2012 Fourth International Conference on Intelligent Networking and Collaborative Systems

A Methodology for Ontology Evaluation using Topic Models

Aryya Gangopadhyay*, Matthew Molek†, Yelena Yesha†, Mary Brady‡, and Yaacov Yesha †
*Information Systems, University of Maryland Baltimore County (UMBC), Baltimore, MD 21250
Email: gangopad@umbc.edu
†Computer Science and Electrical Engineering University of Maryland Baltimore County (UMBC),
Baltimore, MD 21250, Email: {molek1,yeyesha,yayesha}@umbc.edu
‡National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, Email: mary.brady@nist.gov

***Abstract*—The purpose of this paper is to describe a methodology for objectively evaluating ontologies. Our approach involves randomly partitioning the elements of an ontology into disjoints training and test set respectively, generating topic models on the training set, and evaluating how well the model fits the test set. We have tested our methodology on the Translational Medicine Ontology and collected extensive experimental results. The results include the average perplexity score for the entire ontology as well as those for individual elements. Since our methodology provides a numeric score for an ontology it can be used to compare ontologies. Furthermore, elements with high perplexity scores might indicate that either these do not fit well with the rest of the ontology, or that the descriptions for these elements are inadequate. Different perplexity scores among sibling elements indicate the need to revise the structure of the ontology.**

***Keywords*-ontology; evaluation; topic models; Latent Dirichlet Allocation (LDA)**

I. Introduction

Ontologies are formal knowledge representation schemes that capture the knowledge of a domain in terms of elements, their properties, descriptions, and inter-relationships. While ontologies play a very important role, it is unclear how to objectively assess the quality of an ontology [1], [2], given that different ontologies can be created from the same underlying knowledge base. While many ontology building tools are available, there is an inherent subjectivity involved in every ontology. The effectiveness of an ontology depends on the context in which it is used. For example, elements of an ontology can be too specific in one context and too general in another. There could be multiple ways to group the elements of an ontology or to design the semantic hierarchies. Hence, no established standards exist to build a sound ontology that can be used for different purposes.

The inherent subjectivity of an ontology makes it hard to come up with an objective measure in terms of evaluation of a given ontology. Subjective evaluation methods include peer reviews and user ratings. However, these methods do not provide specific metrics for evaluating ontologies. Automatic approaches for ontology evaluation include formal logic-based methods that check for logical consistencies, natural language processing (NLP) methods, and alignment or mapping of ontologies. Formal logic based methods have not been adopted widely, NLP methods are primarily geared towards information retrieval and extraction, and ontology mapping is mainly used for data integration and fusion.

The purpose in this paper is to describe a methodology that can provide an objective measure of a given ontology. We propose to measure the quality of knowledge capture and representation of an ontology. Our goal is to provide an overall score of a given ontology and also identify the discrepancies in its semantic structure by providing quantifiable metrics for individual elements.

The rest of the paper is organized as follows. We discuss the related work in this area in Section II. Our proposed methodology is presented in Section III. We discuss the results of our experiments in Section IV, and in Section V we discuss our conclusions and future plans.

II. Related Work

In this section we provide a brief overview of methods for ontology evaluation. More detailed discussions can be found in [3]. Here we focus only on objective evaluation methods. The majority of automated ontology evaluation approaches fall into one of the following broad categories [4]:

- **Gold Standard Evaluation Methods:** these methods compare a candidate ontology to a "gold standard" ontology, which has been deemed to be an ideal ontology for the given context. Examples include [5] and [6]. A methodology for gold standard based evaluation of ontologies is described in [7]. The major problem with this approach is that any ontology chosen to be the "gold standard" may be deemed to be an arbitrary choice.
- **Application driven Approaches:** these evaluation methods use a candidate ontology to complete a task, the results of which are used to determine the quality of the ontology. Examples of work using this approach include [8], [9] and [10]. These methods are context specific and do not provide a way to measure the effectiveness of an ontology for different purposes.
- **Data driven Approaches:** these evaluation methods compare an ontology to an external source of data, often a related text corpus [11], [9]. When measuring how

978-0-7695-4808-1/12 $26.00 © 2012 IEEE
DOI 10.1109/iNCoS.2012.42

well an ontology covers a corpus, a data driven method may be called a "corpus coverage" evaluation[12].
Existing methods of measuring the fitness of an ontology to a text corpus include [11] and [9]. [11] used statistical text mining measures, comparing the words in the ontology triples with the words of the corpus to obtain scores for coverage, accuracy, precision, and recall. A drawback of their method is that it "does not deal with an actual conceptualization, but rather with its representation or lexicalization in a text, meaning that we cannot directly access the conceptualization (meaning level)" [11]. In addition, no consideration is given to the structure of the ontology. As a result, much of the meaning of the ontology and the corpus are lost before any comparisons are made.
A new method is proposed in [9] that measures the "structural fit" of the ontology with the corpus, which is accomplished by extracting clusters of important terms from the corpus using Latent Semantic Analysis, extending the clusters with hypernyms from WordNet, correlating those terms to ontology elements, and finally measuring the fit with a probabilistic measure of whether elements from the same clusters were close to each other in the ontology. The problem with this approach is finding the best collection of texts that would be representative of the context or tasks for which the ontology is designed.

- **Formal Methods:** one of the best known methods in this category is *OntoClean* [13], [14]. This approach is based on general ontological notions such as *essence*, *identity*, and *unity* to characterize the intended meaning of the properties, classes, and relations in an ontology. These are represented by formal metaproperties that impose constraints on the structure of an ontology. This evaluation method analyzes constraints and cleans the existing ontologies to make them more rigorous.

III. Methodology

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a mixed membership model that has been used to discover hidden themes or "topics" in a document corpus. In the context of a corpus of text documents, a topic model captures the underlying themes or topics that exhibit themselves in different proportions in the documents. The topics themselves are distributions over words or terms that appear in the corpus. Given that the only observable parameters are the words that appear in the documents, the challenge is to estimate the hidden parameters such as the word distributions in topics, the topic proportions in documents, and the word assignments to generate the documents. The LDA is a mixed membership model that generates each document in a corpus as a bag of words given the hidden parameters. The challenge is to estimate the hidden parameters given the observable data. Several methods have been proposed for the parameter estimation of the LDA, and we follow the mean variational methods [15] in this work. The LDA model estimation and inference were done using the lda-c implementation [16].

B. Method

Since an ontology captures the thematic structure of a domain being modeled, its structure should correspond to the descriptions, relationships, definitions, and comments included in the ontology . Our methodology for evaluating ontologies is based on probabilistic topic models [17], [18], [19]. In particular, we demonstrate how the Latent Dirichlet Allocation (LDA) model can be used to evaluate ontologies. The ontology should be structured in such a way that semantically proximal elements should be located in neighboring branches of the ontology. Furthermore, since the ontology is a knowledge presentation scheme of an underlying domain, it should exhibit an overall level of semantic congruence.

Our proposed methodology provides objective measures for both the overall quality of the ontology and the structural proximity of its elements that are semantically similar. In our methodology the elements of an ontology are analogous to the documents in a corpus, the textual annotation of each element, along with its other properties such as relationships with other elements, synonyms, etc. are the terms that capture the semantic essence of each element. The first step in our methodology is to generate the *term-element* matrix for the ontology, which we describe in Section III-C.

The next step in our methodology is to randomly partition ontology elements into two disjoint groups: training and test. The training group is used to generate topic models. The number of topics is normally a user-defined input parameter. Here we have used a varying number of topics as described in Section IV. This step is adapted from [20].

Next the topic model is tested for its predictive performance on the elements held out in the test set. The prediction is measured in terms of the maximum likelihood of test elements. Thus, the maximum likelihood score is computed for every element in the test set. We convert the maximum likelihood scores into *perplexity scores* [18]. The perplexity scores monotonically decrease with the likelihood scores. Lower perplexity scores indicates better performance. More formally the perplexity of an element e is calculated as follows:

$$p(e) = \exp^{-\frac{1}{n}\sum_{i=1}^{n} \ln q(w_i)} \quad (1)$$

where $q(w_i)$ is the probability that the term w_i will be used to describe element e, and n is the total number of terms associated with the *metadata* for the element, which is explained in Section III-C. The overall perplexity score for the test set is the average of the perplexity scores of the individual elements.

C. Term-Element Matrix Generation

Given an ontology, O, we construct a bag of words for each element in O. These bags of words are constructed from a user defined set of features for each element, typically the local name, label, and comments of the element. If there are other useful properties, such as description or definition properties, those should be included as well. The idea is to capture all words that describe the ontology element, especially comments and definitions which are rich in text. We refer to this as the *metadata* corresponding to the element.

In addition, each element's bag of words is affected by related elements in O. All words associated with direct superclasses and subclasses are included in each element's bag of words. The same goes for words associated with elements stated to be equivalent to the current element.

In the approach described in [20], if two classes are stated to be disjoint, all shared words from those elements are removed from their respective bags of words. We do not include this rule, as disjoint classes are often very similar. For example if there are two classes *Man* and *Woman*, it would make sense to declare that they are disjoint. However, we do not want to remove all of the shared words that describe them as adult humans. The greater the number of meaningful words that can be extracted for each ontology element, the better the resulting LDA topic model will represent the meaning of O. The goal of the work presented in [20] is mapping elements of ontologies, while the goal of our work is evaluation of ontologies.

Once the words for all elements of O have been extracted, stopwords are removed, a list of distinct terms is created, and a term-element matrix is constructed. Each row in the matrix represents one element of O, and each column is one of the words extracted from O. To be specific, row i, column j, of the matrix stores the number of times word j appeared in element i's bag of words.

We have implemented a Java program using the Protégé-OWL API [21] that automates the process of building a term-element matrix from an RDF/OWL ontology.

IV. Experimental Results

A. Platform

Our methodology is empirically tested as follows. All experiments are run on a Redhat Linux version 6.2 server with four processors, each with 12 cores, with 1/2 terabyte of RAM, and 14 terabyte of hard disk space. In the first step step we randomly split an ontology into two disjoint parts for training and testing respectively. The training set contained 67% of the data and the test set contained the rest 33% of the data. This is equivalent to 3-fold cross validation. Next, topic models are created with the training data for a varying number of topics ranging from 5 to 200. This process is repeated three times with three different training and test sets. The topic models created in the previous step are then used to assess the likelihood scores for the instances in the test set, resulting in likelihood scores for each of the elements in the holdout sample (test set). Following the literature on probabilistic topic models we convert the likelihood scores into perplexity, which is defined as the probability that the instances in the test data can be generated using the training data. Lower perplexity scores indicate higher probabilities. The results show consistent perplexity scores across the number of topics. In addition, the perplexity scores of neighboring elements such as siblings are compared to identify structural anomalies suggesting possible redesign of certain parts of the ontology.

B. Data

We used the Translational Medicine Ontology (TMO) created by the National Institutes of Health in our experiments [22]. TMO serves as a global schema for data integration as well as a facilitator for queries that span across multiple heterogeneous databases, and thus provides a platform for managing information on personalized medicine. TMO was built from the lexical analysis of topics that are of interest to 16 different user categories including biologists, immunologists, systems physiologists, primary care physicals, health plan providers etc. that form a diverse but overlapping list of entities ranging from molecules proteins, and cell lines to roles such as active ingredients to processes such as diagnosis, study, and intervention to informational entities including dosages, signs and symptoms. These diverse entities are mapped to 75 classes, 223 class equivalence mappings to 201 target classes from 40 ontologies.

C. Analysis

Our experimental results identify several characteristics of the ontology. We present our analysis by discussing the ontology in terms of its overall score, scores in each of the cross validations, identification of elements with moderately high perplexity scores, and extreme anomalies. For all experimental results presented in Figures 1-5, the x-axis denotes the number of topics in the topic model while the y-axis shows the perplexity scores.

1) Overall Score for the Ontology: We took the average perplexity scores over the three cross validations for all elements. The results are shown in Figure 1. The average perplexity scores varied between 2.7 to 4.1, with a standard deviation of 0.46, over the number of topics ranging from 5 to 200. As shown in Figure 1 the overall perplexity scores are fairly stable over different number of topics with a slight dip for the 200 topic model. The overall score can be used as a global measure of an ontology and used to compare the "goodness" of ontologies. Two extreme anomalies, found in the third cross validation, were eliminated in order to avoid bias in the average perplexity scores. We discuss the anomalies below.

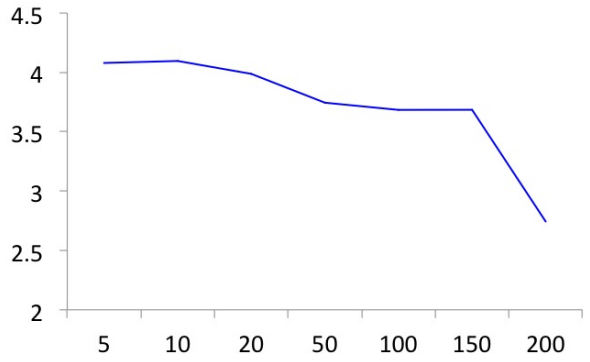

Figure 1. Overall Perplexity Scores

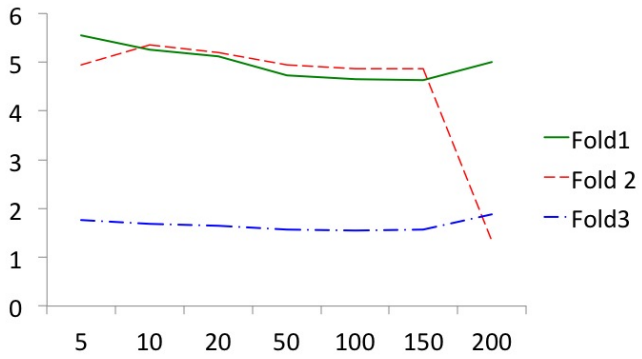

Figure 2. Perplexity Scores over Three Folds

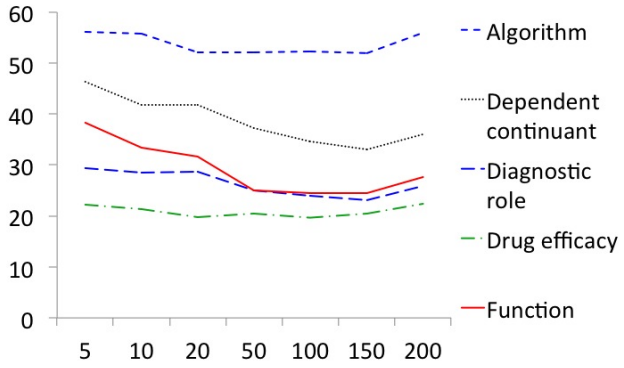

Figure 3. Elements with Moderately High Perplexity Scores (fold 1)

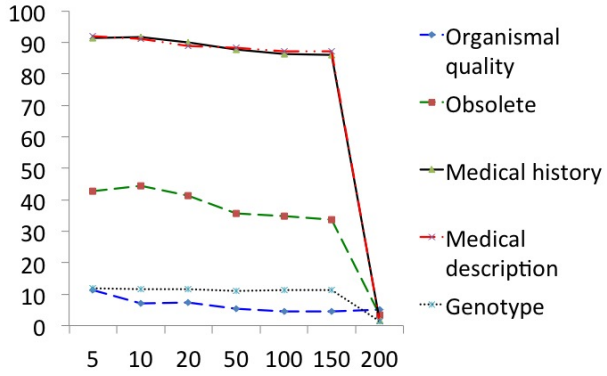

Figure 4. Elements with Moderately High Perplexity Scores (fold 2)

2) Perplexity Scores Across Folds: The average perplexity scores for each of the three folds in our cross validation are shown in Figure 2. Elements in folds 1 and 2 have higher perplexity scores than in fold three with the average scores for folds 1, 2, and 3 being 5, 4.5, and 1.7 respectively. We should note that the anomalies are removed only for fold 3 and not for folds 1 and 2, since the anomalies in folds 1 and 2 are moderate while those in fold 3 are extreme. We further note that the perplexity scores are uniform in folds 1 and 3 with the standard deviations being 0.3 and 0.12 respectively, whereas the standard deviation of the perplexity scores in fold 2 is 1.4. This indicates that elements in fold 3 are semantically closer to the rest of the ontology as compared to those in folds 1 and 2. Furthermore there are elements in fold 2 that have high and low perplexity scores which might need further investigation in the ontological structure for these elements. In addition, Figure 2 indicates a substantial drop in the perplexity scores of the elements when the topic size is 200. For varying number of topics the scores are fairly stable, which might indicate that the model is over-fitting the elements in fold 2 when the topic size is 200.

3) Moderately Anomalous Elements: Our methodology identified 10 elements that are moderately anomalous in the sense that their perplexity scores are significantly higher than the average in their groups. Five of these elements were found in each of fold 1 and fold 2. The perplexity scores for these elements are shown in Figures 3 and 4 respectively. As shown in Figure 3 the element "Algorithm" has the maximum perplexity score with the average of 56.1 over the seven topic ranges. The standard deviation of the perplexity scores are small across topic numbers indicating structural inconsistencies in the ontology in terms

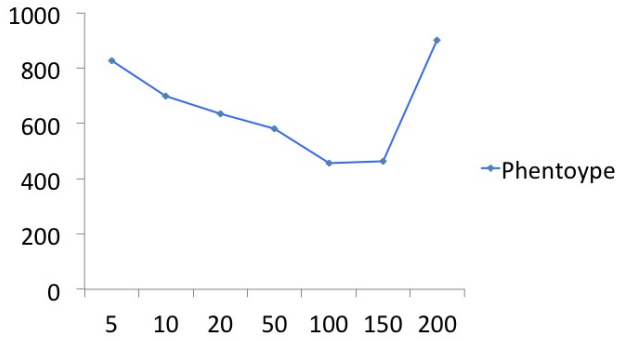

Figure 5. Element with Very High Perplexity Score

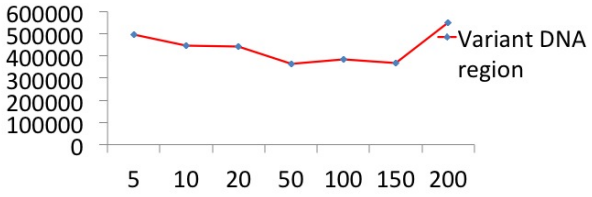

Figure 6. Element with Extremely High Perplexity Score

of nodes in the ontology corresponding to these elements *vis-a-vis* their corresponding descriptions. The anomalous elements in fold 2 are more varied that those in fold 1.Two elements: "Medical History" and "Medical description" have the highest perplexity scores in fold 2, whereas "'Organismal quality" and "Geonotype" have considerable smaller perplexity scores. The standard deviation across number of topics are fairly small indicating the same patterns across topic ranges.

4) Extremely Anomalous Elements: Two elements were found to be "extremely anomalous", indicating that these elements do not fit well with the rest of the ontology. Both of these are found in fold 3. The most extreme anomaly is found in element "Variant DNA region", whose average perplexity score is several orders of magnitude higher than any of the other elements in the ontology. The description of this element in the ontology should be revised. The other extremely anomaly occurred in the element "Phenotype", whose perplexity score is also significantly higher than that of other elements. While the description of these elements may be adequate by themselves, they are substantially different from those of their neighboring elements.

V. Conclusion

In this paper we have described a method for objectively evaluating ontologies. The method consists of creating topic models from the elements and comparing the "fitness" of the topic model by generating a list of elements in the test set. Our method can be used to compare ontologies, identifying elements that are either not a good fit for the ontology or are not located appropriately in the taxonomic structure of the ontology.

One limitation of this method is that it uses the descriptions of the elements to evaluate the quality of the ontology. In some ontologies element descriptions are either too brief or non-existent. In such cases this method can still work by using other sources such as Wikipedia or existing literature that provide reliable descriptions of the elements.

Our future work includes experimenting with additional ontologies and comparing distance measures between elements in an ontology with those in our proposed method. We plan to compare our methodology with other existing work in this area. We intend to suggest ways to "improve" an existing ontology by either changing the descriptions of some of the elements and/or restructuring the elements in the ontology.

Acknowledgment

The authors would like to thank NIST for funding this work. We would also like to thank the anonymous reviewers for their comments on the paper.

References

[1] S. L. Tomassen, "Conceptual ontology enrichment for web information retrieval," PhD, NTNU, 2011.

[2] P. Cimiano, *Ontology Learning and Population from Text - Algorithms, Evaluation and Applications*. Springer, 2006.

[3] D. Vrandecic, "Ontology evaluation," Ph.D. dissertation, KIT, 2010.

[4] J. Brank, M. Grobelnik, and D. Mladenić, "A survey of ontology evaluation techniques," in *Proceedings of the Conference on Data Mining and Data Warehouses SiKDD 2005*, no. a. Citeseer, 2005.

[5] A. Maedche and S. Staab, "Measuring similarity between ontologies," *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pp. 15–21, 2002.

[6] J. Brank, D. Mladenic, and M. Grobelnik, "Gold standard based ontology evaluation using instance assignment," in *Proc. of the EON 2006 Workshop*, 2006.

[7] K. Dellschaft and S. Staab, "On how to perform a gold standard based evaluation of ontology learning," in *Proceedings of the 5th International Semantic Web Conference (ISWC)*, I. C. et al., Ed. Springer Verlag, 2006, pp. 228–241.

[8] R. Porzel and R. Malaka, "A task-based approach for ontology evaluation," in *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer, 2004.

[9] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data driven ontology evaluation," in *Proceedings of Internation Conference on Language Resources and Evaluation*, 2004.

[10] M. Sabou, J. Gracia, S. Angeletou, M. d'Aquin, and E. Motta, "Evaluating the semantic web: A task-based approach," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*. Springer-Verlag, 2007, pp. 423–437.

[11] P. Spyns and M. Reinberger, "Lexically evaluating ontology triples generated automatically from texts," *The Semantic Web: Research and Applications*, pp. 85–97, 2005.

[12] M. Sabou, V. Lopez, E. Motta, and V. Uren, "Ontology selection: Ontology evaluation on the real semantic web," in *Proceedings of the EON'2006 Workshop, "Evaluation of Ontologies on the Web"*, 2006.

[13] N. Guarino and C. Welty, "Evaluating ontological decisions with ontoclean," *Commun. ACM*, vol. 45, no. 2, pp. 61–65, Feb. 2002. [Online]. Available: http://doi.acm.org/10.1145/503124.503150

[14] H. Alani, S. Harris, and B. O'Neil, "Winnowing ontologies based on application use," in *ESWC*, 2006, pp. 185–199.

[15] A. Cherkaev, *Variational Methods for Structural Optimization*, ser. Applied mathematical sciences, 2000, vol. 140. [Online]. Available: http://www.math.utah.edu/book/vmso

[16] http://www.cs.princeton.edu/~blei/lda-c/index.html/.

[17] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *JMLR*, vol. 9, pp. 1981–2014, 2008.

[18] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[19] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, 2011. [Online]. Available: http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

[20] V. Spiliopoulos, G. Vouros, and V. Karkaletsis, "Mapping ontologies elements using features in a latent space," in *Web Intelligence, IEEE/WIC/ACM International Conference on*. IEEE, 2007, pp. 457–460.

[21] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102889/.

[22] J. S. e. a. Luciano, "The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside," *Journal of Biomed Semantics*, vol. 2(Suppl 2), 2011.