RDF graph containing all relations has been created for every dataset. The semantic similarity measures following the approach of [Maedche et al. 2002, 348] are calculated based on this RDF graph as a base for the later classification analysis. The Semantic Similarity distance metric is a weighted combination of three dimensions of similarity: taxonomy similarity *TS*, relational similarity *RS* and the attribute similarity *AS*. We refer to [Maedche et al. 2002, 348] for further details on the calculation.

The other metrics included all ego-centered player/user information for a pairing per tuple. For the feature-vector metrics of dataset one, we further added the sum of the overall league points a player achieved so far (separated into three classes according to different regional leagues) and computed possible interaction variables, namely gender difference and age difference. Finally, each tuple in the feature vector metrics comprises the information about the two pairing players/users and the outcome class. Due to the analysis layout, the outcome variable is dichotomous in the form of *L (Link)* vs. *NL (No Link)*, respectively 1 vs. 0. We further calculated the *common neighborhood* as the number of shared neighbors to compare the introduced metrics with a baseline predictor. To finally create a sample for the prediction task of dataset one, randomly 1,000 player pairings have been chosen from the dataset. The 1,000 player pairings included 500 pairings where the two players actually had a relation and 500 pairings where no relation existed. This forms two groups in our outcome class: *L*, for an existing link and *NL* for no existing link. In the sample both groups are equally distributed. Nonetheless, it should be kept in mind that the network showed only 4.4% of all possible links. In

real world networks, however, the *NL* group will be very likely highly overrepresented. Additionally the procedure has been performed for dataset two, here comprising 1,000 user pairings in each group.

3 Classification analysis

Before performing the classification analysis, it is advisable to comprehensively examine the statistical properties of the Semantic Similarity metric. On this account we performed an ANOVA to assess the discriminative power of the metric. We will later perform a binary logistic regression, a discriminant analysis and use a decision tree to evaluate the predictive power of the different data modeling. We decided not to weight the three similarity measures (*TS, RS, AS*) as proposed by [Maedche et al. 2002, 348] and use all three similarity measures uncombined to get further insight how each element influences the predictive accuracy.

TS = taxonomy similarity
RS = relational similarity
AS = attribute similarit

The class "NL" here represents the class of pairings with no link, whereas "L" represents the class with an existing link between the two pairing.

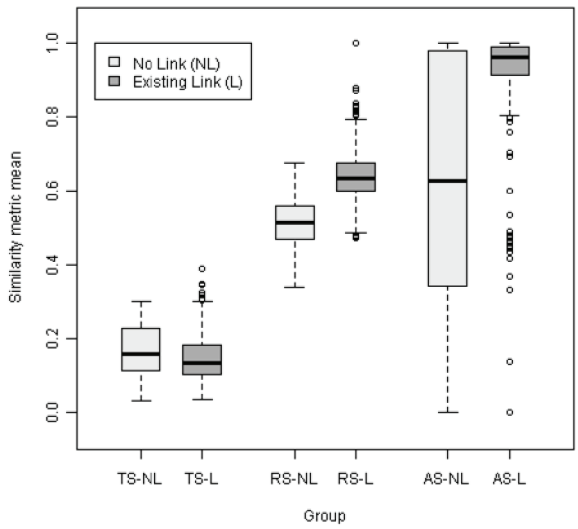
Figure 2 shows boxplots of all similarity measures grouped by the outcome class in dataset one.

The plot highlights obvious mean differences for each similarity measure in the outcome class. The relational similarity in particular shows a distinct mean difference, having comparable standard deviation dimensions. Taxonomy similarity also shows a mean difference whereas the standard deviation of both groups is quite similar. The attribute similarity shows a very high standard deviation, possibly leading to false predictions in some regions, if the prediction would rely on this element only. However, the increased weight for the relational similarity when combining all three measures, as proposed by [Maedche et al. 2002, 348], seems reasonable as it promises the highest discriminative power when visually analyzing the boxplots in

TS = taxonomy similarity
RS = relational similarity
AS = attribute similarit

The class "NL" here represents the class of pairings with no link, whereas "L" represents the class with an existing link between the two pairing.

Figure 2.



TS = taxonomy similarity
RS = relational similarity
AS = attribute similarit

The class "NL" here represents the class of pairings with no link, whereas "L" represents the class with an existing link between the two pairing.

Figure 2: Boxplots of semantic similarity components in outcome class of dataset one players.

The similarities of the Facebook dataset (dataset two) showed a similar result. Taxonomy similarity has been omitted in dataset two as all users showed an identical taxonomy similarity. This is due to the fact that there are practically no taxonomic differences for two users as there are no user classes or categories in dataset two. We will further omit this attribute in all later steps. To statistically prove the visual impressions, a one-way ANOVA on group differences for each similarity measure was performed. The results depict a significant group difference for every measure. Again, relational similarity *RS* (M = .52, SD = .06 in group *no link* and M = .64, SD = .08 in group *link,* dataset one) shows the highest effect size regarding Cohen's *d* (1.76 for dataset one, 1.30 for dataset two) as already assumed by the analysis of the boxplot charts. Worth mentioning is that all similarity measures show a significant mean difference of $p<.001$ in the outcome groups. These findings statistically prove the ability of the semantic similarity measure to discriminate the outcome groups. The high values of Cohen's *d* effect size measure further depict a genuine difference between linked and not linked pairs in terms of their semantic similarity.

After having laid the statistical foundations, the questions of the predictive power of the Semantic Similarity measure is examined. The results are shown in Table 1. First of all, a logistic binary regression (cut value = .5) has been performed for each dataset and all metrics. The regression showed a significant relation for all three predictor variables TS, RS and AS: $p<.001$, $R2 = .86$ (Nagelkerke), $R2 = 0.64$ (Cox & Snell) in dataset one. As seen in Table 1 the logistic binary regression leads to a predictive accuracy with 93.6% in dataset one and 79.4% in dataset two. Furthermore, the predictive accuracy is quite similar in both groups. The analysis of the feature-vector data shows a lower predictive accuracy with an overall accuracy of 80.8%.

		Accuracy (dataset one)		
Method	Data model	TP	TN	Overall
Logistic binary regression	Feature Vector*	81.6%	79.7%	80.8%
	Semantic Similarity measure	94.6%	92.6%	93.6%

	Common neighbors	96%	95%	**95.6%**
Discriminant analysis (cross validated)	Feature Vector*	73.3%	87.1%	79.1%
	Semantic Similarity measure	95.8%	89.2%	**92.5%**
	Common neighbors	98.4%	74.5%	86.4%
Decision Tree (C 4.5) (10Fold cross validated)	Feature Vector	85.3%	81.9%	83.5%
	Semantic Similarity measure	90.0%	90.4%	90.2%
	Common neighbors	98.6%	93.4%	**96.09%**
Accuracy (dataset two)				
Logistic binary regression	Feature Vector*	44.6%	59.8%	52.2%
	Semantic Similarity measure	70.3%	94.6%	**82.4%**
	Common neighbors	60.9%	97.9%	79.4%
Discriminant analysis (cross validated)	Feature Vector*	44.5%	59.6%	52.0%
	Semantic Similarity measure	59.6%	98.7%	79.1%
	Common neighbors	60.9%	97.9%	**79.4%**
Decision Tree (C 4.5) (10Fold cross validated)	Feature Vector*	40.9%	71.0%	55.9%
	Semantic Similarity measure	67.9%	96.8%	**82.35%**
	Common neighbors	60.9%	97.9%	79.4%

* Due to missing values, not all cases could be included leading to slightly unequal class distributions.

Table 1: Results of class prediction using different prediction methods (TP = True Positive Rate, TN = True Negative Rate

Again, true positives and true negatives are quite similar distributed. Interestingly, the baseline predictor *common neighbors* outperforms the Semantic Similarity measure in dataset one, especially when using a decision tree classifier. As it is vital to investigate the nature of the relationship of the predictor variable, in particular the elements of the semantic similarity, a discriminant analysis was conducted. Discriminant analysis is used to examine how to best possible separate a set of groups (here *L* vs. *NL*) using several predictors.

Relational similarity *RS* showed with a structre matrix value of .634, that it is the most important predictor for dataset one in differentiating the two groups, followed by the attribute similarity *AS* with a value of .385. Again, taxonomy similarity *TS* seems to have a low contribution to the group differentiating with a value of -.088. As expected, the prediction results of the discriminant analysis are comparable to the binary logistic regression, since both methods are different ways of achieving the same result. It should be noted that within the classification of the feature vector dataset not all instances could be included as the dataset suffered from missing values in several variables. That is also why a decision tree classifier (C 4.5, [Quinlan 1993]) has been tested on both datasets. This classification method shows a the highest predictive accuracy for the common neighborhood metric in dataset one, compared to the other metrics, whereas the logistic regression on the Semantic Similarity metrics leads the highest accuracy in dataset two, as seen in Table 1. In summary, the metrics lead to quite different results in the two datasets, emphasizing the influence of the underlying data. Clearly, the Semantic Similarity seems to benefits from fuller information, i.e. incorporation of background knowledge, which leads to an advantage in prediction accuracy, especially in dataset two. Here, the dataset is characterized by a higher proportion of relational information.

4 Conclusions and future work

In this paper we investigated one important challenge for performing data mining and machine learning directly on Semantic Web data. We created a RDF graph from existing data and computed similarities between instances of that graph. We further calculated traditional metrics to compare both approaches. We evaluated both methods performing a classification analysis in the form of a link prediction between any two actors of a social network, which the data comprised. The results show that the similarity measure based on the RDF graph performs note worthily. We conclude that the incorporation of all available background knowledge yields to an improvement of existing methods. Nevertheless, the results depict that the result heavily rely on the data structure. The results of the baseline predictor common neighborhood also show that Semantic Web technologies not necessarily outperform traditional methods. However, the ontologies used in this paper were rather simple and served to examine a new approach in SNA. In particular the results of dataset two indicate that analyses on highly interrelated data might benefit from Semantic Web technologies.

Another advantage is the expandability of Semantic Web data modeling. New ontologies can be easily attached to already existing metadata. Furthermore, the accessibility to free and comprehensive ontologies is growing rapidly. These ontologies could enhance the analysis process. To give one example, a similarity analysis of customers with hobbies in the field of music or movies, could be improved by music[6] and movie[7] ontologies, which move the analysis from treating this information in a way of nominal values to a relational graph, where each instance is interconnected. In a traditional layout, these values would be mainly compared in the way *equal* or *not equal*[8], whereas the use of ontologies may offer an answer to *how related* two values are. Additionally, Semantic Web technologies offer possibilities to automatically derive new data from existing one through inference engines.[9] Since this approach could further improve the predictive power of analyses, it should be considered for future work. Again, we would like to emphasize that the main purpose of the analysis in this paper is not to find the optimal predictor for one specific problem but, to investigate the potential of Semantic Web technologies in the field of machine learning.

Nonetheless, the study revealed several unanswered question and issues in this field. A major disadvantage is the question of interpretability. As seen in the result section, no conclusions can be made, what predictor in the original RDF graph of the Semantic Similarity had the biggest influence on its predictive power. In the case when causal explanations are needed, this inevitably leads to further effort for the analyst. In many areas, however, this is an important issue. Reviewing the analysis in this paper, the results of the semantic similarity measure give almost no explanation what exactly made two players similar, despite the fact that their "relation" to each other had the highest discriminative power. Another issue is the weighted combination of the three components of the similarity measure, proposed by [Maedche et al. 2002, 348]. The results of this study show that it is advisable to treat each component separately. If the problem domain, including the ontology, is more or less static, the optimal weights for a single

[6] http://musicontology.com/
[7] http://www.movieontology.org/
[8] We think most people will agree that „Dirty Dancing", „Flashdance" and „Alien" are non-identical movies, but have different „distances" from each other.
[9] Check http://fowl.sourceforge.net/ or http://jena.sourceforge.net/inference/ for examples.

problem (e.g. classification) could be assessed in an optimization process for a representative sample of metadata, if available. The statistical analyses of the components of the semantic similarity also showed some unexpected results. Further research is necessary to resolve this subject. Moreover, the ontology used in this paper was rather simple in comparison to existing ontologies in other domains. The similarity measure should therefore be evaluated upon complex ontologies and very large datasets to gain a more precise picture. Nonetheless, this paper improved upon past research by showing the potential of Semantic Web technologies as a basis for data mining and machine learning.

Literatur

Barabási, A.-L., Albert, R., "Emergence of Scaling in Random Networks," Science, vol. 286, Oct. 1999, pp. 509-512.

Berendt, B., Hotho ,A., Gerd, S., "Towards Semantic Web Mining", In Proceedings of the First International Semantic Web Conference on The Semantic Web, ISWC '02, Springer-Verlag, 2002, pp. 264-278.

Berners-Lee , T., Hendler , J., Lassila, O., "The Semantic Web (Berners-Lee et. al 2001)," Scientific American, vol. 284, May. 2001, pp. 28-37.

Bisson, G., "Why and How to Define a Similarity Measure for Object Based Representation Systems", In Towards Very Large Knowledge Bases, IOS Press, 1995, pp. 236-246.

Brickley, D., Guha, R., McBride, B., RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, 2004.

Delteil, A., Faron-Zucker, C., Dieng, R., "Learning Ontologies from RDF annotations", In IJCAI 2001 Workshop on Ontology Learning, Proceedings of the Second Workshop on Ontology Learning OL 2001, Volume 38 of CEUR Workshop Proceedings, A. Maedche, S. Staab, C. Nedellec, and E.H. Hovy, eds., CEUR-WS.org, 2001.

Emde , W., Wettschereck, D., "Relational Instance Based Learning", In Machine Learning - Proceedings 13th International Conference on Machine Learning, L. Saitta, ed., Morgan Kaufmann Publishers, 1996, pp. 122-130.

Gloor, P.A., Krauss, J., Nann, S., Fischbach, K., Schoder , D., "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", In Computational Science and Engineering, 2009. CSE '09, 2009, pp. 215-222.

Grimnes, G., Edwards, P., Preece, A., "Instance Based Clustering of Semantic Web Resources", In Proceedings of the 5th European semantic web conference on The semantic web: research and applications ESWC'08, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, eds., Springer Berlin / Heidelberg, 2008, pp. 303-317.

Gruber, T. R., "Toward principles for the design of ontologies used for knowledge sharing", In Guarino, N., Poli, R., Publishers, K. A., Substantial, I. P., and Gruber, T. R., editors, In Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, in press. Substantial revision of paper presented at the International Workshop on Formal Ontology, 1993.

Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006, pp. 36-39.

Huang, Y., Tresp, V., Kriegel, H.-P., "Multivariate prediction for learning in relational graphs", In NIPS 2009 Workshop: Analyzing Networks and Learning With Graphs 2009.

Kautz, H., Selman, B., Shah, M., "Referral Web: combining social networks and collaborative filtering," Commun. ACM, vol. 40, 1997, pp. 63-65.

Klyne, G., Carroll, J. J., McBride, B., RDF Primer. W3C Recommendation, 2004.

Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V., "New perspectives and methods in link prediction", In KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA: ACM, 2010, pp. 243-252.

Lula, P., Paliwoda-Pkekosz, G., "An ontology-based cluster analysis framework," In Proceedings of the first international workshop on Ontology-supported business intelligence held in Karlsruhe, Germany, OBI '08. ACM, 2008.

Maedche, A., Zacharias, V., "Clustering Ontology-Based Metadata in the Semantic Web", In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery PKDD '02, Springer-Verlag, 2002, pp. 348-360.

Pulido, J.R.G., Ruiz, M.A.G., Herrera, Cabello, Legrand, R., E., S., Elliman, D., "Ontology languages for the semantic web: A never completely updated review," Know.-Based Syst., vol. 19, 2006, pp. 489-497.

Quinlan, J.R., C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning), Morgan Kaufmann, 1993.

Smith, M. K., Welty, C., McGuinness, D.L., OWL Web Ontology Language Guide. http://www.w3.org/TR/owl-guide/, W3C Recommendation, 2004.

Watts, D.J., Strogatz, S.H., "Collective dynamics of 'small-world' networks", Nature, vol. 393, Jun. 1998, pp. 440-442.